QUERY EXHAUSTIVITY, RELEVANCE FEEDBACK AND SEARCH SUCCESS IN AUTOMATIC AND INTERACTIVE QUERY EXPANSION

Pertti Vakkari* & Susan Jones" & Andy McFarlane" & Eero Sormunen*

* Department of Information Studies, FIN-33014 University of Tampere, Finland
" Department of Information Science, City University, Northampton Square, London EC1V OHB, U.K.

Abstract

This study explored how the expression of search facets and relevance feedback by users was related to search success in interactive and automatic query expansion in the course of the search process. Search success was measured both in the number of relevant documents retrieved and relevance scores of these items based on a four point scaling. Research design consisted of 26 users searching for four TREC topics in Okapi IR system, half using interactive and half automatic query expansion based on RF. The search logs were recorded, and the users filled in a questionnaire for each topic concerning various features of searching. The results showed that the exhaustivity of the query was the most significant predictor of search success, and that interactive expansion led to better search success than automatic one.

1. Introduction

Interactive studies using best-match systems have typically analysed how various means of term suggestion and selection are connected to search success. Success has mainly been related to average features of the process, e.g. average number of iterations or terms used over the whole process (cf. Beaulieu & al. 1996; Over 2001), not how effectiveness evolves in the course of the interaction process. It is evident that the interaction of human capacities like expressing search concepts in terms and giving relevance feedback (RF) contribute throughout the process interactively on search outcome (cf. Brajnik & al. 1996; Fowkes & Beaulieu 2000, Vakkari 2002). Also a more detailed understanding of the search process in addition to outcome helps to create tools for supporting it. The aim of this study is to explore how users' ability to express search facets and RF is related to search success in the course of the search process.

In systems where query expansion is based on RF two capabilities of users are crucial for success in searching: their ability to select search terms and judge the relevance of the items retrieved (Beaulieu & al. 1996; Belkin 1996). Firstly, users' ability to express the conceptual constructs representing their information needs in terms in initial query formulation is critical for searching. The success of the initial search depends on that articulation. The best results are achieved if the chosen terms cover the concepts of the construct, and reflect the terminology used in the potentially relevant documents (Bates 1986; Blair & Maron 1985). In the subsequent search process users' ability to construct queries by selecting and removing terms from the list generated by RF or any other term suggestion device is vital for the success of the retrieval (Beaulieu & al. 1996). Pennanen & Vakkari (2002) have shown that the more exhaustively the users were able to articulate their information problems in query terms in a Boolean system, the more useful references they retrieved in a search.

Secondly, users' ability to assess the relevance of the items retrieved has an impact on terms extracted by RF mechanism. If they are capable of identifying relevant and particularly highly relevant documents in the result set, the terms extracted are likely to reflect pertinently both the content of the conceptual construct representing information need and its expressions in the vocabulary used in the potentially relevant documents (Beaulieu & al. 1996; Vakkari 2002). If the users assess improper documents as relevant, the extracted terms do not represent the information need validly and lead to query drift, i.e. the alteration of the search topic caused by improper query expansion (Mitra & al. 1998).

The degree of automation of query expansion based on RF regulates user involvement in the process (Beaulieu & al. 1996; Efthimiadis 1996). In automatic query expansion (AQE) users formulate the initial query and the system expands it based on RF. In interactive query expansion (IQE) users select or remove terms from a list provided by RF mechanism. Thus, after the initial search in the previous one only users' ability to judge the relevance of documents, and in the latter one both their capability to judge relevance and to choose terms affect retrieval effectiveness. It is an open question to what degree articulation and use of search terms and relevance feedback affect search success in automatic and interactive query expansion. There is scattered evidence either on how RF (Hancock-Beaulieu & al. 1995) or term selection is related to search success (Beaulieu & al. 1996; Brajnik & al. 1996; Fowkes & Beaulieu 2000; Koeneman & Belkin 1996) in IQE, but no studies on the combined contribution of these two factors on search success. The aim of this study is to analyse how search success is related to RF and term selection in successive interactive and automatic query expansions.

2. Related studies

In the few interactive studies on QE research questions vary, producing fragmenting results. There are only a handful of studies comparing AQE and IQE.

