CITY UNIVERSITY

CLUSTERING INFORMATION RETRIEVAL SEARCH OUTPUTS

S. Yasemin Kural

September 1999

Submitted for fulfilment of the requirements for the degree of
PhD in Information Science

I wish to thank Steve Robertson and Susan Jones for their ideas, support and understanding. It is greatly thanks to them that I enjoyed and benefited from the time I spent at City.

I also wish to thank City University for granting me a PhD scholarship.

Abstract

Users are known to have difficulties in dealing with information retrieval search outputs especially if the outputs are above a certain size. It has been argued by several researchers that search output clustering can help users in their interaction with IR systems. Clustering may provide users an overview of the output by exploiting the topicality information that resides in the output but has not been used in the retrieval stage. It can enable them to find the relevant documents more easily and also help them to form an understanding of the different facets of the query that have been provided for their inspection. This project aimed to investigate the viability of using clustering as a way of mediating users' interaction with search outputs and attempted to identify its possible benefits.

Can&Ozkarahan's(90) C^3M algorithm was used to test the effectiveness of clustering as a way of search output presentation. C^3M is a relatively simple, non-hierarchical method that has been shown to give compatible or superior results to best-known hierarchical methods.

The method was implemented in TCL and linked to the department's experimental IR system Okapi. Implementation included a procedure of term selection for document representation which preceded the clustering process and a procedure involving cluster representation for users' viewing following the clustering process. After some tuning of the implementation parameters for the databases used, several experiments were designed and conducted to assess whether clusters could group documents in useful ways.

One group of experiments aimed to assess the ability of the implementation to bring together topically related documents. It was quite difficult to gather data for such an assessment, but the existence of a set of data generated for TREC Interactive track(1996) enabled us to design experiments that at least approximately satisfied our objective. TREC provided a set of queries, and groups of relevant documents with facet assignments made by expert users. It was thus possible to make an inference by measuring the correlation between the clusters relevant documents were assigned to and the facet assignments made for the documents by TREC experts.

The utility of this data set was limited for various reasons discussed in the related chapters, however, it can be concluded that clusters cannot be relied on to bring together relevant documents assigned to a certain facet. While there was some correlation between the cluster and facet assignments of the documents when the clustering was done only on relevant documents, no correlation could be found when the clustering was based on results of queries defined by City participants to the Interactive track.

Another group of experiments was conducted to compare output clustering with relevance ranking as a search output representation method. This comparison was necessary as an immediate consequence of clustering search output would be the loss of relevance ranking. It had to be assessed whether clustering could help users to find the relevant documents more easily than by relevance ranking, before any clustering solution could be proposed as an alternative to relevance ranked output.

For this purpose, two sets of user experiments(n=20 and n=57) were conducted based on the users' own information needs. While changes have been made to the implementation between the first and the second set of experiments, the experimental design was almost the same in both runs. Users were first asked to rank clusters formed from the search output(top 50 documents) and then make relevance judgements for the individual documents for the same output. The precision of cluster(s) marked best by the users were then compared to precision values that would be attained by relevance ranking at comparable thresholds.

The results from the 1st group of user experiments were not conclusive(in some part due to the smallness of the data set), but they drew our attention to the importance of representation of clusters and documents for users' viewing. After some changes to the implementation, mainly related to representation issues, and an intermediate set of 10 experiments to assess two new representation formats, a set of 57 user experiments were conducted to measure and compare precision values attainable by clustering versus relevance ranking.

These experiments revealed no significant precision difference between clustered outputs and ranked lists. The number of cases where one method achieved better than the other was slightly higher for the ranked lists at the top cluster level and slightly higher for the clustered representation at the top two clusters level. However the overall average precision values were higher for the ranked list at both levels.

As such, clustering did not appear to be preferable to ranked lists especially as it also represented overheads in both computing time and resources involved in creation of the clusters, and the time and effort taken by the users to inspect them.

An interesting outcome of the user experiments was the ability of the users to identify clusters that do not include relevant information. There were less relevant documents among the clusters marked last by the users as compared to the documents ranked last at similar threshold levels. This brought out the possibility of using clusters as an exclusion tool to improve the precision of ranked lists. After exclusion of documents from the last cluster, ranked lists performed significantly better than the clusters at the top cluster level.

There was also some evidence (consisting of observation of users during the experiments and a few user comments) that clusters could be used to provide the users with a glimpse of the search results, in order to decide whether to inspect the search results or initiate a new query straight away.

In summary, cumulative experiment results imply that clustering cannot outperform relevance ranking, and seems to deserve only a secondary role in users' interaction with IR systems. However, it should also be noted that the experiment results are not representative of the whole set of possible user types and search situations and it may be possible to identify search situations where clustering can be more beneficial than relevance ranking.

Table of Contents

List of Tables

List of Graphs

1. Introduction

1.1 Problem statement

A user's interaction with the search output is often far from optimal. Especially when the output exceeds a certain threshold, users are inclined to sample just a few documents or abandon the query altogether. If the few sampled documents are not found relevant, it is very likely that the rest of the documents will not be inspected at all.

Various researchers have emphasised the difficulties experienced by users in dealing with large output sets (Tibbo(94); Ashford(95)). Wiberly&Dougherty(88) have found that users, intermediaries and search services employ output size thresholds for deciding whether to inspect an output set or not. The figures are highly variable and most have limited scientific validity as they come from individual sources and small observational studies[1]. On their subsequent research, Wiberley&Dougherty(90) have found a threshold of 15-30 postings for OPAC users. Saracevic et al.'s(90) user studies reported on the same year revealed the 'magic number' to be 150±50.

These findings suggest that users have a subliminal concept of a threshold level even if the figure depends on the individual circumstances. Blair(80) suggests the term 'anticipated futility point' to describe this threshold level. The 'anticipated futility point' has been evidenced to play an important role in users' searching behavior. In their experiments on the discourse between users, intermediary searchers and the system interaction, Saracevic et al.(90) have found the most prominent search style to be reductionism:

> *"...the output size was very much often too large; even without looking at the contents, the tactics were to reduce, reduce, reduce...no user really wants large outputs, no matter how they may profess to 'wanting everything'. Very, very few do. A general trait of human information behaviour seems to be that humans are sensitive to and intolerable of information overload, particularly reducing document overload..."*

Saracevic et al. described the information related to the size of the postings set as 'magnitude feedback' and stated that:

> *"...magnitude feedback involved a far greater proportion of the discourse than the relevance feedback and affected a far greater number of subsequent search tactical decisions and directions. Both the users and intermediaries paid far greater attention*

[1] Some of the figures are: 10-15, 15-35, 50-70, 100±50.

to the issues of sheer magnitude than to the issues of relevance...Manageable magnitude comes first, precision second, recall, if at all a distant third..."

Search output can be reduced by introducing additional search terms, but studies show that a great majority of users do very little or no Boolean searches(e.g. Siegfried et al, 93). Even experienced users may not be willing or able to find the appropriate terms to narrow down the search. Su's(94) finding that recall is more important for users than precision implies that users are not very comfortable with narrowing down searches especially when they do not want losses in recall.

By taking the search output size as a determinant of whether the references will be inspected or not, users seem to confirm Miller's(56) theory of short-term memory. According to Miller, human memory is composed of a short-term memory which has an effective capacity for 7±2 'chunks' of information[2], and a long-term memory of virtually unlimited capacity. The amount of information that we can hold in the 'chunks' depend on our expertise/familiarity of the field. Generally, we cope with the short term capacity constraint by grouping several concepts under a larger category and allowing vivid examples to stand for complex concepts(Morris, 94).

The short term memory constraint has been a focus of the human-computer interaction field which has established some rules of good practice to cater for it. In IR, however, this constraint is largely ignored and the possibility of getting around this problem by imitating human behaviour of "categorising & exemplifying" is left unexploited. The user is left to his own devices to utilize his short term memory and digest a given search output. He is often provided with a ranking of the output according to estimated relevance values and that is, at times, sufficient to get rid of the classification problem. But at other times, such as when one experiences difficulties in expressing his information need, or has a more exploratory approach towards a search output, or when a large number of documents have similar relevance ranks, relevance ranking may not be very helpful.

[2] Also referred as 5±2 elsewhere.

1.2 Basic form of proposed solution

In a 1988 paper, Swanson, in response to Fairthorne's(63) argument on the need to formulate principles of information science, including postulates of impotence(PI), i.e. statements of what could not be done, came up with a number of PIs for information science. The sixth PI is as follows:

> *"Word-occurrence statistics can neither represent meaning or substitute for it. Such data, however, can be used, with occasional success, to signal or point out potentially fruitful areas of text where a human being might then seek meaning or relevance."*

If this postulate is right, we may wish to consider assigning a different function to information retrieval systems. Instead of a fixed focus on 'finding' the relevant documents, we may want them to be more decentralised agents, that primarily seek to stimulate and provide hints to the users to explore the collection and help the user to improve his statement of information need, which is seldom complete or optimally expressed. If users can be presented with topical highlights in a clear and simple fashion, it may all become a hotter-colder game, where the user experiences further control over what he sees and is encouraged to explore, employ his powers of reasoning, and articulate his needs instead of passively reacting to system generated list of documents.

But we cannot just loosen the focus of a retrieval system to achieve this. We do not want the user to wander aimlessly through the collection, but make a series of moves, each of which leads to an increased understanding of the coverage of the topic within the collection, and awareness of the extent to which his information need appears to be in congruence with this coverage. Each move may represent a refinement of his query and also provide the user with some information that is relevant to his query.

One way of achieving this is having a retrieval system to find the documents that are most likely to be relevant, and then analysing the output set to identify different facets represented and presenting the user with representations of these facets instead of individual documents. In the past, various researchers have endavoured to do such a classification for document collections by cluster analysis. But much less effort has been spent on classifying the search output.

This project aimed to investigate the viability of using clustering as a means of improving user interaction with search outputs in information retrieval systems. It was envisaged that the topicality information that resides in the search output could be utilised to show the users some kind of pattern existing in the document set; it could enable them to overview a document set quickly and make judgements on groups of documents simultaneously.

Alternatively, if the coverage of the output was inappropriate for the user's need, the topicality information presented in the cluster representations could give cues to the user that could be helpful in modifying his query.

Clustering the search output could also prove useful in automatic query expansion. Robertson et al.(97a) report that during the TREC experiments, Okapi searchers expressed some dissatisfaction in not being able to retrieve items which reflected different aspects or facets of a particular query. Query expansion was mainly useful for finding more of the same as the relevance feedback information mostly incapacitated diversifications from the documents already judged as relevant. If clustering could achieve a helpful categorisation of the documents, users could be able to base query expansion on certain clusters that seem more likely to lead to aspects so far unrepresented among the retrieved documents.

Can&Ozkarahan's(90) C^3M algorithm was used to test the effectiveness of clustering as a way of search output presentation. C^3M is a relatively simple, non-hierarchical method that has been shown to give compatible or superior results to best-known hierarchical methods.

The method was implemented in TCL and linked to department's experimental IR system Okapi. Implementation included a procedure of term selection for document representation which preceded the clustering process and a procedure involving cluster representation for users' viewing following the clustering process. After some tuning of the implementation parameters for the databases used, several dfferent types of experiments were designed and conducted to assess whether clusters could group documents in useful ways.

One group of experiments aimed to assess the ability of the implementation to bring together topically related documents. These experiments were based on the TREC Interactive track (1996). TREC provided a set of queries, and groups of relevant documents with facet assignments made by expert users. It was thus possible to make an inference by measuring the correlation between the clusters relevant documents were assigned to and the facet assignments made for the documents by TREC experts.

Another group of experiments were conducted to compare output clustering with relevance ranking as a search output representation method. This comparison was necessary as an immediate consequence of clustering search output would be the loss of relevance ranking. It had to be assessed whether clustering could help users to find the relevant documents more easily than by relevance ranking, before any clustering solution could be proposed as an alternative to relevance ranked output.

For this purpose, two sets of user experiments(n=20 and n=57) were conducted based on the users' own information needs. While changes have been made to the implementation between the first and the second set of experiments, the experimental design was almost same in the both runs. Users were first asked to rank clusters formed from the search output(top 50 documents) and then make relevance judgements for the individual documents for the same output. The precision of cluster(s) marked best by the users were then compared to precision values that would be attained by relevance ranking at comparable thresholds.

PART I : BACKGROUND INFORMATION AND LITERATURE REVIEWS

2. Relevance ranking

2.1 Relevance ranking in principle

There are various justifications for ranking documents from a search output. The simplest is that users can only look at documents one at a time, and it's good to try to sort them in a meaningful way rather than presenting them randomly. Actually, ranking is implicated in any act of retrieval:

> *"Often there are only two ranks in the ranking, in which case one may speak of a 'retrieved' set and an 'unretrieved' set." (Cooper, 71).*

The essential justification for ranking is improving retrieval precision. In their study of the effectiveness of ranking algorithms, Noreault et al.(81) have found a 20 per cent improvement in precision of best-ranking algorithms over a random ordering of the output. Ro's(88) evaluative study on the applicability of ranking algorithms to improve the effectiveness of full text retrieval has found that the precision of full-text retrieval improved 2 times at the same level of recall when ranking algorithms were applied. Ranking also enhances the chances of success of the relevance feedback process by bringing the relevant items to the user's attention early in the search (Salton, 83).

Normally, relevance ranking is based on a matching process, and when it comes to justifying the use of a particular process, two different reasons leading to different ranking principles have been given:

1. Relevance is a continuous variable and different texts may satisfy the information need to different degrees,
2. Relevance is a dichotomous variable; but given that a system cannot be certain that any given text will be judged relevant, it can only make a probabilistic statement about relevance.

These two arguments impose different requirements on the matching function to be used, but they have often been confused and used wrongly in justification of chosen matching functions (Robertson, 78). The first argument requires a matching function to be able to treat relevance as a continuous variable. In other words, it requires an assessment of the degree of relevance. The second argument implies that the function should rank the texts in order of their probability of relevance. This probability can be estimated based on the imperfect match between the indexed texts and the stated request.

So far, of the two arguments, only the probability of relevance argument has been utilised formally to determine a matching function (e.g. Sparck Jones&Robertson, 76). The degree of relevance argument has not been used as such, although it is frequently used in justification of chosen methods. This is basically due to the difficulty of formalising the concept into a workable form. The difficulty, according to Robertson(78), arises because the degree of relevance has nothing to do with the request(formal statement of the need) or the IR mechanism. It is about the relationship between the text and the information need, in the judgment of the user. Information need is a psychological state, something whose “structure” cannot be seen(Cooper, 71). Neither the request verbalised by the user, or the query presented to a system are necessarily complete and accurate presentations of the information need.

Cooper makes a useful discrimination to help us conceptualise what can and cannot be reasonably measured about relevance: “logical relevance” versus “utility” of a document. According to him*, logical relevance*, alias “topic-appropriateness”, has to do with whether or not a piece of information is on a subject which has some topical bearing on the information need in question, and *utility* has to do with the ultimate usefulness of the piece of information to the user. But he also acknowledges the relationship between utility and logical relevance by stating that logical relevance is almost the only factor in utility which the designer does know how to deal with very effectively.

Utility depends on a variety of factors, such as the user’s prior knowledge about the information need, particular areas of interest/curiosity, the document’s credibility, source and recency and the documents already inspected/found to be relevant. While it seems quite appropriate to regard usefulness and utility as a matter of degree, it also follows that they are not easy to formulate, as most of these factors cannot be translated to the query and remain solely on the user’s mind.

With all its apparent impracticality, the degree of relevance argument cannot be overlooked. There is a general recognition that relevance should be regarded as a variable of degree rather than a dichotomous property(Robertson, 78), and users confirm this frequently by making comments about documents’ degree of relevance. The assumption that relevance is a dichotomous variable seems to be sustained for the sake of achieving a workable formulation rather than being a well-maintained fact.

This assumption, together with the assumption of term independence that is often made in devising probabilistic relevance estimation procedures, has been an important source of criticism for the probability of relevance argument(Cooper, 94):

> *"The introduction of simplifying assumptions known to be less than universally valid surely compromises to some degree the accuracy of the probability estimates that result. The important question is by how much. The extent to which the accuracy of the estimate is compromised by adopting various assumptions is not yet fully known...[but] in the IR application, it is only the order of the output documents that is crucial, not the retrieval status values that impose the order...Some probabilists find comfort in the hope that there could be considerable inaccuracy in the magnitudes of the probabilities without affecting their order much."*

However, research to date does not seem to help to devise a plausible method that can lead to the implementation of the degree of relevance argument. Cooper points to the lack of competing approaches with well articulated assumptions:

> *"...the computation of the probability estimates proceeds via an unbroken line of deductions whose premises are made clear at every step...This cannot be generally said for competing methods...Though their practitioners usually accept the PRP (Probabilistic Ranking Principle) in principle, they make no attempt to lay out the assumptions by which the probability estimates it calls for are reached. The simple reason for this is that no probability estimates per se are calculated. Rather, what is computed for each document is a 'retrieval coefficient', 'similarity measure', or 'retrieval status value', and the precise relationship of that value to the probability of usefulness is left unexamined. There is an inarticulate leap from the theoretical model to the assumption that the output ranking is what one wants."*

Eventually, it is the effectiveness of various methods, rather than their theoretical models that count most. But it is not easy to achieve a valid comparison of the different retrieval methods. TREC is probably the best platform where different models can be compared for performance, but the multitude of performance factors such as the type of treatment given to the documents, search strategy, and scoring criteria apart from the general form of retrieval model, make it difficult to assess the general implications of the results for particular approaches (Sparck Jones, 95). With all the available evidence, Sparck Jones(97), in her summary performance comparisons, concludes that methods based on various different retrieval models are capable of giving comparable results.

2.2 Use of ranking in different retrieval models

There are various different ways of ranking search outputs. Of the two main natural language retrieval models, the vector space model ranks the documents in descending order of query-document similarity and the probabilistic model ranks them on the probability of relevance.

The case for Boolean models is more complicated. Until recently, a Boolean retrieval system, by definition, would be assumed to have no relevance ranking. This assumption was so established that, Willett(85) described "ranked output" as an alternative name for best match and nearest neighbour searching and Keen(91) assumed ranked output to be a form of non-Boolean retrieval. Opponents of Boolean systems have criticised them generically for absence of relevance ranking (see Radecki, 88; Lu et al., 96). As Turtle(94) points out:

> *"In many experimental studies, Boolean results are assumed to be unordered. In most commercial systems Boolean results are ranked(ordered) using some document characteristic that is query independent (e.g., by date, alphabetically by author or title, or in some classification order)."*

But, Boolean retrieval and relevance ranking do not have to exclude each other although ranking may have a more restricted meaning with the Boolean systems. A Boolean system can rank documents that appear to be more relevant at the top of the list and help users to find the information they need more easily and quickly. However, it cannot (in principle), give precedence to documents that have high probability of relevance, but fail to fulfil some of the Boolean constraints imposed in the query.

There have been various efforts to bring the benefits of ranked output to Boolean systems. Models resulting from these efforts have been variously named as "extended Boolean models" or "soft Boolean models". The implementations based on these models have usually benefited from one or more of the ideas listed below:

1. Relaxing the interpretations of Boolean conditions, but ranking the documents that conform to the conditions fully before those retrieved by a somewhat relaxed set of conditions (Salton, 88).
2. Using document term weights reflecting the importance of individual terms in a document to calculate query-document similarities.
3. Having separate retrieval and ranking stages.

The SIRE (Syracuse Information Retrieval Experiment) is possibly the earliest example for this kind of solution(Noreault et al., 77). It has a two stage processing chain where first a Boolean query formulation is processed in the conventional manner, and the documents retrieved at this stage are further processed by computing a cosine similarity measure between each document and a "flattened" query consisting of all the original query terms connected by Boolean OR operators.

Factors most frequently used in these models are listed as term frequency, inverse document frequency, proximity of terms, document length, terms' position in document, and when available, relevance judgements(e.g. Jacobs, 94; Burnett et al. 96).

It can be said that research on improvements to the conventional Boolean retrieval systems had a low impact on commercial retrieval systems until the 90s. This is, according to Smit&Kochen(88) due to the service providers' lack of knowledge and financial considerations about the costs of changing their existing systems.

During the 90s, online service providers started to show increased interest in incorporation of new techniques into their retrieval systems. This interest was aroused by market growth, development of more cost effective methods for practical implementation and partly the new emphasis on large scale IR experimentation pioneered by TREC which has lifted the providers' concerns about the scalability of the new solutions developed within academia and also proved effective in bringing together research and vendor communities (Koshman& Rasmussen, 94).

Ranking is now a standard feature of state-of-the-art retrieval systems. Almost all of the participants to TREC(Text Retrieval Conference)[3] have used some form of ranking regardless of their searching methods. All leading Web search engines provide some kind of ranking (see Section 2.3 for details).

Some of the leading online database providers also provide ranking facilities. DIALOG, acclaimed to be the world's most comprehensive online information source, giving access to over 450 databases, launched TARGET, a ranking tool that employs partial-match probabilistic retrieval in 1993. For its Web pages DIALOG uses a different search tool, PLWeb[4], that also has a relevance ranking facility. This product was also being used by EBSCO Publishing, Grolier Educational, the Associated Press, Dow Jones & Co., Washington Post Company, and USA Today as of September 1997.

But there are still some database providers who do not provide any ranking facilities. DataStar, Europe's leading online database service, giving access to over 350 databases, has only a crude sorting tool "RANK", that is based on the count of occurrences of unique terms

[3] See section 11.1 for information on TREC conferences.

[4] PLWeb is a product of PLS(Personal Library Software Inc.), a supplier of information retrieval software with features such as relevance ranking, automatic/dynamic concept searching, and distributed

and by no means is an estimate of relevance. BIDS, UK's leading provider of networked information for end-users in the higher education and research communities, and Silverplatter, another internationally acclaimed provider, also lack any mechanism of relevance ranking(as of 1998).

Yet, the trend towards enhancing the features of retrieval systems is apparent and we may expect to see ranked output as a standard feature of retrieval systems in the near future.

2.3 Factors incorporated in relevance ranking

There are many different factors that affect relevance judgements. Various researchers undertook to identify and categorise these factors(e.g Cuadra&Katter, 67; Schamber,94). Shamber's(94) overview brings together 80 factors that have been identified in previous studies. These factors are grouped under the headings:

1. Judges
2. Requests
3. Documents
4. Information system
5. Judgement conditions
6. Choice of scale

Some examples from each category are presented in Table 2-1.

searching. The company states that the most powerful weapon in the PLWeb arsenal is relevance ranking.

Table 2-1 : Factors from Shamber's categories of relevance

Judges	biases, cognitive style, concept of relevance, formal education, intelligence, judging experience, judgement attitude
Requests	diversity of content, difficulty level, functional ambiguity, specificity/ amount of information
Documents	aboutness, accuracy, authorship, credibility, difficulty level, diversity of content, importance, informativeness, novelty, pertinence, publication source, recency, style, subject matter, usefulness, textual attributes
Information system	access, accuracy, browsability, comprehensiveness, convenience of location and hours, cost saving, currency, simplicity, time spent
Judgement conditions	specification of the task, definition of relevance, order of presentation, size of document set, time for judging
Choice of scale	availability of anchors, ease of use, kind of response required, number of rating categories, type of scale

Shamber also reported more recent studies where relevance criteria are elicited from users' own explanations through open-ended questioning(e.g. Schamber(91), Park(92), Barry(93)). Overall, the user-criteria studies yielded criteria similar to those of previous studies, implying the existence of a finite range of criteria for users in all types of information problem situations.

While it is important to acknowledge the multitude of factors affecting relevance decisions, it is hard to envisage an information retrieval system that incorporates many of these factors effectively and without complicating the system-user interaction. Generally information systems focus only on information content and they are supported by the user studies that reveal information content to be the top priority of the users (Schamber, 94; Froehlich,94). But other criteria may also play a critical role. According to Froehlich:

> "*The prototypical core for relevance judgements or the nuclear sense of relevance is topicality...relevance has a nuclear or focal basis around which other criteria gravitate or override...All relevance judgements start with topically relevant materials (which is an appropriate first step of systems), but the diverse criteria come into play operating dynamically in a process in which certain citations are rejected or accepted on one or more criteria. However, there is a collective set that operates for any one user or class of users and the relationship among the criteria that users employ have 'family resemblances'.* "

Generally, information systems try to infer the relevance of information content from a set of quantitative factors. The most commonly used factors are ICF(inverse collection frequency), TF(term frequency), document length, proximity of terms and term position. There seems to be different patterns in the use of various factors between experimental and operational systems. The following tables show the leading factors used in relevance ranking by TREC participants and by Web search engines.

Table 2-2 shows the factors most frequently used by TREC participants in each year of the conference. Figures are based on the number of participants that have provided their system summary and timing information.

Table 2-2: Most frequently used factors in ranking documents in TREC experiments

	Number of orgs.[5]	Term frequency		Inverse collection frequency		Document length		Other term weights[6]		Proximity of terms		Position in document	
		num.	%	num.	%	num.	%	num.	%	num.	%	num.	%
TREC-1	19	14	73	12	63	11	58	6	32	8	42	5	26
TREC-2	28	20	71	16	57	13	46	14	50	5	18	2	7
TREC-3	28	25	89	18	64	18	64	13	46	11	39	2	7
TREC-4	29	27	93	19	66	20	69	13	45	9	31	2	7
TREC-5	29	25	86	22	76	18	62	9	31	9	31	3	10

Table 2-3 shows the factors that four leading Web search engines claim to be using in ranking search outputs (as of September 1997). Another leading search engine Excite declares using the method of Intelligent Concept Extraction(TM)[7] in ranking documents.

[5] Based on number of participating organisations whose system summary and timing information have been published in TREC Proceedings. The number is smaller but quite close to total number of organisations that have participated.

[6] e.g. weights based on available relevance information, query term frequency, within-document term frequency, position of term in topic statement.

[7] This method scores documents based on a correlation of their concepts, as well as actual keywords. It is claimed that the method achieves performances that are comparable to Latent Semantic Indexing, but uses an entirely different technology which is far less demanding in computational power and therefore both rapid and scaleable.

Table 2-3: Factors used by Web search engines for relevance ranking[8]

	ICF	Proximity of terms	% of query terms found in the document	Position of terms in the document	Within document term frequency	Document Length	Generality of matching category in the tree hierarchy
AltaVista	x	x	x	x			
Infoseek	x		x	x			
Yahoo			x	x			x
Lycos Pro[9]	x	x	x	x	x		
Electronic Library		x	x				
HotBot	x			x	x	X	

There is less information available about the factors used by database and software vendors. DIALOG states the factors used by TARGET as TF, ICF, proximity of terms, and query terms found in the document. PLWeb(see Section 2.2) states the factors incorporated in the ranking of its retrieval software as the number of distinct query terms that appear in a document, ICF, TF and document length.

2.4 Relevance ranking vs. clustered representation

By presenting the output as clusters, we will deprive the users of relevance ranking[10], which is itself a very powerful tool of output representation. In order to rationalise this, a comparative evaluation needs to be made.

In the context of this project, it was necessary to establish whether clustering could help users to find the relevant documents more easily than by relevance ranking. But this was not enough. Output clustering represented an overhead both for the system and the user and it was necessary to assess whether benefits, if any, brought about by clustering, could also outweigh the accompanying overhead.

The overhead involved is composed of:

[8] Derived from the help texts of the listed search engines as of September 1997.

[9] Lycos Pro asks users to specify the importance of each factor for their query and customises the ranking according to these weights.

[10] Clusters can still have relevance ranks, in the form of, e.g. average relevance of all documents. But this will mean that a highly relevant document in a cluster of documents with low relevance values will not be noticed, and the averaging process will actually decrease the significance of low and high values.

1. the time and computing resources needed by the system to perform the clustering operation on the retrieved search output, which delays the time the user can be provided with the results,
2. the time users spend trying to assess the cluster representations.

The first overhead can be negligible considering the current processing capabilities of computers. The second overhead is the main point of concern, which actually cannot be described as just an overhead, as it provides the user with some kind of summary information about the output. This information can save the user from inspecting an output, if it implies that the output set may be inappropriate; or it can provide the user with clues to improve his query formulation.

The process of evaluating cluster representations may also be better utilised by allowing the users to mark the representative titles for relevance while they are evaluating the clusters. This may avoid re-evaluating titles/documents that were displayed in the cluster representations and help users with their assessments of the clusters.

3. Cluster analysis

3.1 Introduction

Classification is central to our understanding of any phenomena. The following lines stress its importance:

> *"All the real knowledge which we possess, depends on methods by which we distinguish the similar from dissimilar. The greater number of natural distinctions this method comprehends the clearer becomes our idea of things. The more numerous the objects which employ our attention the more difficult it becomes to form such a method and the more necessary" (*Linnaeus, Genera Plantarum, 1737*).*

Unfortunately, in many cases where classification would improve our understanding, measuring the similarities between the objects and identifying the classes based on these similarities is far from trivial. Many numerical techniques have been devised to assist us with classification in such cases. These techniques have originated largely in the natural sciences such as biology and zoology "*in an effort to rid taxonomy of its traditionally subjective nature and to provide objective and stable classifications*" (Everitt,93).

In parallel with the developments in computer technology, the number of numerical classification techniques and their applications have increased dramatically during the second half of the twentieth century. While various names have been applied to these methods by different disciplines such as numerical taxonomy(biology), Q-Analysis (psychology), and unsupervised pattern recognition(AI), the most common generic term is cluster analysis. Cluster analysis has been employed in a wide range of disciples for the solution of specific classification problems, some examples are psychiatry, medicine, social services, market research, education and archaeology.

3.2 Using cluster analysis

Using cluster analysis is not straightforward. An investigator needs to make decisions based on his data and purpose at every stage of the analysis. This may apply even to the definitions of terms like cluster, group and class. Bonner(64) has suggested that the meaning of these terms should be defined based on value judgment of the user. Later, Cormack(71) and Gordon(81) have defined cluster with properties such as internal cohesion and external isolation.

Main decisions that have to be made by an investigator wishing to use cluster analysis concern the representation of the objects, measures of association, the method to be used and the representation of the clusters. Unfortunately there exists, in general, no sound theoretical basis for most of these decisions. Everitt describes a typical investigation(93):

> *"It is generally impossible a priori to anticipate what combination of variables, similarity measures and clustering techniques are likely to lead to interesting and informative classifications. Consequently the analysis proceeds through several stages with the researcher intervening if necessary to alter variables, choose a different similarity measure, concentrate on a particular subset of individuals etc. The final, extremely important stage, concerns the evaluation of the clustering solution(s) obtained. Are the clusters real or merely artefacts of the algorithms? Do other solutions exist which are better? Can the clusters be given a convincing interpretation?"*

As a rule of thumb, some graphical representation of the data should be obtained prior to application of any method. This is critical to both the understanding of the data and validity of the analysis. According to Jain&Dubes(88):

> *"Cluster analysis is a tool for exploring data and should be supplemented by techniques for visualising data".*

There are various methods for such visualisation (see Everitt, 93). A frequently used method is principal components analysis. This works by transforming original variables to a new set of independent variables each accounting for decreasing portions of variance of the original variables and provides a two-dimensional view of the data. Another method, Andrew's plots, creates a similar transformation by plotting multi-variate data over the range -π and π by a trigonometric function.

The next stage involves deciding whether the raw data or data derived from them should be used as input for the cluster analysis. This depends mainly on the clustering technique to be used and the type of raw data concerned. Some reduction in number of variables is often called for to make the clustering procedure practically feasible.

3.3 Clustering methods

There are many different clustering methods which are neither mutually exclusive nor can be neatly categorised into a few groups. Hierarchical methods are the broadest family to be categorised under one group. Some other categories are optimization methods, clumping methods(creating overlapping clusters), density search techniques and mixture models.

No clustering method can be judged to be 'best' in all circumstances. Particular methods will be best for particular types of data. In many applications, it might be reasonable to apply a number of clustering methods. If all produce very similar solutions the investigator might perhaps have more confidence that the results are worthy of further investigation. Widely different solutions might be taken as evidence against any clear-cut cluster structure.

A non-comprehensive outline of commonly used methods is provided below.

3.3.1 Hierarchical methods

Among various methods, hierarchical clustering methods are the most commonly used ones. A hierarchical classification consists of a series of partitions which may run from a single cluster containing all individuals, to n clusters each containing a single individual. Hierarchical classifications may be represented by a two dimensional diagram known as a dendrogram which illustrates the fusions or divisions made at each successive stage of the analysis.

In the main, hierarchical clustering techniques are divided as agglomerative methods which fuse the individuals in successive steps and divisive methods which successively separate the group of n individuals into finer classes. It is up to the investigator to decide at what point to stop dividing or fusing. The drawback is the possibility of influence from a priori expectations. Both processes are irrevocable in the sense that once two objects are divided they cannot be fused again and vice versa. Divisive methods are far less popular than the agglomerative methods and they will not be detailed here.

As their names suggest, these methods impose a hierarchical structure on the data and it is usually necessary to consider whether this is merited or whether it introduces unacceptable distortions of the original relationships of the individuals, as implied by their observed proximities(Everitt, 93).

Some of the commonly used agglomerative methods are outlined below and their strengths and weaknesses are pointed out. However, it should also be stated that studies comparing the performances of various hierarchical methods have found that performance varied according to input data types used(Cunningham&Ogilvie, 72; Milligan,80).

3.3.1.1 Single link clustering (nearest neighbour technique)

This is one of the simplest agglomerative clustering techniques. It seems to be the archetypal clustering technique as most other techniques are said to be inspired by it. In this method distance between groups is defined as that of the closest pair of individuals; one individual from each group are considered.

The method is deemed to be the best hierarchical method as far as theoretical soundness is concerned(Jardine&Sibson, 71) and there are efficient methods devised for allowing the single linkage method to be applied to large data sets(Sibson, 73; Rohlf 73, 78). With single linkage method, the hierarchy is progressively updated as new similarities become available, regardless of the order in which they are calculated, and there is no need to store an interdocument similarity matrix (Willett, 88). It is also preferred for its invariance under monotonic transformations of the proximity matrix(Krzanowski, 88), making it susceptable to scaling and combining different variables.

However, in many applications, single linkage is the least successful method in producing useful cluster solutions. The weakness of the single link method is its tendency of clustering individuals linked by a series of individuals(chaining effect), this leads to creation of loosely bound clusters with little internal cohesion (El-Hamdouchi&Willett, 89). While this is undesirable for many applications, it is not unreasonable in subjects like taxanomy where evolutionary chain mechanisms are at work (Krzanowski, 88).

Another limitation of this method is that it also lacks any immediate definition of cluster centre or representative (Murtagh, 83).

In Cluster Based Retrieval(CBR), this method has been found to give a poor performance, producing small number of large, ill defined document clusters(Voorhees, 85; Burgin, 95). Griffiths et al.'s (84, 86) experiments on document clustering revealed that complete linkage, group average and Ward's method gave far superior results than those obtained with the single linkage method.

3.3.1.2 Complete linkage clustering (furthest neighbour clustering)

This is the opposite of single linkage clustering in the sense that distance between groups is now defined as that of the most distant pair of individuals, one from each group. The definition of cluster membership is very much stricter than that for single linkage, and the large straggly clusters in the latter case are here replaced by large numbers of small, tightly bound clusters (El-Hamdouchi&Willett, 89), of equal diameter (Krzanowski, 88). In graph theoretical terms, this method corresponds to the identification of the maximally complete subgraphs at some threshold similarity (Willett, 88).

According to Willett(88), complete linkage is probably the most effective method for document clustering, despite also being the method that requires the greatest computational resources.

This method is sensitive to observational errors.

3.3.1.3 Group average clustering

With this method, the distance between two clusters is defined as the average of the distances between all pairs of individuals that are made up of one individual from each group. It represents a mid-point between the two extreme types of linkage methods, i.e. single linkage and complete linkage. This method is known to minimise the distortion imposed on the inter-object similarity matrix when a hierarchic classification is generated (El Hamdouchi&Willett, 89).

Group average has been found to have the best overall performance in a comparative study involving seven hierarchical clustering methods(Cunningham& Ogilvie, 72). In another study, it has been reported that the performance of this technique was affected by outliers(Milligan, 80).

3.3.1.4 Centroid clustering

Unlike the above three methods that operate directly on the proximity matrix and do not need access to the original variable values of the individuals, centroid clustering does require original data.

With this method, groups once formed are represented by their mean values for each variable, that is their mean vector, and inter-group distance is now defined in terms of distance between two such mean vectors. Use of a mean implies that the variables are on an interval scale; the method is however, often used for other types of variables.

A disadvantage of the centroid method is that if the sizes of the two groups to be fused are very different then the centroid of the new group will be very close to that of the larger group and may remain within that group. This may be avoided by assuming that the groups to be fused are of equal size, the apparent position of the new group will then always be between the two groups to be fused.

This method is also biased towards finding 'spherical' clusters.

3.3.1.5 Median clustering

Median clustering is similar to the centroid clustering but it attempts to make the process independent of group size by assuming that the groups to be fused are of equal size; the apparent position of the group will then always be between the two groups to be fused.

Lance and Williams(67) suggest that this method should be regarded as unsuitable for measures such as correlation coefficients, since interpretation in a geometrical sense is not possible. Median clustering also suffers from the chaining effect like the single link clustering method.

3.3.1.6 Ward's method

Ward(63) proposed a clustering procedure seeking to form the partitions in a manner that minimizes the loss associated with each grouping, and to quantify that loss in a form that is readily interpretable. At each step in the analysis, union of every possible pair of clusters is considered and the two clusters whose fusion results in the minimum increase in 'information loss' are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion, ESS.

Ward's method is biased toward finding 'spherical', tightly bound clusters (El-Hamdouchi &Willett, 89) and it is not very good in coping with unequal sample sizes(Kuiper and Fisher, 75). Additionally, it is only defined explicitly when the Euclidian distance is used for the calculation of the interdocument similarities; the use of an association coefficient (e.g. the Dice coefficient) will not result in an exact Ward classification (Willett, 88).

Griffiths et al.(86) have found this method to give better results than single linkage, complete linkage and group average methods in their CBR experiments with small document collections.

3.3.2 Optimization methods

These methods produce a partition of the individuals for a particular number of groups, by either minimizing or maximizing some numerical criterion. They do not necessarily form hierarchical classifications of the data.

The basic idea behind these methods is that associated with each partition of the n individuals into the required number of groups g, is an index, f(n,g), the value of which is indicative of the 'quality' of this particular clustering. Associating a number with each partition allows them to be compared.

Optimization methods suffer from several shortcomings. Blashfield(76) finds that the result from an optimization method can, in some cases, be radically affected by the choice of the starting partition, although with well structured data it is reasonable to expect convergence to the same. They may also impose a 'spherical' structure on the clusters found even when the 'natural' clusters in the data are of other shapes. Finally, selecting the number of groups is a considerable issue with optimization methods. Several methods have been suggested to aid the users in this respect (Beale, 69; Marriott, 71; Calinski and Harabasz, 74).

3.3.3 Mixture models

These models attempt to provide means for deriving inferences from sample to population. The most common approach is mixture distributions.

The mixture approach provides a firmer statistical basis than other methods, and involves no decision over what particular similarity or distance measure is appropriate for a data set. Nevertheless mixture models have their own set of assumptions such as normality and conditional independence, which may not be realistic in all applications.

3.3.4 Density search clustering techniques

These methods search for regions of high density in the data, each such region generally being taken to signify a different group. Some of these techniques have their origins in single link clustering but attempt to overcome chaining. Two example methods are briefly described below.

3.3.4.1 Taxmap method

This method attempts to detect clusters by comparing relative distances between points and searching for continuous, relatively densely populated regions of the space, surrounded by continuous relatively empty spaces. Clusters are formed initially in a way similar to that for the single linkage model, but criteria are adopted for judging when additions to the clusters should be stopped.

3.3.4.2 Mode analysis

This is also a derivative of single linkage clustering which searches for natural sub-groupings of the data by seeking disjoint density surfaces on the sample distribution. The search is made by considering a sphere of some radius, R, surrounding each point and counting the number of points falling in the sphere. Individuals are then labelled as dense or non-dense depending on whether their spheres contain more or less points than the value of the linkage parameter, K, which is preset at a value dependent on the number of individuals in the data set.

A difficulty with mode analysis is its failure to identify both large and small clusters simultaneously.

3.3.5 Clumping techniques

Clumping techniques allow overlapping clusters. They often begin with the calculation of a similarity matrix, followed by the division of the data into two groups by minimizing what is known as a cohesion function.

Algorithms to minimize these functions proceed by successive reallocations of single individuals from an initial randomly chosen cluster center(Bonner, 64). By iterating from different starting points many divisions into two groups may be found. In each case members of the smaller group are noted and constitute a class to be set aside for further examination.

3.4 Measurement of association

All clustering methods need to measure the association between the objects in some way. This can be done using various similarity, dissimilarity and distance measures.

A similarity coefficient indicates the strength of the relationship between two objects given the values of a set of p variates common to both. Most similarity measures are non-negative and are scaled so as to have an upper limit of unity, although some are of a correlational nature so that the coefficient varies between -1 and 1.

Dissimilarity measures are a complement of the similarity measures. There are also several dissimilarity coefficients representing a kind of 'distance' function. These generally satisfy the conditions:

$D(X,Y) \geq 0$ for all X, Y $\in$ P

$D(X,X) = 0$ for all X $\in$ P

$D(X,Y) = D(Y,X)$ for all X,Y $\in$ P

$D(X,Y) \leq D(X,Z) + D(Y,Z)$ for all X,Y $\in$ P

The fourth condition is the triangle inequality. Triangle inequality is the theorem from Euclidian geometry which states that the sum of the lengths of two sides of a triangle is always greater than the length of the third side.

An example of the dissimilarity coefficient satisfying the above four conditions is

$\frac{|X \nabla Y|}{|X| + |Y|}$ where $(X \nabla Y) = (X \cup Y) - (X \cap Y)$ is the symmetric difference of sets X and Y

It is simply related to Dice's coefficient by

$$1 - \frac{2|X \cap Y|}{|X| + |Y|} = \frac{|X \nabla Y|}{|X| + |Y|}$$

and is monotone to Jaccard's coefficient subtracted from 1.

In terms of the method of association, it has been argued that differences in output among different measures are insignificant, providing that these are appropriately normalised. Lerman(70) found that many of the frequently used measures were monotone with respect to each other. Sneath and Sokal(73) agree with Lerman and suggest that the simplest type of coefficient that seems appropriate should be used. Yet some experimental results reveal that coefficients used may affect the outcome(e.g. Willett, 83; Kirriemuir&Willett, 95).

3.5 Assessing clustering solutions

By organizing multi-variate data into subgroups or clusters, clustering may help the investigator discover the characteristics of any structure or pattern present. Applying the methods in practice however, requires considerable care if over-interpretation of the solutions obtained is to be avoided. Much attention needs to be given to questions of cluster validity, although such questions are rarely straightforward and are “full of traps for the unwary”(Dubes&Jain, 79).

Interpreting the results from a clustering algorithm is often dominated by personal intuition and insight. The problem is that clustering algorithms generate clusters even when applied to random data, and it clearly becomes necessary to try to guard against elaborate interpretation of the solutions found in such cases.

4. Use of cluster analysis in IR

4.1 Cluster hypothesis

In the context of IR, a cluster indicates a group of documents that are more closely associated with each other than with those in different groups.

Within the same context, cluster hypothesis states that "closely associated documents tend to be relevant to the same request"(Jardine & van Rijsbergen, 71). If the hypothesis is true, cluster analysis can usefully be employed for automatic classification of documents and clusters of similar documents can be treated as a single unit for retrieval purposes.

Although automatic classification of documents has been suggested much earlier by Fairthorne(47) and Good(58) and hierarchical document classification methods have been employed by Doyle(65), Broffit et al.(66) and Litofsky(69) before the cluster hypothesis was put forward, it was not until Jardine & van Rijsbergen's 1971 paper that CBR came to be thought as a rival to best match searches in both efficiency and effectiveness.

Jardine and van Rijsbergen argued that hierarchically clustered files have the potential to give improved retrieval performance when compared with searches of unclustered files. Using the Cranfield collection, they computed the distances between pairs of relevant documents and, alternatively, between pairs of relevant and non-relevant documents and concluded that the cluster hypothesis holds[11]. They also noted that cluster-based retrieval strategies showed greater effectiveness the greater the relative importance attached to precision, and the effectiveness for recall-based queries could be improved by retrieving more than one cluster.

Later, Voorhees(85) criticised this test on the grounds that distances between relevant and very distant (and very numerous) nonrelevant documents distorts the distribution and argued that it was more important to look at pairs of similarly described documents and determine if they tend to be relevant to the same request. She based her tests on the proportion of nearest neighbours to a document that is relevant. She found that the cluster hypothesis was true for only one of the four test collections she examined. According to Gordon(91), her results could be even worse if her tests had not overlooked the prior probability that two documents will be co-relevant. Combining the results of the two experiments, Gordon(91) concluded that document similarity is a necessary but not sufficient condition of co-relevance.

[11] Griffiths et al(86) have later calculated the overlap between R-R and R-NR distributions for 7 test collections. Values ranged between .43 for the Cranfield collection which has been reported to have the best separation leading to highest effectiveness values for CBR and .83 for the UKCIS collection.

There have been numerous experimental studies comparing the performance of cluster-based searches with non-cluster searches. These studies, some of which are exemplified below, gave mixed results and in time led to decline in enthusiasm for the cluster hypothesis.

Although experiments to date do not allow a consensus on the validity of the cluster hypothesis, there seems to be an agreement that the performance of clustering depends a great deal on the collection and also the decisions on document representation, measures of association and clustering method. The Cranfield collection, for example, gives good results even when some descriptor terms are randomly deleted; Inspec and Keen collections, conversely, do not give encouraging results even when the initial descriptions are modified in supposedly helpful ways(van Rijsbergen & Sparck Jones, 73).

According to Shaw et al.(97), even the studies that have reported good performances for CBR can be questioned as to the practical value of their results. In their recent study, Shaw et al.(97) tested whether the results from cluster based retrieval experiments conducted in the last two decades gave results that were superior to results that could be attained by pure chance. They have set performance standards for 13 test collections by computing retrieval effectiveness values that could be obtained from a random cluster structure. They measured retrieval effectiveness by:

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

and compared the F values with operational levels of cluster-based performance reported in selected sources during the past 20 years. Their comparisons revealed that typical levels of operational cluster-based retrieval could be explained on the basis of chance. This, they suggest, is due to the fact that the structure imposed on a set of documents by topical relatedness may not reliably associate documents relevant to the same query, or more simply, topical relatedness does not necessarily manifest relevance relationships.

Another critical evaluation of the cluster hypothesis was provided by Ottaviani(94). Ottaviani rejected the cluster hypothesis altogether. He cited berry picking (Bates, 89) and citation pearl growing(Markey& Atherton, 80) models to describe users' information seeking process and criticised the cluster hypothesis on the grounds that relevance is a complex and dynamic phenomenon dependent on various interacting dimensions and a system cannot retrieve the relevant set in one simple step. Ottaviani pointed out that users' own clustering process is one where each stage involves some additional information and some loss of redundant information in contrast to a priori clustering where, each fusion means considerable information loss.

4.2 Cluster based retrieval

Assuming that the cluster hypothesis is true, CBR starts by comparing the queries with the clusters, or more accurately cluster representatives, from a database whose documents have a priori been classified into clusters. Detailed query-by-document comparison is then performed within the selected clusters.

A cluster is matched against a query by means of a matching function comparing the query with the cluster representative. In CBR, a cluster is generally either represented by terms ranked according to their cluster frequency(e.g. van Rijsbergen & Croft, 75) or by terms that represent a certain threshold of occurrences in the cluster. The results from CBR may be displayed by ranking clusters(Griffiths et al., 86) or by ranking the individual documents from a pool of relevant clusters (Voorhees, 85).

4.3 A taxonomy of CBR

It is possible to broadly catagorise the work done so far on CBR as:

- term clustering vs document clustering, or as:
- a priori clustering vs interactive clustering.

4.3.1 Term clustering vs. document clustering

In term clustering, the objects to be clustered are indexing terms and clusters are composed of terms that tend to co-occur in documents. Term clustering provides a mechanism for increasing recall of an IR system by allowing each term in a document and/or query to be replaced by the identifier of the cluster containing that term, and adding terms from each of the clusters which contain one of the original query terms(Willett,88).

This mechanism, however, does not consistently improve the search performance. Minker et al.(72) found that query expansion by the addition of clustered terms could reduce search performance in some cases. Sparck Jones(73) has conducted the most comprehensive study of term clustering. She used document occurrences as attributes and found that small clusters of low frequency terms were most effective, regardless of the clustering method used. However, text retrieval effectiveness was found to improve on only one of the three test collections used. According to Peat&Willett(91), term clustering using document occurrences tend to group terms that occur in similar numbers of documents which is generally not desirable.

In document clustering, terms become attributes of the documents rather than being objects of clustering. In comparison to term clustering, document clustering seems to be better justified and more widely used in the context of IR. As this project is involved in document clustering, the following sections will only focus on aspects of document clustering.

4.3.2 A priori clustering vs interactive clustering

A priori clustering involves classification of the documents of a collection without any user input/feedback. Its aim is to improve retrieval efficiency and effectiveness(see Section 4.2 on page 39).

A priori clustering has been the dominant approach in earlier work. But, its mixed performance as a retrieval mechanism, and its complexity and computing requirements has decreased its popularity and caused it to be abandoned in favour of methods like best match and Boolean retrieval.

Recently, wider usage of text retrieval systems (especially the use of the search engines and related precision problems), and diminishing computing time and space constraints seems to have caused a revival of interest in clustering documents. But the interest seems to focus more on interactive clustering, where clustering is seen as a tool improving user's interaction with a retrieval system rather than as a retrieval mechanism.

Among recent studies of interactive clustering, we can list Gordon's (91) user-based clustering, Kirriemur&Willett's(95) study focusing on document duplicates and Xerox' Scatter-Gather method (see Section 4.8 on page 53 for overview of some recent general purpose and Web clustering studies). Scatter-Gather method uses the a priori clustering methods but incorporates user's choices as input to subsequent clustering steps within an IR session(see below).

User-based clustering suggests that user evaluations can provide co-relevance data to form a basis for clustering. However, user-based clustering has important shortcomings, it requires a lot of user feedback which is not always available, and additionally it assumes that:

- the documents will have a relatively long shelf life,
- inquirer's queries will match the descriptions used for retrieving documents or clusters,
- cluster hypothesis will hold for queries, i.e. queries relevant to a given document will be descriptively similar, and documents co-relevant for one query will be co-relevant for other queries.

In addition to these, there are also several facilities provided by search engines, which present the user with term or document groupings created out of his search output(see Section 4.8.2 on page 63). These groupings aim to improve precision by helping the user to discover useful subsets of documents out of the search output.

4.4 Document clustering

Document clustering has been put to use for two non-exclusive purposes. First, as an alternative to linear associative searches as investigators looked for more efficient means of retrieval(e.g. Salton, 71; Yu, 74) and secondly as a component of IR systems which potentially could enable not only higher efficiency but also higher effectiveness(e.g. Jardine & van Rijsbergen, 71).

Salton(68), one of the early commentators of CBR, argued that document clustering would save time but would reduce effectiveness. Jardine&van Rijsbergen (71) and Croft(80) later claimed on the contrary that it had potential for improving the effectiveness since a cluster once found would tend to contain only relevant and no non-relevant items. The rationale for such expectation is that a CBR strategy takes into account the relationships between documents as well as the relationship between the query and the individual documents.

The first approach(clustering for efficiency) has become redundant with advances in best match search methods. The second approach(clustering for effectiveness) remains to be the primary topic of CBR and has been the subject of various comparative experiments(e.g. Jardine & van Rijsbergen, 71; van Rijsbergen & Croft, 75; Croft, 80; Griffiths et al., 84; El-Hamdouchi & Willett, 89). See Sections 4.5.1 on page 43 and 4.5.2 on page 46 for some experimental results.

4.5 Choice of method for document clustering

Clustering methods, initially devised for use in numerical taxonomy with lesser numbers of objects and attributes, were not suitable for the huge numbers of objects and attributes involved in document clustering. Computing time and space constraints were highly restrictive and in order to compensate for this, researchers have had to employ heuristics in creating the cluster structures in the early experiments. These methods required the users to make decisions on various input parameters and the final classification was generally dependent on these decisions. The necessary decisions were not easy to make as these parameters could only be determined empirically, but had to be specified in advance. Van Rijsbergen lists these input parameters(79) as:

- the number of clusters desired,
- a minimum and maximum size for each cluster,
- a threshold value on the matching function, below which an object will not be included in a cluster,
- the control of overlap between clusters.

The use of such parameters could be appropriate when there is a need to impose a certain structure on a group of objects, but in many other cases where one looked for 'natural' classifications, it would be more appropriate for at least some of these parameters to be determined as the result, rather than the input of a model which sufficiently resembles the human thinking. As van Rijsbergen(79) pointed out, the distinguishing characteristics of methods requiring these parameters as input was that they did not seek an underlying structure in the data but attempted to impose a suitable structure in it.

In later years, a considerable amount of research has been aimed at developing efficient means of generating similarity matrices and the subsequent classifications(Griffiths et al., 84). The successful artefacts from these studies and the developments in computing and storage led investigators to become less constrained by efficiency and shift emphasis to theoretical issues.

Theoretical soundness was mainly judged by the three criteria defined by Jardine& Sibson(71):

- the clustering produced by a method should be unlikely to change drastically when further objects are incorporated, i.e. stability under growth,
- the method should be stable in the sense that small errors in the description of the objects lead to small changes in the clustering,
- the method should be independent of the initial ordering of the objects.

These criteria still stand as a benchmark for validating the clustering methods applied.

An additional criterion involved in document clustering decisions is the computational requirements as the number of objects(documents) and attributes(terms) are particularly high. Some clustering methods have $\Omega(n^2)$ running times and $\Omega(n^2)$ space dependency as they need to consider all pairs of similarities. Fortunately, due to the sparse distribution of variables in document clustering, there are various different ways out of this problem[12].

[12] e.g. using an inverted index to skip the zero similarities(e.g. Croft, 77); avoiding looking at all pairwise similarities by considering a nearest neighbour algorithm as an initial, crude clustering (Murtagh, 83; Willett, 84; El-Hamdouchi & Willett, 86; Croft et al, 89); doing a coarse clustering of the collection which has lower time dependency and then applying a finer classification to the resulting clusters(e.g. Buckshot and fractionation algorithms used by Cutting et al, 92).It is also possible to reduce the

Some non-hierarchical clustering methods have rectangular running times, i.e. O(kn) where k denotes the desired number of clusters and n denotes the number of documents(Lewis, 92). These algorithms usually start by choosing seeds for a desired number of clusters and then assign each document to the closest seed. The procedure can be refined by iteratively improving the selection of cluster seeds.

Clustering methods have been classified in various different ways by IR researchers. These classifications highlight the different dimensions of the process:

- monothetic vs. polythetic: in monothetic grouping each individual possesses a large but unspecified number of the properties whereas in polythetic grouping each property in the group is possessed by a large number of the individuals and no property is possessed by every individual in the group[13](Sparck Jones,70),

- the clustering based on a measure of similarity between the objects to be clustered vs. the clustering proceeding directly from the object descriptions[14](van Rijsbergen, 79),

- partitioning/overlapping/hierarchical: classifications where common documents are allowed(overlapping), no common documents are allowed(partitioning/hierarchical), and where clustering structure is represented by a tree(hierarchical) (Can&Ozkarahan, 90),

- single-pass/iterative/graph theoretical algorithms: documents are handled only once in single pass algorithms whereas each iteration improves the value of an objective function in iterative algorithms(Can&Ozkarahan, 90).

Historically, the principal approach in choosing a document clustering method has been deciding which hierarchical method to use, and many comparative studies have focused only on hierarchical methods. So, it may be worth focusing further on these methods, and also their counterparts, the non-hierarchic methods.

4.5.1 Use of hierarchical clustering methods in IR

resource consumption considerably by imposing conditions on the relationships between the documents and the centroids(Voorhees, 86).

[13] Monothetic/polythetic distinction is important in divisive clustering. The distinction has been described differently by Willett(88). According to him monothetic classifications require that a given cluster must contain certain terms, whereas in polythetic classifications there are no specific terms required for cluster membership. In a polythetic classification, we may have non-typical class members.

[14] Given that direct classification of the documents has proved theoretically to be intractable(van Rijsbergen, 79), the overriding approach in CBR has been clustering based on a measure of similarity between the objects to be clustered.

Because hierarchic clusters provide different solutions at various levels, they represent a special case of searching. They can be searched either top-down or bottom-up. Top-down processing involves iteratively taking the highest scoring branch, until some stopping condition is achieved. The subtree at that point is then returned as a result(Cutting et al., 92). Bottom-up searching requires definition of a starting point. Van Rijsbergen and Croft (75) have coped with this problem by simply assuming that a single relevant document is readily available; another way is to find that document by a different search technique. Preferably, all clusters may be scanned via a best match search to identify the cluster most similar to the query (Willett, 88).

IR researchers have put forward various reasons for favoring hierarchical clustering methods:

1. efficient strategies can be devised to search a hierarchic clustering (Jardine & van Rijsbergen, 71),
2. the level of a cluster can be used as a parameter in retrieval strategies analogous to rank position or association measure threshold (Jardine & van Rijsbergen, 71).
3. construction of a hierarchic system is much faster than construction of a non-hierarchic(that is, stratified but overlapping) system of clusters (van Rijsbergen, 79),
4. the storage requirements for a hierarchic structure are considerably less than for a non-hierarchic structure, particularly during the classification phase(van Rijsbergen, 79).

However, some of the points made are platitudes rather than being clearly established benefits and there seems to be reason to question the appropriateness of hierarchical methods for document clustering(the fact that we may like to have a hierarchic representation is not enough).

The basic argument against the use of (agglomerative) hierarchical methods in CBR is that document collections do not have any intrinsic qualities that make them appropriate for hierarchic conceptualisation; on the contrary, due to the variety of aspects a document may cover, an overlapping classification where a document can be a member of more than one group seems more appropriate.

While it has been considered an advantage elsewhere(Jardine & van Rijsbergen, 71) that these methods do not allow overlapping, the fact that a document is usually relevant to more than one query(which are not necessarily closely associated to each other), makes this attribute disadvantageous.

According to van Rijsbergen, hierarchic methods are preferable to non-hierarchic methods as they are less consumptive of computing resources, but Willett claims the opposite:

> *"...non-hierarchic methods are generally much less demanding of computer resources than are hierarchic methods, and thus would seem to be more appropriate for the clustering of large files of chemical structures..."(Willett et al., 86b).*

What's more, research work to date shows that even when a hierarchical structure is imposed, it is typically only the bottom-level clusters which give good retrieval results(see below). Even if we have a hierarchical representation, we do not seem to have much use for its upper levels. Everitt's(93) critical statement about inappropriate uses of hierarchical clustering might be relevant for document clustering as well:

> *"It is [to] biological applications such as the evolutionary trees that hierarchical classifications are most relevant. Hierarchical clustering procedures are, however, now used in many other fields in which hierarchical structures may not be the most appropriate. The danger of imposing a hierarchical scheme on data which is essentially non-hierarchical is clear."*

Hawkins et al.(82) also warn users of hierarchical methods:

> *"Users should be very wary of using hierarchic methods if they are not clearly necessary."*

Even when appropriate to the task, hierarchical methods suffer some shortcomings: ultimately all of them reduce the data to a single cluster containing all the individuals, forcing the investigator to impose a particular stage to stop. Making this decision may not be easy in many cases.

There are also additional shortcomings of particular methods. For example, the single-link method(also known as the nearest neighbour technique), the most extensively used method in document retrieval (van Rijsbergen, 79) has a very loose definition of cluster membership. It only seeks to ensure that any cluster member is more similar to at least one member of the cluster than it is to any member of any other cluster and it may form clusters with little internal cohesion(chaining effect). Although it does satisfy the theoretical soundness and efficiency criteria, it has been found to perform poorly in many experiments(e.g. Dubes & Jain, 76) and many comparative studies of use of hierarchic agglomerative methods for CBR show that the single-linkage method generally gives results that are inferior to those obtainable when other hierarchic agglomerative methods are used(Willett, 82; Griffiths et al., 84; Voorhees, 85; Willett, 88).

It has not been possible to draw conclusions from the comparative studies as to which hierarchical method gives good results under which circumstances. In Griffiths et al.'s(84, 86) studies using various test collections[15], group average method gave better results in recall-oriented searches when only one cluster was retrieved and the Ward's method gave the best overall performance when more than one cluster was retrieved. Allen et al. (93) have also used Ward's method for their interface for navigating clustered search outputs (see Section 4.8.1.7 on page 62). Voorhees(85) found complete linkage to be the best performer in her experiments and El Hamdouchi(87) and El-Hamdouchi & Willett(89) found complete linkage to be worse than both the single linkage and group average[16]. El Hamdouchi and Willett's studies have also found non-cluster searches to have better performance than cluster searches.

Griffiths et al.(86), demonstrated that cluster search could give a performance comparable to and in some cases significantly superior to non-cluster searches when a simple, non-hierarchical overlapping classification containing just the nearest neighbour clusters(NNC) are used. NNC's can be implemented using inverted file lists corresponding to the terms in the query to yield a vector, the *i*th element of which contains the sum of the weights of the terms common to the query and the *i*th document. The largest such element then specifies the nearest neighbour for that query. This algorithm, described by Noreault et al.(77), can be implemented at a relatively low computational cost.

An interesting finding of Griffiths et al. was the retrieval of quite different sets of documents by NNC searches and full search. In a retrieval experiment, the full and NNC searches of the INSPEC database retrieved 233 and 203 documents and only 90 of these documents were common to both types of search[17].

4.5.2 Use of non-hierarchical methods in IR

Non-hierarchical methods proceed by seeking sets which satisfy certain cohesion and isolation conditions defined in terms of the similarity measure and they may result in partitions or clumps.

There seems to be some aversion to non-hierarchical methods among IR researchers. Van Rijsbergen(79) has suggested that non-hierarchical methods are unpopular due to the

[15] Keen and Cranfield in 84; Keen , Cranfield, Evans, Harding, Lisa, Inspec and UKCIS in 86.

[16] Willett(88) questions the validity of the varying results from comparative experiments; Griffiths et al. used Dice's coefficient instead of a distance measure in their application of the Ward method and this makes the comparison with other studies difficult. On the other hand, the algorithm used by El Hamdouchi and Willett in building complete link clusters was a suboptimal algorithm (CLINK) and hence the outcome could in fact be better.

[17] This provides an opportunity to improve the level of performance by combining searches. Unfortunately, it is not easy to predict the proportion of documents to take from each type of search. A previous test by Croft&Thompson(84) on combining different types of search by applying an adaptive mechanism has failed to give results superior to that of a single search strategy applied consistently.

computational difficulties involved. Willett(88) agrees by arguing that non-hiearchic methods need to generate and evaluate all possible partitions to see which of them best satisfies some optimality criteria and this computational burden is infeasible for databases of nontrivial size. This, he continues, has led to the use of heuristic methods resulting in suboptimal partitions.

In early work, the non-hierarchic methods have mostly been used to improve best match searches, employing heuristics and input parameters such as the number of clusters required, minimum or maximum cluster sizes, and threshhold document-cluster similarity levels. However, their retrieval effectiveness was not as good as those of unclustered files and according to Willett(88), with advances in best match search algorithms, there is no need to focus on them further.

According to Lewis(92) non-hierarchical methods are not preferable even when the computational complexity is low. He agrees with Willett(88) that they have only been thought as a supplement to nearest neighbour search and in order for them to be useful for this purpose, the partition must be fairly fine. He argues that non-hierarchical methods have not been aggressively pursued by the IR community due to the large number of clusters required in the partition.

Yet, there seems enough evidence to advocate large numbers of small clusters rather than small/limited numbers of clusters made up of large numbers of documents. Many studies have found a large number of small bottom-level clusters to be most effective in retrievals(e.g. Sparck Jones, 73; Griffiths et al, 84, 86; Shaw, 90[18]) Some others got superior results when they limited searches to bottom-level clusters without needing to navigate the hierarchy at all(e.g. Croft, 80[19], El-Hamdouchi & Willett, 89). Macskassy et al. (98) study on how users clustered documents from Web search results has also found that users generally created rather small clusters.

Given such evidence, building whole hierarchies becomes counter-intuitive, since upper levels are less likely to be useful.

[18] Although Shaw implied that the inverse relationship between retrieval performance and cluster sizes would mean a random cluster structure under the random graph hypothesis.

[19] His findings have led Croft to advocate a network model for document retrieval instead of a hierarchical method.

4.6 Other necessary decisions

Apart from the choice of method to be used, clustering requires decisions to be made on document representation, method of association, matching function and cluster representation.

4.6.1 Document representation

In CBR, two types of variables have been suggested to represent documents:

1. <u>citations</u> used in citation clustering which measures the degree of similarity between a pair of documents by the citations shared in common(Willett,88),
2. <u>terms</u> (index terms, keywords, thesaurus terms).

In a great majority of CBR work, only terms have been used for representing documents. Generally, the vector space model is used for document representation and a document is represented by a document-by-term matrix where individual entries d_{ij} indicate whether term j occurs in document í (or the number of occurrences of term j in document i).

Different document attributes such as the number of citations, document date, type, length, author, journal, and level of treatment may as well be used as additional attributes. But, given the lack of any precedents for their use in IR, it is hard to decide how to utilise them. They represent distinct data types and methods to incorporate them into clustering algorithms may not be easy to devise (e.g. for date or length it would be necessary to establish ranges, and their effect would not necessarily be linear - see Graph 10-1 on page 176 for the relationship between publishing year and relevance judgments for a set of 20 search outputs from the user experiments).

The next question to be considered is the number of variables to be measured for each document. Usually, there are a huge number of variables which might be used to produce a classification and a restriction needs to be applied for a couple of reasons. Firstly, more variables mean increasing computational requirements. Secondly, a more than necessary number of variables may obscure the clustering structure (Everitt, 93).

Number of variables may also affect the number of resulting clusters when that number is not a parameter imposed to the system but rather estimated by the algorithm itself. Can&Ozkarahan's(90) experiments on Inspec and Tods214 databases showed that the number of clusters is directly proportional to the size of indexing vocabulary in a given database.

A final decision on variables concerns whether the variables should be weighted or not and if they are to be weighted, how it should be done. In CBR, this means a choice between binary and weighted indexing. Sneath and Sokal(73) are against the use of weighting schemes in cluster analysis. Weighting schemes are not popular among CBR researchers either; studies by Willett(83) and Can&Ozkarahan(90) revealed that weighting made little difference to the clusterings made by the single link method.

According to Korpimies&Ukkonen(98) term weighting is necessary in output clustering and the focus should be on the term frequencies within the output set; terms which are frequent or too infrequent within the document set should be given small weights. They have formulated "contextual inverted document frequency" as:

$$cidf(Q,t_j) = \frac{1}{\sum_{i=1}^{n} w_{ij} \times rel(Q,D_i)}$$

Where *wij* stands for the weight of term *tj* in document *Di*, n is the number of documents in the collection and rel(*Q, Di*) indicates the degree to which document *Di* matches query *Q*.

Their experiments with the *cidf* weighting revealed that the "clustering potential" (defined as the ratio of average similarity between all document pairs that belong to a cluster divided by average similarity of all pairs within the output set) increased when the output frequencies were incorporated in the clustering; there was an improvement of 18% compared to the use of IDF, and 34% compared with the situation where all terms had equal weights. Further gains were obtained when IDF was used to exclude low frequency terms from the similarity measurements.

The main idea behind the *cidf* is similar to that of the Term Selection Value (Robertson, 90), a measure correlating with the terms' ability to discriminate between documents(see Section 5.3.5.1 on page 111).

4.6.2 Measuring the association between documents

The most common measures of association used in IR are:

1. Simple matching coefficient: number of shared index terms[20],
2. Dice's coefficient: the number of shared index terms divided by the sum of the number of terms in two documents. If subtracted from 1, it gives a normalized symmetric difference of two objects.
3. Jaccard's coefficient: number of shared index terms divided by union of terms in two documents,
4. Cosine coefficient: number of shared index terms divided by multiplication of square roots of number of terms in each document,
5. Overlap coefficient: number of shared index terms divided by minimum of number of terms in each document.

There are also several dissimilarity coefficients, Euclidian distance being the best known among them. However, it has a number of important shortcomings: It is scale dependent, which may cause serious problems when it is used with raw data and it assumes that the variable values are uncorrelated with each other(Everitt, 93). A major limitation in the IR context is that it can lead to two documents being regarded as highly similar to each other, despite the fact that they share no terms in common(but have lots of negative matches). The Euclidian distance is thus not widely used for document clustering, except in Ward's method(Willett, 88).

Finally, there are some attempts to base a measure of association on a probabilistic model(e.g. Maron&Kuhns, 60; Kuhns, 65; Guiliano, 65; El Hamdouchi, 87). Maron&Kuhns, Kuhns and Guiliano measured the association between two objects by the extent to which their distributions deviated from stochastic independence. Most of their measures unrealistically took as a null hypothesis independence and equal probability of occurrence of the terms in each string(Jardine & van Rijsbergen, 71). El Hamdouchi used the probability of two documents being co-relevant to a query as the main criterion for cluster formation.

4.6.3 Matching function

Matching function is the algorithm used for measuring the similarity of the objects to the representative. In CBR context, it measures the similarity of the query to the cluster representative.

Often association measures used in assessing document similarities are used as matching functions, e.g. Jardine and van Rijsbergen(71) have used the normalized symmetric

[20] Everitt(1993) defines matching coefficient in a different way, according to him it denotes the ratio of total number of variables on which the two objects match, to the total number of variables. Hence 'negative' matches where both objects have 0 for the variable contribute to the similarity coefficient.

difference(Dice's coefficient subtracted from 1), Griffiths et al.(86) and El-Hamdouchi & Willett(89) used the cosine coefficient.

