WORLD INTELLECTUAL PROPERTY ORGANIZATION
ORGANISATION MONDIALE DE LA PROPRIÉTÉ INTELLECTUELLE

GENEVA/GENÈVE

Literature survey: Issues to be considered in the automatic classification of patents

C. J. Fall,[1] K. Benzineb

ELCA Informatique, Avenue de la Harpe 22-24, CH-1000 Lausanne 13, Switzerland, and
World Intellectual Property Organization, 34 Chemin des Colombettes, CH-1211 Genève 20,
Switzerland

This document summarizes the state-of-the-art in automated patent classification in mid 2002. We consider a variety of issues relevant to the future development of an IPC categorizer, in the context of the WIPO CLAIMS project.

In particular, we survey the most popular algorithms for text classification currently in use in the research community and in commercial products. A large number of research papers about document categorization have been studied and their main results are summarized. Important authors are identified in the literature.

The IPC is a complex classification scheme and we identify its peculiarities with respect to automated categorization. Tests involving the automatic classification of patents in Europe and the US are reported in detail.

Software products tailored for IPC categorization and patent topic clustering are listed and evaluated. Popular commercial software packages for generic text classification are equally presented, along with freeware alternatives.

We demonstrate that the technical needs of WIPO's categorization system are at the top end of current technological possibilities and will probably only be satisfied by the development of a customized solution and the availability of high-quality training sets.

[1] Email address: cjf@elca.ch, caspar.fall@wipo.int

I. Table of Contents

II. Record of changes

Filename	Version	Date	Description / Author
WIPO-CategorizationSurvey.doc	1.0	29.10.02	Final version / CJF

III. Acknowledgements

The authors of this document would like to thank G. Karetka, A. Farrasopoulos, M. Makarov, and A. Törcsvári for their assistance in preparing this report and in surveying the literature.

This work also benefited from conversations with researchers at the University of Geneva in the ISI group.

IV. Abbreviations

API	Application Programming Interface
CLAIMS	Classification Automated Information System
EELS	Engineering E-Library, Sweden
EPO	European Patent Office
INPI	*Institut National de la Propriété Intellectuelle* (France)
IPC	International Patent Classification
IT	Information Technology
JPO	Japanese Patent Office
OCR	Optical Character Recognition
PCT	Patent Cooperation Treaty
SIC	Standard Industrial Classification
SQL	Structured Query Language
URL	Uniform Resource Locator
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization
XML	eXtended Mark-up Language

V. Glossary

In order to establish a common vocabulary, we present here some definitions of technical terms appearing in this report.

Category	A generic collection of documents related to the same specific topic.
Categorize	To associate a document with one or more predefined categories.
Classify	A synonym for categorize.
Clustering	To separate a set of documents into categories without relying on any predefined taxonomy. A clustering system should automatically identify topics and group documents into them by calculation of topical similarity
Corpus	A collection of documents.
Ontology	A semantic network defining logical relations between concepts.
Polysemic	A polysemic word can have several different meanings depending on its context.
Section	The highest level of category in the IPC.
Class, subclass, group, subgroup	The names of labelled hierarchical subdivisions of the IPC, ordered by decreasing size and increasing number. At each subdivision of the IPC, the number of categories is multiplied by an order of magnitude.
Stemming	The process of extracting from a word the meaningful part, known as the lemma, and ignoring inflection, declension, or plural suffixes.
Synset	A collection of synonyms of words, that could form part of an ontology.
Taxonomy	A set of categories in which documents may be classified.
Term	A word or a phrase present in a document and used for indexing the document's content.
Vocabulary	The full collection of different words appearing in a corpus.

Terms specific to a given paragraph of this document are defined in the body of the text.

1 Introduction

This document relates to the WIPO CLAIMS project, which aims to provide the IT infrastructure for the reform of the International Patent Classification (IPC). We provide here a literature survey examining issues to be considered in the automatic patent categorization CLAIMS project. Approaches, algorithms, software options, limitations, companies, and important authors are discussed.

For this task relevant information from public sources has been collected and an analysis of the state-of-the-art in current categorization software, both in general, and as applied to patent classification is presented.

In this document, we first introduce the CLAIMS project and the need for categorization software. We summarize the technology outlook foreseen in this field. In Chapter 2, we present some categorization generalities about system training and testing. The major algorithms commonly in use are presented briefly, language issues are investigated, as are the uses of ontologies in document categorization. Chapter 3 is dedicated to reviewing applications of automated classification to patents in academic research, in patent office tests, and in commercial products. Chapter 4 lists some commercial and freeware generic categorization packages, while Chapter 5 presents our conclusions.

The Appendices consist of details about our research methodology and a who's-who list, in Appendix A, as well as an extensive bibliography of documents used as a basis for this report, in Appendix B.

1.1 Context of the CLAIMS project

The International Patent Classification (IPC) system, developed and managed by WIPO, is currently under reform [Calvert01]. In particular, its next edition will be divided into a stable core level of classification and an advanced dynamic set of categories that will be frequently updated.

The CLAIMS project aims to provide the technical infrastructure to support this reform. In particular, an automated patent classification system is needed for small and medium-sized national patent offices [Karetka02]. The system should facilitate the attribution of IPC codes to patent applications, promote the use of the IPC in member states, and possibly help with patent reclassification tasks. It may also motivate small patent offices to provide electronic versions of patent applications if this is not already the case.

The minimal goal is to produce a categorizer that is able to predict with at least 80-90% accuracy the correct IPC subclass from a selection of 3 to 4 suggestions—corresponding to about 600 categories—with the final decision being made by a human user through an ad hoc interface [Karetka02]. The categorizer should work in English and French, with a clear possibility to add other languages later.

Accurate patent classification is extremely important because it helps to determine the scope of the subsequent search for relevant prior inventions [Adams00]. The IPC is a large hierarchically-organised and human-designed system for classifying all fields of technological inventions [IPC7]. An automated classification tool will need to reflect the categorization rules currently used by human classifying experts and be adapted to the complexity of the taxonomy.

1.2 Automated categorization needs

A recent report by the Gartner Group suggests that the best method for classifying knowledge begins with a manual process of category selection, rather than by using automatically-generated document clusters [GGKM]. The IPC is a knowledge categorization system designed by experts with several hierarchical levels of detail and refined for over 35 years though seven editions.

A hand-built document taxonomy has several advantages. It reduces complexity and provides an overview of the knowledge contained within the information set [GGtaxonomy]. Hierarchical taxonomies capture information about relationships and are easy to browse.

The IPC is however a complex classification system. Because of its broad scope, non-expert human classifiers have difficulty using the IPC for manually attributing IPC codes. A tool for assisting with patent categorization would thus be of use in national patents offices where experts are lacking and staff members struggle to attribute accurate IPC codes. When a well-established classification scheme is in place, automation can be extremely useful for streamlining categorization and enhancing productivity [GGKM].

The promotion of IPC use by small and medium companies, as well as by individual inventors, would benefit from the provision of better public online services for patent application categorization.

The minimal objective is to provide a tool that would assist users in classifying patent application to IPC subclass level, corresponding to approximately 600 categories, perhaps with deeper IPC support in a subset of fields. More ambitiously, a system supporting main group categorization (with about 8,000 categories) is desired. Facilities for the categorization of documents in several main European languages are also vital [Karetka02].

