The impact of metadata on the accuracy of automated patent classification

Georg Richter[1], Andrew MacFarlane[2]

[1]Senior Database Designer. Thomson Scientific, Middlesex House, 34-42 Cleveland Street, London W1T 4JE, United Kingdom. georg.richter@thomson.com (to whom correspondence should be addressed.

[2] Centre for Interactive Systems Research, Dept. Of Information Science, City University, Northampton Square, London EC1V 0HB, United Kingdom

Abstract

During the last decade, the advance of machine learning tools and algorithms has resulted in tremendous progress in the automated classification of documents. However, many classifiers base their classification decisions solely on document text and ignore metadata (such as authors, publication date, and author affiliation). In this project, automated classifiers using the k-Nearest Neighbour algorithm were developed for the classification of patents into two different classification systems. Those using metadata (in this case inventor names, applicant names and International Patent Classification codes) were compared with those ignoring it. The use of metadata could significantly improve the classification of patents with one classification system, improving classification accuracy from 70.8 up to 75.4 percent, which was highly statistically significant. However, the results for the other classification system were inconclusive: while metadata could improve the quality of the classifier for some experiments (recall increased from 66.0 to 68.9 percent, which was a small but nonetheless significant improvement), experiments with different parameters showed that it could also lead to a deterioration of quality (recall dropping as low as 61.0%). The study shows that metadata can play an extremely useful role in the classification of patents. Nonetheless, it must not be used indiscriminately but only after careful evaluation of its usefulness.

Keywords: Automated classification, Metadata, Inventors, International Patent Classification, Bibliographic data; Classifier committee, Patent classification

1. The development of automated classification tools in the area of patent information

Automated classification of documents has generated a lot of research interest over the last few years. The huge increase in the number of electronically available documents during this period has made intellectual classification and indexing increasingly difficult and costly. For organisations with an interest in the storage, handling and retrieval of documents, automated classification tools are often seen as a useful remedy against the explosion of costs arising from intellectual document indexing and classification.

Patents and patent applications are a typical example of this scenario. There was a six-fold increase in the number of PCT applications between 1990 and 2001 (see Figure 1). Not only patent offices but also commercial patent information providers are struggling to come to terms with the volume of information published in patents. For patent offices, the primary classification task is the association of International Patent Classification codes as well as national or European classifications. These classification systems are too fine to realistically achieve sufficiently high accuracy by automated classifiers so some efforts have focused on using them for the preclassification stage, where patent applications are associated to the appropriate technical unit for the examination phase [4, 5]. Texts of patents are widely and freely available on the world wide web, making patents ideal subjects for automated classification. Consequently, various publications have described attempts of automated patent classification over the last years [2,4,5,6,15,18,19]. Most of these have originated from patent offices, possibly due to the confidentiality that is often applied by commercial organisations with regards to their research .

In addition to that, commercial patent information providers often apply a multitude of indexing and classification systems to patents. This study was undertaken within a larger project to evaluate the merits of automated classification for various commercial patent alerting products by Thomson Scientific. The outcome of the overall study clearly showed that automatic classification still falls short of the standards that can be expected from intellectual classification so that, at best, automatic classification can be used to support the intellectual classification process but not to replace it. While the general trend in the information and communication industries clearly points towards linguistic and statistical automation of document processing, the tools which are currently available do not merit abandoning intellectual processes.

This paper describes the investigation of the usefulness of metadata for automatic classification. It is the abbreviated version of a dissertation prepared in the context of an MSc course at City University, London. For a much more detailed account of the study, the original dissertation [11], is available from the author.

Metadata is commonly described as "data about data". It is usually distinguished from the data itself through suitable mark-up or by being stored in a physically different location. In the case of textual data, common types of metadata would be the author, the date of publication or keywords and descriptors. An important difference between (textual) data and metadata on the linguistic level is the fact that metadata elements become semantically meaningful through the "field" in which they occur, whereas textual data elements obtain their meaning through linguistic rules and their position within their context. Metadata is usually considered as being structured, whereas the body of text that it describes is unstructured. However, many common automatic classification algorithms adopt the so called "bag of words" approach, where classification decisions are made solely on the basis of statistical occurrences of words in certain classes without considering the semantic context of each word or the syntax of the text. For classifiers adopting this approach, the distinction between data and metadata becomes less important.

Nonetheless, many document classification systems apply their classifications only on the basis of the document text. In fact, a fundamental review of the area [14] is explicitly limited to the description of systems where “metadata, such as, for example, publication date, document type, publication source, etc.; is not assumed to be available.” This trend can also be observed in the field of patent classification. For example, Larkey [6] describes an automatic patent categorisation and querying system that only uses the title, abstract, the first twenty lines of the background summary and the exemplary claims as basis for classification (in other words, only textual data). It is stated that this selection “has shown the best performance”, but no details of this performance test are disclosed. While it is stated that “about 50” out of “hundreds of fields” were used by the system, there is no further mention of any specific field, except the aforementioned ones. No mention of the potential contribution of metadata to the overall classifier is made.

The European Patent Office (EPO) has been developing automatic methods for patent categorisation, based on a k-Nearest Neighbour algorithm. To date, the only reported results have been based on the abstract or the abstract plus the full text of the document. In a recent paper about this project [5] , the authors postulate that the addition of bibliographic data (alongside extracted citations) is likely to have a positive impact on the correct classification, but no concrete efforts to incorporate these data have been reported to date.
In this context, the EPO has also held an open competition to assess the state of the art in the field. Of the participating teams, the one with the best results [4] used a classifier based on the Winnow algorithm, and achieved remarkably good results. However, again there was no mention of the use of metadata or any investigation about its actual contribution. Other patent classification studies also do not seem to have incorporated metadata into their classifiers [15, 18].

Another system concerned with patent classification did make use of structured metadata, but only in the form of cited patents [2]. Thus, to enhance and reinforce the categorisation decision, patents cited by this patent as well as patents that cite this patent (both designated as “neighbours”) were also considered by the classifier. This led to a reduction of misclassifications from 36% to 21%.

Why is metadata not commonly utilised for automatic classification ? Some of the most apparent reasons are:

- Especially when compared with patents, most documents do not contain much metadata. Therefore, its use seems to add little value and it is not widely established in the research community.

- In academic settings, automatic classification is often investigated not as an end in its own right but to test and demonstrate the effectiveness of machine learning methods. If text categorisation is conducted for this purpose, optimisation of classifiers through inclusion of metadata only deflects attention from the principal, machine learning issues.

- Commercially available classification software is commonly based on linguistics and cannot easily incorporate linguistic information. Furthermore, the main

business driving this industry is the categorisation of short text documents, such as E-Mails or news stories.

- Human classifiers would usually not use metadata, such as author names, text length or publication year to classify a text into content-specific classes. It is an easy mistake to transfer "human ways of thinking" to software tools and computers. However, especially "bag of words" classifiers do not work by attempting to linguistically analyse the content of a document, but by being "trained" on an intellectually classified set of documents (usually referred to as the "training set"), analysing the relationship between individual terms and intellectually associated classes to deduce rules that are then applied to classify unclassified documents. Whereas a human indexer is unlikely to "memorise", for example, inventor names and their commonly associated classes on a great scale, for a computerised system, it is just as easy to analyse the relationship between author names and classes as it is to do the same with keywords.

Patents are not only very rich in metadata, compared with other documents, there is also a strong likelihood that metadata can give a good indication about the content of the patent. The most obvious example of this are the classifications that are applied to patents by the patent offices, which directly reflect the content. Of course, this category of metadata is only available after the application has been processed by the patent office and can obviously not be used by the office itself to classify the patent in the first place. Thus, it can only be used by secondary patent information providers. However, there are plenty of potentially useful metadata attributes that are available at the time of filing. For example, inventor names and applicant names carry a lot of meaning about the content of the document since companies as well as specific inventor teams tend to file series of applications for inventions that are closely related to each other. Thus, if an inventor or applicant of an unclassified patent appears in the training set, it can be a powerful indicator of the class to which the unclassified document belongs. Other document attributes that can conceivably give some indication to which class a patent belongs include, for example, cited documents, name of the patent attorney, document length, number of claims, number of illustrations or priority country. It is potentially useful to include them in the automated classification procedure. This study examines whether some particular parts of metadata (inventors, applicants and IPC codes) can increase the quality of automated classification.

2. Description of this study

The classification systems used in these study are two independent classification systems used by Thomson Scientific in two of its patent information products, Investigational Drugs Patent Fast Alert and the Current Patents Gazette.
Thus, each patent is classified twice, once for each classification scheme. At present, these classifications are assigned intellectually. They are important for internal processing and appear in some patent information products published by the company.