4.6.4 Cluster representation

Cluster representation is usually similar to document representation when representation is used for retrieval purposes. This similarity is especially evident with the hierarchical methods where clusters themselves become objects of the clustering process in subsequent iterations.

A cluster may simply be represented by a binary string in which a 1 in the *i*th position indicates the occurrence of the *i*th index term in the cluster, or there may be a requirement for a minimum number of occurrences for such representation. This number may be a fixed threshold or an estimate based on the number of documents (e.g. Jardine and van Rijsbergen(71) suggested looking for a minimum of log2|C| documents). The number of distinct terms used in cluster representation is called the maximum length of the centroid.

Cluster representation involves further considerations when a cluster is presented for user's viewing. The form of representation becomes critical as it affects user judgements.

Several studies have concentrated on how various forms of surrogate affect relevance judgements for documents. Some have found that the quality of the surrogate increases with its length(Hagerty, 67; Weis and Katter, 67; Marcus, Kugel and Benenfeld, 78) whereas others found no significant differences between different types of surrogates in terms of quality (Rath, Resnick and Savage, 61; Kazhdan, 79). In his overview of the literature on relevance, Mizzaro(97) concludes that while there seems to be an agreement that titles have the least quality followed by keywords, extracts and abstracts, quality of the words, besides their quantity should also be taken into account.

There are no corresponding studies for representing clusters although it is probably even more critical as the level of abstraction is higher. Usually, clustering studies that involve user judgements adopt a certain representation based on common sense or precedent studies. An example is the Scatter-Gather method where clusters are presented by summaries containing topical terms(terms with high weights in the cluster's profile) and typical titles(see below for details). Another example is a study by Allen et al. (93) where clusters have been represented with ranked document lists in similarity order.

In a couple of recent studies, researchers have used phrases in cluster creation and representation in addition to terms; phrases have more descriptive power, and can make it easier for the users to interprete cluster representations (for details see Paraphrase Interface in Section 4.8.1.5 on page 60 and HuskySearch in Section 4.8.2.4 on page 71). However, user experiments are needed to evaluate whether and how phrases can be used to improve cluster formation and representations.

4.7 Cluster validity

Initially, cluster validity is tested by comparing the clusters with randomly generated clusters. If the clusters created prove to be different from randomly generated clusters, further tests may be undertaken to test the overall structure and the validity of individual clusters. Willett(88) suggests using Ling's(75) random graph test for testing whether document clusters are non-random.

The extent to which the cluster structure represents similarity relationships in the data set can be tested by distortion measures. It is thought that a method that imposes a small degree of distortion on the similarity matrix may identify more natural groupings. However, it has also been argued that distortion of the similarity matrix is not necessarily undesirable, a clustering method should attempt to identify groupings that are more intense than those present in the similarity matrix (Williams et al., 71).

For a hierarchical method, distortion may be measured by comparing the set of interdocument similarities in the input similarity matrix by the output dendrogram. Studies by Griffiths et al.(84) showed that the group average and single linkage methods resulted in the lowest degrees of distortion.

Apart from these mathematical approaches, user feedback on co-relevance is very important for validating document clusters. Although a very cumbersome process when the data is not readily available, many studies undertook to evaluate clusters using relevance data from user evaluations(e.g. van Rijsbergen & Sparck Jones, 73; El-Hamdouchi & Willett, 89; Shaw W.M., 90; Gordon, 91; Shaw R.J. & Willett, 93; Kirriemuir & Willett, 95; Hearst&Pedersen, 96). Most of these studies have used the effectiveness measure E for evaluating the performance(van Rijsbergen, 79). For a given search with a recall value of R and precision value of P:

$$E = 1 - \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

where β is a user-defined parameter reflecting the relative importance attached to recall and precision. Generally, a value of 0.5 is used for precision-oriented searches, and a value of 2 for recall-oriented searches. A value of 1 represents equal importance attached for precision and recall. Van Rijsbergen (in Sparck Jones, 81) acknowledges, however, that this measure involves loss of an unquantified amount of information caused by the conflation of the precision-recall graph to one value.

An underlying weakness of using co-relevance of documents to assess clusters is that it expects all documents found relevant by a user to be similar to each other. However, two co-relevant documents can be more dissimilar to each other than many other non-relevant documents, as they may be responding to different parts of a user's query.

4.8 Examplary clustering appications

A number of projects/applications that have similar purposes to this project are reviewed in this section. The review is divided between general purpose clustering work and clustering focused on Web documents.

Work on clustering Web documents is particularly interesting as it represents a renewed interest in clustering applications stimulated by the need to introduce some order to the vast numbers of Web documents and help users to cope with huge search output results.

4.8.1 General purpose document clustering work

4.8.1.1 Xerox' Scatter/Gather browsing system

Cutting et al.(92) argue that a typical information search session takes place within a spectrum where we have browsing with no well defined goals on one side and a narrowly defined search for particular document(s) on the other. Cluster-based retrieval, like many other information access techniques emphasises the search end and neglects the browsing end. Cutting et al. attempt to fill this gap using clustering as an information access tool in its own right.

In their proposed system, the Scatter-Gather browsing method is envisaged as the table-of-contents of a book and one or more text search methods are also incorporated for index searching. The browsing method is primarily intended for helping the user to formulate a search request, but it can also be used to reach individual documents or organise a huge search output.

Typically, a user is presented with summaries of a small number of document groups resulting from scattering the collection. These summaries consist of topical titles(titles of documents close to the cluster centroid), and typical terms(terms of importance in the cluster).

The first such summary is an outline of the corpus. Based on these summaries, the user selects one or more groups for further study. These groups are gathered together to form a sub-collection and they are further clustered(scattered) into a small number of document groups; the process goes on until the groups become small enough to enumerate individual documents.

The Scatter-Gather method uses two different clustering algorithms: Buckshot, a fast algorithm for online Scatter/Gather of user selected clusters and, fractionation, a more accurate but slower clustering for the static partitioning of the entire corpus which is presented first to the user. Both techniques are drawn from the hierarchical algorithms, but achieve rectangular time bounds.

These methods find the k centroids using the aggregated results from group average agglomerative clustering applied locally to small sets of documents. Buckshot uses a random sample (size $\sqrt{kn}$) to find the centroids, fractionation works on fixed sized groups, iteratively treating each group(cluster) as individuals until k groups remain. Each cluster is represented by the "most central" documents, i.e. the documents whose similarity to the cluster profile are highest, and the documents on the outskirts are excluded from representation of the contents of the group. Typically medium frequency words are used to sort the corpus.

After the centroids are identified, each document is assigned to the nearest cluster. A split and join algorithm is also used to refine the clusters by separating poorly defined clusters into two well separated parts and merging clusters which are too similar based on certain similarity and distinguishability criteria.

Cutting et al. argue that the algorithms are capable of finding natural clusters if the input data has well separated clusters. They prove this by showing that for the Buckshot algorithm the probability of sampling no items from a certain natural cluster is very low. But this may not always mean that one of these items will become a centroid, especially as the number of 'desired' clusters will not necessarily match the number of 'natural' clusters[21]. As for fractionation, they claim that, by definition, it merges only the pairs in the same 'natural' cluster.

In a sense, the Scatter-Gather method is even more ambitious than the cluster hypothesis as it implies that it can bring together and efficiently summarise the contents of similar documents at various hierarchical levels. This idea is embodied in their cluster refinement hypothesis(Cutting et al, 93), which states that

[21]Join and split algorithms should help to reduce this gap to some extent. But their exemplary session shows a constant number of 8 clusters at each iteration.

"documents similar enough to be clustered together in a fine-grained clustering will be clustered together in a coarse-grained clustering".

This is questionable as some of the attributes that have been effective at the finer levels will inevitably be replaced by other attributes that characterise the larger groups, and the new combinations of attributes cannot always be expected to keep the finer groups intact. Additionally, it is difficult to keep on providing coherent and meaningful clusters at upper levels since summarisation will require higher and higher losses of information at those levels.

Scatter-Gather browsing has later been enhanced further by incorporation of the TileBars interface(Rao et al, 1995). Within the TileBars interface each document is represented by a large rectangle and the rectangle is partitioned into squares each representing a coherent text segment or title. The partitioning is done using an algorithm called TextTiling[22].

The first evaluation study of Scatter-Gather consisted of a user survey of four users, each executing 13 queries on a combined query system of Scatter-Gather and the TileBars interface(Hearst et al, 95). The results can be summarised as follows:

- the participants performed well compared to the initial results returned for the other systems participating in TREC-4,
- Scatter-Gather clustering was found interesting but one participant expressed concern about tossing out appropriate documents,
- system performance was a formidable problem keeping the participants from achieving better results,
- participants usually employed Scatter-Gather for narrowing down the set and the TileBars for finer selection.

Later, Hearst&Pedersen(96) conducted further experiments for the evaluation of the Scatter-Gather method. In these studies, subjects were less successful in finding relevant documents for a given query through browsing the hierarchical structure than by employing similarity searches. This has led Hearst&Pedersen to shift focus on applying the Scatter/Gather on search outputs generated by conventional search methods. Further experiments based on TREC/Tipster standard reference collection using 49 TREC-4 topics showed that such

[22] The TileBars interface requires the user to successfully parse the different topics of the query into term sets and enter them on different lines on the interface. (which may be too much to ask from a typical user). On the query output, each row of the rectangle corresponds to hits for one of the term sets from the query. When term sets are discussed simultaneously, their corresponding tiles blend together to cause a prominent dark block to appear. Scattered discussions have lightly colored tiles and large areas of white space. A mouse-click on a dark region brings up a view of the document starting at a passage with a large number of hits(Rao et al, 1995). The user can specify certain conditions on tiles in the query (e.g. that a certain term must occur at a certain percentage of tiles).

clustering gave superior results as compared to ranked titles. Their findings can be summarised as follows:

1. The top-ranked cluster almost always contained 50% or more of the relevant documents retrieved and third, fourth and consecutive clusters usually contained 10% or fewer.
2. Precision values of best clusters were superior to the values from the ranked ordering truncated at the same number of documents for outputs of up to 1000 documents. At 1000 documents, relevance ranking outperformed clustering but this was mainly because number of clusters was fixed at five and clusters became too large at this level. When the best cluster was re-clustered for n=1000 documents, the best cluster out of the best cluster again outperformed the ranked results.
3. Precision comparisons between the original ranked list and the best cluster at small cut-off levels(i.e. top 5, 10, 20 documents) have also revealed superior results for the best cluster.
4. A user study was conducted to find out users' ability to choose the "best cluster". The study consisted of four students each executing 13 queries. Twelve cases where no relevant documents were evident were omitted and in 31 out of the remaining 38 cases, users have chosen the best cluster. In 22 of these cases the best cluster was the only choice, in 10 it was one of two choices and in 6 it was one of three choices. In post-interview users indicated that they used the Scatter-Gather display mainly for narrowing down the output. None of the participants thought having more than five clusters would be a good idea.

These findings are less impressive than they appear as they are based on best clusters and the accompanying assumption that users are able to identify them. The comparison of best clusters with ranked lists loses some of its significance when the best cluster is not the only (or first) cluster but one of two or three choices. In such cases the precision would decrease accordingly, and quite possibly below the level attained by ranked lists.

Hearst&Pedersen give to reasons to explain the success of Scatter/Gather in this study as opposed to many past clustering studies in IR that gave inconclusive or negative results:

1. These experiments were run on large text collections where most documents were full-texts as opposed to earlier studies where documents were composed of titles and abstracts and no large text collections with queries and relevance judgments were available.
2. In earlier studies, clustering was done over the entire collection in advance, in this study clustering is dynamic and the results are highly dependent on the given query.

In the light of their results Hearst&Pedersen argue that dynamic clustering has the potential to be more closely tailored to the characteristics of a query than static clustering and revise their assumptions of cluster hypothesis:

> *"We too assume that documents that cluster together are similar in some ways, and thus relevant documents will tend to cluster near other relevant documents and further away from non relevant ones. However, in contrast with the assumption underlying the strategy of earlier work, we do not assume that if two documents D1 and D2 are both relevant or nonrelevant for query QA, they must also both be relevant or nonrelevant for query QB."*

Another study by Xerox researchers(Pirolli et al.), reported in 1996, assessed Scatter/Gather method's ability to help users in finding relevant documents and facilitate incidental learning about the test collection. They compared Scatter/Gather with a 'word-based' (free text) search interface using the Tipster text collection and 12 TREC topics. Sixteen users participated in the experiments. They were asked to find articles relevant to given topics, draw diagrams representing the topic structure of the collection and generate search terms for given query topics after having used the search systems.

The results revealed that participants using the Scatter/Gather interface to answer queries took a significantly longer time than those using the 'word-based' interface, and found less numbers of relevant documents with substantially more variation in number of documents saved. In 27% of the Scatter-Gather queries (13 of the 48), users chose no documents as relevant, whereas all the 'word-based' interface queries resulted in at least one saved document.

Yet, Scatter-Gather users generated significantly more new terms that were not given in the query description, as compared to the 'word-based' interface users. But, there was no significant difference between the two methods in the number of nodes and links participants used to create diagrams to represent topic structures (except at the highest level; Scatter-Gather subjects used significantly more primary topics in their diagrams).

Pirolli et al. concluded that Scatter/Gather was not as effective as a common word-based search technique as a stand-alone IR tool, but it might be useful "in support of the kind of exploratory sensemaking activities that occur when users encounter large unknown text collections".

They have also analysed the distribution of relevant documents across clusters and concluded that:

- most relevant documents for a typical (TREC) query would tend to be grouped into few clusters by the Scatter-Gather algorithm,

- the users judge most relevant documents to be grouped into few clusters when they make precision estimates for clusters,
- the users' judgments appear to have a well-defined power-law relationship to the actual distribution of relevant documents.

In a subsequent study, Pirolli (97) used the Scatter-Gather method to investigate users' cluster navigation choices. Based on prior study results, he maintained that users select more clusters as the difficulty of the query decreases. This is based on the assumption that users select clusters in order of decreasing profitability (i.e. precision) up to a certain threshold. The threshold is the point where the overall average rate of finding relevant documents will be decreased by adding the next lower ranked cluster.

Pirolli hypothesised that users base their cluster selections on their expected profitabilities derived from the cluster summaries. He compared 151 cluster selection decisions made by users for a previous study against predictions they have formulated using three different kinds of evaluation functions. These functions measured the correspondence ('degree of information scent') between a query and a cluster using simple word overlap, interword correlation and strength of activation (where evaluation is based on clusters that receive user attention). The comparisons revealed a good level of prediction of user decisions for all three evalution functions.

The importance of this study is that it represents a user-centered approach in its attempt to predict user behaviour with a measurement of expected cluster profitability. Although it may not involve a measurement of actual usefulness of the clusters, it points to factors affecting users' decision-making.

4.8.1.2 Dubin's work: document analysis for visualisation

Dubin(95) stressed the importance of good representation of documents for a document visualisation system. Assuming that document clustering supports the user's goal of separating relevant documents from non-relevant ones, he has focused on finding out term sets that display a high power of discriminating clusters.

He used the term 'group discriminatory power' to define the property of two or more variables to separate records into groups, irrespective of the usefulness or interpretability of those groups. The sets that display a non-random distribution over a document space are supposed to have high discriminatory power.

He based his experiments on sets of a hundred documents from four diverse databases and used the centroid method for calculating term discrimination power. For each of the sets he chose six 'worst', six 'good'(highest ranking) and six 'best'(least correlating terms amongst highest ranking) discriminating terms. The 'best' discriminators were used to test the hypothesis that strong discriminators with few high correlations among variable pairs can demonstrate better discriminatory power than strong but highly intercorrelated variables.

The results revealed higher clustering structures when terms with high discriminatory power are used. The output for one of the document sets used was projected on a 2-D display using the VIBE positioning function. This display showed that the clustering structure is more apparent when the correlations are minimised.

4.8.1.3 Kirriemuir & Willett's clustering work for the purpose of identifying duplicate and near duplicate full-text records

Kirriemuir&Willett(95) worked on clustering search outputs using hierarchical cluster analysis. Their purpose was to detect duplicate and near-duplicate fulltext records and save users from inspecting the duplicates. Such a facility may especially be appreciated in multi-database searches.

They took the number of common words and their frequencies in the documents as a measure of similarity between document pairs and applied twenty combinations of five different similarity coefficients and four different agglomerative clustering methods on outputs from 20 searches. The search outputs varied between 80 to 327 documents.

The most laborous part of the study was the generation of user evaluations for each document from each search output in order to compare and evaluate the performance of the clusters generated by each of the twenty coefficient/method couples.

It was expected that a perfect clustering algorithm would group all pairs of documents that were evaluated by the users as "identical/very similar" (defined as the type5 relationship) and then go on with those that were document rewrites and those couples where one item was the abstract of the other(defined as the type4 relationship).

Two different approaches were adopted for the comparison. The first one being rather stringent, looked for type5 and type4 relationships identified before any weaker (types1-3) relationship was encountered; it resulted in a high variability of performance between different similarity coefficients and the best coefficient-method couple amounted to identification of 53% of type5 and type4 relationships before any type3 relationship was encountered.

The second comparison involved the proportion of type5 and type4 relationships in the first 25% of the fusions and generated better results. Ward's method applied to the cosine coefficient performed best by identifying 79% of the type5 and type4 relationships. These results, given that type 4 and 5 relationships represent very high similarity among the documents, reveal the difficulty of identifying document similarities by statistical measures.

4.8.1.4 Willett et al.'s experiments on clustering chemical structures

Willett et al.(86b) have compared various non-hierarchical methods for clustering chemists' search outputs that were too large for individual inspection. As a result of their experiments, they have found that a method that is due to Jarvis&Patrick(73) performs consistently better than the other methods that were tested.

The method involves implementation of a near-neighbour table, an NxK integer array containing the identifiers for the K nearest neighbours for each of the N objects. Jarvis&Patrick suggest that two objects I1 and I2 should be placed in the same cluster if I1 is a near neighbour of I2, I2 is a near neighbour of I1 and I1 and I2 share at least *S* near neighbours in common, where *S* is a user-defined parameter. Orderings of the neighbours can also be incorporated in the formulation for a weighted model[23].

4.8.1.5 Paraphrase interface

Anick&Vaithyanathan's(97) Paraphrase system combines phrasal information and document clustering to achieve context-based retrieval. It is envisaged that larger linguistic units such as noun phrases will make better units for describing cluster contents than a list of single words.

The documents from a collection are reduced in dimensionality using principal component analysis, and then clustered using K-Means clustering algorithm, with a given target cluster size. This is followed by extraction of noun compounds which are in this project defined as any contiguous sequence of words consisting of two or more adjectives and nouns that terminates in a head noun.

Each noun compound (also called a phrasal unit), is then run as a query against the clustered document collection, and the number of documents containing the phrase in each cluster are recorded. The word and phrasal occurrence statistics are then combined to find out "sample topics" for the cluster, which are expected to be good descriptors of its context.

When a user issues a query, the cluster ids for the top n ranked documents are retrieved from the result list. The frequency of result list documents within each cluster is then used to rank

[23] Willett et al.(86b) have found little difference between the use of weighted and unweighted versions.

the contexts and the top m contexts are presented to the user. Each context is depicted by means of four kinds of information: a short list of key terms, sample topics (including both words and phrases), pulldown lists of phrases that documents within the cluster make references to and which contain words in the current query, and finally a sample title.

This paradigm is expected to provide a richer dialog between system and user by introducing a new level of abstraction, higher than individual documents and lower than the entire database. Phrasal information associated with the clusters can also be transformed into interactive *contexts* for the purposes of iterative query refinement and structured browsing.

Authors identify two shortcomings in the model. Firstly, adding a new level of abstraction into an IR system introduces new kinds of user misconceptions; e.g users may not realise that all relevant documents may not be in the same context, regardless of the context's description. Secondly, noun compounds may not capture the useful relationships for certain concepts; these include some proper nouns, common nouns that tend to occur with prepositions rather than modifiers and verbs.

They also point to the need for user experiments, fine-tuning the process and investigation of its scalability and performance across different database domains.

4.8.1.6 Latent Semantic Indexing for improving the efficiency of clustering algorithms

Xerox researchers Schutze and Silverstein (97) have used Latent Semantic Indexing (LSI) and truncation to speed up the distance calculations involved in clustering. Each document is projected onto a small subspace of the total term space, thereby reducing the average number of terms in each document.

In truncation, document vectors are weighted, for each document vector *d* only *c* largest components are kept and all other components are set to zero. For the weighting, term frequency weighting is used where d_t is replaced by 1+log d_t.

In LSI, the documents are converted to LSI space and the *d* lowest dimensions of each document are obtained. LSI space has the advantage of guaranteeing that the documents have the lowest possible least-square distance to the original documents after being projected to a *d* dimensional space., LSI finds an optimal dimensionality reduction, and according to the authors, is less likely than truncation to discard information crucial for clustering.

Their experiments revealed that the term reduction has no measurable effect on cluster quality. No significant difference was found between the effectiveness of the two projection methods at term thresholds of 50 and 150 (excluding the inferior results obtained with truncation with 20 terms). As a result of the comparable performances of LSI and truncation, authors recommend the use of truncation with about 50 terms per documents, as LSI projections require expensive 'compile-time' calculations.

4.8.1.7 Interface for navigating clustered document sets returned by queries

This is one of the earliest studies focusing on clustering as an interactive tool for browsing search outputs. Allen, Obry and Littman (93) designed an interface for exploring the structure of search outputs. The interface uses an interactive dendrogram displaying a hierarchical clustering of the documents.

The interface consists of four major windows: the query window, the interactive dendrogram, the subtree document lists, the text window and lists of 'proximal' documents. Colours are used to indicate the selected subtree and document. The highest similarity document within the selected subtree is highlighted by a green path from the root of the subtree to the leaf representing that document. The length of the bars of the dendrogram is proportional to the article's similarity to the query.

For the implementation, a set of 25,629 articles from the Academic American Encyclopedia that had been preprocessed using the Latent Semantic Indexing tools was used. The top 400 documents returned for a query are clustered using Ward's algorithm. For efficiency, several parallel and distributed environments are used, resulting in an overall response time of approximately one minute. Clusters are represented to the user in the form of ranked document lists in similarity order.

Informal experiments revealed that the usefulness of the interface varied a great deal between queries. Clustering appeared to be useful for retrieval when questions are composed of several parts or include terms with multiple meanings. But the interface was found to be of little benefit when the clusters do not match the categories of interest to the user, or when a single conceptual category with documents ranking at the top of the output list satisfies the users' information need.

4.8.2 Clustering documents on the Web[24]

There is growing interest for clustering/categorising documents on the Web. This interest is aroused by the general dissatisfaction of users with the search results(Zamir&Etzioni, 98; Mackassy et al. 98; Shimamura et al. 98, Zamir et al. 97; Korpimies&Ukkinen, 98; Mechkour et al., 98). The overwhelming problem with the Web searches is low precision. The reasons underlying this problem are multiple:

- A typical user provides only one or two terms in his query and this gives little chance to a search engine to respond to the information need satisfactorily. Pinkerton(94) reports the average number of words in a query submitted to the WebCrawler to be 1.5. According to Kate Hagerty(98) of Northernlight, 50% of queries submitted to the Northern Light search engine have two or more words. This figure tells us that 50% of the queries had single terms, and the average number is at least 1.5.

- Although most search engines now use a more or less similar syntax, there is no standardisation as yet, and users may feel uneasy about using Boolean logic to narrow down their query. Coupled to the small number of terms in the user query, this hampers the search engines' ability to rank the results in a satisfactory way.

- Search engines use different criteria to rank the documents, therefore the quality of ranking and what a user can expect from it is varied.

Various approaches exist to improve the precision of web searches. Conceptually, these can be divided into four groups:

1. Providing the users with a wide range of tools to enable them to specify their queries precisely: users can use best match retrieval with Boolean logic in the same query by requiring certain terms, but entering others as optional. Some search engines also support specification of phrases, private names, location, and currency of the document, and position of the term in the document.
2. Categorising some or all documents of the collection into hierarchical subject groups to let users specify their queries by going down the subject hierarchy. The subject hierarchy is generally determined by human indexers, though the actual application to millions of documents may be via an automatic categorisation process. Groups are not mutually exclusive in terms of the documents they contain. Best known example of this approach is Yahoo, which categorises its whole collection. Yahoo encourages users to browse through the hierarchy, instead of entering a search statement at the highest level of the hierarchy.

[24] Statistics and references on this section are based on investigations made in April 1998.

While a comprehensive subject hierarchy may be invaluable with many typical search topics, it is probably not so functional in addressing specific user needs. For queries which cannot be easily answered via the subject hierarchy, Yahoo automatically sends the query to AltaVista search engine. Most other search engines create subject indexes for a limited number of topics of general interest. For example, InfoSeek provides 16 channels, each providing information for a specific topic area, examples are “automotive”, “health”, “business”, and the “the good life channel”. AltaVista and Hotbot provide a “browse by subject” option where subjects are hierarchically categorised into groups like “computers&internet”, “society&politics” and ”people&chat”.

3. Grouping the documents/terms from the search output based on their similarities, and displaying labels representing the subgroups (generally of documents) which the user can select and reject to narrow down his query. These applications differ in methods and criteria used in grouping, the function of grouping in the search session and the extent of dynamic processing involved. They are provided not as replacement but as optional to the relevance ranked list of documents.
4. Clustering a “representative” document set (that can be a specialist database for a certain area like agriculture or a subset of the collection), to form a pool that can be navigated through topics to build a precise query specification, before searching the web. There are no identified working examples of this approach on the Web yet, but publications about an ongoing project, “The WebCluster Project” at Robert Gordon University.

Of these different approaches, the third one is the most interesting in the context of this project. By definition, this approach brings to mind some kind of clustering algorithm being used to create the groups. However, among the search engines and one meta-crawler that provided output grouping functionality as of April 1998, it was either unlikely or unclear whether any document clustering algorithms were actually being used. In the cases of NorthernLight and Inference Find, site information seemed to be the dominant factor in grouping the documents; with AltaVista terms were being clustered rather than the documents, and there was no access to document subgroups represented by the terms.

It was also observed that, the use of the term “clustering” on the Web had been expanded to refer to any kind of grouping of the documents, irrespective of the method used. For example, Infoseek described the grouping of documents that lead to the same Web site as “clustering the URLs together into one result”. As another example, Kate Hagerty of NorthernLight suggested an unexpected use for clustering. While accepting that their subject folders are created and labelled based on categories that the retrieved documents have been assigned to at the time of inclusion into the collection, she stated that they use clustering algorithms to decide the order in which they show the folders.

A review of the methods employed by AltaVista, NorthernLight, Inference Find and HuskySearch is presented below. A project being developed by the NTTLabs is also briefly reviewed.

Unfortunately, the service providers have been secretive about the processes they use to group search outputs; for this reason the descriptions provided below had to be based on inferences that were reached through experimenting with the applications and the reading material provided on their web pages.

4.8.2.1 AltaVista's Cow9 Technology[25] - "refine" function

Cow9 is a collaborative effort between researchers at Digital Equipment Corporation and Ecole des Mines de Paris. It is activated by choosing the "refine" function after the search results are retrieved. The refine function generates a set of customised topics, groups of "closely related terms", that represent the query results. Users can choose to include or exclude each of these topics to refine their query.

The topics are represented in a list, each including a set of subtopics. Users are invited to include or exclude some of the topics or more specifically subtopics by clicking on this list. There is also a graph view, which shows a map of expandable topics, based on relations between the topics. Users with java-enabled browsers can include or exclude topics on this graph instead of on the list view. In the list view, topics are sorted in order of probable relevance; in the graph view, the relevance probability is indicated by yellow bars to the left of each topic.

After requiring/excluding some of the topics, user can rerun the original query which has now been narrowed down with his specifications; or he may choose to further "refine", getting a new list or graph of the topics based on his require/exclude choices.

Although it is said that each of the customised topics contains a set of related documents, there seems to be no functionality to select and view the documents under a certain topic. From the users' point of view, this makes Cow9 a term clustering method rather than a document clustering method.

The limited information AltaVista provides about the Cow9 technology is that it uses statistical analysis - not human interpretation - to sort search results into general topics. However, results of the experiments with the "refine" facility seem to be implying some kind of lexicon and possibly a subject index at work in corporation with statistical analysis. It is also possible

[25] Based on information provided on AltaVista web pages

that clustering is based solely on statistical analysis, but the analysis is based on a much more comprehensive document set than the search output.

As the developers have been unwilling to make any explanations, we can only rely on examples to gain a vague understanding of the methods they employ. Below are topics generated for 6 queries used to explore the Cow9 functionality:

Q1: For the query “byzantine” some of the categories were[26]:

- greeks, hellenic, hellenism
- ottoman, turkish, turks, turkey
- pope, papal, papacy, popes, infallibility, indulgences
- orthodox, church, liturgy, catholic, liturgical, orthodoxy, catholics, rite, liturgies
- archaelogical, excavations, archaelogy, excavated, excavation
- islamic, islam, muslims, muslim, caliph, ibn, caliphate
- empire, emperor, constantine, rome, emperors

Q2: For the query “london food” some of the categories were:

- ontario, toronto, mississauga, canadian
- food, retailing, sainsbury, foods, retailers, retail, supermarkets, tesco, consumer
- journal, bacteriology, quarterly, microbiology, toxicology, physiology, pharmacology
- sustainable, environmental, unsustainable, sustainability, sustainably, pollution, ecological, ecology
- bosnian, bosnia, serbs, serb, sarajevo, serbia, yugoslavia
- kenya, nairobi, unep, iucn, zimbabwe

Q3: For the query “okapi retrieval” some of the categories were:

- okapi, giraffe, giraffes, animals, necked, calves
- astrophysics, astronomical, extragalactic, astronomy, sternwarte, astrophysical, astronomico
- zoo, conservation, zoos, zoological, rhino, habitat, cincinnati
- mammals, primates, rodentia, marsupial, shrews, tamarins, chimpanzees

[26] The chosen topics are those that seemed to provide better representations as compared to other topics

Q4: For the query “trec okapi” some of the topics were:

- trec, retrieval, indexing, sigir, queries, acm, proceedings, documents, query
- okapi, giraffe, zoo, gorillas, zaire, davic, conservation, talisman, giraffes
- estate, texas, real, broker, seller, license, buyer, disclosure, disclose
- croft, inquery, tipster, ciir, callan
- telescope, cgro, telescopes, astronomers, comptel

Q5: For the query “hair color” some of the categories were:

- scalp, shampoo, conditioner, shampoos, thinning, baldness, conditioners
- dress, dresses, outfit, wardrobe, gown, bride, wedding, blouses
- rooms, suites, concierge
- cosmetics, nails, nail, cosmetic, fragrances, mascara, skincare, haircare, lipsticks
- smiled, sighed, leaned, nodded, glanced, grinned, shrugged, glared, gestured

As seen in these examples, variants of a word stem and words that have close relationships are generally grouped together. There are also topics composed of only verbs or nouns forming quite homogeneous sets that lack the random element characteristic of textual documents.

It seems the extent of such grouping is somewhat beyond what could be achieved by purely statistical analysis that is based on these terms’ co-occurrences in the retrieved documents, or the collection. It is also striking that in the fifth query, the last topic representation is solely based on past tense verbs.

In most queries, there are also a few topics that seem to be formed more arbitrarily, in contrast to the listed topics. Examples are:

For Q1: “byzantine”

- istanbul, birdhouse, tmbg

For Q2: “london food”

- knowhere, pukka, cringing, hallowed, skateboard, entertainers, tweak, mates

For Q5: "hair color"

- favorite, waist, shoe, hips, astrological, monogamous, jewelery

A shortcoming of the method observed with Q3:"okapi retrieval" was that it did not seem sensitive to the ranking of the documents. All topics generated for this query by "refine" function were related to zoology/animals or astronomy/astrophysics although in the ranked list the first 19 documents with highest relevance ranks were about Okapi search engine. This situation did not change when the query was rerun as "okapi retrieval search engine".

When the query was defined as:

Q4: "trec okapi"

some topics focusing on IR came up together with others focusing on astronomy and zoology. Interestingly the term "okapi" again only occurred in a zoological context in topic representations.

Finally several queries with Turkish terms were tried out to test the claim that the method was not language dependent. With these queries, a majority of the topics seemed quite arbitrary. For example:

For Q6: "ankara, mart" {ankara:capital city of Turkey, March} some of the categories were:

- icin, bir, buyuk, uzere, butun, olan, olarak, icinde, veya {for, one, big, just about to, whole, is, being, in, or}
- onemli, cok, zamanda, guzel, daha, soru, ancak, ortak {important, very, in time, beautiful, more, question, but, partner}
- yeni, ikinci, birinci {new, second, first}
- gibi, olacak, insan { like, will be, man}
- konusunda, sonunda, sadece, gereken {about this topic, in the end, only, needed}

The difference in performance with English and Turkish documents may be an indicator of a treatment for the English documents that is lacking in Turkish documents. It may also be due to the small size of Turkish documents on the Web as compared to the English documents, that makes the collection statistics less reliable for the Turkish documents (assuming that the collection statistics are being used in grouping the output results).

4.8.2.2 NorthernLight's Custom Search Folders[27]

NorthernLight's application has more resemblance to a document clustering approach, in that it provides groups of similar documents, but it seems unlikely that any clustering algorithms are employed in the process.

When a query returns more than 25 documents, together with a list of the top ten documents retrieved as the result of a query, NorthernLight also shows a set of folders on the left side of the screen, each of which includes a group of documents with similar characteristics. Each folder is of one of the four different kinds of folders described below:

- subject (e.g. hypertension, baseball, camping, expert systems, desserts)
- type (e.g. press release, product reviews, maps, resumes, recipes)
- source (e.g. commercial Web sites, personal pages, magazines, encyclopedias, databases)
- Language (e.g. English, German, French, Spanish)

Documents not assigned to any folders are grouped under the folder "all others". Folders are divided into subfolders as the user navigates to the next level down, and the user can navigate up and down to find out the document subgroups that he might want to view in detail. Folders are not mutually exclusive. Documents can be classified into multiple folders based on their characteristics, they may also be assigned to multiple subject folders.

The subject folders represent the most important category in the context of this project. These folders are based on a hierarchy of over 200,000 subject terms created by NorthernLight's librarians. Search results are automatically grouped using this manually created hierarchy. This hierarchy is also used in indexing the documents as a company official[28], has been reported[29] to say that new documents are assigned to categories as they are loaded.

The folders are represented as links, with brief titles. The type of the folder is not explicitly displayed to the user, although it can generally be inferred from the title. In the case of subject folders, a suitable topic from the subject hierarchy is probably being used as the title.

As mentioned previously, NorthernLight's application does not seem to involve any clustering process; the hierarchy of 200,000 subject terms created by human indexers helps creation of "sensible" document folders with "sensible" titles; in that sense, it is not very much different from Yahoo's subject index.

[27] Based on information and product reviews provided on NorthernLight's web pages.
[28] Joyce Ward, director on content classification in Northern Light
[29] Source: "Vendors future: NorthernLight - Delivering high-quality content to a large internet audience", Richard Wiggins, System Architect, NEM Online

NorthernLight claims that the folders are unique for each search, but their creation and labeling based on the pre-defined criteria limits the extent to which the process can adopt to the user's situation and makes it less of a dynamic process (Kate Hagerty of NorthernLight accepts that the subject folders reflect categories that the retrieved documents have been assigned to at the time of joining the collection).

In order to explore the distribution of the documents among folders, five different queries[30] were tried out with the NorthernLight search engine. At the highest level of the hierarchy, the proportion of documents in:

- subject folders ranged between 0.5% and 7.1%, with an average of 3.7%,
- source folders ranged between 27% and 73%, with an average of 50%,
- type folders were available for only Q1 and Q4 , 1% and 4% respectively,
- no language folders were available,
- "all others" category ranged between 31% and 78%, with an average of 52%

In the initial folder lists representing the whole output set, the average share of subject folders was small as compared to the source folders and "all others" categories. The overlap of documents between folders were between 0% and 11%, with an average of 5%.

It was interesting that for the query "okapi retrieval", "city.ac.uk" sites were presented as 3 different folders (ftp, web.soi, web.cs), whilst there were 89 documents that were placed into the folder "all others".

4.8.2.3 Inference Find (meta-crawler)[31]

Inference Find works by passing the queries to a number of search engines[32] in parallel, merging the results, removing the redundancies and clustering the results into 'neat' 'understandable' groupings. Engines returning fewer documents(e.g. 10 at a time) are called more than once to increase the number of documents retrieved.

Once the documents are retrieved, Inference Find clusters the related items together. The clusters are represented by a title which can be a subject or a site, and a list of links that are document titles.

[30] Q1:"hair color", Q2:"constantinople", Q3: "okapi retrieval", Q4: "funding university education in UK", Q5: "cost of living in tokyo"

[31] Based on information provided on Inference Find's pages as of April 1998: http://Web.inference.com/infind/about.html

Generally a majority of cluster titles refer to sites rather than subjects. Site information in cluster titles may refer to :

- a certain organisation: e.g. "City Site", for documents on "okapi retrieval" found at City University;
- a geographic location: e.g. "Misc. Asian Sites";
- a type of location: e.g. "Misc. Non-Profit Sites" or
- a combination of them: e.g. "Misc. European Educational Institution Sites."

The links under each title do not necessarily appear to be topically related. For example for the query "okapi retrieval", under the cluster title "Misc. Educational Institution Sites" were the links titled:

- Cello: an Internet Browser
- German-Russian Geneology
- High-Performance Knowledge Retrieval
- Onion Patch(sm): New Age Public Access Systems
- STO's Internet Patent Search System
- UCR High Energy Astrophysics
- World Heritage Sites in Zaire

It was also interesting that although City University was listed as the first cluster with 8 links to SOI pages, the link "Professor Stephen Robertson" was listed in a separate cluster with the title "Misc. European Educational Sites".

4.8.2.4 HuskySearch

This is a prototype search engine developed by Zamir&Etzioni(98) from the University of Washington. It is based on an incremental, linear time algorithm called Suffix Tree Clustering(STC) which creates clusters based on phrases shared between documents. Various studies have shown that phrases improve precision without hurting recall. STC makes use of proximity information between words by treating a document as a string.

STC starts by transforming documents to text strings by use of a light stemming algorithm. A set of base clusters are then identified in a way that can be described as the creation of an inverted index of phrases using a data structure called a suffix tree. Finally, base clusters with a high overlap in their document sets are merged using the equivalent of a single-link

[32] WebCrawler, Yahoo, Lycos, AltaVista, InfoSeek and Excite as of April 1998

clustering algorithm where a pre-determined minimum similarity betweeen base clusters serves as the halting criterion.

Zamir&Etzioni have conducted some experiments to evaluate their method. They have defined 10 query topics, generated collections of 200 snippets and documents for each query, and made relevance judgments for each of these documents. They have then clustered the documents using six different clustering methods: STC, Single-Pass, K-Means, Buckshot, Fractionation, and GAHC. Best clusters were chosen for each method up to the 10% of the result set. The STC algorithm generated higher precision values than all other methods and also the ranked lists.

According to the authors, the success of STC is due to the use of phrases and also that it allows for overlapping clusters. The precision values for all 6 methods are slightly lower when snippets are used instead of full documents, however this may be acceptable as using snippets considerably reduces the computing resource requirements.

In order to test this algorithm several queries were run with the prototype provided on the Web. The method always generated 20 clusters regardless of the number of documents retrieved. The overlap between clusters was generally around 90-100%. In some queries the overlap between shared phrases(/terms) was also high and this made it difficult to differentiate between clusters. For example, for the query "telework employee internet", "telecommuting" appeared on 9 of the 20 clusters. However, in queries where the algorithm has succeeded in extracting good numbers of phrases, the results seemed meaningful.

Authors have also tried using multi-word phrases as additional attributes on the performance of the vector-based algorithms, however, their comparisons gave mixed results.

Prototype of this algorithm is available at Web.cs.washington.edu/research/clustering.

In a previous paper related to the suffix-tree clustering, Zamir&Etzioni (97) describe a method for evaluating the quality of clustering, the *Global Quality Function* (GQF).

GHQ is used to generate a score for each cluster and also for the clustering solution. The score *s(c)* of a cluster is described as the product of its size and its "dampened cohesion". The cohesion of a cluster is the number of words common to all the documents in the cluster. The score of a singleton cluster is zero.

For a clustering solution, the GQF is a product of three components:

1. $f(C)$ – a function proportional to the fraction of documents in non-singleton clusters (in experiments, the ratio of the number of documents in non-singleton clusters to the overall number of documents has been used),
2. $1/g(|C|)$ a function proportional to the number of non-singleton clusters (in experiments the number of non-singleton clusters raised to the power of 0.5 has been used) and
3. $\sum_{c \in C} s(c)$ the sum of the scores of all clusters in the clustering, giving the formula:

$$GQF(C) = \frac{f(C)}{g(|C|)} \sum_{c \in C} s(c)$$

The GQF is expected to provide a trade-off between having a small number of large clusters of low cohesion versus having many small clusters of high cohesion. Zamir&Etzioni use the GQF as the heuristic to guide the hierarchical clustering algorithm and as the halting criterion. At each iteration of the algorithm, the two clusters whose union would result in the highest increase in the GQF is merged and the algorithm terminates when no merge increases the GQF.

It would be interesting to see whether this measure can also be used to evaluate a clustering solution during a search session. Such a measure could be employed to decide whether it is worth presenting the clustered output for the user's viewing, or whether the user would be better off with a ranked list. This would require collection specific benchmarks for low and high scores.

4.8.2.5 Human performance on clustering Web pages

Macskassy et al's (98) user study focuses on how users group documents from search outputs; assessing the similarities and differences between users' behaviour in grouping documents is important to appreciate whether web clustering implementations can actually create meaningful and effective clusters.

Ten subjects participated in the study, each separately clustering the results of five most frequent queries directed to the Rutgers webWatcher search engine, which indexes all Rutgers University web pages. Each of the queries returned between 10-16 web links.

It is probably not a big surprise for any IR researcher that any two subjects participating in this study had little similarity in their clusters.

Interestingly, users have created rather small clusters, and those who had access to the full text of the documents created more clusters than those who only saw the URLs. With the full texts, the overlap between clusters and the percentage of documents clustered were also higher.

Another important finding was that, on average, more than a third of the documents were not grouped with any other documents.

The small number of users and documents in this study limit the generalisability of the results. The authors acknowledge that it is necessary to conduct another experiment with higher numbers of users and documents, where a second group of subjects can also be incorporated in the study to evaluate the clusters created by the first group. It is important to see whether this second group would assess all clustering solutions created by different groups as acceptable, or if none of these solutions would be judged appropriate by all subjects.

4.8.2.6 NTTLabs - Customizable multi-engine search tool with clustering[33]

This project combines two ideas to create a comprehensive search tool. The first is the distribution of indexing load to specialised search engines in different geographic locations. The second is helping the user to refine his query by:

- classifying search results to provide the user an overview of the output and get his feedback on clusters to refine the query,
- finding out words related to those extracted from the search results, which can help the user to identify the domain of his query.

A hierarchical clustering method is used to group the documents that are represented by a combination of their “characteristics of hypertexts”, title and heading words, anchors, list items, and formatted text(bold, italic). In addition to keywords, common IP addresses and links between documents are perceived to be good indicators of association and used for initial assignment of documents. The method and measures used in initial assignment are not specified.

[33] Based on 6th International WEB Conference, Hyper Proceedings

In order to provide partial results to the user as documents are retrieved, an iterative approach is used to cluster documents retrieved in different phases. After the initial clustering of documents, the clusters are iteratively grouped according to their similarities to construct a hierarchical structure. Similarity between cluster c(i) and c(j) is calculated as:

$$R(c(i),c(j)) = \sum TF(i,t) * TF(j,t) * \left(\frac{DF(t,q)}{QF(t)} \right)$$

where

$TF(i,t)$ is the relative frequency of term *t* in cluster c(i)

$DF(t,q)$ is the document frequency of term t in the whole collection of query q,

$QF(t)$ is the frequency of term t appeared in queries so far.

$DF(t,q)$ and $QF(t)$ values are updated after user feedback: for terms from topics selected by the user, $DF(t,q)$ is doubled to increase the term's weight and for terms that are not found in topics selected by the user, $QF(t)$ is doubled to reduce the term's weight.

To customise the search tool for individual users, query terms and terms identified as related terms after user's feedback are stored as a concept space.

The prototype of this system is available at: http://ails4.csie.ntu.edu.tw8080/iragent.html

4.8.2.7 WebCluster Project – Robert Gordon University

Harper et al.'s (99) study seeks to mediate access to the Web documents through query formulation in a 'relatively small, well structured, pre-clustered collection covering a particular subject domain'. Authors name such collections as source collections and assume them to be up-to-date, comprehensive and homogeneous in content.

The intention is to reveal to the user the semantic and topical structure of the domain by means of cluster representations. The user is asked to choose documents/clusters of interest, which are processed by the system to formulate a query to be submitted to the target collection, e.g. a subcollection of the Web, indexed by a search engine. Users are allowed to edit the query, although it has been observed to result in loss of 'good' terms as well as 'bad' terms for query formulation.

The process is akin to a routing task, and is expected to help users especially when they have vague information needs or when they need to access information in an unfamiliar domain.

Harper et al. are in the process of deciding various possible ways of access, methods of clustering and document and cluster representation. They have conducted small scale tests to gather data that can help to make various implementation decisions before proceeding with a more extensive evaluative study to test their hypothesis that cluster-based mediated retrieval can be more effective than traditional retrieval in an interactive setting.

As a result of these experiments they have found that most users felt comfortable with the idea of mediated access, although they seemed to prefer generating their own queries when they were familiar to the topic. But even in such cases, users appreciated the improvement in the results brought about by mediated access. Users often needed to view sample documents from clusters as the cluster representations were not adequately informative.

The query topics used in these experiments were selected by the authors from a collection of Reuters articles. The queries forwarded to the Web resulted in a high proportion of news articles, which is interpreted by the authors as evidence that different characteristics of documents such as the structure, the genre, the language, the formatting style may be acting as features in clustering. This explanation is somewhat dubious as it is likely that the chosen query topics were quite typical news items probably not discussed in many other types of documents. But if they were actually discussed in other types of documents, it would be rather alarming that the query sent to the target system had been only capable of retrieving documents of the type that are already available in the host system. This issue actually brings to the surface a crucial weakness of this study; that is the unrealistic assumption that the small, domain-specific collection will actually be representative of a huge heterogeneous collection such as the Web.

4.8.2.8 NEC's Domain cluster interface for Web search

Shimamura et al.s (98) domain cluster idea involves grouping documents from a search output based on their URL domain names. The method is implemented on NEC's Intranet, which includes over 50,000 HTML documents.

When a query is defined, matching documents are grouped with respect to their URL's. An organization name database which stores pairs of domain names and the corresponding organization names is used to find the appropriate name for each URL group. The groups created are labelled with the organisation names corresponding to the URLs and displayed to the users in a hierarchical view that reflects the organisational structure of their company.

Researchers assume that grouping based on the URLs can help users to find relevant documents in all types of search situations, and claim that using such a method eliminates the difficulties involved in browsing large sets of output sets. It is assumed that documents relevant to a user's query will be concentrated in one or two clusters, which users will recognise easily based on the organisation where the document has been created, and thus will be able to instantly reduce the output set to a fraction of the total number of retrieved documents.

Authors recognise that such an approach can be useful only if the users have some knowledge of the organizations, and for an Internet-wide application the organization name database would be more difficult to maintain (it would also be less meaningful for the users).

PART II. DESIGN AND IMPLEMENTATION

In this section, various design and implementation decisions taken at different stages of the project are summarised and discussed.

Chapter 5 discusses the various design aspects that needed to be considered in the implementation, and the decisions regarding these aspects. Decisions regarding some of the these aspects were uniform among different experiments throughout the project. For some other aspects there were variations between the parameters used, or methods adopted among different experiment sets. The reasons for such variation are outlined in detail in the following sections.

It will be useful here to briefly introduce the groups of experiments undertaken in this project, as the design and implementation chapters often refer to a particular experiment group to explain certain decisions.

The two main groups of experiments were user experiments with the Inspec database and experiments with the TREC interactive queries for 1996. User experiments tried to assess whether clustering was capable of helping users find relevant documents more easily as compared to relevance ranking. After some tuning and testing of the implementation, a group of 20 user experiments were undertaken to compare the precision values of the first and first two clusters chosen by the users against the precision values of ranked lists at comparable thresholds. This group is referred to as the 1st group of user experiments.

The results of these experiments suggested that the implementation and the experimental setup could be improved in several ways. The implementation was modified accordingly and an intermediate group of 10 experiments were realised to compare two alternative cluster representation formats that had been developed. This group is referred to as the Intermediary user experiments.

Finally a set of 55 user experiments were carried out in similar setup to the 1st group of user experiments, with the purpose of generating adequate data to assess our hypotheses. This group of experiments is referred to as the 2nd group of user experiments.

TREC experiments were realised without any direct user involvement. These experiments aimed to exploit the facet evaluations that are provided for the TREC interactive queries each year, based on expert judgments. The experiments were based on 7 query topics from the 1996 experiments.

The first group of TREC experiments were based on the set of relevant documents for each query. Most or all relevant documents were clustered and the correlation between documents facet and cluster assignments was investigated. This group of experiments are referred to as TREC-REL.

It was also of interest to find out if a similar correlation would exist when we clustered the documents that were retrieved in response to users' query statements. For this purpose, the initial query statements of City participants were run and the top 50 documents retrieved were clustered. This group of experiments are referred to as TREC-USER.

Finally, we looked at whether users could benefit from basing query expansion on chosen clusters rather than relevant documents. Using the clusters generated for user queries in the previous experiment group(TREC-USER), we expanded users' queries on each cluster and investigated whether these inspections would help the user to retrieve more relevant documents with the same facet assignment as found in the documents within the clusters. This group of experiments are referred to as TREC-EXP. Each of these experiments are reported in detail in Part 3 Experiments section.

Chapter 6 includes the research question/hypotheses. Chapter 7 and 8 summarise the development phase; the prototype development in VB and the final implementation in TCL.

All programs used in this implementation were written by the student, with some support from Okapi programmer Thien Do, where necessary. This provided flexibility in modifying programs for different experiments, but also brought some inefficiencies during the TREC experiments that would not be experienced had the programs been written by a C++ programmer. The student depended on data created by Okapi's C++ programs, and she was not able to interact directly with the Okapi search engine. For example, in the TREC experiments, it had been necessary to run comprehensive queries to retrieve the set of relevant documents instead of directly creating a file that includes all relevant documents. This was because document identifiers were not separately indexed in the TREC database, and it was also not possible to use the document identifiers as query terms as the Enquire interface was designed to accept up to 20 search terms. As aresult some unnecessary time had to be spent in trying to retrieve the relevant set, and in some cases a few of the relevant documents were missed despite all efforts.

5. Using cluster analysis in this project

Until recent years, studies on clustering search outputs have been far less common than studies on clustering document collections in the IR literature. However, in the last few years, there has been a renewal of interest in using clustering in the IR field, and frequently with the purpose of grouping and presenting search outputs to the users' evaluation.

Collection clustering and output clustering are differentiated both by their purpose and computational requirements. Collections are clustered for retrieval purposes whereas output clustering is attempted as a tool for the next step, i.e. the process of user's viewing/browsing the documents regardless of how they were retrieved. Although output clustering involves much less numbers of documents, it still requires efficient algorithms in order to be able to support interactive sessions.

It can be argued that both processes seek to discriminate relevant documents from irrelevant ones. However, this is not exactly true. Even when a majority of the documents from an output are relevant, clustering may be able to help the user to digest/analyse them in a more systematic and intuitive manner by creating distinct groups that represent different facets of the query.

So far, there have been only a few studies that have focused on clustering search outputs; at the beginning of this project, Allen, Obry and Litzman's(93) interface for navigating clustered document sets (see Section 4.8.1.7 on page 62) Xerox's Scatter-Gather method (see Section 4.8.1.1 on page 53) and Kirriemuir&Willett(95)'s clustering to identify document duplicates (see Section 4.8.1.3 on page 59) were the prominent examples. In these studies, workers confined themselves to evaluating the methods with user relevance feedback without focusing on the user needs, preferences or requirements arising in such a process. More recently, several Web search engines have implemented various methods to group search outputs(see Section 4.8.2 on page 63). However, these methods are generally kept as trade secrets and there are no reports available regarding their effectiveness or usability.

With the literature on clustering search outputs being far from abundant, it has been necessary in this project to adopt an exploratory approach and turn to IR/CBR applications for illumination where relevant.

5.1 Choice of method

The wide variety of clustering methods that could be employed for clustering search outputs made it infeasible to make comprehensive evaluations for each possible method. It was more practical to focus on the methods that had some history of being successfully used for document clustering.

The main requirements for this implementation were satisfying the theoretical soundness criteria (see Section 4.5 on page 41), efficiency, and intuitiveness. Our focus was on search outputs of around 50 documents. This figure was large enough to make it worthwhile to apply clustering, and small enough to cluster without excessive information reduction. The review of the cluster analysis methods and the CBR literature had led us to believe that a non-hierarchical method would be more appropriate than a hierarchical method for the task at hand (see Section 4.5.1 on page 43 and Section 4.5.2 on page 46). An additional reason for preferring a non-hierarchical method was that we simply did not need a hierarchy with only about 50 documents.

Since the majority of the methods so far used in document clustering were hierarchical methods, the range of methods that were to be considered was greatly reduced. Among the non-hierarchical methods, Can&Ozkarahan's(90) C^3M algorithm was found to be intuitive and relatively simple. It conformed to the theoretical soundness criteria, and provided a means of estimating the number of clusters as well as identifying cluster centroids and forming clusters around the centroids. The algorithm was of partitioning type, but it was also possible to convert it to the overlapping type if required. It also had a history of good performance with the Inspec database that was available in the department for use in IR experiments.

5.1.1 Can&Ozkarahan's cover-coefficient-based clustering (C^3M)

Can&Ozkarahan(90) have proposed a novel but simple clustering method for document clustering. In this method, the relationship between documents is represented in a document-by-document matrix C where entries are computed through a two-stage probabilistic experiment. Each entry, c_{ij}, in the matrix shows the probability of selecting any term of d_i from d_j for a document pair. The measure of association is asymmetric. Off-diagonal entries represent the extent to which d_i is covered by d_j (coupling of d_i by d_j) and diagonal entries represent the extent to which d_i is covered by itself (decoupling of d_i from the rest of the documents).

The sum of off-diagonal entries of the *i*th row indicates the extent of coupling of d_i with the other documents of the set and is referred to as the coupling coefficient (Ψi) of d_i. The extent to which a document is covered by itself is called the decoupling coefficient (δí) of di and is equal to 1-Ψí. δí takes higher values when di contains fewer terms that are common with other documents.

Authors have hypothesised that the number of clusters within a database should be high if individual documents are dissimilar and low otherwise. A database with similar documents will have a low sum of δ's and few clusters and a database with dissimilar documents will have a high sum of δ's and many clusters.

They have proposed an estimate of the number of clusters comprising the sum of δ (decoupling coefficient) values:

$$n_c = \sum_{i=1}^{m} \delta_i$$

They reported that, in their experiments with binary indexing, the nc values have generated very close values to the estimate of the average number of clusters computed as the ratio of the size of the indexing vocabulary to average indexing exhaustivity. The experiments with weighted indexing showed 19% greater values than the estimated nc.

As far as the selection of centroids(cluster representatives) are concerned, Can&Ozkarahan have argued that the documents to be chosen as centroids must be neither too general(containing many terms) or too specific(containing a few terms). To satisfy these constraints they introduced the seed power Pi:

$$P_i = \delta_i \times \psi_i \times \sum_{j=1}^{n} d_{ij} \quad \text{for binary D matrix}$$

$$P_i = \delta_i \times \psi_i \times \sum_{j=1}^{n} \left(d_{ij} \times \delta_j' \times \psi_j'\right) \quad \text{for weighted D matrix}$$

where

δ_i represents the intercluster dispersion

ψ_i represents the intracluster cohesion

and $\sum$ provides the normalisation.

δ_j', ψ_j' are the corresponding values of the term - to - term matrix.

The first nc documents with the highest seed power are selected as centroids. There is a possibility of having two very similar documents both with high seed powers but according to Can&Ozkarahan such documents can be identified easily as they would have very close seed power values. Can&Ozkarahan suggest grouping the candidate documents that are in close range of similarity and checking for each couple from the group that:

$cii - cjj > \varepsilon$

$cii - cij > \varepsilon$

$cjj - cji > \varepsilon$

$cij - cji > \varepsilon$

where ε is a small real constant - set as 0.001 in Can&Ozkarahan's experiments with the Inspec database (this solution did not work well in this project, and a different approach was adopted to avoid very similar documents from becoming seeds simultaneously, see Section 5.3.2 on page 92 for details).

Once the cluster centroids are determined, clustering proceeds as follows:

1. for each d_i (if it is not a centroid, find the cluster seed that maximally covers d_i, if more than one centroid meets this condition, assign d_i to the cluster whose seed power value is the greatest,
2. put the documents which do not find any centroids into a ragbag cluster.

It is possible to adapt the algorithm to allow for overlapping by relaxing the rules of step 1. We may choose to assign a document to all centroids that maximally cover the document or to all centroids that meet a certain coverage threshold.

Can&Ozkarahan's tests for cluster validity have produced significant results confirming its validity. CBR performance of C^3M was also compared with full-text retrieval and several other clustering algorithms. It outperformed the results obtained by Hamdouchi& Willett (89) for the hierarchical algorithms single-link, complete linkage, average link and the Ward's method, and the results obtained by Voorhees(86) for single-link, average link and complete linkage. However, the comparability is limited as these researchers did not use the same matching functions, and the results could be different with different matching functions.

5.2 Choice of database

We needed a database that would be appropriate for user experiments. It was also necessary to have the database on a search engine that would relevance rank the retrieved documents for comparison with the clustered output. The availability of Inspec database on the department's experimental probabilistic retrieval system Okapi fulfilled both of these requirements for the user experiments.

In addition to the Inspec database, the news database of the 1996 TREC Interactive track was used for experiments aimed at assessing the capability of the algorithm to discriminate between different facets of an output set. TREC data was especially pertinent to this project as it provided queries and sets of relevant documents for these queries together with facet assignments for each relevant document.

5.2.1 Inspec database

The Inspec database focuses on worldwide literature on physics, electronics and electrical engineering, computers and control and information technology. Its coverage of the information retrieval and information technology subject areas allowed us to target students from the department's MSc courses (Information Science and Information Systems and Technology courses) for the user experiments.

One drawback of the database was that the records were confined to the 1989-92 period. This meant that it would not fulfil a user's information need satisfactorily; users would still need to search the database for more recent documents by other means. This probably reduced the experiments' attraction for prospective participants. It was also not suitable for some of the queries where participants needed up-to-date information about leading-edge technologies and applications.

Fortunately, before the 2nd group of the experiments, the database providers agreed to supply an update that covered the period 1993 to end of 1996. This improved the general user satisfaction with the search results.

5.2.2 TREC (Text REtrieval Conference)

TREC is an international project led by the US based National Institute of Standards and Technology (NIST) with the stated goals of:

1. Encouraging research in text retrieval based on large scale text collections,

2. Increasing communication among industry, academia and government by creating an open forum for exchange of research ideas,
3. Speeding the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real world problems,
4. Increasing the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

TREC is the major experimental programme of recent years (Robertson et al.; 97a). The number of participating systems has grown from 25 in 1992 to 36 in 1995, including most of the major text retrieval software companies and most of the universities doing research in text retrieval (Harman, 96).

The conference involves testing and comparing information retrieval systems under various experimental conditions. Every year, since 1992, NIST specifies a number of search topics and distributes them to various participating academic and commercial institutions from around the world. NIST also provides a large quantity of textual data to be used for the searches. Participants then undertake a series of searches and return the resulting output to NIST for relevance assessment and performance evaluation.

Each topic is developed from a genuine need for information brought in by the assessors. Each assessor constructs his/her own topics from some initial statements of interest and performs all the relevance assessments on these topics (Harman, 96).

Topics describe an information need with many aspects[34] – an aspect being roughly one of many possible answers to a question which the topic in effect posed. For the interactive track assessors make two sorts of assessments: the traditional relevance assessment and the aspectual assessment. For each topic, the aspect assessor (who has –normally- also performed the relevance assessment) is asked to read each of the documents to be judged and gradually:

1. create a list of the aspects found somewhere in the documents,
2. select and record a short phrase describing each aspect found,
3. determine which documents contain which aspects,
4. bracket each aspect in the text of the document in which it was found(Over, 98).

Type of information required varies from topic to topic. On one topic it may be instances of a certain event, on another it may be examples of certain sorts of people, places or things. For

[34] The word 'aspect' has been replaced by the word 'instance' as of TREC-7 (1999).

example, for a query on “the available treatments for high blood pressure”, aspects (instances) might be:

- use of calcium,
- regular exercise,
- biofeedback,
- the Mediterranean diet.

Relevance judgements precede the aspectual judgements, and it is possible to have a relevant document that does not cover any aspects. It is also likely that a document will cover multiple aspects.

The pooling of the outputs makes the experimental collection usable for other experiments after the event. As a new search on one of the topics is unlikely to yield relevant documents that have not been assessed, TREC collections are deemed usable for experiments outside the framework of the TREC programme and “constitute by a long way the closest approach yet to the ‘ideal’ test collection envisaged by the UK research community in the 1970’s” (Robertson et al., 97a).

The database distributed by NIST is generally a news database that contains documents which show a much higher variation in size than the Inspec documents, its vocabulary is also much less controlled.

5.2.2.1 TREC queries and the usability of the TREC data for evaluating the implementation

TREC queries aim to simulate real life information need situations. They generally require answers to specific questions rather than topical information on a certain area. This means that users need to form rich sets of query terms that are capable of differentiating not only between topical and non-topical documents, but also among the topical documents that are capable of answering the given question and others that discuss the topic from a different perspective or level of detail.

Results from interactive (user-based) experiments are generally poorer than results from the other tracks that do not involve users. This is not surprising when we consider that users typically initialise their queries with no more than a couple of terms. This means they generally need to make much iteration to reach the term sets capable of satisfying TREC queries. This can be a frustrating experience, given the time constraints and the possible necessity to not only find the optimal term set to express the given requirement, but also to steer the query in various directions to combine information about different facets of the query.

TREC queries can also be ambiguous. For example, one of the 1996 topics (no:256) was:

> *"Negative reactions to reduced requirements for college undergraduate core studies". "To be relevant, a document will provide negative opinion/facts concerning the fact that colleges have reduced their basic requirements for the granting of degrees to undergraduates."*

Some of the documents[35] that were deemed relevant to this topic by TREC experts discussed changes made in required subjects to include non-western, ethnic, 3rd World, and women's studies, in order to reflect the multicultural student bodies of the colleges and make the curriculum more responsive to the changing interests of society. They did not mention any reductions in requirements or lowering of standards, but criticised these changes as attacks on the Western civilisation and liberal education principles. It appears that, in accepting these documents as relevant, TREC experts focused only on the "negative opinion/fact" part of the query and neglected the part about "the fact that colleges have reduced their basic requirements for the granting of degrees to undergraduates". It requires a certain line of reasoning or a priori knowledge to conclude that these documents are relevant to the given query. Certainly, the proponents of these changes would not agree with TREC experts that these documents were relevant.

With a few other documents, it was even more difficult to understand why the TREC experts judged them relevant. One such document discussed that top-rated colleges should not be forced to focus on undergraduate teaching for financial reasons, but be able to utilise their strengths in research[36]. It had no mention of reductions or requirements at all. Another document focused on the details of current and planned MBA programs in Oxford and Cambridge universities, again with no mention of requirements or any changes regarding them[37]. A third discussed affirmative action in the work place, and quoted a white worker saying "We were told if you went to college you could write your own ticket. But affirmative action in [the] workplace has lowered standards to the point where education counts against you"[38]. Here again, the link is quite inexplicit. Finally, a document inviting applications for FIPSE grants was deemed relevant probably because it invited proposals that "improve the quality of undergraduate education by raising academic standards"[39].

Actually, with this query, ambiguity stems from the documents deemed relevant rather than the query itself. Similar ambiguity also exists with topic 264. While the topic is defined as:

[35] WSJ920323-0128, WSJ910924-0104, AP880912-0024, WSJ910731-0098, WSJ910423-0160, AP880827-0026, AP880418-0188, AP880401-0125.

[36] FT931-15789

[37] FT922-13879

[38] WSJ911111-0102

[39] FR88812-0046

> *"U.S. Citizens in Foreign Jails"*
> *"With international travel becoming more commonplace, many travellers, whether due to ignorance of the law or with criminal intent, are ending up in foreign jails. Of interest is the country involved, the reason for the seizure, and jailing, and the sentence, if convicted. U.S. intervention, if any, would also be of interest."*

several of the documents deemed relevant were about American hostages kidnapped and held by groups such as the Islamic militias in Lebanon[40]. These documents have all been assigned the aspect "American hostages" but the relevance of this aspect to the given topic was not explicit.

TREC organisers have improved facet evaluation procedures to minimise such ambiguities in later years and although this project used the 1996 data, the experiments were not directly affected by these peculiarities.

These peculiarities have been mentioned here as they reveal the subjectivity involved which is almost impossible to avoid in a text categorisation process. This situation cannot be assumed to be a oddity of the TREC experiments. Within any document set, there will be various legitimate meanings/facets depending on different points of view, and no solution ignoring such multiplicity can be deemed valid.

However, there are a few other issues that limit the usability of TREC data to evaluate clusters. Firstly, TREC facet evaluations do not intend to answer the question of what a document is about, but assess which fact or pre-defined aspect about a question is mentioned in the document. For example, the description of the query topic 255 is:

> *"Name countries that do not practice or ignore environmental protective measures."*

Among the set of relevant documents, there may be three documents discussing lack of measures to control pollution caused by passenger cars. From the perspective of clustering, It could be desirable to see these three documents within the same cluster. However, if these documents each mention a different country, the TREC experts would assign them to different facets, regardless of their high level of contextual similarity. Conversely, documents discussing different aspects of environmental pollution would all be assigned the same facet if they all mentioned the same country name.

Another limitation of using the TREC facet information for evaluating the implementation was that, for some topics, there were a considerable number of relevant documents that were not assigned to any aspects. The proportion ranged between 9% and 74% with four of the 12 queries having more than half of their relevant documents without any facet assignments. Those documents, and any others that were assigned to an aspect with no other document being assigned to the same aspect, would not contribute to the evaluation of clusters. There were also topics in which a considerable number of documents were assigned a combination of two or more aspects (in four of the 12 queries, more than half of the documents with facet evaluations had more than one facet assignment, the average figure for all 12 queries was 37%), and in such cases it was not evident how they should have been distributed among clusters. While it was possible to cluster the documents first based on these aspects and then to compare them with clusters formed based on the term occurrences, the value of such a practice would be questionable as clustering results would be validated with respect to another set of clustering results, instead of a reference source independent of clustering.

Finally, related again to the multiplicity of facets that could be covered in a document, was the consequence that "cluster-ability" of the documents was limited. TREC documents can be rather long and some of them can be referred to as composite documents, as they discuss various different issues. While discussion of more than one facet in a document does not necessarily make it a 'composite' document in other contexts, it is significant for this project as the implementation is expected to discriminate documents at facet level and evaluated against the facet assignments.

The clustering implementation was designed to assign a document to more than one cluster if it is equally (maximally) covered by more than one cluster seed. To cater for the multiplicity of facets that could be covered by TREC documents, the algorithm could be further relaxed to assign documents to all cluster seeds that covered them to an extent that is close to the maximally covering seed or above a certain rate. But this would increase the rate of overlap between clusters that was already much higher than with the Inspec experiments. Relaxing the assignment criteria would also be inappropriate for documents that had narrow scope.

5.2.3 Indexing the databases

In the first group of user experiments, the existing index for the Inspec database was used in the searches. This index incorporated a small number of stopwords, and a list of synonyms, and phrases. The stopword list generated for use in the experiments was incorporated in the implementation and the terms from the retrieved documents were filtered using this list before selection of the list of representative term candidates (see Section 5.3.4.3 on page 100 for information on the stopword lists used).

[40] AP881024-0066, AP880815-0071, AP881004-0065, AP880414-0024, AP880506-0008, AP880506-0165, AP881008-0050, AP880214-0084, AP881125-0162.

At some point during the first group of experiments, the Inspec database was re-indexed based on the final (expanded) list of stopwords and synonyms, in order to increase efficiency of the implementation. This was also necessary for the effectiveness of the clustering algorithm. In a clustering implementation synonyms should be treated as a single variable to ensure that:

1. No two equivalent terms are incorporated in a representation separately, creating a duplication(double-weighting) without any extra informativeness,
2. Two synonyms with different distributions among the document set, do not conceal a similarity that would be more clear and accentuated, had a single identifier been used in all documents.

We needed to have a unique representative from the list of synonyms to be returned as the term, regardless of whichever form it occurred in the document. CISR programmer Thien Do, made the necessary modifications to Okapi programs to enable this.

The indexing was realised by comparing the documents in the collection with a go-see-list (GSL). A GSL includes lists of stopwords, phrases and synonyms. In indexing a document, terms that match any of the stopwords in the GSL are not indexed. Terms matching any of the items from a list of synonyms are indexed by the preferred term(one of the synonyms), which ensures that all synonyms generate matches in the retrieval stage when any of the group members are used in a query. TF and ICF values are also calculated collectively for a group of synonyms. Phrase lists enable the search engine to recognise phrases, and weight them specifically as a phrase, rather than as individual terms which are parts of a phrase.

The indexing also presented us with an opportunity to define the present and past tense forms of most of the irregular verbs as synonyms, as it was not done by the stemming algorithm used (see Section 5.3.4.2 on page 99).

The TREC database was similarly indexed before the facet evaluation experiments, using the GSL generated for these experiments (see Section 5.3.4.3 on page 100).

5.3 Other implementation decisions

Decisions regarding document representation, cluster representation and various other parameters of the algorithm were arrived at by a combination of hypothesising, prototypical implementations and experimentation. A rule of thumb was to keep the implementation as simple as possible, so that the process remains sufficiently transparent and allows diagnostic evaluation. The considerations and final decisions related to the various aspects of the implementation are summarised below.