Most of the studies show that QE benefits users by producing a better retrieval performance (Belkin & al. 2000), (Hancock-Beaulieu & al 1995), (Koenemann & Belkin 1996). Short and partially matching initial searches seem to gain most from QE (Hancock-Beaulieu & al 1995), (Fowkes & Beaulieu 2000). The few comparisons between AQE and IQE suggest that IQE is more efficient (Koenemann & Belkin 1996). A problem with AQE seems to be the systems' poor ability to recognise facets and structure the query accordingly (cf. Jones & al. 1995, Kekäläinen 1999, Mitra & al 1998).

Koenemann and Belkin (1996) found that in IQE searchers used on average fewer terms and made fewer iterations than in AQE. Beaulieu & al. (1996) showed that in IQE compared to automatic one searchers' term manipulations were beneficial to search success. They found that terms removed by the searchers were justified due to

their low improvement of performance. On the other hand, they found that relevance feedback terms as opposed to user terms had greater retrieval effectiveness (cf. Spink 1995).

The quality of the selected terms had an impact on the success of QE. Difficulty to understand terms provided by Local Context Analysis in IQE caused inconvenience to users in term selection (Belkin & al. 2000). IQE was successful if the terms suggested seemed to be potentially useful (Hanckock & al. 1995)and if the terms selected had clear and central semantic relations to the search topic (Koenemann & Belkin 1996).

The effectiveness of QE seems also to depend on users' ability to identify search facets and on the complexity of the search topic. In a study by Brajnik & al. (1996) 45 users searched for two topics in an experimental Boolean system by using IQE. Term suggestions were based on a thesaurus. The results showed that in an easier topic, where the users were able to reach an appropriate conceptualisation (i.e. to identify facets), there was a statistically significant correlation between the average number of terms per query and search performance, whereas no correlation in the difficult topic. They conclude that IQE has a synergetic effect in easier topics, but does not help searchers if they fail to identify enough facets for the query. Thus, they suppose that users' ability to articulate facets of a topic is vital for successful selection of search terms from the list of terms provided in IQE, and consequently for the success of the search.

Fowkes & Beaulieu (2000) studied how 24 students searched for six TREC-8 topics on two versions of the Okapi system. They found that in terms of identifying instances IQE was more productive for more difficult and complex topics as perceived by searchers, whereas AQE was more effective in simple topics. Simple topics were precisely defined and instances could be easily identified by the outset. Thus, it seems that RF in these cases was good, producing pertinent query terms for AQE contributing to search success.

Complexity referred to users' difficulty in understanding the scope and content of the topic. Fowkes & Beaulieu (2000) mention that these topics required either defining

the scope of the topics, e.g. what is a tropical storm, or effort in making sense of the documents themselves. Thus, it seems that the authors refer by the complexity of the topic in the previous case to the selection of search terms, and in the latter one to the relevance feedback. Thus, IQE seems to work better compared to AQE if users have difficulties in selecting search terms or giving RF. Especially the user-generated terms during the search seemed to explain the success to a certain extent.

In sum, IQE seems to produce a better search performance with fewer search terms and iterations compared to AQE. Some studies show that QE is successful if searchers are able to articulate search topics in query terms. However, IQE works better than AQE if they have difficulties in identifying and expressing facets and giving RF. AQE works better if facets of the topic are easier to identify or if assessing relevance is easy.

3. Research design

The research question of this study is how searchers' selection of search terms to express the aspects of the topic and their relevance assessments for feedback are related to the number of relevant documents and the degree of relevance of the documents retrieved by successive automatic and interactive query expansions.

3.1. Experimental setting

The study subjects were 26 students of information science, electronic journalism and computer science at City University, London using the Okapi system. They were not familiar with the system. All students searched for four TREC topics (338, 403, 427 and 442), half using automatic and half interactive query expansion facilities of Okapi. The topics were randomised and rotated completely so that each student searched the topics in different order.

Both versions of Okapi had a common user interface consisting of three display areas (Figure 1). The top frame provides a space for query entry and control buttons for selecting search functions. The narrow left frame present a list of terms comprising the current query and the wider frame switches between hit-list and document display.

(insert figure 1 about here)

In AQE the searchers formed the initial query. Each subsequent query expansion was based on 15 best terms extracted from the documents indicated relevant by searchers. In IQE the searchers also formed the initial query, but the system allowed them to remove terms from a list of 15 best terms suggested by the system based on RF.