With the forthcoming frequent updates of the advanced part of the IPC, categories will be modified or created to account for new fields of technology, thus preventing IPC subgroups from becoming overloaded with patents. It will thus be necessary to regularly form new categories of patents. In this context, automated document clustering could serve to guide the creation of new groups and subgroups by collecting similar patents together.

1.3 Technology outlook

Automatic document categorization typically uses statistical models or hand-coded rules to rate a document’s relevancy to certain subject categories. From the early 1990s, the effectiveness of automated text categorization has been rising thanks to advances in theoretical research and the development of new algorithms [Sebastiani02]. Although the accuracy will no doubt reach a plateau below 100%, automated systems are expected to achieve at least as high effectiveness as human classification.

With the advent of cheap computing power, categorization algorithms become more sophisticated and provide better performance. For example, the response time of a system at the USPTO has been reduced from 12 minutes to 10 seconds per query since it was installed [Smith02].

The Gartner Group suggests that text categorization is rapidly gaining popularity. By 2005, at least 70% of major corporations will deploy automatic text categorization for analysis and improved organization of internal documents with 0.8 probability [GGhypecycle].

This automation does not come without error. In 2001, the Gartner Group rated the typical accuracy of statistical techniques applied to non-overlapping categories to be in the 80 to 95% range [GGhypecycle]. More precise distinctions require hand-coded rules, such as those provided by ontologies.

Ontologies provide a formalized view of certain fields of knowledge. They explain concepts within domains, their attributes, and their relationships with other concepts.

Taxonomies are a form of simplified ontologies. Such a machine-understandable semantic layer has far-reaching implications for improving search capabilities and categorization tasks. Creating ontologies is a complex task, particularly for a domain as vast as that covered by the IPC, but the potential benefits are vast. The Gartner Group predicts that by 2005 ontologies will become a standard technology for marking up electronic product catalogues with 0.6 probability [GGhypecycle].

The Gartner Group recently published a research note about the increasing importance of ontologies. Gartner recommends that enterprises should begin to develop the needed semantic modelling and information management skills within their integration competence centres. By 2010, ontologies using strong knowledge representations will be the basis for 80% of application integration projects with 0.8 probability [GGontologies].

According to Gartner research, automatic category creation, also known as clustering, will fail to produce good results in creating complex taxonomies fully automatically until 2006 with 0.8 probability [GGtaxonomy]. Assisting human taxonomy creation is however a realistic proposition today.

2 Categorization generalities

2.1 Categorization methodology

Document categorization consists in associating a document with one or more predefined categories. Each category is a labelled group, and usually collects documents related to the same topic. The collection of categories for a single application is known as the taxonomy.

Automated classification is the assignment by a computer system of documents to categories on the basis of the information contained in the documents.

A different and perhaps more difficult task is that of clustering, where the system must first define a collection of categories and then assign documents to them.

The steps involved in all automated categorization systems are illustrated in Figure 1. First, a document to be classified must be pre-processed. This step may involve scanning the document and performing OCR to extract the full text if an electronic version is already available. Conversion to a format suitable for the categorizer input, such as XML, may be necessary.

The words in the documents are then listed and indexed. The vocabulary is extracted, phrases and terms may be recognised, and word stemming is optionally performed.

As document vocabularies tend to be large, it is advantageous to restrict the terms used for discriminating between categories. Common stopwords may be removed and the most significant words retained. If a module for recognising semantic values of words is implemented, ontologies may be used to expand or disambiguate terms in the document (see paragraph 2.4).

Finally, a discriminating algorithm is invoked to distinguish between the categories. A huge variety of techniques have been developed for this purpose, but they all rely on the terms selected from the document. A selection of common algorithms is presented in paragraph 2.2.

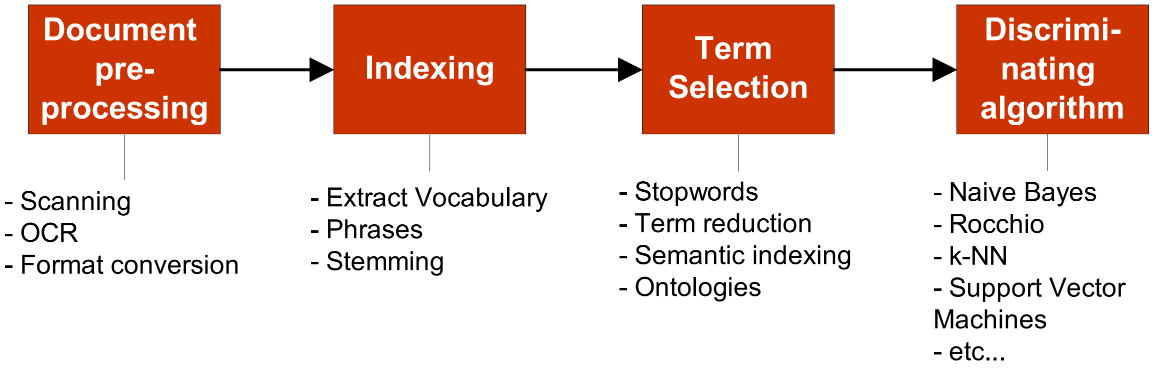


Figure 1: Categorization steps and options

2.1.1 Training

To implement an automated categorizer, most systems require training. During this phase, a collection of manually-classified documents are presented to the system, following the steps shown in Figure 1. From this information, the system learns to recognise category signatures according to the specifics of its algorithms.

In order to allow statistical categorization algorithms to develop a well-formed understanding of each category, it is necessary to ensure that the set of training documents is well distributed among the categories [Gey99]. The influence of the skew of the training set distribution on the resulting accuracy is dependent on the taxonomy details and is still a subject of active research where few publications exist.

If the training set consists of documents of highly-variable length, statistical algorithms exploiting word distribution may have difficulty learning to classify documents accurately, particularly if longer documents are preferentially located in some categories. The classifier might then unnaturally favour such categories.

Some algorithms require a validation step after the training phase, in which a set of thresholds are tuned to allow the system to distinguish between documents relevant to a category and those that are not. Such a procedure is commonly required for multiclassification tasks, when a variable number of categories must be associated with each document. The document collection used for validation is ideally different from that used for training.

2.1.2 Testing

Automated categorization systems can be tested by presenting unassigned documents to the tool and examining which categories are suggested. These automated assignments must then be compared with the results of human classification. It is important to test an automated categorization tool with a collection of documents that were not used during training, to avoid biasing unfairly the results.

When multiclassification tasks are required, a number of common indicators of accuracy of used. Amongst these are:

- **Precision:** indicates the system's ability to retrieve categories that should be attributed, and is defined as the ratio between correct categories found and total categories found.
- **Recall:** indicates that system's ability to retrieve all relevant categories, and is defined as the ratio between correct categories found and total correct categories.
- **F1-score:** In general, there is a trade-off between precision and recall. By means of thresholds, most automated categorization systems can be tuned to provide high precision but then suffer low recall, or vice-versa. The F1-score is a harmonic average between precision and recall that attempts to account for this choice. Average or maximum F1 values are reported.
- **Break-even**: Another attempt to account for precision and recall variation is provided by the break-even value, which corresponds to the value where precision and recall are equal.

Mathematical definitions of these measures are provided in [Sebastiani02].

2.2 Categorization algorithms

A variety of different algorithms for text categorization have been developed. We list here the main approaches, and present them in a non-mathematical manner. The mathematical details are reviewed extensively in the literature, see for example [Sebastiani02] and references therein.