5.3.1 Partitioning vs. overlapping clusters

The algorithm is originally of partitioning type, but it was also possible to utilise it as overlapping type(see Section 5.1.1 on page 81). Overlapping clusters can help to avoid forcing documents into a single cluster when they cover different facets represented by different clusters. They can also increase users' confidence in selecting or rejecting clusters.

Yu(74), in his studies on user-based clustering (more precisely, query-based clustering) preferred overlapping clustering suggesting it is usual that some documents should appear in more than one group. Methods used for clustering/grouping search output results on the Web also allow overlap between clusters(see Section 4.8.2 on page 63).

Actually, the choice between overlapping and partitioned clusters has to be made depending on the purpose of clustering. If output clustering is seen as a way of viewing the search output, overlapping clusters may not be very desirable. They might cause confusion, and users may not like to see documents repeatedly in different clusters if that happens with a considerable proportion of documents. But, as mentioned above, it may be preferable if the clusters are used as a narrowing down or query expansion tool.

With an online implementation, a sub-optimal solution to this problem can be to mark or "de-emphasise" documents that have been viewed by the user in a previous cluster. Alternatively, it may be left to the user to choose between partitions or overlapping clusters.

With our implementation of the C^{3}M algorithm, it seemed appropriate to allow overlapping in cases where a document is maximally covered by more than one cluster seed. Can&Ozkarahan(90) suggest assigning such documents to the seed that has the highest seed power, but, in the context of this project, it would be better to assign the document to both seeds since the document has equal similarity to both.

This arrangement resulted in an average of 1.6(3%) overlapping documents in the 2nd set of user experiments, and did not create any recognisable problems. However, with TREC experiments, the situation was more problematic. When clustering the documents relevant to a TREC query(TREC-REL), the overlap rate varied between 0% and 36% depending on

- the query(i.e the document set retrieved),
- document representation (best passages(BP) vs. full documents(FD)) and
- method of seed selection used (checking seed candidates' coverage by previous seeds and skipping them if they are maximally covered by any of the previous seeds(C), versus

checking seed candidates only for whether they are identical to a previous seed (NC) using a small threshold value of difference between seed powers),

- method of choosing representative terms (choosing the terms with respect to their weights within the search output vs. choosing them from individual documents (again with respect to their weights within the search output) in order to ensure a more balanced distribution of representative term occurrences among documents.

(See Section 5.3.4.7 on page 104 on using full documents versus best passages in TREC experiments, 5.3.2 on page 92 on seed selection methods, and 5.3.4 on page 95 for selection of representative terms).

The outputs from the queries were clustered in various combinations of representative term and seed selection methods, and document representation formats to evaluate the combined effects of these factors on the rates of overlap. On average, using best passages and checking the seed candidates' coverage by previous seeds (labelled BP-C) gave the lowest overlap rates, and using full documents without checking the seed candidates' coverage by previous seeds (FD-NC) produced the highest overlap rates.

The overlap rates of BP-C combination were further reduced when the terms to be used in clustering were chosen from the individual documents, instead of the search output, to ensure that each document includes a minimum number of representative terms(see Section 5.3.4.8 on page 106 for details). This is to be expected as documents with only one or two representative terms are more likely to be equally similar to more than one seed, than documents that at least have 3 or 4 representative terms.

See Appendix 1.5 for overlap values resulting from different combinations of the factors outlined above.

5.3.2 Selection of seed documents

Seed documents are chosen from among documents with highest seed powers. Seed power is a composite measure; it is found by multiplying the coupling and de-coupling coefficients and the number of positive entries for the document in the document-by-term matrix. This measure favours documents with a medium number of terms that are neither too general to be considered as cluster seeds nor too specific to be unable to attract documents to their clusters once they are assigned as seeds.

A serious problem with choosing the seeds according to their seed powers is the risk of having two similar documents with high seed powers and assigning both of them as seeds, thus creating two very similar clusters. Can&Ozkarahan(90) were concerned about this possibility, and have suggested threshold testing as a means of avoiding identical seeds (see Section 5.1.1 on page 81).

During our tests for tuning the algorithm, Can&Ozkarahan's suggested threshold level for identifying identical seeds was increased tenfold to 0.01, as generally seed power values seemed to vary between 0.3 and 1.5 between two consecutive candidates in the sorted list of seed power values.

However, with one of the TREC queries, namely query 256, it was evidenced that even this level could be too low, as two seeds with a difference of 0.08 between their values had 17 common documents in their clusters. Their coverage of each other was found to be quite high (0.14 whereas overall average coverage value was 0.018) and it was also observed that they had 48 common terms out of the 60 terms selected for use in the algorithm. It seemed clear that these seeds were highly similar, although they had passed the threshold test.

As we were not just interested in avoiding identical seeds in this implementation, the threshold testing Can&Ozkarahan have put forward to avoid having identical documents could as well be too permissive for our purposes. A candidate could pass the test of thresholds, but still be covered by a previous seed document more than any other document in the output set. If such a document is assigned as a seed, it would mean that that seed would have a lesser resemblance to any member of its cluster than to the seed that maximally covers it.

Since some of the clusters formed in the initial trials with the implementation appeared to be too similar to each other, we found it useful to try a different method of choosing the seed documents to avoid the situations described in the above paragraph. This method consisted of checking, while assigning seeds, whether the candidate document is maximally covered by one of the previous seeds and rejecting that candidate if that is the case. It would ensure that a candidate would not become a seed if its similarity to another seed was higher than any other document. The method also avoided the need to determine a certain threshold level to avoid identical seeds.

However, this method had a certain drawback. Although the total number of seeds would still be the same (as another candidate would be used to replace the rejected one), most of the documents that would normally be clustered around the rejected seed would now be assigned to seeds to which they were less similar than the rejected seed. It was also likely that a majority of them would be assigned to the seed that covered the rejected seed maximally, hence creating a large cluster and distorting the balance between cluster sizes.

In order to assess the consequences of using this method, six Inspec query results were clustered both by using this method(labelled C), and by assigning all top n candidates as seeds, unless they are identical to another seed(labelled NC).

As expected, the method decreased the average similarity (to the assigned seed) of documents that were maximally covered by the rejected seed in all but one of the cases. But, the overall average coverage values of the documents by the assigned seeds has increased for all but one of the queries, when the seeds were checked for whether they were covered maximally by any other seed(labelled C). This was because by rejecting a seed candidate which had a fair alternative, i.e. the seed that covered it maximally, we could make place for another seed which could be more differentiated than the rest of the seeds and capable of attracting some documents to itself more strongly than the rest of the seeds.

It was also important to find out the extent to which the method distorted the cluster distributions. As far as maximum cluster sizes were concerned, the method did not seem to create inflated clusters that distorted the distribution. In fact the maximum cluster sizes were smaller for 4 of the six queries and unchanged in one query when the method was used (Table 5-1). As a result of these findings, it was decided that this method could be employed for both Inspec and TREC experiments without undesirable side effects.

Table 5-1 : Inspec queries - Maximum cluster sizes

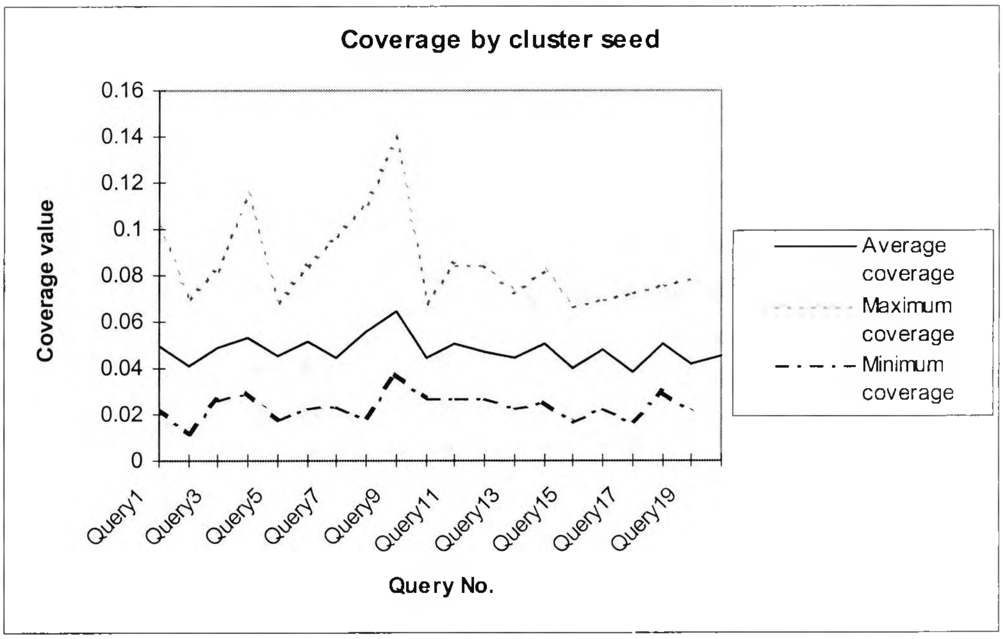
query	C	NC
Q1	18	14
Q2	19	21
Q3	20	21
Q4	14	17
Q5	17	17
Q6	18	20

5.3.3 Assignment of documents to clusters

In the C^{3}M method, the documents are assigned to the seeds that cover them maximally; in the case of more than one seed covering a document maximally, the document is assigned to the seed with the highest seed power (As mentioned in the previous section, in this project, the documents have been assigned to all maximally covering seeds).

It was also worth considering setting a coverage threshold for assignment of documents to seeds, to avoid assigning documents to seeds to which they are only peripherally associated to. Since the Inspec queries seem to generate a stable range of values for coverage (Figure 5-1), setting a threshold level for assigning documents to seeds would probably not be difficult.

Figure 5-1 : Average, minimum, maximum and median coverage values by query



However limiting documents' assignment to seeds would also mean that we would have a hybrid output made up of a few clusters and a ragbag. This would increase the complexity of the output and also raise the question of how to arrange the output set. The individual documents could be sorted within themselves by their relevance rank values, but it would be hard to decide their position relative to the clusters. The hybrid representation was also likely to introduce difficulties in comparing ranked retrieval and clustered output. Therefore, a coverage threshold was not used in this implementation.

5.3.4 Document representation

Representation of the documents has the utmost effect in determining the way they are distributed among the clusters. In a typical CBR application, documents are represented by term-by-document matrices, where term selection is based on some term occurrence thresholds.

As mentioned before(see Section 4.6.1 on page 48) , it is possible to complement this matrix with additional variables such as the source, author, date of publication, obtainability, cost and various other information that may be incorporated in the document. However such information is generally non-standardised and difficult to process. It is also unclear how these factors should be used in conjunction with terms.

For these practical reasons, in this implementation, document representation was based only on terms. Factors that should be considered in selection of terms for the matrix were determined as:

- number of terms to be included in the analysis,
- use of stemming,
- use of stopwords,
- use of term weights for term selection,
- weighted vs. binary variables,
- search terms and method of searching.

In addition to these factors, two other factors emerged during the course of the TREC experiments:

- use of full documents vs best passage,
- number of representative terms in documents.

5.3.4.1 Number of terms to be included in the analysis

The number of different terms to be obtained from a given search output depends primarily on the indexing exhaustivity and the type of document representation(i.e. whether abstracts, full-text or just titles and keywords are used).

It has been mentioned before that (see Section 4.6.1 on page 48) there is not a monotonic relationship between number of features and quality of representation. Extra features provide additional information, but the more that are used, the greater the risk of obscuring the cluster structure. Therefore, with databases including abstracts or full texts, it may be advantageous to filter the term set so as to end up with a compact set composed of terms that are likely to be good discriminators.

In IR, the two most broadly used factors in term selection are ICF(inverse collection frequency) and TF(term frequency); in their experiments with the C^{3}M algorithm, Can&Ozkarahan(90) have used both upper and lower ICF thresholds for selecting terms to be used in document representation. But, in this study, as our domain, i.e. the search output, was only a small subset of the collection, low-level, more specific factors were needed for selecting terms that had discriminatory value within this subset. A two-step selection process was employed to select the terms to be used in clustering:

1. Elimination of terms that have too low or too high occurrence frequencies in the search output,
2. Selecting terms that have high discriminatory values from within the remaining set of terms. (This process is detailed on Section 5.3.4.4 on page 102).

Terms that occur in only a few or in almost all documents have very small discriminatory value and their exclusion can greatly reduce the size of the term set. In Inspec database, a typical search output of 100 documents, generated about 1800 terms, and around 75-80% of these terms occurred only once or twice in the output set. As long as these terms were distributed randomly, we would be safe in thinking that they contributed little to the description of the search outputs. Their elimination would result in substantial computational benefit and little loss of information. This loss was in fact necessary, as classification is impossible(except when it is trivial) unless some information is ignored. It would also be reasonable to exclude the terms that occur in a majority of the documents as they would also have little discrimination power.

These two operations, respectively, work to increase and decrease the communalities in the clustering structure. It follows that the number of clusters and average similarity between documents are affected extensively by such measures.

It is not desirable to have a crowded matrix(i.e. one having too many “1” entries), as a dense matrix points to higher similarities between documents on average and it becomes difficult to differentiate between documents; conversely, a sparse matrix may yield too many clusters that diminishes the value of clustering as a categorisation tool. Term frequency thresholds can be used as a tool to manipulate the density of the term-by-document matrix to yield more balanced clusters.

In order to determine the minimum/maximum thresholds to be used in the experiments, several queries were run with various combinations of threshold levels. As expected, higher thresholds increased the matrix density and average resemblance between documents. This in turn led to smaller numbers of clusters. Resulting number of terms and clusters for four queries are presented below(Table 5-2).

Table 5-2 : Inspec queries - Number of representative terms and clusters with different term occurrence threshold levels

	With the stopwords used by CISR		With the stopwords used by CISR plus the semi-stopwords composed by Steve Walker									
	T: 3 T+A: 5		T: 3 T+A: 5		T: 4 T+A: 6		T: 5 T+A: 7		T:6 T+A: 8		T:7 T+A: 9	
Query	T.num	Cl.num	T.num	Cl.num	T.num	Cl.num	T.num	Cl.num	T.num	Cl.num	T.num	Cl.num
Q1	352	11	287	12	191	9	146	8	118	7	89	6
Q2	170[41]	8	128	8	86	6	59	5	42	4	28	4
Q3	355	11	286	12	196	9	151	8	122	7	103	7
Q4	334	11	267	11	188	9	141	8	122	7	99	6

T: minimum occurrence threshold for title terms
T+A: minimum occurrence threshold for abstracts + titles
T.num: number of terms
Cl.num: number of clusters

The experiments revealed that some of the terms used in clustering had very little contextual value, and the higher number of such terms used in a clustering solution, the lower would be our chances of having meaningful clusters. Eventually, the decision was made to set the minimum occurrence threshold at 4, but expand the list of stopwords to eliminate a good proportion of terms with low contextual value(see Section 5.3.4.3 on page 100). An upper threshold of 40(80% of the documents in the output set) was also used to exclude terms that occur on almost all documents and hence have very small discriminatory value.

When clustering the relevant documents for the TREC queries(TREC-REL), as the number of relevant documents was different for each query, it seemed reasonable to determine the number of terms to be used in the algorithm based on the number of documents retrieved for the query. Trials showed that this could help to ensure more balanced distribution of documents among the clusters. Basically, 40 terms seemed appropriate for queries that retrieved less than 30 documents. Fifty terms were used for queries retrieving between 30 and 49 documents and 60 terms were used for sets containing 50 or more documents.

When clustering the documents retrieved in response to City users' initial query statements in the TREC experiments (TREC-USER), as the number of documents retrieved were equal for each query, 60 terms were used to represent the documents.

For the TREC experiments, minimum term frequency threshold was set to 4 as with the Inspec experiments, but the maximum term frequency threshold was set to 60% as compared to 80% with Inspec experiments, as the average number of common terms among the documents was much higher with TREC data, and the upper term occurrence threshold had to be reduced to achieve well balanced clusters (see Section 11.1 on page 191).

5.3.4.2 Use of stemming

Stemming is commonly used in IR systems as it improves recall and helps performance by reducing storage requirements.

Stemming improves recall by combining morphological variants with similar semantic interpretations. However, as with any other natural language processing problem, it is not very easy to come up with a comprehensive set of rules for stemming and studies on usefulness of stemming have not produced clear results. A recent comparative study has found an average absolute retrieval improvement of only 1-3% due to stemming(Hull, 96).

In this project, we were interested in arriving at a compact set of terms for document representation, and use of stemming could help us achieve this purpose. Some kind of stemming seemed necessary as no stemming would bring the possibilities of:

1. double counting two variates of a term with almost identical meaning and
2. non-inclusion of a term which ranked below the threshold level with its two or more variates but would exceed the threshold if they were combined.

There were two stemming algorithms used within the department: weak and strong. Both were based on Porter's suffix stripping algorithm(80) which works by treating complex suffixes as compounds made up of simple suffixes, and removing the simple suffixes in a number of steps. There are 5 different steps in the algorithm and in each step the removal of the suffix is made to depend upon the form of the remaining stem, which usually involves a measure of its syllable length.

Weak stemming is based on Step 1(but it leaves 'ed' endings) and Step 5 and is used for the Inspec database. It reduces nouns into singular forms, discards "ing"s of verbs but leaves the past tense verb forms as is.

[41] Q2 retrieved 57 documents, therefore it generated lesser number of terms and clusters. The remaining queries all generated more than 100 documents and first 100 were selected for inclusion in the clustering.

Strong stemming uses all five steps and is used for the TREC databases. In addition to what is described for weak stemming, it transforms all verb forms to present tense (except irregular ones), and gets rid of many suffixes such as "-ational", "-alism" "-icate", "-ative", "-alize", "-ical", "-ful", "-ness", "-ably", "-al", "-er", "-ism", "-tion", "-ment"(e.g. measurement reduces to measur-, worker to work-, consistently to consist).

For both, there is some conflation of British and American spellings, and an attempt to deal with some non-standard plural forms (e.g. terminal 'ae' becomes 'a', terminal 's' preceded by 'l' or 'u' is retained).

Both types were tried on six Inspec queries and weak stemming was chosen for both user and TREC experiments as it appeared to result in more meaningful clusters.

5.3.4.3 Use of stopwords

Most IR systems use a generic list of stop words to get rid of the terms which are not expected to play a role in retrieval. It is generally considered proper to be cautious in assigning terms as stopwords and keeping the set of stopwords small, as morphological variants and even the most typical stopwords can play an important role in retrieval(e.g. Riloff, 95). Bookstein et al.(95) point out that whether a term bears content or not depends on the collection. He has found that the pronoun "her" bears some content within the Bible and may be a good keyword for queries about women in the Bible. Dubin(95) has also reported similar findings.

In this project, stopwords are used extensively as a way of ensuring a compact, meaningful set of representative terms. Their extensive use is a compensation for using low term frequency thresholds, which increases the number of candidate terms to be included in the algorithm considerably.

In initial trials with the algorithm, only the small stopword list that was created by CISR researchers was used(see Appendix 1.1). As the trials revealed that the expansion of the stopword list could be beneficial to the meaningfulness of the clustering solutions, terms defined by Steve Walker as semi-stopwords for experiments with the Okapi system[42] were added to this initial list (see Appendix 1.2).

Trials with this expanded list revealed that some improvement had been achieved, but there was still some scope to expand the list further to eliminate more non-contextual terms. It was clear that the algorithm would produce more satisfactory clustering solutions, if we could feed into it only the terms that had a high likelihood of pertaining to some aspect of the query. While it was impossible to achieve this, we could expand the list of stopwords and try to eliminate more of the terms that would not have much contextual value with respect to most queries. But it was hard to describe and rationalise the process of such elimination. Inspection of most frequent terms of the database resulted in a vague definition of the terms that could be added to the stopword list:

1. words that are the nuts and bolts of documents: i.e. words which link contextual terms and help the author with the flow of the text rather than conveying meaning themselves, e.g. expect, called, often, otherwise, perhaps, together, supposedly, especially, initially, nevertheless, etc. These terms are probably stopwords in some retrieval systems,
2. words that have various meanings: since we are after a concise representation of documents, it is preferable to leave out a term that could be helpful only in a limited number of cases instead of keeping it so that it distorts the matrix in many other cases. Examples are: express, face, matter.

In order to come up with a comprehensive list of stopwords based on these rough definitions, terms that had collection frequencies greater than 1000 in the Inspec database were inspected and further stopwords were identified.

Ideally, it would be desirable to review the list with a lower threshold but even this list was quite big (including 2898 items) and it would be much more difficult if the threshold was further lowered. Since the most frequent terms were the ones most likely to come up in the search outputs, there would also be decreasing returns from lower thresholds.

It was not easy to go through the list and pick the terms to be treated as stop-words. The process was highly subjective and far from being routine. After several iterations on the frequency list, a final set was formed and when this was added to the existing stopword list, the cumulative figure reached 900 (see Appendix 1.3). The GSL also included 950 synonym groups, but they were mostly taken from the GSL prepared in the CISR for the TREC experiments, and a majority of the items from this list were redundant for the Inspec data.

The queries were re-run after this stage using various term occurrence threshold levels using weak or strong stemming. On the whole, it may be said that the final clustering solutions showed higher intra-cluster resemblance.

Later on, for the facet evaluation experiments with the TREC database(D25_96), a similar inspection of the terms with the highest collection frequencies was made and about 800 stopwords were identified. The synonym list of the GSL for this database was also expanded to include 900 synonym groups(Appendix 1.4).

[42] Walker used a semi-stopword list to indicate terms that are not excluded from the indexing, but will

5.3.4.4 Use of term weights for term selection

During the 1st group of user experiments, terms to be included in the algorithm, i.e. terms that are to represent the documents in the term-by-document matrix, were chosen with respect to some lower and upper output frequency thresholds. After the clustering process, terms to represent the clusters were chosen out of this term set, based on their weights within the cluster(see Section 5.3.5 on page 111).

Throughout this process, collection frequencies were never used. That sometimes resulted in “unrepresentative” representative terms that could somehow shine out within the output set due to chance. Search output statistics were of limited statistical value and not always adequate to differentiate between contextual terms and non-contextual terms that accidentally happened to have relatively high frequencies in the output

After the 1st group of user experiments, in order to put the term selection process into a wider perspective, we decided to use the collection statistics in conjunction with the output statistics for selecting terms to represent the documents. By including the more reliable collection statistics into the selection process, we could choose terms based on their ability to discriminate between the output set and the whole collection. A term set formed in this way is more likely to convey contextual information about the output than a term set formed solely on the basis of output statistics. The TSV (Term Selection Value) measure, which correlates with a term's ability to discriminate between document groups, seemed to be an appropriate measure for this purpose.

In his 1990 paper, Robertson suggests that an appropriate criterion for term selection is different from an appropriate criterion for weighting of a term that has been already selected. He argues that for term selection (for query expansion) it is necessary to strike a balance between high-precision and high-recall, since neither very infrequent terms or very frequent ones are very useful.

This formula could be used to select terms for cluster representation as well, as the second part of the equation favours terms that occur frequently in the cluster but less frequently in the whole output. But we needed to modify the formula parameters, in order to find the values for terms in clusters in comparison to the search output. With such modification, TSV would be calculated as:

$a_t = w_t (p_t - q_t)$

not be used as additional terms in query expansion.

where w is calculated in the same way as described above in this section.

$$w = \log \frac{(t_c + 0.5)(T - C - t_T + t_c + 0.5)}{(C - t_c + 0.5)(t_T - t_c + 0.5)}$$

$$p = \frac{t_c}{C} \quad \text{and} \quad q = \frac{t_T - t_c}{T - C}$$

T is the number of documents in the search output,

t_T is the number occurrences of the term in T,

C is the cluster size and

t_c is the number of term occurrences within that cluster.

Upper and lower frequency threshold levels were still used as the initial criteria to extract a set of candidate terms that had the potential to discriminate between the documents. After this initial selection, Term Selection Value's(TSV)s were calculated for each term that falls within the threshold values to find out the terms that have highest discriminatory power.

After the completion of clustering, TSV's were again used to find out terms to represent the clusters; at this time TSV's were calculated only using search output and cluster statistics in order to estimate the discriminatory powers of terms within a cluster as compared to the search output set.

5.3.4.5 Weighted vs. binary indexing

Another decision on implementation concerned whether the variables should be weighted or not and if they are to be weighted, how this should be done. In CBR, this generally means a choice between binary and weighted indexing.

Sneath and Sokal(73) are against the use of weighting schemes in cluster analysis, and studies by Willett(83) and Willett et al.(86b) revealed weighting made little difference to the clusterings made by the single link and the non-hierarchical Jarvis&Patrick methods, respectively. Can&Ozkarahan(90) also report that the weighted version of C^3M does not give better results.

Given this evidence, the decision was made to use binary matrices in this project.

5.3.4.6 Search terms and method of searching

Search terms may be excluded from the document representation if all documents contain the term in the same form. But if the search is a best match search then the search terms may become good discriminators. Variants of a search term may also have a significance for discriminating documents.

In this project, it seemed reasonable to include search terms in the term-by-document matrix, but because documents were retrieved using these terms, they were highly likely to be candidates for inclusion anyway. As their contribution to clustering would depend on their output frequency like any other term, the decision was made to leave the selection completely to the algorithm.

5.3.4.7 Full documents versus best passage

This factor is only relevant to the TREC database which includes very long documents and documents which may be focusing on more than one subject. It is logical that if some documents have parts that discuss different topics, and these parts can be recognised in some way, considering only the relevant parts will be useful for analysis of relevance and query expansion using relevance feedback.

Okapi is capable of differentiating between passages of a document by utilising query and relevance feedback information. For documents longer than 10K, best passages are marked while displaying the documents to the users and users are able to make relevance judgements on full documents(FD) or best passages(BP).

Initially, in this implementation, full documents were used in the TREC experiments. But when clustering the documents retrieved by the users(TREC-USER), some queries resulted in unbalanced cluster sizes despite all modifications made to the algorithm to avoid the problem(see Section 11.1 on page 191). Using best passages instead of full documents came into consideration as a possibility to alleviate the problem.

In order to find out whether best passages could improve the balance of clusters, TREC queries for both the first and second set of experiments were re-run and results were clustered using both full documents(FD) and best passages(BP) and the two alternative seed selection methods(C and NC[43]) . Of the four alternative methods BP-C produced the most

[43] C involved checking seed candidates' coverage by previous seeds and skipping them if they are maximally covered by a previous seed, NC involved checking only for whether a seed candidate is identical to a previous seed; see Section 5.3.2 on page 92 on seed selection methods

balanced clusters. It also resulted in lower overlap rates and higher coverage rates (see Appendices 1.5 and 1.6).

A problem that was experienced when clustering the documents retrieved by the users (TREC-USER) was having documents that had zero or negative similarity to all seeds and hence could not be assigned to any clusters[44]. Using best passages also seemed better in reducing such situations as compared to using full documents (Table 5-3). But even with the best passages, in TREC-USER experiments there were more documents containing a very high number of the representative terms when compared to the documents from TREC-REL experiments, and this led to less balanced cluster sizes. The average standard deviation in cluster sizes was 4.26 for the TREC-REL experiments and 5.98 for the TREC-USER experiments (See Appendix 1.8 for comparison of number of terms in documents and the standard deviations in cluster sizes in TREC-USER versus TREC-REL).

Table 5-3 : Number of documents with zero similarity to all seeds (TREC-USER)

query	BP-C	BP-NC	FD-C	FD-NC
254	3	3	3	4
256	1	0	3	3
258	0	0	0	0
264	0	0	2	2
284	0	0	1	2
286	0	0	0	0
292	0	0	0	0
average	0.57	0.43	1.29	1.57

In the TREC-USER experiments, the unretrieved relevant documents were checked for their similarity to the clusters formed out of the top 50 retrieved documents. With some queries, many documents had zero or negative similarity[45] to all clusters. But here again, the problem was resolved when best passages were used. Number of unretrieved relevant documents whose maximum similarity to the clusters is zero or negative are presented in Table 5-4.

[44] These documents included no or only a few of the representative terms.

[45] In the Robertson&Sparck Jones (76) relevance weight formula, negative weights can occur although terms with negative weights would not be considered as candidates for inclusion in a query. In this context, a negative term weight would indicate that the term's absence has a significance, e.g. in a search output of 50 documents, if a term occurs only once in a cluster of 8 documents, but 12 times in the whole set, it will have a weight of -.27 for that cluster.

Table 5-4 : Number of unretrieved relevant documents that have zero or negative similarity to all clusters (TREC-USER)

query	BP-C	BP-NC	FD-C	FD-NC
254	0	0	0	0
256	2	2	23	28
258	0	0	1	1
264	1	1	35	32
284	0	0	0	0
286	1	1	0	0
292	0	0	0	3
average	1	1	8	9

In Table 5-4, the high numbers of relevant documents that have zero similarity to all clusters for queries 256 and 264 can be explained by the scarcity of relevant documents in the search outputs. For these topics, user queries returned only 3 and 6 relevant documents respectively (as compared to an average of 11 relevant documents for the 7 queries), which led to clusters that were not representative of the whole set of relevant documents.

All these improvements can be attributed to improved focus of the documents when best passages are used. In some documents, the best passage may be only a minor part of a very long document, and eliminating document parts that have no relation to the query certainly decreases the chances of selecting non-contextual representative terms.

5.3.4.8 Number of representative terms in documents retrieved

In this implementation, terms are selected with respect to their frequencies in the output set and their TSVs. The number of terms to be included in the analysis is limited to a certain number that is determined with respect to the number of documents retrieved and the database used(see Section 5.3.4.1 on page 96).

The criteria used in term selection is supposed to bring out the set of terms that is most capable of describing the documents in the search output. With Inspec experiments, these term sets yield clustering solutions that are well balanced in terms of cluster sizes. However, with the TREC experiments cluster sizes had a much higher level of variation(see Section 11.1 on page 191). This is perceived as a problem in the context of this project, as the utility of clustering as a search output representation method can be considerably reduced by unbalanced clusters. Clustering is expected to help the users by allowing them to make judgments on groups of documents instead of having to inspect individual documents one by one, and letting them to focus on promising subsets more easily. However, the utility of selecting a cluster that includes 80% of the documents or the utility of rejecting a cluster that includes only a few documents is lower than the utility of making judgements on 4-6 clusters each of which represent a roughly equal proportion of the documents from the output set.

While we aimed to bring out an underlying structure in the document set rather than imposing a superficial grouping, it was also obvious that the final judgment of the value of clustering would be based on the usefulness of the resulting distribution. Therefore, it was necessary to find a means of creating balanced groups while allowing for some flexibility to accommodate variations resulting from the level of diversity of the output sets.

The algorithm was modified in various ways to avoid having a very skewed distribution of cluster sizes in TREC experiments. None of these modifications ensured balanced clusters for all of the queries(see Section 11.1 on page 191). Eventually, it became a necessity to make an elaborate comparison between the Inspec and TREC data to find out why one consistently produced reasonably balanced clusters and the other performed unreliably in that respect despite all modifications made to avoid the problem.

The clustering solutions were based on the C-matrices, which in turn were based on the document-by-term matrix used as input to the algorithm. Therefore, the source of the variation had to lie within these matrices, or more precisely distribution of entries in these matrices. Rows represented terms and columns represented documents in the term-by-document matrices. If a document included a representative term, the intersection of the row representing the term and the column representing the document would have an entry of “1”, otherwise it would have “0”.

Minimum/maximum frequencies of terms that affected the row distributions was determined by thresholds applied during the term selection process. However, their distribution could still be different from those generated with the Inspec experiments. The number of repersentative terms in documents was uncontrolled, and was more likely to be a source of variation.

For comparison, users’ search statements for the seven TREC queries used in the experiments (TREC-USER) and six exemplary Inspec queries were run and results were clustered using equal output frequency and number of representative terms thresholds. For the TREC queries, clusters were created both with best passages and full documents.

For each document and each term row/column totals were calculated. Terms were sorted by their number of occurrences and the documents were sorted by the number of representative terms they included. The sorted lists were then graphically analysed to see not only the difference between mean or median values but also the differences between the distributions. Graphs were created for both individual documents and averages over queries[46].

The distribution of the term frequencies seemed to be very similar between Inspec and TREC experiments. When the distributions were averaged over the queries and the average values were graphed, the lines representing TREC and Inspec were almost overlapping(See Appendix 1.7). Distributions of the TREC results with best passage(BP) vs full documents(FD) were also highly similar. The range of average and median values over all queries and the overall average and median values are presented in Table 5-5.

Table 5-5 : Term frequencies within the search output - Inspec vs TREC

	Inspec	TREC - FD	TREC - BP
Average range	8.2-10	7.9-9.2	7.3-10.5
Average	8.8	8.5	8.7
Median range	6-7	5-7	5-7
Median	6	6	6

However, the distribution of the number of representative terms in each document were somewhat different between the Inspec and the TREC results. Graphically, TREC experiments seemed to yield curves that were closer to the x-axis in the low frequency areas, but steeper in high frequency areas as compared with the results of the Inspec experiments (Graph 5-1). The average difference between TREC and Inspec distributions was somewhat decreased when best passages were used instead of full documents. But this was due to changes in distributions for some of the queries, rather than an all encompassing effect (see Section 5.3.4.7 on page 104 for details) .

As Graph 5-1 shows, compared to Inspec queries, TREC queries result in a more extreme distribution; they have more documents that include none or just a few of the representative terms, and more documents that include more than 1/3 of the terms, a figure representing the maximum number for most Inspec experiments.

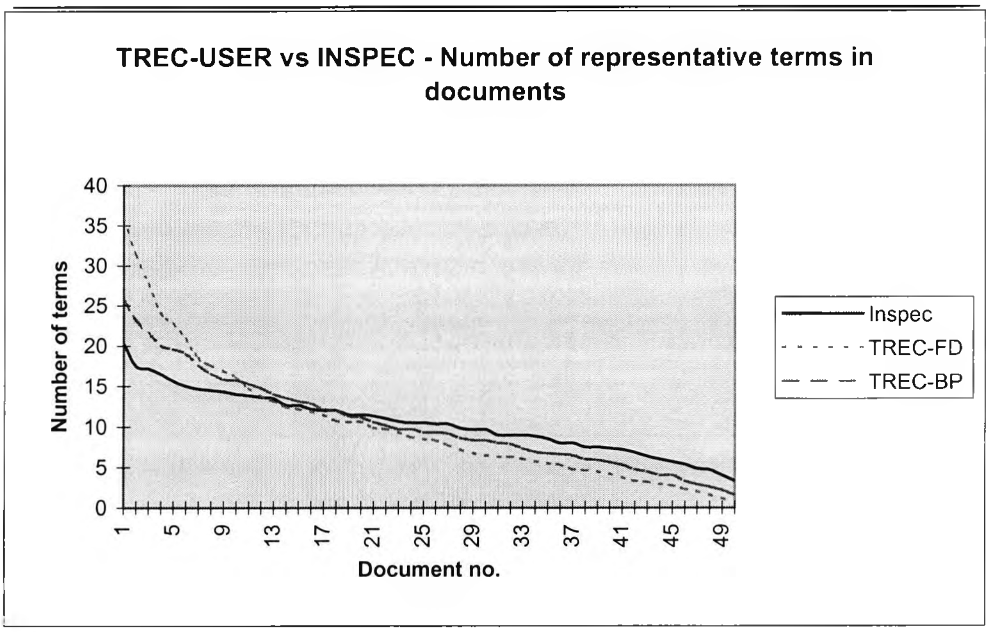
The differences between average and medium number of terms in documents are not very significant, but there seems to be higher variation between the median values than between the average values. The range of average and median values over all queries and the overall average and median values are presented in Table 5-6.

[46] For instance, for the number of representative documents in each document for the 2nd run of the TREC experiments, figures from each of the 7 queries were sorted, and the figures from the documents

Table 5-6 : Number of representative terms in documents - Inspec vs TREC

	Inspec	TREC - FD	TREC - BP
Average range	9.6-12	9.5-11	8.8-12.58
Average	10.5	10.2	10.4
Median range	9-12	5-11	8-12
Median	10	8	9

Graph 5-1 : TREC user queries (TREC-USER) vs Inspec - number of representative terms in documents as averaged over queries



The benefits of using best passages seemed to depend basically on its effect on the distribution of number of representative terms in the documents. This was best demonstrated by the individual graphs of the user query results for the TREC topics(TREC-USER)[47].

For instance, in queries 254 and 286 where full documents and best passages had the most similar standard deviations in cluster size, the distribution of number of terms in documents were almost overlapping(see Appendix 1.7 for the distribution for query 254).

In the four queries (256, 258, 264, 284), in which the balance of cluster sizes were improved, similarity of distributions of number of terms in documents to distributions from Inspec queries

with the lowest number of representative terms were averaged to create the first entry, figures from the documents with the second lowest number were averaged to create the second entry and so on.

was improved when best passages were used; the curves were less close to the x-axis in low frequency areas and/or more close to it at high frequency areas[48] (see Appendix 1.7 for the distribution for queries 256 and 264).

It is intuitive that the steeper distribution of TREC query results is due to unbalanced cluster sizes. The documents with no representative terms are not assigned to any clusters but listed in a ragbag. Those with just a few terms are assigned to clusters, but without strong evidence of resemblance. Hence the distribution in the low frequency area increases the possibility of having documents that cannot be assigned to any seeds. They may also increase the overlap between clusters, as a document with just one or two representative terms is more likely to be equally covered by more than one seed than a document that has at least three or four representative terms.

On the other hand, documents with a lot of representative terms increase the average similarity measure and make it difficult for the algorithm to discriminate between the documents. These documents increase the possibility of having large clusters.

Analysis of these results made it clear that a smoother frequency distribution was needed to avoid very small and very large clusters and documents that cannot be assigned to any seeds. It was envisaged that the term selection process could be modified to create a smoother distribution.

Normally, for document representation, terms with highest TSV's were selected from among the set of terms that fulfilled the threshold criteria. In the light of these findings, this method was changed to select terms from individual documents, iteratively until a certain number of terms were found. Again terms with highest TSV's were chosen from each document, unless that term was already chosen as a representative term from another document. In such cases, the next document was investigated, and these common term candidates ensured that on average 5-6 rounds could be made to come up with about 50-60 terms. This resulted in a minimum of 7-11 representative terms when clustering the sets of relevant documents (TREC-REL), and 3-7 representative terms when clustering the documents retrieved in response to user search statements(TREC-USER).

It was envisaged that this method could also smooth the skewness of the distribution curve in the high frequency area, as the terms chosen from documents that would otherwise have few or no representatives could reduce the number of terms that could be chosen from documents

[47] The difference between full documents and best passages were less pronounced in the 1st run of TREC experiments.

[48] In query 284, the curve from FD was different from the rest of the queries in that it was similar to Inspec distributions in low and high frequency areas but it had a bulge in the middle frequency area. With BP, the bulge disappeared, and the graphical distribution became more similar to distribuions from Inspec queries.

already being represented by a fair number of terms. However, when this method was applied to TREC-REL and TREC-USER experiments, the curve was usually moved up at both ends, rather than moving up in the low frequency area and moving down at the high frequency area(see Appendix 1.7 for average values over 7 queries). Unbalanced clustering solutions persisted in most cases.

One suggested way of smoothing the distribution at the high frequency area was checking the terms occurring in this area, and trying to exclude some of the terms that occurred in this area but not at the low frequency area. This would lower down the curve in this area without affecting the improved situation at the low frequency area. This suggestion was not pursued as it seemed to complicate the term selection process further, and without providing a good rationale for why certain terms should be excluded.

5.3.5 Cluster representation

Decisions on cluster representation were highly important for the user experiments. A set of 10 experiments following the 1st group of user experiments (see Section 9.7 on page 157) were solely designed and administered for evaluating two cluster representation methods.

Basically, the clusters were represented by a combination of representative titles and representative terms; number of documents in each cluster was also indicated.

In fact, titles are not the ideal document representatives as they are not always chosen by the authors with the sole aim of informing the user about document contents. In his overview of the literature on relevance, Mizzaro(97) states that there seems to be an agreement that titles have the least quality followed by keywords, extracts and abstracts. However, their conciseness and compactness fitted our requirements, and it was thought that with the support of some representative terms they would generally be able to give a sufficient idea to the users about the cluster contents.

5.3.5.1 Representative terms

Most simply, representative terms could be selected as the top n terms with highest occurrences in each cluster. But this would result in representative terms with little contextual value among more meaningful ones. Therefore, during the 1st group of user experiments, representative terms were chosen with respect to their relevance weights calculated using Robertson&Sparck Jones (76) relevance weight formula (F4, point-5 version):

$$w = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)}$$

where:

N = collection size

n = number of postings of term

R = total known relevant documents

r = number of postings in known relevant documents

The formula parameters were modified in the following way in order to find weights for terms in clusters in comparison to the search output:

$$w = \log \frac{(t_c+0.5)(T-C-t_T+t_c+0.5)}{(C-t_c+0.5)(t_T-t_c+0.5)}$$

T is the number of documents in the search output,

tT is the number occurrences of the term in T,

C is the cluster size and

tc is the number of term occurrences within that cluster.

Weights were calculated for each term used in document representation and the top ten terms with the highest weights were shown to the user together with the representative titles.

The results from the 1st group of user experiments revealed that the representative terms were not quite helpful for the users in discriminating between the clusters. They did not always convey meaningful information and sometimes discrimination became even more difficult due to having the same representative term in more than one cluster.

As we considered the representation problem further, it became evident that choosing terms based on their relevance weights was not the best approach to be used. The TSV (Term Selection Value) measure, which correlates with the terms' ability to discriminate between documents, seemed to be a more appropriate measure for our purposes(see Section 5.3.4.4 on page 102 for description of TSV).

With this formula, not only terms that occur in a certain set more frequently than in the rest of the documents, but also terms that occur frequently within the rest of the documents but not in the chosen set get high values (A term may discriminate both by occurrence and non-occurrence). Since we would use the measure for representing individual clusters, it seemed reasonable to choose only terms that occur more frequently within a cluster as compared to the whole search output. Therefore, we needed to filter out high values that were derived by multiplication of a negative weight value and a negative (pt - qt) value.

Another point of concern was having some common representative terms among different clusters. The cluster representative terms were chosen from among the document representative terms for consistency; and sometimes the same term could be chosen for more than one cluster. A control was introduced to avoid such occurrences that could confuse the users. This meant that some clusters could be represented by less than 10 terms.

It was also likely that the 10 representative terms were found cryptic by at least some of the users. Presented in a string separated by hyphens, they had little context by themselves. A similar problem is mentioned by Robertson et al.(97a) about the query terms; searchers have much difficulty in assessing or predicting the appropriateness of query terms in isolation and need to be provided with some context. However, there is a trade-off between presenting the user with more information and keeping their cognitive burden to a minimum and it is necessary to find a solution that strikes a good balance.

One way to provide context without increasing the searcher's cognitive load would be using the representative terms in selection of the titles to be used in the cluster representation. This option has been tested in the intermediary cluster representation experiments (see Section 9.7 on page 157), and found to be a viable option in representing clusters (see next section for implementation details).

5.3.5.2 Representative titles

The two meaningful measures that were calculated for each document were their coverage by other documents and their seed powers. The former determined which seed the document would be assigned to, while the latter determined whether the document itself was capable of representing other documents by becoming a cluster seed. It was possible to choose representative titles according to their coverage by the seed document or their own seed powers. The former would ensure a more consistent picture of the cluster, while the latter would bring up a more diversified picture. Since the seed is central to the characterisation of the cluster, it was considered more appropriate to use its coverage of the other documents as the basis for choice of the representative titles.

During the 1st group of user experiments, the seed documents were assigned as the first representative title of each cluster, and additional representative titles were chosen from among the titles that had highest coverage by the seed document. The number of representative titles depended upon the size of the cluster; the number was set as 15% of the cluster size plus one for the seed. In this way, larger clusters were given more space, providing an indication of the cluster sizes to the user, and giving the clusters an emphasis proportionate to their share of the output set.

The results of the 1st group of user experiments (see Section 9.6 on page 140) revealed that there was a need to improve the meaningfulness of the cluster representations to help the users to discriminate the clusters more easily. One idea was to change the way representative titles were chosen, to enable them to convey more contextual information to the users.

Choosing the titles based on their coverage by the seeds could sometimes give counter-productive results, as the coverage values, based on the title and the abstract of the document; did not always ensure that the title, on its own, would convey useful information to the user. It was more likely that a document would have a high similarity to the seed document due to the terms that occur in its abstract (rather than its title), which is not accessible to the user at the stage of evaluating the clusters.

One way of ensuring some context would be to see whether any of the query terms described by the user, existed among the title terms. The more query terms there were in the title, the higher would be the likelihood that the user could get an idea of the way the document approached his topic of interest.

This idea was augmented by using the existence of representative terms in the titles as an additional criterion in choosing the representative titles. If users could see some of the representative terms among the title documents instead of only viewing them as a list of terms without any context, they could probably assess the clusters more easily.

In order to compare this idea with choosing representative titles by coverage values, two alternative cluster representation formats (Rep-B and Rep-C) were developed, and a series of experiments were administered as a follow up to the 1st group of experiments to compare these formats(see Section 9.7 on page 157 for details). For Rep-B, the representative titles were chosen with respect to the coverage values (i.e. the extent to which they were covered by the seed document), for Rep-C they were chosen with respect to the number of query terms and representative terms incorporated in the titles; except for the cluster seed title which was included as the first representative title regardless.

These experiments revealed that titles were more important in assessing clusters than the representative terms and choosing them according to the amount of query and representative terms they included seemed to be a viable option. However, there was not a significant difference between the two representations, and the decision was made to continue the 2nd group of user experiments with both formats. Users would be asked to rank the clusters in each representation separately, and at the end of the experiment, compare them in their perceived usefulness.

For the 2nd group of the user experiments, Rep-C evolved into Rep-E with the modification of the representative title selection process to choose all the titles with respect to their inclusion of query and representative terms, instead of using the seed document's title by default and applying the criteria to the other candidate titles. Representative term list was also incorporated into Rep-E after the list of representative titles for each cluster. Rep-B evolved into Rep-D, with the small alteration of placing the representative terms list after the titles, instead of before.

For the 2nd group of the user experiments, a variation was also made to the number of titles to be used in the cluster representation; instead of making this number dependent on cluster sizes, it was fixed to three. While proportionate representation could make sense under normal circumstances, it could affect users' ranking during the experiments. The more titles were used to represent a cluster, the less ambiguous it would be to the user and hence have higher likelihood of being ranked better than a similar but smaller cluster.

5.4 Validation

Conducting valid information retrieval experiments is no easy task. This is, Sparck Jones(81) suggests, due to the fact that the characteristics of retrieval systems are difficult to determine and their implication for experiment difficult to identify. According to Tague (in Sparck Jones, p.59, 81) theoretically optimum design can never be achieved, and the art of information retrieval experimentation is to make the compromises that will least distract from the usefulness of results.

The majority of the IR studies cannot be confined to a pure naturalistic or a pure scientific approach and this study was no exception. In a recent paper Robertson&Beaulieu(97b) have argued that the qualitative/quantitative distinction of research methods can be misleading; the highly individual and very qualitative judgement of relevance results in quantitative measures of performance when cumulated over large numbers of documents and requests.

Essentially, our focus on system-user interaction implied the necessity of a naturalistic approach for evaluation, however, it was also necessary to assess the method's performance in quantitative terms, such as its capacity to enable users to find higher numbers of documents compared to ranked output in similar time spans.

A problem situation that has both naturalistic and scientific approaches may be 'partitioned' and different approaches can be used to deal with different aspects of the situation or even with the same aspect differently construed. The basis for such endeavor lies in the 'complementarist approach' which holds that it is possible to use different system methodologies in a complementary way to deal with the different situations for which each is most suited. Complementarism is one of three positions underlying Flood and Jackson's(91) paradigm of 'critical systems thinking'.

The nature of the task at hand in this project was highly subjective. It is generally accepted that clustering is an iterative process where the success of the techniques and measures used depend on the data at hand, the interpretation that can be given to the solutions by the investigator, and the use, if any, that can be derived from the clustering solutions. In order to assess the implementation with some confidence, we needed some kind of benchmark to help us evaluate the use that could be derived from clustering.

It seemed appropriate to set that benchmark as the precision levels attainable by relevance ranking, as relevance ranking was the main rival of clustering as a search output display method. Comparing the precision values of ranked output with precision values of clusters chosen by the users at equivalent threshold levels could provide us with data that could be statistically analysed and used to assess the implementation with a degree of confidence.

But, it was necessary to acknowledge that using relevance data to assess clusters was not the perfect solution in terms of experimental validity. As Shaw et al.(97) suggested, the structure imposed on a set of documents by topical relatedness may not reliably associate documents relevant to the same query, or more simply, topical relatedness does not necessarily manifest relevance relationships(see Section 4.1 on page 37). In terms of this project, this meant that, the capability of the clusters to bring together topically related documents could be underestimated when assessed with respect to co-relevance information.

The existence of TREC data with facet information for the set of identified relevant documents enabled us to address this issue to some extent. We were not bound to relevance information with this data, and could assess the clusters based on topical similarity of the documents they included. However, TREC facet evaluations did not always depend on general topicality, and they had some other shortcomings limiting their usability(see Section 5.2.2.1 page 86).

With all their limitations, these benchmarks enabled us to make certain quantitative measurements. But, the amount of data that could be gathered for analysis was far from abundant. Especially with the TREC data, the amount of data available did not allow us to conduct regular statistical analysis; and methods designed specifically for small amounts of data had to be employed (see Section 12.4.2 on page 208).

Throughout the experiments, the data was also diagnostically analysed for improving the implementation and to gain insight to the interplay between its various parameters. Many queries were run repeatedly with different settings of the parameters, in order to understand how these parameters affected the clustering solutions.

5.4.1 Assessment of relevance

It is widely accepted that assessment of relevance is a highly subjective practice. So much so that Gebhardt(75), in building his probabilistic model for the relevance assessment of documents, finds it appropriate to consider the relevance assigned by a juror to be a random variable. He states:

> *"The source of random variation lies not only in the choice of juror (who is assumed to have sufficient knowledge to understand the problem) but also in the time; at different times, the same juror might judge differently, partially caused by fatigue, external interruptions or by documents he has just read, but partially also because without any recognisable cause."*

In addition to factors detailed in Section 2.1 on page 17, an important source of variation to be elaborated here is the users' deductive powers. As Cooper explains (71):

> *"...the stored information available for satisfying an information need with the help of a reference retrieval system must be viewed as the user's own knowledge plus the system-stored data, and the deductive apparatus is the user's own deductive power plus whatever deductive power has been programmed into the system."*

Users, at times can be incapable of making the necessary deductions and may not be aware that a certain document is relevant. It has been suggested that individuals are different in their openness to information and the more cognitively complex an individual is, the greater the number of relationships he can find between his cognitive categories, and the more likely he is to call a message as relevant (Davidson, 77). But, to cite Cooper again:

> *"It is the philosophy of this definition of relevance,..., that it is simply a fact of life that retrieval system patrons will often be unable to recognise immediately the relevance of a relevant statement, and perhaps never recognise it."*

Researchers often try to control the variance that may stem from users' searching behaviour by trying to draw from a population of similar search experience and specifying the expected commitment and time to be spent on the search task; for example, the model for a user in TREC is a dedicated searcher, not a novice searcher, with the ability to do both high precision and high recall searches and willing to look at many documents and repeatedly modify queries in order to get high recall (Harman, 93). It is also common to provide individual decision-makers with rules for making decisions about which information is relevant (Davidson, 77).

Users' evaluations of clusters similarly depend on their knowledge about the topic they are searching for, their deductive powers and various other factors. But, in the context of this project, the most important point was whether they had a consistent frame for making relevance decisions while evaluating the cluster representations and the individual documents. To some extent, this depended on users' affinity with their areas of stated information request. Users familiar with the query topic could have a more stable frame of evaluation and could feel more confident in their judgements. Users who had no or little knowledge of the topic were likely to be hesitant and could possibly alter their frame of evaluation during the experiment and make inconsistent judgements.

The implication for this project was to acknowledge the possibility of users making inconsistent document and cluster evaluations, and try to achieve as big as possible a sample set to limit the effects of such variation. In order to monitor this, at the end of each experiment, users were asked to reveal if their frame of evaluation had changed between the ranking of the clusters and relevance assessment of the documents[49].

5.4.1.1 Representation of individual documents for assessment of relevance

During the 1st group of the user experiments, users were presented with a printed list of the titles of the top 50 documents and asked to mark them as relevant or non-relevant. In order to avoid any possible biases related to the ranking of the documents, ten different random orders were generated, and the documents were ranked using one of these in each session.

In addition to the list of titles, document abstracts were printed separately and users were invited to ask for the abstract if a title seemed ambiguous. But in practice, most users did not ask for any abstracts, although it was very likely that at least a few of the titles would be ambiguous in each output set.

A concern at this point was whether the amount of information provided to the users affected their judgements in a particular direction. There is evidence that users are less likely to be stringent in their relevance judgements when the amount of information given to them is low (Rees&Schultz, 67). As precision values seemed quite high (averaging 34%), it was likely that users preferred to err on the positive side if they were uncertain, rather than asking for further information. Thus, it was decided that for the second group of user experiments, title and abstract information would be merged and users would skim through the merged list to make relevance judgements. This also ensured that the information available to the implementation versus the users would be more compatible. Additionally, year, author and source information

[49] This issue was overlooked during the first set of the user experiments; during the 2nd set, users were asked in the post-questionnaire whether they would make different evaluations if they were asked to go over the same clusters and documents again.

was provided for each document. Users were not asked to read all abstracts in full, but skim through them especially for the ones whose titles seemed ambiguous.

5.4.2 Statistical tests used

Nonparametric statistical tests were used to assess the significance of experiment results for both TREC facet evaluation experiments and Inspec user experiments. This type of tests was more suitable for our purposes as they were less demanding in their assumptions compared to parametric tests and were suitable for small sample sizes.

Fexact test was used to assess the results of the TREC experiments and Wilcoxon Signed ranks test was used to assess the results of the Inspec experiments. Each of these tests are outlined below.

5.4.2.1 Fexact test

In TREC experiments, it was necessary to test whether a relationship existed between the facets of documents and the clusters they were assigned to (or were most similar). The chi-square goodness-of-fit test could not be used with the data in hand as the data sets were small. A test statistic more appropriate for small, sparse data sets is Fisher's exact test.

This test does not use the Chi-square approximation, but employs the exact probability distribution of the observed frequencies assuming marginal row and column totals to be fixed. Its value indicates the probability of having the given or a more "extreme" (i.e. higher discrepancies in proportions between different groups) table by chance. This value is then compared to the commonly used significance levels (0.01 or 0.05).

The computational effort required to use Fisher's exact test for contingency tables larger than 2x2 has severely limited its use in the past (Everitt, 92). However, in the 80's, the advent of faster algorithms and availability of inexpensive computing power has extended the bounds where the exact test is feasible. Mehta & Patel's(83) algorithm is the best known work to date.

Currently, the leading statistical software providers provide the test only for 2x2 tables in their standard packages[50]. SPSS offers an optional module, StatXact for analysis of small samples. This module is developed by Cytel Software, a company whose president is Dr. Cyrus Mehta, the developer of the algorithm mentioned above.

[50] S-Plus, SPSS, Minitab as of 1997

StatXact was not available within the university computing facilities, but Mehta&Patel's algorithm was available as a Fortran subroutine on the digital library of the ACM, as it was the topic of an article published in the ACM Transactions on Mathematical Software in 1986. Since we needed to use the algorithm for a limited number of tables, it was more practical to use this Fortran subroutine for the analysis. The algorithm was downloaded, compiled and run on Fortran-4.2. Some of the table sizes were larger than the sizes recommended as feasible by Mehta&Patel, but possibly due to further developments in computing power since their paper was published, results were generated in seconds for all of the tables.

5.4.2.2 Wilcoxon Signed Ranks Test

Wilcoxon Signed Ranks Test was used to assess the significance of the differences between the precision values achieved by clustered outputs and ranked lists in the user experiments. This test only assumes that the variable under consideration has a continuous distribution, and is applicable to the case of two related samples. It considers not only the direction of differences but also the relative magnitude of the difference (Siegel&Castellan, 88).

The null hypothesis considers the treatments X and Y as equivalent; i.e. they are samples from populations with the same means and same continuous distribution. If one treatment achieves higher ranks in many more cases than the other, the null hypothesis is rejected.

This test has a special table of critical values, however when the sample size is larger than 15, the normal distribution probabilities can be used instead. If the experimenter has predicted the direction of differences in advance, a one-tailed test is appropriate; for a two-tailed test, the probability value needs to be doubled.

5.5 Pre and post-questionnaires

5.5.1 Pre-questionnaire

In all user experiments, users were asked to describe their information need/query terms on a pre-questionnaire sheet. They were also asked if they had made any searches for this information need before and if they did whether they had found some relevant documents (see Appendix 2.3).

5.5.2 Post-questionnaire

After each experiment users were asked to state whether the experiment has helped them to find any relevant documents, and to comment about usefulness of the clusters.

During the 2nd group of the user experiments, as each user reviewed and ranked clusters in two different cluster representations, they were asked to state which one they preferred. Additionally, they were asked if there had been any changes in their perceived information need during the experiment that could make their ranking of the clusters versus the relevance judgements for the documents uncomparable. They were also asked to state whether factors like author, publication date and journal had affected their relevance decisions.

6. Research question

The basic research question of this project was assessment of the capability of search output clustering to help end-users in dealing with search outputs. Within the limits of feasibility, this question was approached from two different, but complementary angles.

Firstly, as relevance ranking is the main rival for output clustering in helping the users in dealing with search outputs, it was necessary to assess the comparative performance of output clustering against relevance ranking. For this assessment, the hypotheses were set as follows:

> **Null hypothesis(1)** : if the documents that are retrieved for a user's query are clustered, and users are asked to rank the clusters in order of expected usefulness, there will be no difference between the precision values of clusters ranked highest by the users, and the precision values of ranked lists at comparable thresholds.
>
> **Alternative hypothesis(1)** : if the documents that are retrieved for a user's query are clustered, and users are asked to rank the clusters in order of expected usefulness, clusters ranked highest by the users will have significantly higher precision values, compared to the precision values of ranked lists at comparable thresholds. [51].

In order to test these hypotheses, two groups of user experiments based on real user information needs were designed and conducted(See Section 9 on page 135 and Section 10 on page 163).

The results of the 1st group of user experiments led us to question various aspects of the implementation used in the experiments. One such aspect was related to representation of clusters for users' inspection(see Section 9.6.5.1 on page 153). To find out the optimum way to represent the clusters, an intermediate experiment was designed to compare two alternative cluster representations that were envisaged as the two most viable representation modes. The hypotheses for these experiments were similar to those used in the 1st group of experiments, except that comparison was between the two cluster representations rather than cluster representation and ranked representation:

> **Null hypothesis(1a)** : There is no difference between the two cluster representations; namely Rep-B and Rep-C with respect to their ability to show the users recognizable "relevant clusters" that include significantly high proportions of relevant documents.

Alternative hypothesis(1a) : There is a significant difference between Rep-B and Rep-C with respect to their ability to show the users recognizable "relevant clusters" that include significantly high proportions of relevant documents.

A second way of approaching the basic research question was to investigate whether clustering was capable of creating groups of documents representing distinct aspects from a query. Although seemingly more pertinent to the basic research question, the significance of this question was undermined by the difficulties involved in its assessment.

Firstly, testing this hypothesis with real user needs was not easy. It would require more time and effort on behalf of the users and participation would consequently be low. Even if reasonable numbers of participants could be persuaded to group documents into distinct clusters, their document grouping behavior would not necessarily be consistent with their document searching behavior.

The availability of data from the TREC interactive track(see Chapter 5.2.2 on page 84), i.e. a number of queries and sets of relevant documents with aspect assignments for each query, and query statements specified by the users that took part in the experiments, presented a unique opportunity to assess the algorithm in this respect without involving users.

Three different groups of hypotheses were set to assess the facet discrimination ability of the implementation based on the available data. The first was aimed to make maximum use of the aspect information, and focused on the relationship between the relevant documents' facet assignments and the clusters they would be assigned to:

Null hypothesis(2) : if the documents that are relevant to a query are clustered, there will be a random distribution of facet evaluations among clusters.

Alternative hypothesis(2) : if the documents that are relevant to a query are clustered, documents with similar facet evaluations will be gathered together in clusters in significant proportions.

Findings from the experiments designed to test these hypotheses are presented in Section 12 starting on page 197.

Hypotheses(2) were tested utilising the available facet information almost fully for the sample queries. Only the relevant documents for each query were retrieved and clustered, in order to

[51] This would also mean that a clustered output contains "irrelevant clusters" that include lower

avoid interference from non-relevant documents for which we had no facet evaluations. But it was unlikely that a user would be able to retrieve such high proportions of relevant documents with high precision, especially with their initial search statement. In order to find out what clustering could do for the output from a user query, two additional sets of hypotheses were developed.

The first of these, Hypotheses(3) aimed to test whether a relationship existed between the facets represented in a cluster from the search output and facets of the unretrieved relevant documents that were most similar to this cluster.

> **Null hypothesis(3)** : if the documents that are retrieved for a user's query are clustered, and the unretrieved relevant documents are checked for their similarity to each of these clusters, there will be no relationship between the clusters to which the documents are most similar and their facet evaluations.

> **Alternative hypothesis(3)** : if the documents that are retrieved for a user's query are clustered, and the unretrieved relevant documents are checked for their similarity to each of these clusters, a significant relationship will be found between the clusters to which the documents are most similar and their facet evaluations.

The clusters from user queries were also evaluated to find out whether there were any relationships between the retrieved documents' facet assignments and the clusters they were assigned to.

> **Null hypothesis(3a)** : if the documents that are retrieved for a user's query are clustered, there will be a random distribution of facet evaluations among clusters.

> **Alternative hypothesis(3a)** : if the documents that are retrieved for a user's query are clustered, documents with similar facet evaluations will be gathered together in clusters in significant proportions.

Findings from the experiments designed to test these hypotheses are presented in Section 13.5 starting on page 214.

Finally, as an extension to Hypotheses(3), we asked the question of whether a user could run a query, then select certain cluster(s) as relevant, and expand his query automatically on this feedback to retrieve further documents that are assigned to similar facets as the documents from the chosen clusters:

proportions of relevant documents than the ranked output at comparable threshold levels.

Null hypothesis(4) : if the documents that are retrieved for a user's query are clustered, and users expand their query based on a chosen cluster, automatic query expansion will not be capable of providing significant proportions of further relevant documents, that have similar facet evaluations to the documents in the selected cluster.

Alternative hypothesis(4) : if the documents that are retrieved for a user's query are clustered, and users expand their query based on a chosen cluster, further relevant documents, that have similar facet evaluations to the documents in the selected cluster, can be retrieved by automatic query expansion.

Findings of the experiments designed to test these hypotheses are presented in Section 14 starting on page 223.

These three hypotheses complemented each other to give us an idea of the potential of the clustering algorithm to discriminate between the facets represented by documents and also the practical issues arising in such a process.

7. Prototype implementation (with VB)

C^{3}M implementation started with a prototype based on Can&Ozkarahan's(90) article detailing the algorithm. The prototype was built on Visual Basic Version3. It was designed so as to accept a binary document-term matrix from a 'txt' file and to give as result the C^{3}M matrix, the number of clusters, cluster seed documents and the cluster assignments.

Once completed, the prototype was tested with the demo example in Can&Ozkarahan's article to ensure that its results were in conformance. It was also tested with several dummy matrices with explicit clustering structures, and these simulations also produced reasonable outcomes.

The next step was trying out the algorithm on actual search outputs using the Okapi search engine, but further programming was necessary to convert these outputs to document-by-term matrices. Unix scripting language Tcl (tool command language) was used for this purpose. Tcl is a powerful programming tool with a simple and consistent syntax. It supports dynamic data structures and regular expression handling and offers interactive development and testing that facilitates rapid and easy development. Especially its various list manipulation utilities have proved handy in this implementation.

The search system of Okapi(BSS) was written in C and there was a necessity to use some of its commands and variables for the retrieval of the search output data. CISR programmer Thien Do has provided assistance for this 'communication'. He also contributed to the implementation in later stages by writing C programs for:

1. the retrieval of a stemmed version of the documents,
2. selective retrieval of document fields, and
3. matrix multiplication (for improved performance)[52].

7.1 Testing the VB prototype with a query

The VB prototype was tested using the Inspec database on Okapi search engine. The search output used for testing was the result (top 100 documents) of the query "information retrieval search output"(Q1).

[52] As Tcl is an interpreted language, it executes more slowly than compiled C code. This proved to be an obstacle in case of matrix multiplication, a necessary stage of the algorithm. It was estimated that for a typical search output Tcl would take 8½ hours to multiply two document-by-term matrices consisting

For this query, terms were not stemmed and the lower frequency threshold for term selection was set to 5. For terms that are found in the titles, a threshold of 3 was used. The upper threshold of frequency was set to 80% of the documents both for the titles and the titles+abstracts. In practice the upper limit amounted to elimination of just a few terms in case of titles+abstracts and they were either stop words or query terms. For titles, it was generally redundant as none of the terms occurred in more than 80% of the titles.

Once these terms were gathered separately, they were combined to a single list and unique terms were chosen. Matrix formulation, then, consisted of checking this list against the term list of each document and assignment of an entry of 1 if the term occurs at least once in the title and/or the abstract and assignment of an entry of 0 if the term does not occur in the title or the abstract of a given document.

353 terms were chosen out of the output set of Q1 based on the thresholds defined above. The resulting binary matrix was converted to a Notepad 'txt' file and fed into the C^3M prototype in VB. This matrix resulted in 12 clusters.

As a means of evaluating the significance of the clusters, the documents were inspected and clustered manually before they were clustered by the C^3M algorithm for comparison with the C^3M results. This was obviously a subjective experiment of no scientific value and revealed only partial overlaps between the manual and automatic groupings. The algorithmically generated clusters had their own themes and members and there also seemed to be some arbitrariness in their structures. Some clusters were a subset of a manual cluster and some other clusters took their members from several manual clusters. There were also clusters whose members looked dispersed among and outside the manual clusters(see Figure 7-1).

of real numbers with up to six decimal digits, with C a similar multiplication could be completed in about ½ minute.

Figure 7-1 : Distribution of the clusters formed by the implementation among the clusters that were generated manually for Q1

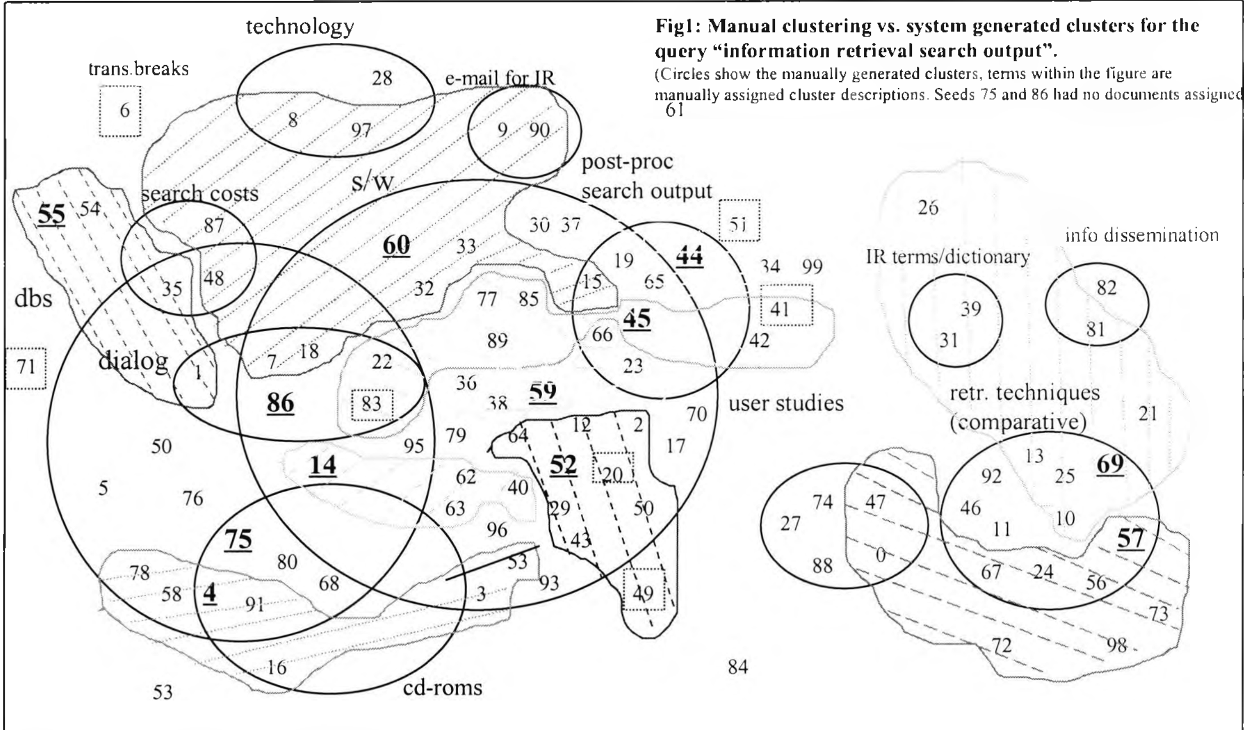


Fig1: Manual clustering vs. system generated clusters for the query "information retrieval search output".
(Circles show the manually generated clusters, terms within the figure are manually assigned cluster descriptions. Seeds 75 and 86 had no documents assigned.