In the beginning of the session the system, the aim and the procedure of the trial was briefly introduced and demonstrated to the searchers. Also written instructions were distributed to the participants. Before searching they filled in a short questionnaire measuring their search experience. A twenty minutes time slot was allocated for each topic, during which they were asked to find as many relevant documents as possible. The transaction logs of the searches were recorded. After each topic they filled in a questionnaire measuring familiarity with the topic, experiences with the searching task like selection of terms, assessing the relevance of the items, and controlling the direction of the search, and satisfaction with the results and time allotted for the search.

3.2. Measurement

A facet analysis was conducted for the four TREC topics by two members of the research team. First both of them made the analysis separately. The results were compared and the differences solved by discussion in a group, which consisted of the two researchers and an expert in facet analysis.

No	Topic description	Facets
388	Identify documents that discuss the use of organic fertilizers (composted sludge, ash, vegetable waste, micro-organisms, etc.) as soil enhancers.	• Activity: the process of soil enhancement • Mechanism: organic fertilizers • Components: composed sludge, ash, vegetable waste, micro-organisms
403	Find information on the effects of dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone	• Condition: osteoporosis • Causes: intake potassium, magnesium, fruits, vegetables • Effects: bone mineral density • Person: elderly

	decay).	• PREVENTION
427	Find documents that discuss the damage ultraviolet (UV) light from the sun can do to eyes.	• Condition: Eye damage, diseases, cataracts, ocular melanoma • Causes: Sun, UV, ultraviolet light
442	Find accounts of selfless heroic acts by individuals or small groups for the benefit of others or a cause.	• Activity: Heroic acts (particular) • Person: Individuals, small groups

Table 1. Facets of the topics 388, 403, 427 and 442.

A facet is a concept (or a family of concepts) identified from and defining an exclusive aspect of a topic. It may contain one or more search terms (Sormunen 2002b). The result of the analysis is shown in table 1.

The exhaustivity of a query denotes to the number of facets covered by query terms (Sormunen 2002). In this study query exhaustivity is the ratio (%) between the number of facets expressed by query terms per the number of facets in a topic. Also the number of terms per a facet was calculated. It can be called query extent (Sormunen 2002b). The variables of the study are described in table 2.

We measured the number of terms generated by users in as well as the terms extracted by the system be they added automatically or selected by the users to the query. The maximum number of terms extracted per iteration was 15.

Variable	Measurement
Query exhaustivity	100 x # of facets expressed by query terms / # of all facets
Terms per facet	# of terms / facet in a query
User (U) terms	# of terms generated by a user
Extracted (X) terms	# of terms extracted by the system
Relevant documents	# of retrieved documents judged relevant by formal assessors
Relevance scores	Sum of gradings of relevant documents
Feedback	# of documents judged relevant consistently by a user and formal assessors – # of documents judged relevant only by a user

Table 2. Variables of the study

In order to get a more detailed picture of the search success the documents for the four TREC topics were analysed by using a four-level relevance scale reflecting the degree of relevance of the documents. The point of departure of the analysis was the pool of relevant documents for each topic produced by the TREC judges. These documents were reassessed by using procedures and criteria described in Sormunen (2002a). The topics were assessed by one assessor each. The judges were familiar with the criteria and the procedure used due to their earlier reassessing experience. The final pool consisted of those documents judged as relevant by TREC assessors and some extra assessed relevant by our judges. TREC and our judges are called "formal assessors".

Search success was measured in two ways. Firstly, we calculated the number of retrieved documents that were judged relevant by formal assessors. Secondly, we calculated relevance scores - or cumulative gain as called by Järvelin & Kekäläinen (2000) - by adding up the gradings of all found documents assessed as relevant by TREC or our assessors. The grading was as follows: Those documents which were judged as relevant by TREC, but not relevant by our assessors = 1. Marginally relevant documents = 2; relevant documents = 3; highly relevant documents = 4.

In those cases when a searcher had judged a document as relevant, but it was not judged by TREC or our assessors, a member of the research team read the article and estimated its relevance according to the criteria in Sormunen (2002a). Only very few of those documents were relevant.