2.2.1 Naïve Bayes

The Naïve Bayes categorization approach is a simple probabilistic technique. The probability that a particular document belongs to given class is determined by relying on the assumption that word distributions are independent variables, *i.e.* that the presence of one word has no effect on the distribution or presence of other words in the same document, an assumption that is most likely not seen in practice. From estimated probabilities that words belong to different categories, the probabilities that documents belong to various categories are determined [Sebastiani02]. Documents are often represented by vectors that account only for the presence or absence of words, not for their frequency in each document. Overall, the Naïve Bayes algorithm is simple to implement, but usually outperformed by the other techniques explained below [Yang99b].

2.2.2 Rocchio

The Rocchio technique builds for each category a single prototypical document. The category profile consists of a weighted list of words or terms formed from the word distribution within the category. To decide which categories a new document belongs to, its word distribution is compared to those of the prototypical category documents. When the similarity is high enough, the new document is assigned to the category in question. [Sebastiani02]

While extremely simple to implement, difficulties arise with this algorithm when the documents in a category are bunched in several disjoint groups. The algorithm will then compare a new document to an average prototypical document that may be quite different from those in each bunch of the category, leading to obvious categorization errors.

Refinements to improve the Rocchio algorithm consist in including negative training examples, taken as documents just outside each category, and using them as negative indicators when computing prototypical document vectors [Sebastiani02].

2.2.3 k-NN

In the k-NN approach, when a new document is to be classified, it is compared to the existing set of pre-classified documents to locate those that are most similar. In this, it is an example-based approach where all categorization effort is deferred until new documents are presented. Similarities between documents are computed by comparing word distributions. k-NN stands for k nearest neighbours, where k indicates the number of neighbouring documents examined, in practice between 20 and 50. The suggested category of the new document can then be estimated from those of neighbouring documents by weighting their contributions according to their distance [Sebastiani02]. By sorting the scores of suggested categories, multiclassification tasks can also be performed.

The k-NN algorithm is usually found to be very accurate [Yang99]. It does not divide the document space in a simple linear way, and therefore outperforms the Rocchio approach [Sebastiani02]. It scales well to huge numbers of training documents and categories [InxightWP].

The k-NN approach has been applied to a vast number of categories (14,321 categories), where the F1-score was found to be 51% on a corpus of medical abstracts [Yang99].

Furthermore, this algorithm naturally provides a set of documents similar to the document to be classified. Such a feature is advantageous in the case of patent IPC categorization, as it provides a basis for the search for prior art.

2.2.4 Support vector machines

Support vector machine algorithms are extremely powerful at text categorization, but rather abstract in their description. The objective of the algorithm is to find the decision surface in the space of all possible documents that best separates documents relevant to a category from those that are not. In this respect, it relies not on all documents in the category but only on the outlying documents that delimit the edges of the category. The advantages of the technique are that there are no parameters to configure and that term selection is often not required as huge vocabularies can be supported with ease [Sebastiani02]. The disadvantage is that such a system behaves essentially as a black box, where categorization decisions can only be explained to the user with difficulty.

2.2.5 Neural networks

An artificial neural network consists of a network of many simple units, usually positioned in successive layers. Communication channels that carry numeric data connect the units, with varying connection strengths, on the model of biological networks.

A network layer receives input, in the form of a collection of terms and weights representing a document, intermediate layers process the weights, and an output layer suggests a relevant category [Sebastiani02]. A large number of variations exist in this architecture [Wermter99].

2.2.6 Decision rules

Decision rules algorithms classify a document by following a set of classification directives or rules. The rules indicate when a word, or a collection of words, or the absence of a word, is a good indicator that the document belongs to a given category. The rules may be combined in the form of a complex decision tree [Sebastiani02, Schapire00].

Such category decision rules can be learned automatically, by examining which words discriminate between categories, or experts can formulate them manually. In this respect, this approach is unique as it allows document categorization without the availability of a training set of pre-classified documents, provided experts can formulate the categorization task accurately [Johnson02].

When no explicit knowledge about categorization rules is available, machine-learning algorithms that rely of word statistics, such as those presented in the preceding paragraphs are eminently suitable. In the case of the IPC, a large amount of prior classification rules and knowledge is available, making rules-based categorization an attractive option. Such a technique appears not to have been extensively tested on patent classification.

2.2.7 Categorizations into hierarchical taxonomies

When the taxonomy is hierarchical, it is interesting to consider whether the categorization algorithm could or should exploit this fact. Indeed, one can imagine directly classifying incoming documents among a large number of taxonomy nodes without taking the hierarchy into account, or one can imagine categorizing documents at each level of the taxonomy tree separately, traversing the tree and selecting from a small number of subcategories at each step, as illustrated in Figure 2. In the latter case, it might then be possible for a discriminating keyword at one level to become a stopword at another. The complexity of a hierarchical classifier is higher than that of a flat scheme, as a number of separate classifiers often need to be trained for each level of the hierarchy.