CLUSTERS AND THEIR REPRESENTATIVE TERMS

cl.1(seed doc:4) : controlled, wai, serv, resorc, publication, position, cover, combin, year, book
cl.2(seed doc:14) : medline, med, grateful, issu, offer, engin, wordprocess, IR, fulltext, address
cl.3(seed doc:44) : factor, come, take, size, manag, industri, i, expansion, relevenc, related
cl.4(seed doc:45) : link, intuitiv, describ, driven, download, necessari, maual, wordproc, IR, fulltext
cl.5(seed doc:52) : d, faciliti, described, small, databank, chemical, updat, simpl, manual, grafic
cl.6(seed doc:55) : proceed, - , international, topic, publish, deliveri, chemical, industri, onlin, abstract
cl.7(seed doc:57) : best, compared, selection, proposed, paragraf, effectiveness, support, fulltext, rank
cl.8(seed doc:59) : host, pc, softwar, requir, bibliografi, numeric, grafic, i, allow, wordprocess
cl.9(seed doc:60) : e-mail, internet, alternativ, telecommunication, standard, cost, electronic, list, hard, group
cl.10(seed doc:69) : syntax, free, context, content, experimental, ranked, natural, exist, called, shown
cl.11(seed doc:75) : testdriv, performenc, corporat, acquisition, simpl, financial, qualiti, wordprocess, IR, fulltext
cl.12(seed doc:86) : post, descriptor, statistical, section, operat, individual, wordprocess, IR, fulltext, includ

8. TCL implementation

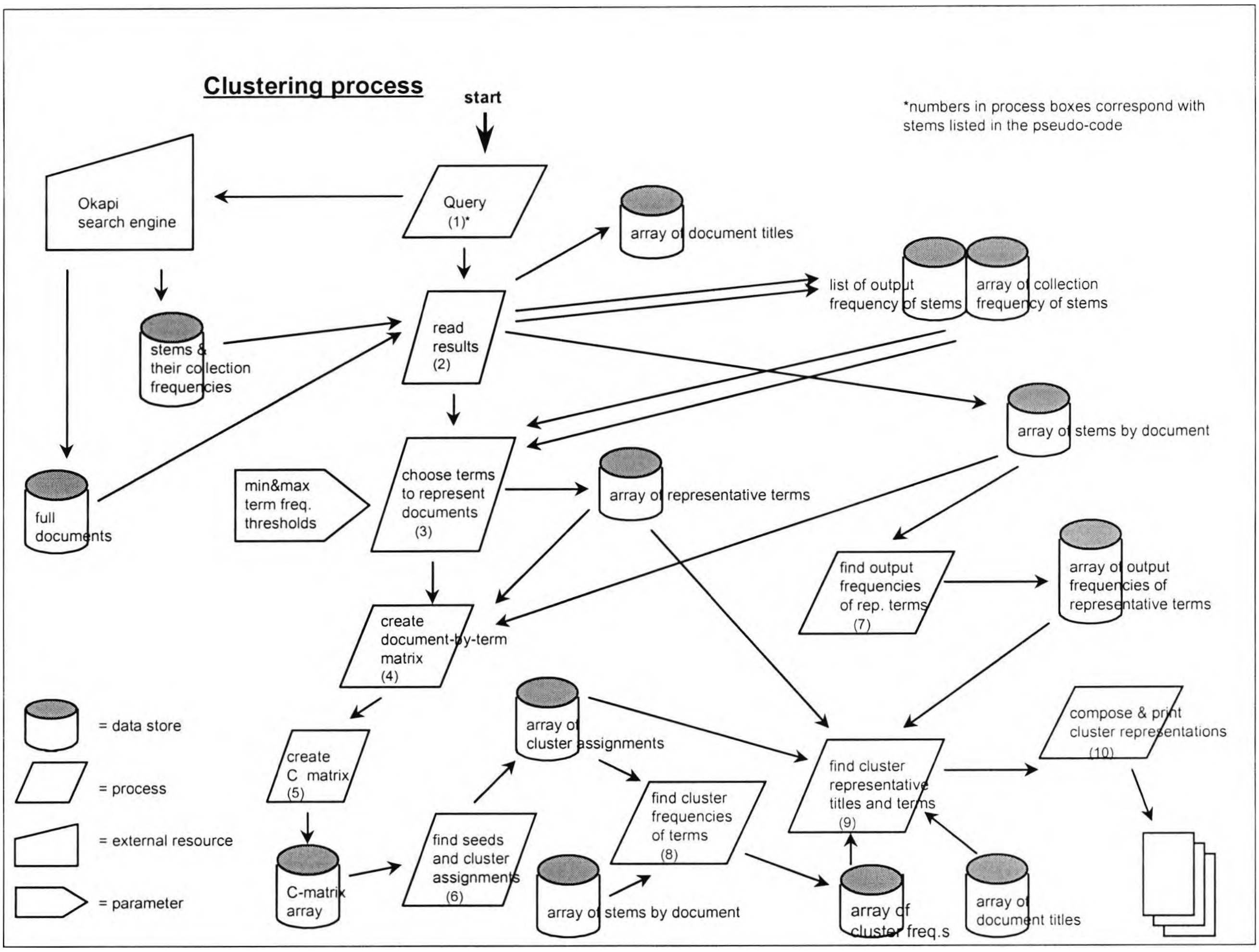
Before experimenting on different queries and on possible variations of the algorithm, it became a practical necessity to bring together the two parts of the implementation. The VB prototype was converted to TCL code in a rather straightforward way and the additional coding for checking whether seed candidates are maximally covered by a previous seed was implemented at this point(see Sectin 5.3.2 on page 92).

The only problem with the Tcl implementation turned out to be the time requirements of the matrix multiplication step and once this task was handled in C, the program returned the cluster list in about 6-7 minutes after the query terms were fed. This duration was reduced to around 2 minutes when the stopwords were indexed instead of being checked for within the algorithm (see Section 5.3.4.6 on page 104).

8.1 Pseudo-code for the TCL implementation

There is little variation between the TREC and Inspec implementations, except for the additional requirements brought about by the user involvement in the Inspec experiments. The pseudo-code below is based on the Inspec implementation, for which cluster representations was also an important issue. A data flow diagram is also provided on the next page for graphical representation.

Graph 8-1 : Pseudo-code - data flow diagram


Clustering process
start
*numbers in process boxes correspond with stems listed in the pseudo-code
Okapi search engine
Query (1)*
array of document titles
list of output frequency of stems
array of collection frequency of stems
stems & their collection frequencies
read results (2)
array of stems by document
full documents
min&max term freq. thresholds
choose terms to represent documents (3)
array of representative terms
find output frequencies of rep. terms (7)
array of output frequencies of representative terms
create document-by-term matrix (4)
= data store
= process
= external resource
= parameter
create C matrix (5)
C-matrix array
find seeds and cluster assignments (6)
array of cluster assignments
find cluster frequencies of terms (8)
array of stems by document
find cluster representative titles and terms (9)
array of cluster freq.s
array of document titles
compose & print cluster representations (10)

1. **Query:** The user's query is sent to Okapi search engine via the Enquire interface, top 50 documents are retrieved (in case of TREC experiments top 50 or all relevant documents depending on purpose) and three files are created: history, full_docs and stemmed_docs.

   History: this file contains query terms, list of retrieved documents and some other unused statistical information
   Full_docs: this file contains the whole documents, in the case of Inspec searches, it contains document identifier, title, date, author, issn number and the abstract. This file is used to extract titles for cluster representation, and to show the user document details at the final stage of the user experiments, i.e. relevance judgements for individual documents.
   Stemmed_docs: this file is a list of unique stems from each document, and their collection frequencies. Document representation is based on data from this file.

2. **Read results:** stemmed_docs and full_docs files processed to:

   - create an array of collection frequencies of the terms occurring in the documents(for use in TSV calculations) (named frq(j)),
   - create a list for each document that contains the unique terms in the document (named tok(j)),
   - extract the titles from the full_docs file and create an array of document titles (named tit(j)),
   - extract the query terms and query stems from the history file and create cur_query list and checklist lists from the query terms and query stems respectively.

3. **Choose terms to represent documents:** the parameters used for term selection are: search output frequencies, collection frequencies, upper and lower output frequency thresholds, collection size and maximum number of terms to be used for representation.

   - select the terms that fall between the maximum and minimum output frequency thresholds,
   - calculate the TSV(see Section 5.3.4.4 on page 102 for details about TSV) for each of the terms that falls within the output frequency thresholds,
   - if the list of these terms is greater than n (n=70 in the 2nd set of user experiments; it took different values for the TREC experiments depending on the number of documents retrieved) choose the top n terms with highest TSVs for document representation, otherwise use all terms in the list (array of representative terms named types(j)).

4. **Create document-by-term matrix:**

    - compare the list of representative terms against terms from each document and create entries of 1 for occurrence and 0 for non-occurrence on a matrix (two-dimensional array) with dimensions of: number of representative terms x number of documents(named mat(i,j)).

5. **Create C-matrix:** see Section 5.1.1 on page 81 for details of the C^{3}M algorithm and the C-matrix:

    - create an array containing the inverse of the number of 1 entries in each row(named xaxis),
    - create an array containing the inverse of the number of 1 entries in each column(named yaxis),
    - create a multi-dimensional array by multiplying xaxis values with each entry of mat(i,j) of the corresponding row (named int1(i,j)),
    - create a multi-dimensional array by multiplying yaxis values with each entry of mat(i,j) of the corresponding column (named int2(i,j)),
    - multiply int1(i,j) with int2(i,j) to create the C-matrix (named cmat(i,j)).

6. **Find seeds and cluster assignments:**

    - extract the decoupling (diagonal entries) and coupling (1-decoupling coefficient) coefficients of the documents from the C-matrix,
    - find out the number of clusters estimate by summing the decoupling coefficients(variable named clnum), take the integer value of clnum,
    - calculate seed powers of each document and sort them in decreasing order
    - assign the top “clnum” candidates as seeds, if two consecutive candidates are very similar, skip one of them and consider the next candidate(the similarity comparison was done by checking whether a candidate was maximally covered(highest cmat(i,j) value) by any of the documents assigned as seeds and if so, the candidate was not assigned as seed, see Section 5.3.2 on page 92 for details).
    - for each of the non-seed documents find the seed that maximally covers the document and assign the document to that seed; if more than one seed has the highest cmat(i,j) value, assign the document to all maximally covering seeds; if the document is not covered by any seeds (cmat(i,j)=0 for all seeds), put the document into a ragbag
    - create lists of documents contained in each cluster (named j.mem)

7. **List output frequencies of representative terms:**

   - list the number of occurrences of representative terms (types(j)) among the terms from the retrieved documents (tok(j) array)

8. **Find cluster frequencies of terms:**

   - find the number of occurrences of representative terms (types(j)) among the terms from each cluster (for each cluster j, check tok(i) array for each member in j.mem).

9. **Find cluster representative titles and terms:**

   - calculate the TSV value for each term from each cluster using cluster frequency and output frequence statistics
   - choose 10 terms with highest TSV values as representative terms, exclude representative terms that have been chosen for more than one cluster
   - for Rep-D, choose two documents (apart from the seed document) that have the highest coverage by the seed, use the seed title and titles of these two documents as representative titles
   - For Rep-E, check each title from each cluster for the number of query (checklist) and representative terms they include, choose three titles with the highest number of representative and query terms as representative titles.

10. **Compose and print cluster representations:**

    - Create Rep-D and Rep-E files for users evaluation. Files contain the query, and the resulting cluster representations. Each cluster representation shows the number of documents in the cluster, three representative titles and up to 10 representative terms.

In addition to Rep-D and Rep-E files, a file containing the details of the retrieved documents is created for user's relevance judgements for individual documents. The documents are sorted in random order.

8.2 Modifying the implementation for a different version of Enquire

Before the start of the TREC user experiments (TREC-USER), Okapi programmer Thien Do left the department. This meant that the version of the Enquire interface that was used in the experiments would not be supported anymore. Given this situation, and the fact that some extra modifications in the currently used version would certainly be needed for the planned TREC experiments, it was decided that the implementation should be modified to work with a different version of Enquire, the interface developed by Mike Gatford for the TREC experiments.

This meant some duplication of effort, as Mike Gatford had to create programs from scratch to provide the data required for the implementation. The resulting data was structured somewhat differently, therefore the implementation needed to be modified to cater for this.

The main requirements for the TREC experiments were:

1. For all TREC experiments, creating separate files for the best passages and the full documents (see Section 5.3.4.7 on page 104),
2. For TREC-REL experiments, retrieving top 1000 documents and returning as the query result only those that are among the list of relevant documents for a given query (see Section 12.3.1 on page 199),
3. For TREC-EXP experiments, enabling query expansion based on a group of documents marked as relevant (see Section 14.3 on page 224).

PART III : EXPERIMENTS

Experiments conducted in this project can be mainly divided into two groups; user experiments with Inspec database and facet experiments with the TREC database.

After the first 20 user experiments, the results and user feedback were evaluated, and various decisions were made to improve the algorithm and the experimental set-up. An intermediary group of 10 experiments were conducted to decide on the cluster representations to be used in future experiments. The changes made to the implementation and the experimental set-up made the previous and future experiments incomparable. Hence the 55 experiments conducted after the changes were not consolidated with previous data and the final assessments and conclusions for the user experiments were based solely on these 55 experiments.

Three different groups of experiments were conducted with TREC data. Each of these experiments was designed to test a different hypothesis. The results from these experiments gave us an overall idea about the ability of the algorithm to group documents based on the facets they represented, and the potential of cluster representations as a relevance expansion tool.

The experiment details and results are reported separately for each group of experiments in the following sections. Sections 9 and 10 outline the user experiments, Section 11 gives some general information related to TREC experiments and Sections 12,13 and 14 provide details on each of the 3 different groups of facet experiments.

9. First group of user experiments

After the preliminary trials and resulting adjustments, it was necessary to evaluate the implementation for its potential utility and usability. Evaluation requires experimentation either under laboratory or operational settings, or more realistically, in a compromise between the two. An implementation that is aimed at improving user-system interaction obviously requires the scrutiny of real users for conclusive results and has to be positioned as much towards the operational type end of the spectrum as possible.

Evaluative studies and in particular comparative studies, seem to involve less complication than explanatory studies where the high level of interdependence between variables and their non-deterministic nature makes it difficult to draw valid conclusions. In a comparative study, it may be easier for an experimenter to isolate a variable or to allow for interaction between two variables to the same level so that their effect can be controlled. The trade-off is that comparative studies are limited in what they can tell us.

9.1 Research question

The primary question that we tried to answer at this stage was this: can clustering of the search output be more effective than relevance ranking? In a typical session, can it enable users to view and select a higher percentage of the relevant documents than with relevance ranking?

Although the main research question of this project was the plausibility of clustering as a search output display tool, it was practically not possible to purely test an idea and the question had to be narrowed down to assessing the performance of a specific implementation that is built upon the idea. The assumptions and decisions made during the implementation would unavoidably have an important effect on the results.

9.2 Hypotheses

Null hypothesis(1) : if the documents that are retrieved for a user's query are clustered, and users are asked to rank the clusters in order of expected usefulness, there will be no difference between the precision values of clusters ranked highest by the users, and the precision values of ranked lists at comparable thresholds.

Alternative hypothesis(1) : if the documents that are retrieved for a user's query are clustered, and users are asked to rank the clusters in order of expected usefulness, clusters ranked highest by the users will have significantly higher precision values, compared to the precision values of ranked lists at comparable thresholds[53].

9.3 Experimental design

Various different user experiments could be envisaged to evaluate the effectiveness of the method. As with any other study involving users, it did not seem efficient or feasible to create a fully controlled laboratory environment, or a perfect operational setting free of any artificialities. It was therefore important to acknowledge the strengths, weaknesses and biases of any chosen solution.

In a recent study of similar context, Hearst&Pedersen(96) tested their search output clustering method by a series of experiments based on TREC/Tipster Standard Reference Collection and TREC-4 queries. Initially they focused on precision and compared the best clusters (those including highest percentage of relevant documents) with the list of relevance ranked documents at the cut-off point of size of the best cluster.

Following that, they devised a user study in which they gave 13 TREC-4 interactive track queries to each of 4 participant users. The participants were given the familiar task of "finding as many good documents as you can for a topic, in around 30 minutes, without collecting too much rubbish". Based on this study, Hearst&Pedersen measured how often, after issuing a search and clustering the results with the scatter/gather algorithm, the users chose the cluster with the highest proportion of relevant documents.

This type of experiment seemed also appropriate for the preliminary evaluation of the C^3M implementation. But there would be some differences in both gathering and analysis of the data.

First of all, users would be requested to attend the experiments with their own information needs. The documents would be retrieved ad hoc and naturally they would not come along with an attached relevance judgement made by another person. Adhering to the distinctions generally made (e.g. Belkin in Sparck Jones, 81), the focus would be on situational relevance (pertinence) which is the judgement by the user of the appropriateness of text to need based not upon the question put to system but rather on the user's entire desire and need state at the time of receiving the text.

This meant that users would have to evaluate the full set of ranked documents for relevance(after evaluation of clusters). Yet, the burden could be worth the benefits of this route, as:

1. the experiments woud be more realistic as they focused on real needs,
2. the two phases of the experiment (selecting clusters from the cluster list and judging documents in the relevance ranked list) would be more comparable as the source of both would be the same person.

This design enabled us to bypass the possible complication that would stem from differences between searchers or requests which is a typical problem in comparative 'matched pair' experiments. But, in exchange, it introduced the risk of users' altering their perceptions of information need halfway through the experiment, and making inconsistent judgements. If after ranking clusters and seeing a few documents, a user changed his frame of evaluation

[53] As mentioned previously, this would also mean that a clustered output contains "irrelevant clusters"

dramatically, the relevance judgements and the cluster rankings would no longer be comparable. In fact, this risk is part of every IR experiment as independence of judgement cannot be assumed even for a single set of judgements over a set of documents and unfortunately no evaluation method that takes this into account has been devised yet (Robertson in Sparck Jones, p.17, 81).

Another consequence of basing experiments on real user needs was the necessity of attracting a larger number of users as a participant would typically have only one or two information needs.

9.4 Experimental setup: user task and the flow of the experiments

In this experiment, users were not able to freely iterate both displays for more naturalistic sessions. We did not have a user interface to facilitate such interaction, and even if such an interface existed, we would have to limit the users in this respect for comparability of the results.

The subjects were presented with top 50 documents instead of the top 100 that has been used in previous rudimentary tests. This was envisaged as an upper limit, considering that it would be too much of a burden for a subject to evaluate a bigger set. (It was likely that many of the subjects would possibly not evaluate a full set of 50 documents in a real situation). Even with 50 documents, some users found it exhausting to review and make decisions about the whole set and a couple of users actually complained about this difficulty in the post-questionnaire, when they were asked about perceived weaknesses of the experiment(see Table 10-10).

Typically, a user started by reading an introductory text that specified his task and filling a pre-questionnaire that asked him about his current information need and search experience. The stated need was turned into a search statement with as little editing as possible. The top 50 documents retrieved were clustered and the clusters were printed out for the user's examination. This took about 6-7 minutes before the indexing of the database with the final set of stopwords and about 2 minutes after the database was indexed for improving performance.

The participant was allowed to modify his search statement if he decided, after viewing the output, that the statement was not a good representation of his intended query. Otherwise, he was invited to rank the clusters in order of preference.

that include lower proportions of relevant documents than the ranked output at same threshold levels.

After ranking the clusters, the participant was asked to make relevance judgements on the 50 documents on a separate list. The 50 titles were presented to the participant in random order; he was asked to mark his judgments on this list and also encouraged to inspect the printed list of abstracts when necessary.

Upon completion of the experiments, users were asked to fill in a short post-questionnaire with open-ended questions. The post-questionnaire included questions regarding the subject's general satisfaction with the retrieval results and any problems encountered during the session. His opinion as to the informativeness/usefulness of clusters was also explored (see Appendix 2.1, Appendix 2.3 and Appendix 2.4 for the introductory leaflet and pre- and post-questionnaires and Appendix 3.1 for an example of the cluster representation of a search output).

9.5 Sampling

Attracting participants for experimental studies is not easy. In information retrieval experiments, investigators usually have to suffice with subjects who agree to take part, rather than having a pool of users from which they can make random or stratified selections.

Providing some sort of incentive may help to find participants more easily. In this experiment, the main incentive was thought to be that the experiments would be based on users' own information needs. While this incentive worked to some extent, it had to be complemented by various other compensations such as tutoring help and library tokens, to attract a sufficient number of participants.

It is not difficult to understand users' reluctance, given the uncertainties involved, and the time and mental commitment required. Even if resolving a real information need could be of value to a user, it might not be perceived as worth the degree of co-operation required of them and existence of experimental constraints may mean that the need may not be resolved as well as it could be resolved by a different source under more flexible conditions.

The Inspec database used in the experiments was not fully up to date. In order to increase the utility of the experiments for the users, after each experiment, users were provided with query results for the uncovered period.

For practical reasons, students from the Information Science or Information Systems and Technology MSc courses were invited to participate in the experiments with their information needs. These students had some experience of information retrieval in using online and print media. Therefore, they were not complete novices, and were in a better position to evaluate the clustered output while answering the post-questionnaire.

The purpose was to conduct at least 20 sessions. It was hoped that the results would be indicative enough to enable some tentative conclusions.

9.6 Analysis of the results

The analysis of data from the user experiments was based on the distribution of the relevant documents among the clusters/ranked lists. Since we envisaged clustering as an output presentation method rather than a retrieval method (the clusters were based on the output of the ranked retrieval), there was no question of comparing cumulative recall and precision values.

Comparison of a clustered versus ranked document set is not straightforward. In a ranked list, the measurement unit is a document and its position within the list is significant. In a clustered output, the displayed units are clusters, but they are also gateways to documents and the two units may be considered for analysis in a variety of ways.

The following comparisons, all of which provide an hypothetical stopping rule, seemed viable for analysis of the user experiment results:

- Number of relevant documents in clusters ranked highest by the user vs. number of relevant documents at the same threshold level in the ranked output,
- Number of relevant documents in first half of documents in both outputs,
- Number of documents necessary to inspect in order to find half of the relevant documents (with clustered documents sorted in order of user choice for determining the threshold point).

The first method was accepted as the main evaluation method for the user experiments. Clusters with the highest precision values ('best clusters') were also evaluated against ranked lists. Although the 'best clusters' may not be attainable (in the sense that users will not be able to recognise them in every case) a dramatic difference between the two types of evaluation could alert us to the need to improve the cluster representation[54].

Wilcoxon Signed Ranks Test was used to assess the significance of the differences between the precision values achieved by clustered outputs and ranked lists in the user experiments (see Section 5.4.2.2 on page 120).

[54] The flaw is that by defining best cluster as the one with the highest precision value we are giving the clustered output several chances(as many as number of clusters) against a single chance for the ranked output. It is hard to balance this situation as ranked output cannot provide us with alternatives. This issue is elaborated further in Section 9.6.5.1 on page 153.

9.6.1 Users and queries

Finding users proved to be more difficult than expected. Full time MSc students from the Information Science and Information Systems and Technology courses were approached, but the response was initially weak. Eventually, within a span of 4 months 20 experiments could be completed. A few of these were done by other research students within the department and the rest by the students from the MSc courses.

In a few cases, the candidates could not do the experiment, as their information needs required quite recent data. A list of the queries used in the experiments is presented in Table 9-1.

Table 9-1 : 1st group of user experiments - Queries

User	Query no	Query	Number of clusters	Number of relevant documents
1	1	"parallel, computing, sorting, algorithms, parallel, methods"	4	17
2	2	"visual, basic, programming, learning, Microsoft, windows, application"	4	4
3	3	"formal, models, specifications, inter-operable, distributed, heterogeneous, systems, middleware"	4	12
4	4	"children, multimedia, education, child, retrieval, information, age"	5	20
5	5	"virtual, reality, visualisation, design"	4	22
6	6	"information, technology, democracy, implications, government"	5	12
7	7	"electronical, on-line, access, user, study, archival, archives, primary, materials, images, image, retrieval"	4	13
8	8	"internet, business, strategy"	5	11
9	9	"information, resources, molecular, biology, evaluation"	5	17
10	10	"distributed, database, systems, parallel, techniques, data, warehousing, decision, support"	4	21
11	11	"IT, information, technology, cost, benefit, analysis, system, implementation, design"	5	22
12	12	"procedural, programming, languages, comparison,	4	22

		non-procedural, packages, 4gls, application, development”		
13	13	“user, interface, evaluation, criteria”	4	29
13	14	“disability, information, technology, IT”	5	35
14	15	“user, interface, evaluation, expert, hci, human, computer, interaction”	4	9
15	16	“compare, applications, development, programming, packages, 4gl”	4	19
1	17	“object, oriented, techniques, neural, network, applications”	4	11
11	18	“decision, support, systems, financial, institutions, trading, marked, equities, trend, use”	5	16
16	19	“ thesaurus, retrieval, mapping, mesh, controlled, vocabulary, end, users”	4	18
17	20	“community, information, system, cis, communities, on-line, voluntary, sector, not, for, profit, local, government, electronic, democracy”	4	13

9.6.2 The need for query editing

It was the purpose of the experiment to keep the investigator’s role to a minimum. The Okapi system, being based on best match probabilistic retrieval, enabled natural language queries to be directly given to the system as the search statement.

However, in some cases the search terms provided by the users were inadequate and failed to retrieve a reasonable amount of relevant documents. Output sets with very few relevant documents would have limited value for comparative analysis. Therefore, in the few cases where the users expressed dissatisfaction with the results, they were prompted to come up with additional terms or leave out terms that were over-represented in the outputs and queries were re-run. E.g. in Query4:

> *“children multimedia products navigation retrieval expertise quality information suitability age”*

retrieved almost no documents mentioning children, therefore another query was formed:

> *“children multi-media education child retrieval information age”*

which eventually retrieved 20 relevant documents.

9.6.3 User remarks

After each session, users were asked three questions. The first question was:

> *"Has this session provided you with any useful documents that you would like to inspect?"*

All users responded positively to this question, most of them briefly saying "yes". The responses to the second question:

> *"Did you feel that the clusters provided a useful representation of the documents in the search output?"*

were more mixed(see

Table 9-2). About half of the users mentioned some dissatisfaction with the clustered output. Some commented that clusters were not well differentiated and clear, some others complained about misleading titles. Other users appeared to find the clusters meaningful and useful.

Table 9-2 : Answers to the 2nd question in the post-questionnaire

"Overall there was useful information in the clusters." (Q1)

"Yes - although possibly adding one or two more names of representative documents would make it clearer."(Q2)

"Yes - although not much difference between each cluster it did help focus a little."(Q3)

"Yes - but difficult to identify core topics for each cluster and cluster boundaries."(Q4)

"Yes."(Q5)

"I was surprised that so many relevant documents were found – considering the initial cluster documents listed"(Q6)

"Not necessarily; the titles are of course misleading and within one cluster I might find one title that might be relevant, when I consulted the abstract I found that it was irrelevant."(Q7)

"Yes."(Q8)

"Yes, in fact the best match cluster was very good. But I found it hard to judge the relevance in the clusters judged 2-3-4."(Q9)

"I like the idea of clusters as it takes less time to traverse through one cluster and get rid of the 'noise'. On the other hand the clusters didn't cluster the documents together in the way I'd hoped, i.e. it put documents into two clusters where I'd have them into one (based on the query)."(Q10)

"I am not sure, because the documents I found relevant were not attributable to any identified cluster."(Q11)

"Yes, it allowed me to target information far quicker."(Q12)

"Yes, because it indicated which area was better suited to my enquiry".(Q13)

"Yes, although sometimes part of a cluster was interesting, which therefore meant the whole cluster was not as useful."(Q14)

"A fairly good representation with some much closer to the mark than others." (Q16)

"*Only two clusters have appeared to have relevant documents, but more were found to be relevant on inspection of the retrieved set."* (Q17)

"Not sure." (Q18)

"The clusters grouped similar things together. I am not very satisfied with the documents retrieved though. And that could affect my relevance assessment." (Q19)

"Not very - many more useful 'hits' in the full output." (Q20)

Finally to the question:

> "Have you encountered any problems during the session? or have you observed any weaknesses in the experiment design?"

The users gave the following answers:

Table 9-3 : Answers to the 3rd question in the post-questionnaire

"Some of the descriptions of the presented documents were too broad for me to make relevance judgements."(Q1)

"Yes - I should have specified more clearly the exact area of my interest."(Q2)

"No problems - although I would add an interface layer asking rankings of the terms that make up the query."(Q3)

"Only weakness was that one document cluster in each session contained non or only peripherally relevant information, but no observed weakness in experiment design."(Q4)

"Many of the titles were not self explanatory and many of the documents were quite centralised and only an aspect was perhaps relevant."(Q5)

"No." (Q6)

"No." (Q7)

"No, I haven't. However if I could see the abstracts of the documents, I could make better ranking." (Q8)

"No. Just a very minor point: it would be useful at the end when you are given the titles and abstracts to judge the relevance of each individual document to have both sets in the same order." (Q9)

"Nop!" (Q10)

"The titles of the documents were not exemplary of the article. I needed to review the abstract to make an evaluation of the relevance." (Q11)

"It is difficult to make a relevancy judgement in the absence of full information." (Q12)

"Yes, in that some of the titles were not relevant to my research". (Q13)

"Yes, the ability of search engine to understand certain words." (Q14)[55]

"No problems during the session - very straightforward and in terms of speed very good" (Q16)[56]

"Perhaps it was a little difficult to see the surprising number of relevant documents from clusters." (Q17)

"would have been able to narrow the search (refine) if able to use NOT." (Q18)

"Making relevance judgements without abstracts is hard. Titles are very general or I don't understand them. The documents retrieved are very general, they maybe useful as background information. Sometimes I feel I have said documents were relevant within the set (some kind of relative relevance) but they are not what I expected." (Q19)

"I wasn't clear about the purpose of the clusters." (Q20)

9.6.4 Tabular and graphical analysis of the results

[55] The problem here was not being able to get results for certain queries if the output contained terms that could not be parsed by the BSS. The problem, in that case, was bypassed by excluding that document from the output.

[56] These comments are not really applicable as the user had in mind the performance of web search engines when evaluating the implementation.

The data from the experiments did not seem to have an explicit direction and it was not likely that it could be adequate for statistical analysis. However, there were some patterns that pointed to the need for improving the implementation and the experimental design, and it appeared reasonable to evaluate the existing data, identify ways of improvement and feed these back to the implementation before proceeding with further experiments.

Clusters were ranked and analysed in various ways for comparison against ranked lists. Primarily, they were considered in the order given by the user. Then, in order to see their potential effectiveness, they were compared using their precision order[57]. Finally, to check if the seed powers have any significance for relevance, clusters were sorted according to the seed powers of the cluster seeds and the clusters with highest and highest two cluster seed powers were compared to performance of ranked lists.

The following comparisons were made between the clusters and the ranked lists for each of the three different rankings of clusters:

1. Number of relevant documents in top cluster vs. number of relevant documents in the ranked list at the corresponding threshold level,
2. Number of relevant documents in top 2 clusters vs. number of relevant documents in the ranked list at corresponding threshold level,
3. Number of documents that had to be reviewed until half of the relevant documents are found.

The results can be summarised as follows:

1. The number of relevant documents in the top cluster was higher than those at the comparable threshold levels in the ranked list in a slightly higher number of queries than vice versa. However the difference seemed significant only in comparison of the ranked list with clusters sorted by precision (see Table 9-4, Table 9-5 and Table 9-6). Graphical representation of precision values by queries is provided in Figure 9-1 to Figure 9-3.
2. User sorted clusters seemed to perform equally well as ranked lists at top 2 cluster threshold level. System sorted and precision sorted clusters both performed better than the ranked list (see Table 9-5). Graphical representation of precision values by queries is provided in Figure 9-6.
3. When the clusters and the ranked lists were compared as to which one provided half of the relevant documents with the user having to inspect lesser numbers of documents, clusters performed slightly better than the ranked lists, the difference being most significant in the

[57] While sorting clusters by precision, if two clusters had equal precision, the number of documents and the number of relevant documents from these clusters were averaged to determine the thresholds for comparison with the ranked list.

case of clusters sorted by precision(Table 9-6). Graphical representation by queries is provided in Figure 9-7 and Figure 9-8.

Table 9-4 : Top cluster vs. ranked list - comparison of number of relevant documents

ranked list vs.:	Number of queries where clusters provided more relevant documents	Number of queries where ranked list provided more relevant documents	Number of queries where both performed equal
user sorted clusters	9	6	5
system sorted clusters	9	8	3
clusters sorted by precision	14	2	4

Table 9-5 : Top 2 clusters vs. ranked list - comparison of number of relevant documents

ranked list vs.:	Number of queries where clusters provided more relevant documents	Number of queries where ranked list provided more relevant documents	Number of queries where both performed equal
user sorted clusters	7	7	6
system sorted clusters	13	6	1
clusters sorted by precision	14	0	6

Table 9-6 : Number of documents to be reviewed until half of relevant documents are found

ranked list vs.:	Number of queries where clusters performed better	Number of queries where ranked list performed better	Number of queries where both performed equal
user sorted clusters	10	8	2
system sorted clusters	13	7	0
clusters sorted by precision	17	3	0

Figure 9-1: Precision at top user cluster vs. in ranked list at similar threshold

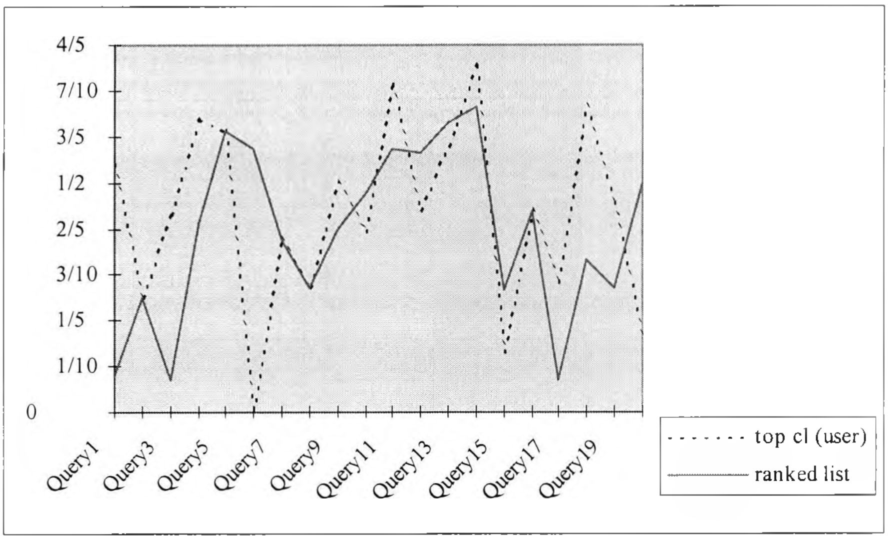


Figure 9-2 : Precision at cluster with highest seed power vs. in ranked list at similar threshold

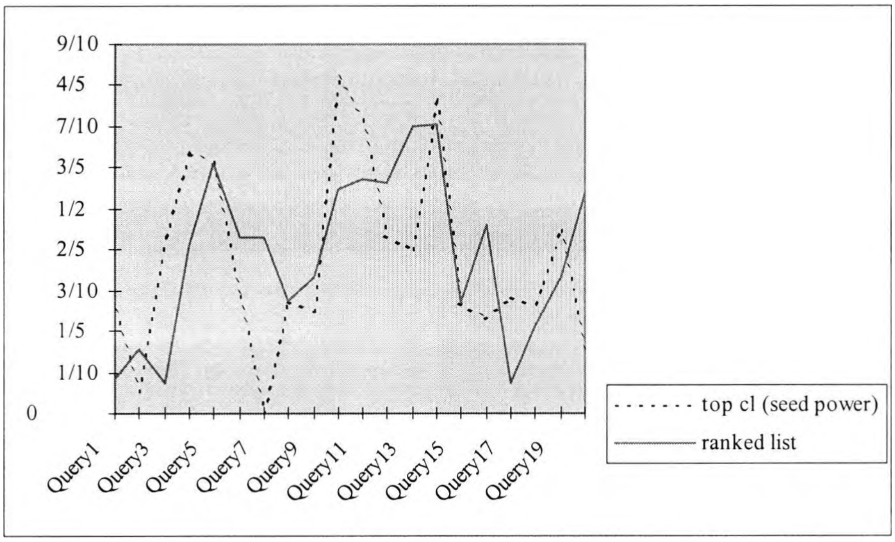


Figure 9-3 : Precision of the cluster with the highest precision vs. ranked list at similar threshold

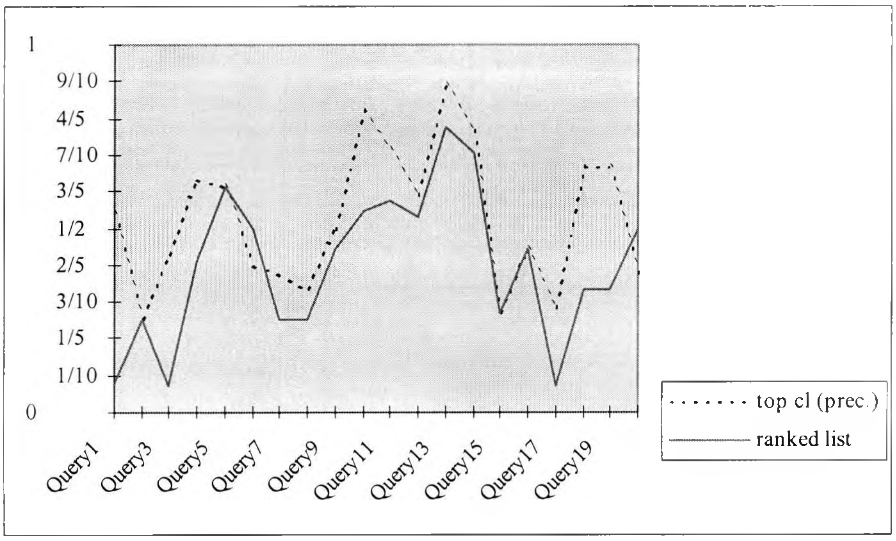


Figure 9-4 : Precision at top 2 user clusters vs. in ranked list at similar threshold

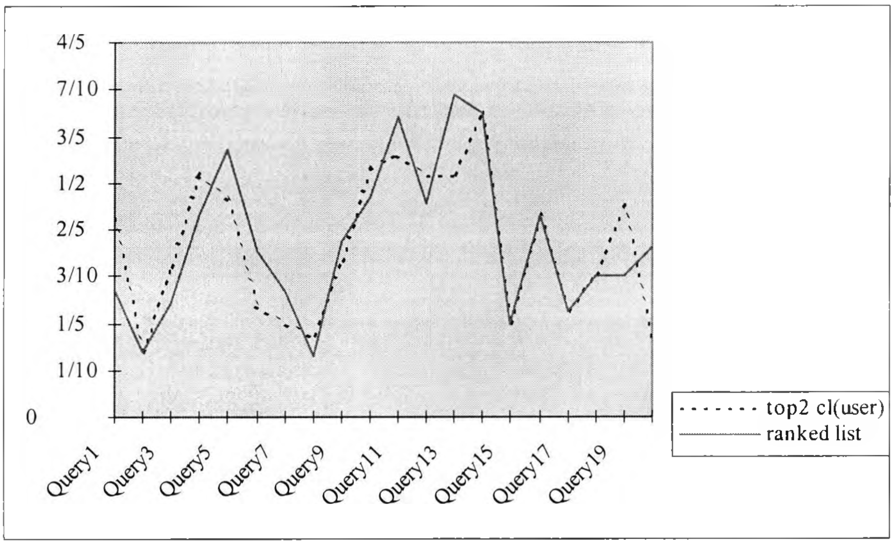


Figure 9-5 : Precision of clusters with highest cluster seed power vs. ranked list at similar threshold

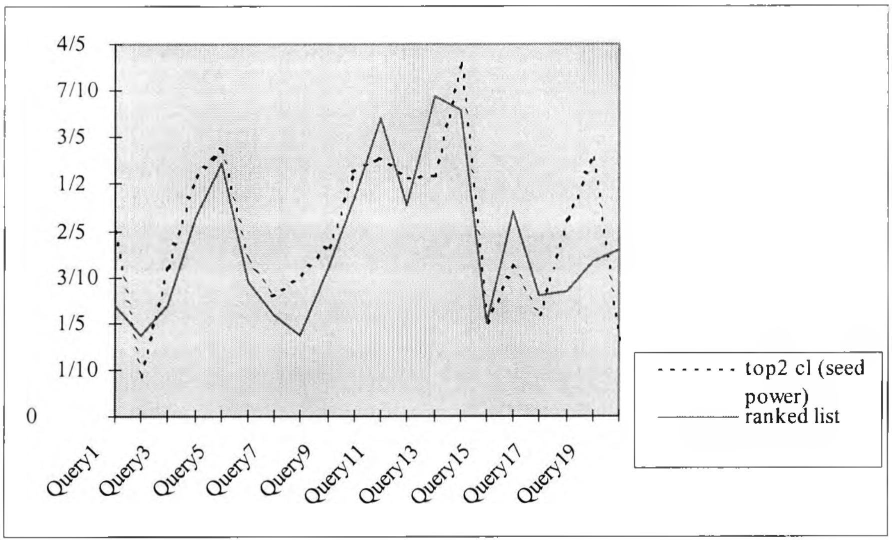


Figure 9-6 : Precision at top two clusters with the highest precision vs. in ranked list at similar threshold

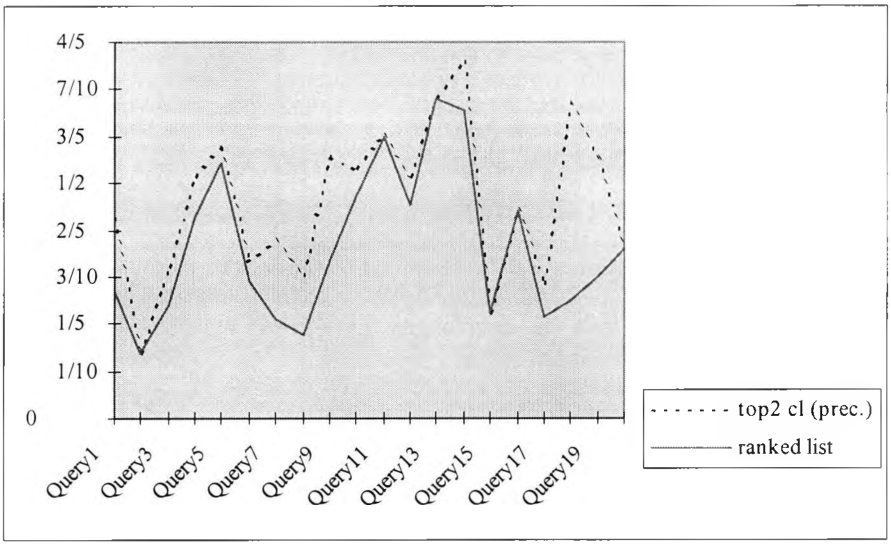


Figure 9-7 : Number of documents to be viewed until half of the relevant documents are found - Ranked list vs. user clusters

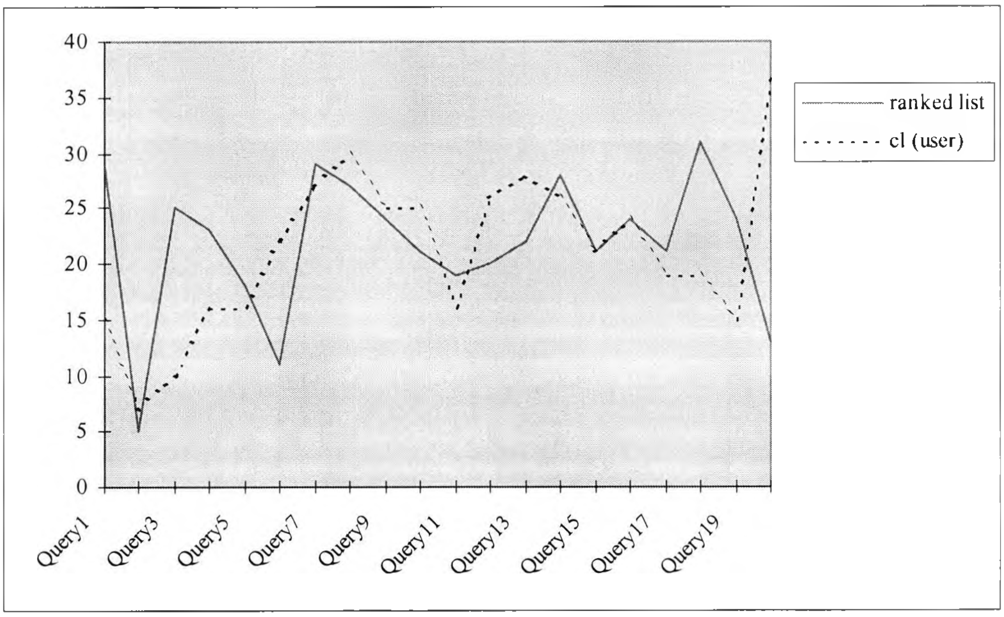
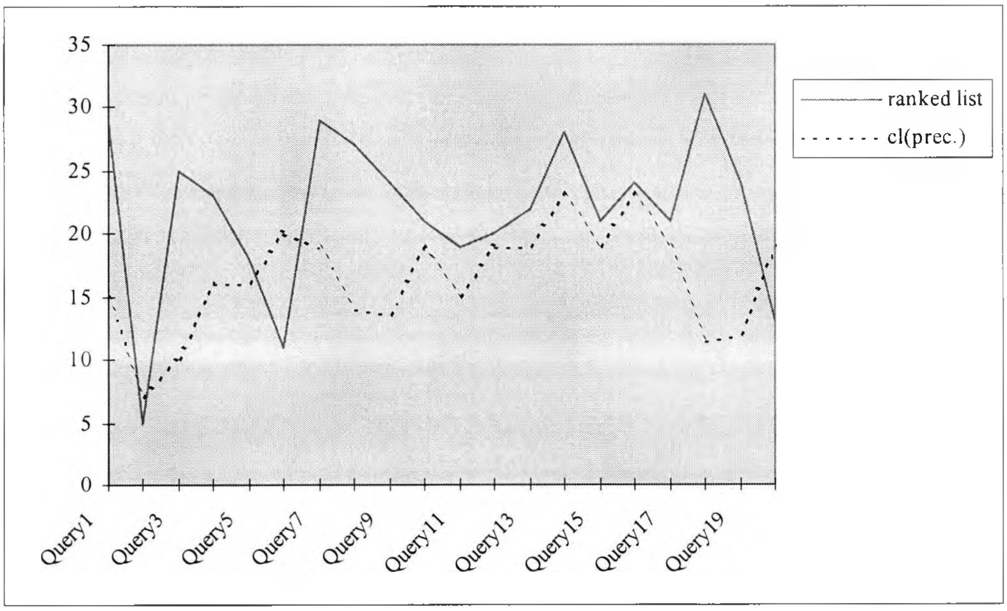


Figure 9-8 : Number of documents to be viewed until half of the relevant documents are found - Ranked list vs. clusters sorted by precision



9.6.5 Discussion of results

9.6.5.1 Significance of clusters sorted by precision

Overall, clusters performed a little better than ranked lists in retrieving more relevant documents in comparable thresholds when clusters were sorted by the users or by the cluster seed powers. The differences were bigger in the case of the top cluster sorted by the user and top 2 clusters sorted by the cluster seed power.

User and seed power sorted clusters performed better also when we compared the number of documents to be viewed until half of the relevant documents were found.

Overall, however, significant results were obtained only when clusters were sorted by precision. This could mean that the implementation was capable of creating meaningful clusters, but they were not recognisable to the users. In fact, a quarter of the users mentioned problems with judging documents by their titles, and some of them spent considerable time looking for abstracts when they were evaluating individual titles for relevance. These could be taken to indicate that the cluster representation used in these experiments (consisting of the top 10 representative terms and a few titles) might not be appropriate. In order to improve the representation, two new alternative representations were developed based on experiment results and user feedback(see Section 5.3.5 on page 111 for details). These representations were evaluated with a small number of users (see Section 9.7 on page 157), and were used in subsequent experiments with minor modifications.

While recognising the importance of representation of clusters, the variance between the performance of best clusters and clusters chosen by the users could not be explained simply as a matter of inefficient representation. Likewise, it seemed that the comparative performance of best precision clusters against the ranked lists could not be accepted at the face value. Such a comparison gave the best precision clusters more than a fair chance of outperforming the ranked list and the clusters marked first/second by the users (there is a variety of document sets from which to choose the clusters with best precision: 4-5 groups for the top cluster and 6-10 groups for the top 2 clusters). The results needed to be corrected allowing the probability of creating significant clusters randomly.

For a rudimentary test of the likelihood of having random clusters outperform the ranked list or the clusters sorted by precision, a set of clusters were generated randomly for each query. For each query the random set consisted of the same number of clusters containing the same number of documents as the original clustered output. The clusters were then sorted according to their precision values and compared against the top and top 2 clusters with best precision from the original clustered output and the precision values of corresponding ranked lists.

As expected, the significance of the clusters was dramatically reduced at both top and top 2 clusters levels. In the top cluster level, the best precision cluster outperformed the random clusters and the ranked list only in 10 of 20 cases. In a further 2 cases, it outperformed the random list and performed equally well as the ranked list. The random list outperformed the ranked list and the clustered output in 5 out of 20 cases.

When the top 2 clusters were considered, the situation got even worse. The random list outperformed the top 2 clusters and the ranked list in 10 cases whereas the clustered output gave the best results only in 6 of the 20 cases. Here again there were further 2 cases where the clustered list outperformed the random list and performed equally well as the ranked list (see Table 9-7).

Table 9-7 : Comparison of random clusters with the highest precision clusters and the relevance ranked lists

	top cluster	top 2 clusters
clusters outperform ranked and random lists	10	6
random list outperforms clusters and the ranked lists	5	10
ranked list outperforms clusters and the random list	1	0
ranked list and the clusters perform equally well and outperform the random list	2	2

The significance of the clusters sorted by precision is dramatically reduced in the light of these results. It may be concluded that best precision clusters have little meaning within a set of 50 documents and 4-5 clusters. They could be more meaningful if the document set and cluster sizes were considerably bigger as the variation in precision values would be reduced(approaching the average values as the sets grow bigger).

The significance of best precision clusters is further investigated in Chapter 10 using data from the 2nd group of user experiments. Findings from these experiments confirm the dubiousness of best precision clusters as a measure of performance for small document sets.

It is also clear that the best precision clusters have no utility unless the users can identify them. The ultimate test of any clustering structure has to be based on clusters chosen by the users.

9.6.5.2 Correlation between users' ranking of clusters and the cluster seed powers

There seems to be some kind of correlation between users' ranking of the clusters and the seed powers of the clusters' seeds. In 7 of the 20 experiments users ranked the 2 clusters with highest cluster seed powers as the best 2 clusters; and in 10 other cases, users ranked the cluster with highest or second highest cluster seed power as their first or second cluster. Only in 2 cases users chose their top 2 clusters solely from among clusters whose cluster seed powers were not among the highest two seed power values.

Clusters with highest cluster seed power values also gave good results when the comparisons were based on the number of documents to be viewed until half of the relevant documents were found.

Clusters were presented to the users in order of cluster seed power value and the order of presentation might have influenced the users' choices. But there could also be a correlation between the magnitude of cluster seed powers and the quality of clusters. In such a situation, seed power information could be utilised to decide whether to present users with clustered outputs or ranked lists. This possibility is investigated in the next section.

9.6.5.3 Predicting cluster performance

It would be useful to have a means of assessing the likelihood of clusters outperforming the ranked list for a certain output. Some candidates for such a role were envisaged as:

1. the average coverage values of documents by the seeds of the clusters they belong to,
2. the average seed powers,
3. the number of different relevance weight levels from the ranked list.

Coverage values can be thought of as a measure of intra-cluster similarity. High values may signify clusters with high coherence. Seed power is a composite measure; it favours documents with a medium number of terms that are neither too general to be considered as cluster seeds nor too specific to be unable to attract documents to their clusters once they are assigned as seeds. Low seed powers may be taken to mean that documents are clustered around a seed that is not a good representative. These two measures can tell us something about the quality of the clusters independent from the ranked list. The number of different relevance weight levels, on the other hand, may indicate the ability of relevance ranking to discriminate between the documents. If all documents from the output set have the same relevance weight, the order of the documents might not be very meaningful.

In order to have an indication of the potential of these measures in assessing cluster performance, the experiment results were divided into groups where each representation method outperformed the other, and cluster performances were compared. For assessing the effect of relevance weight levels, the experiment results were divided into groups with low/high number of different weight levels.

Analysis of the results revealed that, the first candidate, average coverage value, did not seem to differ between outputs where ranking outperformed clustering and vice versa (see Table 9-8).

Table 9-8 : Average coverage values

	within queries where ranked list performed better	within queries where cluster(s) performed better
Cluster ranked first by the user	0.045	0.049
Cluster with highest precision	0.048	0.048
Cluster with highest seed power	0.048	0.044
Clusters ranked top 2 by the user	0.050	0.047
2 clusters with highest precision	0.048	0.048
2 clusters with highest seed power	N/A	0.048

Higher average seed powers, on the other hand, seemed to correlate somewhat better with clusters outperforming ranked lists (see Table 9-9).

Table 9-9: Average seed powers

	within queries where ranked list performed better	within queries where cluster(s) performed better
Cluster ranked first by the user	1.22	1.43
Cluster with highest precision	1.13	1.39
Cluster with highest seed power	1.24	1.33
Clusters ranked top 2 by the user	1.25	1.43
2 clusters with highest precision	N/A	1.37
2 clusters with highest seed power	1.35	1.37

In order to assess whether variability of relevance weight levels in the ranked list had any correlation with the success of clustering, it was necessary to determine a threshold point above which outputs could be labelled as having high variation in their document weights. Given that outputs from the experiments had single to eight different levels, outputs with one to three different weights were described as having low variation in their relevance weights, and outputs with four or more different weight levels were described as having high variation. Based on these values, 6 cases were defined as having low variation and 13 were defined as having high variation in their relevance weight levels. Percentage of cases where clusters outperformed ranked lists in each group are presented in Table 9-10.

As expected, clusters seem to be more useful when the variance is low although the number of cases do not allow us to draw any conclusions.

Table 9-10 : Variance in number of relevance weight levels

	low variance	high variance
top cluster chosen by user outperforms ranked list	50%	43%
cluster with the highest cluster seed power outperforms ranked list	50%	43%
cluster ranked first by precision outperforms ranked list	83%	64%
top 2 clusters chosen by the user outperforms ranked list	33%	35%
2 clusters with the highest cluster seed powers outperforms ranked list	67%	64%
clusters ranked top 2 by precision outperforms ranked list	83%	64%
	n=6	n=14

9.7 Intermediary user experiments: Evaluating the alternatives for cluster representation before the 2nd group of user experiments

The experiment results and user feedback from the first group of user experiments lead us to consider ways of improving the document and cluster representation formats. Two new formats were developed to improve clarity and informativeness of the cluster representations. One of these formats used coverage values to decide which documents should be used in cluster representation; it also used 10 representative terms to complement the title information. The other format chose titles with respect to the number of query terms and representative terms included in the title. This format did not use an additional representative terms list(see Section 5.3.5 on page 111 for details).

These formats were evaluated with a small number of users who also took part in the first group of experiments.

9.7.1 Research question

A practical way of testing the possible changes to the cluster representations would be to go back to the previous users who took part in the experiments and present them with alternative cluster representations for their original queries. If they ranked the clusters in these new representations and also commented about what they thought about each different format, we could use this information in combination with their original relevance judgements for the documents and compare each alternative on both performance and user feedback.

A drawback of this idea was the possible changes in the users' state of mind since the several months that passed after their actual evaluations. Although they could be asked to try to revive the same need, as they would not be asked to make relevance judgements for the individual documents once more, there could be some discrepancy in their evaluation frame that would be used at this stage versus their original assessments made during the initial experiment.

After a few experiments with the users, it became evident that we could not rely on previous relevance judgements, as there was considerable variation between the decisions made initially and during these experiments (see Section 9.7.3 on 159). Therefore, in each experiment, users were asked to make judgements for both clusters and individual documents.

9.7.2 Hypotheses

Null hypothesis(1a): In terms of cluster representation for end-users' relevance decisions; there is no difference between:

> using titles of documents whose coverage by the seed document are highest and a set of representative terms

or

> using titles that include the highest proportions of query and representative terms without any representative terms

in enabling the users to find the most relevant (highest precision clusters).

Alternative hypothesis(1a): In representing the clusters to the users, choosing titles that contain high proportions of query and representative terms may help users to identify high precision clusters more easily than using titles of documents whose coverage by the seed document are highest and a set of representative terms.

9.7.3 Experimental design

Briefly, the experiments were conducted in the following way:

1. Users were presented with an introduction (see Appendix 2.2) similar to the one given to them in their first session and asked to try to think of their original query as a current information need.
2. The two pre-prepared output representations, one with and one without a representative term list were shown to the user one by one and he was asked to rank clusters for relevance. The representative titles were chosen with respect to the coverage values(i.e. the extend to which they are covered by the seed document) in the first representation(hereafter Rep-B) and with respect to the number of query terms and representative terms incorporated in the titles in the second representation(hereafter Rep-C); except for the seed document which was displayed as the first document regardless.
3. After completing the rankings, the user was requested to fill in a brief questionnaire about the perceived effectiveness of clustered output and comparison of the two representations regarding their usability and informativeness(see Appendix 2.5).

To test the assumption that the users would make reasonably similar judgements to those made during the 1st group of experiments, the first few users were asked to make relevance judgements for the individual documents once more and evaluate the cluster representation they were given during the 1st group of experiments along with the two new representations. This revealed that there could be considerable changes in evaluations made during the initial experiments and these experiments; users sometimes made different relevance judgements for a given document, and often ranked the clusters from the initial experiment in a different way than they did originally.

One user altered his relevance judgements for five of the documents inspected, and also marked his first cluster second and the second cluster first. The same user, for a different query, altered judgements for 4 documents and once more ranked the previously second cluster as first and the first one as second. Another user marked seven documents differently in her second experiment and ranked the clusters previously ranked second and third as first. There were also slight variations in the document sets retrieved, due to the latest changes in the GSL(see Section 9.1). These changes affected the ranking of documents and led to some documents listed in the initial sessions falling below the top 50 threshold.

Actually, given the dynamic nature of relevance, this was an expected outcome. Users' relevance judgements are known to change depending on various factors like new documents seen, changes in the perceived information need and specification of the task and the time allocated to making the relevance judgements. Given this evidence, it seemed necessary to continue with the rest of the respondents in the same way, i.e. asking them to make evaluations for the original representation and individual documents as well as the new representations.

9.7.4 Cluster representation evaluation results

Users who took part in the 1st group of user experiments were invited to attend these evaluative experiments with their original query. Among the 18 users, 9 responded positively and 10 sessions (with one user doing two sessions) were repeated.

The discrepancies between the judgements for the original representations in the 1st group versus these experiments hampered the comparability of the original representation against the new representations. If we were to use the new document judgements with the old cluster judgements, or the old document judgements with the new cluster representations in comparing the initial representation with the new ones, that would mean ignoring the time span and the changes in user's evaluation frame. On the other hand, the new document judgements could not be used with the new cluster judgements of the initial cluster representations, because the document sets from the first and current runs were not identical.

Therefore, reliable comparisons could only be made between the two new representations and even these were not conclusive as the number of cases was quite small.

The comparisons of the representations were based both on performance (i.e. precision) of top and top two clusters and the feedback provided by the users in the post-questionnaire.

Among the top ranked clusters, Rep-C gave the best results in four cases, Rep-B gave the best results in one case and the ranked list gave the best result in two cases. Among the top two clusters, Rep-B gave the best results in three cases, Rep-C gave the best results in one case, and the ranked list gave the best results in three cases. In the remainder of the cases, more than one representation gave top performance (see Table 9-11).

Table 9-11 : Performance of the two cluster representations and ranked lists

Highest precision values provided by	Top cluster	Top 2 clusters
Rep-B	1	3
Rep-C	4	1
Ranked list	2	3
Rep-B and Rep-C	1	2
Ranked list and Rep-B	0	1
All	2	0
n=10		

The post-questionnaire used in this run included two questions that could help to decide about the new format of output presentation to be used in later experiments. First was:

> *"If you compare the three different cluster representations that are provided for your inspection, which one do you think, gives you a better, more informative representation of your query results?"*

Both Rep-B and Rep-C were specified as the preferred representation in four cases. In two cases, users showed a preference for the initial representation.

The second question was:

> *"Were the representative titles and the representative terms (where available) equally useful in helping you to evaluate the clusters?"*

In 8 out of the 10 cases, users found titles to be more useful in evaluating the clusters. In the other two cases they were regarded as equally useful. Users replies to this question are given in Table 9-12.

Table 9-12 : Inspec cluster representation experiments - comments about usefulness of titles versus the representative terms in the cluster representations

"Titles give you a more immediate idea of what specific items will contain. Representative terms are necessarily general and more abstract, so they are not a very good indication of what the cluster will contain. I only used representative terms to rank two clusters that were similarly relevant."

"Some of the terms for a document I know to be what I required, were not particularly representative of the documents contents."

"[I based my decisions more on titles] but terms are useful to see how the system interpreted my query."

"Some of the representative terms could have been useful in an interactive search as ideas for reformulating the search."

"The terms are confusing, relevance decision making is achievable based merely on the titles."

"[I based my decisions more on titles], I should have based decisions on representative terms too, but chose in haste."

It seems that representative terms were generally utilised as complementary to the information conveyed by the titles. There was also some evidence that, representative terms may prove useful even when users comment otherwise. For example, one user stated that he based his decisions more on titles, and found the representative terms general and more abstract; but when asked which representation seemed to depict the query results better, he chose Rep-B which includes representative terms. Another user stated a preference for Rep-C, but actually the ranking he made for Rep-B gave better precision values.

The original purpose of this run was to decide whether to use representative terms or not and how to chose representative titles; i.e. from among documents with highest resemblance to the seed (Rep-B) or from among documents that contain the highest amount of query and representative terms in their titles (Rep-C).

It was easier to give an answer to the first question than the second one. Representative terms appeared to have a complementary role, although they were not as "user friendly" as the titles. But it could be better to list them after the titles in accordance with the secondary role ascribed to them by the users.

It was more difficult to decide about the second question, i.e. the question of how to choose the titles. The performance results seemed to be slightly better for Rep-C although the number of cases were not enough to draw reliable conclusions. It was hard to interpret the users preferences; firstly because equal numbers of users showed a preference for each of the representations, and additionally, even if there was a difference in these figures, we did not know, when a user chose one of the representations over the other, if this was due to lack/presence of representative terms or the way the titles were chosen.

Given the lack of sufficient evidence to prefer one type of representation over the other, the decision was made to use both types of cluster representations in the subsequent user experiments. This would mean continuing to ask users to evaluate two representations instead of one. But the two representations to be used would be somewhat different from the ones that have been evaluated. Since the representative terms appeared to be helpful in evaluating the clusters, both of the new representations would include terms, but they would be listed after the representative titles. The representation including titles from documents with highest seed powers would be referred to as Rep-D (successor to Rep-B) and the representation including titles with the highest amount of query and representative terms would be referred to as Rep-E (successor to Rep-C).

Rep-E also differed from Rep-C in that, while creating cluster representations for Rep-E, all the representative titles would be chosen from among those that have the highest proportion of query and representative terms, instead of the hybrid approach in its predecessor Rep-C where the seed document was listed as the first representative document regardless of how many query and representative terms it included. In Rep-E, seed documents would only be listed if their titles had a high proportion of query and representative terms that made them eligible for being listed.

10. Second group of user experiments

The research question and hypothesis for the 1st group of user experiments were valid for this group as well (see Sections 9.1 and 9.2 on page 136). We wanted to assess whether clustering the search output could be more effective than relevance ranking in helping the users to view and select more relevant documents in comparable time periods. Some perceived weaknesses of the first group of experiments were eliminated and the implementation was modified based on results and user feedback from the previous experiments(see Sections 5.3.4 and 5.3.5 starting on pages 95 and 111). Our target was to carry out as much as 50 user experiments, in order to be able to test our hypothesis and produce statistically meaningful results.

10.1 Experimental design

The experimental design was similar to the design used in the 1st group of user experiments (see Section 9.3 on page 136). One of the main differences from those experiments was having two new cluster representations (Rep-D and Rep-E) instead of the one representation used initially. As the results of the experiments to compare two alternative cluster representations were not conclusive, the two alternative formats, Rep-D and Rep-E were used simultaneously in these experiments (see Section 9.7). The order of their evaluation by the users was alternated between experiments in order to avoid a possible bias. The other major difference was showing the users abstracts together with the titles at the stage of relevance ranking of individual documents (during the first group of experiments, users were presented with lists of titles at this stage; they were encouraged to ask for abstracts where titles were inadequate to make relevance decisions, but many of them did not ask for abstracts).

Before starting this group of experiments, IEEE provided an update to the existing Inspec database (to cover until end of 1996) which improved our chances of attracting users. In many cases, the results from this database were complemented by identical searches on the online Inspec database for the uncovered period. Together with the abstracts retrieved as a result of the experiments, abstracts from the more recent period were printed and given to the users after the completion of the experiment. In addition to this, participants who wanted to base their search on their dissertation topics were given five library tokens in order to be able to order some of the relevant documents which were not available in the library.

10.2 Analysis of the results

10.2.1 Users and queries

A total of 57 experiments were carried out with 48 users, spanning a time period of 11 months. Six users did more than one search at different times. All users employed 2 or more terms/phrases to specify their queries. The average number of terms/phrases used was 6.4 and the median was 6. The terms used, number of relevant documents retrieved among the top 50 documents, and the number of clusters created are presented in Table 10-1.

Two experiments were excluded from evaluation (query numbers 18 and 46) as users marked over 80% of documents as relevant. Such high ratios of precision signified that users were interested in all aspects of the query, hence differentiating between aspects would not be useful and clustering solutions could not be assessed in these cases.

Table 10-1 : 2nd group of experiments - Queries

User	Query no	Query	Number of clusters	Number of relevant documents
1	1	"information technology, artificial intelligence"	6	23
2	2	"virtual reality, virtual, reality, implications"	4	32
1	3	"virtual reality, implications"	5	25
1	4	"virtual reality, social, medical"	5	14
3	5	"electronic commerce, electronic, commerce, information technology, trend, implications"	7	29
4	6	"teleworking, telecommuting, homeworking, home, employment, implications"	5	13
5	7	"learning organisation, learn, organisation"	6	22
6	8	"knowledge management, information system, implement, benefit, implication, organisation"	5	9
7	9	"networked organisations, implications"	5	32
8	10	"context learning, information retrieval"	7	20
9	11	"developing countries, information technology, government policy, government, policy, information, technology"	4	15
7	12	"learning organisation, learning, organisation, local, government, public, private, management"	4	15
10	13	"information management, implications, benefits, industry, current, development"	5	27
11	14	"information management, implementations, implications, case studies"	5	15
12	15	"networked organisations, network, organisation, current, development, implication, benefit"	5	21
7	16	"information management, business, companies, organisations, benefits, competition"	4	40
7	17	"information retrieval, system, evaluation, performance, compare, criteria, comparative"	4	35
13	18	"groupware, groupworking, bscw, networked, organization" (query excluded from analysis)	4	43
14	19	"search, engines, compare, pros, cons, yahoo, excite"	6	10

15	20	"torque, motors, capacitive, displacement, sensors, autotransformers, magnetoencephalography, meg, linear, step"	5	18
16	21	"annual report and accounts, paperless office, financial information, printed, material, user needs, behavior, screen, interaction, cost benefit analysis, future of the book, legacy systems"	5	11
17	22	"intranet, design, local, government"	5	17
18	23	"computer, based, training, adults, work, systems"	5	20
19	24	"variable, transformer, auto, finite, element, modelling, electromagnetic, fields, 3d, fe, opera, carbon, brushes, commutator, mathematical, computer, simulation, contact, resistivity"	4	28
20	25	"design, methodology, virtual, reality, environment"	5	25
21	26	"cognitive, interactive, user, centred, model, search, information, seeking, behaviour"	4	24
22	27	"journalism, newspaper, research, impact, internet"	5	28
19	28	"autotransformer, stabiliser, alternating, current, apparatus, adjustable, transformer, stabilized, contact, track"	5	15
23	29	"document, clustering, information retrieval"	5	29
24	30	"non-boolean, search, vector, space, model, probabilistic, best, information retrieval, system"	4	25
25	31	"local, authorities, councils, internet, intranet, web, sites, government"	4	13
26	32	"intranet, usage, survey, uk, organisations, companies"	4	18
27	33	"information, visualisation"	5	30
28	34	"electronic, shopping, home, barriers, online, implications, disadvantages, shortcomings, commerce"	5	36
29	35	"uk, information, broking, broker, role"	6	21
21	36	"clinical, information, retrieval, systems, design, computer, based, medical, records"	4	24
30	37	"software, evaluation, evaluate, methods"	6	16
25	38	"citation, index, information, interdisciplinary, science, asci, ahci, bradfords, law, export, cross"	5	13
1	39	"wizard, database, normalisation, procedure, programming, development, tool"	5	6
31	40	"social, technical, design, functional, non-functional,		

		case, specification, capture, scenario, coupling, system, new, output, user”	4	25
32	41	“graphic, manipulation, java, thresholding, binary, morphological, procedures, histograms, graylevel, smoothing, high, pass, low-pass”	4	20
33	42	“online, banking, implications, benefits, limitations, electronic, pc, home, internet, disadvantages, home-banking”	6	37
34	43	“thesauri, query, formulation, enhancement, formulate, enhance, end-user”	5	23
35	44	“internet, teleworker”	5	23
36	45	“stemming, algorithms, implementations, european, languages”	5	11
37	46	“NLP, natural, language, processing, information, retrieval” (query excluded from analysis)	4	44
38	47	“electronic, publishing, ep, extranet, intranet, groupware, systems”	5	14
39	48	“nuclear technology, nuclear information systems”, developing, countries, agriculture, hydrology, information, systems, behavior”	5	37
40	49	“computer, security, social, daily, lives, implications”	5	4
41	50	“intelligent, home”	6	17
42	51	“information, technology, developing, countries”	6	17
43	52	“london, underground, developments, commuter, current, future”	6	16
44	53	“telebanking, online, electronic, home, banking”	5	7
45	54	“e-mail, oganisations, research”	6	28
46	55	“information, technology, developing, countries”	6	18
47	56	“information, technology, skills, shortage”	5	19
48	57	“inis, database, nuclear, united, nations, vienna”	6	15

10.2.2 User remarks

Before the experiments, users were given a pre-questionnaire, where they were asked to describe their information need and state whether they have searched for this topic previously. They were also asked to give an indication of their online searching experience.