The nature of relevance feedback depends on how validly the searcher is able to assess the relevance of the retrieved items. We measured the relevance feedback by subtracting the number of documents assessed relevant only by a user from those which the user assessed as relevant consistent with formal assessors. Thus, we suppose that the previous ones do not reflect the topic validly, and produce improper query terms. The higher the feedback score, the better the feedback for term extraction.

4. Results

4.1. First two iterations

The results of the study will be calculated over the four topics, because the number of searchers per topic (26) were too small for statistical elaboration of the results. This approach also needs to be adopted because of the small number of participants who

expanded the initial queries. The number of expanded searches varied from 18 in the topic 442 to 23 in the topic 427.

The students had difficulties in starting searches and perhaps also in formulating initial queries. Many of them restarted the search many times. This reflected their difficulties in grasping the expansion mechanism of the system. In the initial query formulation they had to first enter the terms and press the "return" button, and after that press "search the database" button. A similar procedure was required after relevance assessments. They were used to systems where searching began immediately after the return button was pressed, which difference made the searching clumsy. This was also reflected in the few iterations (on average 2. 8) they made per topic.

In the initial search significantly more terms, and also slightly more facets were generated by those who used IQE (table 3). Their queries were slightly more exhaustive than among those who used AQE. The former also retrieved slightly more relevant documents and their relevance score was higher. On average, the documents retrieved in both groups were only marginally relevant (2.3), which did not differ from the mean relevance scores of documents (2.3) over all topics.

Variables	AQE (41)	IQE (43)	p.
User-terms	4.6	5.8	.005
Facets	2.0	2.3	.12
Query exhaustivity (%)	73	80	.14
Terms / facet	2.5	2.6	.82
Relevant documents	3.1	3.8	.30
Relevance scores	7.0	8.9	.25
Scores / document	2.3	2.3	.95
Feedback	0.4	1.0	.43

Table 3. Results of the initial search

There was no statistically significant difference in the number of those documents, which the users identified relevant consistent with the formal assessors (2.4 vs. 2.3; p = .67), whereas IQE retrieved significantly (p. = .05) more documents compared to AQE, which were classified relevant by official judges, but not recognised by the users.

Perceived familiarity with the topics did not differ in the two user groups (3.5 vs. 3.7; p = .50). This is in line with the finding that in both groups the number of those documents, which were identified relevant consistent with formal judges were about the same. Thus, the familiarity with the topic does not explain the difference in the number of relevant documents.

The difficulties experienced in discovering terms for the initial query did not differ in both groups (2.7 vs. 2.6; p = .94). Thus, the perceived difficulty did not explain the difference in the number of terms or facets between the groups in the initial query.

Correlation coefficients show that the number of user generated terms explains to a certain extent the difference in the number of facets (r = .38; n = 84; p. < .05, r > .26). Query exhaustivity correlates most strongly with the number of relevant documents (r = .37), and relevance scores (r = .32). Thus, users who generated most terms for the initial query were able to cover most extensively the facets by query terms, which was vital for the search success.

Variables	AQE (41)	IQE (43)	p.
User-terms	2.8	2.7	.62
Terms in total (U+X)	15	9.3	.0001
Facets	2.0	2.1	.39
Query exhaustivity (%)	71	69	.77
Terms / facet	8.4	5.2	.0001
Relevant documents	1.5	2.1	.16
Relevance scores	3.3	4.4	.27
Scores / document	2.3	2.2	.69
Feedback	-0.2	0.2	.40

Table 4. Results of the second iteration (first expansion).

In the second iteration AQE produced a significantly higher number of terms and terms per facet (table 4) mainly due to the removal of system suggested terms by those using IQE. Compared to the initial search query expansion reduced significantly the number of user generated terms (Difference = 2.4; p. = .0001) and query exhaustivity (Difference = 6; p. = 003). Especially interactive expansion decreased the number of User-terms and query exhaustivity reducing the difference between the groups. A closer examination revealed that both the system and users removed initial query terms. This removal also affected the reduction of query exhaustivity.

Also in the second iteration IQE produced a slightly greater number of relevant documents and higher relevance scores than AQE. Again, there were no differences in average scores per document between the groups. Neither did these scores differ from the mean relevance scores over the topics.