2.1 The classification systems

Both classification systems are very coarse (i.e. there is only a very small number of categories). The first system is designated as "Gazette classification" because it classifies patents into categories for the "Current Patents Gazette", which is a weekly alerting service that summarises newly published patents in the areas of pharmaceuticals and biotechnology. It comprises the following classes:

A	New compounds
B	New uses, formulations and methods of treatments
C	Chemical processes and combinatorial technology
D	Biotechnology
E	Devices and Equipment
F	Electrotherapy and other non-chemical treatments

The second classification system is designated as "PFA classification" because it classifies patents into categories used by the "Patent Fast Alert (PFA)" product. Here, the classes are based on the therapeutic application of the invention, rather than the type of technology:

AC	Oncologic, Endocrine & Metabolic Patents
AG	Pulmonary-allergy, Dermatological, Gastrointestinal &

	Anti-inflammatory Patents
AM	Anti-infective Patents
BT	Biological & Immunological Patents
CN	Central & Peripheral Nervous System Patents
CV	Cardiovascular & Renal Patents

A fundamental difference between the classification systems is that only one class from the Gazette classification can be assigned to each patent, whereas several categories from the PFA classification can be assigned to a single patent. Note also that both classification systems have an additional implicit class, consisting of "rejected" records. These are records that match the initial seletion criteria but are intellectually rejected as being outside the scope of the information products. This class is particular significant as, from a business perspective, false rejections of records are more critical errors than records that were classified into the wrong class.

2.2 Methodology

2.2.1 The classifier

The automatic classification was performed with k-Nearest Neighbour classifier developed for this particular study. k-Nearest Neighbour classifiers belong to the machine-learning discipline of “memory based reasoning”, which was introduced by Stanfill and Waltz [16]. These authors argued convincingly that machine learning methods based entirely on the induction, creation and subsequent application of rules are flawed because “we consider the phenomenon of reasoning from memories from specific episodes ... to be the foundation of an intelligent system, rather than an adjunct to some other reasoning method” and “it is difficult to conceive thought without memory”. Thus, memory based reasoning algorithms simply work by looking at similar problems or entities from the past to solve a new problem or to process a new entity. The initial problem to be solved was to deduct the pronunciation of a word from its spelling and they created a system that could compute the pronunciation of a word by comparing it with similarly spelled words [16]. This concept was successfully applied to text categorisation to classify free-text fields on American census forms [3] and to classify records from the MEDLINE database [20].

Thus, k-Nearest Neighbour classifiers do not contain an explicit learning step but take an unclassified document, identify *k* classified documents that are most similar to it (where *k* is set to an arbitrary or empirically optimised value) and apply the same class (or classes) to the document, to which these nearest neighbours belong. In this study, similarity is defined as the cosine value of the vectorised new document and its vectorised neighbour:

$$sim(D, D_j) = \frac{\sum_{i=1}^{n} t_i \in (D \cap D_j)\ d_i \times d_{ij}}{\| D \|_2 \times \| Dj \|_2}$$

where the variables have the following meanings:
D is the document to be classified
D_j is the document to which similarity is to be determined
t is a word (in this case a word shared by both documents)
n is the total number of distinct words shared by both documents
$||D||_2$ (and $||D_j||_2$ respectively) is the "norm" of document, a normalisation factor required to prevent long documents being chosen over short documents, simply on the basis of having many words :

$\| D \|_2 = \sqrt{\sum_{i=1}^{n} d_i^2}$ (where n is the total number of distinct terms in the document).

(see also [20]).
d_i and d_{ij} are the weights of the word t_i in documents *D* and D_j, respectively (the "weights" of words indicate their significance for the particular document).
The most common term weight function in the field of information retrieval and automated text classification is undoubtedly *term frequency/inverse document frequency (tfidf)*, which was developed by Salton and Buckley [15] as

$$tfidf(t_k, d_j) = \#(t_k, d_j) \times \log \frac{|Tr|}{\#_{Tr}(t_k)}$$

Here, $\#(t_k,d_j)$ denotes the number of times term t_k occurs in document d_j *(term frequency)* and $\frac{|Tr|}{\#_{Tr}(t_k)}$ represents the total number of documents divided by those in

which t_k occurs *(inverse document frequency)*. Inverse document frequency is based on the intuition that "the more documents a term occurs in, the less discriminating it is" [14].

In practice, it has been found that the term frequency component of the above formula becomes too dominant for terms that occur with high frequency within a document. Therefore, it is commonly modified to a non-linear function where the slope levels off with increasing term frequency. Several variations were tested for a k-Nearest Neighbour classifier [21] and the "ltc" variant known from the SMART retrieval system [12] showed the best performance. Ltc uses $1 + log_2 \#(t_k,d_j)$ as term frequency component [1].

For overviews of other text classification algorithms, the dissertation on which this paper is based [11], as well as papers by Sebastiani [14] and Yang [22] are recommended.

2.2.2 Test set and training set

In this study, the k-Nearest Neighbour algorithm was implemented in a relational database (using Oracle) that contains more than 40,000 intellectually classified patent records. These records include all PCT applications published in the years 2001 and 2002 to which the pre-selection criteria, used to identify potential candidates for inclusion in the patent alerting services, apply.

These database records have been divided into two sets, a training set, containing the documents whose classes are known to the classifier, and a test set, containing documents whose classes the classifier has to determine (by finding the nearest neighbours from the training set). By comparing the assigned classes to the intellectually applied classes, quality indicators, such as precision and recall, can be obtained. These were used to compare the results of various classification experiments, in particular those using only text and those using additional metadata, with each other.

2.2.3 Types of metadata used in this study

As outlined above, there is a considerable number of metadata types that can conceivably contribute to the accuracy of automated classification. In this study, only three attributes were considered: IPC codes, inventors and applicants. For the purpose of this classifier, inventors and applicants for each record are considered as one single combined attribute. This means that the resulting data is three-dimensional, and therefore far easier to visualise and interpret than four-dimensional data. It also reduces errors caused by polysemous inventors (different individuals with similar names).

The only problem with this approach is the question of how to calculate the term weights for IPC classifications and for inventors. With regard to inventors, there is no reason to believe that less frequent inventors are somehow more significant than more frequent inventors. With regard to IPCs, it has been pointed out that *tfidf* is generally less, or not at all, applicable to controlled vocabulary [8].

Therefore, to create weights for inventors and IPCs, advantage was taken of the fact that term weights are normalised by dividing each weight through the document norm and, consequently, $\sqrt{\sum_{i=1}^{n} d_i^2}$ always has a value of one (see also the previous section). IPC and inventor weights were created with the notion that, for each set of attributes, the sum of its squared term weights should also have the value of one. This method of weighting inventors and IPCs has the advantage that, initially, the text component of a patent record has the same weight as all combined inventors and the same weight as all combined IPCs. During classification, a factor can be applied to inventor data (and IPC data) to adjust the relative weight of an attribute type upwards or downwards. Furthermore, the same method can easily be applied to different document domains and to other types of metadata.

For inventors, the resulting weights were distributed equally, so that for a given patent, the same weight value was assigned to each inventor. However, international rules for applying IPCs to a patent state that they have to be subdivided into one primary classification (which is considered to be most important) and, optionally, one or more secondary classifications. To use this intellectually assigned weighting, primary IPCs were weighted twice as high as secondary IPCs but, like with inventors, it was ensured that the sum of all squared IPC weights was 1.

Therefore, inventor weights (d_{inv}) were calculated as

$$d_{inv} = \sqrt{\frac{1}{\#inv}}$$

where *#inv* is the total number of inventors of the patent and IPC weights (d_{IPC}) were calculated as

$$d_{ipc} = \sqrt{\frac{1}{\#ipc + 1}}$$

(or twice this amount for primary IPCs, as outlined above).

2.2.4 Threshold values

The initial result of a k-nearest neighbour classifier is not a particular category or a set of categories that can be assigned to the patent in question. It is a matrix assigning to each patent / category pair a value which is calculated by adding together the similarity scores between the patent and its nearest neighbours that belonged to the particular category. Each value is assumed to be roughly proportional to the likelihood of the respective patent belonging to the respective category. For the Gazette classification, the final assignment of the category to a patent is easy since with this classification scheme, each patent belongs to exactly one category, so the obvious method of assignment is to choose the category for which the patent received the highest score.

For the PFA classification, the assignment is much more difficult, as the number of categories per patent varies. Therefore, it is necessary to identify some kind of threshold value that determines whether a patent / category score is large enough to justify assignment of the patent to the category. This has to be handled carefully as "the thresholding method used in a categorization system … can influence its results significantly" [23].

The method used here is the so-called PCut algorithm [9], based on the assumption that the overall distribution of categories in the training set is similar to that of the test set (or the set that needs to be classified). Thus, for each category, a number of documents that have to be classified (in our case the test set) are ranked by score and the category is assigned to the *n* top ranked documents. *n* is category-specific and proportional to this

category's share of documents in the training set. The obvious problem with PCut is that the number of documents for each class is defined by the class distribution of the training set. If the class distributions differ between the training set and the test set, the classifier has an inborn minimum error rate. Furthermore, PCut is only applicable to domains where a considerable number of documents is classified at the same time (which is the case for this application, as the documents are supplied on a weekly basis, each weekly intake comprising several hundred patent records). It cannot be used for applications where each document has to be classified independently from other documents. The suitability of PCut for the domain described in this paper has been evaluated in detail and it was concluded that it compares well to other algorithm. A more detailed discussion is provided in the underlying dissertation [14].