Most users stated that they had done more than 10 online searches during the past 12 months, which implied that they had a reasonable level of search experience (Table 10-2).

Table 10-2 : Participants' search experience

Number of searches done during the past 12 months:	Number of cases	Ratio
Less than 3	3	5%
3 - 10	4	7%
More than 10	42	76%
No answer	6	11%

Users' answers also revealed that most of them have searched for their described information need previously, although without much success in some cases(Table 10-3).

Table 10-3 : Whether previous searches were made for the described information need

Has the participant searched for this topic previously?:	Number of cases	Ratio
Yes	35	64%
No	17	31%
No answer	3	5%

After the experiments, users were asked to fill in a post-questionnaire composed of 6 questions (see Appendix 2.6).

The first question asked whether the search session has provided the user with any useful documents. All users responded positively to this question. The second question was:

> *"Did you feel that the clusters provided a useful representation of the documents in the search output?"*

In twenty-nine of the 55 experiments, users responded positively to this question. In twelve further cases, responses were positive but qualified with terms like "somewhat", "to an extent", "in some cases". In 11 experiments users gave negative responses to the question, and in 3 experiments they were uncertain (Table 10-4). Individual user responses to this question are presented in Table 10-5.

Table 10-4 : Summary of responses to the second question in the post-questionnaire

Responses	Number of cases	Ratio
Positive	29	53%
Mildly positive	12	22%
Negative	11	20%
Uncertain	3	5%

Table 10-5 : Answers to the 2nd question in the post-questionnaire

"Yes. Categorisation quickly gives you an insight into the type of documents that it will hold" (U1,Q1)

"Yes, to some extent" (U2,Q2)

"Yes" (U1,Q3)

"Yes" (U1,Q4)

"Not particularly. It is difficult to see the relationship between the title in a particular cluster" (U3,Q5)

"Yes" (U4,Q6)

"Yes, but it could be better" (U5,Q7)

"On some occasions yes, but mostly no. I did not feel the clusters provided a useful representation" (U6,Q8)

"Yes" (U7,Q9)

"Yes. Though would sometime like to go back and check which document was in which cluster" (U8,Q10)

"Yes, although only three documents were [in] list, the key words helped to determine the ranks" (U9,Q11)

"Yes" (U7,Q12*)*

"Yes it did" (U10,Q13)

"Not entirely" (U11,Q14)

"Somewhat" (U12,Q15)

"Yes" (U7,Q16)

"Yes" (U7,Q17)

"Well, not quite,...since none of the clusters have achieved three relevant articles. However, some of them have achieved an interesting two" (U14,Q19)

"Yes" (U15,Q19)

"Yes, although many of the clusters were not relevant to my information need" (U16,Q21)

"Yes" (U17,Q22)

"To a certain extent" (U18, Q23)

"No" (U19, Q24)

"In some cases" (U20, Q25)

"The clusters did group together the documents that didn't look interesting (Cluster 4 in both representations). I feel, however, that clusters are too closely based on words which is not always useful as for example the word "model" means different things in different contexts." (U21, Q26)

"I don't know" (U22, Q27)

"Not much" (U19, Q28*)*

"Some clusters contained highly similar documents, most of them relevant or irrelevant. Some clusters seemed to have leftovers - "miscellaneous documents" , e.g. Cluster 2 in Rep-E" (U23, Q29)

"Sort of. [I] prefer more information and the chance to browse" (U24, Q30)

"Yes" (U25, Q31)

"Overall, yes" (U26, Q32*)*

"Yes" (U27, Q33)

"Yes" (U28, Q34)

"Yes - but not in all cases" (U29, Q35)

"I don't know enough about the subject to assess this" (U21, Q36*)*

"Yes, although sometimes there was too little information about the content of the article"(U30, Q37*)*

"No, results more useful"(U25, Q38)

"Not really" (U1, Q39)

"No, there were many outputs that I've found useful but was not represented in the clusters" (U31, Q40)

"Yes" (U32, Q41)

"Yes" (U33, Q42)

"Yes" (U34, Q43*)*

"Fairly useful" (U35, Q44)

"Yes, though it was sometimes hard to tell by looking at the titles whether the cluster was relevant or not - e.g. Cluster 5 in Rep-D." (U36, Q45)

"Yes" (U38, Q47)

"In general, yes. But some of the titles are vague, and there is a necessity to read the abstract" (U39, Q48)

"Less satisfied than expected" (U40, Q49)

"Yes" (U41, Q50)

"Yes" (U42, Q51)

"Yes" (U43, Q52)

"Yes" (U44, Q53)

"Sometimes, but I got frustrated - if one title from one cluster appeared in another cluster, then it would have been spot on." (U45, Q54)

"Yes" (U46, Q55)

"Some of them" (U47, Q56)

"Not enough, they may hide something important, and highlight something else, which can mislead the user; especially for a first search; where you know nothing, you learn by searching" (U48, Q57)

The 3rd question asked users to compare the two alternative cluster representations they were presented with during the experiment. The representations Rep-E(where representative titles are chosen with respect to the number of query and representative terms they include) and Rep-D(where representative titles were chosen based on their similarity to the seed) were provided for their inspection once more at this stage to help with their assessment. In 29 cases(53%) Rep-E was preferred and in 19 cases(35%) Rep-D was preferred; in the remaining 7 cases, users found them to be very much similar.

As users have reviewed some kind of representation of their search outputs twice during this experiment, it was important to assess whether their perceptions of their information needs have changed from one representation(clusters) to another(ranked list). A dramatic change could affect their relevance evaluations and hamper the validity of comparisons between the clusters and the ranked lists. The 4th question aimed to assess whether this has been the case:

> *"Has the way you would describe your information need changed during your inspection of the clusters and documents? And if you were asked to go over the same clusters and documents again now, would you make different evaluations for some of them?"*

In twenty-five cases (45%), users gave negative answers to this question and in 9 cases(16%) they gave positive answers. In seven cases(13%) users stated that the information need would remain the same but there could be changes to evaluations made so far; in seven cases (13%) they stated the reverse: the description of the information need would be different but the assessments made would remain unchanged. In the remaining seven cases, users either have not responded to the question or their responses were not relevant (Table 10-6).

Table 10-6 : Summary of responses to the fourth question in the post-questionnaire

Responses	Number of cases	Ratio
"No.."	25	45%
"Yes.."	9	16%
"Information need unchanged, but evaluations would differ"	7	13%
"Information need would change, but evaluations would remain unchanged"	7	13%
No relevant response	7	13%

Users' individual responses to this question are given in Table 10-7.

Table 10-7 : Answers to the 4th question in the post-questionnaire

"I did find some documents which I thought would be irrelevant but turned out to be useful." (U1,Q1)
"Yes, viewing the abstract allows me to pin-point my information needs more accurately. It is likely there will be a variation but it will be by the margin." (U2, Q2)
"No my information needs remain the same. I might make different evaluations now since I've read through the abstracts." (U1, Q3)
"No. Maybe." (U1, Q4)
"My information need has not changed. However, if I had to go over the clusters again, I would probably make different evaluations." (U3, Q5)
"Yes, having read the abstracts, and selecting or narrowing down to areas of greater interest within this topic(s)." (U4, Q6)
"Probably." (U5, Q7)
"I would describe my information need differently but, no I would not make different evaluations of the clusters provided." (U6, Q8)
"Yes, the way I would describe my information need has changed. No, I would make the same evaluation of the clusters and documents but re-phrase the way I requested information." (U7, Q9)
"[I would] include NOT neural networks and NOT education." (U8,Q10)
"No - I would not make different evaluations if I had to *do it again."* (U9, Q11)
"Yes changed, I would use more terms like ... development, training, life long learning, re-training; yes, I would make different evaluations of the clusters now." (U7, Q12)
"Probably not." (U10, Q13)
"My information need has not altered. However, I would be more specific about what I'm looking for." (U11, Q14)
"Yes." (U12, Q15)
"No, No." (U7, Q16)

"No. No." (U7, Q17)

"Not at the moment." (U14, Q19)

"No." (U15, Q20)

"No." (U16, Q21)

"How they are grouped together - some clusters had more relevant documents. Some titles didn't give enough information about the subject area." (U17, Q22)[58]

"No change to information needs; might make a slightly different evaluation." (U18, Q23)

"No." (U19, Q24)

"No, No." (U20, Q25)

"Definitely YES. As soon as I saw the initial clusters, I would have reformulated my query, possibly looking for more documents within Cluster 1! By iteratively creating clusters I would have eventually get better results. I would say no [to second part of the question], probably reading some of the papers would change my relevance assessment as sometimes the title/abstract is not informative enough(especially for collective work)." (U21, Q26)

"No. I am quite satisfied about how I evaluated the documents and the clusters." (U22, Q27)

"Yes." (U19, Q28)

Not answered (U23, Q29)

"Yes. No." (U24, Q30)

"No. No." (U25, Q31)

"In the main, I don't think my cluster evaluations would change. However, looking at the document results set - would need to find an alternative to simply - 'survey' as this brought up surveys not related to intranets.." (U26, Q32)

"Information need has changed. The evaluation will be same." (U27, Q33)

"No, Yes." (U28, Q34)

"No." (U29, Q35)

"Yes, as this is the first time I search on this subject; the titles and abstracts have given me pointers to other areas I should search for. 2) I don't think so. Actually reading some of the documents would change it." (U21, Q36)

"Yes, I think I need to be more specific in my searching in the first place. After reading a summary of each article, I realised that some of them didn't have the information I expected them to have." (U30, Q37)

"No, No." (U25, Q38)

"No. It has not changed. My evaluation would be the same." (U1, Q39)

"No, my information need is still the same. Probably because I know more after reading the abstracts." (U31, Q40)

"No." (U32, Q41)

"No." (U33, Q42)

"I would add some more facets into my formulation of information need. The evaluation of clusters would not change if I went through them again." (U34, Q43)

[58] Probably explaining her answer to the previous question.

"Yes, probably." (U35, Q44)

"a)Not changed. b)Yes." (U36, Q45)

"No, in both questions." (U38, Q47)

"The information need is not changed. However, after reading the abstracts thoroughly, I felt that I may change the ranking given to the clusters." (U39, Q48)

"I think it might be better if I change the description." (U40, Q49)

"No." (U41, Q50)

"Information need doesn't change. So also the rating." (U42, Q51)

"Perhaps I need to be more precise about what I want." (U43, Q52)

"No." (U44, Q53)

"Looking at abstracts was more useful than looking at clusters alone. I would need to look at abstracts having been given clusters, but not vice versa." (U45, Q54)[59]

"No." (U46, Q55)

"No." (U47, Q56)

"Yes, I would describe it in a different way. Yes." (U48, Q57)

These answers are important as they show that users' needs and evaluation criteria can change during the evaluation process. Such changes may hamper the comparability of results from clustered outputs and ranked lists. However, it is difficult to assess and accommodate the effect of such change or pinpoint a single point in the process where the change has occurred. This is actually an inherent problem of IR user experiments (see Section 9.3 on page 136); independence of judgement cannot be assumed even for a single set of judgements over a set of documents and unfortunately no evaluation method that takes this into account has been devised yet (Robertson in Sparck Jones, p.17, 81).

Taking the users responses at face value, we might assume that in a majority of cases where users gave a relevant answer to this question, they have stated that the evaluations would not be altered (67%) while in 33% of the cases they have stated that they would make different evaluations if they were asked to do it again. We do not know the extent and consequences of the change in those cases and excluding these experiments from the analysis would reduce our data set considerably. Hence, despite acknowledging the complication introduced by this situation, the analysis was carried out on the whole data set.

The fifth question asked whether users paid attention to a document's author(s), publication year, and the journal/availability in their assessment of relevance. In 58% of the cases, users stated that one or more of these factors were influential in their decision making. In 11 cases,

[59] In all experiments users were given the two cluster representations, and after completion of their ranking(separately), they were given the list of abstracts. They did not have access to the cluster representations while evaluating the abstracts. The user's statement is confusing here, as no user was asked to rank clusters after reviewing the abstracts.

users stated that the author of the document played a role in their relevance decisions. Publication year and journal/availability were marked as important factors in 21 and 18 cases respectively (Table 10-8).

Table 10-8 : Factors involved in relevance judgement

Factors involved in relevance judgement:	Number of cases	Ratio
Author	11	(20%)
Publication year	21	(38%)
The journal/its availability	18	(33%)
Overall:		
One or more of the above	32	(58%)

These responses introduced further complexity to the analysis of the results, as these factors were not incorporated in the algorithm. Both the clustering and ranking algorithms considered only the term occurrences, whereas users considered additional factors. More importantly, users did not have access to the author/journal/date information while ranking the clusters, but saw them while making relevance judgements for individual documents.

In order to get a rudimentary indication of the extent to which these additional factors might have affected the results; output sets from twenty experiments were analysed for the publishing year of both relevant and non-relevant documents. Ten of these experiments were chosen from among those where users stated that the publishing year had been a factor in their decisions and ten were chosen from those where the users maintained that the publishing year has not been a factor in their decisions.

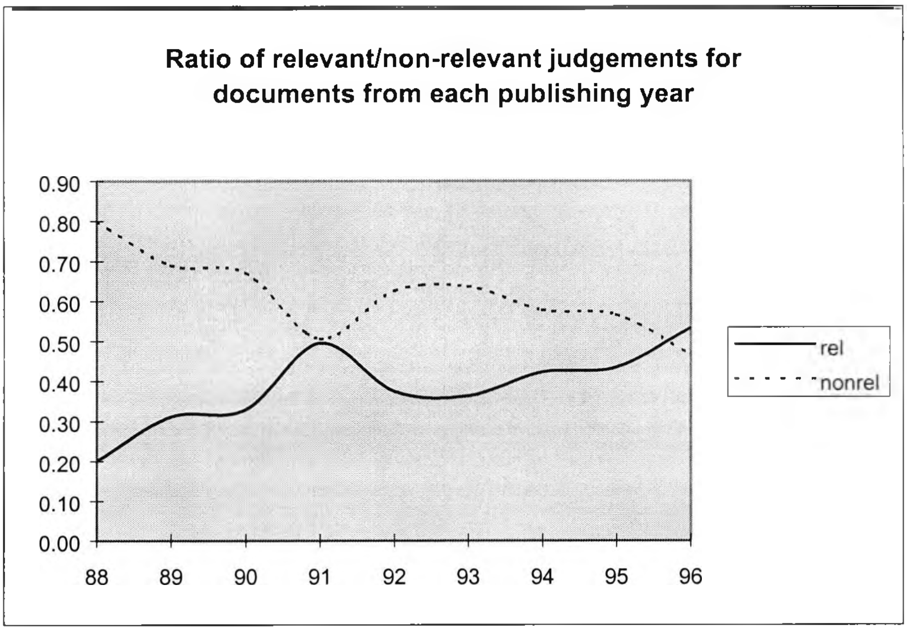
Interestingly, no difference was found among the two groups as far as the median publishing years for relevant and non-relevant documents were concerned(Table 10-9). The documents spanned the period between 1988 and 1996. The median publishing year was 1994 for the relevant documents and 1993 for the non-relevant documents for both groups. Actually, the difference between the average values was lower among the users who stated that the publishing year was a factor in their relevance judgements (93.4 versus 93 for this group as opposed to 93.3 versus 92.5 for those who stated that the publishing year decisions did not matter). This might be taken to mean that although some of the users did not particularly pay attention to the publishing year information, they took into consideration some kind of recency factor, which was conveyed within the document context.

Table 10-9 : Effect of publishing year on relevance judgements

	Relevant documents		Non-relevant documents	
	Mean year	Median year	Mean year	Median year
Publishing year affects relevance decisions (n=10)	93.4	94	93	93
Publishing year does not affect relevance decisions (n=10)	93.3	94	92.5	93

Although the average and median publishing year values were higher among the relevant documents, the graphical depiction of the proportion of relevant/non-relevant documents for each year (cumulative for all 20 output sets) revealed that there was not a linear relationship between the probability of being relevant and publishing year(Graph 10-1). The values for 1988 were negligible as they were based on a very small number of documents60. Still, it appeared that it was more likely that a document dated before 1991 would be marked non-relevant rather than relevant. For the following years, the difference between the probabilities showed less variation.

Graph 10-1 : User experiments - ratio of relevant/non-relevant documents for documents from each publishing year



[60] Only a total of 8 documents were retrieved from this year as opposed to a range of 83-212 documents from each of the other years.

An additional factor users could consider in their evaluations was the size of the clusters. The number of documents in each cluster was indicated in cluster representations, but it was not supposed to be a factor to be considered by the users in evaluating clusters. A question was added to the questionnaire after the 26th experiment, to ask users whether they considered the cluster sizes while ranking the clusters. In all but two of the 31 subsequent experiments, users stated that they did not consider the cluster sizes in evaluating the clusters.

Finally, users were asked whether they had any problems during the session, or observed any weaknesses in the experiment design. In some cases users briefly answered "No" to this question. Some of them gave more elaborate answers, these answers are listed in Table 10-10.

Table 10-10 : Answers to the 6th question in the post-questionnaire

"1.Pre-questionnaire very open. 2.Questionable value of results if query is general." (U4, Q6)
"I will stress the fact that documents presented were a bit too much to read at a go. Probably if I had taken it away, I might be able to make a better comment/more informed." (U5, Q7)
"I may want one article from one cluster which looks brilliant but not the others in that cluster. A ranking of how many useful titles per cluster might be useful. I would want to pick individual titles not clusters". (U7, Q9)
"Partial relevance" (U8, Q10)
"It took me a few minutes before I understood what I had to do. More initial explanation would be helpful." (U9, Q11)
"One document I wasn't sure if it would be useful or not." (U7, Q12)
"Most of the information provided was suited for professional research" (U11, Q14)
"No problems as such except the fact that there are quite a number of materials to sift from. This could result in some amount of indecision, and a little confusion." (U12, Q15)
"Maybe if I was given the full list first then presented with the clusters (U13, Q18)
"Maybe to run the search again - this would enable a refining of the search queries." (U16, Q21)
"My main concern is the lack of interactivity. Also sometimes its difficult to score the clusters only based on representative documents. There is no indication of whether the terms or the titles should be favoured." (U21, Q26)
"The user reads Rep-E immediately after Rep-D, so already has an idea on the clusters and may be influenced by this." (U23, Q29)
"My main difficulty (not to do with the experiment design) was comparing cluster representations D and E." (U26, Q32)
"My ranking would probably have been slightly different if it was based on a continuum rather than a clear cut yes or no for relevance." (U28, Q34)
"Clusters were not easy to evaluate." (U29, Q35)

"When having to rank clusters, I think that just having titles wasn't useful enough, especially because I don't know much about the subject. Besides, it wasn't clear whether I should also inspect the terms or only the titles to rank the clusters." (U21, Q36)

"Lack of up to date data impeded the results of the queries." (U32, Q41)

"Yes. Could there have been a mid-point? [to be able to mark a document as 'relevant' 'somewhat relevant' and 'not relevant']. (U33, Q42)

"See Q2. The selected documents in the different representations were possibly not always significantly differentiated, e.g. clusters 1 and 2 have 2/3 documents the same in both representations though the order is different". (U36, Q45)

"No problems encountered. The experiment tried to provide relevant source of information" (U39, Q48)

"It is uncertain to me that whether the available information is not enough or the experiment design should be improved" (U40, Q49)

"No problem, but I think the 2 cluster representations are the same" (U41, Q50)

"It will be useful to restate the search terms I initially gave you" (U42, Q52)

"Ambiguity over number of documents in cluster and number of titles displayed from cluster. Journal availability raises another set of issues. Usually you search then find out how available journals are. You don't start out from available journals and do a search (referring to a previous question in the post-questionnaire) (U45, Q54)

10.2.3 Tabular and statistical analysis of results

Two alternative cluster representations were used in this experiment. Users ranked clusters for both representations in turn, and proceeded with relevance judgements for individual documents on a randomly sorted list.

For comparison of ranked list with clustered output, precision values of clusters marked best and second best by the users were compared to precision values at comparable threshold levels in the ranked list. But before this we needed to decide which cluster representation should be used for comparison. Of the two alternatives, Rep-E (the representation where existence of query and representative terms played a role in choice of representative titles) was preferred in 29 cases(53%) and Rep-D (where representative titles were chosen based on their similarity to seed document) was preferred in 19 cases (35%). The difference was considerable but not statistically significant.

While Rep-E was preferred by the users in more cases, it gave poorer precision values in comparison to Rep-D. Both at the top cluster and top 2 clusters levels, Rep-D gave higher precision values than Rep-E in more cases than vice versa (Table 10-11). However, the differences were not significant.

Table 10-11 : Comparison of precision values achieved by Rep-D vs Rep-E

Higher precision provided by:	Top cluster	Top 2 clusters	Total
Rep-D	17(31%)	22(40%)	39(35%)
Rep-E	13(24%)	12(22%)	25(23%)
Equal	25(45%)	21(38%)	46(42%)

Rep-D outperformed Rep-E also when the comparison incorporated the precision values achieved by relevance ranked lists. Rep-E gave higher precision values than Rep-D and ranked lists in only 14 of the 110 measurements (at thresholds of top and top 2 clusters), whereas Rep-D outperformed Rep-E and ranked lists in 24 measurements (Table 10-12). Hence, despite Rep-E's higher popularity with the users, Rep-D was used as the preferred representation in the analysis of the results.

Table 10-12 : Comparison of ranked list with Rep-D and Rep-E cluster representations

Number of cases where highest precision is provided by:	Top cluster level	Top 2 clusters level	Total
Relevance ranked list	18	17	35(32%)
Rep-D	10	14	24(22%)
Rep-E	6	8	14(13%)
All equal	7	4	11(10%)
Ranked list and Rep-D	3	4	7(5%)
Ranked list and Rep-E	3	0	3(3%)
Rep-D and Rep-E	8	8	16(15%)

Rep-E was excluded from the analysis and new comparisons were made for the cases where Rep-E outperformed the ranked list and Rep-D, or performed equally well. The final results for Rep-D and ranked list were then evaluated at top cluster and top two clusters levels.

Ranked lists and the clustered representations seemed to give comparable results at both top and top 2 clusters levels (Table 10-13). The sign test was insignificant at both levels. However, although the number of cases where each method yielded higher precision were almost the same, at the top cluster level there was a 10% difference between the average precision values in favour of the ranked lists (55% as compared to 50% for the clustered representation). The Wilcoxon test(see Section 5.4.2.2 on page 120) gave 2-tailed probability of 10% at this level. At the top two clusters level, the differences between the precision values were smaller(49% for ranked lists vs. 47% for clustered representation), and number of cases where clustering outperformed ranked lists was actually higher than vice versa; the Wilcoxon test gave 2-tailed probability of 92%.

Table 10-13 : Comparison of ranked list with Rep-D

Number of cases where highest	Top cluster level		Top 2 clusters level	
precision is provided by:	number	ratio	number	ratio
Relevance ranked list	22	40%	20	36%
Rep-D	21	38%	23	42%
Both	12	22%	12	22%
Total	55		55	

These results implied that clustered representation and ranked lists were almost comparable in helping the user find relevant documents. However, the difference between precision values at the top cluster level suggested that ranked lists could be preferable, especially when we take into account the time overhead involved in selecting good clusters.

It was also of interest to assess the extent to which users succeeded in identifying high precision clusters. Users identified the clusters with highest precision in only 16 of the 55 experiments (29%). In 19 experiments they ranked the cluster with the second best precision value as first. In 13 of these 19 cases, the best cluster was ranked second and the second best cluster ranked first. Only in 5 cases users ranked the 4th or 5th best cluster as the first(Table 10-14).

Table 10-14 : Users' ranking of highest precision clusters

User ranks the highest precision cluster:	Number	Ratio
1st	16	29%
2nd	19	35%
3rd	13	24%
4th - 5th	5	9%
several clusters are ranked 1st by the user, or have equal precision	2	3%

Interestingly, users seemed to be better in identifying non-relevant clusters than the relevant ones. In 19 out of 51 cases(37%) where all clusters were ranked, users have ranked the lowest precision cluster as the last[61].

When the precision values for the clusters marked last by the users were compared to the precision values for the ranked lists at the same threshold level, it was found that these clusters often had lower precision than the ranked list; i.e. considering that cluster marked last by the user had n documents, the last n documents in the ranked list had more relevant documents than those found in that cluster. The comparisons were made both at the last cluster and last two clusters levels. At the last cluster level, there was a significant difference between the average precision values (Wilcoxon test gave 2-Tailed P of 3.4%), but the differences at the last 2 clusters level was insignificant(Table 10-15).

Table 10-15 : Comparative precision values for cluster(s) marked last vs. ranked list

	Ranked list	Clusters	Equal
Average precision:			
at the last cluster threshold	40%	33%	
at the last 2 clusters threshold	37%	36%	
Number of cases where precision is higher:			
at the last cluster threshold	24(47%)	15(29%)	12(24%)
at the last 2 clusters threshold	25(49%)	17(33%)	9(18%)
n=51 (number of cases where users have marked all clusters)			

Users' ability to identify the lowest precision cluster led us to the idea of using cluster representations as a rejection tool. Experiment results implied that there would be more relevant documents among the lowest ranked documents than in the cluster ranked last by the user. If we excluded this cluster from the ranked list, the low ranked relevant documents could rise to higher ranks and improve the precision of the ranked list. In order to test this idea, ranked lists were modified to exclude the documents from the lowest ranked cluster and re-evaluated for precision for the 51 cases where users have identified a last cluster. The results were analysed both at top and top 2 clusters level.

The effect of excluding the documents from the lowest marked cluster did not always increase the precision of the ranked list. Even though some irrelevant documents were removed from higher ranks (in addition to some relevant ones), they were not always replaced by other relevant documents. At the top cluster level, the precision of the ranked list was higher in 16 cases and lower in 13 cases after this process. At the top 2 clusters level precision was higher in 18 cases and lower in 14 cases(Table 10-16). The precision of ranked lists was better than

[61] In four of the 55 cases users marked only best clusters or marked all clusters as first or second.

the precision of top clusters in 24 cases after exclusion of worst clusters as compared to 20 cases before exclusion (Table 10-17).

Table 10-16 : Precision of the ranked list after exclusion of documents from lowest ranked cluster

Effect of exclusion on precision of ranked list	Top cluster level	Top 2 clusters level
Higher	16	18
Lower	13	14
Unchanged	22	19

Table 10-17 : Precision of ranked lists before/after exclusion of lowest ranked cluster from the ranked list

Number of cases where precision is higher with:	Top cluster	Top 2 clusters
Ranked list after exclusion of last cluster	24	19
Ranked lists before exclusion of last cluster	20	18
Equal	7	14
n=51		

When the ranked lists' precision values (after excluding the documents from the lowest ranked cluster) were compared with the performance of clusters marked first and second by the users, it was found that ranked lists performed as good as or better than the clusters in 76% of the cases at top cluster level(Table 10-18). But, at the top 2 clusters level, clustered representation seemed to perform slightly better than the ranked lists.

Table 10-18 : Precision of clusters vs ranked lists after exclusion of lowest ranked cluster from the ranked list

Number of cases where precision is higher with:	Top cluster	Top 2 clusters
Ranked list	24(47%)	19(37%)
Clusters	12(24%)	23(45%)
Equal	15(29%)	9(18%)

This represented some improvement in the ranked lists over their previous performance (Table 10-13) against the clustered output at the top cluster level. After exclusion of the clusters marked last by the users, precision values provided by the ranked lists were significantly higher compared to those achieved by the clusters marked first by the users (Wilcoxon test gave a 2-tailed P value of 2.4%).

It was also of interest to see how the ranked lists performed compared to best precision clusters. Although results of these experiments did not suggest that users could actually identify these clusters (see Table 10-14); we wanted to measure the potential performance of clusters, and assess the significance of such performance.

When the performance of best clusters was compared to the performance of ranked lists at top and top 2 clusters levels, it was found that best clusters outperformed ranked lists in a majority of the cases at both levels(Table 10-19).

Table 10-19 : Performance of best clusters vs ranked lists

Number of cases where higher precision is provided by:	Top cluster level	Top 2 clusters level	Total
Best cluster(s)	33(60%)	39(71%)	72(65%)
Ranked lists	9(16%)	8(15%)	17(15%)
Equal	13(24%)	8(15%)	21(19%)
n=55			

The same calculations were carried out once more with the exclusion of the documents from the lowest marked clusters (in order to improve the performance of the ranked lists - see above). This provided only a slight improvement in the ratio of cases where ranked lists performed better than the best cluster(s); and again the ranked lists were outperformed by best precision clusters in performance (Table 10-20).

Table 10-20 : Performance of best clusters vs ranked lists - after clusters ranked last by the users are excluded from the hitlists

Number of cases where higher precision provided by:	Top cluster level	Top 2 clusters level	Total
Best cluster(s)	28(55%)	34(67%)	62(61%)
Ranked lists	11(22%)	9(18%)	20(20%)
Equal	12(24%)	8(16%)	20(20%)
n=51			

In both cases, best precision clusters significantly outperformed ranked lists at top and top 2 clusters levels. The remarkable performance of best precision clusters were in line with the results obtained for the first set of user experiments (see Table 9-4 and Table 9-5 on page 148). However, as discussed there (see Section 9.6.5.1 on page 153), this had little practical significance, and was largely due to chance. Even if we had randomly created clusters, they were likely to outperform ranked lists when sorted in precision order. The cluster sizes were not big enough to have a distribution of relevant documents that converge to the average expected figures, and divergence from averages meant getting low and high precision clusters. The smaller the cluster sizes, the higher the chances of outperforming the ranked list would be as:

- the effect of divergence would be more punctuated: one extra relevant document makes a bigger difference to precision in a set of 6 documents as compared to a set of 20 documents, and
- there would be more choices to select from, i.e. as many as the number of clusters (given a fixed number of documents).

Hence, the best clusters were not considered to be a valid benchmark, although in several studies researchers have assumed that users can identify those clusters successfully (e.g. Hearst&Pedersen,96; Zamir&Etzioni, 98).

However, it was of interest to assess the extent to which clustering formed groups of documents with higher precision values than those which could be expected under a random distribution. For each of the 55 queries, 100 random cluster distributions were created, with cluster sizes matching those originally created. Precision values were calculated for each of the clusters, and highest values were averaged to generate an approximate expected precision for the best clusters. These values were then compared to actual best precision values achieved in the experiments.

In 30(55%) out of the 55 cases, original best clusters were outperformed by the average best precision value expected under random distribution. In the remaining 25 cases, original best clusters gave higher precision values. However, although the difference in number of cases was in favour of the random distribution, original best clusters had on average 7% better precision than the values expected under random distribution.

Finally, to investigate whether the seed powers could be an indicator of cluster quality, as implied by the results of first group of user experiments(see Section 9.6.5.3 on page 155), average seed power values for the cluster seed documents were calculated for each query. The values for queries where clustered representation performed better were compared with values for queries where the relevance ranked list performed better. The differences were insignificant (2.219 for clusters, and 2.206 for ranked lists), therefore the possibility of using seed powers as an indicator of cluster quality had to be abandoned.

10.2.4 A subjective qualitative evaluation of clusters and their representations

While the clusters have been evaluated for their precision values in the above pages, their ability to group together documents representing a certain aspect of a query and to convey this aspectual information to the users have not been assessed. However, it was possible to make some qualitative evaluations in this respect, by looking at the composition of the clusters, i.e. documents included and representative terms.

The overall impression created by the cluster solutions was that they could not be relied on to consistently produce meaningful document groups that were easily recognisable by the users. At times, the distinctions between the clusters were discernible and cluster representations appeared meaningful, but at other times, documents representing a certain aspect seemed to be dispersed among different clusters and representative titles or terms made little sense. It was also evidenced that, within clusters from a certain output, some were more distinct in their facet coverage while others seemed to contain leftovers.

Some examples are provided below:

Query 17

This query involved evaluation of information retrieval systems. Query terms used were "information retrieval, system, evaluation, performance, compare, criteria, comparative".

Four clusters were created for this query. All documents related to information retrieval were brought together in clusters 1 and 2. These two clusters, between them, shared all occurrences of "computerised information retrieval"(14), "document"(9), "literature"(8), "interactive"(4), "on line"(4), "recall"(6), "relevance (9), "retrieval" (8). Cluster 1 also included all occurrences of "research"(9), "environment"(7), "academic"(5), "future"(4), "implication"(4) and 8 out of 9 postings of "search". Cluster 2 had all postings of "probabilistic"(5), and 4 out of 5 postings of "query" and "relevant".

Clusters 3 and 4 included documents that focused on performance studies/comparative evaluations of computer hardware, networks, algorithms, filters, etc. There was only one posting of the term "information" among the 16 documents found in these clusters. None of these documents discussed information retrieval.

Query 19

This query was about comparisons between Internet search engines. Query terms were "yahoo, excite, compare, search, engines, pros, cons".

Six clusters were created for this query. Clusters 4 and 6 included documents about computer architecture, networks and programming. The user marked no documents from these clusters as relevant. The 25 documents included in these two clusters had no occurrences of "information", "internet", "web", or "Web", while these words had some occurrences in each of the clusters 2, 3 and 5. Search engine names "Yahoo" and "Lycos" also occurred only in clusters 1, 2 and 3.

Of all clusters, Cluster 3 had the highest precision followed by clusters 1 and 5. These clusters were the only clusters that included any representative terms that could be related to the query (Table 10-21).

Table 10-21 : Query 19 - cluster representative terms

Cluster	Representative terms
1	electronic, future, scheme, ourselves, method, index, library, experience, retrieval, they
2	test, multiple, new, world, network, value, procedure, speed
3	effective, type, on line, world wide, web, media
4	case, cost, control, compare, application, communication, line program
5	site, engine, large, alternative, Web, task, based
6	software, design, versus, help, computer, product, user

Query 31

Query 31 involved local government web sites. Query terms used were "local, authorities, councils, internet, intranet, web, sites, government".

Most documents retrieved in response to this query seemed to be only partially relevant. Some documents discussed information systems in local governments, but did not mention use of Internet/Intranets; some others discussed Internet/Intranets, but in contexts different from local government.

Four clusters were created for this query. A majority of the documents related to local government were in Cluster 4. This cluster had 10/13 occurrences of "council", 9/14 occurrences of “authority”, 4/5 occurrences of “tax”, 4/6 occurrences of “district” and 7/20 occurrences of “government”. However, terms “intranet” and “internet” had only one occurrence, and terms “world wide”, “web”, and “Web” had no occurrence in this cluster.

Cluster 3 also had some documents related to local government, but these seemed to involve provision or searching of information on Internet as opposed to Cluster 4, which focused on local government IT systems.

Clusters 1 and 2 had the highest number of postings for terms “internet”, “intranet”, “world wide”, “web” and “Web”. The two clusters had combined postings of 25/32, 26/27, 13/16, 31/37 and 4/5 for each of these terms, respectively.

Each cluster included some documents that were marked as relevant by the user. Although no document seemed to discuss “intranets” and “local government” together, the user marked some documents as relevant for their coverage of either Intranets or local government IT systems.

Query 32

This query was about Intranet usage surveys in organisations in UK. Query terms were “intranet, usage, survey, UK, organisations, companies”.

Four clusters were created for this query. Of these four clusters, three had occurrences of Intranet in two out of three representative titles. Only one cluster, namely Cluster 3, had no occurrences of “Intranet” among its representative titles or terms. It had the lowest number of postings for the terms “intranet”(2/21) and “internet”(2/14) and no postings of “web” or “world wide” while it included the highest postings of the terms like “usage”, “survey”, “questionnaire”, “satisfaction” and “trend”. This cluster was ranked last by the user and none of its 19 documents were marked as relevant. All the other clusters achieved 44% or higher precision.

10.2.5 Discussion of results

The null hypotheses for the user experiments (see Section 9.2 on page 136) could not be rejected. Clustering did not provide the users with more relevant documents than relevance ranked lists at comparable thresholds. The precision values were quite similar, and the differences were statistically insignificant. Actually, at top cluster level, average precision values provided by relevance ranked lists were 10% better than the clustered outputs (55% versus 50% respectively).

However, the experiments did lead to an interesting and potentially useful finding:

- there were more relevant documents at the bottom of the ranked lists compared to the clusters ranked last by the user (2-Tailed P=3.4%).
- ranked lists had higher precision as compared to the clusters ranked first by the user when the clusters ranked last were excluded from the ranking (2-Tailed P=2.4%).

Thus, clustering seemed more efficient as a rejection tool, than a selection tool. While at the top 2 clusters level, clusters have yielded higher precision values in more cases (23/55 as compared to 20/55 cases for the ranked lists), the difference was far from being significant.

It is necessary to mention here some limitations of the experiments that may have affected our findings:

- In 16(29%) cases, users stated that they would make different evaluations if they went over the clusters and documents again. These users have altered their evaluation criteria (more clearly, description of their information need) at some point during their evaluation, and hence their evaluations of clusters versus documents might not be fully comparable. As mentioned before, this was an almost unavoidable consequence of user experiments of this type.
- In 32(58%) cases, one or more of the factors - author, publication year, journal/availability of the document played a role in the users' relevance decisions. These criteria were not included in the clustering or the relevance-ranking algorithm. But they still caused a bias, as the users did not have access to year/author/journal information while ranking clusters, yet considered these criteria while making relevance judgements for individual documents. This, again, affected the comparability of their evaluations of clusters versus the individual documents. However, it should also be taken into account that the contents of the documents would generally be the most important factor, and the other factors could only affect a limited number of documents (e.g. generally there would be a limited number of known authors, and limited knowledge of availability information for journals).

One further shortcoming was the use of phrases and adjacency searches in some queries. This gave the ranked list an advantage over the clustered output, as the algorithm could only recognise and use phrases if they were indexed as such. However this did not have much effect on the results as in many cases the number of phrases/adjacency terms were limited to one or two and most of the phrases used were indexed in the database (e.g. "information retrieval", "artificial intelligence"). Un-indexed phrases were employed in only 12 of the 55 experiments, and in 10 of these cases there was only one such phrase. Using one un-indexed phrase was unlikely to alter the clustering results significantly, although it would have considerable effect on relevance ranking.

11. Facet evaluation experiments

With the user experiments, the comparison was mainly based on density of relevant documents in clusters. But a set of clusters could each represent a distinct relevant aspect[62] of a query and have equal numbers/ratios of relevant documents. With the evaluation measures applied for the Inspec user experiments, such an output would have no significance, even if it perfectly discriminated and grouped documents representing each facet. A more meaningful evaluation could be made if different facets of a query topic could be identified and the resulting clusters analysed based on this facet information.

Some kind of facet information is provided by TREC(Text Retrieval Conference) experts for the documents judged relevant for the TREC interactive topics(see Section 5.2.2 on page 84 for details on TREC experiments). This valuable data set was utilised in the facet evaluation experiments in this project to assess the ability of the algorithm to group documents that represent distinct facets of a query.

As City University participates in these experiments, the databases and a GUI to access the data(Enquire interface) were already available in the department. Yet some modification of the interface was necessary to generate the files to be used by the clustering implementation (see Section 8.2 on page 133).

At the time facet experiments were started, the most recent query sets and results were from 1996. Hence, the 1996 database and the interactive queries from that year were used for the experiments.

Three different sets of experiments were planned to evaluate the algorithm in different but complementary ways.

The first set of experiments were based solely on relevant documents for each of the 7 chosen queries, and looked for whether the algorithm could cluster these documents in conformance with their aspect assignments. This experiment was designed to utilise the available facet information to the maximum, without any interference from non-relevant documents; but it was not a realistic design in the sense that output sets would not normally be composed of only relevant documents.

The second and third sets of experiments were aimed to be more realistic. They were based on the original queries (initial statements) specified by City users participating in the TREC

[62] Terms “facet” and “aspect” are used interchangeably throughout this chapter.

interactive experiments in 1996. For the second set, documents retrieved by the user queries were clustered, and each of the unretrieved relevant document's similarity to each of the clusters was measured. Then, these similarities were analysed to see if there existed a relationship between the aspects assigned to these documents and the clusters they were most similar.

The third set of experiments took the second set one step further. The user queries were run and documents from one of the clusters were marked as relevant. The query was expanded on this feedback, and the resulting search output was compared to the initial search output to see the extent to which documents representing different aspects moved in/out or up/down the ranks of the output list.

11.1 Difficulties with clustering TREC documents

TREC documents are less structured and more variable in length as compared to Inspec documents. Some structural modifications were needed in the implementation to cater for these differences. For example, while with the Inspec database the title and abstract fields were easily selected and used as input, TREC documents did not feature any field delimiters and titles needed to be extracted with some approximation[63]. Source information was also an indistinguishable part of the documents and in one query[64], several document representative terms that were actually part of source information had to be specifically excluded from the analysis.

A more serious issue was the recurring problem of having unbalanced clustering solutions. Often there was a very large cluster containing a majority of the documents from the output set and a number of very small clusters, which not only contradicted the purpose of clustering, but also made the analysis of results more difficult.

One evident cause of this problem was the higher level of similarity between TREC document representations as compared to Inspec document representations. With a typical Inspec search output(of top 50 documents), the terms used for clustering had a mean output frequency of 9 and a median output frequency of 6.5. With typical TREC queries, with similar search output sizes, the mean was 16 and the median was 15. On average, terms from TREC sets occurred in 2-3 times higher numbers of documents than Inspec sets. This increased the average similarity between documents and made it more difficult for the algorithm to partition the documents into distinct sets in a balanced way.

[63] Extracting the titles was not crucial as the experiments would not involve users and the document identifiers were available for identification purposes.
[64] Query 258, TREC-USER.

Initially, attempts were made to try out different settings for the implementation parameters to find out the best combination that would minimise the problem. Maximum term frequency threshold was reduced from 80% of the size of the output to 60%. This reduced the average level of similarity between the documents by excluding terms that occurred in more than 60% of the documents, and closed the gap between the mean/median term occurrence statistics of TREC and Inspec data. But the problem of unbalanced clusters persisted.

Following that, the representative term selection criteria was changed; term weights were replaced with TSV's(see Section 5.3.4.4 on page 102) which provided some improvement in the balance of cluster sizes for some of the queries. Additionally, the seed selection criteria was altered back to its original form suggested by Can&Ozkarahan (see Section 5.3.2 on page 92), which seemed to bring the balance of clusters to an acceptable level for the first set of TREC experiments.

Despite the above mentioned modifications, cluster sizes seemed to regress to an unbalanced state again, when the documents retrieved by the user queries were clustered (TREC-USER). With most of the queries, when the similarity of unretrieved relevant documents to the clusters from the user query were computed, almost all unretrieved documents seemed most similar to the same cluster.

The main difference between the two sets of TREC experiments was that the search outputs retrieved by the user query statements (TREC-USER) were less focused compared to the sets of relevant documents (TREC-REL). To improve the focus of the documents, the analysis was shifted from full documents to best passages(see Section 5.3.4.7 on page 104 for details).

This change improved the balance of clusters for some queries. Using best passages (labelled BP) seemed preferable to full documents (labelled FD), but still the performance was not reliable, and we could not precisely explain why they performed better than the full documents. We also did not know whether the previous modifications, especially changing the seed selection method, still represented the best approach to be used in combination with best passages (see Section 5.3.2 on page 92 for details).

In order to get a better grasp of the situation, both the sets of relevant documents (TREC-REL) and the user documents (TREC-USER) were clustered using the four combinations for each of the 7 TREC queries chosen for these experiments:

FD-NC: using full documents and the original seed selection method

FD-C: using full documents and checking seed candidates' coverage by previous seed and rejecting them if they are maximally covered by a previous seed

BP-NC: using best passages and the original seed selection method

BP-C: using best passages and checking seed candidates' coverage by previous seed and rejecting them if they are maximally covered by a previous seed

The results were then compared for:

- the average coverage values by assigned seeds(higher values would imply higher intra-cluster similarity)(see Section 5.3.2 on page 92),
- the standard deviation in cluster sizes(to see the extent to which each combination relieved us from the problem of having unbalanced clusters) (see Section 5.3.4.7 on page 104),
- the rate of overlap between clusters which was especially high in some queries in the 2nd set (see Section 5.3.1 on page 91) and
- the significance of Fexact test results(see Section 12.4.2 on page 208).

Evaluation of the results showed that BP-C (using best passages and checking seed candidates' coverage by previous seed and rejecting them if they are maximally covered by a previous seed) gave the best overall results (see Section 5.3.4.7 on page 104).

In parallel with these comparisons, the results from the TREC experiments based on user queries (TREC-USER) were also compared with results from 6 exemplary Inspec queries. Since Inspec queries created no problems as far as getting balanced clusters was concerned, we could use them as a benchmark and find out the ways in which TREC data differed from the Inspec data (see Section 5.3.4.8 on page 106).

These comparisons provided a good grasp of the cause of the problem. The inspection of document-by-term matrix distributions revealed that with the TREC data there was a tendency to have a more extreme distribution of the number of representative terms in documents, yielding a concave curve as opposed to flatter curves from Inspec queries. This distribution was the main cause of having a skewed clustering solution, and using best passages improved the balance of clusters only in cases where the curve representing the number of terms in documents was made flatter by using best passage.

This evidence led us to the idea of changing the method of selection of representative terms. Instead of choosing the terms only according to their search output frequencies and weights or TSVs within the output set, terms could be chosen in a way to ensure that each document has at least a minimum number of the representative terms. A new method of term selection was developed (see Section 5.3.4.8 on page 106) and the TREC experiments were conducted using this method in conjunction with BP-C, i.e. best passages and the method of checking seed candidates for whether they are maximally covered by any previous seeds.

The new method did ensure having a certain number of representative terms in each document. This minimum number varied between 7 and 11 when clustering the sets of relevant documents(TREC-REL), and 6 and 7 when clustering the user documents(TREC-USER) (excluding two exceptional documents among the clusters based on user documents that ended up with 3 and 5 representative terms, as they had no more terms satisfying the output occurrence threshold).

However, we did not have a similarly effective method to smooth the curve at high frequency areas. It was expected that the increasing number of terms in the low frequency area could result in lesser numbers of terms in the high frequency area as we had a fixed number (upper limit) of terms to be used in document representation. But this expectation was confirmed with only some of the queries, with other queries the new terms in the low frequency area also had occurrences in the high frequency area, and the overall effect of the new term selection method had been to shift the curve upwards in all regions. With those queries unbalanced clusters persisted, although overlaps were greatly reduced and documents with zero similarity to all seeds were eliminated.

A consequence of shifting this curve up was increasing the average similarity between the documents from a query. This in turn led to lesser numbers of clusters as compared to the previously used term selection method.

One practical way of decreasing the occurrences in the high frequency area could be to bring down the maximum term frequency threshold further. It was set to 80% for Inspec queries and 60% for TREC queries. By bringing it down to 50%, we could reduce occurrences in the high frequency area in an indirect way. The expectation was that terms occurring in 50 to 60% of the documents were likely to have more occurrences in the high frequency area than terms with lower frequencies, thus by excluding such terms we were likely to reduce the occurrences in the high frequency area. Since the selection of terms from documents would be done after the filtering of the term set based on minimum and maximum frequency thresholds, the minimum number of term occurrences in the low frequency area would not necessarily be affected from the lower maximum frequency thresholds[65].

In order to see the effects of reducing the maximum frequency threshold to 50%, all 7 queries from both first and second sets of TREC experiments were re-run using both threshold settings. Using 50% threshold improved the cluster distributions in some cases, but 60% threshold gave better distributions in some other cases (Table 11-1, Table 11-2). With such mixed results, we decided to keep the 60% threshold, as it represented a less restrictive threshold and resulted in clusters that were better in bringing together documents with similar aspect assignments (yielding more significant results in the Fexact test).

Table 11-1 : Standard deviation in cluster sizes with different term frequency thresholds - clusters based on sets of relevant documents (TREC-REL)

	50%	60%
254	3.83	5.20
256	1.73	5.00
258	2.59	1.91
264	2.99	1.00
284	8.50	8.02
286	9.01	6.65
292	10.28	9.00
average	5.56	5.25

Table 11-2 : Standard deviation in cluster sizes with different term frequency thresholds - clusters based on documents retrieved by the users (TREC-USER)

	50%	60%
254	10.37	11.56
256	6.14	8.58
258	3.56	4.69
264	2.88	2.63
284	5.59	6.16
286	7.41	6.18
292	6.04	6.50
Average	6.00	6.62

11.2 Queries excluded from experiments

Five of the 12 topics provided for the 1996 TREC interactive task were excluded from the experiments due to the distribution of aspects in their set of relevant documents (see Section 5.2.2.1 on page 86).

Topic 274 was not used because 82% of its documents were assigned to two to five aspects simultaneously, and this would deter us from having straightforward expectations as to which document should be grouped with which other documents in the clusters.

[65] Some candidate terms from documents in the low frequency area could be excluded, but there would generally be other candidates to represent the document.

Topics 260 and 293 were left out, because they had few relevant documents (7 and 8 respectively) that had facet assignments; they were not adequate sources of information for analysis.

Topic 299 seemed to suffer both problems; its set of documents with facet information was small compared to other topics, and 88% of these documents had multiple facet assignments.

Finally, topic 255 was excluded because half of its relevant documents were not assigned to any aspects, half of the remaining documents were assigned to multiple aspects and the aspect assignments of the remaining documents were too dispersed to allow any meaningful evaluation.

12. TREC queries - Clustering relevant documents (TREC-REL)

Tables and graphics related to this section are denoted “TREC-REL” to stand for TREC experiments with relevant documents.

12.1 Research question

In these experiments, the purpose was to find out the relationship between the documents' facet assignments and the groups they would be assigned to if they were clustered.

As we had facet information only for the documents deemed to be relevant to the queries by the TREC experts, these experiments were based only on the relevant documents. Non-relevant documents would make no contribution to the analysis, but they would affect the overall results by introducing their own particular facets and terms and thereby diluting the effect of facet information from the relevant documents.

Although these experiments represented no proximity to a real life situation, where both relevant and non-relevant documents would be retrieved in response to a query, it was the best approach to utilise the available information to assess the implementation's ability to differentiate between different facets of a query. This approach would not be necessary if there had been facet assignments for both relevant and non-relevant documents.

Actually, the experiments were not purely based on documents with aspect assignments. Documents that were deemed relevant but not assigned any aspects were also included in the analysis. Depending on the query, these documents formed 11-49% of the document set (Table 12-1). They were likely to reduce the strength of the measurable aspect information within the output set as they did not reinforce the official aspect descriptions but introduced information that was superfluous and possibly somewhat misleading for the purpose of this experiment. In that sense, the experiments were not undertaken under fully controlled conditions, but incorporated some level of redundant information that could be expected in a real life situation.

It is interesting that the two queries (256 and 258) for which the relationship between the aspect and cluster assignments was significant at the level of 5% -as compared to 1% for the rest of the queries- (see Section 12.4.2 on page 208) were among the three queries with the highest proportion of documents with no aspectual assignments.

Table 12-1 : Proportion of relevant documents with no aspect assignments (TREC-REL)

	Proportion	Percentage
254	6/42	14%
256	13/27	48%
258	26/54	48%
264	4/36	11%
284	15/54	28%
286	14/58	24%
292	29/59	49%

12.2 Hypotheses

Null hypothesis(2) : if the documents that are relevant to a query are clustered, there will be a random distribution of facet evaluations among clusters.

Alternative hypothesis(2) : if the documents that are relevant to a query are clustered, documents with similar facet evaluations will be gathered together in clusters in significant proportions.

12.3 Experimental design

As mentioned before, in these experiments, in order to make as much use of the available facet information as possible, only the documents that were deemed relevant for the related query were retrieved and clustered to find out if any relationships existed between the aspects of the documents and the clusters they were assigned to.

In order to carry out the experiments without extensive programming effort(see page 80 for an explanation), an attempt was made to compile the relevant document sets by sending queries to the Okapi search engine as done with the user experiments. But, even with long, comprehensive queries, it was not easy to retrieve a good proportion of such documents among the top 50-60 documents. Hence, instead of retrieving the top n documents as the output set, the output set was compiled by going through the top 1000 documents and selecting the documents that were deemed relevant by the TREC experts.

The number of relevant documents for the seven queries ranged between 39 and 72 and the number of documents that were selected and retrieved from the top 1000 documents ranged between 27 and 59 (see next section for details).

12.3.1 Compiling the set of relevant documents

Initially, query statements were composed using all good candidate terms/phrases from the topic titles and descriptions in the queries. However, these statements did not enable us to retrieve a good proportion of the 'relevant' documents among the top 1000 documents for each of the seven queries. The actual figures for the initial query statements were as follows:

Table 12-2 : Number and proportion of relevant documents retrieved with the initial query statements (TREC-REL)

Topic Number	Number of query terms/phrases	No. of relevant documents retrieved	% of relevant documents retrieved	Number of documents too long to retrieve[66]
254	11	36/42	86 %	3
256	12	21/39	54 %	1
258	13	47/57	82%	7
264	15	20/60	33 %	0
284	14	40/59	68 %	0
286	7	42/72	58%	0
292	11	34/71	48%	0

To improve the recall figures, queries were expanded with further terms obtained from aspect descriptions given by TREC experts and from the unretrieved relevant documents. But there was the limitation that the Enquire interface was configured to work with a maximum of 20 query terms, therefore it was not always possible to include all good candidate terms. After a few iterations the best results obtained were chosen for clustering. The final figures were as follows:

[66] In TREC interactive experiments City University used a maximum document size threshold of 10240 bytes; for the TREC-REL experiments, the threshold parameter was increased to 30720 for queries where aspectual relevant documents were missed because of their length.

Table 12-3 : Number and proportion of relevant documents retrieved with the final query statements (TREC-REL)

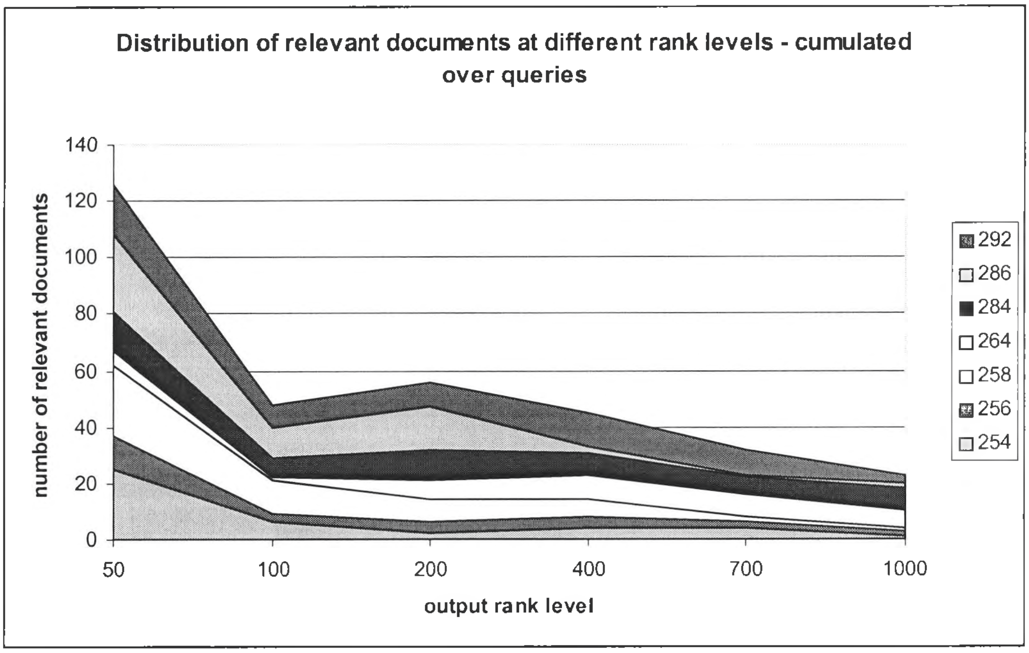
Topic Number	Number of query terms/phrases	Number of relevant documents retrieved	% of relevant documents retrieved
254	13	42/42	100 %
256	17	27/39	69 %
258	13	54/57	95%
264	20	36/60	60 %
284	20	54/59	92 %
286	20	58/72	81%
292	20	59/71	83%

The distribution of the relevant documents among the top 1000 documents is presented in Table 12-4. The cumulative values over the 7 queries are shown in Figure 12-1.

Table 12-4 : Distribution of relevant documents in the top 1000 documents retrieved (TREC-REL)

Topic Number	0-50	51-100	101-200	201-400	401-700	701-1000	Recall rate
254	25	6	2	4	4	1	100 %
256	12	3	4	4	2	2	69 %
258	25	12	8	6	2	1	95%
264	5	1	7	9	8	6	60 %
284	14	7	11	8	6	8	92 %
286	27	11	15	2	1	2	81%
292	18	8	9	12	9	3	83%

Figure 12-1 : Distribution of relevant documents at different rank levels



12.4 Analysis of results (TREC-REL)

Results are analysed qualitatively for each query in the sections below. This is followed by quantitative analysis for statistical significance. The narrative provided by TREC for each query and the query terms used to retrieve the documents are provided in Appendix 4.3. Tables showing the maximum concentration of the aspects in clusters are provided in Appendix 4.4.

12.4.1 Analysis of results by query

12.4.1.1 Query topic 254

Topic : Non-invasive procedures for persons with heart ailments

Description: What kinds of medications or procedures other than heart surgery have been used to treat heart ailments?

Recall rate: 100 % (42/42)

Number of clusters: 3

9 different aspects were identified by the TREC experts for this query. There seemed to be some significance in the distribution of aspects 1, 3, 6 and 9 (Table 12-5).

Table 12-5 : Q254 - Distribution of aspects among clusters

Aspect No	Cl-1	Cl-2	Cl-3
1	1	12	3
2	0	1	2
3	6	1	2
4	0	1	0
5	0	1	0
6	2	0	0
7	0	2	2
8	0	0	1
9	0	2	0
Cluster size	12	21	12

12.4.1.2 Query topic 256

Topic :	Negative reactions to reduced requirements for college undergraduate core studies
Description:	Colleges for a long time have been reducing their requirements in such core topics as history, literature, philosophy and science. Criticisms of this trend has occurred. What are those criticisms.
Recall rate:	69 % (27/39) , 100%(14/14) for documents assigned to aspects that are represented by more than one document.
Number of clusters:	3

TREC experts have identified 7 aspects for this topic. There was some conformance between the clusters and the facet evaluations. Eight out of 10 documents assigned to Asp-1 and both documents assigned to Asp-2 were in Cl-1. Three out of 4 documents assigned to Asp-7 were gathered in Cl-2 (Table 12-6).

Table 12-6 : Q256 - Distribution of aspects among clusters

Aspect No	Cl-1	Cl-2	Cl-3
1	8	0	2
2	2	0	0
4	1	0	1
7	1	3	0
Cluster size	9	14	4

12.4.1.3 Query topic 258

Topic : Computer Security

Description: Identify instances of illegal entry to sensitive computer networks by non-authorized personnel.

Recall rate: 95 % (54/57), 97%(28/29) for documents assigned to aspects that are represented by more than one document.

Number of clusters: 4

TREC experts have identified 24 aspects for this topic. Of these aspects, 11 have been assigned to more than one document. A peculiarity with this query was the high proportion of multiple assignments of aspects to documents. One document was assigned 8 aspects and the average number of aspects assigned to a document was 3[67].

As discussed in Section 5.2.2.1 on page 86, multiple assignments reduce the “cluster-ability”, or more directly, our ability to measure the significance of the clustering structures of the document sets. In this output set, the distribution of documents with facet assignments seemed to confirm the existence of the “cluster-ability” problem. Certain groups of aspects were brought together mainly in two clusters. The seed document in Cl-3 was assigned to aspects 1, 2, 3, 5, 6, 7, 8 and 9, and the seed document in Cl-1 was assigned to aspects 1, 5 and 13. The other two seeds were not assigned to any aspects. Almost all documents assigned to aspects 2, 3, 7 and 8 were in Cl-3. Likewise, documents assigned to 4, 13 or 21 were in Cl-1. Common aspects from seeds in Cl-1 and Cl-3 (aspects 1 and 5) seemed to have equal distribution of documents among these clusters, while two documents assigned simultaneously to aspects 1, 5, and 13 were drawn to Cl-1, whose seed was also assigned

[67] Among documents assigned to any aspects.

these aspects. Cl-2 seemed to have no discernible aspects, but actually it brought together all 5 documents discussing the trial of Paul Bedworth.

Table 12-7 : Q258 - Distribution of aspects among clusters

Aspect No	Cl-1	Cl-2	Cl-3	Cl-4
1	6	0	5	0
2	1	0	4	0
3	0	0	4	0
4	4	0	1	0
5	3	0	3	0
6	3	2	4	1
7	1	0	4	0
8	0	0	2	0
9	2	0	1	0
13	5	0	0	0
15	0	0	0	2
17	1	1	1	0
21	2	0	0	0
Cluster size	16	14	12	12

12.4.1.4 Query topic 264

Topic :	U.S. Citizens in Foreign Jails
Description:	Identify instances where U.S. citizens have been or are being held in foreign jails since the year 1900.
Recall rate:	60 % (36/60), %65(32/49) for documents assigned to aspects that are represented by more than one document.
Number of clusters:	3

For this query, 17 different aspects were identified by the TREC experts. There was some conformance between the documents' aspects and the clusters they were assigned to. Eight of the 10 documents representing Asp-1 and 3 out of 4 occurrences of Asp-8 were assigned to Cl-3. Eight out of the 13 documents representing Asp-4 were retrieved and all were assigned to Cl-2. Cl-2 also included both occurrences of Asp-6 and Asp-15. Both documents representing Asp-3, and both documents representing Asp-11 and 3 of the 4 documents representing Asp-13 were grouped together in Cl-1(Table 12-8).

Table 12-8 : Q264 - Distribution of aspects among clusters

Aspect No	Cl-1	Cl-2	Cl-3
1	2	0	8
2	0	1	0
3	2	0	0
4	0	8	0
6	0	2	0
8	1	0	3
9	0	1	0
11	2	0	0
13	3	1	0
15	0	2	0
Cluster size	13	11	12

12.4.1.5 Query topic 284

Topic : International drug enforcement co-operation

Description: Identify instances where it is shown that international co-operation is taking place in an effort to combat the worldwide drug problem.

Recall rate: 92 % (54/59), %93(39/42) for documents assigned to aspects that are represented by more than one document.

Number of clusters: 4

25 different aspects were identified by the TREC experts for this query. The distribution of documents among the clusters exhibited some conformance to these aspects. All four documents representing Asp-9 were in Cl-1, all three documents representing Asp-7 and both documents representing Asp-4 were in Cl-2, all four documents representing Asp-21 and both of the documents representing Asp-18 were in Cl-3. Documents representing Asp-1 and Asp-19 also showed some conformance, with each having their 3 out of 4 documents gathered together in Cl-3 (Table 12-9).

Table 12-9 : Q284 - Distribution of aspects among clusters

Aspect No	Cl-1	Cl-2	Cl-3	Cl-4
1	0	0	3	1
2	2	2	2	0
4	0	2	0	0
5	0	1	0	0
7	0	3	0	0
9	4	0	0	0
12	1	0	1	1
14	1	1	3	0
16	0	0	1	1
18	0	0	2	0
19	1	0	3	0
20	0	2	1	0
21	0	0	4	0
Cluster size	12	12	25	6

12.4.1.6 Query topic 286

Topic :	Paper cost
Description:	Identify the factors which have led to the cost of paper rising.
Recall rate:	81% (58/72), 88% (44/50) for documents assigned to aspects that are represented by more than one document.
Number of clusters:	3

TREC experts have identified 9 aspects for this topic. All aspects were assigned to more than one document. The “cluster-ability” problem discussed in Section 5.2.2.1 on page 86 was once again evident with this output. The high numbers of occurrences of Asp-1 and Asp-4 created a high level of similarity among the documents in the output set. Of the 44 documents assigned to any aspects, 35 were assigned to either or both of these aspects. There was conformance between the distribution of documents into clusters and their assigned aspects for aspects 2, 5, 6, 7 and 8 (Table 12-10).

Table 12-10 : Q286 - Distribution of aspects among clusters

Aspect No	Cl-1	Cl-2	Cl-3
1	4	8	6
2	1	9	0
3	1	0	1
4	7	11	8
5	0	3	0
6	0	0	2
7	2	0	0
8	0	0	3
9	0	2	1
Cluster size	16	27	15

12.4.1.7 Query topic 292

Topic :	World-wide welfare
Description:	Identify social programs for poor people other than the U.S.
Recall rate:	83% (59/71), 81% (30/37) for documents assigned to aspects that are represented by more than one document.
Number of clusters:	4

TREC experts have identified 32 aspects for this topic. Aspects were solely based on country names, and subdivided in a couple of cases where a country had two distinct programs. Therefore, the correlation between the cluster assignments and document aspects depended greatly on the geographic information that could be derived from the documents.

There were no representative terms that conveyed geographic information. Still particular problems and policies of countries reinforced some conformance between the clusters and the facet assignments. For example, 7 of the 8 documents assigned Asp-8 and 6 of the 7 documents assigned Asp-12 were clustered together in Cl-1 (Table 12-11).

Table 12-11 : Q292 - Distribution aspects among clusters

Aspect No	Cl-1	Cl-2	Cl-3	Cl-4
4	1	0	0	0
8	7	1	0	0
11	1	0	2	2
12	6	0	0	1
13	0	0	2	0
14	0	1	0	0
15	2	1	0	0
16	1	0	0	0
17	1	0	0	0
18	1	1	0	0
19	0	1	0	0
21	1	2	0	0
22	1	2	0	0
Cluster size	26	18	8	7

12.4.2 Statistical significance of the TREC-REL experiments

In order to assess the significance of the facet information/cluster relationships evidenced in these experiments, we needed to test our hypotheses with an appropriate test statistic. The null hypothesis underlying these experiments was:

> **Null hypothesis(2) :** if the documents that are relevant to a query are clustered, there will be a random distribution of facet evaluations among clusters.

and the alternative hypothesis was defined as:

> **Alternative hypothesis(2)** : if the documents that are relevant to a query are clustered, documents with similar facet evaluations will be gathered together in clusters in significant proportions.

Given these hypotheses, we needed to estimate the probability, if the null hypothesis is true, of obtaining a test value at least as extreme as the one observed. If this probability was small enough, the null hypothesis could be rejected.

To assess the associations between facet assignments and the cluster distributions, all seven query results were tabulated into aspect-by-cluster tables. These tables could be likened to contingency tables, a statistics term for data arising when a sample from some population is classified with respect to two or more qualitative variables (Everitt, 92).

Assessing the statistical significance of these tables was not easy as the data sets were rather small. The Chi-square (χ^2) test, which could show us whether there was a relationship between the cluster distributions and the facet assignments, could not be reliably used for these tables as a majority of the cell frequency expectations were less than five. Hence, the Fexact test, which was more suitable for small, sparse or skewed data sets, was employed (see Section 5.4.2.1 on page 119).

It is worth noting a few points about the Fexact test before going on with the results. This test does not give significant results only when all occurrences of an aspect are found in one cluster. If an aspect is represented by 12 documents, and there are 6 occurrences in each of two clusters and no occurrences in the other two/three clusters, this represents some significant information for the test and if combined with other significant information for other aspect occurrences the overall result may emerge as significant. On the other hand, aspects with one or two occurrences make little contribution to assessment of significance. If an aspect is only represented by two documents, having them both in the same cluster does not help much, as the probability of this occurring by chance is not low enough to be deemed significant.

Another situation is where all aspects are concentrated in one cluster. While it is quite a significant distribution when we look at the row distributions (representing distribution of each aspect), it is not significant for the Fexact test as it leaves some columns(clusters) with zero occurrences, for which it is impossible to think of more extreme distributions. This is in conformance with our purpose, as we wish to see the aspects being distributed among clusters in a meaningful way, rather than all accumulating in one big cluster.

The Fexact test gave significant results for all of the cluster solutions based on relevant documents sets (TREC-REL). The distribution of aspects among clusters was significantly different from those that would be expected under a random distribution for:

- topics 254, 264, 284, 286 and 292 at 1% level and
- topics 256 and 258 at 5% level(Table 12-12).

Table 12-12 : Fexact results (TREC-REL)

Query	Fexact value
254	0.2%
256	3.6%
258	3.4%
264	0%
284	0.2%
286	0.8%
292	1.1%

The above results were achieved using the final settings of the algorithm, i.e. using best passages, checking seed candidates' coverage by previous seeds, and choosing representative terms in a way to ensure that each document includes a minimum number of terms (see Section 11.1. on page 191). Fexact test was also used to evaluate the clustering solutions resulting from different combinations of implementation parameters. Overall, the test results were less significant with these combinations(Table 12-13).

Table 12-13 : TREC-REL : Fexact results obtained with TO (old term selection method)

Query	full-text	best passage	best passage and checking seed candidates for coverage by previous seeds
254	0.4%	0%	0%
256	22.7%	1.6%	1.9%
258	3.4%	2.1%	0.1%
264	0%	0%	0%
284	3.4%	1.4%	0.2%
286	n/a[68]	n/a	n/a
292	18.2%	19.1%	0.1%

[68] With topic 286, the algorithm gave an error message, and the suggested course of action did not resolve the problem. However, computation was possible with a reduced set, when the two sparsest rows, representing two aspects that were assigned to two documents each, were excluded from the analysis. With this data, the table was only significant at 7% level, so the null hypothesis could not be rejected. This problem was not experienced in the final run, as representative terms were chosen differently; in a way that increased the average similarity between documents and this resulted in lesser numbers of clusters which made the calculation of the Fexact value easier.

13. TREC queries - Clustering documents retrieved by the users (TREC-USER)

Tables and graphics related to this section are denoted “TREC-USER” to stand for TREC experiments based on user queries.