It is interesting to note that although there were no differences in the number of user generated terms and query exhaustivity between the groups in the second iteration, those using IQE found slightly more relevant documents and their relevance scores were slightly higher than those using AQE. Would it be that in IQE searchers were able to delete system generated terms that did not represent the topic pertinently, and that led to a higher number of relevant documents found compared to AQE? The feedback from the first iteration was not significantly better in the interactive group that it would have explained the difference in the term provision and search results.

The variables that correlated significantly both with the number of relevant documents and relevance scores were query exhaustivity ($r = .43$; $r = .35$) and relevance feedback from the first iteration ($r = .31$; $r = .29$). The type of query expansion was not significantly associated with these variables($r = .16$; $r = .13$). Thus, both the exhaustivity of the query and the correctness of the relevance feedback were positively related to the search success.

In order to elaborate how these three factors contribute to the search success an analysis of variance was calculated. As Fowkes & Beaulieu (2000) have shown and as the exploration of the data of this study confirms, the complexity of the topic has an impact on searchers' ability to articulate search facets and on search results. Because the topics are not homogeneous, the between topic variation of factors may cause spurious associations. In order to control and reduce this between topic variation, each explaining variable, the type of query expansion, query exhaustivity and relevance feedback, were divided into two categories within the topics using median as cutting point. This was also necessary so cells with too few cases were not formed. To reduce the between topic variation in the variables to be explained - the number of relevant documents and relevance scores - the values of these variables within each topic were reclassified into three categories so that the number of cases in each category were about even, and that the within category variation was relatively small. The idea was to form even and homogeneous categories within the topics, but avoid using these factors as dummy variables. This reclassification transformed the variables from interval to ordinal scales. The scores of these variables indicate the

relative position of the units of observation in the scale within each topic, and thus make the factors comparable over the topics.

Relevance feedback	Good				Poor				Totals (n)	
Query exhaustivity	Low	(n)	High	(n)	Low	(n)	High	(n)		
AQE	1.6	(7)	2.3	(12)	1.9	(12)	1.7	(10)	1.9	(41)
IQE	1.3	(6)	2.6	(12)	1.8	(16)	2.1	(9)	2.0	(43)
Totals	1.5	(13)	2.4	(24)	1.8	(28)	1.9	(19)	2.0	(84)

Table 5. The means of the number of relevant documents (relative values) according to QE type, feedback and query exhaustivity in the first expansion

The type of query expansion did not have a significant direct effect on search success in the second iteration (table 5). However, it had a statistically significant impact on search results depending on how exhaustively the query covered the facets of the topic ($F = 3.53$; $p = .05$). Overall, the more exhaustive the query was, the more documents were retrieved ($F = 10.49$, $p = .002$). Especially in the most favourable situations when both relevance feedback was good and query was exhaustive, most relevant documents were retrieved ($F = 8.01$, $p = .006$). In these situations IQE led to the best results compared to automatic one.

Also in the case when relevance feedback was poor, but the query exhaustive, IQE led to a better result than automatic one. It seems that if the user was able to keep the query exhaustive, the opportunity to remove weak terms obviously compensated the poor feedback, and increased the number of retrieved relevant items compared to AQE.

Thus, IQE retrieved more documents compared to automatic one independent of RF, if the query was exhaustive. The exhaustivity of the query was vital for the success of IQE.

In the cases when query was unexhaustive, search success was at its weakest independent of RF. When queries were not exhaustive, AQE seemed to lead to a better search result if the relevance feedback was good. If the user was not able to cover facets exhaustively, but able to give good feedback, the search result was better when using automatic expansion. The feedback evidently produced relevant query terms. The strong elimination of suggested terms by this interactive group (8 terms compared to 5-6 in other interactive groups) deteriorated the query, and led to poorer

search results compared to automatic expansion which included all the extracted terms. It is evident that in some cases the removal of terms removed also a facet, which decreased the exhaustivity of the query leading to a weaker search success (cf. Sormunen 2002b).

In sum, IQE was more effective compared to automatic one, if the searcher was able to express the facets of the topic exhaustively especially if the relevance feedback was good. AQE worked better, if the user could not express the facets of the topic exhaustively, but was able to generate good feedback. However, the most successful search was a result of exhaustive facet articulation combined with good relevance feedback in interactive expansion. The analysis of variance of relevance scores produced roughly similar results.