3. Classification results for the Gazette classification

3.1 Results for single classifiers based on a combination of attributes

3.1.1 Defining the baseline

To obtain meaningful results, it is important to define a baseline, i.e. to determine the classification accuracy for a classifier based only on text. This is the benchmark against which the more sophisticated classifiers have to be compared. Therefore, an experiment with a classifier based only on text was conducted first. The resulting classification accuracy was 70.8%. To compare the "goodness" of text with the goodness of the other two attributes, classifiers solely based on IPC and inventors, respectively, were also tested and the share of correctly determined classes was 70.0% for IPCs and 31.2% for inventors. Both of these figures are expressed as percentage of *all* test set documents, even though classifiers based on these attributes only work on the subset of the test documents that have actually an IPC or inventor, respectively, in common with at least one document of the training set. For IPCs, this restriction is almost negligible, as only 111 patents (1.54% of the entire test set) do not have an IPC in common with any training set document. For inventors, it is very significant: here, 4010 (55.79%) test set patents do not have an inventor in common with training set patents. Expressed as percentage of the eligible documents only, the accuracy of the IPC-based classifier was 71.0% (making it slightly superior to the text-based classifier) and that of the inventor-based classifier was 70.4%. These "baseline results" are summarised in table 1.

3.1.2 Results for single classifiers based on a combination of attributes

In the next set of experiments, it was investigated how classifiers, that are based on two or three attributes (without applying particular weighting), perform in comparison with the text-only based classifier. Table 2 summarises the results. The first conclusion is that the inclusion of metadata in the classification process does indeed improve classification accuracy. All combination classifiers performed better than single attribute-based classifiers and the classifier that took all three attributes into account performed best, with an accuracy of 75.1%, corresponding to an increase of over 6% over the text-only classifier. If one considers the classification trials as statistical experiment, one finds that this improvement in classification accuracy is highly statistically significant ($p < 0.001$). Thus, it is unlikely that the improvement in accuracy has only occurred by chance.

It has now been established that all three attributes (text, IPC and inventor data) can contribute to the quality of the classifier. As described in the previous section, the classifier based on all three attributes outperforms all classifiers based on two or less.

However, to optimise this classifier further, experiments with different relative attribute weights were conducted. More information on the weighting process can be obtained from the underlying dissertation [11]. In principle, text weights have been kept constant to the relative value of 1 and only the other attribute weights were changed.
A total of 400 experiments were conducted for the Gazette classification, with IPC and inventor weights independently taking all values of $n / 5$ (with $n = \{1, 2, 3, \dots, 20\}$. The accuracy of each classifier was calculated and the results were plotted in a three dimensional chart (Figure 2).

Classification accuracy peaks, with a value of 75.4%, at the relative weights of 1 for IPCs and 3.6 for inventors. Accuracy decreases steadily with increasing deviation from these optimal weights with a surprisingly steep decline between IPC weights 0.4 and 0.2. Unsurprisingly, the worst classifier was the one where inventor and IPC weights deviated strongest from the optimum, with an accuracy of 72.9%, at the relative weights

of 4 for IPC and 0.2 for inventor. However, this classifier was still superior to any single-attribute based classifier, including the text-only based classifier.

To visualise the impact of varying inventor and IPC weights independently from each other, the average of classification accuracy for each distinct IPC and inventor weight, respectively, was plotted against the weights. These illustrations are presented in Figures 3 and 4.

It can be seen that, on the basis of average classification accuracy, the optimal inventor weight is even higher than 3.6, with the best average being 74.445% at the relative weight of 4, compared with 74.44% at 3.8 and "only" 74.425% at 3.6 (the value with the highest individual combination of IPC and inventor weights). At weights below 3.6, accuracy decreases steadily and, towards zero, more and more sharply.

For IPC weights (Figure 4), there is a significant peak between the weights 0.4 and 1, with a very sharp decline to 0.2 and another quite sharp decline for increasing values above 1. The "window" of good performance is much narrower for IPC weights than for inventor weights.

3.2 Classification committees based on metadata-specific and text-specific classifiers

Another approach for using both text and metadata as basis for a single classifier consists in the creation of independent classifiers, based on text, IPC and inventor data, to classify the records and form a "committee" where the independent classifiers "vote" on the best class. The concept of committees of independent classifiers has already been widely explored with mixed results [see for example 7, 10, 14].

To explore the usefulness of this approach, the three experiments using only text, only IPC and only inventor data for classification were reviewed and an illustration of the overlap between the automated classification results is presented in figure 5.

This figure shows how many records were correctly classified by all, two, one or no classifier and by which particular combination of classifiers. For example, 21.6% of

patents were correctly classified by all three classifiers, 4.4% were correctly classified by the inventor- and text-based classifiers but not the IPC-based classifier, 11.1% were correctly classified by the IPC-based classifier but none of the other classifiers and so on (note that 12.9% of patents were misclassified by every single classifier).

If the committee approach would require at least two classifiers to agree on a class, only 63.2% of records would be correctly classified, but usually, the rules would stipulate to choose arbitrarily the output of a single classifier for classifying those records where no agreement between at least two classifiers is obtained. From all patents that were correctly classified by only one classifier, the biggest share was correctly classified by the IPC-based classifier (11.1% of the total, compared with 10.5% for the text-based classifier and 2.3% for the inventor classifier). This suggests that, if there is no agreement between at least two classifiers, the class suggested by the IPC classifier should be selected. However, it should be taken into account that in some of these cases, the correct classifier would be overruled because the two other classifiers actually agree on a (wrong) class. The numbers in brackets indicate (as percentage of the total document set) for how many documents this would happen for the respective subset. For example, 2.1% of all documents were correctly classified only by the IPC-based classifier but the agreement of the inventor- and text-based classifiers on a different (wrong) class would overrule the correct result. If this is taken into account, it is better to choose the text-based classifier as final instance if no agreement can be reached, as this would increase the number of correct outcomes by 9.3 percentage points, compared with 9.0 percentage points for the IPC classifier.

However, even this strategy can be further improved. It should be considered that there are 111 patents (1.54%) for which the IPC is not applicable (as there is no common IPC with a training set document). Therefore, an alternative strategy is to choose the IPC-based classifier as final instance if no agreement is reached but to use the text-based classifier's verdict if the IPC-based classifier has no opinion. This approach will finally bring classification accuracy up to 72.9% (this cannot be deducted directly from Figure 5 but was calculated from the original experimental results).

3.3 Classification committees based on combination classifiers

In an additional experiment, a classifier committee was formed not by the classifiers based on IPC, inventor and text data, but by the classifiers based on the combination of IPC/inventors, IPC/text and inventors/text (in each case with the same weighting for both components). This approach is not optimal, as these classifiers are less independent of each other than those which are based only on one attribute. Therefore, it can be expected that the correlation between these classifiers is higher (i.e. more patents are classified identically). It has been shown that classifier committees perform best when the correlation between them is low [17]. However, on the other hand, the individual classifiers based on a combination of two attributes performed much better than the classifiers based on only one attribute (see Table 2). Therefore, it is conceivable that the structural weakness (i.e. decreased classifier independence) is compensated by the increase in quality of the individual classifiers (this potential trade-off has also been recognised in the earlier study [17].

In a similar style to Figure 5, the overlap between the three combination classifiers is shown in Figure 6. The illustration confirms the assumptions which were made on the basis of [17], in particular:

- The overlap between the three classifiers is much higher. The total number of cases where at least two classifiers agreed is at least 84.9% (56.7 + 11.4 + 1.2 + 5.3 + 8.0 + 1.6 + 0.7, plus an additional share of the 13.9 percent where all classifiers were wrong). The corresponding figure for the previous experiment is 67.9%.

- The amount of patents that were correctly classified by all three classifiers was far more than twice as high as that for the previous experiment (56.7% compared with 21.6%), reflecting the superiority of these individual classifiers over those used in the previous experiment.

- Unsurprisingly, the downside was that, if only one classifier calculated the correct result, it was very likely that the other two would agree on the wrong result and overrule the correct classifier. In particular, the text + inventor-based classifier was the sole correct classifier in 8.4% of all cases, but was overruled in almost all of them (8.0%).

- Furthermore, the amount of patents that were correctly classified by none of the classifiers was higher than in the previous experiment (as could be expected), but only slightly (13.9%, compared with 12.9%).

- Most importantly, assumed that the text + inventor-based classifier casts the deciding vote if there is no agreement between at least two committee members, the total amount of accurately classified documents is 75.0%, which is only marginally worse (and not statistically significant) than the best result obtained by a single classifier based on a combination of three attributes. It is conceivable that this could be improved by a more sophisticated voting system, for example one of those described by Li and Jain [10].

It should be emphasised that, in the business context of this application, even accuracy values as high as 80 or 90 percent would not be sufficient to have documents exclusively classified by the automated classifier, without human intervention. The main purpose of the automated classifier can only be to assist and cross-check intellectual classification.