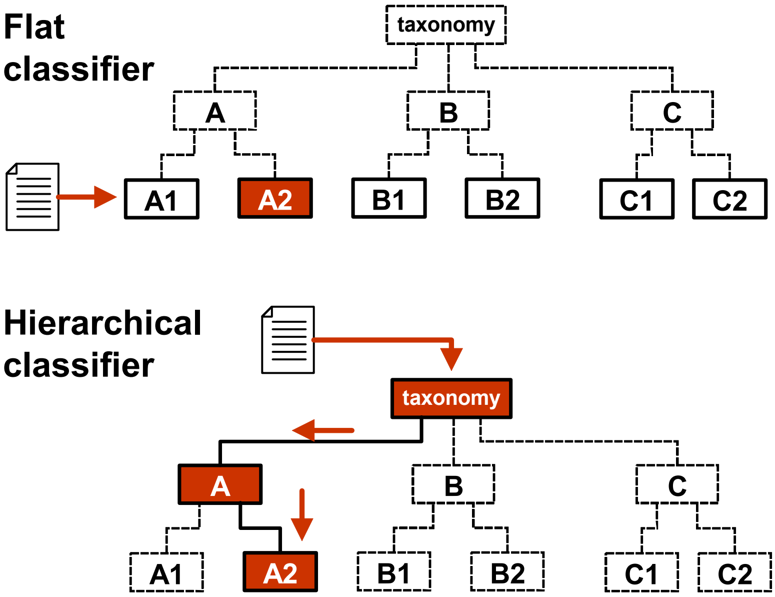


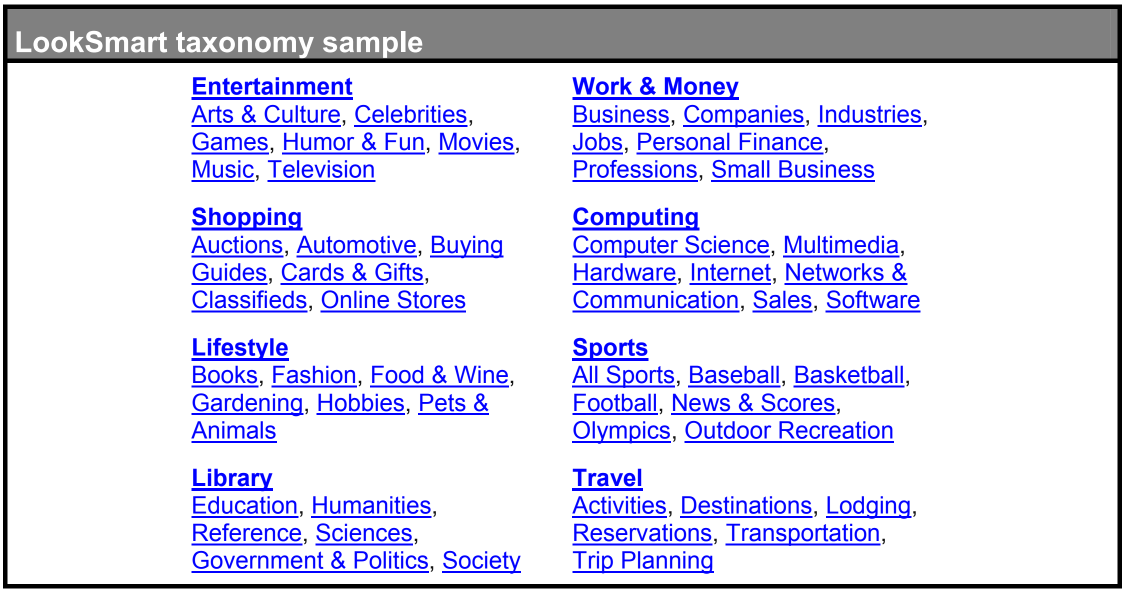
Figure 2 : Flat classifiers enter a document directly into the lowest taxonomy nodes, while hierarchical schemes follow the tree structure during classification

When using a hierarchical classifier, one must decide how to train the system. It may be advantageous to train sub-classifiers only with negative examples from within the same parent category, rather than from the whole corpus. This may provide better sub-category discrimination and smaller training times.

Furthermore, various rules for combining the results of each level of classification can be imagined. One might classify a document by following the hierarchical classification tree and accepting each classification level separately (Boolean combination), or one might multiply the results of all levels of classification and choose the final category from a global score (multiplicative scoring). The latter of these two schemes is of course much less efficient for classification, as all branches of the hierarchy have to be explored.

An automated hierarchical categorization of web pages in the LookSmart directory (www.looksmart.com, see Table 1) into 13 top levels and 150 sub-levels has been tested recently [Dumais00]. The categorization was based only a digest of the full-text of the web page. Using a support vector machine algorithm, a small improvement in the accuracy of a hierarchical scheme (49% average F1 score for 150 categories) was found over a flat classifier (47%). It was found that the accuracy of the hierarchical classifier varied widely across the taxonomy sub-branches. No difference between Boolean combination and multiplication schemes for combining the sub-categorizers was noticed, the former thus being recommended for efficiency reasons.

In [Dalessio00], the authors compare flat and hierarchical schemes and test several variations of hierarchical classifiers. Using a 135-category Reuters dataset of news stories, they obtain slight improvements in precision and recall when using a hierarchical scheme compared to bulk categorization. In the best case, after training and validating, the F1-score for the categorization of test documents rose from 79% (direct global categorization) to 83% (exploiting the hierarchy). It should be noted that the Reuters taxonomy is not intrinsically hierarchical, but that the authors imposed a variety of artificial category hierarchies for the purpose of their work.



LookSmart taxonomy sample

Entertainment Arts & Culture, Celebrities, Games, Humor & Fun, Movies, Music, Television	**Work & Money** Business, Companies, Industries, Jobs, Personal Finance, Professions, Small Business
Shopping Auctions, Automotive, Buying Guides, Cards & Gifts, Classifieds, Online Stores	**Computing** Computer Science, Multimedia, Hardware, Internet, Networks & Communication, Sales, Software
Lifestyle Books, Fashion, Food & Wine, Gardening, Hobbies, Pets & Animals	**Sports** All Sports, Baseball, Basketball, Football, News & Scores, Olympics, Outdoor Recreation
Library Education, Humanities, Reference, Sciences, Government & Politics, Society	**Travel** Activities, Destinations, Lodging, Reservations, Transportation, Trip Planning

Table 1 : A sample hierarchical classification of web pages, from www.looksmart.com

An earlier similar study was made on the same Reuters corpus using a variety of extended Bayesian classifiers in a hierarchy [Koller97]. In this case, a similar improvement in accuracy was noted when using a hierarchical classification employing very small numbers of discriminating words, which differed at each level of the categorization hierarchy.

Another related system was built by researchers at IBM, implementing a multitude of Bayesian classifiers at each level of a hierarchy [Chakrabarti97, Chakrabarti98]. Between 5 and 10% of the vocabulary is used to distinguish between documents at each level of the taxonomy. Their system is claimed to be trained much faster than the previous one [Koller97], and details about required database optimisations have been published [Chakrabarti98]. When compared to a flat classifier, the hierarchical scheme is found to be much faster and to modestly increase accuracy.

A difficulty in training systems to recognize the correct category from a large number of choices is that the overwhelming number of negative examples for each category sometimes leads algorithms to over-reject documents from relevant categories. Thus some researchers have proposed using a special technique to select training documents at each level of a hierarchical scheme [Ruiz02]. In a detailed investigation specifically tailored to classifying medical abstracts related to heart disease, a hierarchical set of neural network classifiers was able to produce more accurate results than a single neural network attempting a flat classification. The average F1 score was improved from 48% to 51.2% when using the hierarchical scheme for 119 categories. However, a well-trained flat Rocchio classifier outperformed the hierarchical scheme and obtained a 51.6% average F1. Neural networks were found to perform better for low-frequency categories, while the Rocchio classifier was more accurate for medium and high-frequency categories. The Rocchio classifier was successfully trained by including positive examples for each category and negative examples just outside the category (rather than the full set of negative examples).

2.2.8 Classifier committees

A classifier committee (also known as a metaclassifier) is a categorization tool that seeks to combine several individual categorization tools or algorithms into a single one, thus aiming to preserve the best features of its constituents, and so obtain a better performance [Sebastiani02]. While an appealing approach at first sight, the challenge is to allow the committee to known when each of its constituents is performing well and not to combine each one's faults.

By providing several alternative point of view, each relying on a different technology or approach, for example with a statistical algorithm and a rule-based ontological approach in parallel, the chances of confusing either the classifier committee or the user exist. By increasing the number of parallel approaches, one does not necessarily increase the final accuracy, as the task of deciding when each technique should be preferred becomes harder. Furthermore, the implementation complexity increases when several techniques must be developed and trained simultaneously.

Research has suggested that the efficiency of a majority-voting scheme, where each individual classifier suggests a best category and the final one is selected from a majority vote, may not be adequate [Bennett02]. The suggestion of the best-performing categorizer in a specific setting may indeed be watered down by votes from less efficient schemes in that context, leading it to being rejected. A better solution is to evaluate during training the efficiency of each classifier for each category and perform voting combinations based on this information.

Experiments have shown that by suitably combining algorithms of a very different nature—such k-NN, Rocchio, and Naïve Bayes—improvements over single classifying algorithms can be obtained [Sebastiani02]. A weighted linear voting system was used in this work, whereby the classifiers' proposals are combined according to their confidence levels in their categorization suggestions.

The so-called boosting method seeks to train a committee of classifiers simultaneously, so that one categorizer may focus on correcting the mistakes made by the others. Experiments have again shown that for several different test collections, some improvements over single categorizers can be obtained [Sebastiani02].