4.2. Last iterations

About half of the expanded searches were finished after the second iteration. Out of the 40 searches, which were expanded further, 29 were interactive and 11 automatic. The average number of iterations by those who continued was the same (3.6) in both groups. They did significantly more iterations compared to those who stopped after the first expansion ($F = 175.04$; $p = .0001$).

In the first expansion the users of interactive mode found slightly more relevant documents than the users of automatic mode (table 4). It is interesting to analyse how was the search success related to the expansion type among those who continued searching after the first expansion. Table 6 shows that the increased number of expansions led to a slight increase in the number of relevant documents retrieved in favour of the interactive mode. Although these differences were not statistically significant, there was a tendency that the difference in the number of found relevant documents increased in favour of interactive mode in the course of the search process.

QE type	1st expansion	2- expansions	Totals
AQE (11)	1.6	1.9	3.6
IQE (29)	2.0	2.7	4.7

Table 6. The average number of relevant documents retrieved in the first expansion and in the expansions after it among those who continued searching.

Comparison of the increase in the number of relevant documents and relevance scores over all expansions reveals that interactive expansion produced better search results than automatic one (table 7.) Users of automatic mode found on average about two relevant documents whereas users of interactive mode found nearly five relevant documents by expanding the initial queries. This difference is statistically significant. The difference in relevance scores was clear although not statistically significant. There was no difference in the average relevance scores per found document.

	AQE (41)	IQE (43)	p.
Relevant documents	2.1	4.0	.02
Relevance scores	4.7	8.1	.10

Table 7. The increase in the number of relevant documents and relevance scores after the first iteration.

In sum, it seems that on average over all expansions interactive mode gives a better search result than automatic one.

4.3. Joint effect on search success

In order to analyse in more detail the joint effect of query expansion type, average relative query exhaustivity, average relative feedback and the number of iterations to the average relative increase in the number of relevant documents retrieved after the first iteration in the whole process, a regression analysis was calculated (table 8).

R = .325 R_2 =.106 Adjusted R_2 = .057 DF(4, 77) = 2,157 p. = .082

Variable	Beta	St. error	B	t-value	p.
Intercept	.518				
QE type	.151	.197	.092	.766	.45
Average feedback	.092	.184	.056	.501	.62
Average exhaustivity	.358	.182	.218	1.965	.05
# of iterations	.158	.104	.183	1.328	.13

Table 8. Multiple regression for the increase in the average relative number of retrieved relevant documents after the first iteration

The only variable in this linear model which correlated significantly with search success was the exhaustivity of the query (p. = .05), even though the number of

iterations seems to be statistically indicative (p. = .13). The type of query expansion or feedback was not associated with the search success. The results of the regression analysis concerning average relative relevance grades showed that none of the four explaining variables were associated with the improvement of the relevance scores after the initial querying. Thus, as a result of the whole query expansion process the average exhaustivity of the query has a significant impact on the increase in the number of relevant documents retrieved.

5. Discussion and conclusions

This study explored how the expression of search facets and relevance feedback by users was related to search success in AQE and IQE. It sought to connect search outcome to these features in the course of the search process. Search outcome was measured both in the number of relevant documents retrieved and relevance scores of these items based on a four point scaling. Research design consisted of 26 users searching for four TREC topics in Okapi, half using interactive and half automatic query expansion based on RF.

The study was based on a smallish number of topics and searchers. This limited the elaboration of the results between the successive searches. It also forced to collapse the searches of the four topics, and use in the elaborations transformed measures, which indicated the relative values of variables in the topics.

The results showed that in the first expansion round IQE helped the users to retrieve slightly more relevant documents than AQE. However, the average relevance scores of the items found did not differ. Thus, the type of query expansion was not associated with the relevance grade of the documents retrieved in the beginning of searching. When users continued searching after the first expansion IQE produced slightly more relevant items than AQE. Again, there was no difference in the relevance gradings of the retrieved documents between the modes of query expansion. Neither of the QE mechanisms were able to support the users in finding exceptionally highly relevant documents. The average relevance scores of the documents retrieved reflected the average scores of the relevant items in the topics.