In the experiments described in the previous sections, it has been shown that the use of metadata for automatic classification can improve classification accuracy with high statistical significance for the Gazette classification system. It is clear that this is not necessarily true if a different classification scheme would be used. Other schemes may assign categories in a way that is completely independent of inventor or IPC data. More light on the question of the general usefulness of including metadata in automated classifiers is shed in the next section, where the same approach is used for the PFA classification, a different classification system.

4. Classification results for PFA classification

4.1 Indicators of classification quality

The Gazette classification system, for which the results were analysed in section 4, assigns exactly one class to each record. Therefore, classification accuracy (the share of all correct associations out of all associations made by the classifier) is sufficient to

evaluate the "goodness" of the classifier. However, in the PFA classification system, it is possible to assign several classes to the same record and no "simple" parameter, like accuracy, exists. For example, the classifier could assign a patent belonging to classes AC and AM only to AC, which is neither completely right nor completely wrong. Therefore, the results are more complex than those for the Gazette classification, making them somewhat more difficult to analyse.

During the experiments, three parameters are generated to evaluate the quality of a classifier: recall, precision (micro-averaged) as well as utility (which counts the number of misclassifications but weighs those that lead to wrong rejection of a relevant patent higher, to reflect the "greater damage" caused by this misclassification in contrast to other misclassifications). It has been found that the correlation between these parameters was extremely high (see Figures 7 and 8) so that it was acceptable to simplify the result by looking at one parameter only (in this case, recall was chosen).

4.2 Baseline results and statistical significance of the results obtained with combination classifiers

In similar fashion to the Gazette classification experiments, baseline studies with a single attribute classifiers and unweighted combination classifiers were conducted and the results are presented in Table 3.

When comparing these values with those for the Gazette classification (Table 2), the most important observation is that the combination classifiers do not offer much benefit over the text-only based classifier. This seems to be mainly due to the poor contribution made by IPCs to classifier quality. All classifiers using IPC values perform worse than the corresponding classifier that does not use IPC values. Thus, it seems that IPCs are not very useful for determining the PFA classification of a patent. In fact, the only combination classifier that achieved a slight (not statistically significant!) increase in recall over the text-only based classifier is the one based on text and inventors.

Even the inventor-only based classifier performs considerably worse than the text-only based classifier. This is in stark contrast to the Gazette classification, where IPC-only, inventor-only and text-only based classifiers performed almost equally well. The

preliminary conclusion is that, for the PFA classification, metadata does not necessarily increase the performance of the classifier.

Like for Gazette classification, the next step consisted of conducting a plurality of experiments to investigate how different combinations of attribute weights perform. Again, 400 experiments were conducted, with a constant text weight of 1 and IPC and inventor weights independently looping through all values of n / 5 (with n = {1, 2, 3, … , 20}). The recall of each classifier was calculated and the results were plotted in a three dimensional chart (Figure 9). Note that the axis representing the IPC weights has been inverted because otherwise the high recall values for low IPC weights would have completely obscured the rest of the chart.

The chart shows that the best performing classifier was the one where the inventor weight was set to 4 and the IPC weight was set to 0.4. This classifier achieved a recall of 68.9%, which is statistically highly significant ($p < 0.01$), even though the nominal increase in recall over the text-only based classifier is still quite small (4.4 percent). Given the apparent detrimental impact of IPC values on classification accuracy outlined above (and it can be easily deducted from the illustration that, overall, IPC weight is inversely proportional to recall), it is surprising that the best performing classifier has the non-minimum IPC weight of 0.4 and that the lower IPC value of 0.2 reduces recall, rather than increasing it further. It can also be seen that, for low inventor weights, an IPC weight of 0.2 performs better than a value of 0.4 but Figure 10 shows that, on average, classifier performance peaks at an IPC weight of 0.4.

It should also be noted that recall is strongly inversely proportional to IPC weights when they vary between 0.6 and 2.0, but outside these boundaries the impact of varying IPC weights on recall is less dramatic (although, especially for higher values, there is still a clear inverse correlation). The most likely explanation for this is the fact that once a quite high or quite low weight is reached for a particular attribute, it becomes very dominant or very obscured, so that its impact either obliterates that of other attributes or is obliterated by other attributes. In these cases, increasing or decreasing the weight even further does not have so much impact on overall class assignments anymore. This assumption is also confirmed by figures 3 and 4 for the Gazette classification, where it can be seen that the "steepest slopes" in the graphs plotting the impact of these attributes on classification accuracy occur between values of 0.5 and 2.

The impact of inventor weights on classifier performance is illustrated in figure 11. It is interesting that even though the inventor-only based classifier performs considerably worse than the text-only based classifier, the best-performing combination classifier has a maximum inventor weight of 4 (further experiments showed that similar results could be obtained for an inventor weight of 4.2 but recall decreased again for even higher inventor weights). Having said that, the overall impact of varying inventor weights on recall performance appears to be relatively low. Different inventor weights have little or no impact on recall if IPC and text weights are kept constant. The fact that the best result was achieved with a high inventor weight is more likely to have occurred by chance than to be particularly meaningful.

The low impact of inventor weight on recall performance can be partially explained by the fact that inventor weights only affect the subset of test set patents that have one or more inventors in common with the training set. However, the impact of inventor weight on classification accuracy is much more pronounced in the Gazette classification (see figures 2 and 3), so this cannot be the only explanation for this phenomenon.

Overall, it can be seen that, for the PFA classification, it is much more questionable whether the inclusion of metadata offers real improvements for the quality of automated classification. While statistically significant improvements in classification quality were achieved for some attribute weight combinations, the magnitude of these achievements was quite small. Even more important is the factor that out of 400 experiments conducted with different attribute weight combinations, only 95 showed any improvement over the text-based classifier, the 305 remaining classifiers actually performed worse. The two worst performing classifiers (with an IPC weight of 4 and inventor weights of 0.2 and 0.4) showed a recall of only 61.0%. In stark contrast, for the Gazette classification, every single classifier based on three attributes was clearly superior over the text-only based classifier ($p < 0.001$ for all of them). The main conclusion to draw from both experiments is the confirmation of the assumption that empirically derived findings cannot necessarily be applied to different classification schemes, even though they are both based on content and use the same data set.

5. Conclusions

The main purpose of this project was to investigate whether the use of metadata by automated classifiers, as opposed to the exclusive use of text only, can improve classification quality. It should be emphasised again that limited empirical experiments like those described in this article cannot provide the ultimate answer, due to the heterogeneity of classifiers (and their parameters), classification systems and document sets. In fact, there can be no ultimate answer because something useful for one classification system may not be useful for another. This view was also confirmed by the experiments, showing that the use of metadata significantly and consistently improves automated classifiers for the Gazette classification whereas it is far from clear whether it is beneficial for PFA classification.

It is difficult to make a definite statement on the causes for this different behaviour of different classification systems. The most obvious explanation is that metadata (in comparison with text) is more useful to classify patents for the Gazette classification than it is for classifying patents for the PFA classification. Investigating this hypothesis beyond the experiments described here would be very difficult and not within the scope of this study. It is certainly true, however, that there exist certain keywords (in particular those describing illnesses) that, even when taken in isolation, allow almost unambiguous assignment of a PFA class to a patent record. For example, the term "tumour" is an almost certain indicator of the category AC. For the Gazette, there are no such unambiguous keywords but there are (albeit only a few) IPC classifications that indicate certain classes with a likelihood of 99 percent or more.

Inventors, too, seem to be more useful for the Gazette than for the PFA classification. One possible conclusion is that it is apparently more common for inventors to work "across boundaries" of therapeutic areas (for example to be involved in cardiovascular as well as neurological and/or cancer research) than "across boundaries" of technology (for example process chemistry, formulation technology and/or biotechnology).

In any case, and this may well be the single most practical conclusion of this study, the experiments conducted with the Gazette classification show that the use of metadata is something that should be seriously considered for any classifier. Ignoring it means ignoring a potentially useful and significant optimisation tool. However, the experiments conducted with the PFA classification show that this should not be done

without critical investigation of the benefits. Even though metadata could improve classification quality for the PFA classification system with some attribute weight combinations, more than three quarter of metadata / text weight combinations led to a deterioration of classification quality. Thus, it is very important not to believe that any form of metadata will improve the classifier, but to carefully adjust the weight of metadata and metadata components in relation to text.

It should also be noted that only two types of metadata were used in this study: inventor/assignee combination and IPCs. As described in section 1, many automated patent classification systems (in particular those developed by patent offices) try to classify patents into the IPC system so in these cases, it is obviously not possible to use IPCs as the basis for the classifier. However, patent documents have several other types of metadata that can be potentially used by classifiers. Firstly, it is clearly possible to separate inventors from applicants, so a classifier could use these attributes independently from each other. It would certainly lead to more matches between test and training set documents (since the inventor/applicant combination is far more specific than individual inventors or applicants) but the matches would be potentially more misleading and the neighbours identified in this way would be less close. There exist other bibliographic details that are unique to patents, for example priority and application details, although these are unlikely to have any relation to the content (and the class) of the patent. One bibliographic detail that could be investigated is the legal representative (patent attorney) since patent attorneys usually specialise in certain technical fields.