2.2.9 Comparisons between algorithms

It is difficult to extrapolate the effectiveness of a classification algorithm from one corpus and taxonomy combination to another. For example, an algorithm selecting for a category defined as "documents with 50 or fewer instances of the letter A" can easily produce 100% precision and recall. If the category is defined as "articles about companies whose stocks will go up by $5/share tomorrow", precision and recall probably won't be very high whatever the algorithm and corpus [Lewis02].

The influence of document pre-processing techniques on the accuracy of classifiers has been systematically explored for web page categorization [Mase98b]. The application consisted in classifying web pages in 15 categories, with over 10,000 training documents. In this context, pre-processing is important because web pages are of strongly varying length and have a diverse vocabulary. In this particular case, it was shown that word stemming did not increase accuracy noticeably, that stopword deletion was important, and improved accuracy by 8%, that truncating the full text of the pages reduced the accuracy by only 2% when a maximum of 3,000 characters per page were used, and that deleting very low frequency words reduced the accuracy by 3%. Various techniques for normalizing keyword weights with respect to their distribution across the categories showed little differences. It is unclear whether these results can be generalized to other corpuses of documents.

Extensive comparisons between algorithms have been published in the literature [Hearst98, Yang99, Yang99b, Sebastiani02, Lewis02]. Only controlled experiments on the same corpus and taxonomy can be used to distinguish well between algorithm accuracies. Usually, such tests have been performed on standard Reuters test collections of newswires, as these are large and readily available online. The number of categories tested is seldom above a hundred. As the number of categories increases, the effectiveness of all algorithms drops. The rate at which the accuracy diminishes is strongly dependent on the corpus distribution, the definitions of the categories, and the textual content of the documents.

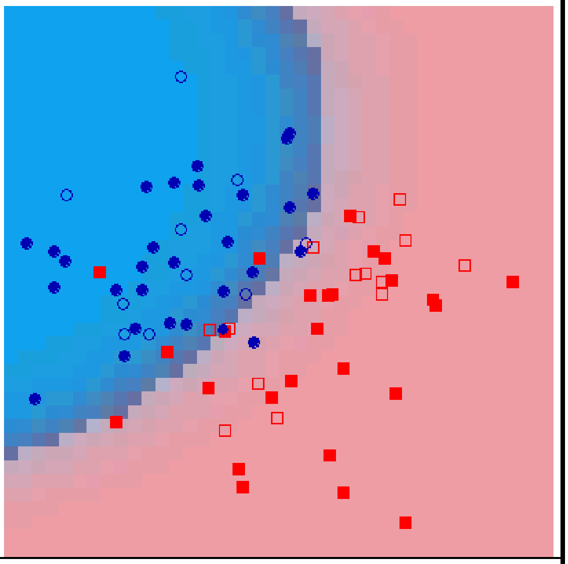
In general, studies have shown that support vector machines and k-NN algorithms outperform neural networks, all of which are more accurate than Rocchio and Naïve Bayes approaches. Tests on several different corpuses support this view [Sebastiani02, Lewis02]. Classifier committees generally deliver good performance as well.

Little information exists about the accuracy of decision rules algorithms, although one comparative study found decision-tree accuracy between that of support vector machines and Naïve Bayes algorithms [Hearst98].

We illustrate the differences between the algorithms using an online categorization demonstration package (www.cs.technion.ac.il/~rani/LocBoost/index.html). This applet provides a number of categorization algorithms, applied to a two-dimensional categorization problem. In Figure 3, three categorization algorithms are applied to the same data. Red squares and blue circles represent a random set of points (which would correspond to documents in a text categorization task) to classify in two distinct categories. Solid points represent the training set, open points show the test set. Shaded areas give the outlines of the two categories, which are seen to differ strongly between the various algorithms.

Naïve Bayes

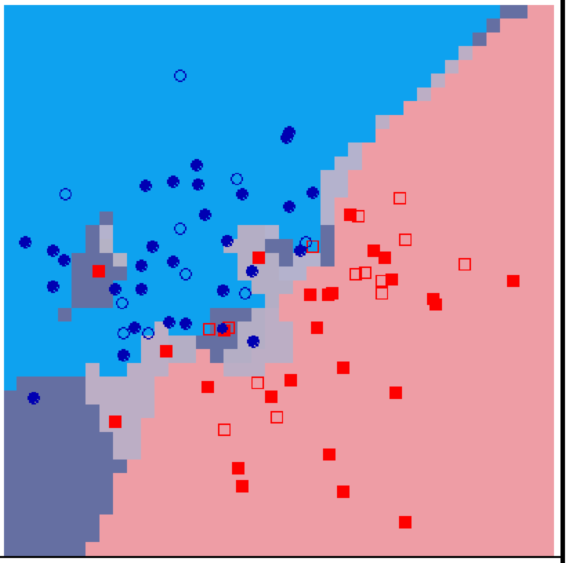
In this two-dimensional example, the naïve Bayes categorization relies only on the statistical distribution of points along the x- and y-axes.



k-NN

The k-NN categorization is performed here by examining the two nearest neighbours.

If a larger number of training points were provided, a correspondingly larger number of nearest neighbours might be required for best efficiency.



Support Vector Machine

This application of a support vector machine algorithm uses a so-called linear kernel, where the decision surface between the two categories is constrained to be a plane.

More complex implementations allow the decision surface to have a more complicated shape, but these have not been extensively used in text categorization tasks.

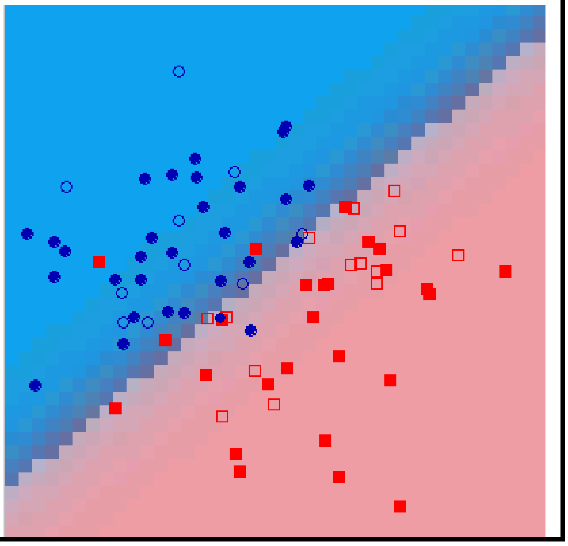


Figure 3: Comparison between various categorization algorithms applied to a random two-dimensional classification task, from www.cs.technion.ac.il/~rani/LocBoost/index.html

2.3 Language issues

Most categorization research is performed with English language documents. While discriminating algorithms are expected to perform equally well with documents in any language, the extraction of word or term vocabularies from documents in other European languages must be carefully considered. Lists of common stopwords to eliminate will obviously differ from language to language.

Issues related to word stemming can complicate word indexing. In German, for example, long words such as *Lebensversicherungsgesellschaftsangestellter* or "life insurance company employee" cannot be stemmed as simply as English words. Instead a more complex word tokenization algorithm can be implemented to extract the components of compound words. Stemming algorithms in non-English languages have been described [Gaustad01, Porter02]. French is said to be particularly difficult to stem [Porter02].