13.1 Research question

The TREC-REL experiments were designed to make use of the maximum amount of facet information that was available to help us evaluate the algorithm. Output sets were generated so as to include only the documents that were deemed relevant by the TREC experts. The queries used to generate these sets were quite artificial, based not only on the query descriptions provided to the users for the experiments, but also terms extracted from the facet definitions of the TREC experts and terms found by inspecting the relevant documents.

Those experiments enabled us to see the potential of the algorithm to differentiate between documents that correspond to different aspects of a query. But, with artificially generated queries and outputs with no non-relevant documents, they were far from revealing what would happen if and when the algorithm was used in real search situations.

With a typical user query, only a fraction of relevant documents would be retrieved at the top 50 documents level. In such a case, it could be desirable for a user to pick one or more promising cluster representations (instead of individual documents), expand his query based on these clusters and retrieve further documents similar to those in the chosen clusters (that are likely to be relevant given his feedback).

In order to see the performance of the algorithm in such a situation, two different experiments were designed. The first of these experiments did not utilise relevance feedback, but computed the similarities of the unretrieved relevant documents to each of the user clusters and checked whether there was a relationship between these documents’ aspects and the clusters they were most similar. It also included an analysis of the cluster/aspect distribution of the documents retrieved for the user queries. The results of this experiment are reported in this section.

The second experiment investigated whether a user could run a query, select a certain cluster out of the clusters formed from his search output as relevant and expand his query on this relevance feedback, in order to retrieve further relevant documents with aspects similar to those found in the selected cluster.

Details and results for the first experiment are given below; details and results for the query expansion experiments are reported in the next chapter.

13.2 Hypotheses

Null hypothesis(3) : if the documents that are retrieved for a user's query are clustered, and the unretrieved relevant documents are checked for their similarity to each of these clusters; there will be no relationship between the clusters to which the documents are most similar and their facet evaluations.

Alternative hypothesis(3) : if the documents that are retrieved for a user's query are clustered, and the unretrieved relevant documents are checked for their similarity to each of these clusters; a significant relationship will be found between the clusters to which the documents are most similar and their facet evaluations.

The clusters from user queries were also evaluated to find out whether there were any relationships between the retrieved documents' facet assignments and the clusters they were assigned to:

Null hypothesis(3a) : if the documents that are retrieved for a user's query are clustered, there will be a random distribution of facet evaluations among clusters.

Alternative hypothesis(3a) : if the documents that are retrieved for a user's query are clustered, documents with similar facet evaluations will be gathered together in clusters in significant proportions.

13.3 Experimental design

The seven query topics used in the TREC-REL experiments were again used in this experiment, but this time with the initial search statements of City users in 1996 TREC Interactive experiments.

The purpose was to retrieve the same set of documents as the users did, and it was envisaged that this could be achieved by simply running the queries on the Okapi search engine. However this was not quite the case. The sets of documents retrieved during the actual user search sessions were considerably different to the sets of documents retrieved for the same queries in the experiments. Although the query terms were the same, due to the modifications made to the search engine during the past two years, the documents were ranked differently. The rate of overlap for the seven queries ranged between 26 and 88% (see Appendix 5.2 for details).

The extended GSL used for indexing the database for the experiments was also a likely source of variation. The intended purpose of using an extensive GSL was to avoid non-contextual terms from becoming document representatives during the clustering stage. Describing them as stopwords during indexing enabled us to exclude them in a practical and efficient way without affecting the performance of the implementation(see Section 5.3.4.3 on page 100). But this could also affect the ranks of documents retrieved.

The variation caused by the extended GSL could simply be avoided by using the default GSL that had been used in 1996 and excluding the stopwords of the extended GSL after the retrieval stage. But the variation caused by the changes to the search engine were more difficult to tackle. Our purpose was to retrieve the same documents among the top 50 as the users did in 1996, but we did not have the possibility to run the queries with the version of the search engine used at that time.

The problem was resolved by an indirect solution. We retrieved the user documents using the interface that was designed for clustering the relevant documents from the TREC queries (TREC-REL). This interface went over the top 1000 documents retrieved for a query and picked the ones that matched the list of relevant documents for that query(see Section 12.3 on page 198). The documents retrieved by the users were listed as the set of relevant documents for hypothetical queries. And the documents matching this list were returned as search results.

Even with this method it was not possible to retrieve the full 50 documents for each of the 7 queries, on average 96% were retrieved (see Appendix 5.2 for details). But since there seemed to be no other practical way to retrieve a full list[69], the outputs retrieved by this method were used for testing hypotheses 3 and 3a (TREC-USER).

The retrieved documents were clustered and for each of the unretrieved relevant documents, the similarity between each cluster and the document was measured to find out whether there existed any relationships between the clusters the documents were most similar and the aspects they had been assigned to. The resulting distributions were checked for statistical significance using the Fexact test (see Section 12.4.2 on page 208).

[69] As mentioned previously, the student did not have skills in C programming, which was the language used for developing the Okapi search engine; thus it was not possible to generate a file of these documents with a simple C program and it had been necessary to find a workaround for this requirement.

For measuring the similarity between a cluster and an unretrieved relevant document, each representative term from the cluster was checked against the terms found in the document; and for those occurring in the document, the terms' within cluster term weights were summed to reach a cumulative measure of document-cluster similarity.

In order to test the Hypotheses(3a), the distribution of retrieved relevant documents among the clusters was also analysed to see if there was a relationship between the clusters that the documents were assigned to and their aspect assignments.

13.4 Success of user queries in retrieving relevant documents

The number of relevant documents retrieved in response to users' search statements was a crucial factor affecting the potential of the resulting clusters to represent the aspects, and to exhibit sufficient similarities to unretrieved relevant documents that have identical aspect assignments.

Most simply, if no relevant documents are retrieved, there will be no data to analyse for assessing the distribution of relevant documents among the clusters. But the essential importance of the proportion of relevant/non-relevant documents is that, the more non-relevant documents there are in an output set, the higher their contribution will be to the description of the document set. They will bring in their own aspect information, and dilute the aspect information present in the relevant documents that are retrieved. The clusters will be less representative of the relevant documents they include and hence there will be a more vague relationship between the resulting clusters and the unretrieved relevant documents with similar aspect assignments. Although topicality may remain to be strong with inclusion of non-relevant but topical documents, it is less likely to be so in comparison with a group of relevant documents all representing the same aspect.

In this experiment, the number of relevant documents retrieved was quite varied among the seven queries used (see Appendix 5.2). Queries 256, 292 and 264 had respectively, only one two and five documents with aspectual information in their output sets. The remaining queries had 9 to 18 aspectual documents, and were the only queries actually capable of displaying any relationship between the distribution of documents among the clusters and the aspect assignments (Hypotheses 3a). They would also be more capable of representing the overall set of relevant documents and would be better candidates to evaluate the main hypotheses (Hypotheses 3) of this experiment set.

13.5 Analysis of results (TREC-USER)

Results are analysed qualitatively for each query in the below sections. This is followed by quantitative analysis for statistical significance.

13.5.1 Analysis of results by query

13.5.1.1 Query topic 254

Topic :	Non-invasive procedures for persons with heart ailments
User query:	heart ailments, angioplasty, stents, arthrectomy, roto router, non-invasive surgery, heart medication
Recall rate:	19/42 (45%), 18/36 (50%) for aspectual documents
Number of clusters:	4
Number of documents retrieved:	50

A good proportion of relevant documents was retrieved by the user's query statement. However, only two of the aspects were represented by more than two documents, diminishing the chances of a significant relationship between the cluster and aspect assignments. Unsurprisingly, Fexact test result revealed no significant relationships between these documents' aspects and the clusters they have been assigned to.

When the URR(unretrieved-relevant) documents' similarity to each cluster was computed, it was found that the documents were divided between Cl-1 and Cl-4, clusters containing the highest number of relevant documents and representatives of different aspects that were found in the output set.

There was some relationship between the clusters to which the URR documents were most similar and the aspects these documents were assigned to. But the relationship was not strong (Fexact test result was 11%). See Appendix 5.1 for distribution of aspects among clusters and unretrieved relevant documents' similarity to clusters from the user query.

13.5.1.2 Query topic 256

Topic :	Negative reactions to reduced requirements for college undergraduate core studies
User query:	undergraduate, degree, reduced, requirements, standards, criticism, negative
Recall rate:	3/39 (8%), 1/14 (7 %) for aspectual documents
Number of clusters:	4
Number of documents retrieved:	50

Three relevant documents were retrieved in response to the user's initial query statement. Two of the relevant documents were in Cl-2 and one was in Cl-1; only one of these documents had an aspect assignment(Table 13-1). Therefore, there could be no analysis of the distribution of retrieved documents among the clusters.

Table 13-1 : Q256 - Distribution of aspects among clusters

	Cl1	Cl2	Cl3	Cl4
Asp 7	0	1	0	0
Cluster size	11	25	6	8
Documents with aspect assignments	0	1	0	0

Cl2 contained the documents that had the most topical similarity to the query and Cl-4 seemed to be the second best in topicality. When URR documents' similarities to the user clusters were measured, it was found that all of them were most similar to either one of these topical clusters. All documents assigned to Asp-7 were most similar to Cl-2, which included the only aspectual document which was assigned to Asp-7 (Table 13-2). The relationship between the clusters the URR documents were most similar to and their aspect assignments were significant at 3.6% level.

Table 13-2 : Q256 - Unretrieved relevant documents' similarity to clusters from the user query

	Cl1	Cl2	Cl3	Cl4
Asp 1	0	2	0	8
Asp 2	0	0	0	2
Asp 4	0	1	0	1
Asp 7	0	3	0	0
Number of documents	0	6	0	11

13.5.1.3 Query topic 258

Topic :	Computer Security
User query:	computer crime, illegal entry, hackers, spying, sensitive information, computer networks, unauthorised entry
Recall rate:	15/57 (26%), 11/28(39%) among aspectual documents
Number of clusters:	4
Number of documents retrieved:	48

Fifteen relevant documents were retrieved for the user's query. Only 11 of these documents were assigned to any aspects. There was no statistically significant relationship between the distribution of the documents among the clusters and their aspect assignments.

While there seemed to be some relationship between the clusters URR documents were most similar and their aspect assignments, it was not statistically significant. This was in part due to the sparsity of the data. Of the 11 aspects represented by the unretrieved relevant documents, only 3 was represented by more than 2 documents. See Appendix 5.1 for distribution of aspects among clusters and unretrieved relevant documents' similarity to clusters from the user query.

13.5.1.4 Query topic 264

Topic :	U.S. Citizens in Foreign Jails
User query:	U.S. citizens, Americans, foreign jails, goals, prisons, foreign, overseas, international, criminal, conviction, sentenced, broken the law, drugs, murder
Recall rate:	6/60 (10%), 5/49 (10%) among aspectual documents
Number of clusters:	4
Number of documents retrieved:	47

Six relevant documents were retrieved in response to the user's initial query statement and 5 of these documents had aspect assignments(Table 13-3). All aspectual documents were divided between Cl-1 and Cl-4. Since we had only five aspectual documents in the output set, there could be no analysis of the distribution of retrieved documents among the clusters.

Table 13-3 : Q264 - Distribution of aspects among clusters

	Cl1	Cl2	Cl3	Cl4
Asp 4	3	0	0	2
Asp 6	0	0	0	2
Asp 9	0	0	0	1
Asp 15	1	0	0	0
Cluster size	14	12	13	8
Documents with aspect assignments	3	0	0	2

When the URR documents' similarity to the clusters was inspected, it was found that a significant relationship (at 2.4% level) existed between these documents' aspect assignments and the clusters they were assigned to (Table 13-4).

Table 13-4 : Q264 - Unretrieved relevant documents' similarity to clusters from the user query

	Cl1	Cl2	Cl3	Cl4
Asp 1	0	1	2	8
Asp 2	0	0	1	1
Asp 3	0	0	0	3
Asp 4	6	0	0	3
Asp 7	0	0	0	1
Asp 8	2	0	5	4
Asp 11	2	0	0	0
Asp 13	1	0	1	3
Asp 14	0	0	1	0
Asp 15	2	0	0	2
Asp 16	0	0	0	1
Number of documents	12	1	8	23

13.5.1.5 Query topic 284

Topic :	International drug enforcement co-operation
User query:	narcotics, drugs, prosecution, international Cupertino, international borders, drug shipments, law officers, police officers
Recall rate:	14/59 (25%), 9/42 (21%) among aspectual documents
Number of clusters:	4
Number of documents retrieved:	48

Fourteen relevant documents were retrieved in response to the user's query and nine had aspect assignments (Table 13-5). There seemed to be some relationship between cluster and facet assignments, but the distribution was too sparse to reveal a significant relationship. Fexact test result was 10%.

Table 13-5 : Q284 - Distribution of aspects among clusters

	Cl1	Cl2	Cl3	Cl4
Asp 1	0	0	1	0
Asp 4	1	0	1	0
Asp 5	0	0	1	0
Asp 7	2	0	0	0
Asp 9	0	2	0	0
Asp 12	1	0	0	0
Asp 19	0	0	0	1
Asp 20	0	0	0	1
Cluster size	21	10	7	10
Documents with aspect assignments	3	2	2	2

The relationship between the clusters URR documents were most similar and their aspect assignments were significant at 0.6% level. The distribution was significant on its own, but there was not necessarily a good conformance between the aspects represented in the user clusters, and the aspects assigned to the URR documents that were most similar to that cluster(Table 13-6).

Table 13-6 : Q284 - Unretrieved relevant documents' similarity to clusters from the user query

	Cl1	Cl2	Cl3	Cl4
Asp 1	1	0	2	1
Asp 2	5	1	0	0
Asp 5	1	0	0	0
Asp 7	0	0	1	0
Asp 9	1	1	0	0
Asp 12	0	0	2	0
Asp 14	4	0	0	1
Asp 16	0	0	1	1
Asp 18	0	0	0	2
Asp 19	1	0	0	1
Asp 20	0	0	0	1
Asp 21	0	0	1	3
Number of documents	13	2	7	9

13.5.1.6 Query topic 286

Topic : Paper cost

User query: paper cost, cost of paper, paper, cost, paper pulp, cost of raw materials, inflation, labour costs, increase, expense, rising, price rise, price increase, cost of processing, increased taxes, higher taxes, factors contributing to, shortage of materials, increased wages

Recall rate: 18/72(25%), 17/49 (35%) among aspectual documents

Number of clusters: 4

Number of documents retrieved: 49

Eighteen relevant documents were retrieved for the user query, and 17 of them had aspect assignments. Asp-4 occurred in a majority (13/17) of relevant documents, and there were only three documents that were assigned to any aspects without being assigned to Asp-4. No relationship was evident between the aspects of these documents and the clusters they were assigned to. See Appendix 5.1 for distribution of aspects among clusters and unretrieved relevant documents' similarity to clusters from the user query.

13.5.1.7 Query topic 292

Topic : World-wide welfare

User query: social program, welfare program, social security, poor, needy, disabled, Europe, Africa, Asia, Far East, Middle East

Recall rate: 3/71(4%), 2/37(5%) among aspectual documents

Number of clusters: 4

Number of documents retrieved: 44

Only 44 of the 50 user documents could be retrieved for this query and only two of these documents had an aspect assignment. Hence there was no case of analysing the relationship between the aspect and cluster assignments.

There was no significant relationship between the aspect assignments of the unretrieved relevant documents and the clusters they were most similar. See Appendix 5.1 for distribution of aspects among clusters and unretrieved relevant documents' similarity to clusters from the user query.

13.5.2 Statistical significance of TREC-USER experiments

As mentioned for each query individually in the above sections, the Fexact test results for the distribution of aspects among user clusters were insignificant.

However, the relationship between URR documents' aspect assignments and the clusters they were most similar was significant for three of the seven queries: Q256 at 3.6% level, Q264 at 2.4% level, and Q284 at 0.6% level.

There were two main factors that adversely affected the results of these experiments. The first, which applied only to the cluster distribution of documents retrieved for the user query (Hypothesis 3a), was having inadequate data; inadequate even for the Fexact test which is specifically designed for small data sets. Fexact test looks for distributions that are more extreme than the distribution in the data set, but when there are only a few occurrences of an aspect, there is no case for extremity, and likelihood of any distribution is relatively high.

Hence the distribution of documents among the clusters did not represent a case for analysis for queries 256, 264 or 292 which had only one to five aspectual documents, and for query 284 which had a higher number of relevant documents but all aspects were represented by only one or two documents. See Appendix 5.2 for the proportion of aspects represented by less than three documents in the search outputs.

Another explanation for the insignificance of distributions (for both hypotheses 3 and 3a) was the effect of non-relevant documents on the clustering solutions. Non-relevant documents formed 64 to 93 percent and non-aspectual documents formed 68 to 95 percent of the search outputs (See Appendix 5.2 for details). When in small ratios, aspectual documents could not have significant influence on the representation of clusters and failing that they could not lead to significant relationships between the clusters URR documents were most similar and their aspectual assignments.

14. TREC queries - Relevance expansion on clusters created using documents retrieved by the users (TREC-EXP)

Tables and graphics related to this section are denoted “TREC-EXP” to stand for TREC-Relevance Expansion.

14.1 Research question

TREC-EXP experiments aimed to see the potential of clustering as a relevance feedback tool. What would happen if, instead of inspecting and making relevance decisions on individual documents, users made decisions on cluster representations and expanded their queries by marking a whole cluster as relevant?

Conceivably, expansion on different groups of documents would result in hitlists that have some common and also some unique documents. However, it was necessary to assess the extent of differentiation created by different clusters. But more importantly, it was necessary to find out whether such expansions would help the users to easily find more relevant documents that are similar to those found in the cluster.

Search output clustering could be offered as a viable query expansion tool, if after such an expansion new relevant documents similar to those represented in the cluster entered the search output, and the relevant documents already retrieved in the first iteration ascended in the hitlist.

14.2 Hypotheses

Null hypothesis(4) : if the documents that are retrieved for a user’s query are clustered, and users are asked to provide relevance feedback on clusters, automatic query expansion based on this feedback will not be capable of providing significant proportions of further relevant documents, that have similar facet evaluations to the documents in the selected cluster.

Alternative hypothesis(4) : if the documents that are retrieved for a user’s query are clustered, and users are asked to provide relevance feedback on clusters, further relevant documents, that have similar facet evaluations to the documents in the selected cluster, can be retrieved by automatic query expansion.

14.3 Experimental design

As with the TREC-USER experiments, TREC-EXP was based on the initial query statements of City users for the 1996 TREC Interactive track. The same seven query topics used in the previous experiments were employed to test the hypotheses of TREC-EXP.

A different version of the Enquire interface was developed for these experiments, with the added functionality of relevance feedback and query expansion. Originally, it was envisaged that user query statements would be run on this version, the top 50 documents would be retrieved, documents from a cluster (based on the clusters created for the TREC-USER) would be marked as relevant and each query would be expanded on this feedback.

But, due to the problem of being unable to retrieve the same documents as the users did, it was not possible to follow this plan. This problem had first emerged while clustering the documents retrieved by the users (TREC-USER) and was due to the modifications made to the search engine since the original TREC experiments took place (see Section 13.3 on page 212). For those experiments, the problem was tackled by defining the hitlists that were presented to the users during the original experiments, as the set of relevant documents for a hypothetical query, and using the version of Enquire that went over the top 1000 documents of the search output to pick the set of relevant documents for a query. Unfortunately, that version did not have a relevance feedback facility, and the version with the relevance feedback facility was not programmed to go over the hitlist to pick certain documents.

The only way of tackling this problem seemed to be to increase the hitlist size of the new version considerably, in order to retrieve a good proportion of documents that had been retrieved in the 1996 experiments. Since we had already retrieved and clustered the documents from 1996 hitlists for the previous experiment (TREC-USER), we no longer needed to retrieve or cluster these documents as a homogeneous set. All we needed was to find them in the hitlists and mark documents from a certain cluster as relevant. In order to achieve this the hitlist size was increased from 50 to 400. This ensured that we retrieved all or a majority of the documents that were among the top 50 documents presented for the user's viewing in the original experiments.

For each of the 7 topics, the users' initial query statements were run as many times as the number of clusters created for that query during the previous experiment(TREC-USER); in each run, documents from a different cluster were marked as relevant, and the query was expanded based on this feedback.

In these experiments, no new clustering processes were involved and the analyses were based solely on the clusters that had been created for the TREC-USER experiments and the "history" files generated during the retrieval process.

History files were one of the standard outputs of the Enquire interface; they listed the query terms, hitlists from each iteration, documents viewed by the user, and the relevance judgements. It was easy to extract from these files the identifiers of the documents from different iterations, and their ranks within the hitlist. The lists of documents were then compared with the list of relevant documents, and for each relevant document, the position within the initial hitlist and the position after the query expansion was marked.

The results of query expansion with each cluster from each query were analysed to find out the extent to which relevant documents moved up/down or in/out of the top 400 list. As we were more interested in substantial movements rather than moves of a few ranks in the hitlists, the analysis was based on moves across certain pre-defined thresholds. The thresholds were set as the top 30 documents, 31-100 documents and 101-400 documents. As long as a document remained within one of these thresholds, the difference between its ranks before and after query expansion would not be taken into account. Only the documents moving from one threshold to another, or newly entering (or dropping from) the top 400 list would be reported.

Additionally, in order to obtain an overall indication of the performance of each expansion, the number of documents ascending to a higher threshold and the number of documents newly entering the top 400 list were added and these figures were subtracted from the number of documents descending to a lower threshold and the number of documents dropping from the top 400 hitlist. The final figure was reported as "total gain/loss" value after listing all kinds of moves for a certain query expansion.

For each of the clusters, the moves were analysed both for the whole set of relevant documents and for the documents that were assigned to aspects similar to those represented in the cluster. Although the latter was more pertinent for testing the hypotheses of the experiments, the moves for the whole set of relevant documents could also provide some useful information.

Additionally, for each query, the overlap between hitlists retrieved before and after each query expansion and between hitlists from different clusters were tabulated to see the extent to which different clusters would lead the users to different results. The overlap figures at top 30 and top 400 were tabulated both for the whole set of documents and the set of relevant documents. These tables are presented in Appendix 6.4, while a consolidated summary table is presented in the following section.

14.4 Analysis of results (TREC-EXP)

The proportions of relevant documents retrieved in response to the users' query statements were highly variable among different topics. The precision values for the aspectual documents ranged between 2% and 36% (see Appendix 5.2). The initial query statements, number of documents and number of aspectual documents in each cluster, and the working queries after expansion are presented in Appendix 6.1.

There were a total of 28 clusters formed for the 7 query topics. However, it was not possible to test the hypotheses of this group of experiments on most of the clusters. Eight of the 28 clusters included no aspectual documents, and 4 included only one aspectual document. Aspectual documents had limited effect in these 4 low-precision clusters. Fourteen clusters included documents representing more than one aspect of a query. These clusters, even if they had good precision values, would not help us to test our hypothesis.

This left only 2 clusters where only one aspect was represented by more than one document; one of these clusters had two and the other had three documents representing the same aspect.

There was no significance in the moves(i.e. changes in the ranks) of aspectual documents for any of the clusters after query expansion. The actual figures are reported in Appendix 6.2 and Appendix 6.3. Appendix 6.2 includes overall results of expansion on each cluster, regardless of availability of aspectual documents. Appendix 6.3 includes results from clusters representing more than one aspect of a topic; such clusters are of limited value, as the effect of expansion can also be attributed to the fact that these clusters contained relevant documents, rather than documents representing a certain aspect.

Given these results, it is not possible to suggest that there is a relationship between the aspects represented in clusters, and the aspects of documents that enter or ascend in the hitlists after query expansion. However, the data we have is also insufficient to reject the hypotheses of this experiment. There were only two clusters where an aspect was represented in a cluster by two or three documents and in three other clusters aspects were represented by only one document, limiting the effect they have on cluster descriptions.

There was considerable variation between the overlap rates of hitlists generated for different query topics. However, the average values suggested that query expansion on clusters was capable of retrieving different sets of documents. The average overlap rate between the original hitlist and the hitlists generated after query expansion was 40%, and the average overlap rate between hitlists generated by query expansion on two different clusters was 40%

at the top 400 documents level. The same figures at the top 30 documents level were 16% and 20% respectively(Table 14-1).

Table 14-1 : Average overlap between hitlists (TREC-EXP)

Query	between original hitlists and hitlists generated by query expansion		between two hitlists generated by query expansion	
	top 400	top 30	top 400	top 30
254	23%	4%	17%	12%
256	45%	22%	55%	32%
258	57%	3%	46%	17%
264	59%	34%	51%	27%
284	31%	11%	66%	27%
286	25%	10%	6%	4%
292	42%	25%	40%	24%
average	40%	16%	40%	20%

15. Cumulative evaluation of results

In this section, the experimental results are analysed cumulatively and an attempt is made to evaluate their implications. The evaluation takes into consideration the statistical significance of the results and feedback from users, as well as the experimental conditions/limitations and their consequences for the validity of the results.

15.1 User studies

Altogether, 87 user experiments were completed. After the first 20 experiments, results were evaluated and several areas of improvement were identified. The main finding of this evaluation was the importance of cluster and document representations both as an input to the algorithm, and for the users' decision-making. Some modifications were made to the implementation(see Section 10.1 on page 164) and 10 experiments were carried out to compare two alternative cluster representation formats developed to be used in further experiments. However, results from this comparison were unconclusive, and a decision was made to use both representation formats in the following experiments.

After all envisaged improvements to the implementation and the experimental setup were realised, 57 user experiments were carried out to test the research hypotheses using two different cluster representation formats. Due to the experimental setup differences between the first 20 and last 57 experiments, the results were not combined for analysis. The first 20 experiments were statistically inconclusive, but the final 55 experiments(two of the 57 experiments were excluded from analysis) did give some significant results, although the findings did not confirm our alternative hypotheses (see Section 15.1.2 on page 230).

15.1.1 Complications and shortcomings of user studies

Attracting users to take part in the experiments was a serious bottleneck of this project. Several different methods were used: noticeboard and classroom announcements, e-mails, and word of mouth. In exchange for participation, students were offered the printed results of their searches, and additionally 5 library tokens (if they searched for their MSc project topic) or some tutoring help. Despite the fact that the experiments were based on participants' own information needs and the target population, i.e. the postgraduate students, were likely to do some online searching for their courseworks and projects anyway, the participation hardly reached the targeted number after repeated announcements.

Generally, students were not enthusiastic about participating in experiments; they probably saw it as a time-consuming and ambiguous thing to do. The majority of the participants did not have a good awareness of the Inspec database, and tended to limit their online searches to the Web. The announcements tried to stress the value of the Inspec database, but the attendance still depended mostly on the timing and the possible benefits of the experiments for the users.

The experimental task required a good level of user concentration. There was some variation between the time and effort users put into the experiments, possibly depending on the importance of the information need, relevance of the results, personal differences and also the amount of time the user could allocate to the task.

On the whole, it can be said that the experimental design did not have crucial flaws that could invalidate the results. Yet, it involved compromises/limitations some of which could be described as inherent difficulties of IR experiments:

1. The main premise of the user experiments was that the documents relevant to a user's query would be concentrated around one or two topical areas. Unfortunately, the experimental results showed this expectation to be unfounded for the user profile that took part in the experiments. In most cases, the documents found relevant by the users seemed to represent a variety of different aspects rather than being concentrated on one or two topical areas. This is a subjective observation of limited value, as it is difficult to determine the topicality of documents. However, the high precision values reinforced the idea that most users were open to discovering various different aspects related to a topic. With a high precision output, even if the clustering made a meaningful grouping of the documents, it would be of limited value to the users and the difference between the precision values of the clusters would not be very significant.
2. All participants were postgraduate student volunteers from the Department of Information Science looking for information related to their courses. The median number of query terms defined by these users was 6, above the average figure for a typical Web search(see Section 4.8.2 on page 63). As such, the experiments were not representative of the whole range of possible user types and needs and the results could not be generalised.
3. There was a difference between the modes of evaluation of clusters versus documents: participants were asked to rank the clusters and to make "yes/no" relevance decisions for the individual documents. If a user is generous about marking documents as relevant, the comparison between the two presentations will not be of much value as the relation between the cluster rankings and document judgements becomes diluted.
4. The experiment expected the users to make consistent judgements while evaluating clusters and individual documents; however, it was also recognised that users'

perceptions of their information needs could evolve during the experiment. This learning effect, an inherent limitation of IR user experiments(see Section 5.4 on page 115), could hamper the comparability of evaluations made for clusters versus those made for individual documents. The alternative would be to have different users to evaluate each output format. But as the experiments were based on users' own information needs, this was not feasible. Additionally, as we compared clustered output as a rival to relevance ranked list, it was more appropriate to have a user to evaluate a set of clusters, and then see whether the information the user gathered from the cluster representations was actually in conformance with the information he would derive from individual documents contained in the clusters.

5. Neither the clustering algorithm nor the relevance ranking method included any factors other than the term occurrences within the documents. However, users responses to the post-questionnaire(Table 10-8 on page 175) revealed that more than half of them have considered (one or more of the) factors like documents' authors, journal and publishing year in making relevance judgments(see Graph 10-1 on page 176 for the variation between the relevance judgements with respect to the publishing year). The clustering representation displayed no information on these details, whereas the document lists provided to the users for relevance judgments included all three types of information. This situation might have affected the comparability of cluster rankings and relevance judgements; however, it was beyond the scope of this project to devise methods to incorporate these factors in cluster formation and representation and additionally in the relevance ranking method for comparability reasons.
6. A weakness related to the cluster representations was that the same or very similar titles could be chosen as representative titles twice, due to existence of documents that are almost identical except for small details, such as the source. However this did not occur too often: in only 9 of the 55 experiments one cluster contained the same title twice. In two of these cases, both Rep-D and Rep-E representations included such a cluster, in 5 cases only Rep-D included such a cluster and in 2 cases only Rep-E contained one. This situation was not foreseen, if it had been anticipated beforehand, it could have been avoided by side-stepping the titles that are almost or completely identical to a previously chosen representative title.

15.1.2 User study results and their significance

User experiments revealed no significant precision difference between clustered outputs and ranked lists. In the first group of user experiments(n=20), clusters performed slightly better than ranked lists at the top cluster level; at the top 2 clusters level, the performances were equal (see Table 9-4, Table 9-5 on page 148).

In the final group of user experiments(n=55), the number of cases where one method performed better than the other was slightly higher for the ranked lists at the top cluster level and slightly higher for the clustered representation at the top two clusters level. However, the overall average precision values were higher for the ranked list at both levels. As such, clustering did not appear to be preferable to ranked lists especially as it also represented overheads in both computing time and resources involved in creation of the clusters, and the time and effort taken by the users to inspect them.

An interesting outcome of the user experiments was the ability of the users to identify clusters that did not include relevant information. There were less relevant documents within the clusters marked last by the users as compared to the documents ranked last at similar threshold levels. This brought out the possibility of using clusters as an exclusion tool to improve the precision of ranked lists. After exclusion of documents from the last cluster, ranked lists performed significantly better than the clusters at the top cluster level(Table 10-18 on page 182).

There was also some evidence (consisting of observation of users during the experiments and a few user comments) that clusters could be used to provide the users with a glimpse of the search results, in order to decide whether to inspect the search results or initiate a new query straight ahead.

15.1.3 User's attitudes towards clustering search outputs

In ¾ of the cases, users were positive about the usefulness of the cluster representations (see Table 10-4 on page 169). However, their responses also pointed towards some serious weaknesses of the clustered representation:

1. Cluster representations are not always adequately informative:

   "...possibly adding one or two more names of representative documents would make it clearer."(1st group, Q2)
   "Many of the titles were not self explanatory..."(1st group, Q5)
   "The titles of the documents were not exemplary of the article. I needed to review the abstract to make an evaluation of the relevance." (1st group, Q11)
   "Making relevance judgements without abstracts is hard. Titles are very general or I don't understand them." (1st group, Q19)
   "Some titles didn't give enough information about the subject area." (2nd group, Q22)
   "... [I] prefer more information and the chance to browse" (2nd group, Q30)

"When having to rank clusters, I think that just having titles wasn't useful enough, especially because I don't know much about the subject..." (2nd group, Q36)
""...it was sometimes hard to tell by looking at the titles whether the cluster was relevant or not..." (2nd group, Q45)
"...some of the titles are vague, and there is a necessity to read the abstract" (2nd group, Q48)
"It is uncertain to me whether the available information is not enough or the experiment design should be improved" (2nd group, Q49)

2. Cluster representations may be misleading:

"... the titles are of course misleading and within one cluster I might find one title that might be relevant, when I consulted the abstract I found that it was irrelevant." (1st group, Q7)
"...there were many outputs that I've found useful but was not represented in the clusters" (2nd group, Q40)
"...[clusters] may hide something important, and highlight something else, which can mislead the user; especially for a first search; where you know nothing, you learn by searching" (2nd group, Q57)

Before the final group of experiments an attempt was made to improve the cluster representations, but as the comments from the users reveal, the representation was still far from being reliably informative.

15.2 TREC - facet evaluation experiments

Three different sets of experiments were carried out based on 7 queries for the 1996 interactive track. The first set, TREC-REL, was based on relevant document sets for each query and assessed the ability of the algorithm to group together documents with the same aspect assignments.

The second and third sets were based on the top 50 documents retrieved as a response to the City participants' initial query statements for the queries. The TREC-USER experiments aimed to assess whether there was a correlation between the aspects of the relevant documents that were not retrieved among the top 50 documents and the aspects found in the clusters (formed out of the top 50 documents) to which these documents showed the highest similarity. This was expected to give us an indication of whether we could expect to retrieve these unretrieved relevant documents by expanding the query based on the cluster they were most similar.

In the TREC-EXP experiments, we went one step further and actually expanded the queries based on each of the clusters formed out of the initial search outputs. The search results were then analysed to find out whether previously unretrieved relevant documents that had similar aspect assignments to a cluster would rise to higher ranks when queries were expanded based on that cluster.

15.2.1 Complications and shortcomings of facet evaluation experiments

TREC experiments aimed to assess the ability of the implementation to bring together topically related documents. It was quite difficult to gather data for such an assessment, but the existence of a set of data generated for TREC Interactive track(1996) enabled us to design experiments that at least approximately satisfied our requirements.

TREC documents were less structured and more variable in size as compared to Inspec documents. Hence, it had been necessary to tune the algorithm considerably to achieve reasonably balanced clusters with the TREC document sets. But more importantly, the utility of the available aspect information was limited/inadequate for some of our objectives.

Firstly, the quantity of documents assigned to a certain aspect was not always adequate to come to a conclusion about the significance of their distribution. This was especially the case with the TREC-USER and TREC-EXP experiments where the proportions of relevant and aspectual documents in the output sets were quite low (varying between 2% and 36%, see Appendix 5.2). The aspects of the retrieved but non-relevant documents were unknown. This meant that we could not know whether most documents from a cluster covered the same aspect although only a few were relevant, or whether there were a variety of different aspects in each cluster which reduced the potential effect of the aspectual documents as descriptors of the clusters. Similarly, we could not know if the non-relevant documents retrieved after query expansion actually represented the same aspects as the aspectual documents found in the cluster.

This situation seriously affected our ability to interpret the experiment results. If a cluster included a high proportion of documents representing a certain aspect, although some of these documents were not relevant, we could expect to retrieve further documents representing this aspect when the query is expanded based on this cluster. But if all other documents for which we had no aspectual knowledge represented different aspects, we could not have high expectations, as the aspectual information would be diluted.

The available aspect information was further reduced by the multiple aspect assignments that dominated some of the queries. Clusters including documents with multiple assignments were not useful for assessing cluster-document similarities or query expansion results. In such cases, we did not know whether a document was retrieved just because there were some other relevant documents in the cluster, or it was actually due to representation of a certain aspect in the cluster.

Another point of concern was whether the facets defined by the TREC experts represented the only or most informative way of labeling the documents. There are many different ways of categorising a group of documents, and the aspects defined by TREC experts were not necessarily the only valid ones; a good example is query 292 where the aspects were based solely on country names. It is not difficult to think of other ways of identifying aspects for this query which has been defined as "*identify social programs for poor people other than in the U.S*".

Finally, TREC experiment results could be contested on the grounds that they did not represent judgements of real users with real information needs. However, it would be very difficult to gather information in such detail (although it was still less than sufficient for our purposes) within more naturalistic settings.

15.2.2 Facet evaluation results and their significance

In the TREC-REL experiments, the clusters were based on relevant document sets, with aspect assignments for almost all of the documents. With all 7 queries, there was significant correlation between the distribution of the documents among the clusters, and the aspects represented by these documents (see Section 12.4.2 on page 208).

In the TREC-USER and TREC-EXP experiments, the proportion of documents with aspect assignments were much smaller, and this made it difficult to draw conclusions from the results(see above section).

The only interesting result derived from these experiments was the correlation between the unretrieved relevant documents' aspect assignments and the clusters they were most similar (TREC-USER). This relationship was statistically significant for three of the seven queries: Q256 at 3.6% level, Q264 at 2.4% level, and Q284 at 0.6% level.

The inadequacy of TREC facet information was most acutely felt with the TREC-EXP experiments, where the experiment hypotheses could be tested with only 6 of the 28 clusters created for the 7 queries (see Section 14.4 on page 226). The number of unretrieved relevant documents that were assigned to the aspects represented in these clusters were generally quite small. Query expansion did not appear to lead to retrieval of further relevant documents (with the same aspect) for any of the 6 clusters. Even when a few other relevant documents were retrieved after query expansion, similar results were achieved also after expansion based on other clusters that did not represent the aspect in any of their documents.

16. Conclusions and recommendations

This section seeks to put the findings of this project into some perspective, and attempts to identify areas of future research.

The conclusions we can derive from the results of this project will be limited as the experiments conducted are not representative of the whole set of possible user types and search situations. It is also possible that there may be other combinations of clustering, term selection and cluster representation methods that are more effective than those used in this project. Hence, it should be kept in mind that some of the limitations or weaknesses of clustering discussed in this chapter may at least partly be a consequence of the particular methods adopted in this project.

16.1 Factors affecting clustering performance

There are two main requirements for achieving useful cluster solutions for search outputs. First, it is necessary for the algorithm to bring together topically related documents. This depends on:

1. the way the algorithm works,
2. the quality of the terms/variables chosen for representing the documents, and
3. the number and distribution of facets represented within the document set.

It can be expected that given a certain set of representative terms, most clustering algorithms will produce similar solutions unless they have particular features or limitations. On the other hand, Everitt(93) suggests that no clustering method can be best in all circumstances and in many applications, it might be reasonable to try a number of clustering methods. CBR research in the past has also shown that performance of clustering methods vary between databases/applications, so it may be worth trying a few different methods with a given database.

The quality of the terms/variables chosen is very important as clustering needs to be based on a limited number of features to avoid obscuring the clustering structure. Ideally, it would be desirable that each term/variable correlates with existence/non-existence of a certain facet represented in the output set. However, this would require a prior knowledge of what the facets are and there would be little need to perform clustering if we had this information.

What we might aim instead could be to come up with a term set that has a high ability to differentiate between the documents in the output set, with the hope that they will lead to meaningful, coherent document groupings. The TSV, used in this project, is one such measure.

The effectiveness of a measure like TSV can be maximised by ensuring that:

- all synonyms of a term/phrase are treated as one variable in the term selection process,
- all meaningful text strings are considered in the term selection process; e.g. phrases, terms occurring in close proximity[70].

In addition to using such a measure, using a large number of stopwords can help to produce more satisfactory clustering solutions by excluding terms that would not have much contextual value with respect to most queries. While the contextual terms are likely to be different for different databases, the stopwords (the generic terms) may be similar for many databases; therefore building lists of stopwords may not require customised efforts for each database. This argument is supported by the high rate of overlap between the list of stopwords for the two databases used in this project.

Once the document representatives are chosen, the potential usefulness of clustering will depend on particular characteristics of the document set at hand. There is no guarantee that a set of documents can always be divided into a small number of neat groups. The distribution of facets among the documents can be highly overlapping, resulting in many documents with multiple facets; in such a case, it is less likely that the clustering solution will provide the users with clusters that are easily recognisable.

Alternatively, similarities between the documents may be weak, or the similarities may be high, but distinctions negligible. If the average level of distinction between the documents is low, we may end up with an unbalanced clustering solution, e.g. with one big and plenty of small clusters. This is not as useful as a clustering solution where the clusters are of approximately equal size. With a well-balanced clustering solution, rejecting or selecting each cluster has a certain utility for the user whereas with an unbalanced clustering solution, the utility of selecting a large cluster, or the utility of rejecting a small cluster is low.

In this project, it was evidenced that unbalanced clusters are caused by a skewed distribution of the number of representative terms found in each document (see Section 5.3.4.8 on page 106); it is likely that this distribution also affects other clustering algorithms in similar ways. The likelihood of having a skewed distribution can be reduced by manipulating the term selection process; in this project, term selection process was modified for the TREC experiments to ensure that all documents contain a minimum number of the representative terms, it could also be possible to take additional measures to avoid the situation where some documents contain almost all of the representative terms and lead to unbalanced clusters. However, such measures imply further computational complexity and stipulate the unrealistic assumption that any set of documents can neatly be divided into a limited number of equal sized groups.

The second requirement for achieving a useful clustering solution is conveying concise but sufficient information to the users about the contents of the clusters. As with the first requirement outlined above, ensuring a good choice of representatives is again crucial(for a detailed discussion, see Section 5.3.5 on page 111). The size of clusters is also important, as the information reduction can be expected to increase proportionately with the number of documents included in a cluster.

In addition to these, users' previous knowledge/expectations/openness to information play an important role in utilising the provided information. It is likely that as users get more experienced in making decisions based on cluster representations, they will get better in predicting the contents of the clusters.

Unfortunately, there is not enough research on the problem of cluster representation (for users' viewing) and cluster representations remain primarily based on title information, which is notorious for its unreliability.

16.2 Feasibility of search output clustering as an interactive IR tool

In this project, output clustering did not prove to be a strong rival to the ranked lists as an output representation method. However, it is conceivable that clustering can be preferable to relevance ranking in certain situations, e.g. when users have a more exploratory approach, or find it difficult to narrow down their queries. It may also be a fruitful approach to make the use of clustered output dependent on some indication of the quality/expected utility of the clusters.

Selective use of clustering is actually more in line with the general methodology of clustering and applications of clustering in other fields. In fields like biology and zoology, clustering involves a more sophisticated procedure than in the applications in IR. A researcher in those fields is likely to apply a variety of different methods to the same data, employing different similarity measures and variable sets; he might wish to visualise the data before deciding whether to attempt clustering at all, and choose to focus on a subset of the individuals instead of the whole group. After the clustering process he tests the validity of clusters using methods

[70] It is also desirable to include document attributes like recency, source, author, document length as

developed for this purpose, and decides whether the results provide a useful/informative grouping of the data(Everitt,93).

The researcher's primary concern in these decisions is to find out whether the data is suitable for classification, and if it is, which combination of methods will yield the most useful results.

In IR applications, clustering decisions are made a priori, based on more mechanical aspects like the speed of producing results, or the ability of the method to produce balanced clusters. It is not customary to consider the contextual characteristics of a particular data set, or employ any means of validating the cluster solutions before presenting them to the users. There is an unpronounced assumption that clustering will be equally applicable to any group of documents, whether it be a whole database or search results from a query.

However, it appears that the potential usefulness of clustering is not uniform against queries. If measures that are capable of indicating quality of clustering can be devised, clustering can be used selectively.

For the C3M algorithm, with the first set of user experiments, it appeared that, the magnitude of seed powers correlated with the probability of the clustered representation to perform better than the ranked lists (see Table 9-9 on page 156). However, with the second group of user experiments, where we had a higher number of users, the differences between average seed powers between queries where clustering performed better versus those where ranked lists performed better were only marginal (2.206 for queries where relevance ranking performed better, and 2.219 for queries where clustered representation performed better). Hence the possibility of using seed powers as an indicator of cluster quality had to be ruled out.

Several measures have been described by IR researchers to assess the quality of a clustering solution. A recent contribution is *Global Quality Function* (GQF) of Zamir&Etzioni (97) (see Section 4.8.2.4 on page 71), which is actually envisaged as a method of forming clusters rather than evaluating them.

Another related measure is Korpimies and Ukkinen's (98) *clpot* which is meant to measure the potential of a term weighting scheme for output clustering:

$$clpot(tw) = \frac{csim(tw)}{csim'(tw)}$$

additional variables into the clustering process. This is discussed in more detail in Section 16.4 below.

where *csim(tw)* represents the average similarity between every possible pair of documents that are in the same cluster, and *csim'(tw)* represents the average similarity between every possible pair of documents in the output.

However, there seems to be no measures specifically designed to measure the quality of a clustering solution during a search session or experiments which use a quality measure to decide whether it is worth presenting the user with a clustering solution instead of a ranked list.

There is also scope for improving the efficiency of clusters by being selective in cluster assignments. During the user experiments, for the sake of comparability, all the documents from an output set were clustered in the user experiments. But, it was evidenced that cluster representations do not provide a consistent level of informativeness even among the clusters from the same query. Several users pointed out this problem when asked about whether the clusters provided a useful representation:

> *"On some occasions yes, but mostly no.. "* (Q 8)
> *"In some cases"* (Q 25)
> *"Yes - but not in all cases"* (Q 33)
> *"Sometimes.."* (Q 54)
> *"Some of them"* (Q 56)

According to another user (2^{nd} group – Q29)

> *"Some clusters contained highly similar documents, most of them relevant or irrelevant. Some clusters seemed to have leftovers - "miscellaneous documents.."*

In operational situations, the clustering can be limited to documents that show a reasonable amount of similarity to a seed document. This is actually a common practice in the case of output groups created for Web search engine queries. The outliers can be listed between the clusters, with respect to their relevance weights.

While it is not clear to what extent we can differentiate good clusters from bad ones, it is conceivable that numeric indicators of cluster coherence and distinctiveness may help us to weed out 'outliers'.

Finally, it is possible to make use of clusters as a way of improving the performance of ranked lists by using it as an exclusion tool. The most significant finding from the user experiments was that users were good at spotting the worst clusters. There were significantly less relevant documents in the worst marked clusters than the documents ranked last at the same threshold level (see Table 10-15). When the worst clusters were excluded from the ranked lists, it was found that the performance of ranked lists was significantly better than the clusters (see Table 10-18). This situation confirmed the observation made by User 21:

> *"The clusters did group together the documents that didn't look interesting.."*

Worst clusters can also be used as a source of terms for negative relevance feedback. Belkin et al.(98) suggest that terms which appear only in negatively judged documents may be indicators that the context in which the good terms appear in these documents is inappropriate to the searcher's problem, or that the topic which they represent is treated only peripherally, or from an inappropriate point of view. Their model of relevance feedback expands queries by both the important terms in the positively judged documents and by the important terms in the negatively judged documents(with negative weights) which do not appear in the query or the positively judged documents.

Belkin et al. point out an advantage of negative feedback over positive feedback as not increasing the RSV of other documents which treat the same aspect of the topic; this attribute of positive relevance is undesirable because it results in demoting documents treating different aspects of the topic in the output ranking, and promotes documents that cover the same aspect, giving the user 'more of the same'.

Their user experiments using negative and positive relevance feedback as a term-suggestion device for relevance feedback have found that it gave better aspectual recall compared to use of only positive feedback; however the difference was not significant, in part due to the small number of cases. However, subjects did not perceive negative feedback as useful as positive feedback and were less likely to use it.

It is necessary to conduct further experiments with higher numbers of users, and probably experiment with using negative-weights in an implicit way. Within that context, using worst clusters (as identified by the users) may serve as a quick way of finding negative weight terms.

While considering all these possible utilisations of clustering, its overhead costs - especially in terms of time and effort required from the users to assess the clusters - and possible level of acceptance by the users also need to be put into the equation. Most users seem to avoid complexity as much as possible and keep away from using sophisticated tools offered to them in many search systems. As pointed by Anick&Vaithyanathan(see Section 4.8.1.5 on page 60) introducing a new level of abstraction may also lead to some user misconceptions that can affect the usefulness of a search session, e.g. users may expect the cluster representations to be definitive descriptions of clear cut document groupings.

Clustering may be more appropriate for people who use search systems on a regular basis, building an understanding of the way clustering works and how its results can be interpreted.

16.3 Comparison of findings with results from other recent studies and general assumptions about document clustering

Since the clustering hypothesis was put forward and supported by experimental evidence by Jardine&van Rijsbergen(71), various researchers conducted studies to test the hypothesis with various different methods and databases(see Section 4.1 on page 37). These studies produced mixed results and the cluster hypothesis has lost its credibility due to these results and some strong theoretical criticisms (Voorhees,85; Gordon,91; Shaw,97)

Recently, wider usage of text retrieval systems (especially the use of the Web search engines and related precision problems), and diminishing computing time and space constraints have caused a revival of interest in clustering documents. However, the focus is no longer on clustering collections for retrieval purposes, but on using clustering in a variety of different ways to enhance users' interaction with search systems.

Although the cluster hypothesis continues to be used as an assumption in some recent studies, it is generally not pursued as the main purpose of a research project. An interesting revision to the hypothesis has been put forward by the Xerox researchers Hearst&Pedersen (96) based on the findings from their clustering experiments:

> *"We too assume that documents that cluster together are similar in some ways, and thus relevant documents will tend to cluster near other relevant documents and further away from non relevant ones. However, in contrast with the assumption underlying the strategy of earlier work, we do not assume that if two documents D1 and D2 are both relevant or nonrelevant for query QA, they must also both be relevant or nonrelevant for query QB."*

Almost all new work employs clustering in an interactive way, instead of adhering to a static cluster solution. In a considerable number of studies, it is utilised as a search output presentation method; leading examples are:

- Allen, Obry & Litzman's (93) interface for navigating clustered search outputs (see Section 4.8.1.7 on page 62),

- Xerox' Scatter-Gather algorithm(see Section 4.8.1.1 on page 53), which has been envisaged for both collection and output clustering, with the focus shifting to output clustering in recent years,
- Zamir&Etzioni's (97,98) Suffix Tree Clustering which applies clustering to search results and creates overlapping clusters based on phrases shared between documents (see Section 4.8.2.4 on page 71),
- Eguchi et al.'s (99) experiments on employing users' relevance judgements on the clustered search results for incremental query expansion,
- Shimamura et al.'s (98) clustering method which groups Web search results by organisation name (Shimamura does not actually use any clustering algorithms, although he describes his method as a clustering method; in similar fashion NorthernLinght and InferenceFind describe their output grouping methods as clustering methods even though their groupings appear to be greatly based on site information, for details see Section 4.8.2 on page 63).

Clustering is also being put to use in query formulation, refinement, expansion and routing, some examples are:

- Mechkour et al.'s(98) WebCluster project which uses document clustering for routing, 'to mediate access to the Web documents': users are presented with cluster representations formed from a well structured collection covering a particular subject domain, and their feedback is used to formulate a query to be forwarded to the Web search engine (see Section 4.8.2.7 on page 75),
- Anick&Vaithyanathan's (97) Paraphrase interface for context-based information retrieval for query refinement and expansion, where clustering is employed as a static retrieval tool as in many older studies, but the user is allowed to manipulate key terms of a chosen cluster to launch a new query(see Section 4.8.1.5 on page 60),
- Yiming-Yang et al.'s (98) study on use of text retrieval and clustering techniques for event detection from a temporally ordered stream of news stories,
- Bhatia&Doegun's (98) techniques of developing cluster characterizations and clusters by employing user viewpoint, where this viewpoint is elicited through a structured interview,
- Ebert et al.'s (97) document corpus management system which seeks to identify trends, discover common links and find clusters of similar documents within a collection, by combining the results of many single queries in various ways.

Unfortunately, most recent projects focus more attention on development than evaluation. Even if some kind of evaluation is made, this usually involves relevance measures as in the user experiments of this project(e.g. Zamir&Etzioni(98)). Furthermore, it may also involve the assumption that users are able to select best clusters(e.g. Hearst&Pedersen, 96;

Zamir&Etzioni, 98), which, according to empirical results of this project, is an unrealistic assumption. It is meaningful to employ precision measures when comparing clustered output against relevance ranked lists; however, a more realistic evaluation of clustering would require assessment of topicality rather than precision. Actually, even topicality is not an ideal measure as we need the information that differentiates between documents rather than information that may be common to all. This type of information is more difficult to collect and analyse compared to yes/no type relevance judgements.

It is surprising that although search output clustering is generally envisaged as a tool to improve users' interaction with search systems, there is little work that focuses on user needs or behaviour. Pirolli's (97) study on predicting users' cluster selection decisions(see Section 4.8.1.1 53) and Macskassy et al.'s (98) investigation on whether different subjects create similar groupings of a set of Web documents are exceptional studies in this sense. In their study, Macskassy et al. have found that each subject had different behaviour in terms of overlap, size and number of clusters, and there was little similarity between the clusters created by any two subjects. These results can be valuable input to future studies as they give us an indication of the difficulties involved in evaluating clustering methods.

16.4 Proposals for future research

As mentioned in Section 16.1, in order to produce a successful clustering solution, it is necessary to:

1. Achieve a meaningful grouping of documents, and
2. Ensure that the cluster representations convey sufficient information about cluster contents to the user.

Although there are numerous recent studies that involve clustering and use cluster representations, there is less effort specifically aimed at improving meaningfulness of clusters or informativeness of cluster representations. A number of threads can be followed to improve our understanding on each of these issues.

In grouping the documents from an output set, it is important to be able to represent the aspects contained within a document set in a concise and balanced way where each term has a unique meaning and no document attributes are over-represented. During this project, a considerable level of effort went into defining a large group of stopwords and synonyms, in order to minimise the selection of non-contextual terms as document representatives, and to avoid under/over representation of aspects that may be caused by treating synonyms separately. The significance of the results of the TREC-REL experiments revealed that a representation capable of capturing the aspectual information contained within a document set was attainable. It is also conceivable that the selection of representative terms can be improved further with a more systematic approach, for example by using distributions of terms over the collection as a factor in selection. Terms with non-random distribution are supposed to have high discriminatory power, and this can be measured by using the skewness and elongation diagnostics(see Dubin's work on Section 4.8.1.2 on page 58). According to Dubin the terms that are strong discriminators with few high correlations among variable pairs have the best discriminatory power.

The value of using phrases for document representations has already been shown in University of Washington's Husky Search project (see Section 4.8.2.4 on page 71), Anick&Vaithyanathan's(97) Paraphrase Interface (see Section 4.8.1.5 on page 60) also uses phrases for cluster descriptions. As phrases have higher predictive and representation value than single terms, capturing and considering all phrases (in addition to the indexed phrases) in the term selection and cluster representation processes can improve performance and usability of clusters. However, the probability of a phrase reoccurring literally is low, and this diminishes their utility. Koster et al. (99) stress the value of utilising both noun and verb phrases in text categorisation tasks and suggest applying a number of normalizations and matching techniques to offset the low reoccurrence problem.

Using other document attributes like document length, type, source, availability, recency, author information, number of citations, level of treatment may also help us to achieve more meaningful groupings. Users' responses to the post-questionnaire (Table 10-8 on page 175) revealed that more than half of them have considered (one or more of the) factors like documents' authors, journal and publishing year in making relevance judgments. It could be beneficial to investigate these factors further to determine the extent of their importance in user decisions, and whether this importance varies between different user types and search situations. However, these factors represent distinct data types and methods to incorporate them into clustering algorithms may not be easy to devise (e.g. for date or length it would be necessary to establish ranges, and their effect would not necessarily be linear - see Graph 10-1 on page 176 for the relationship between publishing year and relevance judgments). Furthermore, some factors would affect only some of the documents (e.g. generally there would be a limited number of known authors, and limited knowledge of availability information for journals) and some factors would require eliciting information from the users other than the query terms, and this would complicate the system-user interaction.

Despite their limitations and the difficulties involved in incorporating these additional factors, their importance in users' decision-making process has been confirmed in various studies (Schamber, 91; Park, 92; Barry,93) and investigation into how these factors can be utilised remains an untapped area of research. Users' information needs can be met more effectively if methods can be devised to incorporate some of these factors in cluster formation and representation and also in relevance ranking algorithms.

Ensuring that the cluster representations convey sufficient information about the contents of the clusters is the second big challenge. Clustering needs to overlook a considerable amount of information in order to be able to make categorisations. The amount of reduction is proportionate to the number of documents being clustered, given a fixed number of clusters. Although we attempted to keep the information loss to a minimum in this project by limiting the number of documents to be clustered to 50, the effects of the information loss was felt by at least 20% of the users, who thought that the clusters did not provide a useful representation, and possibly a further 22% who were only mildly positive about the usefulness of the cluster representation(see Table 10-4 and Table 10-5 on page 169). Several users have expressed concern about this problem (see Section 15.1.3 on page 231).

This is understandable, as there is a huge amount of information reduction from 10-15 abstracts contained in a cluster to cluster representations containing three titles and up to ten representative terms. A second layer of abstraction is being placed between the users and the actual documents, if we consider that the titles/abstracts themselves represent an initial phase of information reduction. Cluster representation is a representation of representations.

It may be fair to say that the representation used in this project was moderately successful in representing the clusters, as it achieved results that were comparable to the ranked lists. However, users' comments also made it clear that the titles are not reliable as indicators of topicality.

It may help to choose the representative titles from among the documents that have the highest relevance weights, or to avoid short titles on the grounds that they would be less informative than longer titles, or to minimise the overlap/similarity between the representative titles to avoid duplication of information. On the other hand, it may be possible to come up with more radical solutions by investigating new/innovative ways of representing clusters. For example, titles may be replaced by more meaningful sentences from the document abstracts. Alternatively, cluster representations may consist of distinct concordance lists for the query terms. A comparative study of representation methods could help to determine strengths and weaknesses of different approaches.

As discussed in detail in Section 16.2 on page 238, it is also worth investigating selective use of clustering, both in terms of deciding whether to present the user with clusters or not, and in terms of assigning to clusters only documents that are above a similarity threshold. Using clusters to support ranked retrieval, for example by using them as a rejection tool or as a source of negative weight terms are also interesting areas for further research.

Finally, it is worth mentioning that output clustering research could benefit from new evaluation measures designed specifically for measuring cluster quality (using factors like topicality, coherence, distinctiveness). It is also necessary to have data sets that are generated specifically for clustering research with topicality/aspect information both for relevant and non-relevant documents. While TREC interactive topics and aspects have been used in this project to assess the ability of the algorithm to group together documents representing similar aspects, this data set only approximately fulfilled the requirements for testing our hypotheses. The aspect definitions and assignments were based on relevance and did not seek to define the documents; additionally, for most queries the number of aspectual documents were less than sufficient for statistical analysis.

Data sets generated specifically for clustering research are needed to help researchers to compare different clustering and term selection methods before proceeding with user studies, even if these cannot be as realistic as results from user studies.

REFERENCES

ALLEN RB; OBRY P; LITZMAN, M (1993) An interface for navigating clustered document sets returned by queries, Proceedings of the ACM Conference on Organizational Computing Systems, p. 166-71.

ANICK PG; VAITHYANATHAN S (1997) Exploiting clustering and phrases for context-based information retrieval, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelpia, PA, USA, 27-31 July 1997, p. 314-21.

ASHFORD J H (1995) Full text retrieval in document management, Information Management &Technology, V.28 N.1, 28-32.

BARRY C (1993) The identification of user criteria and document characteristics: beyond the topical approach to information retrieval. Ph.D. dissertation. Syracuse, NY: Syracuse University, School of Information Studies, 1993.

BATES MJ (1989) The design of browsing and berrypicking techniques for the online search interface, Online-Review (UK), vol.13, no.5, p.407-24.

BEALE EML (1969) Cluster Analysis, London: Scientific Control Systems Ltd.

BELKIN NJ; CARBALLO JP; COOL C; LIN S; PARK SY; RIEH SY; SAVAGE P; SIKORA C; XIE H; ALLAN J (1998) Rutgers' TREC-6 Interactive track experience, The Sixth Text Retrieval Conference (TREC-4), NIST, Gaithersburg.

BHATIA SK; DOEGUN JS (1998) Conceptual clustering in information retrieval, IEEE-Transactions-on-Systems-Man-and-Cybernetics-Part-B (Cybernetics) (USA), Vol.28, No.3, p.427-36, published IEEE.

BLAIR DC (1980) Searching biases in large interactive document retrieval systems, JASIS, V.31, 271.

BONNER RE (1964) On some clustering techniques. IBM J. Res. Dev., 8, 22-32.

BOOKSTEIN A (1978) Fuzzy requests: An approach to weighted Boolean searches, JASIS, Vol.29, No.3, 156-8.

BOOKSTEIN A; KLEIN ST; RAITA T (1995) Detecting content bearing words by serial clustering - extended abstract, Proceedings of the 18th International ACM-SIGIR Conference on Research and Development in Information Retrieval, 319-327.

BROFFIT JD; MORGAN HL; SODEN JV (1966) On some clustering techniques for information retrieval, Report ISR 11 to the National Science Foundation, Sect. IX, 1966, Cornell University, Department of Computer Science.

BURGIN, R (1995) The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity, JASIS, Vol. 46, No. 8, p. 562-72.

BURNETT M; FISHER C; JONES R (1996) In TEXT Precision Indexing in TREC-4, The Fourth Text Retrieval Conference (TREC-4), NIST, Gaithersburg, 287-294.

CALINSKI T & HARABASZ J (1974) A dendrite method for cluster analysis, Communications in Statistics, 3, 1-27.

CAN F & OZKARAHAN E (1990) Concepts and effectiveness of the cover-coefficient based clustering method for text databases, ACM Transactions on Database Systems, Vol.15 N.4, 483-517.

CAN F & OZKARAHAN E (1984) Two partitioning type clustering algorithm, JASIS, 268-276.

CHANG C; HSU C (1996) Customisable multi-engine search tool with clustering, 6th International World Wide Web Conference, Hyper Proceedings, htpp://www6.nttlabs.com/HyperNews/get/PAPER53.html

CUNNINGHAM KM & OGILVIE JC (1972) Evaluation of hierarchical grouping techniques. A preliminary study, Comp. J., 209-213.

COOPER W S (1994) The formalism of probability theory in IR: A foundation or an encumbrance? Proceedings of the 17th International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, 242-247.

COOPER W S (1971) A definition of relevance for information retrieval, Information Storage And Retrieval, Vol.7, 19-37.

CORMACK RM (1971) A review of classification. J. R. Statist. Soc., Series A, 134, No.3, 321-367.

CROFT B (1977) Clustering large files of documents using single-link method, JASIS, V.28 , 341-344.

CROFT W B (1980) A model of cluster searching based on classification, Information Systems, Vol.5, 189-195.

CUADRA CA & KATTER RV (1967) Experimental studies of relevance judgements final report. Volume 1: Project summary. Santa Monica, CA: System Development Corp.

CUNNINGHAM KM & OGILVIE JC (1972) Evaluation of hierarchical grouping techniques. A preliminary study. Comp. J., 15, 209-213.

CUTTING D R; KARGER D R; PEDERSEN J O (1993) Constant interaction time scatter-gather browsing of very large document collections, Proceedings of the 16th International ACM-SIGIR Conference on Research and Development in Information Retrieval, 126-134.

CUTTING D R; KARGER D R; PEDERSEN J O; TUKEY J W (1992) Scatter-gather: a cluster based approach to browsing large document collections, Proceedings of the 15th International ACM-SIGIR Conference on Research and Development in Information Retrieval, 318-329.

DAVIDSON D (1977) The effect of individual differences of cognitive style on judgements of document relevance, JASIS, Vol.28, 273-284.

DOYLE LB (1965) : In Statistical association methods for mechanised documentation (Eds. M. E. Stevens, V.E. Guiliano, and L.B. Heilprin), pp 15-24, U.S. Department of Commerce, Washington D.C.

DUBES R & JAIN AK (1979) Validity studies in clustering methodologies, Pattern Recognition, Vol.11, 235-254.

DUBES R; JAIN AK (1976) Clustering techniques: the user's dilemma, Pattern-Recognition (UK), vol.8, no.4, p.247-60.

DUBIN D (1995) Document analysis for visualisation, Proceedings of the 18th International ACM-SIGIR Conference on Research and Development in Information Retrieval, 199-204.

EBERT DS; ZWA A; MILLER EL; SHAW CD; ROBERTS DA (1997) Two-handed volumetric corpus management, IEEE-Computer-Graphics-and-Applications (USA), Vol.27, No. 4, p.60-2, July-Aug. 1997, Published: IEEE.

EGUCHI K; ITO H; KUMAMOTO A; KANATA Y (1999) Adaptive and incremental query expansion for cluster-based browsing, Proceedings of the 6th International Conference on Database Systems for Advanced Applications, Hsinchu, Taiwan, 19-21 April 1999, p. 25-34.

EL-HAMDOUCHI A; WILLETT P (1986) Hierarchical document clustering using Ward's method, SIGIR 86, 149-156.

EL-HAMDOUCHI A; WILLETT P (1987) Techniques for the measurement of clustering tendency in document retrieval, J of IS, Vol.13 N.6, 361-5.

EL-HAMDOUCHI A; WILLETT P (1989) Comparison of hierarchic agglomerative clustering methods for document retrieval, Computer Journal, Vol.32 N.3, 220-7.

EVERITT B S (1993) Cluster analysis, Hodder&Stoughton, 3rd ed.

EVERITT B S (1992) The analysis of contingency tables, Chapman&Hall, 2nd ed.

FAIRTHORNE R (1963) “Mathematics, mechanics, and statistics for the information science cirriculum or what mathematics does an information scientist need?” In: Luhn, H.P., ed. Automation and Scientific Communication. Washington D.C.: American Documentation Institute 26th Annual Meeting, Part 1, 39-40.

FAIRTHORNE RA (1947) Proc. Brit. Soc. Int. Bibl., 9, 35.

FLOOD R; JACKSON M C (1991) Creative problem solving - total systems intervention, John Wiley & Sons, Chichester.

FROEHLICH T J (1994) Relevance reconsidered - Towards an agenda for the 21st century: introduction to special topic issue on relevance research, JASIS, Vol.45, No.3, 124-134.

GEBHARDT F (1975) A simple probabilistic model for the relevance assessment of documents, Information Processing and Management, Vol.11, 59-65.

GOOD I J (1958) Speculations concerning IR, Research report:PC-78, IBM Research Centre, Yorktown Heights,NY.

GORDON M D (1991) User-based document clustering by re-describing subject descriptions with a genetic algorithm, JASIS, Vol.42 N.5, 311-322.

GORDON AD (1981) Classification, Chapman and Hall, London.

GRIFFITHS A; ROBINSON L; WILLETT P(1984) Hierarchic agglomerative clustering methods for automatic document classification, Journal of Documentation, Vol.40, N.3, 175-205.

GRIFFITHS A; LUCKHURST; WILLETT P (1986) Using inter-document similarity information in document retrieval systems, JASIS, Vol.37, 3-11.

GUILIANO VE (1965) In: Statistical association methods for mechanised documentation (Eds. M.E. Stevens, V.E. Guiliano and L.B. Heilprin), pp. 25-32, U.S. Department of Commerce, Washington D.C.

HAGERTY K (1967) Abstracts as a basis for relevance judgement, Master’s thesis, Graduate Library School, University of Chicago, Chicago, IL.

HARMAN D (1996) Overview of the fourth Text Retrieval Conference (TREC-4), The Fourth Text Retrieval Conference (TREC-4), NIST, Gaithersburg.

HARMAN D (1993) Overview of the first Text Retrieval Conference (TREC-1), The First Text Retrieval Conference (TREC-1), NIST, Gaithersburg, 1-20.

HARPER DJ; MECHKOUR M; MURESAN G (1999) Document clustering for mediated information access, IRSG 99 - 21st Colloquium on Information Retrieval. Glasgow, Scotland, 19-20 April 1999, Draft Proceedings, ed. Monica Landoni.

HEARST M; PEDERSEN I (1996) Xerox TREC-4 site report. In: Harman, DK, ed., The Fourth Text REtrieval Conference. Gaithersburg, MD: NIST.

HEARST MA (1995) Visualisation of term distribution information in full text information retrieval. In Human factors in Computing Systems, CHI '95 Conference Proceedings, 59-66, Denver, ACM.

HULL D A (1996) Stemming algorithms: a case study for detailed evaluation, JASIS. V.47, N.1, 70-84.

JACOBS P S (1994) GE in TREC-2: Results of a Boolean approximation method for routing and retrieval, The Second Text Retrieval Conference (TREC-2). NIST, Gaithersburg, 191-199.

JAIN AK & DUBES RC (1988) Algorithms for clustering data, Prentice Hall.

JARDINE N; VAN RIJSBERGEN CJ (1971) The use of hierarchical clustering in information retrieval, Information Storage & Retrieval. Vol.7, 217-240.

JARDINE N; SIBSON R (1968) The construction of hierarchic and non-hierarchic classifications, Comp. J.. 11, 117-184.

JARDINE N; SIBSON R (1971) Mathematical taxonomy, Wiley, London&NY.

JARVIS RA; PATRICK EA (1973) Clustering using a similarity measure based on shared nearest neighbours, IEEE Transactions on Computers. C-22:1025-1034.

JONES S; WALKER S; GATFORD M; DO T (1997) Peeling the onion: Okapi system architecture and software design issues, Journal of Documentation. Vol.53 N.1, 58-68.

KAZHDAN T V (1979) Effects of subjective expert evaluation of relevance on the performance parameters of a document-based information-retrieval system, Nauchno-Tekhnicheskaya Informatsiya. Seriya 1, 13, 21-24.

KEEN E M (1991) The use of term position devices in ranked output experiments, Journal of Documentation. Vol.47, No.1, 1-22.

KIRRIEMUIR E; WILLETT P (1995) Identification of duplicate and near-duplicate full-text records in database search-outputs using hierarchic cluster analysis, Program. Vol.29 N.3, 241-256.

KORPIMIES K; UKKINEN E (1998) Term weighting in query-based document clustering, Advances in Databases and Information Systems. Second East European Symposium, ADBIS'98 Proceedings. Poznan, Poland, 7-10 Sept. 1998, p. 151-3.