In all, it seemed that interactive mode produced over the whole expansion process more relevant documents than automatic one. This finding is in line with the results by Koeneman & Belkin (1996) that IQE was more efficient than AQE. Our results

confirm also their finding that on average interactive expansion consisted on fewer terms than automatic one.

The small number of searches prevented the deeper elaboration of results in the successive searches after the first expansion. However, the analysis of the first query expansion showed that search success was related both to how exhaustively search facets were covered by query terms and to the relevance feedback. This result is consistent with the findings of Brajnik & al (1996) and Fowkes & Beaulieu (2000) which suggest that query expansion is successful if searchers are able to articulate the facets of topics in query terms.

The major finding from the first expansion was that the number of relevant documents retrieved was the result of an interaction between the share of facets in the topic articulated by query terms (exhaustivity of the query) and the successful feedback and the type of query expansion. Thus, the success of QE type depends on users' ability to express facets and relevance feedback.

The exhaustivity of the query was the strongest predictor of search success in the first expansion. The more exhaustive the query was, the more documents were retrieved. When the searchers were able to produce exhaustive queries, interactive expansion was more successful than automatic one. Especially in the most favourable cases, when the the query was exthaustive and the feedback good, most relevant documents were retrieved by both means of expansion, and IQE helped in finding most items.

Also in cases when feedback was poor, but query exhaustive, IQE helped to find more relevant documents than AQE. It seemed that if the user was able to keep the query exhaustive, the opportunity to remove inappropriate terms obviously compensated the poor feedback, and increased the number of retrieved relevant items compared to AQE.

In the cases when query exhaustivity was low, search success was at its weakest independent of the RF. If the user was not able to cover facets exhaustively, but able to produce a good feedback, the search result was better when using automatic expansion due to the heavy elimination terms by the interactive group.

The main result of this study was that compared to the query expansion type or the goodness of RF or the number of iterations, the exhaustivity of the query was the most significant predictor of search success. It seemed also that an exhaustive query compensated to a certain extent weak feedback, but not vice versa. Thus, the success

of expansion was essentially connected to users' opportunity and ability to cover the facets of the topic. This finding of the importance of query exhaustivity to search success in a best-match system is consistent with the results by Sormunen (2002b). He has shown that in laboratory conditions increasing the exhaustivity of optimised unstructured queries led to a significant increase in precision especially in low levels of exhaustivity. Thus, increased covering of facets of the topic by query terms is vital for search success.

Sormunen's (2002b) findings refer also to that that increase in exhaustivity improves more precision of structured than unstructured optimised best-match queries. This finding combined with our results suggests that best-match IR systems should facilitate the formulation of structured queries for the users. This is in line with the findings from laboratory experiments, which show that structured queries perform better than weakly structured ones (Hawking & al. 1997; Kekäläinen 1999; Kekäläinen & Järvelin 1998). This suggestion fits also with the finding that highly relevant documents benefit essentially more from the concept-based query expansion in ranking than marginally relevant ones (Sormunen & al. 2001), although our study did not show a connection between the exhaustivity of the query and relevance scores.

Neither type of expansion based on RF support structuring queries. The result of the exercise depends primarily on users' ability to find terms for articulating facets and secondarily on feedback. As the results of this and other studies (Pennanen & Vakkari 2002) suggest, users' ability to articulate facets vary. Also their ability to eliminate in IQE inappropriate terms varies producing varied search results (Beaulieu & al. 1996). A point of departure for improving best-match systems would be providing users in the initial search with an option to explicate the facets of the topic. This would at least give a more successful starting point for successive expansions.

As our results suggested a good feedback does not compensate an unexhaustive query. Thus, supporting users to keep the query exhaustive would be desirable. A means for preserving exhaustivity and increasing the extent of the facets is to expand the initial query automatically by using thesaurus. In laboratory experiments this has produced successful results compared to unstructured queries (Kekäläinen 1999). As far as we know there are not yet methods for automatically expanding queries in a structured way based on initial query terms and RF (cf. Kekäläinen 1999; Mitra & al. 1998). Our results are in favour of developing such methods.

References

Bates, M. Subject access to online catologs: A design model. Journal of the American Society for Information Science 37 (6) (1986) 357-376.

Beaulieu, M. & Robertson, S. & Rasmussen, E. Evaluating interactive systems in TREC. Journal of the American Society for Information Science 47(1) (1996) 85-94.