Some researchers have suggested disregarding tokenization in favour of indexing sequences of letters, also known as N-grams of characters [Biskri02, Huffman95]. In this approach, words are discarded, and all 4-letter combinations in the document are instead used to represent it. This approach can be used equally well in all languages.

Additionally, some researchers have converted accented characters to pseudo-sequences of English letters [Banik01]. For example, Hungarian characters have been converted using the following replacements: é→ee, á→aa, ú→uu, õ→oeoe, ö→oe, û→ueue, etc… This particular problem of accented characters, and more generally of extended and non-latin alphabets, may also be solved if the tokenizer and the indexer support the Unicode UTF-8 or UTF-16 standards.

2.4 Use of ontologies in categorization

Ontologies are semantic networks of concepts and relationships, generally organized in a hierarchical structure. They allow one to represent and process information which would be difficult to capture by other means, for instance by brute force methods

2.4.1 Definitions

There are a large number of definitions of ontologies and most of them vary widely, depending on the point of view or application. Thomas Gruber gives a popular definition: “An ontology is an explicit specification of a conceptualization. […] In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names are meant to denote, and formal axioms that constrain the interpretation and well-formed use of these terms." [Gruber93].

John Sowa gives another good definition, certainly not narrower in scope: "The subject of ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D. The types in the ontology represent the predicates, word senses, or concept and relation types of the language L when used to discuss topics in the domain D." [Sowa01]

These broad and open definitions by recognized gurus of the field show that ontologies are abstract representations of a given domain which not only allow domain concepts to be captured, but also their features and functions, as well as their relationships with other concepts. By their very structure, they can easily be computerized, although their implementation may be time-consuming.

From a practical point of view, ontologies offer a handy and powerful way of representing any type of information and of using such a representation to process that information. Ontologies allow in particular metadata to be integrated to describe the content or use of given documents or objects.

Applications domains are extremely wide. The most ambitious one is probably the Semantic Web project (www.w3.org/2001/sw), which aims to integrate metadata in the World Wide Web to allow for more efficient information searches. Tim Berners-Lee, father of the World Wide Web and director of the Semantic Web project, declared in 2001 that "Projects from the areas of knowledge representation and ontologies are coming together […] There's a clearly understood need for ontologies in a large number of industries" (www.xml.com/pub/a/2001/03/21/timbl.html) A large fraction of the research performed in the context of the Semantic Web will have a direct incidence on the use of ontologies in categorization applications, in particular the research concerning automatic or semi-automatic document annotation and metadata creation [EKAW02].

Other applications include various knowledge management tools, including knowledge maps, semantic search engines, intelligent catalogues, and automatic categorization systems.

2.4.2 Examples of ontologies

Since the tree of Porphyry drawn by Peter of Spain in 1329, and probably even before, humans have used trees to represent concept types and hierarchies. However, ontologies do not necessarily have a tree structure, or a broad coverage: there are generic ontologies such as WordNet (www.cogsci.princeton.edu/~wn)

[Fellbaum98], intermediate ontologies such as in the CYC system [Kingston01], and domain ontologies such as the Drug Ontology (www.cs.man.ac.uk/mig/projects/old/drugontology). Some systems integrate all three levels, such as ONIONS (ONtological Integration Of Naive Sources) [Ding01]. Some ontologies integrate information at a very detailed level, as for instance in the furniture domain, as shown in Figure 4. This ontology was implemented at the University of Geneva by the ISI Group and may be browsed online at cui.unige.ch/isi/cterm/english.html.

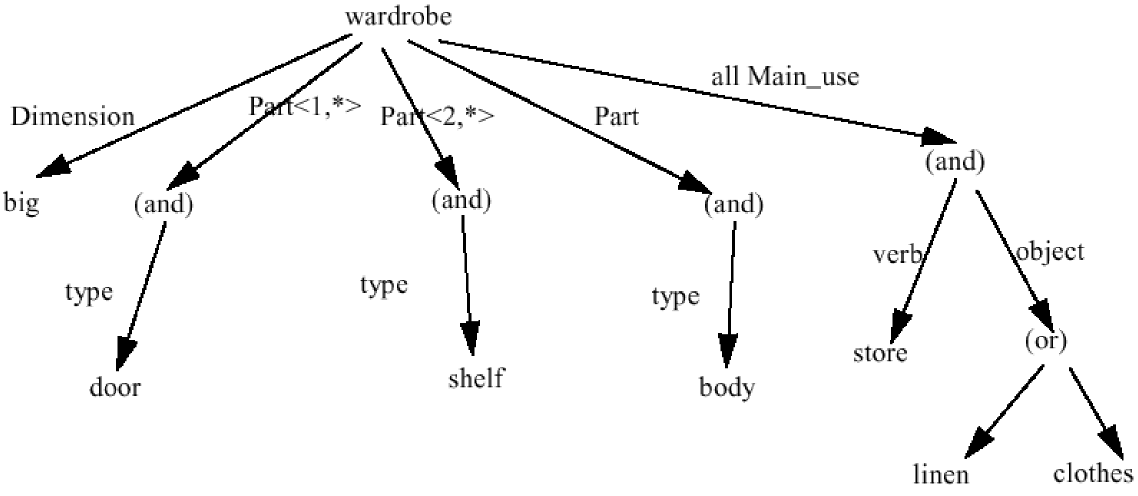


Figure 4: Detail of a furniture ontology, from ISI Group, Geneva University

Thus ontologies may be used to represent information, and even knowledge, in a way that then allows search and inference rules, sorting functions, and other utilities for knowledge management applications to be built.

2.4.3 Some ontology-based applications

Few companies seem to be using ontologies for knowledge management applications so far, probably because this technology's important evolutions are recent. Two companies have been working in the field for some time.

Ontoprise GmbH (www.ontoprise.de/home.htm) is a spin-off from the University of Karlsruhe and is considered one of the most important European research centers on natural language processing in general, and on ontologies in particular. Its flagship product, the SemanticMiner, is a Knowledge Retrieval platform that combines semantic technologies with conventional retrieval approaches. It is said to improve navigation by enabling the user to easily define semantic queries to all kinds of information sources – especially unstructured documents. Ontoprise does not seem to have applied its ontological technology to categorization projects so far.

Ontology Works (www.ontologyworks.com) have developed a specific environment called the Integrated Ontology Development Environment (IODE) to generate, edit, load, browse, and manage ontologies. According to them, "These tools are used to engineer, and create ontologies reflecting knowledge in a business domain and automatically generate databases and application software components directly from this business knowledge. Furthermore, the tools work with any commercially available platform (Oracle, Microsoft SQL Server, DB2, Java, etc.), regardless of the type of database (object oriented, relational)." It seems that IODE is dedicated to overall knowledge management tasks. We do not know about any application of IODE ontologies to a categorization project.

2.4.4 Principles of an ontological support to categorization

Our interest in ontologies, in the context of the IPC categorization project, lies in the fact that they offer a very powerful way of representing information, especially when information is already classified in a hierarchical structure.

Ontologies are not simply a way to organize information. They allow semantic descriptors to be included, detailing what a concept is composed of, or included in, or what other words or phrases may be used to designate it, including in other languages. Therefore ontologies may be used not only to organize and retrieve information or knowledge, but also to improve the power of an automatic categorizer, for example by:

- Contributing to disambiguation, through the use of synonym sets, descriptors, and relationships
- Allowing human-oriented categorization rules to be represented
- Allowing words in other languages to be added to describe existing concepts.