KOSHMAN S; RASMUSSEN E (1994) IR research and innovation in commercial online systems: an exploratory survey, 22nd Annual Conference on the Information Industry in Transition. Montreal, Canada, 127-48.

KOSTER CHA; DERKSEN D van de Ende; POTJER J (1999) Normalization and Matching in the DORO system, IRSG 99 - 21st Colloquium on Information Retrieval. Glasgow, Scotland, 19-20 April 1999, Draft Proceedings, ed. Monica Landoni.

KRZANOWSKI WJ (1988) Principles of multivariate analysis – A user's perspective, Oxford University Press, New York, USA.

KUHNS JL (1965) In: Statistical association methods for mechanised documentation (Eds. M.E. Stevens, V.E. Guiliano and L.B. Heilprin), pp. 33-39, U.S. Department of Commerce, Washington D.C.

KUIPER FK & FISHER L (1975) A Monte Carlo comparison of six clustering procedures, Biometrics, 31, 777-783.

LANCE GN & WILLIAMS WT (1967) A general theory of classificatory sorting strategies: 1. Hierarchical Systems. Comp. J., 9, 373-380.

LERMAN HN (1970) A planar solution to 3-dimensional plotting, Software-Age (USA), vol.4, no.7, p.16-19.

LEWIS D (1992) An evaluation of phrasal and clustered representations on a text categorisation task, Proceedings of the 15th International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, 37-50.

LITOFSKY B (1969) Utility of automatic classification systems for information storage and retrieval, PhD thesis, U of Pennsylvania.

LU X A; HOLT J D; MILLER D J (1996) Boolean system revisited: its performance and its behaviour, The Fourth Text Retrieval Conference (TREC-4), NIST, Gaithersburg, 459-473.

MACSKASSY SA; BANERJEE A; DAVISON BD; HIRSH H (1998) Human performance on clustering Web pages: a preliminary study, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 27-31 Aug. 1998, p. 264-8.

MARCUS R S; KUGEL P; BENENFELD A R (1978) Catalogue information and text as indicators of relevance, JASIS, Vol.29, p.15-30.

MARON ME; KUHNS JL (1960) J. ACM, 7, 216-244.

MARKEY K; ATHERTON P; NEWTON C. (1980) An analysis of controlled vocabulary and free text search statements in online searches, Online-Review (UK) , Vol.4, no.3, p.225-36.

MARRIOT FHC (1971) Practical problems in a method of cluster analysis, Biometrics, 30, 499-505.

MECHKOUR M; HARPER DJ; MURESAN G (1998) The WebCluster project. Using clustering for mediating access to the World Wide Web, Proceedings of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Vic., Australia, 24-28 Aug. 1998, p. 357-8.

MEHTA C R & PATEL N R (1986) Algorithm 643 - FEXACT: A FORTRAN subroutine for Fisher's exact test on unordered *r* x *c* contingency tables, ACM Transactions on Mathematical Software, Vol.12, No.2, June 1986, 154-161.

MEHTA C R & PATEL N R (1983) A network algorithm for performing Fisher's exact test in rxc contingency tables, Journal of American Statistical Association, Vol.78, No.382, 427-434.

MILLER G A (1956) The magical number seven, plus or minus two. Some limits on our capacity for processing information, Psychological Review, V.63, 81-97.

MILLIGAN GW (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms, Psychometrica, 45, 325-342.

MINKER J; WILSON GA; ZIMMERMAN BH (1972) An evaluation of query expansion by the addition of clustered terms for a document retrieval system, Information Storage and Retrieval, 8, 329-348.

MIZZARO S (1997) Relevance: the whole history, JASIS, V.48, N.9, 810-832.

MORRIS R C T (1994) Toward a user-centred information service, JASIS, V.45, N.1, 20-30.

MURTAGH F (1983) A survey of recent advances in hierarchical clustering algorithms, Computer Journal, Vol.26, 354-359.

NOREAULT T; McGILL MJ; KOLL MB(1981) A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment, in: RN; Robertson SE; van Rijsbergen CJ; Williams PE, Information Retrieval Research, Butterworth & Co., 1981.

NOREAULT T; KOLL MB; McGILL MJ (1977) Automatic ranked output from Boolean searches in SIRE, JASIS, Vol.28, No.6, 333-339.

OTTAVIANI JS (1994) The fractal nature of relevance, JASIS, Vol.45 N.4, 263-72.

OVER P (1998) TREC-7 Interactive Track Report, The Seventh Text Retrieval Conference (TREC-7), NIST, Gaithersburg.

PARK T K (1992) The nature of relevance in information retrieval: an empirical study. Ph.D. dissertation. Bloomington, IN: Indiana University.

PEAT HJ; WILLETT P (1991) The limitations of term co-occurrence data for query expansion in document retrieval systems, Journal-of-the-American-Society-for-Information-Science (USA), vol.42, no.5, p.378-83.

PINKERTON B (1994) Finding what people want: experiences with the WebCrawler, Second International WWW Conference '94, July 1994, Chicago, USA (http://info.webcrawler.com/bp/WWW94.html)

PIROLLI P (1997) Computational models of information scent-following in a very large browsable text collection, CHI 97, Atlanta GA USA, 22-7 March 1997.

PIROLLI P; SCHANK P; HEARST M; DIEHL C (1996) Scatter-Gather browsing communicates the topic structure of a very large text collection, CHI 96, April 96, Vancouver, BC Canada.

PORTER M F (1980) An algorithm for suffix stripping, Program, Vol.14,No.3, 130-7, July 1980.

RADECKI T (1988) Trends in research on IR - the potential for improvements in conventional Boolean systems, IP&M, Vol.24, No.3, 219-27.

RAO; PEDERSEN (1995) Rich interaction in the digital library, Communications of the ACM, Vol.38 N.4, 29-39.

RATH G J; RESNICK A; SAVAGE T R (1961) Comparisons of four types of lexical indicators of content, American Documentation, Vol.12, N.2, 126-130.

REES A M; SCHULTZ D G (1967) A field experimental approach to the study of relevance assessments in relation to document searching (2 vols., NSF Contract No. C-423). Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University, Cleveland, OH.

RILOFF E (1995) Little words can make a big difference for text classification, Proceedings of the 18th International ACM-SIGIR Conference on Research and Development in Information Retrieval, 130-136.

ROBERTSON, SE; WALKER S; BEAULIEU M(1997a) Laboratory experiments with Okapi: participation in the TREC programme, Journal of Documentation, Vol.53 N.1, 20-34.

ROBERTSON, SE; BEAULIEU M (1997b) Research and evaluation in information retrieval, Journal of Documentation, Vol.53 N.1, 51-57.

ROBERTSON, SE (1990) On term selection for query expansion, Journal of Documentation, Vol.46 N.4, 359-364.

ROBERTSON, SE (1981) The methodology of information retrieval experiment. In: K. Sparck Jones, ed., Information Retrieval Experiment., Butterworth&Co., London, 9-31.

ROBERTSON, SE (1978) Ranking in principle.Journal of Documentation,Vol.34, No.2, 93-100.

ROBERTSON, SE; SPARCK JONES K (1976) Relevance weighting of search terms, JASIS, Vol.27, 129-46.

ROHLF FJ (1978) A probablilistic minimum spanning tree algorithm. Information Processing letters, 7, 44-48.

ROHLF FJ (1973) Hierarchical clustering using the minimum spanning tree. Computer J., 16, 93-5.

RU J S (1988) An evaluation of the applicability of ranking algorithms to improve the effectiveness of full text retrieval II. On the effectiveness of ranking algorithms on full-text retrieval, JASIS, Vol.39, No.3, 147-60.

SALTON G (1988) A simple blueprint for automatic Boolean query processing, IP&M, V.24, No.3, 269-280.

SALTON G; McGILL (1983) Introduction to Modern Information Retrieval, McGraw-Hill, Singapore.

SALTON G (1971) Cluster search strategies and the optimization of retrieval effectiveness. In G. Salton, editor, The SMART Retrieval System, 223-242. Prentice-Hall, Englewood Cliffs NJ.

SARACEVIC T; MAKROS H; SU L (1990) Nature of interaction between users and intermediaries in online searching: A qualitative analysis, Proceedings of 53rd ASIS Annual Meeting, V.27, 47-54.

SCHAMBER L (1994) Relevance and information behaviour In: Williams, Martha E. ed. Annual Review of Information Science and Technology, Vol.29, p 3-48.

SCHUTZE H; SILVERSTEIN C (1997) Projections for efficient document clustering, Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 74-81.

SHAW RJ; WILLETT P (1993) On the non-random nature of nearest-neighbour document clusters, IP&M, Vol.29 N.4, 449-52.

SHAW WM (1990) An investigation of document structures, IP&M, Vol.26, 339-348.

SHAW WM; BURGIN R; HOWELL P (1997) Performance standards and evaluations in IR test collections: cluster-based retrieval models, IP&M, Vol.33 N.1, 1-14.

SIBSON R (1973) SLINK: An optimally efficient algorithm for the single-link cluster method, Computer J., 16, 30-34.

SIEGEL S; CASTELLAN NJ (1988) Nonparametric statistics for the behavioral sciences, McGraw-Hill, International edition.

SIEGFRIED S; BATES MJ; WILDE DN (1993) A profile of end-user searching behaviour by humanities scholars: the Getty online searching project report #2, JASIS, V.44 N.5, 273-291.

SHIMAMURA H; TAKANO H; KAMBA T; KOSEKI Y (1998) A domain cluster interface for WWW search, Proceedings 9th International Workshop on Database and Expert Systems applications. Vienna, Austria, 26-28 August 1998, p.455-60.

SMIT PH; KOCHEN M (1988) Information impediments to innovation of on-line database vendors, IP&M. Vol.24, No.3, 229-41.

SNEATH PHA & SOKAL RR (1973) Numerical Taxonomy, W. H. Freeman and Co, San Francisco.

SPARCK JONES K (1997) Summary Performance Comparisons TREC-2, TREC-3, TREC-4, TREC5, http://www-nlpir.nist.gov/TREC/trec5.papers/sparckjones.ps

SPARCK JONES K (1981) Information Retrieval Experiment, London, Butterworths.

SPARCK JONES K (1973) Collection properties influencing automatic term classification performance, Information-Storage-and-Retrieval (UK), vol.9, no.9, p.499-513.

SPARCK JONES K (1970) Some thoughts on classification for retrieval, Journal of Documentation, V.26, 89-101.

SU L T (1994) On the relevance of recall and precision in user evaluation, JASIS, V.45 N.3, 207-217.

SWANSON D R (1988) Historical note: information retrieval and the future of an illusion, JASIS, Vol.39, No.2, 92-8.

TAGUE J M (1981) The pragmatics of information retrieval experimentation, in: Information Retrieval Experiment, ed. K. Sparck Jones, Butterworth&co., London, p. 59-102.

TIBBO H (1994) Indexing for the humanities, JASIS, V.45 N.8, 607-619.

TURTLE H (1994) Natural language vs. Boolean query evaluation: A comparison of retrieval performance, Proceedings of the 17th International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, 212-221.

V RIJSBERGEN C J; SPARCK JONES K, 1973, A test for the separation of relevant and non-rel docs in experimental test collections, Journal of Documentation, V.29, 251-257.

VAN RIJSBERGEN (1979) Information Retrieval, Butterworths, London.

VAN RIJSBERGEN (1981) Retrieval Effectiveness. In: Jones K Sparck, ed., Information Retrieval Experiment, Butterworth&co., London, p. 32-43.

VAN RIJSBERGEN C J; CROFT W B (1975) Document clustering: An evaluation of some experiments with the Cranfield 1400 collection, IP&M, V.11 , 171-182.

VOORHEES E M (1986) Implementing agglomerative hierarchical algorithms for use in document retrieval, IP&M, Vol.22 N.6, 465-476.

VOORHEES E M (1985) The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval, PhD thesis, Cornell U, Ithaca, NY.

WARD JH (1963) Hierarchical grouping to optimize an objective function. J. Am. Statist. Ass., 58, 236-244.

WEIS R L; KATTER, R V (1967) Multidimensional scaling of documents and surrogates (Tech. Memorandum SP-2713, 29 p.), Santa Monica, CA: Systems Development Corporation.

WEISS R; VALEZ B; SHELDON MA; NAMPREMPRE C; SZILAGYI P; DUDA A; GIFFORD DK (1996) Hypursuit: a hierarchical network search engine that exploits content-link hypertext clustering. Proceedings of Hypertext '96. Washington, DC, USA, 16-20 March 1996, p. 180-3.

WIBERLY S E; DAUGHERTY R A; DANOWSKI J A (1990) User persistence in scanning postings of a computer driven information system, LCS. Library & Information Science Research (Bay38). V.12, 341-353.

WIBERLY SE; DAUGHERTY RA (1988) User's persistence in scanning list of references, College & Research Libraries. V.49 N.2, 149-156.

WILLETT P (1988) Recent trends in hierarchic document clustering: a critical review, IP&M. Vol.24 N.5, 577-597.

WILLETT, P; WINTERMAN V; BAWDEN D (1986b) Implementation of non-hierarchical cluster analysis methods in chemical information systems. Selection of compounds for biological testing & clustering of substructure search output, Journal of Chemical Information &Computer Science. V.26, 109-118.

WILLETT P (1985) Ranked output searching in textual and structural databases, 9th International Online Information Meeting. London, UK, 3-5 Dec. 1985, 343-53.

WILLETT P (1983) Similarity coefficients and weighting functions for automatic document classification: an empirical comparison, International Classification. 10, 138-42.

WILLIAMS WT; LANCE GN; DALE MB; CLIFFORD HT (1971) Controversy concerning the criteria for taxonometric strategies, Computer-Journal (UK), vol.14, no.2, p.162-5.

WILLIAMS WT; CLIFFORD HT; LANCE GN (1971) Group-size dependence: a rationale for choice between numerical classifications, Computer-Journal (UK), vol.14, no.2, p.157-62.

YIMING-YANG; PIERCE T; CARBONELL J (1998) A study on retrospective and on-line event detection, Proceedings of 21st Intl. ACM SIGIR Conference on R&D in IR. Melbourne, Australia, 28-36.

YU C T (1974) A clustering algorithm based on user queries, JASIS. Vol.25, 218-226.

ZAMIR, O; ETZIONI, O (1998) Web document clustering: a feasibility demonstration, Proceedings of 21st ACM SIGIR Conference on R&D in IR, Melbourne, Australia, 24-28 Aug. 1998, p. 46-53.

ZAMIR O; ETZIONI O; MADANI O; KARP RM (1997) Fast and intuitive clustering of Web documents, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining – KDD 97. Newport Beach, CA, USA, 14-17 Aug. 97, p. 287-90.

Appendices

Appendix 1 : Lists of stopwords

Appendix 1.1 : Stopwords incorporated in the Okapi system

a, after, al-, along, also, although, among, an, are, as, at, be, been, before, behind, being, between, beyond, but, by, can, d', das, de, der, do, don't, down, een, for, from, go, had, has, have, he, her, here, hers, het, him, his, how, however, if, into, is, its, l', la, le, me, more, my, myself, no, off, on, our, over, s', so, some, than, that, the, their, them, then, there, these, they, this, those, through, throughout, thus, to, toward, under, up, upon, vol, was, we, were, what, whatever, when, whence, where, wherein, whether, which, who, whose, why, with, within, y, you, your, yourself

about, and, et, in, not, of, or, since, und (function words)

Appendix 1.2 : Semi-stopwords as defined by S. Walker

address, advance, advanced, al, all, another, any, application, approach, area, aspect, because, beginning, book, both, brief, briefly, ca, came, cant, century, circa, collected, collection, come, compared, concept, condition, conference, consider, considered, consist, consisted, contemporary, contribution, could, course, criticism, critique, currently, demonstrate, demonstrated, detail, development, die, different, difficult, discuss, discussed, done, during, each, early, easy, easily, edited, edition, editor, effect, eg, el, elementary, en, especially, essay, essential, etc, evaluation, example, exercise, few, finally, foundation, further, give, given, global, good, guide, handbook, happen, happened, held, ie, illustrated, impact, implication, importance, including, influence, inside, intermediate, international, introduce, introduction, introductory, investigating, investigation, involving, issue, iv, ix, journal, kind, kingdom, la, le, lecture notes, lecture, library, like, lo, made, main, make, making, manual, many, method, modern, monograph, most, much, must, new, now, num, number, obtain, obtained, often, old, only, organisation, original, other, out, outline, outlined, outside, overall, overview, pamphlet, paper, paperback, particular, particularly, party, past
perspective, possible, possibility, practical, practice, presented, previous, previously, principle, problem, proceeding, process, project, proposed, publication, published, publisher, rather, reader, reading, recent, recently, reference, related, relating, report, respect, review, role, same, seen, selected, seminar, series, service, several, short, should, similar, simple, source, sourcebook, special, state of the art, studies, study, subject, such, suggest, suggested, suitable, supplement, survey, system, take, technique, text, theme, theories, theory, therefore, today, together, tomorrow, topic, towards, understand, united, up to date, use, used, useful, using, very, via, view, volume, while, without, workbook, would, year

Appendix 1.3 : Final list of stopwords

a, able, about, above, absolute, absolutely, accept, access, accessed, accommodate, accomplish, according, accordingly, achieve, achieved, acquire, acquired, across, act, action, actual, actually, adapt, adapted, add, addition, address, addressed, adjust, adjusted, adopt, adopted, admit, admitted, admittedly, advance, advanced, advantage, affect, afterward, again, against, ago, agree, agreed, aim, al, al-, albeit, alike, all, allow, allowed, almost, alone, along, alongside, already, also, alter, alternatively, although, altogether, always, am, among, amount, an, analyze, analyse, analysis, annual, another, answer, answered, any, anyhow, anymore, anyway, apart, appear, apply, applied, approach, appropriate, appropriately, approximate, approximately, apr, april, arbitrary, arbitrarily, are, area, arent, argue, argued, argument, arise, around, arrange, arranged, article, as, aside, ask, asked, aspect, assume, assumed, assumption, attempt, attention, attract, attribute, aug, august, author, available, average, avoid, aware, away, b, back, base, basis, be, became, because, become, been, before, began, begin, beginning, begun, behind, being, belief, believe, believed, belong, below, beneath, beside, better, between, beyond, book, both, bottom, break, brief, briefly, bring, broad, broadly, brought, but, by, c, ca, call, came, can, cannot, cant, carry, cause, central, certain, certainly, change, character, check, circa, claim, clear, clearly, close, closely, colloquium, come, comment, common, complete, completely, compose, composed, comprehensive, comprise, comprised, conceive, conceived, concern, concerned, conclude, concluded, conclusion, condition, conference, confirm, congress, consequently, consequtive, consider, considered, consist, consisted, constitut, construct, contain, contained, continue, convention, correspond, corresponded, could, couldnt, course, cover, covered, create, created, d, das, date, day, de, deal, dealt, dec, decemb, december, decide, decided, decision, define, defined, definite, definitely, demonstration, demonstrate, demonstrated, depend, depended, der, description, describe, described, desirable, desire, despite, detail, determine, determined, did, didnt, differ, difference, different, direct, directly, discuss, discussed, distinguish, do, doesnt, don't, done, dont, down, draw, drawn, drew, driven, due, duration, dure, during, e, each, earlier, early, een, effect, eg, either, el, emerge, emfasis, emphasis, emphasize, employ, employed, en, enable, end, enough, ensure, enter, entire, equal, especial, especially, essential, essentially, establish, etc, even, ever, every, everything, exact, exactly, examine, example, exceed, except, exhibit, exist, existed, expect, expected, explain, explained, explanation, explicit, explicitly, express, expressed, extent, extremely, f, face, faced, fact, factor, fairly, fall, far, feb, february, fell, few, figure, figured, final, finally, find, fine, first, fit, focus, follow, for, foremost, form, former, forth, forward, found, foundation, four, frequently, friday, from, full, fully, further, furthermore, g, gave, general, get, give, given, glance, glimpse, go, gone, good, got, grasp, h, had, hadnt, hand, handl, happen, happened, hardly, has, hasnt, have, havent, held, hence, here, hereafter, hereinafter, hereby, herein, het, highly, hold, hope, how, however, hug, i, ibid, identify, identified, ie, iee, if, illustrate, illustrated, immediately, imply, implied, inc, include, included, including, indeed, indicate, indicated, initial, initially, inside, insignificant, insofar, instead, instance, instruction, intend, into, involve, involved, involving, is, isnt, issu, it, item, its, iv, ix, j, jan, january, jul, july, june, just, k, keep, kept, key, kind, knew, know, known, knowingly, l', l, la, last, lately, later, latter, le, lead, least, led, left, less, let, li, like, likely, little, lo, look, lot, lower, m, made, main, mainly, maintain, major, make, making, manifest, many, mar, march, markedly, massive, match, matter, may, maybe, mayjun, me, mean, meant, meanwhile, meet, mention, mere, merely, met, might, miss, mon, monday, more, most, mostly, move, moved, much, must, n, name, namely, near, nearly, necessarily, need, neither, never, nevertheless, new, next, no, none, nonetheless, nor, normal, normally, notably, note, noted, nothing, notice, notwithstand, nov, novemb, november, now, num, number, numerous, o, obtain, obtained, obvious, occur, occurred, oct, octob, october, off, often, old, once, one, ongo, ongoing, only, onto, open, opt, or, order, originally, other, otherwise, ought, our, out, outline, outlined, outside, over, overall, overview, own, p, part, partial, partially, particular, particularly, partly, per, perhaps, permit, place, placed, play,

Appendix 1.3 : Final list of stopwords – continuing from previous page

Played, plus, point, possibly, possible, possibility, precisely, prepare, prepared, present, presented, previous, primary, primarily, principal, principle, prior, priori, probable, probably, problem, proceed, proceeding, process, produce, produced, progress, project, proper, properly, proportional, propose, proposed, provide, provided, purpose, put, q, question, quick, quickly, r, raise, rather, reach, read, readily, reading, realise, realisation, realised, realize, realization, realized, really, reason, reckon, refer, referred, reflect, reflected, regard, regarded, regardless, relate, related, relating, relatively, rely, relied, remain, remark, render, report, reported, represent, represented, requirement, require, required, respect, respectively, result, rest, retain, retained, return, returned, reveal, revealed, review, reviewed, right, run, s', s, said, same, sat, saturday, say, saying, second, see, seek, seem, seemed, seemingly, seen, seminar, sent, sep, separately, sept, septemb, september, service, set, several, shall, short, should, shouldnt, show, shown, side, similar, simple, simply, since, so, socalled, sold, some, something, sometime, somewhat, soon, special, specially, specify, specific, specifically, specified, spite, st, state, still, straight, straightforward, strong, studies, study, subject, subsequent, such, suggest, suggested, suitable, summarise, summary, sunday, support, suppose, supposed, symposium, t, take, taken, talk, tell, tend, than, that, the, then, there, thereafter, thereby, therefore, these, thing, third, this, those, though, three, through, throughout, thurs, thursday, thus, to, today, together, told, tomorrow, too, took, topic, total, toward, towards, treat, treated, tried, true, try, tues, tuesday, turn, twice, two, u, under, underlie, underlying, understand, understood, undertake, undertaken, undertook, undoubtedly, unlike, until, up, upon, use, used, useful, using, usual, v, valid, variety, various, verlag, very, via, view, vis, vol, volume, von, w, want, was, wasnt, way, wednesday, well, went, were, werent, what, whatever, when, whence, where, whereas, wherein, whether, which, while, who, whom, whole, whose, why, wide, widely, will, with, within, without, wont, workbook, worst, would, wouldnt, x, y, year, yesterday, yet, yield, z, zero.

Appendix 1.4 List of synonyms

(abide, abode),(absence, absent),(academy, academic), (accounting, accountancy),(accru, accrual), (accurate, accuracy, accurately),(active, actively),(adequacy, adequate, adequately),(administrative, administration),(adverse, adversely),(advertising, advertisement, adverts),(advice, advise),(advisor, advisory),(aeroplane, airplane, aircraft),(aesthetic, esthetic),(afghan, afghanistan),(african, africa, afro),(ageism, agism, ageist, agist),(aggress, aggressively),(agricultural, agriculture),(air lines, airlines), (airport, aerodrome),(alaska, alaskan),(albania, albanian),(alcoholism, drunkeness),(algerian, algeria), (alps, alpine),(aluminium, aluminum),(american, america),(amplified, amplify),(analyse, analysis, analyze),(analytical, analytic),(angola, angolan),(anthropological, anthropology),(anti nuclear, antinuclear, anti-nuclear),(anti smoking, no smoking, no-smoking, nonsmoking, non smoking, antismoking), (anxiety, anxious),(anybody, anyone),(appendix, appendice),(apr, april),(arabia, arab, arabic, arabian),(argentina, argentinian),(arise, arose),(aristotle, aristotelean, aristotelian),(armenia, armenian),(asia, asian, asiatic),(athens, athenian),(au-pair, au pair),(aug, august),(australia, australian, australasia, australasian),(austria, austrian, austro),(auto immune, autoimmune),(autonomy, autonom), (awake, awoke),(back up, backup),(bad, badly),(bankrupt, bankruptcy),(bavarian, bavaria),(bbc, british broadcasting corporation),(bear, bore),(beat, beaten),(beautiful, beautifully),(belgium, belgian),(bend, bent),(big, bigger),(bind, bound),(bite, bitten),(bleed, bled),(blow up, blowup),(blow, blew, blown), (bluecollar, blue-collar, blue collar),(bolshevik, bolshevism),(bosnian, bosnia),(brazil, brazilian),(break down, breakdown),(break off, breakoff),(break up, breakup),(break, broke, broken),(brecht, brechtian), (breed, bred),(bright, brighter),(broad, broader, broadly),(bt, british telecom),(build up, buildup),(build, built),(built in, builtin),(bulgarian, bulgaria),(bureau, bureaus),(burn, burnt),(buses, omnibus, bus), (businessman, businessmen),(buy back, buyback),(buy in, buyin),(buy, bought),(buyout, buy out),(cad, computer aided design),(cal, cai, computer aided instruction, computer aided learning, computer assisted learning),(california, calif, californian),(cambodia, cambodian),(canada, canadian),(capital punishment, death penalty),(catch, caught),(cba, cost benefit analysis),(cd-rom, cd rom),(center, centre, central),(ceo, chief executive officer),(certify, certified),(chairman, chairmen, chairperson),(cheap, cheaper, cheaply),(check out, checkout),(check up, checkup),(chief, chiefly),(child, children, childhood),(chile, chilean),(china, chinese),(choose, chose, chosen),(chunnel, channel tunnel),(cia, central intelligence agency),(clarify, clarified),(classify, classification),(clean, cleaner),(cleanup, clean up),(cling, clung),(close up, closeup),(closer, closely),(cnd, campaign for nuclear disarmament),(co generation, cogeneration),(co, company),(community charge, poll tax),(compare, compared, comparatively, comparison),(compete, competition, competitive),(complain, complaint, complained), (complem, complementary),(computation, computational),(concur, concurred, concurring), (congressman, congressmen),(conservative party, tory party, conservative and unionist party),

Appendix 1.4 List of synonyms – cont.

(consortia, consortium),(conspire, conspired, conspiracy),(constant, constantly),(constrain, constraint), (converse, conversely),(correct, correctly),(corsica, corsican),(cost, costly),(cost-of-living, cost of living),(count down, countdown),(creep, crept),(criterion, criteria),(croatia, croatian),(cry, cried),(cttee, committee),(cuba, cuban),(curriculum, curricula, curricular),(cut back, cutback),(cut off, cutoff),(cut out, cutout),(cyprus, cypriot),(czar, csar, tsar, tzar),(czechoslovakia, czech, czechoslovak, czechoslovakian), (data base, database),(dbms, database management system),(debug, debugged), (dec, decemb, december),(deep, deeper, deeply),(delay, delayed),(deliberate, deliberately), (demonstrate, demonstrated, demonstration),(denmark, danish, dane),(deploy, deployed),(design, designed),(desper, desperately),(destined, destiny),(destroy, destroyed),(deter, deterred, deterring), (develop, developed, development),(dhss, department of health and social security),(diagnose, diagnosed, diagnosis),(dial up, dialup),(die, dying, death),(diet, dietary),(difficult, difficulti),(dig, dug),(dirt, dirty),(disagreement, disagree, disagreed),(disc, disk),(discontinue, discontinued),(discretion, discretionary),(display, displayed),(disproportion, disproportionately),(disqualify, disqualified),(diversify, diversified, diversification),(dm, dm1, dm2, dmark),(dod, department of defense),(doi, department of industry),(dream, dreamed, dreamt),(drink, drank, drunk),(drive, driven, drove),(drop out, dropout),(dropin, drop in),(dry, dried),(dwell, dwelt),(dwi, driving while intoxicated),(east, eastern),(easy, easier, easily),(eat, ate, eaten),(economic, economy),(education, educational),(eec, european economic community, eu, european union, ec),(egypt, egyptian),(elder, elderly),(email, electronic mail),(empowerment, empower),(enduser, end-user, end user),(england, english),(enhance, enhence),(enjoy, enjoyed),(enrol, enrolled),(enthusiasm, enthusiastic),(entire, entirely, entirety),(environment, environmental),(epa, environmental protection agency),(equit, equity),(eritrean, eritrea),(erosion, erode),(ethiopia, ethiopian),(eurodollar, euro-dollar, euro dollar),(europe, european),(everybody, everyone),(exclus, exclusion),(expedit, expedition),(expence, expense),(experiment, experimental, experimentally, experimentation),(expire, expired, expiry),(explicit, explicitly),(extensive, extensively),(extra-terrestrial, extraterrestrial, extra terrestrial),(extraordinary, extraordinarily),(fade out, fadeout),(failure, fail),(falklands, falkland islands, malvinas),(fall out, fallout), (fall, fallen, fell),(fao, food and agriculture organisation),(fast, faster),(fbi, federal bureau of investigation),(fda, federal drug administration),(feb, februari),(feed, fed),(feel, felt),(fema, federal emergency management agency),(few, fewer),(ffr, ffr1, ffr2, ffr3),(fiber, fibre),(fierce, fiercely),(fight, fought),(financial, finance),(finland, finnish),(firmer, firmly),(flanders, flemish),(flee, fled),(fling, flung),(florence, florentine, firenze),(fly, flew, flown, flied),(fold up, foldup),(followon, follow on),(foot, feet),(forbid, forbidden, forbad),(foreclose, foreclosure),(foretell, foretold),(forget, forgotten, forgot),(forgive, forgave, forgiven),(former, formerly),(formul, formula),(forsake, forsook, forsaken),(france, french, francais),(franchised, franchise),(free, freely),(freeze, frozen, froze),(freudian, freud, sigmund freud),(friend, friendly),(front-end, frontend, front end),(fuel-cell, fuel cell),(fulltext, full-text, full text),(gaol, jail),(gases, gassing, gaseous, gas),(gatt, general agreement on tariffs and trade),(gdp, gross domestic product),(gec, general electric company),(gentleman, gentlemen),(genuine, genuinely),(geographical, geography),(geological, geology),(geometrical, geometry),(germany, german, germanic, germano, deutsch, deutschland),(get out, getout),(gilt-edge, giltedge, gilt edge),(gingrich, newt gingrich),(gloom, gloomy),(gnp, gross national product),(goose, geese),(gov, govt, governmental, government),(graphical, graphic),(great, greater, greatly),(greece, greek, grecian),(grow, grew, grown), (grown up, grownup, adult),(guarante, guaranteed),(guilt, guilty),(guncontrol lobby, guncontrol, gun control),(gunman, gunmen),(haemophiliac, haemophilia),(haiti, haitian),(half, halv, halved, halves),(hand out, handout),(handicapped, disadvantaged),(hang out, hangout),(hang, hung),(happy, happily),(hard, harder),(harmon, harmony),(hawaii, hawaiian),(he, him, his, hed, he'd, he's, hes, himself),(hear, heard),(heavy, heavier, heavily),(hebrides, hebridean),(hegel, hegelian),(her, hers, she, she'd, she's, shes, herself),(hero, heroic),(herts, hertford, hertfordshire),(hide, hidden, hid),(hierarch, hierarchy, hierarchical, hierarchically),(high school, highschool),(high, higher),(hightech, hitech, high tech), (himalayas, himalayan),(hold out, holdout),(hold up, holdup),(holland, dutch, netherlands),(hondura, honduran),(hong kong, hongkong),(hook up, hookup),(hope, hopefully, hopeful),(hungary, hungarian), (hunger, hungry),(i'd, i'll, i'v, i'm, myself, my),(iaea, international atomic energy authority),(iba, independent broadcasting authority),(iceland, icelandic),(ici, imperial chemical industries),(ideolog, ideology),(imf, international monetary fund),(impact, effect),(impede, impediment),(implement, implemented),(improper, improperly),(in house, inhouse),(incorrect, incorrectly),(increase, increasingly), (indebt, indebted),(indefinite, indefinitely),(india, indian, indo),(indices, index),(indirect, indirectly),(indo-european, indo european, indoeuropean),(indonesian, indonesia),(industrial, industry),(infect, infective), (inflation, inflationary),(influence, influential),(infra red, infrared),(injure, injury, injured),(instant, instantly),(intence, intense),(internationally, internationalism, international),(intervene, intervention), (introduce, introduced, introductory, introduction),(iq, intelligence quotient),(ira, irish republican army), (iran, iranian, persia, persian),(iraq, iraqi),(ireland, irish),(isdn, integrated services digital network), (israel, israeli),(italy, italian, italia, italiana, italiano),(itn, independent television news),(jan, januari), (japan, japanese),(jeopard, jeopardy),(jerk off, jerkoff),(jewelry, jewellery),(jewish, judaism, jew),(joint, jointly),(jordan, jordanian),(judgement, judgment),(judicial, judiciary),(jul, july),(jun, june),(jung, jungian),

Appendix 1.4 List of synonyms – cont.

(justified, justification, justify),(kenya, kenyan),(keynes, keynesian),(kg, kilogram),(kibbutz, kibbutzim), (klein, kleinian),(km, kilometer),(kneel, knelt),(knock down, knockdown),(knock out, knockout),(know, knew, known, knowingly),(koran, quran, q'ran),(korea, korean),(kurd, kurdish),(kuwaiti, kuwait),(labour party),(lan, local area network),(large, largely, larger),(lay off, layoff),(lay out, layout),(lay, lai, laid),(lb, lbs),(lean, leant),(leap, leapt),(learn, learnt),(lebanon, lebanese),(legend, legendary),(lend, lent),(libya, libyan),(licence, license),(lifesupport, life support),(lift off, liftoff),(light, lit),(line up, lineup),(lithuania, lithuanian),(lock out, lockout),(lock up, lockup),(longterm, longer term, long term),(look up, lookup),(loos, loosen),(low down, lowdown),(lowcost, low cost),(lse, london school of economics),(ltd, limited),(luck, lucky),(lunacy, lunatic, mad, madness, insane, insanity),(macedonian, macedonia),(machiavelli, machiavellian),(macro economics, macroeconomics),(maff, ministry of agriculture fisheries and food), (make up, makeup),(malaysia, malaysian),(malta, maltese),(man, men),(maneuv, maneuvr),(mar, march),(marcuse, marcusian),(mark up, mark-up, mark up),(marx, marxian, marxist),(math, maths, mathematics),(matrix, matrices),(mddx, middlesex),(mediumterm, medium term),(medical, medicine), (melt down, meltdown),(memo, memorandum),(meter, metr),(mexican, mexico),(micro computers, microcomputers, micro-computers),(micro economics, microeconomics),(microelectronics, micro electronics, micro-electronics),(middle ages, medieval, mediaeval),(milan, milanese),(milit, military), (minimal, minim),(minnesota mining and manufacturing, mmm),(mis, management information systems), (miserable, miser, misery),(mislead, misled),(mistake, mistaken, mistook),(misunderstand, misunderstood),(mit, massachusetts institute of technology),(mock up, mockup),(moderately, moderate),(modest, modestly),(modify, modification),(mold, mould),(mongolia, mongolian), (monogamous, monogamy),(monopolies and mergers commission),(month, monthly),(mortgaged, mortgage),(moslem, muslim, islamic, islam),(mouse, mice),(movies, moving pictures, cinematic, cinema),(mp, memberof parliament),(mt, mount),(napoleon, napoleonic),(narrow, narrowly),(nasa, national aeronautical space agency),(nato, north atlantic treaty organization),(nazi, nazis),(ncb, national coal board),(near, nearby, nearer),(near-death, near death),(network, networked),(new york, ny),(new zealand, nz),(new, newer, newly),(nf, national front),(nhs, national health service),(nicaragua, nicaraguan),(nigeria, nigerian),(nlp, natural language processing),(northeast, northeastern),(northwest, northwestern),(norway, norwegian, norse),(notified, notify),(nov, novemb, november),(nspcc, national society for the prevention of cruelty to children),(o and m, organisation and methods),(oap, senior citizen, old age pensioner),(occupy, occupied),(occur, occurred, occurring),(oct, octob, october),(oecd, organisation for economic cooperation and development),(offence, offense),(old, older),(on line, online, on-line),(opec, organisation of petroleum exporting countries),(operatic, opera),(operations research, operational research),(optimum, optimal),(ordinary, ordinarily),(os 2, os/2, os-2, os2),(ounc, oz), (pakistan, pakistani),(palestine, palestinian),(panama, panamanian),(papua, papuan, new guinea, papua new guinea),(paraguay, paraguayan),(paris, parisian, parisien, parisienne),(part-time, part time),(partially sighted, blind),(paste up, pasteup),(pavlov, pavlovian),(pay off, payoff),(pay out, payout), (pay, pai, paid),(pembs, pembroke, pembrokes, pembrokeshire),(per cent, percent, percentage),(perfect, perfectly),(persons, people),(pertain, pertinent),(peru, peruvian),(petrol, petroleum), (ph d, phd),(phenomenon, phenomena),(philosophical, philosophy),(phone, telephone), (phonein, phone-in, phone in),(physiological, physiology),(pick up, pickup),(pigs, swine),(pinup, pin-up, pin up),(plag, plagu),(play off, playoff),(plo, palestine liberation organisation),(poland, polish), (policeman, policemen),(polygamous, polygamy),(polymerisation, polymer),(polyphonic, polyphony), (poor, poorer, poorly),(porn, porno, pornography, pornographic),(portray, portrayed),(portugal, portuguese), (post war, post-war, postwar),(predominant, predominantly),(preliminary, preliminarily), (presidentially, presidential),(prestige, prestigious),(prevail, prevalent),(print out, printout),(private, privately),(prolif, proliferate),(prompt, promptly),(proportion, proportionate),(prove, proven),(province, provincial),(prussia, prussian),(psbr, public sector borrowing requirement),(psychiatric, psychiatry), (psychological, psychology),(psychology of language, psycholinguistics),(purchase, buying, bought, buyer, buy),(qualify, qualified),(quantit, quantity),(quarter, quarterly),(quick, quicker, quickly),(quiet, quietly),(racism, racist, racialism, racialist),(rail road, railroad, railway),(rapid, rapidly),(rare, rarely), (ratify, ratified),(read out, readout),(reagan, ronald reagan, mr reagan, president reagan, president ronald reagan),(reconsider, reconsidered),(recovery, recover, recovered),(recur, recurr, recurred), (reenter, reentry),(regular, regularly),(religious, religion),(reluctant, reluctantly),(repay, repai, repaid, repayment),(repeat, repeatedly),(reproduce, reproduction),(response, responce),(retir, retire, retired), (revise, revision),(revolution, revolutionary),(rich, richer),(right, rightly),(ring, rang, rung),(rip off, ripoff),(rise, risen),(risk, risky),(rival, rivalry),(romania, romanian),(rough, roughly),(round off, roundoff), (round up, roundup),(routine, routinely),(rspb, royal society for the protection of birds),(rspca, royal society for the prevention of cruelty to animals),(rub down, rubdown),(run off, runoff),(run, ran), (russia, russian, soviet, soviet union, ussr, union of soviet socialist republics),(sacrify, sacrifice),(sad, sadly), (safe, safely, safer),(salop, shrops, shropshire),(salt 2, salt ii),(satisfactory, satisfactorily),(saudi, saudiarabia, saudis),(saw, sawed, sawn),(scandinavia, scandinavian),(scarce, scarcely),(sci fi, sci-fi, scifi, science-fiction, science fiction),(scottish, scots, scotland),(scrutinize, scrutinise, scrutiny, scrutinised, scrutinised),(secret, secretly),(seek, sought),(self sufficiency, selfsufficient, self sufficient), (sell out, sellout),(sell, sold),(send, sent),(sep, sept, september, septemb),(separate, separately), (serbia, serbian),(serbocroat, serbo-croat, serbo croat, serbocroatian, serbo-croatian, serbo croatian),(set up, setup),(severe, severely),(shake, shaken, shook),(sharp, sharply),(shear, sheared, shorn),(shelf, shelv, shelves),(shine, shone),(shoot, shot),(short, shorter),(shortterm, shorter term, short term),(shrink, shrank, shrunk),(shut down, shutdown),(shut off, shutoff),(shut out, shutout),(significant, significantly),(silence, silent),(simple, simpler),(sincerE, sincerely),(sing, sang, sung),(sink, sank, sunk),(sit in, sit ins, sitin),(sit, sat),(slay, slew, slain),(sleep, slept),(slight, slightly),(sling, slung),(slink, slunk),(slovak, slovakia),(slow down, slowdown),(slow, slower, slowly),(slowdown, slow down),(smell, smelt),(smooth, smoothly),(social science, sociology),(social work, social worker),(societ, society),(sole, solely),(somali, somalian, somalia),(somebody, someone),(soon, sooner),(south african, south africa),(south america, latin america, south american, latin american),(south, southern),(southeast, southeastern),(southwest, southwestern),(sow, sowed, sown),(soy, soya, soybean, soya bean, soyabean),(spain, spanish, hispanic, espana),(speak, spoke, spoken),(speed up, speedup),(speed, sped),(spell, spelt),(spend, spent),(spill, spilt),(spin off, spinoff),(spin, spun),(spit, spat),(spoil, spoilt),(spokesman, spokesmen, spokeswoman),(spring, sprang, sprung),(stagnate, stagnant),(stake out, stakeout),(stand off, standoff),(stand, stood),(standard deviation),(standard of living, living standards),(star wars, sdi, strategic defense initiative),(start up, startup),(state-of-the-art, state of the art),(stats, statistics, statistical),(statut, statutory),(stay, stayed),(std, sexually transmitted diseases, vd, venereal disease),(steady, steadily),(steal, stolen, stole),(stick, stuck),(stimulate, stimulus),(sting, stung),(stink, stank),(stockmarket, stock market),(strategic, strategy),(strew, strewed, strewn),(strict, stricter, strictly),(stride, strode, stridden),(strike, struck),(strong, stronger, strongly),(submit, submitt, submitted),(suburb, suburban),(success, successfully, successful),(sudden, suddenly),(suffice, sufficient),(supplement, supplem, supplementary),(sure, surely),(surprise, surprised, surprisingly),(survey, surveyed),(suspicion, suspicious),(swear, swore, swor, sworn),(sweden, swedish, sverige),(sweep, swept),(swell, swollen),(swift, swiftly),(swim, swam, swum),(swing, swung),(switzerland, swiss, suisse),(swp, socialist workers party),(symposia, symposium),(syria, syrian),(taiwan, taiwanese),(take off, takeoff),(take up, takeup, take-up),(takeover, take over, take-over),(tanzania, tanzanian),(tax, taxation, taxable, taxability),(teach, taught),(tear, tore, torn),(technic, technique),(technology, technological),(telecommun, telecom),(temporary, temporarily),(terrible, terribly),(texan, texas),(thai, thailand),(thatcher, mrs thatcher, margaret thatcher, mrs margaret thatcher, thatcherite, thatcherism),(theses, thesis),(they, theyd, theyll, theyv, them, theyre, their, themselves),(third world, 3rd world, under developed countries, underdeveloped areas, underdeveloped countries, under-developed countries, developing countries),(thorough, thoroughly),(thrive, throve),(throw, threw, thrown),(tibet, tibetan),(tight, tighter, tightly),(tigray, tigre),(time out, timeout),(tooth, teeth),(touch down, touchdown),(tough, tougher),(toys, toying, toy),(trade off, tradeoff),(trading on inside information, trading by insider, trading by an insider, insider trading),(transfer, transfere, transferred, transferr),(transmit, transmitt, transmitted),(trash, garbage),(tread, trod, trodden),(tribute, tributary),(trick, tricky),(true, truly),(tuc, trade union congress),(turkey, turkish),(turn off, turnoff),(turn out, turnout),(tv, television),(uk, united kingdom, britain, great britain, british isles, gb, british, gt brit),(ukrain, ukrainian),(ultimate, ultimately),(undergo, undergone),(underwrite, underwritten),(undu, unduly),(unesco, united nations educational scientific and cultural organization),(unexpect, unexpectedly),(unfair, unfairly),(unfortunate, unfortunately),(unidentified flying object, ufo),(unify, unified),(unnecessary, unnecessarily),(unreli, unreliable),(unsuccess, unsuccessfully), (up to date, up-to-date),(urgence, urgent),(uruguay, uruguayan),(usa, us of a, us, united states, united states of america),(usda, united states department of agriculture),(userfriendly, user friendly),(utd, united),(uv, ultra violet, ultraviolet),(vast, vastly),(vat, value added tax),(vdu, vdt, visual display unit, visual display terminal),(venezuelan, venezuela),(venice, venetian, fenice),(verify, verified, verification),(viable, viability),(vienna, viennese),(vietnam, vietnamese),(violence, violent, violently),(vlsi, very large scale integration),(voluntary, voluntarily),(wake, woke, woken),(wales, welsh),(walk out, walkout),(warm up, warmup),(we, we'd, we're, we've, our, ours, ourselves),(weak, weaker),(wear, wore, worn),(weave, wove, woven),(weep, wept),(west indies, west indian),(whitecollar, white-collar, white collar),(wide, widely, wider),(wife, wives),(win, won),(withdraw, withdrawn, withdrew),(withhold, withheld),(withstand, withstood),(wolf, wolves),(woman, women),(wood, wooden),(wordprocessing, word processing),(world health organization),(world war i, world war one, world war 1, first world war, 1st world war, 1914-18 war, world war 1914-1918, european war 1914-1918, premiere guerre mondiale),(world war ii, world war two, world war 2, second world war, 2nd world war, 1939-45 war, world war two 1939-45, world war 1939-1945, seconde guerre mondiale, deuxieme guerre mondiale),(worldwide, world wide),(wring, wrung),(write down, writedown),(write off, writeoff),(write up, writeup),(write, written, wrote),(wsj, wall street journal),(x 11, x windows, xwindows, x11),(x ray, x-ray, xray),(year book, yearbook),(yorks, yorkshire),(you, yorself, youd, you'd, youll, youve, you'll, you've, you're, youre, yourself, yourselves),(young, younger),(yugoslavia, yugoslav, yugoslavian),(zcars, z-cars, z cars)

The following abbreviations are used in the tables in Appendix 1.5 aand Appendix 1.6:

Document representation formats: **BP** denotes best passages, **FD** denotes full documents
Seed selection method: **C** denotes checking seed candidates' coverage by previous seeds and skipping them if they are maximally covered by any of the previous seeds, NC denotes checking seed candidates only for whether they are identical to a previous seed using a small threshold value of difference between seed powers),
Selection of representative terms: **TO** denotes the method of choosing the terms with respect to their weights within the search output, **TN** denotes choosing them from individual documents (again with respect to their weights within the search output) in order to ensure a more balanced distribution of representative term occurrences among documents.

Appendix 1.5 : Overlap between clusters in TREC queries with different combinations of term and seed selection methods, and document representation formats

Cluster overlap rates for TREC relevant document sets (TREC-REL)

query	BP-C	BP-NC	FD-C	FD-NC
254	5%	8%	3%	5%
256	0%	4%	8%	0%
258	6%	13%	11%	6%
264	3%	8%	11%	36%
284	2%	0%	4%	15%
286	16%	19%	37%	37%
292	7%	10%	3%	15%
average	6%	9%	11%	16%

Cluster overlap rates for TREC user query results (TREC-USER)

query	BP-C	FD-C	BP-NC	FD-NC
254	8%	0%	4%	8%
256	8%	46%	2%	52%
258	18%	14%	14%	22%
264	6%	48%	4%	70%
284	4%	8%	40%	36%
286	0%	2%	26%	6%
292	2%	16%	2%	20%
average	6.57%	19%	13%	31%

Overlap rates with the new method(TN) of choosing representative terms and BP-C

query	TREC- relevant documents (TREC-REL)		TREC-user documents (TREC-USER)	
	TO	TN	TO	TN
254	5.13%	3.33%	8%	4%
256	0.00%	0%	8%	0%
258	6.38%	0%	18%	2%
264	2.78%	0%	6%	0%
284	1.92%	1.85%	4%	0%
286	15.79%	0%	0%	0%
292	6.78%	0%	2%	0%
average	5.54%	0.74%	6.57%	0.85%

Appendix 1.6. Comparison of average coverage values with respect to the seed selection method used

(see Appendix 1.5 for abbreviations used)

Inspec queries - Coverage by assigned seed - for documents maximally covered by the rejected seed

query	C	NC
Q1	0.034	0.044
Q2	0.045	0.052
Q3	0.040	0.042
Q4	0.049	0.042
Q5	0.038	0.043
Q6	0.035	0.037

Inspec queries - Coverage by assigned seed - for all non-seed documents

query	C	NC
Q1	0.047	0.050
Q2	0.052	0.051
Q3	0.045	0.042
Q4	0.049	0.045
Q5	0.056	0.055
Q6	0.043	0.039

Coverage by assigned seed - for all non-seed documents (TREC-REL)

query	BP-C	BP-NC	FD-C	FD-NC
254	6.44%	6.40%	5.79%	5.42%
256	7.24%	7.05%	6.04%	6.38%
258	6.60%	5.90%	6.62%	7.46%
264	6.90%	7.13%	8.02%	4.26%
284	5.58%	4.76%	5.58%	5.50%
286	5.07%	5.04%	4.96%	5.09%
292	4.21%	3.59%	4.32%	3.68%
average	6.01%	5.70%	5.91%	5.40%

Coverage by assigned seed - for all non-seed documents (TREC-USER)

query	BP-C	FD-C	BP-NC	FD-NC
254	7.14%	7.52%	7.08%	6.91%
256	6.18%	7.76%	5.94%	7.87%
258	6.15%	7.01%	7.46%	8.15%
264	6.32%	7.33%	5.68%	7.45%
284	6.68%	6.29%	5.53%	4.71%
286	6.71%	7.29%	5.98%	6.52%
292	4.94%	6.09%	4.83%	6.22%
average	6.30%	7.04%	6.07%	6.83%

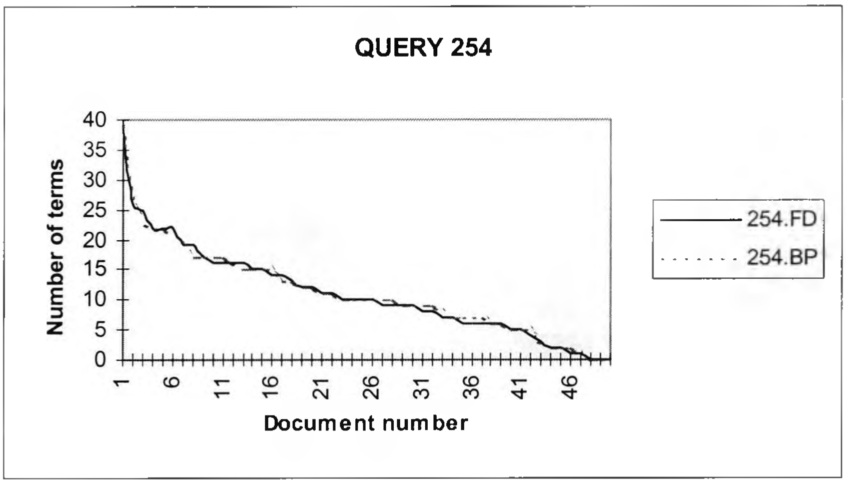
Coverage by assigned seed : old(TO) vs new method(TN) of term selection

	TREC - relevant documents (TREC-REL)		TREC - user documents (TREC-USER)	
	TO	TN	TO	TN
254	6.44%	5.01%	7.14%	5.36%
256	7.24%	6.76%	6.18%	6.03%
258	6.60%	4.26%	6.15%	4.83%
264	6.90%	5.68%	6.32%	4.28%
284	5.58%	4.63%	6.68%	4.76%
286	5.07%	3.03%	6.71%	4.89%
292	4.21%	3.03%	4.94%	4.85%
average	6.01%	4.67%	6.30%	5.00%

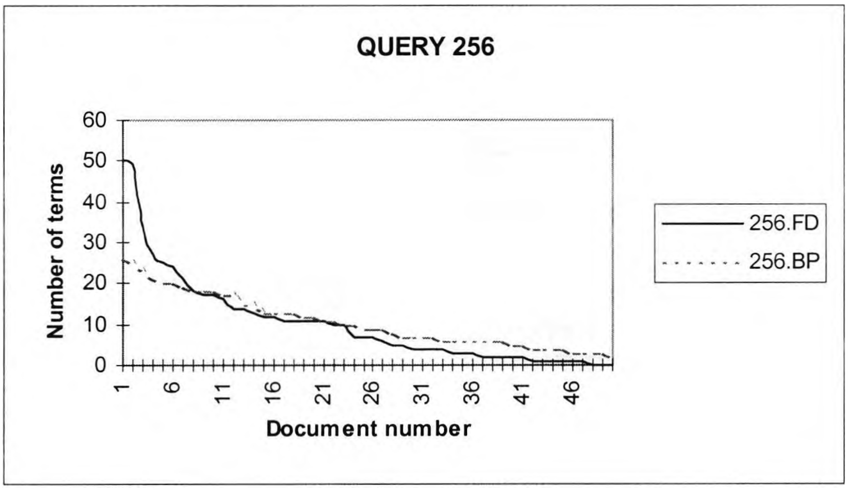
Appendix 1.7 Number of representative terms in documents

FD denotes full documents, BP denotes best passages

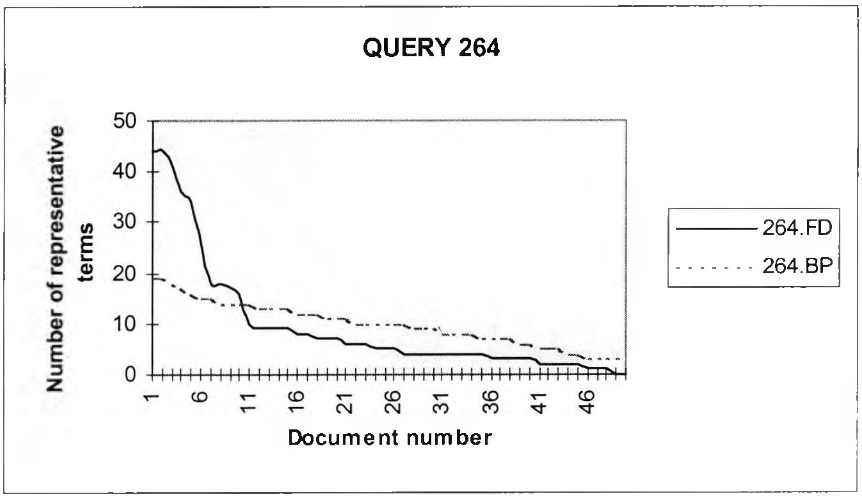
Q254 - number of representative terms in documents (TREC – USER)



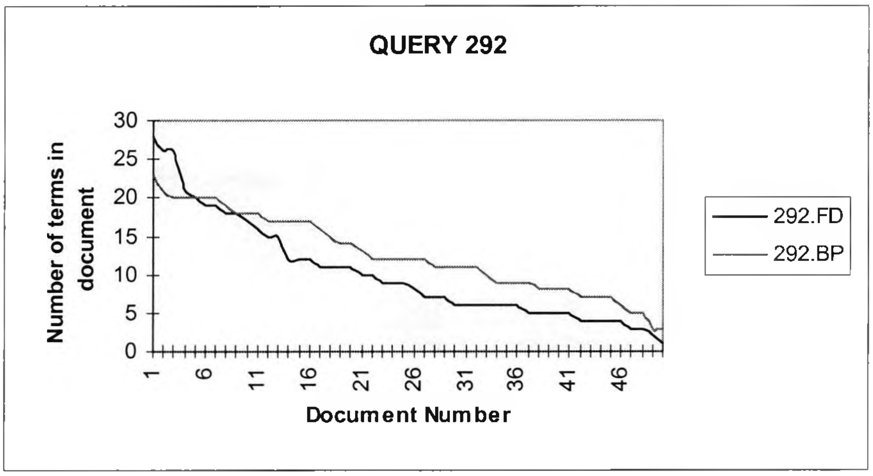
Q256 - number of representative terms in documents (TREC – USER)



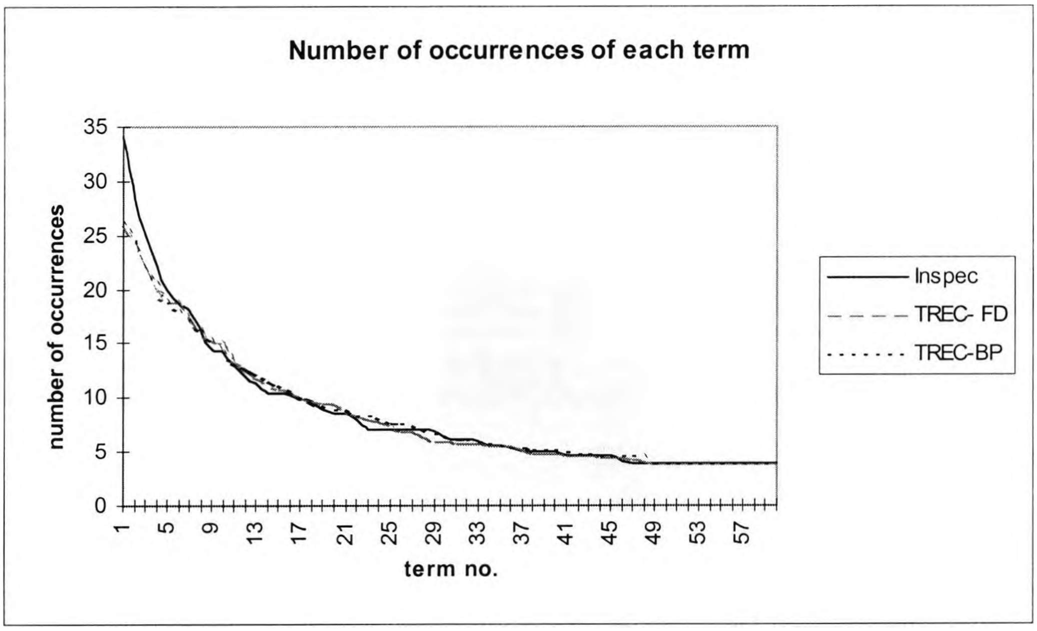
Q264 - number of representative terms in documents (TREC – USER)



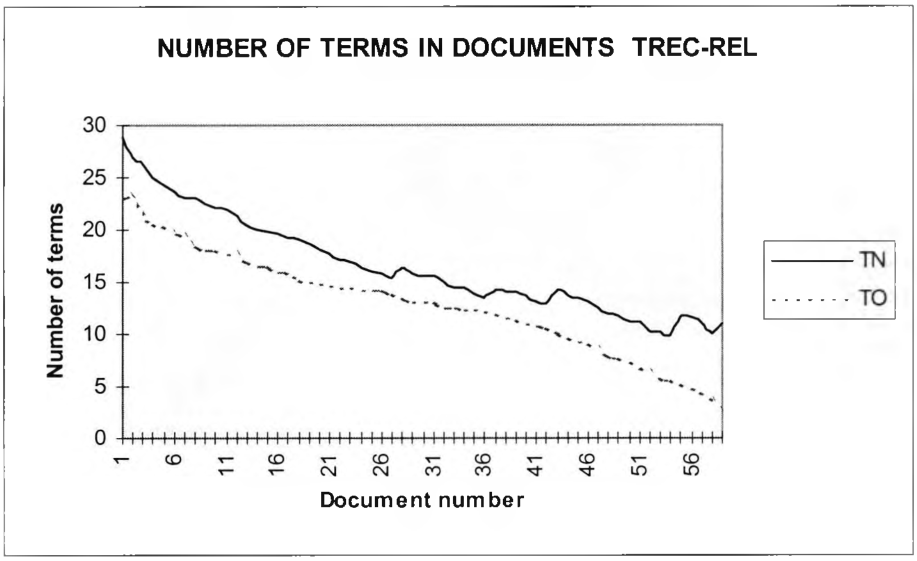
Q292 - number of representative terms in documents (TREC – USER)



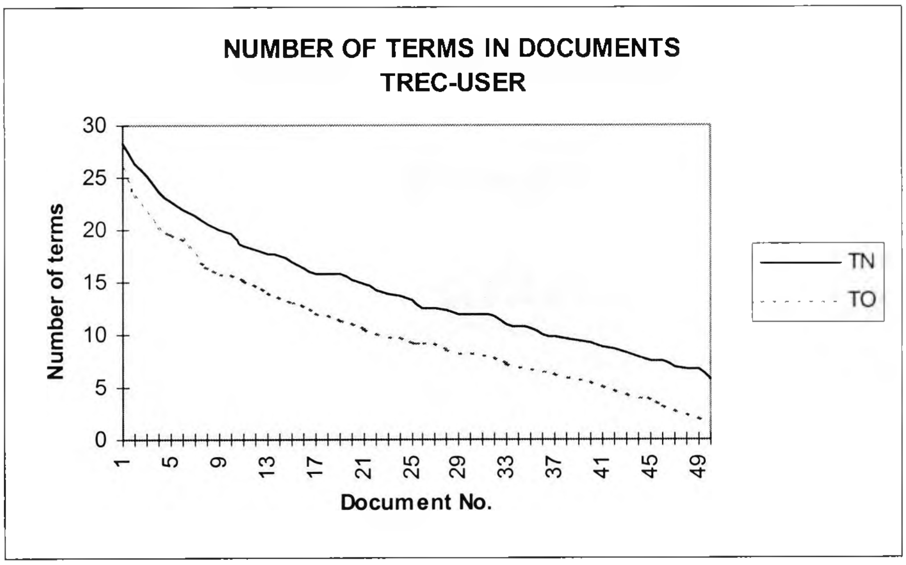
Number of occurrences of each term averaged over queries - Clusters based on TREC user documents(TREC-USER) vs Inspec query results



TREC-REL Experiments - Number of terms in documents when clustering sets of relevant documents: the final term selection method(TN) vs previous term selection method(TO)



TREC-USER Experiments - Number of terms in documents when clustering documents retrieved by the users: with the final term selection method(TN) vs previous term selection method(TO)



Appendix 1.8: Standard deviations in cluster sizes

(see Appendix 1.5 for abbreviations used)

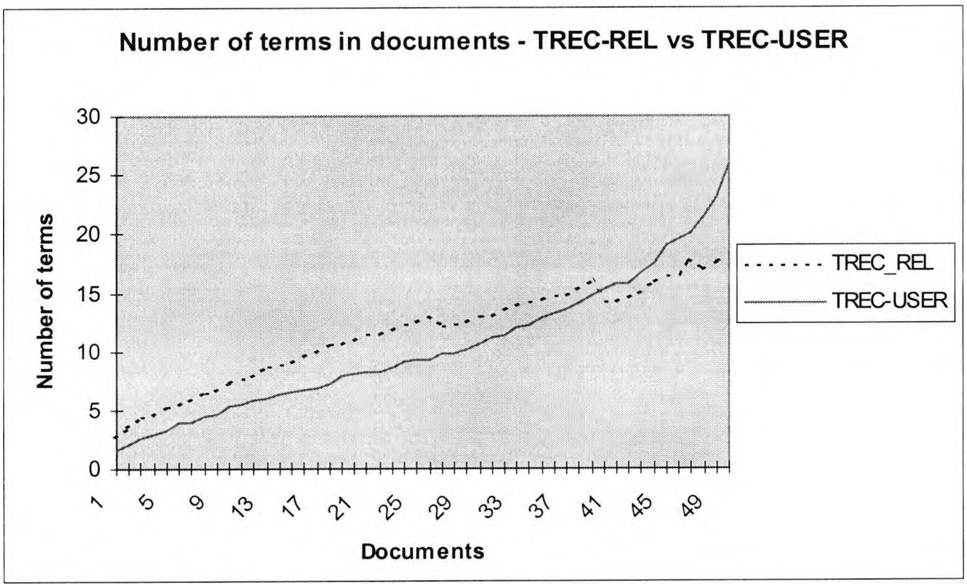
Standard deviations in cluster sizes - (TREC-REL)

	BP-C	BP-NC	FD-C	FD-NC
254	4.92	3.32	4.40	7.27
256	2.08	2.64	6.11	3.78
258	2.50	2.56	3.93	3.98
264	2.75	2.36	0.71	5.76
284	5.18	6.39	3.83	4.74
286	7.13	6.77	4.34	5.51
292	5.22	7.78	4.44	4.72
average	4.26	4.55	3.97	5.11

Standard deviations in cluster sizes - (TREC-USER)

	BP-C	FD-C	BP-NC	FD-NC
254	10.85	11.07	11.22	10.23
256	3.63	12.66	4.27	11.30
258	3.87	8.92	3.53	6.62
264	2.79	12.46	5.99	9.08
284	5.43	7.92	5.57	9.71
286	6.35	5.17	4.51	5.42
292	8.96	6.50	8.76	3.67
average	5.98	9.24	6.26	8.01

Number of representative terms in documents – Averaged over 7 queries



Appendix 2 : Introductory sheets and questionnaires employed in user experiments

Appendix 2.1 : Introduction to the experiment:

INFORMATION RETRIEVAL EXPERIMENT

This study aims to evaluate the efficiency of clustering search output as an output presentation method.

You are expected to participate with a real information need. Your description will be used as a query and will be run on the ENQUIRE system which contains the Inspec database as of 1993.

The top 50 documents will be retrieved and clustered, i.e. grouped for your inspection. Each group will be represented by 10 representative terms and a subset of its documents' titles. Number of documents actually included in the group will also be given.

Based on this information you will be invited to rank the groups according to their expected usefulness to your information need.

You are welcome to make explanatory or descriptive notes for each group on the spaces provided.

You will then be invited to indicate, on a separate sheet, which if any of these 50 documents are relevant to your information need. Again document titles will be provided but you are also welcome to ask for abstracts of any of these documents.

There are brief pre and post-questionnaires for gathering information that could be of use for analysis of results.

Thanks for your cooperation.

Appendix 2.2 : Introduction to the intermediary experiments for evaluating the alternative cluster representations repeating 1st runs queries

INFORMATION RETRIEVAL EXPERIMENT

This study aims to evaluate the efficiency of clustering search output as an output presentation method. You have participated to the study before with a certain information need, and at this stage you are requested to consider that information need and once more make evaluations for clusters and documents for relevance. The documents will generally be same as the ones retrieved before, but the cluster representations will be different.

Again the top 50 documents will be retrieved and clustered, i.e. grouped for your inspection. There will be alternative representations to evaluate for the same output. Two sets will include clusters represented by a subset of documents titles plus representative terms. One set will include clusters that are only represented by titles. Number of documents included in each cluster will also be given.

Based on these representations you will be invited to rank the groups according to their expected usefulness to your information need.

You are welcome to make explanatory or descriptive notes for each group on the spaces provided.

You will then be invited to indicate, on a separate sheet, which if any of these 50 documents are relevant to your information need. Again document titles will be provided but you are also welcome to ask for abstracts of any of these documents.

There is a brief post-questionnaire for gathering information that could be of use for analysis of results.

Thanks for your cooperation.

Appendix 2.3 : Pre-questionnaire:

Information Retrieval Experiment

Pre-questionnaire

Date: ....
Name: ....
E-mail: ....

1. What is the approximate number of information retrieval searches you have done in the past 12 months? (using online systems, CD-ROMs etc).
   ....

2. Please describe your current information need
   ....
   ....
   ....

3. Have you done any previous searches about this need? If yes, did that search lead you to some useful documents?
   ....
   ....

Appendix 2.4 : Post-questionnaire used in the first run of the user experiments

Information Retrieval experiment

Post-questionnaire

Date: ....
Name: ....

1. Has this session provided you with any useful documents that you would like to inspect?
   ....
   ....

   If not, why:
   ....
   ....
   ....

2. Did you feel that the clusters provided a useful representation of the documents in the search output?
   ....
   ....
   ....

3. Have you encountered any problems during the session? or have you observed any weaknesses in the experiment design?
   ....
   ....
   ....

Thanks

Appendix 2.5. : Post-questionnaire used in the intermediary experiments for evaluating the alternative cluster representations

POST-QUESTIONNAIRE

Date: ....

Name: ....

1. Did you find the clusters useful in representing the results of your query?

   ( ) Very useful ( ) Somewhat useful ( ) Not useful at all

Comments: ....

2. Do the clusters represent different aspects/facets of the query or are they not differentiated at all?

   ( ) Clusters represent different aspects/facets of the query.
   ( ) Only some of the clusters represent an identifiable aspect/facet.
   ( ) Clusters are not differentiated from each other.

Comments: ....

3. If you compare the three different cluster representations that are provided for your inspection, which one do you think, gives you a better, more informative representation of your query results?

   ( )A ( ) B ( ) C

Comments: ....

4. Were the representative titles and the representative terms(where available) equally useful in helping you to evaluate the clusters?

   ( ) Yes, they were both useful
   ( ) I based my decisions more on representative terms

( ) I based my decisions more on titles

Comments: ....

Thank you.

Appendix 2.6 : Post-Questionnaire used in the 2nd run of the user experiments

Information Retrieval experiment

Post-questionnaire

Date: ....
Name: ....

1. Has this session provided you with any useful documents that you would like to inspect?
....
....

If not, why:
....
....
....

2. Did you feel that the clusters provided a useful representation of the documents in the search output?
....
....
....

3. If you compare the two different cluster representations that are provided for your inspection, which one do you think, gives you a better, more informative representation of your query results?