However, the search for suitable metadata attributes should not stop with bibliographic details. Document attributes that could potentially give hints to the correct class include the length of the document, number of claims, number of illustrations, number of inventors, country of origin of the inventors, patent citations and so on. Including metadata like these in isolation would be unlikely to increase classification accuracy significantly, but taken in combination, they may lead to real improvements.

Thus, a single sentence answering the main question of the study could be: "It is definitely worth trying to make metadata part of the decision-making process of a classifier, but it should not be taken for granted that it leads to a real benefit and can make things worse if it is not carefully tested and implemented."

It cannot be emphasised too strongly even combination classifiers did not achieve sufficiently high accuracy to replace human classifiers with these machine learning tools. There are undoubtedly more sophisticated classifiers available, in particular those based on true linguistic analysis but experience shows that even those have a long way to go until human indexers and classifiers could be consigned to history.

References

[1] Chai, K.A.M., Ng,H.T. and Chieu,H.L. (2002) Bayesian online classifiers for text classification and filtering**.** In: M.Beaulieu (ed.), Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere 11-15 August 2002. New York: ACM. pp 97-104.

[2] Chakrabarti, S., Byron, D., Indyk, P. (1998) Enhanced hypertext categorization using hyperlinks. In: L.M.Haas and A.Tiwary (eds.), Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle 2-4 June 1998. New York: ACM, pp 307-318.

[3] Creecy, R., Masand, B., Smith, S. and Waltz, D. (1992) Trading MIPS and memory for knowledge engineering: Classifying census returns on the connection machine. Communications of the ACM, Vol 35 (8), pp 48-63.

[4] Koster, C.H.A., Seutter, M., and Beney, J. (2001) Classifying patent applications with Winnow. In: V.Hoste, G. De Pauw (eds.), Benelearn 2001: Proceedings of the 11th Dutch-Belgian Conference on Machine Learning, Antwerp 21 December 2001. Antwerp: CNTS, pp 19-26.

[5] Krier, M. and Zacca, F. (2002) Automatic categorisation applications at the European Patent Office. World Patent Information, Vol 24 (3), pp 187-196.

[6] Larkey, L.S (1999) A patent search and classification system. In: E.A. Fox, N. Rowe (eds.), Proceedings of the 4th ACM Conference on Digital Libraries, Berkeley 11-14 August 1999. New York: ACM, pp179-187.

[7] Larkey L.S. and Croft, W.B. (1996) Combining classifiers in text categorization. In: H.P.Frei (ed.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich 18-22 August 1996. New York: ACM, pp 289-297.

[8] Larson, R.R. (1992). Experiments in automatic Library of Congress Classification. Journal of the American Society for Information Science, Vol 43 (2), pp 130-148.

[9] Lewis, D.D. (1992) An evaluation of phrasal and clustered representations on a text categorization task. In: N.J. Belkin, P.Ingwersen and A.M.Pejtersen (eds.), Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen 21-24 June 1992. New York: ACM, pp 37-50.

[10] Li, Y.H and Jain A.K (1998) Classification of text documents. Computer Journal, Vol 41 (8), pp 537-546.

[11] Richter,G. (2003) Implementation of an automated system for the categorisation of patent applications and investigation of the suitability of metadata to improve categorisation quality. MSc Dissertation, Department of Information Science, City University,. London

[12] Salton, G. (1989) Automatic text processing: The transformation, analysis and retrieval of information by computer. Reading, PA, United States: Addison-Wesley.

[13] Salton, G. and Buckley, C. (1988) Term-weighting approaches in automatic text retrieval. Information Processing & Management, Vol 24 (4), pp 323-328.

[14] Sebastiani, F. (2002) Machine learning in automated text categorization. ACM Computing Surveys, Vol 34 (1), pp 1-47

[15] Smith, H. (2002) Automation of patent classification. World Patent Information, Vol 24 (4), pp 269-271.

[16] Stanfill, C. and Waltz, C. (1986) Toward memory-based reasoning. Communications of the ACM, Vol 29 (12), pp 1213-1228.

[17] Tumer, K. and Gosh, J. (1996) Error correlation and error reduction in ensemble classifiers. Connection Science, Vol 8 (3-4), pp 385-403.

[18] World Intellectual Property Organisation (2000). Presentation of the OWAKE ("Primary automatic classification system"). WIPO Report IPC/CE/29/11 on the 29th Session of the Committee of experts of the IPC Union. Section 59. Geneva: World Intellectual Property Organization. Available: http://www.wipo.org/classifications/en/ipc/ipc_ce/29/11.pdf [14-September-2003]

[19] World Intellectual Property Organisation (2002). Claims Project Status Report. Standing Committee on Information Technologies (SCIT). SCIT/Claims/PR.2 Available: http://www.wipo.org/scit/en/project/reports/2002/doc/scit_claims_pr2.doc [14-September-2003]

[20] Yang, Y. (1994) Expert network: effective and efficient learning from human decisions in text categorisation and retrieval. In: W.B. Croft and C.J. van Rijsbergen (eds.), Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin 3-6 July 1994**.** New York: ACM, pp13-22.

[21] Yang, Y. and Pedersen, J.O. (1997) A comparative study on feature selection in text categorization. In: D.H. Fisher (ed.) Proceedings of the 14th International Conference on Machine Learning, Nashville 8-12 July 1997. San Francisco: Morgan Kaufmann, pp 412-420.

[22] Yang, Y. and Liu, X. (1999) A re-examination of text categorization methods. In: M. Hearst, F. Gey and R. Tong (eds.) Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley 15-19 August 1999**.** New York: ACM, pp 42-49.

[23] Yang, Y. (2001) A study on thresholding strategies for text categorization. In: W.B. Croft, D.J. Harper, D.H. Kraft and J.Zobel (eds.) Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans 9-13 September 2001, New York: ACM, pp137-45.

Acknowledgements

The authors would like to thank Dr Stephen Robertson, Microsoft Corporation, for his friendly and generous support and advice on various aspects of this study, in particular term weight and term space reduction algorithm.

Figure captions

Figure 1 Number of annually published PCT applications between 1990 and 2002

Figure 2 Gazette classification accuracy for 400 classifiers with IPC and inventor weights (relative to text) varying between 0.2 and 4.

Figure 3 Average classification accuracy over inventor weight for the Gazette classifier

Figure 4 Average classification accuracy over IPC weight for the Gazette classifier

Figure 5 Overlap between the correct classification results for inventor, text and IPC-based classifiers. Each circle represents the correctly classified patents for one classifier and the percentages of documents (out of all documents) inside each subset (where each distinct overlap area is regarded as a distinct subset) are expressed by the numbers inside the different areas. The numbers in brackets indicate the percentage of documents (from the entire set) where the correct result of the classifiers would have been overruled by two identical, but wrong, results of the two other classifiers.

Figure 6 Overlap between the correct classification results for, text/IPC, text/ inventor and IPC/inventor -based classifiers. See Figure 5 for additional information.

Figure 7 Scatterplot of recall over precision for a range of PFA classification experiments (demonstrating that, using PCut algorithm, they correlate very strongly)

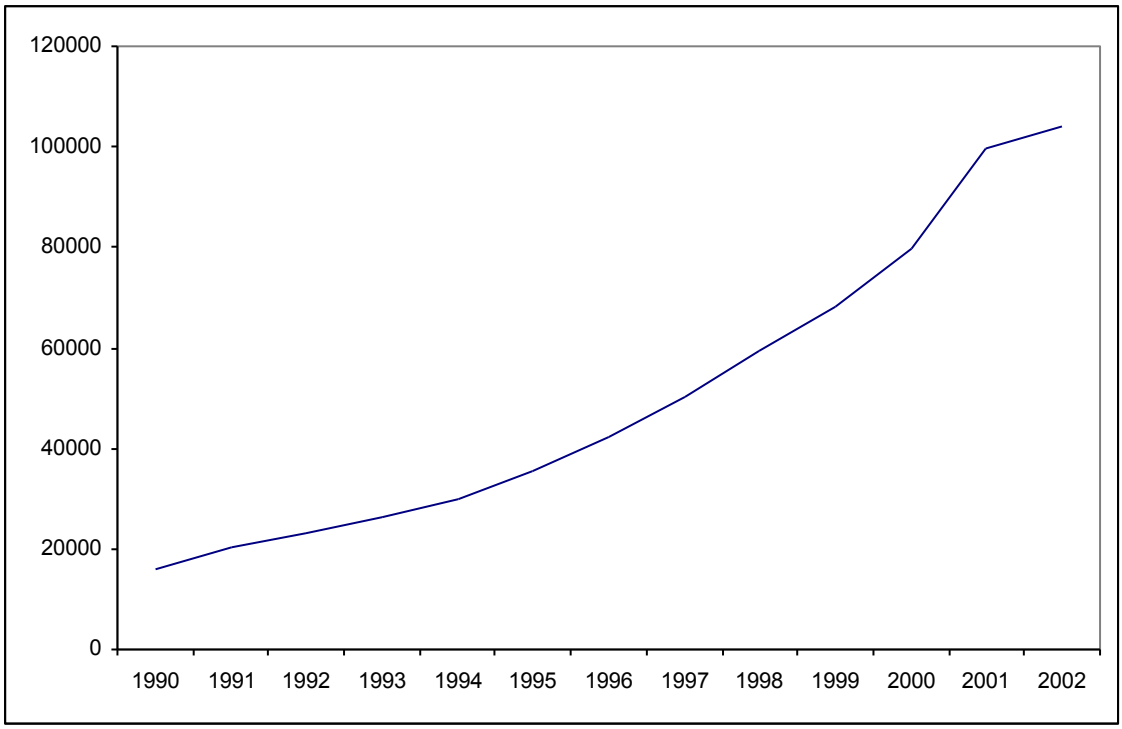
Figure 8 Scatterplot of recall over utility for PFA classification