Research results concerning word disambiguation for guideline dissemination and word classification according to context are published in [EKAW00]. These results validate the approach of using ontologies for this type of application.

2.4.5 Research in ontology support for categorization

Research is not very extensive in this area and tends to produce mixed results. We present some published literature in order of increasing number of taxonomy categories.

Wermter and Hung [Wermter02] have implemented an approach based on neural-network self-organizing maps similar to those of Kohonen, presented below in paragraph 3.2.4. The public WordNet ontology was used to support a Reuters news story classification in 8 topics. Without the ontology support, the results of the system applied to the full text produced a 92.77% accuracy score. With the help of WordNet, the accuracy score rose to 98.95%. When headlines only were used instead of full text for the test set, 85.70% accuracy was achieved without WordNet, against 94.21% accuracy with ontology support. Although no data related to recall is available, it is very encouraging to see that the improvement factor is 6-9%, on a range above 85%, where improvements may be the most difficult to achieve. It should be pointed out that such a neural-network mapping approach is not expected to scale well to large numbers of categories.

Yamazaki and Dagan [Yamazaki97] develop a mistake-driven learner called Winnow, which accounts for variations in document length, and apply it to categorize Japanese news stories into 13 categories. A Japanese thesaurus containing semantic categories is included to improve the scores. The best scores obtained without the thesaurus are 70% precision and 55% recall. With the help of the thesaurus, these scores improve slightly to 72% and 61% respectively. It can be seen from these figures that such an approach allows categorization scores to be improved to some extent, although in absolute terms the results are not impressive.

Scott and Matwin's experiment [Scott98, Scott99] uses WordNet for its synsets and for its hypernymy (*i.e.* its upward hierarchy). Additionally, WordNet is used to change the representation of the text itself, by taking into account the density and frequency of synsets. The test is performed on 3 corpuses, namely Reuters news classification in 90 categories, folk song lyrics categorization in 33 categories [Scott99], as well as some small binary classification tasks with newsgroup contributions [Scott98]. The categorizer is based on the Ripper learning algorithm, a decision rules implementation. For the Reuters and song lyrics corpuses, results including WordNet were invariably poorer than by using a simple set of words. This may result from the fact that these texts contain technical words that do not appear in WordNet.

Rodríguez and Esteban [Rodríguez97, Esteban98] describe the use of the WordNet ontology to improve a Rocchio and a Widrow-Hoff algorithm similar to Winnow. WordNet is used only for synonymy: its synsets are employed for category expansion to include closest synonyms but there is no use of the conceptual relations provided in the ontology. The expanded categories are included in the training set and fed to the algorithms in a Vector Space Model context. The test is performed Reuters news stories in 93 topics. The results are the following: without WordNet, the system achieves 29%. With the support of WordNet, a dramatic improvement to 50.2% precision is recovered. Although the improvement due to the use of WordNet synsets is spectacular in relative terms, the absolute categorization score remains low.

The overall conclusion from this research is that using a general ontology such as WordNet allows one to improve, sometimes significantly, the results of most categorization algorithms. We see two main conditions to improve classification results using ontologies:

- All features—synsets, descriptors, and relationships—of the ontology must be carefully considered.
- The ontology should be completed with domain-specific information tailored to the corpus at hand.

3 Patent categorization

3.1 Peculiarities of the IPC with respect to automated categorization

The IPC is a large hierarchical patent classification taxonomy divided and labelled in sections, classes, subclasses, groups and subgroups. At each sublevel of the taxonomy, the number of categories is multiplied by about 10, from 8 sections to approximately 69,000 subgroups.

Following the reform of the IPC, the classification will be divided into a stable core level of categories, and a frequently-updated advanced level. It is understood that small and medium-sized patent offices will classify patents using only the core set of categories.

An ideal categorization tool would classify each patent or patent application down to subgroup level, corresponding to around 69,000 categories, although this is a target currently beyond the state-of-the-art in fully automated categorization. However, a system that reliably assisted classification of documents to main group or subclass level would still be of vital importance, and would present a significant step forward.

	IPC taxonomy sample	
Section	**A SECTION A — HUMAN NECESSITIES**	
Subsection	**AGRICULTURE**	
Class	**A01 AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING**	
Subclass	**A01B SOIL WORKING IN AGRICULTURE OR FORESTRY; PARTS, DETAILS, OR ACCESSORIES OF AGRICULTURAL MACHINES OR IMPLEMENTS, IN GENERAL**	
References	(making or covering furrows or holes for sowing, planting or manuring A01C 5/00; machines for harvesting root crops A01D; mowers convertible to soil working apparatus or capable of soil working A01D 42/04; mowers combined with soil working implements A01D 43/12; soil working for engineering purposes E01, E02, E21)	
Index	Subclass Index	
	HAND TOOLS	1/00
	IMPLEMENTS USABLE EITHER AS PLOUGHS OR AS HARROWS OR THE LIKE	7/00
	OTHER MACHINES	27/00 to 45/00, 49/00, 77/00
	ELEMENTS OR PARTS OF MACHINES OR IMPLEMENTS	59/00 to 71/00
	TRANSPORT IN AGRICULTURE	51/00, 73/00, 75/00
	PARTICULAR METHODS FOR WORKING SOIL	47/00, 79/00

Table 2 : Portion of the IPC classification at the start of Section A

3.1.1 Technical difficulties in IPC categorization

Accurate patent categorization in the IPC is complicated by the following technical factors related to the nature of the categorization system [Adams00, IPC7]:

1. **References**: Many IPC categories contain references, which serve to guide the classification procedure, as illustrated in Table 2. There are two main types of references:
   - Limitation of scope: These references serve to restrict the patents classified in the category, and indicate related categories where some patents should be preferably placed.
   - Guidance: Lists related categories where similar patents are placed.

The references may list similar categories that are far-removed in the IPC hierarchy. In Table 2, for example, a reference to class E01 exists in subclass A01B. The IPC thus does not form a well-separated hierarchical tree, but one with a multitude of hyperlinks [Arcanum]. It is therefore probable that similar patents are associated with very different IPC codes.

After the reform of the IPC, a set of definitions is expected to provide better scope of the IPC categories than the current references, although their final number is yet unknown.

2. **Placement rules**: Patent classification is governed by additional placement rules. In certain parts of the IPC, a last-place rule governs the classification of documents relating to two adjacent categories (see for example C07), and indicates that the second of two categories should always be selected if two are found to concord. In other parts of the IPC, different specific rules hold (see for example B32B, where a first-place rule holds). Such ad-hoc rules may only be incorporated manually into a custom-built classification tool.
3. **Secondary codes**: Some patents do not have a single IPC code, but are associated with a set of secondary classification codes, relating to minor aspects expressed in the patent. Thus patent classification requires algorithms that go beyond simply selecting one pigeonhole from a given collection. They must be able to decide when a concept is relevant for each category individually.
4. **Indexes**: Portions of the IPC contain indexes, representing a separate classification of patents according to special aspects of the invention, such as the technique or technology employed. Some IPC classes and subclasses serve the double purpose of classification and indexing. A fully-automatic categorization tool would therefore need to support indexing and classification separately.
5. **Taxonomy updates**: Following the reform of the IPC, the taxonomy will be updated regularly, probably on a quarterly basis [Karetka02]. Although this is not foreseen to affect the core level of the IPC, where a classification tool would be most useful, a fully-functional classifier would need to support updates in the taxonomy and reclassification tasks.
6. **X-notation**: When a patent cannot be placed in a current subdivision of the IPC, the X-notation convention is currently used, indicating an insufficiency of the IPC. An automated tool would ideally detect when a patent belongs to a given category of the IPC, but not any of its children in the hierarchy, and then assign it an X-code.