( ) Rep-D ( ) Rep-E

Comments: ....

4. Has the way you would describe your information need changed during your inspection of the clusters and documents? And if you were asked to go over the same clusters and documents again now, would you make different evaluations for some of them?

....
....
....

5. Were any of the following a factor in your evaluating the documents for relevance?

Its author(s)	( ) Yes	( ) No
Its publication year	( ) Yes	( ) No
The journal/its availability	( ) Yes	( ) No

6. Have you encountered any problems during the session? or have you observed any weaknesses in the experiment design?
....
....
....

Thanks

Appendix 3.1 : An example of clustered search output - from the 1st group of user experiments

```
/////////////////////////////////////////////////////////

Thu Feb 27 12:25:47 GMT 1997

QUERY= compare applications development programming packages 4gl

OUTPUT= 50 documents

4 clusters have been formed.

/////////////////////////////////////////////////////////

CLUSTER 1  (includes  13 DOCUMENTS)                    RANK(    )
----------------------------------

REPRESENTATIVE TERMS
oriented - object - department - engineer - business
work - case - complex - solution - integrated

REPRESENTATIVE DOCUMENTS

SEED:Doc: 36   Technology (CASE)

Doc: 21   Refining the process of project control

Doc: 48   Software Engineering in Medical Informatics Proceedings
of the IMIA Working Conference, Amsterdam, Netherlands, 8-10
Oct 1990


COMMENTS:


==================================================================

CLUSTER 2  (includes  12 DOCUMENTS)                    RANK(    )
----------------------------------

REPRESENTATIVE TERMS
product - file - benefit - data - 4gl
offer - complex - solution - computer - forthgeneration languag

REPRESENTATIVE DOCUMENTS

SEED:Doc: 13   Moving to open systems with ALLY

Doc: 49   DBase goes Vax VMS

Doc: 12   In the shadow of big brother
(business PC use)


COMMENTS:
```

```
==============================================================

CLUSTER 3  (includes  16 DOCUMENTS)                    RANK(    )
-----------------------------------

REPRESENTATIVE TERMS
designed - compar - time - developed - relational
experienc - network - management - forthgeneration languag - databas

REPRESENTATIVE DOCUMENTS

SEED:Doc: 47   From a 'well designed' database to
AD/Cycle tools: a reengineering experience

Doc: 10   Fourth generation languages and integrated information
systems for small manufacturing companies

Doc: 42   AUTO-QUIZ: an automatic ACCESS statement generator
for QUIZ query


COMMENTS:

==============================================================

CLUSTER 4  (includes  9 DOCUMENTS)                     RANK(    )
-----------------------------------

REPRESENTATIVE TERMS
control - capabiliti - programmer - packag - forthgeneration languag
languag - work - prototyp - computer - softwar

REPRESENTATIVE DOCUMENTS

SEED:Doc: 3   Computer languages increase capabilities, add tools

Doc: 41   Automated software generation approaches for the
design and development of guidance and control systems
software


COMMENTS:

==============================================================
```

Appendix 3.2 : An example of clustered search output - from the 2nd group of user experiments

```
                    - Rep-D  -


////////////////////////////////////////////////////////////

Fri Oct 30 10:52:05 GMT 1998

QUERY TERMS:
inis
database
nuclear
united
nations
vienna
OUTPUT= 50 documents

6 clusters have been formed.

////////////////////////////////////////////////////////////

CLUSTER 1  (includes  9 DOCUMENTS)                  RANK(    )
----------------------------------



REPRESENTATIVE DOCUMENTS

21: Networks in medicine: a heterogeneous networking example.

15: Client server medicine.

18: A maintenance management system for highways &
urban streets.

REPRESENTATIVE TERMS
interfac - importent - necessari - relational - pc - softwar - united - network

COMMENTS:


==============================================================

CLUSTER 2  (includes  10 DOCUMENTS)                 RANK(    )
----------------------------------



REPRESENTATIVE DOCUMENTS

0: Information strategy stands the test of time.

47: UtiliCorp' bright idea (information technology applications).

1: KLASTER-a suit of application programs for bibliographical
databases.
```

REPRESENTATIVE TERMS
member - application - inis - scienc - countri - established - organisation - c$

COMMENTS:

===

CLUSTER 3 (includes 4 DOCUMENTS) RANK( )

REPRESENTATIVE DOCUMENTS

48: Towards continuous operation of hospital systems.

6: How three banks approach satellite video training.

35: Distributed logic programming.

REPRESENTATIVE TERMS
user - languag - hardwar - technical -

COMMENTS:

CLUSTER 4 (includes 10 DOCUMENTS) RANK( )

REPRESENTATIVE DOCUMENTS

43: Online to the nation' library: Kentucky' experience
with the Library of Congress information system.

11: A database of United Kingdon statue law.

45: Information retrieval/RDB.

REPRESENTATIVE TERMS
on line - librari - universiti - gov - type - united state - prototyp - faciliti

COMMENTS:

===

CLUSTER 5 (includes 7 DOCUMENTS) RANK( )

REPRESENTATIVE DOCUMENTS

27: Improving agricultural information services of small island
nations in the Pacific region.

28: NUBASE: a database of nuclear and decay

28: NUBASE: a database of nuclear and decay
properties.

46: Nuclear criticality parameter study data base.

REPRESENTATIVE TERMS
recent - national - utilis - technologi - nation - information - context - lite$

COMMENTS:

==

CLUSTER 6 (includes 13 DOCUMENTS) RANK( )

REPRESENTATIVE DOCUMENTS

41: Nuclear materials database.

16: An image analysis system for automatic measurements
in solid state nuclear track detectors.

17: Description of a general computer program for
RCS computation of complex shapes described by a
CAD system, founded on GTD/UTD.

REPRESENTATIVE TERMS
nuclear - role - power plant - plant - operat - futur - field - plan - build -

COMMENTS:

Appendix 4 : Trec aspect descriptions and assignments to documents

Appendix 4.1 : Aspect descriptions for the 7 topics used in the TREC experiments

254	1	laser with catheter and angioplasty clear clogged coronary artery
254	2	TPA (tissue plasminogen activator) dissolves clots in heart's arteries
254	3	drug (streptokinase) dissolves clots and aspirin prevents blood clotting
254	4	heart shock machines revive heart victims
254	5	fish oil as an agent to keep newly opened arteries (via angioplasty) from collapsing
254	6	use of DNA, genetic markers, to find cause of inherited and cardiovascular dissease
254	7	3-D scanners access damage to heart and blood flow problem areas
254	8	gentically engineered protein
254	9	stainless steel mesh stent inserted into clogged arteries to keep open
256	1	Reduction was occurring due to pressure/desire of elements considering core subjects had too much "western" emphasis
256	2	higher education losing credibility due to reduction
256	3	substitute topics have little meaning outside of academic world
256	4	truth, beauty, excellence, questions of intellectual aesthetic quality are wrongfully considered irrelevant and dismissed
256	5	foreign language needed to compete against rivalry from other countries
256	6	"broadening" the horizons of conventional curriculum is a strategy to cope with educational deficiencies.
256	7	apparently critical but no reason given for criticism
258	1	U.S. military computers (MILNET)
258	2	Air Force Systems Command
258	3	Jet Propulsion Laboratory
258	4	Lawrence Livermore Laboratory
258	5	ARPANET (Pentagon's Defense Advanced Research Projects Agency)
258	6	NASA
258	7	Lawrence Berkeley Laboratory
258	8	Patuxent Naval Air Station
258	9	Defense Department Computer Network
258	10	NATO
258	11	Robins Air Force Base, Georgia
258	12	U. S. Nuclear Defense Agency
258	13	MITRE Corporation
258	14	UNIX Network
258	15	British Telecommunications
258	16	Secretary of Air Force
258	17	National Security Agency (NSA)
258	18	U.S. Research Laboratories
258	19	DEC Computers
258	20	JANET (Joint Academic Network)
258	21	Redstone Army Depot (Anniston, Ala.)
258	22	Central Intelligence Agency (CIA)
258	23	Navy Research Labs
258	24	Army Communications Center (Okinawa)
264	1	American hostages
264	2	encitement
264	3	detained by foreign military
264	4	drug-related arrests
264	5	news reporter arrests
264	6	possession of alcohol
264	7	free-lance journalist
264	8	illegal entry
264	9	customs violations
264	10	fraud and forgery
264	11	terrorist activities
264	12	espionage
264	13	demonstrators
264	14	subversive activities
264	15	smuggling

264	16	vandalism
264	17	religious activities
284	1	U.S. Mexico meet re drugs and law enforcement
284	2	USDEA agents aid and train local forces to target drug traffickers
284	3	U.S. Army troups raid labs in Bolivia
284	4	Far East countries and U.S. work to cut opium production, track drug shipments and dealers
284	5	U.S. provides aircraft and herbicides to spray opium fields in Burma and Pakistan
284	6	U.S.-Italy crack importation of Asian heroin to U.S. via Italy
284	7	U.S.-Soviet cooperation on drugs
284	8	U.S. Special Forces train Bolivian anti-drug unit and give helicopters to aid in interdiction effort
284	9	Honduran and Dominican Republic assist in capture and extradition of drug kingpin to the U.S. for trial in murder of DEA agent in Mexico
284	10	U.S. Attorney General tours Latin America nations to assess anti-drug trafficking effort
284	11	Bahamian government joined U.S. in combined drug raids on Gun Cay, south of Bimini
284	12	U.S.-European offensive on the international narcotics trade and money laundering
284	13	Soviet-Canadian cooperation on drugs from Kabul-Leningrad-Montreal
284	14	U.S.-Latin-American countries (30) cooperate in drug seizure, destruction, etc.
284	15	Latin-American countries cooperate to curb narcotics trafficking, arrests, seizure of drugs
284	16	U.S. civilian piloted gunships with Peruvian narcotics police raid and destroy cocaine labs, warehouses, and airstrips
284	17	U.S. grant to Haiti to purchase drug-detection equipment
284	18	U.S.-Columbia anti-drug programs ineffective
284	19	International cooperation against drug trafficking, abuse, illegal cultivation, etc.
284	20	U.S.-Cuba cooperate in narcotics interdiction
284	21	U.N. to coordinate worldwide fight against drugs
284	22	U.S. gathered drug information used by Columbian authorities as evidence
284	23	Hong Kong joins 26 nation group to combat money laundering
284	24	U.S.-Panama cooperation in ship boarding for drugs and extradition
284	25	U.S.-Europe-U.N. plan private investment to deter cocoa [sic] growing crop in Peru
286	1	heightened demand
286	2	environmental pressures
286	3	rise in European currencies
286	4	high labor, production, raw material costs
286	5	drop in Canadian dollar
286	6	higher transportation costs
286	7	production cuts
286	8	cost cutting, lower interest rates, strong dollar
286	9	supply disruptions, stock rebuilding, fears of strikes
292	1	rehabilitation of homeless children, Manila, Philippines
292	2	Bangladesh - social programs
292	3	Brazil - abandoned children
292	4	Dutch welfare system
292	5	European Community anti-poverty program
292	6	South Africa - Operation Hunger
292	7	South Africa - government welfare
292	8	Sweden - welfare state
292	9	Sri Lanka - poverty elimination program
292	10	Sri Lanka - government social programs
292	11	France - welfare state
292	12	British welfare state
292	13	Uruguay - pensions, government employment
292	14	India - guarenteed jobs, education, health care
292	15	Germany - "conservative" social insurance
292	16	Italy - "conservative" social insurance
292	17	Australia - "liberal" welfare
292	18	New Zealand - "liberal" welfare
292	19	Spain - cradle to grave
292	20	Ireland - universal health care

292	21	Canada - comprehensive social security system
292	22	Denmark - welfare state
292	23	Indonesia - basic education
292	24	South Korea - basic education
292	25	Thailand - basic education
292	26	Taiwan - health, labor, retirement insurance
292	27	Norway - Nordic womb to tomb welfare
292	28	Finland - Nordic womb to tomb welfare
292	29	Hungary - bloated welfare system
292	30	Pakistan - social action program
292	31	China - rudimentary social security system
292	32	Poland - high social security obligations

Appendix 4.2 : Aspect assignments for documents deemed as relevant for the 7 topics used in the Trec experiments

254 AP880310-0095	100000000
254 AP880331-0092	000000000
254 AP880401-0036	100000000
254 AP880420-0098	000000000
254 AP880422-0083	100000000
254 AP880428-0041	010000000
254 AP880812-0033	001000000
254 AP880829-0083	100000000
254 AP880915-0050	000100000
254 AP880922-0063	000010000
254 AP881025-0145	001000000
254 AP881114-0086	100000000
254 AP881114-0216	100000000
254 AP881115-0113	110000000
254 FT923-453	000000000
254 FT934-1290	000001000
254 FT934-15411	001000000
254 FT942-16930	110000000
254 FT944-7280	001000000
254 FT944-8844	001000000
254 WSJ900511-0074	000000100
254 WSJ900511-0111	100000000
254 WSJ900511-0168	001000000
254 WSJ900713-0178	000000010
254 WSJ900829-0020	100000000
254 WSJ900914-0168	000001000
254 WSJ901228-0127	000000100
254 WSJ910103-0037	000000001
254 WSJ910624-0021	000000000
254 WSJ910725-0117	000000100
254 WSJ910822-0065	100000000
254 WSJ910906-0112	100000000
254 WSJ910917-0093	000000000
254 WSJ910919-0084	100000000
254 WSJ911003-0062	000000001
254 WSJ911105-0075	000000000
254 WSJ911111-0136	001000000
254 WSJ920203-0182	100000000
254 WSJ920224-0089	001000000
254 WSJ920225-0057	100000000
254 ZF207-116-376	100000000
254 ZF207-336-317	000000100

256 AP880304-0153	0000000
256 AP880401-0125	1100000
256 AP880418-0188	1100000
256 AP880622-0278	0000000
256 AP880827-0026	1000000
256 AP880912-0024	1011000
256 AP881017-0135	0000100
256 CR93E-3519	0000001
256 FR88812-0046	0000001
256 FR88922-0078	0000000

256 FR940124-1-00108	0000000
256 FR940429-0-00120	0000000
256 FT922-13879	0000000
256 FT931-15789	0000000
256 FT931-5591	0000000
256 FT932-16735	1000000
256 WSJ900420-0124	0000000
256 WSJ900502-0082	0000001
256 WSJ900518-0117	0000000
256 WSJ900725-0106	0000000
256 WSJ900927-0149	0000000
256 WSJ901226-0037	0000000
256 WSJ910122-0119	0000000
256 WSJ910328-0160	1000000
256 WSJ910423-0027	0000000
256 WSJ910423-0084	0000000
256 WSJ910423-0160	0000000
256 WSJ910510-0021	1000000
256 WSJ910628-0197	1001000
256 WSJ910730-0131	0000000
256 WSJ910731-0098	0000010
256 WSJ910820-0119	0000000
256 WSJ910906-0110	0000000
256 WSJ910924-0104	1000000
256 WSJ910924-0110	0000000
256 WSJ911111-0102	0000000
256 WSJ920117-0185	0000001
256 WSJ920323-0128	1000000
256 ZF207-646-886	0000000

258 AP880418-0123	111100100000000000000000
258 AP880418-0168	111000100000000000000000
258 AP880421-0240	000000000000000000000000
258 AP880616-0002	011011110000000000000000
258 AP880617-0066	111011111000000000000000
258 AP880809-0226	000000000110000000000000
258 AP881023-0020	000001000001000000000000
258 AP881201-0014	100010000000100000000000
258 AP881201-0137	100010000000100000000000
258 AP881202-0013	100010000000100000000000
258 AP881210-0108	000100000000000000000000
258 AP881213-0051	000100000000010000000000
258 AP881217-0091	000100000000000000000000
258 FT931-2806	000000000000000000000000
258 FT931-2937	000000000000000000000000
258 FT931-2938	000000000000000000000000
258 FT931-7090	000000000000000000000000
258 FT931-7553	000000000000000000000000
258 FT932-14128	000000000000000000000000
258 FT932-7637	000001000000000000000000
258 FT932-7643	000001000000000000000000
258 FT933-8055	000000000000000000000000
258 FT942-578	000000000000000000000000
258 FT944-15438	000000000000000000000000
258 FT944-7040	000000000000001000000000
258 FT944-7649	000000000000001000000000
258 WSJ900502-0063	000000000000000000000000
258 WSJ900507-0106	000010000000000000000000
258 WSJ900710-0081	000000000000000000000000
258 WSJ900921-0017	000000000000000100000000
258 WSJ910315-0028	100000000000000000000000
258 WSJ910325-0056	000000000000000000000000
258 WSJ910507-0109	000000000000000000000000
258 WSJ910719-0172	000000000000000000000000
258 ZF207-108-660	100000001000000001000000
258 ZF207-240-704	000000000000000000000000
258 ZF207-241-836	000000000000000000000000
258 ZF207-248-587	000000000000000010000000
258 ZF207-249-496	000001000000000000000000
258 ZF207-253-613	000000000000000010000000
258 ZF207-268-525	000000000000000000000000
258 ZF207-326-060	000000000000000000000000
258 ZF207-330-232	000000000000000000100000
258 ZF207-331-212	000000000000000000000000
258 ZF207-331-214	000000000000000000000000

258 ZF207-373-036	000000000000000000000000
258 ZF207-376-279	000001000000000000000000
258 ZF207-496-812	000000000000000000000000
258 ZF207-534-568	000001000000000000010000
258 ZF207-549-986	000000001000000000000000
258 ZF207-555-480	100100000000100000001000
258 ZF207-569-222	100000000000000000000000
258 ZF207-661-666	110001100000100010001111
258 ZF207-668-992	000000000000000000000000
258 ZF207-760-158	000001000000000000000000
258 ZF207-897-498	000000000000000000000000
258 ZF208-220-083	000000000000000000000000
264 AP880214-0084	10000000000000000
264 AP880228-0072	01000000000000000
264 AP880228-0091	01000000000000000
264 AP880318-0207	00100000000000000
264 AP880325-0168	00010000000000000
264 AP880326-0104	00010000000000100
264 AP880328-0004	00001000000000000
264 AP880405-0102	00010000000000000
264 AP880405-0160	00010100000000000
264 AP880406-0015	00010100100000000
264 AP880407-0027	00000000000000000
264 AP880407-0159	00000000000000000
264 AP880407-0216	00010000000000100
264 AP880411-0192	00010000000000000
264 AP880414-0024	10000000000000000
264 AP880505-0074	00000000000000000
264 AP880506-0008	10000000000000000
264 AP880506-0165	10000000000000000
264 AP880523-0165	00000011000000000
264 AP880525-0072	00000001000000000
264 AP880531-0157	00000001000000000
264 AP880623-0097	00100000000000000
264 AP880629-0111	00100000000000000
264 AP880727-0167	00010000000000000
264 AP880815-0071	10000000000000000
264 AP880824-0005	00000000001000000
264 AP880824-0077	00000000001000000
264 AP880825-0004	00000000000100000
264 AP880901-0100	00000000000000000
264 AP880902-0194	00010000000000100
264 AP880905-0080	00010000000000100
264 AP880915-0056	00000000000010000
264 AP880915-0088	00000000000010000
264 AP880916-0199	00000000000010000
264 AP880917-0152	00000000000010000
264 AP880920-0195	00000000000011000
264 AP880923-0128	00000000000000000
264 AP881004-0065	10000000000000000
264 AP881008-0050	10000000000000000
264 AP881010-0032	00000001000000000
264 AP881014-0058	00000000000000100
264 AP881017-0067	00000001000000000
264 AP881020-0204	00000001000000000
264 AP881024-0066	10000000000000000
264 AP881029-0041	00010000000000000
264 AP881108-0044	00010000000000000
264 AP881114-0162	10000000000000000
264 AP881115-0058	00000001000000000
264 AP881116-0103	00000001000000000
264 AP881122-0178	00000000010000000
264 AP881123-0082	00010000000000000
264 AP881125-0162	10000000000000000
264 FR940304-2-00172	00000001000000000
264 FT921-16444	00000001000000000
264 FT922-11368	00000000000000000
264 FT924-13948	00000001000000000
264 FT934-15265	00000000000000010
264 FT941-9246	00000000000000001
264 FT942-14889	00000000000000010
264 WSJ900906-0074	00000000000000000
284 AP880213-0180	1000000000000000000000000

284 AP880222-0072	1000000000000000000000000
284 AP880302-0015	0000000000000000000000000
284 AP880308-0062	0100000000000000000000000
284 AP880310-0205	1000000000000000000000000
284 AP880312-0121	0110000000000000000000000
284 AP880328-0187	0001100000000000000000000
284 AP880331-0165	0000010000000000000000000
284 AP880401-0280	0000001000000000000000000
284 AP880404-0215	0000000100000000000000000
284 AP880406-0185	0000000010000000000000000
284 AP880409-0127	0000000001000000000000000
284 AP880411-0168	0000100000000000000000000
284 AP880415-0060	1000000000000000000000000
284 AP880415-0165	0100000000000000000000000
284 AP880519-0173	0000000010000000000000000
284 AP880519-0201	0000000010000000000000000
284 AP880520-0105	0000000010000000000000000
284 AP880531-0055	0000000000100000000000000
284 AP880604-0095	0000000000000000000000000
284 AP880617-0255	0000000000010000000000000
284 AP880620-0173	0000000000010000000000000
284 AP880625-0016	0100000000000000000000000
284 AP880720-0114	0000000000000000000000000
284 AP880720-0202	0000001000000000000000000
284 AP880721-0023	0001001000000000000000000
284 AP880811-0234	0000000000001000000000000
284 AP880830-0089	0000000000000100000000000
284 AP880830-0166	0000000000000100000000000
284 AP880831-0021	0000000000000100000000000
284 AP880906-0099	0000000000000010000000000
284 AP880927-0103	0000000000000100000000000
284 AP881011-0041	0000000000000001000000000
284 AP881012-0094	0000000000000101000000000
284 AP881020-0232	0000000000000000100000000
284 AP881112-0073	0000000000000000010000000
284 AP881125-0168	0000000000000000001000000
284 AP881125-0175	0100000000000000000000000
284 AP881208-0008	0000000000000000000000000
284 AP881215-0126	0000000000000000000100000
284 AP881220-0086	0000000000000000001000000
284 AP881221-0056	0000000000000000000100000
284 AP881224-0124	0000000000000000000000000
284 FR88922-0050	0000000000000000001000000
284 FT911-2671	0000000000000000000010000
284 FT921-3061	0000000000000000001000000
284 FT921-6988	1000000000000000000000000
284 FT922-9654	0000000000000000000010000
284 FT924-10092	0000000000000000000010000
284 FT931-9535	0000000000000000000010000
284 FT932-3130	0000000000000000000000000
284 FT934-10491	0000000000000000010000000
284 FT942-10977	0100000000000000000000000
284 WSJ900409-0138	0100000000000000000000000
284 WSJ910529-0062	0000000000010000000000000
284 WSJ910628-0153	0000000000000000000001000
284 WSJ920113-0038	0000000000000000000000100
284 WSJ920204-0078	0000000000000000000000010
284 WSJ920214-0116	0000000000000000000000001
286 AP880307-0242	100000000
286 AP880606-0274	000000000
286 AP881031-0020	000000000
286 AP881128-0213	000000000
286 AP881212-0082	000000000
286 FT911-2295	110000000
286 FT911-2476	000000000
286 FT921-12618	001000000
286 FT921-14279	010000000
286 FT921-16600	010100000
286 FT922-8911	010000000
286 FT923-13050	000000000
286 FT923-6751	000000000
286 FT924-2255	100000000
286 FT924-2257	001100000
286 FT924-2258	100110000

286 FT924-2259	100001000
286 FT924-2260	000100000
286 FT924-2261	000100000
286 FT924-2262	000101000
286 FT931-3233	010010000
286 FT931-7614	000000000
286 FT931-8311	000000000
286 FT932-10615	000000000
286 FT932-10617	000100000
286 FT932-10624	000100000
286 FT932-54	010000000
286 FT933-10479	000000000
286 FT933-11622	000000100
286 FT934-1021	000000010
286 FT934-15536	000000000
286 FT934-15540	000000000
286 FT934-15914	000000000
286 FT941-13510	010100000
286 FT942-11471	000000011
286 FT942-11747	100000010
286 FT942-15324	000000000
286 FT942-6065	000000000
286 FT942-6252	100000000
286 FT942-8846	100000000
286 FT942-9111	000000000
286 FT942-9114	100000000
286 FT942-9116	000000000
286 FT942-9117	010000000
286 FT942-9121	010100000
286 FT943-10931	000110000
286 FT943-11401	000100000
286 FT943-14367	100100000
286 FT943-14907	110100000
286 FT943-16095	000000000
286 FT943-16097	100100001
286 FT943-3587	000000000
286 FT943-4260	000100000
286 FT943-5017	100000000
286 FT943-828	100100000
286 FT944-10659	100100000
286 FT944-11489	000000000
286 FT944-15304	000100000
286 FT944-15305	100000000
286 FT944-1604	100100000
286 FT944-16358	100100000
286 FT944-17523	110100001
286 FT944-3424	000100000
286 FT944-6872	000100000
286 FT944-7454	000100000
286 FT944-8665	000100000
286 FT944-9640	000100000
286 WSJ910716-0039	000000000
286 WSJ910916-0088	010000000
286 WSJ911018-0203	000000000
286 WSJ920124-0029	000100000
286 WSJ920318-0047	000000100
292 AP880215-0207	10000000000000000000000000000000
292 AP880412-0086	00000000000100000000000000000000
292 AP880501-0053	00000000000000000000000000000000
292 AP880519-0263	00100000000000000000000000000000
292 AP880621-0144	00000000000000000000000000000000
292 AP880803-0238	00010000000000000000000000000000
292 AP880929-0181	00000000000000000000000000000000
292 AP880930-0143	00000000000000000000000000000000
292 AP881014-0025	00000000000000000000000000000000
292 CR93H-3525	00000000000000000000000000000000
292 CR93H-9843	00000000000000000000000000000000
292 FT922-12153	00001000000000000000000000000000
292 FT922-4425	00000110000000000000000000000000
292 FT923-1453	00000001000000000000000000000000
292 FT923-1495	00000001000000000000000000000000
292 FT924-10895	00000001000000000000000000000000
292 FT924-11314	00000000110000000000000000000000
292 FT931-10916	00000000000100000000000000000000

292 FT931-737	00000000001000000000000000000000
292 FT932-10159	00000000001000000000000000000000
292 FT932-10165	00000000001000000000000000000000
292 FT932-2023	00000000000000000000000000000000
292 FT932-3491	00000000000100000000000000000000
292 FT932-4859	00000000000010000000000000000000
292 FT932-4864	00000000000010000000000000000000
292 FT932-7223	00000000000100000000000000000000
292 FT933-1192	00000000000001000000000000000000
292 FT933-13307	00000000000000000000000000000000
292 FT933-5512	00000000000000000000000000000000
292 FT934-10072	00000001000000000000000000000000
292 FT934-1035	00000001000000000000000000000000
292 FT934-10921	00000001001100111000000000000000
292 FT934-11014	00000000000000000000000000000000
292 FT934-12800	00010000000100110011000000000000
292 FT934-2265	00000000000000000000100000000000
292 FT934-2293	00000000000000000000000000000000
292 FT934-5692	00000000000000000000010000000000
292 FT934-5693	00000000000000000000010000000000
292 FT934-5921	00000000000100000000000000000000
292 FT934-6695	00000000001000000000000000000000
292 FT934-8227	00000000000000000000000000000000
292 FT941-11779	00000000000000000000100000000000
292 FT941-12426	00000000000001000000001110000000
292 FT941-8918	00000000000100000000000000000000
292 FT942-10229	00000000000000000010000000000000
292 FT942-10729	00000000000000000000000000000000
292 FT942-3324	00000000000000000000000000000000
292 FT942-3413	00000000000000000000000000000000
292 FT942-5248	00000001000000100000100000000000
292 FT943-13402	00000000000000000000000001000000
292 FT943-14531	00000000000000000000000000000000
292 FT943-1684	00000001000000000000010000110000
292 FT943-2556	00000000000000000000000000000000
292 FT943-3430	00000001000000000000010000000000
292 FT943-7	00000000000000001100000000000000
292 FT943-8923	00000000000000000000000000000000
292 FT944-10159	00000000000000000000000000001000
292 FT944-10925	00000000000001000000000000000000
292 FT944-15385	00000000000000100000000000000000
292 FT944-18177	00000000000000000000100000000000
292 FT944-642	00000000000000000000000000000000
292 FT944-6725	00000000000000000000000000000100
292 FT944-9190	00000001000000000000010000110000
292 FT944-9682	00000000000000000000000000000010
292 WSJ900413-0126	00000000000000000000000000000000
292 WSJ910104-0103	00000000000000000000000000000000
292 WSJ910612-0115	00000000000000000000000000000000
292 WSJ910808-0068	00000000000000000100000000000000
292 WSJ911030-0168	00000000000000000000000000000000
292 WSJ911106-0085	00000001000000000000000000000000
292 WSJ920324-0078	00000000000000000000000000000001

Appendix 4.3: TREC queries: narratives provided by TREC and the query terms used to retrieve the relevant documents

Query 254

Narrative:

A relevant document will report/discuss those cases in which persons diagnosed with heart ailments were treated with medications and/or techniques such as angioplasty, stents, lasers, arthrectomy (roto router) etc., in place of surgery. Also advantages of non-invasive procedures over surgery and comparative studies which show any disparity over longevity when either procedure is used.

Query terms used:

non-invasive - heart - ailment - attack - medication - stent - arthrectomy - laser - longevity – angioplasty - disease - heart attack - heart disease

Query 256

Narrative:

To be relevant, a document will provide negative opinion/facts concerning the fact that colleges have reduced their basic requirements for the granting of degrees to undergraduates.

Query terms used:

college – undergraduate - granting - degree - criticism - history - philosophy - science - literature - liberal arts* - liberal* - western* - nonwestern* - credibility - pressure - multicultural* - multiethnic* - ethnic* - black* - grant degree

(*) terms derived from facet explanations provided by TREC experts

Query 258

Narrative:

Illegal entry into sensitive computer networks is a serious and potentially menacing problem. Both ‘hackers’ and foreign agents have been known to acquire unauthorized entry into various networks. Items relevant to this subject would include but not be limited to instances of illegally entering networks containing information of a sensitive nature to specific countries, such as defence and technology information, international banking, etc. Items of a personal nature (e.g. credit card fraud, changing of college test scores) should not be considered relevant.

Query terms used:

illegal entry - illegal - entry - computer - network - computer networks - non-authorized - unauthorized - hackers - agents - information -defense - technology -

Query 264

Narrative:

With international travel becoming more commonplace, many travellers, whether due to ignorance of the law or with criminal intent, are ending up in foreign jails. Of interest is the country involved, the reason for the seizure, and jailing, and the sentence, if convicted. U.S. intervention, if any, would also be of interest.

Query terms used:

foreign - jail - prosecute* - citizen - american - convict - arrest* - detain* - prison* - hostage* - captive* - kidnap* - protester* - free* -traffick* - release* - terrorist* - embassy* - vandalism - smuggle

(*) terms derived from facet explanations provided by TREC experts and the unretrieved relevant documents

Query 284

Narrative:

Drugs are a critical problem being experienced throughout the world. International co-operation will be required if this problem is to be alleviated. Examples of relevant documents would be those related to shared information regarding shipments across international borders as well as sharing of information with various prosecutors and other law enforcement personnel.

Query terms used:

drug - cooperation - enforce - border - worldwide - shipment - prosecute - police - opium* - heroin* - train* - raid - narcotic - cultivation* - aid* - local - share information - international cooperation - drug trafficking* - drug problem

(*) terms derived from facet explanations provided by TREC experts

Query 286

Narrative:

In the last year or so the publishing industry is said to have faced a 40% increase in the cost of paper. What factors have led to this price rise? Is there a shortage of materials? Has the cost of processing risen? Have higher taxes been imposed?

Query terms used:

paper - cost - increase - shortage - demand* - environmental* - currency* - labor* - production* - european* - forest* - recycle* - tree* - capacity* - pulp* - package* - company* - wood* - paper industry - price rise

(*) terms derived from facet explanations provided by TREC experts and unretrieved relevant documents

Query 292

Narrative:

To be relevant a document would identify a welfare program in a foreign country and explain how it works to aid citizens with little or no income. It would include those who can't work because of a disability and people who have extra burden of small children. The document should indicate how these people are supported or not supported. A relevant document should identify the source of the monies used to support such welfare programs.

Query terms used:

social - program - poor - citizen - disabled - children - education* - welfare program - welfare - support - fund - poverty* - government - welfare state* - health care* - insurance* - social security* - pension* - employment* - unemployment*

(*) terms derived from facet explanations provided by TREC experts and unretrieved relevant documents

Appendix 4.4: Concentration of aspects in clusters

Q254 - Concentration of aspects in clusters

Aspect Number	Number of documents assigned the aspect	Number of retrieved documents assigned the aspect	Highest number and proportion of documents assigned to the aspect in one cluster	
1	16	16	12/16	75%
2	3	3	2/3	66%
3	8	8	6/8	75%
6	2	2	2/2	100%
7	4	4	2/4	50%
9	2	2	2/2	100%

Q256 - Concentration of aspects in clusters

Aspect Number	Number of documents assigned the aspect	Number of retrieved documents assigned the aspect	Highest number and proportion of documents assigned to the aspect in one cluster	
1	10	10	8/10	80%
2	2	2	2/2	100%
4	2	2	1/2	50%
7	4	4	3/4	75%

Q258 - Concentration of aspects in clusters

Aspect Number	Number of documents assigned the aspect	Number of retrieved documents assigned the aspect	Highest number and proportion of documents assigned to the aspect in one cluster	
1	11	11	6/11	55%
2	5	5	4/5	80%
3	4	4	4/4	100%
4	5	5	4/5	80%
5	6	6	3/6	50%
6	10	9	4/9	44%
7	5	5	4/5	80%
8	2	2	2/2	100%
9	3	3	2/3	66%
15	2	2	2/2	100%
17	3	3	1/3	33%

Q264 - Concentration of aspects in clusters

Aspect Number	Number of documents assigned the aspect	Number of retrieved documents assigned the aspect	Highest number and proportion of documents assigned to the aspect in one cluster	
1	10	10	8/10	80%
3	3	2	2/2	100%
4	13	8	8/8	100%
8	13	4	3/4	75%
11	2	2	2/2	100%
13	5	4	3/4	75%
15	5	2	2/2	100%

Q284 - Concentration of aspects in clusters

Aspect Number	Number of documents assigned the aspect	Number of retrieved doc.s assigned the aspect	Highest number and proportion of documents assigned to the aspect in one cluster	
1	5	4	3/4	75%
2	7	6	2/6	33%
4	2	2	2/2	100%
7	3	3	3/3	100%
9	4	4	4/4	100%
12	3	3	1/3	33%
14	5	5	3/5	60%
16	2	2	1/2	50%
18	2	2	2/2	100%
19	4	4	3/4	75%
20	2	2	2/2	100%
21	4	4	4/4	100%

Q286 - Concentration of documents in clusters

Aspect Number	Number of documents assigned the aspect	Number of retrieved doc.s assigned the aspect	Highest number and proportion of documents assigned to the aspect in one cluster	
1	19	18	8/18	44%
2	12	10	9/10	90%
3	2	2	1/2	50%
4	28	26	11/26	42%
5	3	3	3/3	100%
6	2	2	2/2	100%
7	2	2	2/2	100%
8	3	3	3/3	100%
9	3	3	2/3	66%

Q292 - Concentration of aspects in clusters

Aspect Number	Number of documents assigned the aspect	Number of retrieved documents assigned the aspect	Highest number and proportion of documents assigned to the aspect in one cluster	
8	11	8	7/8	88%
11	5	5	2/5	40%
12	8	7	6/7	86%
13	2	2	2/2	100%
15	4	3	2/3	66%
18	3	2	1/2	50%
21	4	3	2/3	66%
22	5	3	2/3	66%

Appendix 5.1 : Aspect Distribution and document similarity tables for TREC-USER experiments

Q254 - Unretrieved relevant documents' similarity to clusters from the user query

	Cl1	Cl2	Cl3	Cl4
Asp 1	0	0	0	4
Asp 3	2	0	0	3
Asp 4	1	0	0	0
Asp 6	2	0	0	0
Asp 7	2	0	0	1
Asp 8	1	0	0	0
Number of documents	8	0	0	8

Q254 - Distribution of aspects among clusters

	Cl1	Cl2	Cl3	Cl4
Asp 1	6	3	0	2
Asp 2	1	0	0	2
Asp 3	0	0	0	2
Asp 5	1	0	0	0
Asp 7	1	0	0	0
Asp 9	2	0	0	0
Cluster size	29	13	5	4
Documents with aspect assignments	11	3	0	4

Q258 - Distribution of aspects among clusters

	Cl1	Cl2	Cl3	Cl4
Asp 1	4	0	1	0
Asp 2	3	0	0	0
Asp 3	3	0	0	0
Asp 4	0	1	0	0
Asp 5	4	0	0	1
Asp 6	3	0	1	0
Asp 7	3	0	0	0
Asp 8	2	0	0	0
Asp 9	1	0	1	0
Asp 13	2	0	0	0
Asp 17	0	0	0	1
Cluster size	16	9	16	7
Documents with aspect assignments	6	1	2	2

Q258 - Unretrieved relevant documents' similarity to clusters from the user query

	Cl1	Cl2	Cl3	Cl4
Asp 1	3	0	1	0
Asp 2	1	0	0	0
Asp 3	1	0	0	0
Asp 4	1	2	0	0
Asp 5	1	0	0	0
Asp 6	2	0	0	2
Asp 7	1	0	0	0
Asp 9	0	0	1	0
Asp 13	1	0	0	0
Asp 15	0	0	2	0
Asp 17	0	0	0	1
Number of documents	5	2	4	3

Q286 - Distribution of aspects among clusters

	Cl1	Cl2	Cl3	Cl4
Asp 1	3	2	1	1
Asp 2	2	0	1	0
Asp 3	1	0	0	0
Asp 4	8	2	2	1
Asp 7	1	0	0	0
Asp 8	0	2	0	0
Asp 9	1	1	1	0
Cluster size	21	9	12	7
Documents with aspect assignments	10	4	2	1

Q286 - Unretrieved relevant documents' similarity to clusters from the user query

	Cl1	Cl2	Cl3	Cl4
Asp 1	7	4	1	0
Asp 2	6	2	0	0
Asp 3	1	0	0	0
Asp 4	8	6	1	0
Asp 5	3	0	0	0
Asp 6	2	0	0	0
Asp 7	0	1	0	0
Asp 8	1	0	0	0
Total	19	10	2	0

Q292 - Distribution of aspects among clusters

	Cl1	Cl2	Cl3	Cl4
Asp 12	0	1	0	0
Asp 21	0	0	1	0
Cluster size	20	8	12	5
Documents with aspect assignments	0	1	1	0

Q292 - Unretrieved relevant documents' similarity to clusters from the user query

	Cl1	Cl2	Cl3	Cl4
Asp 4	0	0	1	0
Asp 8	0	0	7	3
Asp 11	0	0	5	0
Asp 12	0	0	6	0
Asp 13	0	0	1	1
Asp 14	0	0	3	0
Asp 15	0	0	3	0
Asp 16	0	0	1	0
Asp 17	0	0	2	0
Asp 18	0	0	3	0
Asp 19	0	0	1	0
Asp 21	0	0	2	1
Asp 22	0	0	3	1
Asp 27	0	0	1	1
Asp 28	0	0	1	1
Total	0	0	28	5

Appendix 5.2

Overlap between documents retrieved by the users vs documents retrieved by the current version of the search engine (TREC-USER)

	number of common documents	overlap rate
254	32	64%
256	44	88%
258	23	46%
264	18	36%
284	23	46%
286	29	58%
292	13	26%
Average	26	52%

Number of user documents retrieved (TREC-USER) through checking the top 1000 documents retrieved

	Number(ratio) of documents
254	50/50
256	50/50
258	48/50
264	47/50
284	48/50
286	49/50
292	44/50
Average	48/50

Percantage of relevant documents in the output set (TREC-USER)

	Relevant number	percentage	aspectual number	percentage
254	18	36%	16	32%
256	3	6%	1	2%
258	15	31%	11	23%
264	6	13%	5	11%
284	14	29%	9	19%
286	18	37%	17	35%
292	3	7%	2	5%

Number of relevant documents with aspect assignments (TREC-USER)

Query	Number of relevant documents	Number of documents with aspect assignments
254	19	18
256	3	1
258	15	11
264	6	5
284	14	9
286	18	17
292	3	2

Proportion of aspects represented by less than three documents (TREC-USER)

	ratio	percentage
254	4/6	66%
256	1/1	100%
258	5/11	45%
264	3/4	75%
284	8/8	100%
286	3/7	43%
292	2/2	100%

Appendix 6 : Query details and results of expansion for the TREC-EXP experiments

Appendix 6.1 : Query term sets - Original user queries and working queries after expansion on each cluster

Topic 254

	user	query expansion on:			
	query	CL1	CL2	CL3	CL4
Number of documents marked relevant for query expansion		29	13	4	4
Number of relevant documents marked relevant for query expansion		12	3	0	4
Number of relevant documents in the top 400 hitlist	27	20	25	32	31
Query terms					
heart ailments	+	+	+	+	+
angioplasty	+	+	+	+	+
stents	+	+	+	+	+
arthrectomy	+	+	+	+	+
non-invasive surgery	+	+	+	+	+
heart medication	+	+	+	+	+
heart		+	+	+	+
coronary		+		+	+
surgery			+		+
balloon			+		+
arteries		+			
catheter, Bard C.R., arterial			+++		
undergoing, mother, ailment				+++	
hospitals				+	
patients, lancet, TPA, treatment					++++
clot, procedure, small, widens					++++
medicine					+

Topic 256

	user	query expansion on:			
	query	CL1	CL2	CL3	CL4
Number of documents marked relevant for query expansion		11	25	6	8
Number of relevant documents marked relevant for query expansion		1	2	0	0
Number of relevant documents in the top 400 hitlist	8	1	n/a[1]	0	8
Query terms					
undergraduate	+	+	+	+	+
degree	+	+	+	+	+
reduced	+	+	+	+	+
requirements	+	+	+	+	+
standards	+	+	+	+	+
criticism	+	+	+	+	+
negative	+	+	+	+	+
baccalaureate			+		+
academic			+		+
graduate			+		+

[1] Relevance expansion on CL2 caused segmentation fault, hence it could not be reported in this table.

certainty, stochasticity, resulted,		+++			
regulatory, continuum, considered,		+++			
addition, duration, NPEs, 7(a)(2		++++			
students, enrolled, educational,			+++		
programs, institution, eligible, HEA			++++		
PACM, TSI, paragraph, inspection,				++++	
maintenance, methods, control,				+++	
ACM, enclosure, glove, OSHA, 9)(4				++++	
study, course, fellowship, master's					++++
2400.43(b, 2400.53, 2400.54					+++
2400.55, 2400.56, 2400.60					+++

Topic 258

	user	query expansion on:			
	query	CL1	CL2	CL3	CL4
Number of documents marked relevant for query expansion		15	8	14	7
Number of relevant documents marked relevant for query expansion		7	1	4	3
Number of relevant documents in the top 400 hitlist	49	43	48	52	43
Query terms					
computer crime	+	+	+	+	+
illegal entry	+	+	+	+	+
hackers	+	+	+	+	+
spying	+	+	+	+	+
sensitive information	+	+	+	+	+
computer networks	+	+	+	+	+
unauthorised entry	+	+	+	+	+
computer		+	+	+	+
Arpanet		+			
crimes		+	+	+	+
security		+			+
virus		+	+		+
network		+			+
1989			+		+
penetrated, passwords, detectable		+++			
prosecuting, perpetrators, v23		+++			
wrist, bequai, slap, fulltext			++++		
v6, copyright, unauthorised			+++		
ziff, system			++		
hacking, electronic				++	
internet, government, law, jail					++++

Topic 264:

	user	query expansion on:			
	query	CL1	CL2	CL3	CL4
Number of documents marked relevant for query expansion		12	12	7	8
Number of relevant documents marked relevant for query expansion		1	0	0	2
Number of relevant documents in the top 400 hitlist	6	7	5	3	5
Query terms					
U.S. citizens	+	+	+	+	+
Americans	+	+	+	+	+
foreign jails	+	+	+	+	+
gaols	+	+	+	+	+
prisons	+	+	+	+	+
foreign	+	+	+	+	+
overseas	+	+	+	+	+
international	+	+	+	+	+
criminal	+	+	+	+	+
conviction	+	+	+	+	+
sentenced	+	+	+	+	+
broken the law	+	+	+	+	+
drugs	+	+	+	+	+
murder	+	+	+	+	+
crime		+		+	+
year		+	+		
life		+	+		
terms, judge, prosecution		+++			
paroled, marijuana, slaying,			+++		
smuggled			+		
arrest, lenient, incarcerated					+++
real, justice					++

Topic 284:

	user	query expansion on:			
	query	CL1	CL2	CL3	CL4
Number of documents marked relevant for query expansion		21	10	6	10
Number of relevant documents marked relevant for query expansion		7	2	2	3
Number of relevant documents in the top 400 hitlist	37	38	26	35	41
Query terms					
narcotics	+	+	+	+	+
drugs	+	+	+	+	+
prosecution	+	+	+	+	+
international cooperation	+	+	+	+	+
international borders	+	+	+	+	+
drug shipments	+	+	+	+	+
law officers	+	+	+	+	+
police officers	+	+	+	+	+
cocaine			+		+
traffickers, enforcement		++			
airstrip, medellin, lehder			+++		
indicted, miami, cartel			+++		
trafficking, seized, seizures,					+++
south american, shipments					++
drug lords					+

Topic 286:

	user query	query expansion on:			
		CL1	CL2	CL3	CL4
Number of documents marked relevant for query expansion		21	9	12	7
Number of relevant documents marked relevant for query expansion		10	4	2	1
Number of relevant documents in the top 400 hitlist	32	38	36	50	37
Query terms					
cost	+	+	+	+	+
cost of paper	+	+	+	+	+
cost of processing	+	+	+	+	+
cost of raw materials	+	+	+	+	+
expense	+	+	+	+	+
factors contributing to	+	+	+	+	+
higher taxes	+	+	+	+	+
increase	+	+	+	+	+
increased taxes	+	+	+	+	+
increased wages	+	+	+	+	+
inflation	+	+	+	+	+
labour costs	+	+	+	+	+
paper	+	+	+	+	+
paper cost	+	+	+	+	+
paper pulp	+	+	+	+	+
price increase	+	+	+	+	+
price rise	+	+	+	+	+
rising	+	+	+	+	+
shortage of materials	+	+	+	+	+
pulp		+	+	+	+

Topic 292:

	user query	query expansion on:			
		CL1	CL2	CL3	CL4
Number of documents marked relevant for query expansion		17	8	11	4
Number of relevant documents marked relevant for query expansion		0	1	1	0
Number of relevant documents in the top 400 hitlist	7	0	24	14[2]	24
Query terms					
social program	+	+	+	+	+
welfare program	+	+	+	+	+
social security	+	+	+	+	+
poor	+	+	+	+	+
needy	+	+	+	+	+
disabled	+	+	+	+	+
europe	+	+	+	+	+
africa	+	+	+	+	+
asia	+	+	+	+	+
far east	+	+	+	+	+
middle east	+	+	+	+	+
welfare			+	+	+
east, middle		++			
medicare, SSI			++		
recipients, medicaid, program				+++	
assistance, administer, AFDC				+++	
children, payments				++	
bush, bureaucracy, taxes					+++
elderly, freeze, formula					+++
zones, preserve					++

[2] Due to a segmentation fault, the full hitlist could not be retrieved after query expansion on Cl3. 270 documents were retrieved instead of 400, hence the number of relevant documents were probably higher than the figure displayed here.

Appendix 6.2 : Overall effect of query expansion based on each cluster

Topic 254:

	query expansion on:			
	CL1	CL2	CL3	CL4
Ascends:				
31-100 to top 30	2	1	4	5
101-400 to top 30	1	-	2	-
101-400 to top 100	3	5	1	3
Total	*6*	*6*	*7*	*8*
Descends:				
Top 30 to 31-100	2	9	4	1
Top 30 to 101-400	-	3	3	-
Top 100 to 101-400	-	1	-	-
Total	*2*	*13*	*7*	*1*
Ascend/descend difference	4	-7	0	7
New entries:				
New entry to top 30	-	-	-	1
New entry to 31-100	2	-	1	3
New entry 101-400	3	3	4	4
Total	*5*	*3*	*5*	*8*
Drop-outs:				
Drop from 101-400	-	2	-	-
Total	0	2	0	0
Entry/drop difference	*5*	*1*	*5*	*8*
Remain within top 30	6	-	8	10
Remain within top 100	-	3	1	-
Remain within 101-400	1		4	4
Total	7	3	13	14
Total gain/loss	*9*	*-6*	*5*	*15*

Topic 256:

	query expansion on[3]:		
	CL1	CL3	CL4
Descends:			
Top 30 to 101-400	-	-	1
Top 100 to 101-400	1	-	1
Total	1	0	2
Ascend/descend difference	*-1*	*0*	*-2*
New entries:			
New entry to 31-100	-	-	1
New entry 101-400	-	-	3
Total	*0*	*0*	*4*
Drop-outs:			
Drop from top 30	-	1	-
Drop from 31-100	1	3	1
Drop from 101-400	4	4	3
Total	*5*	*8*	*4*
Entry/drop difference	*-5*	*-8*	*0*
Remain within top 100	-	-	1
Remain within 101-400	-	-	1
Total	*0*	*0*	*2*
Total gain/loss	*-6*	*-8*	*-2*

[3] Relevance expansion on CL2 caused segmentation fault, hence it could not be reported in this table.

Topic 258:

	query expansion on:			
	CL1	CL2	CL3	CL4
Ascends:				
31-100 to top 30	3	1	3	2
101-400 to top 30	1	2	3	1
101-400 to top 100	2	-	5	1
Total	*6*	*3*	*11*	*4*
Descends:				
Top 30 to 31-100	5	4	2	3
Top 30 to 101-400	4	4	6	5
Top 100 to 101-400	12	10	11	16
Total	*21*	*18*	*19*	*24*
Ascend/descend difference	*-15*	*-15*	*-8*	*-20*
New entries:				
New entry to top 30	1	1	1	1
New entry to 31-100	2	1	3	1
New entry 101-400	3	4	5	2
Total	*6*	*6*	*9*	*4*
Drop-outs:				
Drop from 31-100	4	4	1	2
Drop from 101-400	1	2	1	5
Total	*5*	*6*	*2*	*7*
Entry/drop difference	*1*	*0*	*7*	*-3*
Remain within top 30	1	1	1	1
Remain within top 100	3	10	7	4
Remain within 101-400	6	10	5	6
Total	*10*	*21*	*13*	*11*
Total gain/loss	*-14*	*-15*	*-1*	*-23*

Topic 264:

	query expansion on:			
	CL1	CL2	CL3	CL4
Ascends:				
101-400 to top 30	-	-	-	1
Total	*0*	*0*	*0*	*1*
Descends:				
Top 30 to 31-100	1	1	1	-
Top 30 to 101-400	1	1	1	-
Total	*2*	*2*	*2*	*0*
Ascend/descend difference	*-2*	*-2*	*-2*	*1*
New entries:				
New entry to 31-100	1	-	-	-
New entry 101-400	2	2	-	3
Total	*3*	*2*	*0*	*3*
Drop-outs:				
Drop from 101-400	1	3	3	2
Total	*1*	*3*	*3*	*2*
Entry/drop difference	*2*	*-1*	*-3*	*1*
Remain within 101-400	2	1	1	1
Total	*2*	*1*	*1*	*1*
Total gain/loss	*0*	*-3*	*-5*	*2*

Topic 284:

	query expansion on:			
	CL1	CL2	CL3	CL4
Ascends:				
31-100 to top 30	1	-	2	2
101-400 to top 30	9	3	6	1
101-400 to top 100	2	1	4	8
Total	*12*	*4*	*12*	*11*
Descends:				
Top 30 to 101-400	-	1	-	-
Top 100 to 101-400	1	1	2	1
Total	*1*	*2*	*2*	*1*
Ascend/descend difference	*11*	*2*	*10*	*10*
New entries:				
New entry to top 30	2	-	-	2
New entry to 31-100	1	3	-	4
New entry 101-400	7	5	3	9
Total	*10*	*8*	*3*	*15*
Drop-outs:				
Drop from 31-100	-	3	-	1
Drop from 101-400	2	14	3	7
Total	*2*	*17*	*3*	*8*
Entry/drop difference	*8*	*-9*	*0*	*7*
Remain within top 30	1	-	-	1
Remain within top 100	3	-	-	-
Remain within 101-400	11	12	18	13
Total	*15*	*12*	*18*	*14*
Total gain/loss	*19*	*-7*	*10*	*17*

Topic 286:

	query expansion on:			
	CL1	CL2	CL3	CL4
Ascends:				
31-100 to top 30	2	4	1	3
101-400 to top 30	8	8	-	4
101-400 to top 100	1	4	9	8
Total	*11*	*16*	*10*	*15*
Descends:				
Top 30 to 31-100	-	1	-	4
Top 100 to 101-400	1	1	1	1
Total	*1*	*2*	*1*	*5*
Ascend/descend difference	*10*	*14*	*9*	*10*
New entries:				
New entry to top 30	2	-	4	1
New entry to 31-100	7	3	5	2
New entry 101-400	9	7	14	5
Total	*18*	*10*	*23*	*8*
Drop-outs:				
Drop from top 30	4	1	-	-
Drop from 31-100	1	1	2	1
Drop from 101-400	1	-	1	1
Total	*6*	*2*	*3*	*2*
Entry/drop difference	*12*	*8*	*20*	*6*
Remain within top 30	6	6	7	-
Remain within top 100	-	-	5	-
Remain within 101-400	2	2	4	-
Total	*8*	*8*	*16*	*0*
Total gain/loss	*22*	*22*	*29*	*16*

Topic 292:

	query expansion on:			
	CL1	CL2	CL3[4]	CL4
Ascends:				
31-100 to top 30	-	-	1	2
101-400 to top 30	-	-	-	1
101-400 to top 100	-	4	2	2
Total	*0*	*4*	*3*	*5*
New entries:				
New entry to top 30	-	-	1	-
New entry to 31-100	-	-	5	2
New entry 101-400	-	19	2	15
Total	*0*	*19*	*8*	*17*
Drop-outs:				
Drop from 31-100	2	1	-	-
Drop from 101-400	5	-	1	-
Total	*7*	*1*	*1*	*0*
Entry/drop difference	*-7*	*18*	*7*	*17*
Remain within top 100	-	-	1	-
Remain within 101-400	-	1	1	2
Total	*0*	*1*	*2*	*2*
Total gain/loss	*-7*	*22*	*10*	*22*

[4] Due to a segmentation fault, the full hitlist could not be retrieved after query expansion on Cl3. 270 documents were retrieved instead of 400, hence the number of relevant documents were probably higher than the figure displayed here.

Appendix 6.3 : Effect of query expansion on documents with the same aspects as those present in cluster used for expansion

Topic 254 - Cluster 1

	Aspects represented in cluster[5]	
	Asp1	Asp7
Ascends:		
31-100 to top 30	1	-
101-400 to top 100	1	2
Total	*2*	*2*
Descends:		
Top 30 to 31-100	2	-
Total	*2*	*0*
Ascend/descend difference	*0*	*2*
New entries:		
New entry to 31-100	1	-
Total	*1*	*0*
Entry/drop difference	*1*	*0*
Total gain/loss	*1*	*2*

Topic 254 – Cluster 2

	Aspects represented in cluster
	Asp1
Ascends:	
31-100 to top 30	1
101-400 to top 100	2
Total	3
Descends:	
Top 30 to 31-100	6
Top 30 to 101-400	1
Total	7
Ascend/descend difference	-4
New entries:	
New entry 101-400	1
Total	1
Entry/drop difference	1
Total gain/loss	-3

[5] Aspects 2, 5 and 9 were also represented in this cluster, but were not present in the hitlists(or were present in the same ranges in both hitlists) apart from the documents that were marked relevant for query expansion.

Topic 254 - Cluster 4

	Aspects represented in cluster		
	Asp1	Asp2	Asp3
Ascends:			
31-100 to top 30	2	1	1
101-400 to top 100	1	-	-
Total	*3*	*1*	*1*
Descends:			
Top 30 to 31-100	1	-	-
Total	*1*	-	*0*
Ascend/descend difference	*2*	*1*	*1*
New entries:			
New entry to top 30	-	-	1
New entry to 31-100	1	-	1
New entry 101-400	1	-	2
Total	*2*	*0*	*4*
Entry/drop difference	*2*	*0*	*4*
Total gain/loss	*4*	*1*	*5*

Topic 258 - Cluster 1

	Aspects represented in cluster[6]							
	Asp 1	Asp 2	Asp 3	Asp 5	Asp 6	Asp 7	Asp 9	Asp 13
Ascends:								
31-100 to top 30	1	-	-	-	-	-	1	-
101-400 to top 30	-	-	-	1	-	-	-	-
101-400 to top 100	1	-	-	1	-	-	-	1
Total	*2*	*0*	*0*	*2*	*0*	*0*	*1*	*1*
Descends:								
Top 30 to 31-100	1	1	-	-	2	1	-	1
Top 30 to 101-400	2	-	-	-	-	-	-	-
Top 100 to 101-400	2	1	1	-	1	1	-	-
Total	5	2	1	0	3	2	0	1
Ascend/descend difference	*-3*	*-2*	*-1*	*2*	*-3*	*-2*	*1*	*0*
New entries:								
New entry to top 30	-	-	-	-	-	-	-	-
New entry to 31-100	1	-	-	-	-	-	-	1
New entry 101-400	-	-	-	-	-	-	-	-
Total	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*1*
Drop-outs:								
Drop from 31-100	-	-	-	-	1	-	-	-
Drop from 101-400	-	-	-	-	1	-	-	-
Total	*0*	*0*	*0*	*0*	*2*	*0*	*0*	*0*
Entry/drop difference	*1*	*0*	*0*	*0*	*-2*	*0*	*0*	*1*
Total gain/loss	*-2*	*-2*	*-1*	*2*	*-5*	*-2*	*1*	*1*

[6] Aspect 8 was also represented in this cluster, but there were no other documents assigned to this aspect apart from those in the cluster.

Topic 258 – Cluster 2

	Aspects represented in cluster
	Asp-4
Descends:	
Top 30 to 31-100	1
New entries:	
New entry to 31-100	1
Total gain/loss	0

Topic 258 : Cluster 3

	Aspects represented in cluster		
	asp1	asp6	asp9
Ascends:			
31-100 to top 30	1	3	-
101-400 to top 30	-	1	-
101-400 to top 100	-	1	-
Total	1	5	0
Descends:			
Top 30 to 31-100	1	1	-
Top 30 to 101-400	-	1	-
Top 100 to 101-400	2	1	1
Total	3	3	1
Ascend/descend difference	-2	2	-1
New entries:			
New entry to 31-100	1	-	-
Total	1	0	0
Drop-outs:			
Drop from 101-400	1	-	-
Total	1	0	0
Entry/drop difference	0	0	0
Total gain/loss	-2	2	-1

Topic 258 : Cluster 4

	Aspects represented in cluster	
	asp5	asp17
Descends:		
Top 30 to 101-400	1	1
Top 100 to 101-400	1	1
Total	*2*	*2*
Ascend/descend difference	*-2*	*-2*
Drop-outs:		
Drop from 101-400	1	-
Total	*1*	*0*
Entry/drop difference	*-1*	*0*
Total gain/loss	*-3*	*-2*

Topic 264 : Cluster 1

	Aspects represented in cluster	
	Asp4	Asp15
Descends:		
Top 30 to 31-100	1	-
Top 30 to 101-400	1	-
Total	*2*	*0*
Ascend/descend difference	*-2*	*0*
New entries:		
New entry to 31-100	1	-
New entry 101-400	2	1
Total	*3*	*1*
Entry/drop difference	*3*	*1*
Total gain/loss	*1*	*1*

Topic 264 : Cluster 4

	Aspects represented in cluster
	Asp4
Ascends:	
101-400 to top 30	1
Total	*1*
New entries:	
New entry 101-400	3
Total	*3*
Drop-outs:	
Drop from 101-400	1
Total	*1*
Entry/drop difference	*2*
Total gain/loss	*3*

Topic 284 : Cluster 1

	Aspects represented in cluster	
	Asp7	Asp12
Ascends:		
101-400 to top 30	1	-
Total	*1*	*0*
Ascend/descend difference	*1*	*0*
New entries:		
New entry to top 30	-	1
Total	*0*	*1*
Entry/drop difference	*0*	*1*
Total gain/loss	*1*	*1*

Topic 284 : Cluster 2

	Aspects represented in cluster
	Asp9
Ascends:	
101-400 to top 30	1
Total	1
Total gain/loss	1

Topic 284 - Cluster 4

	Aspects represented in cluster
	Asp19
Drop-outs:	
Drop from 31-100	1
Drop from 101-400	1
Total	*2*
Total gain/loss	*-2*

Topic 286 : Cluster 1

	Aspects represented in cluster[7]				
	Asp1	Asp2	Asp4	Asp7	Asp9
Ascends:					
31-100 to top 30	1	-	1	-	1
101-400 to top 30	4	-	3	-	-
101-400 to top 100	-	-	1	-	-
Total	*5*	*0*	*5*	*0*	*1*
New entries:					
New entry to top 30	-	1	1	-	-
New entry to 31-100	2	1	4	-	-
New entry 101-400	-	1	3	1	-
Total	*2*	*3*	*8*	*1*	*0*
Drop-outs:					
Drop from 31-100	1	-	-	-	-
Drop from 101-400	-	1	-	-	-
Total	*1*	*1*	*0*	*0*	*0*
Entry/drop difference	*1*	*2*	*8*	*1*	*0*
Total gain/loss	*6*	*2*	*13*	*1*	*1*

Topic 286 : Cluster 2

	Aspects represented in cluster[8]	
	Asp1	Asp4
Ascends:		
31-100 to top 30	1	3
101-400 to top 30	3	1
101-400 to top 100	3	4
Total	*7*	*8*
Descends:		
Top 30 to 31-100	-	1
Total	*0*	*1*
Ascend/descend difference	*7*	*7*
New entries:		
New entry to 31-100	-	2
New entry 101-400	2	3
Total	*2*	*5*
Drop-outs:		

[7] Asp3 was also represented in this cluster, but was not present in the hitlists(or was present in the same ranges in both hitlists) apart from the documents that were marked relevant for query expansion.

[8] Aspects 8 and 9 were also represented in this cluster, but were not present in the hitlists(or were present in the same ranges in both hitlists) apart from the documents that were marked relevant for query expansion.

Drop from 31-100	1	-
Total	*1*	*0*
Entry/drop difference	*1*	*5*
Total gain/loss	*8*	*12*

Topic 286 : Cluster 3

	Aspects represented in cluster			
	Asp1	Asp2	Asp4	Asp9
Ascends:				
31-100 to top 30	1	1	1	-
101-400 to top 100	6	1	3	-
Total	*7*	*2*	*4*	*0*
Descends:				
Top 100 to 101-400	1	-	-	-
Total	*1*	*0*	*0*	*0*
Ascend/descend difference	*6*	*2*	*4*	*0*
New entries:				
New entry to top 30	-	-	3	-
New entry to 31-100	2	-	1	1
New entry 101-400	-	3	6	-
Total	*2*	*3*	*10*	*1*
Drop-outs:				
Drop from 31-100	-	1	1	1
Drop from 101-400	-	-	-	-
Total	*0*	*1*	*1*	*1*
Entry/drop difference	*2*	*2*	*9*	*0*
Total gain/loss	*8*	*4*	*13*	*0*

Topic 286 : Cluster 4

	Aspects represented in cluster	
	Asp1	Asp4
Ascends:		
31-100 to top 30	2	2
101-400 to top 30	1	-
101-400 to top 100	5	5
Total	*8*	*7*
Descends:		
Top 30 to 31-100	-	4
Total	*0*	*4*
Ascend/descend difference	*8*	*3*
New entries:		
New entry to top 30	-	1
New entry to 31-100	-	1
New entry 101-400	2	4
Total	*2*	*6*
Drop-outs:		
Drop from 31-100	1	-
Total	*1*	*0*
Entry/drop difference	*1*	*6*
Total gain/loss	9	9

Query 292 – Cluster 2

	Aspects represented in cluster
	Asp12
Ascends:	
101-400 to top 100	1
Total	1
New entries:	
New entry 101-400	2
Total	2
Total gain/loss	3

Query 292 – Cluster 3

	Aspects represented in cluster
	Asp21
Ascends:	
101-400 to top 100	1
Total	1
Total gain/loss	1

Appendix 6.4 : Rate of overlap between hitlists resulting from query expansion on each cluster and the initial hitlists

Topic 254:

Rate of overlap between hitlists resulting from relevance expansion on each cluster

	Cl1-Cl2	Cl1-Cl3	Cl1-Cl4	Cl2-Cl3	Cl2-Cl4	Cl3-Cl4
top 30						
common documents	3	10	10	4	1	15
common relevant documents	0	4	3	1	1	10
top 400						
common documents	170	227	135	148	180	111
common relevant documents	14	19	16	24	21	28

Rate of overlap between the initial hitlists vs hitlists resulting from relevance expansion on each cluster

top 30	Cl1	Cl2	Cl3	Cl4
common documents	9	0	10	11
common relevant documents	6	0	8	10
top 400				
common documents	207	144	223	99
common relevant documents	15	22	27	23

Topic 256:

Rate of overlap between hitlists resulting from relevance expansion on each cluster

	Cl1-Cl2	Cl1-Cl3	Cl1-Cl4	Cl2-Cl3	Cl2-Cl4	Cl3-Cl4
top 30						
common documents	n/a	3	1	n/a	n/a	0
common relevant documents	n/a	0	0	n/a	n/a	0
top 400						
common documents	n/a	42	36	n/a	n/a	0

common relevant documents	n/a	0	1	n/a	n/a	0

Rate of overlap between the initial hitlists vs hitlists resulting from relevance expansion on each cluster

top 30	Cl1	Cl2	Cl3	Cl4
common documents	4	n/a	2	3
common relevant documents	0	n/a	0	0
top 400				
common documents	154	n/a	46	97
common relevant documents	1	n/a	0	4

Topic 258:

Rate of overlap between hitlists resulting from relevance expansion on each cluster

	Cl1-Cl2	Cl1-Cl3	Cl1-Cl4	Cl2-Cl3	Cl2-Cl4	Cl3-Cl4
top 30						
common documents	15	3	14	0	14	2
common relevant documents	4	0	3	0	3	0
top 400						
common documents	269	249	295	263	272	246
common relevant documents	38	39	33	44	39	39

Rate of overlap between the initial hitlists vs hitlists resulting from relevance expansion on each cluster

top 30	Cl1	Cl2	Cl3	Cl4
common documents	3	3	5	2
common relevant documents	1	1	1	1
top 400				
common documents	117	123	138	114
common relevant documents	37	42	44	39

Topic 264:

Rate of overlap between hitlists resulting from relevance expansion on each cluster

	Cl1-Cl2	Cl1-Cl3	Cl1-Cl4	Cl2-Cl3	Cl2-Cl4	Cl3-Cl4
top 30						
common documents	5	14	10	6	2	12
common relevant documents	0	0	1	0	1	1
top 400						
common documents	169	260	220	181	144	260
common relevant documents	5	2	3	2	3	1

Rate of overlap between the initial hitlists vs hitlists resulting from relevance expansion on each cluster

top 30	Cl1	Cl2	Cl3	Cl4
common documents	12	4	15	10
common relevant documents	0	0	0	0
top 400				
common documents	236	194	283	228
common relevant documents	4	3	3	2

Topic 284:

Rate of overlap between hitlists resulting from relevance expansion on each cluster

	Cl1-Cl2	Cl1-Cl3	Cl1-Cl4	Cl2-Cl3	Cl2-Cl4	Cl3-Cl4
top 30						
common documents	2	5	6	5	3	9
common relevant documents	1	1	4	2	0	2
top 400						
common documents	127	232	214	143	191	206
common relevant documents	18	26	28	17	25	25

Rate of overlap between the initial hitlists vs hitlists resulting from relevance expansion on each cluster

top 30	Cl1	Cl2	Cl3	Cl4
common documents	1	0	1	2
common relevant documents	1	0	0	1
top 400				
common documents	249	140	318	198
common relevant documents	28	20	32	27

Topic 286:

Rate of overlap between hitlists resulting from relevance expansion on each cluster

	Cl1-Cl2	Cl1-Cl3	Cl1-Cl4	Cl2-Cl3	Cl2-Cl4	Cl3-Cl4
top 30						
common documents	17	4	12	8	11	5
common relevant documents	9	3	9	6	7	4
top 400						
common documents	298	197	268	111	354	84
common relevant documents	26	35	26	33	31	34

Rate of overlap between the initial hitlists vs hitlists resulting from relevance expansion on each cluster

top 30	Cl1	Cl2	Cl3	Cl4
common documents	5	7	9	5
common relevant documents	5	6	7	4
top 400				
common documents	189	211	111	208
common relevant documents	21	26	30	29

Topic 292:

Rate of overlap between hitlists resulting from relevance expansion on each cluster

	Cl1-Cl2	Cl1-Cl3	Cl1-Cl4	Cl2-Cl3	Cl2-Cl4	Cl3-Cl4
top 30						
common documents	0	0	0	8	10	4
common relevant documents	0	0	0	0	0	0
top 400						
common documents	2	0	5	136	185	91
common relevant documents	0	0	0	10	16	10

Rate of overlap between the initial hitlists vs hitlists resulting from relevance expansion on each cluster

top 30	Cl1	Cl2	Cl3	Cl4
common documents	5	0	0	0
common relevant documents	0	0	0	0
top 400				
common documents	79	140	55	95
common relevant documents	0	5	5	7