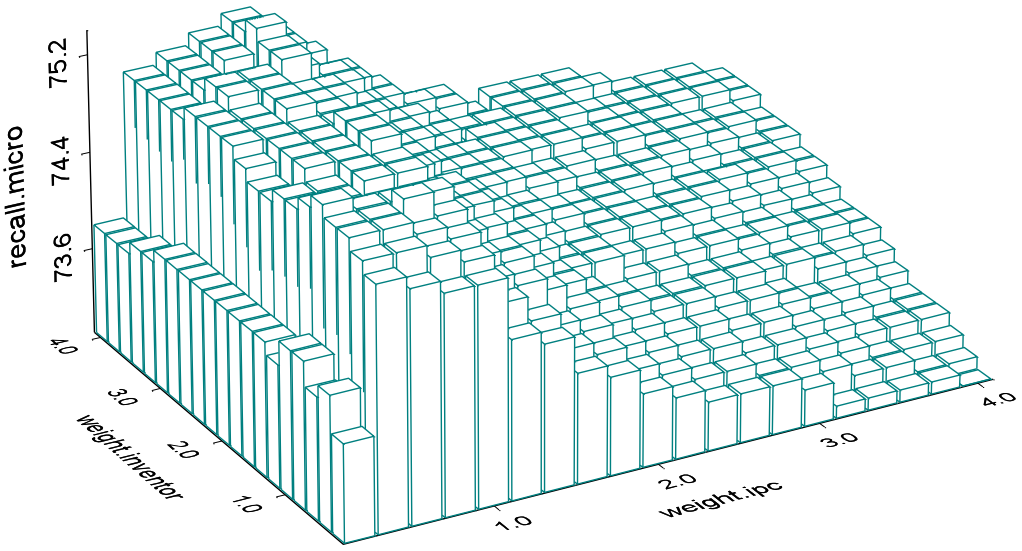
Figure 9 PFA Classification accuracy for 400 classifiers with IPC and inventor weights (relative to text) varying between 0.2 and 4.

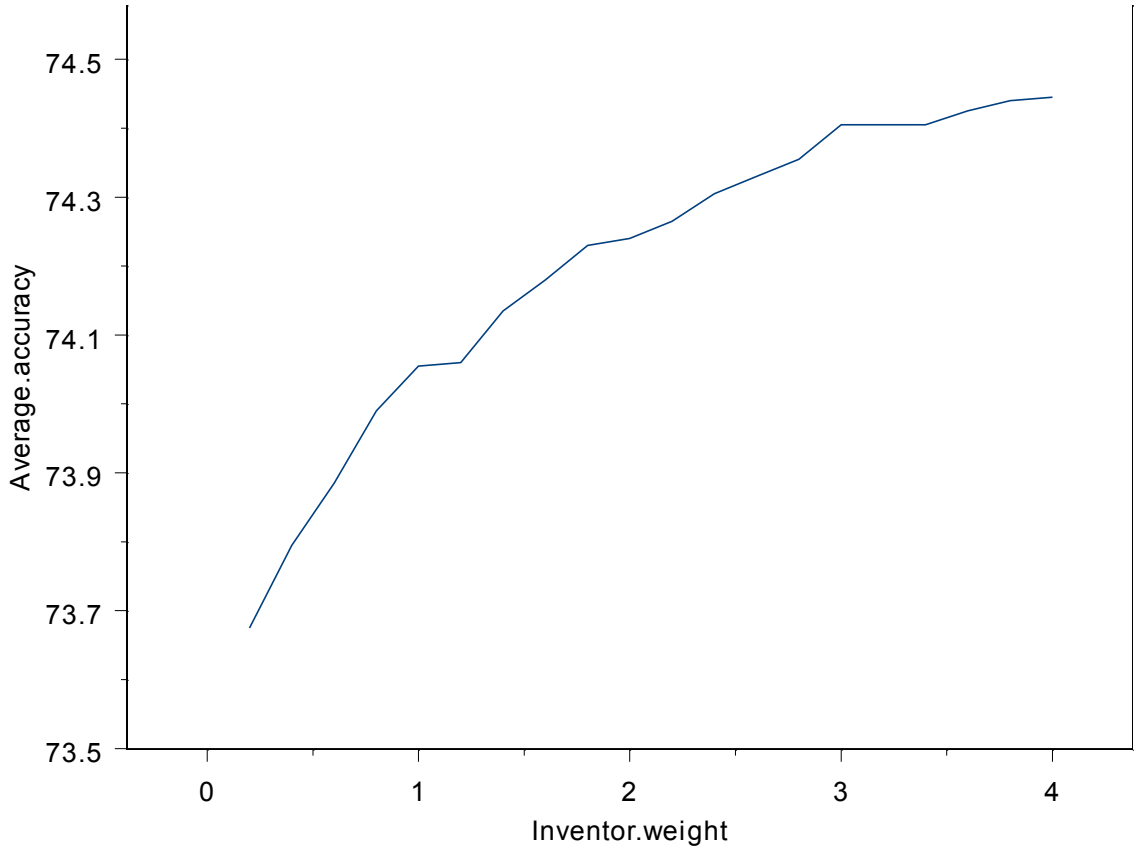
Figure 10 Average recall over IPC weight for the PFA classifier

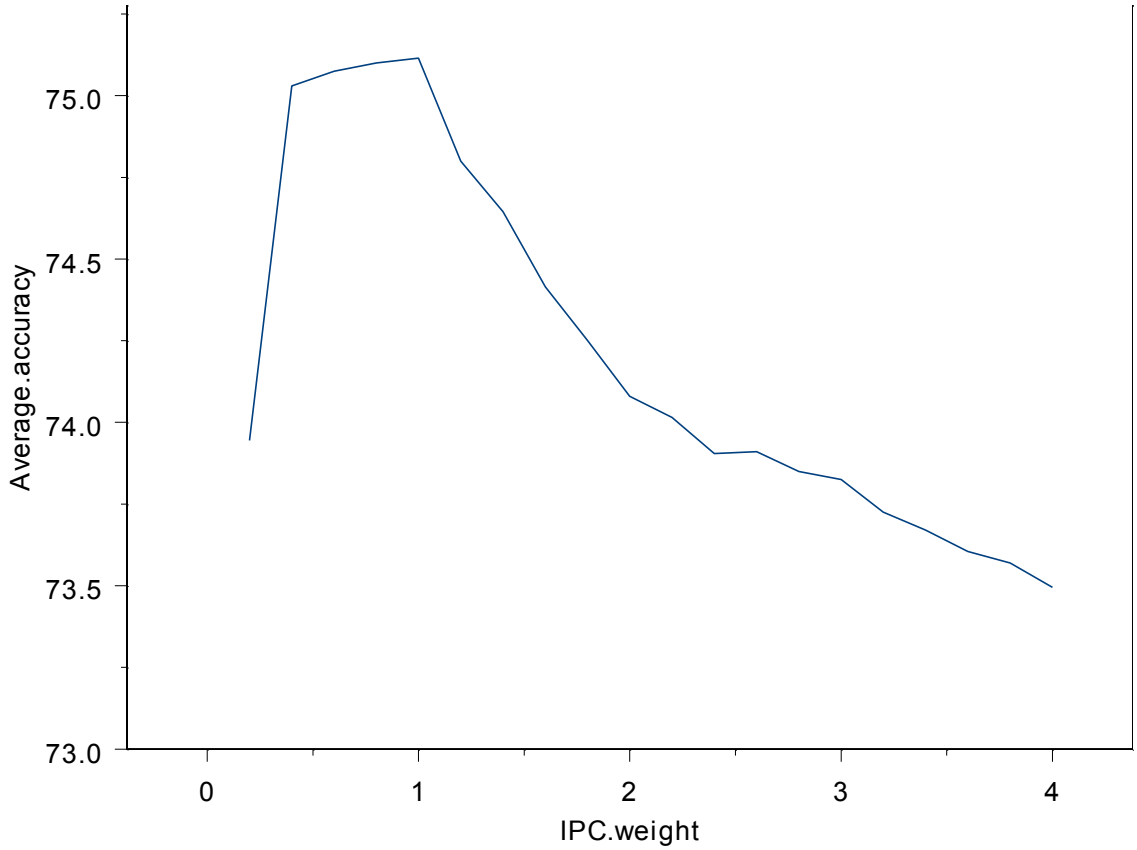
Figure 11 Average recall over inventor weight for the PFA classifier. Note that the values on the y-axis stretch only from 0.63 to 0.645, so the "real" slope of the graph is actually very low, compared with figures 4, 5 and 11.