Belkin, N. Intelligent information retrieval. Who's intelligence? In: ISI'96: Proceedings of the Fifth International Symposium for Information Science. Konstanz: Universitätsverlag Konstanz (1996) 25-31.

Belkin, N. & Cool, C. & Head, J. & Jeng, J. Kelly, D. & al., Relevance Feedback versus Local Context Analysis as term suggestion devices: Rutgers' TREC-8 interactive track experience. In: Proceedings of TREC-8. (2000).

Blair, D. & Maron, M. An evaluation of retrieval effectiveness for a full-text document retrieval system. Communications of the ACM. 28 (3) (1985) 289-299

Brajnik, G., Mizzarro, S. & Tasso, C. Evaluating user interfaces to information retrieval systems. A case study. In Proceedings of the SIGIR'96. ACM, New York (1996) 128-136.

Efthimiadis, E. Query expansion. In: M.E. Williams (ed), Annual review of information science and technology, vol 31. Information Today, Medford, N.J. (1996) 121-187.

Fowkes, H., & Beaulieu, M. Interactive searching behavior: Okapi experiment for TREC-8. In Proceedings of the BCS-IRSG: 22nd Annual Colloquium on Information Retrieval Research. Cambridge (2000), 47-56.

Hancock-Beaulieu, M. & Fieldhouse, M. & Do, T. An evaluation of interactive query expansion in an online library catalogue with graphical user interface. Journal of Documentation. 51 (3) (1995) 225-243.

Hawking, D. & Thistlewaite, P. & Bailey, B. ANU/ACSys TREC-5 experiments. In: E. Voorhees & D. Harman (eds), Information technology: The fifth text retrieval conference (TREC-5). MD, Gaithersburg (1997) 359-375.

Jones, S. & Gatford, M. & Robertson, S. & Hancock-Beaulieu, M. & Secker, J. Interactive thesaurus navigation: Intelligence rules OK? Journal of the American Society for Information Science. 46 (1) (1995) 52-59.

Järvelin, K. & Kekäläinen, J. IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the SIGIR 2000. ACM, New York (2000) 41-48.

Kekäläinen, J. The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. Doctoral Dissertation. Tampere University Press, Tampere (1999).

Kekäläinen, J. & Järvelin, K. The impact of query structure and query extension on retrieval performance. In: Proceedings of the SIGIR´98. ACM, New York (1998) 130-137.

Koenemann, J. & Belkin, N. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In: Proceedings of the Human Factors in Computing Systems Conference (CHI'96). ACM Press, New York (1996) 205-212.

Mitra, M. & Singhal, A. & Buckley, C. Improving automatic query expansion. In Proceedings of the SIGIR´98. ACM, New York (1998) 206-214.

Over, P. The TREC interactive track: an annoted bibliography. Information Processing & Management 37(3) (2001) 368-381.

Pennanen, M. & Vakkari, P. Students' cognition and information searching while preparing a research proposal. In: H. Bruce & al. (Eds.) Emerging frameworks and methods. Proceedings of the CoLIS4. Libraries Unlimited, Greenwood, Col. (2002), 33-48.

Sormunen, E. Liberal relevance criteria of TREC – counting on negligible documents? In: Proceedings of the SIGIR´02. ACM, New York (2002a) 324-330.

Sormunen, E. A retrospective evaluation method for exact-match and best-match queries applying an interactive query performance analyser. In: F. Crestani & al. (Eds.) Advances in Information Retrieval. Proceedings of the 24th European Colloquium on IR Research. Berlin & Heidelberg: Springer (2002b) 334-352.

Sormunen, E. & Kekäläinen, J. & Koivisto, J. Järvelin, K. Document text characteristics affect the ranking of the most relevant documents by expanded structured queries. Journal of Documentation. 57 (3) (2001) 358-376.

Spink, A. Term relevance feedback and mediated database searching. Information Processing & Management 31(2) (1995) 161-171.

Vakkari, P. Subject knowledge, source of terms and term selection in query expansion. An analytical study. In: F. Crestani & al. (Eds.) Advances in Information Retrieval. Proceedings of the 24th European Colloquium on IR Research. Berlin & Heidelberg: Springer (2002) 110-123.