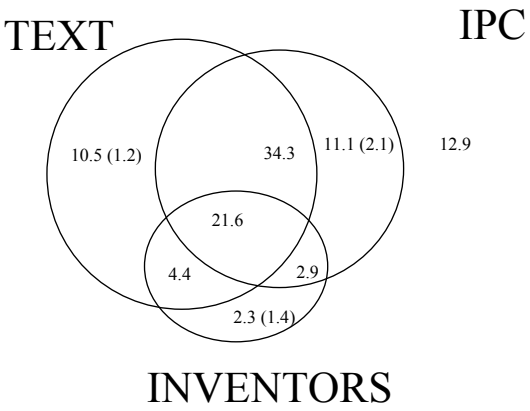
Figures


120000
100000
80000
60000
40000
20000
0
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002


recall.micro
73.6
74.4
75.2
weight.inventor
1.0
2.0
3.0
4.0
weight.ipc
1.0
2.0
3.0
4.0

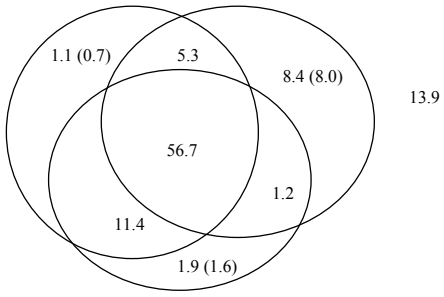

74.5
74.3
74.1
73.9
73.7
73.5
Average.accuracy
0
1
2
3
4
Inventor.weight


75.0
74.5
74.0
73.5
73.0
Average.accuracy
0
1
2
3
4
IPC.weight

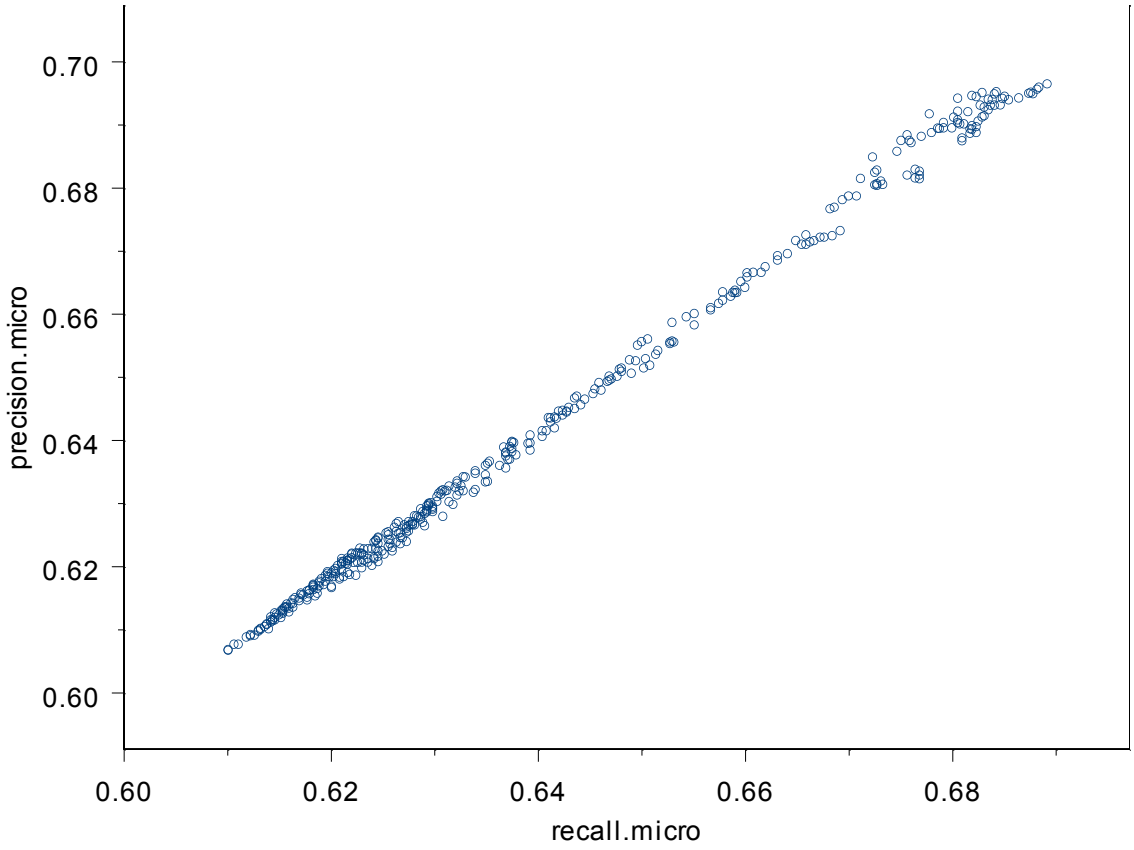

TEXT
IPC
10.5 (1.2)
34.3
11.1 (2.1)
12.9
21.6
4.4
2.9
2.3 (1.4)
INVENTORS

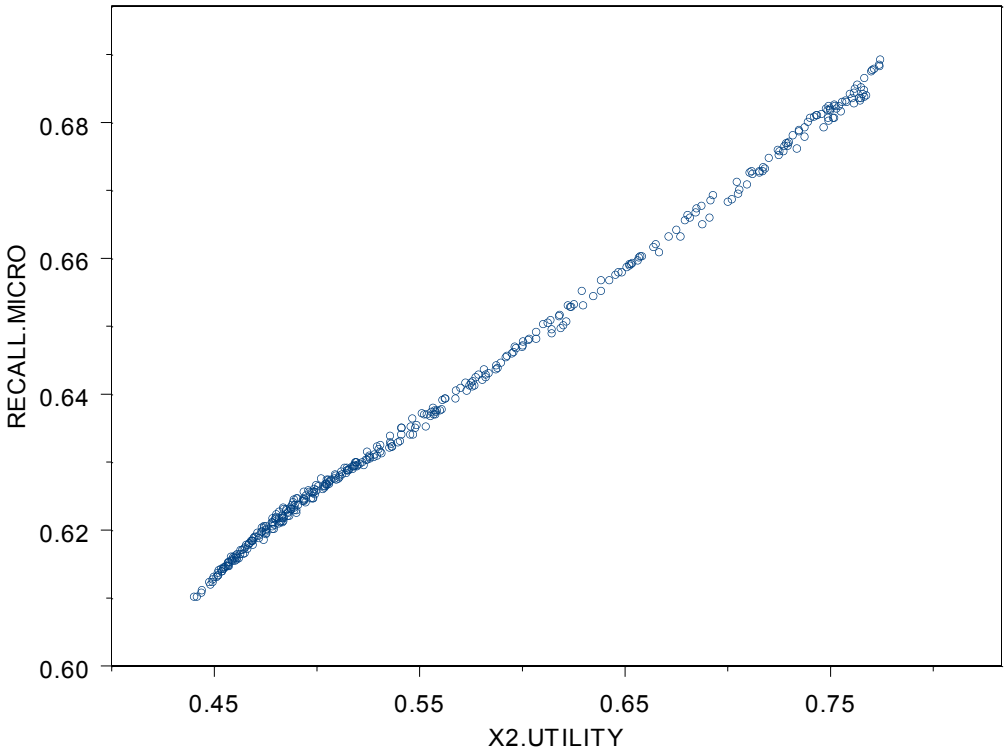
TEXT + IPC

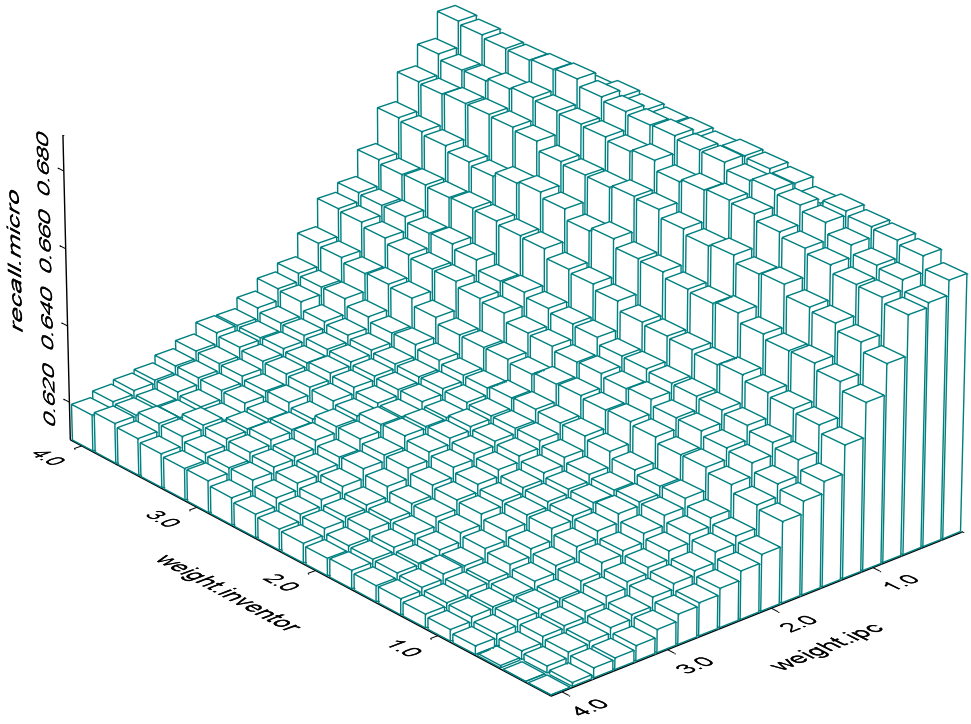
TEXT + INVENTORS

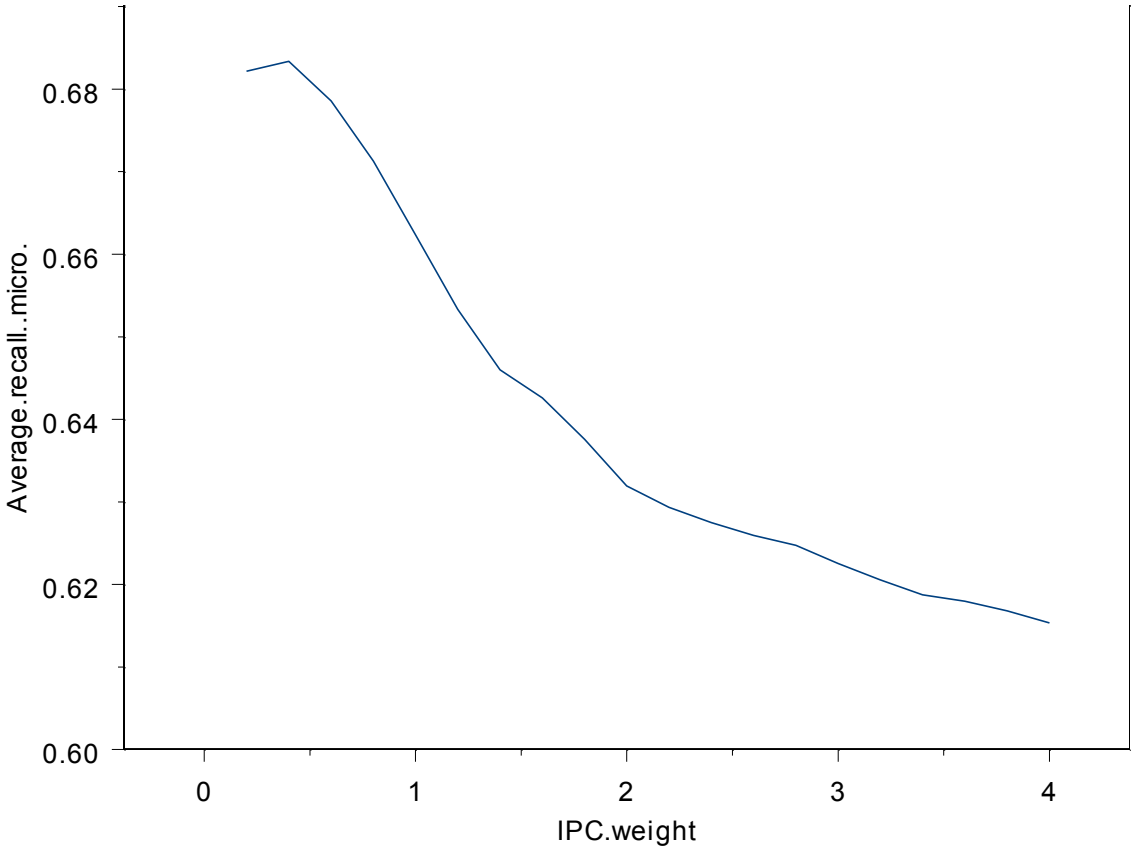


IPC + INVENTORS


0.70
0.68
0.66
0.64
0.62
0.60
precision.micro
0.60
0.62
0.64
0.66
0.68
recall.micro


RECALL.MICRO
0.68
0.66
0.64
0.62
0.60
0.45
0.55
0.65
0.75
X2.UTILITY


recall.micro
0.620
0.640
0.660
0.680
4.0
3.0
2.0
1.0
weight.inventor
4.0
3.0
2.0
1.0
weight.ipc


0.68
0.66
0.64
0.62
0.60
Average.recall..micro.
0
1
2
3
4
IPC.weight

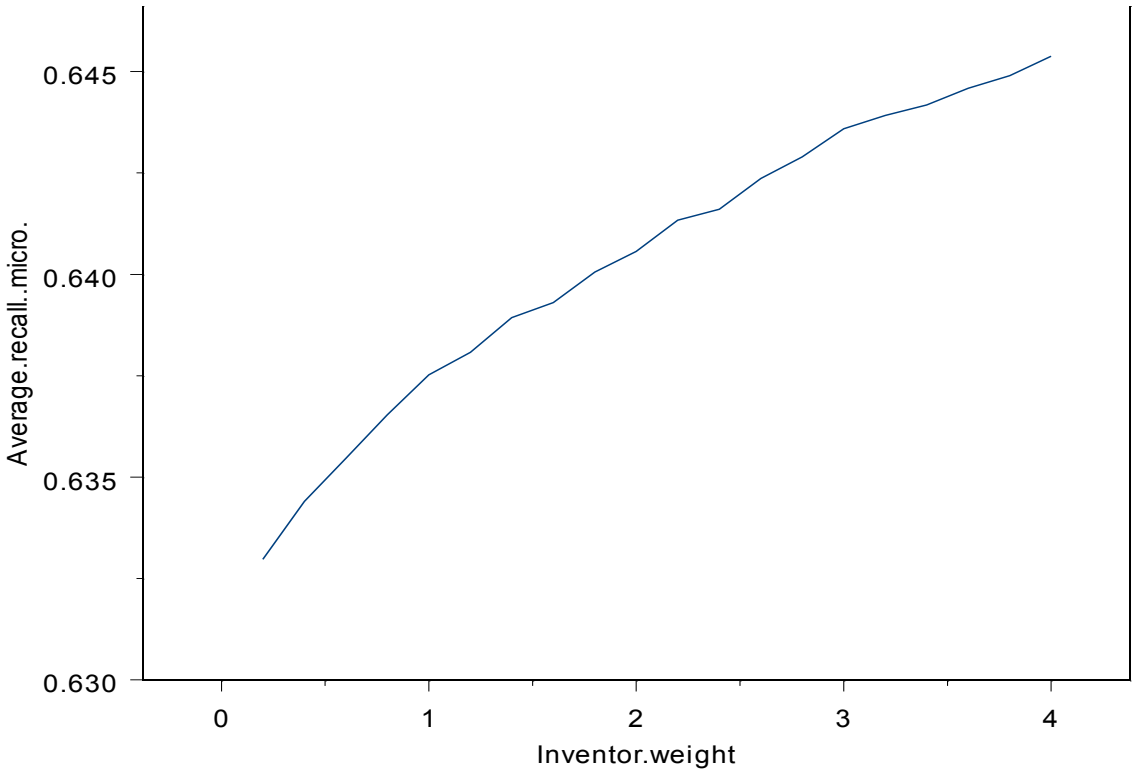

0.645
0.640
0.635
0.630
Average.recall..micro.
0
1
2
3
4
Inventor.weight

Table captions

Table 1 Classification results for the Gazette classification experiments based only on text, inventor and IPC. The percentage figures in brackets indicate classification accuracy on the document subset for which the classifier is actually "eligible", for example excluding the test set documents that have no inventor in common with the training set document for the classifier only based on inventor data.

Table 2 Classification results for the Gazette classification experiments based only on text, inventor and IPC as well as two or three of these attributes with equal weights. See table 1 for an explanation of the meaning of the figures in brackets.

Table 3 Classification results for the PFA classification experiments based only on text, inventor and IPC as well as two or three of these attributes with equal weights.

Tables

Text weight	Inventor weight	IPC weight	Classification accuracy
1	0	0	70.8%
0	1	0	31.2% (70.4%)
0	0	1	70.0% (71.0%)

Text weight	Inventor weight	IPC weight	Classification accuracy
1	0	0	70.8%
0	1	0	31.2% (70.4%)
0	0	1	70.0% (71.0%)
1	1	0	71.6%
1	0	1	74.5%
0	1	1	71.2% (71.9%)
1	1	1	75.1%

Text weight	Inventor weight	IPC weight	Micro-averaged recall
1	0	0	66.0%
0	1	0	62.0%
0	0	1	55.7%
1	1	0	66.6%
1	0	1	65.2%
0	1	1	56.2%
1	1	1	65.8%

Ref: Richter 271