Further issues that must be addressed are related to the nature of the documents that must be classified, namely patent applications:

7. **Language support**: The IPC is translated into several languages, and is used by a large number of patent offices around the world. The ideal categorization tool would support the categorization of patents in various languages with similar accuracy. In particular, English and French are of first importance, with other major European languages to be supported as a second priority [Karetka02]. Training sets in various languages will therefore need to be provided and tested. The availability of large training sets must be carefully considered, particularly for languages where few patents are published annually. If statistical algorithms are used, a few tens of documents will be required—at a minimum—in each category. It is doubtful whether a system trained in one language will be able to classify patents in another language, even if a translation dictionary is incorporated in the system.

8. **Vocabulary**: The terms used in patents are quite unlike other documents such as newspaper or scientific articles. Many vague or general terms are often used in order to avoid narrowing the scope of the invention. Combination of general terms may have a special meaning that it is important to identify. Patent documents also include acronyms and much new terminology [Kando00]. The claims section of a patent is often obscure by design, and set in a legalistic language. As this section is sometimes crucial for classifying patents, an automated system may have difficulty achieving the same level of understanding as a human classifier. In some countries, only the claims section of the patent is translated into English [Karetka02], which would probably not be sufficient for accurate automated classification if the system were trained in English only.
9. **Size variations**: The full text of a patent application varies strongly in length. US patents range in size from a few kilobytes to 1.5 megabytes [Larkey98]. The classification algorithm will have to account for this fact, possibly by truncating the full-text or selecting parts of it for indexing. If abstracts only are used for classification, such length variations could be avoided, although poorer discrimination may result on account of a less extensive vocabulary.

3.1.2 Help for IPC categorization

External sources of information are available to guide IPC classification. In particular:

1. **Catchword index**: A catchword index for the IPC is available in several major European languages. It lists relevant IPC categories for a large number of keywords and currently serves as a guide for manual classification. The French-English catchword index includes around 20,000 terms, while that developed by the German patent office in English and German contains over 120,000 terms. This information could prove most useful for a categorization tool, and appears not to have been exploited previously.
2. **Definitions**: Following the reform of the IPC, a set of definitions, similar to those in use in the US classification system, will be written for the IPC. This forthcoming information could be exploited to guide the classification.
3. **Bibliographic information**: A patent application contains bibliographic information, such as the name of the inventor and his/her company. By examining patents submitted by the same inventor or company, and already associated with an IPC code, an indication about the field of the new application can be guessed and may be of interest for an automated classifier. This could perhaps be simply automated by including the inventor and company in the full-text index of the patent. In the context of US patent classification, a tool was developed to assess the subclass distribution of referenced patents, and of further patents referenced in this set [Eshler01].

3.1.3 IPC categorization technical needs

Automatically categorizing patents requires some specific features that may not be necessary for general web page or news feed automated classification tools. In particular, the following points must be considered [Krier02]:

1. **User interface**: A patent categorization system will most likely suggest IPC codes rather than impose them, so categories must be ranked and confidence levels must be provided [Karetka02]. In this respect, the interface must be well designed. Links to the latest online IPC hierarchy should be provided to allow the user to browse IPC codes in the neighbourhood of the categories suggested by the system.

2. **Human assistance**: The categorization system must be able to recognise when its prediction are likely to be inaccurate. The possibility of discarding a fraction of patents from automated categorization and flagging them for purely manual categorization is an option that should be considered.
3. **Recall is more important than precision**: It is more important for a categorization assistance system to retrieve all possible categories (*i.e.* high recall), at the expense of suggesting some irrelevant ones (*i.e.* low precision), than to risk missing the ideal classification.
4. **Multiclassification**: As patents require several secondary IPC codes, classification cannot be performed by selecting a single category from a multitude of options. Instead, a more difficult task must be performed: each category must be considered separately, and judged whether relevant to the patent under review. Ideally, a main classification should be suggested and several additional ones proposed, if thought important.
5. **Document-based approach**: A classification system that works on the level of individual documents—rather than by relying on pre-built generic category descriptions—may be preferred because a list of similar published patents will be retrieved as part of the classification task. These documents may provide a starting point for a prior art search. Should a system relying on generic category digests be used, such as a Rocchio method, it would be useful for the user to additionally display similar pre-classified patents. One should consider whether similar patents should be retrieved only from the system training set, or also from subsequently categorized patents.
6. **OCR issues**: If patent applications are to be categorized automatically, electronic versions of the text must be available. If patents are not submitted electronically, as is often the case today, OCR must be performed to extract simple text versions from scanned sheets. Such character recognition is not without error. The categorization system should thus be tolerant towards to occasional garbled word [Hull01]. By training a categorizer with documents similarly affected by OCR errors, only little reductions in effectiveness were found in past studies [Ittner95].

3.2 Past research

In the following paragraphs, we summarize the results obtained by past investigations of patent categorization that have been reported in freely-available literature. Particular attention is paid to the resulting accuracies obtained with respect to IPC classification. As these accuracies depend strongly of the total number of categories in the taxonomy, these are also reported in each case.

One must take care not to compare accuracy numbers too quickly. In most cases, differences in the requested automated task explain much of the spread of results. For example, requiring single classification or multiclassification of patents can affect precision and recall strongly. In published literature, it is not always absolutely clear what all the details of the test were, leading to difficulties in meaningful comparisons between papers.

3.2.1 The EPO patent categorization tests

The European Patent Office has performed by far the most comprehensive test of patent categorization software [Krier02]. The EPO evaluated a number of generic classification tools applied to IPC patent classification tasks. The objective was to build a pre-classification tool designed to route patent applications to the correct EPO technical team of experts who search for prior art. The technical experts are divided into 44 directorates and 549 teams, each of which processes a range of IPC codes. It was found earlier that classification tasks at directorate level were 81.2% accurate when performed manually by administrative staff. The objective for the automated classifier was to perform at least as well.

Classifier evaluation was performed by supplying training sets and blind test sets to several commercial and academic partners, who performed the tests themselves. The results were reported back to the EPO, who compiled the results. Full comparative details have not been published, but a summary is given in [Krier02]. One participant reported his results separately [Koster01]. The tests are believed to have involved statistical algorithms only, none of which made any specific use of classification directives, such as those available in the IPC limiting definitions and references.

The tests were performed both with patent abstracts, and with the full text of published patents, where OCR errors had already been manually corrected. For each category, 2,000 training and 1,000 test documents were provided. The tests involved selecting a single directorate or team from those on offer (a mono-classification task). In this respect, the test differed from that of assigning IPC codes.

The following conclusions can be drawn from the exercise [Krier02]:

1. When using the full text of the patents, the precision was 2-9% higher than when using abstracts only.
2. At 100% recall, the precision was 72% at directorate level (44 categories) and 57% at team level (549 categories) for the Inxight Categorizer. These results refer only to the top-classified category, and would be higher if the suggested second choice was included as well.
3. At the requested level of 81.2% precision, a recall of 78% was achieved. In this case 22% of patents were not assigned to a directorate, and were flagged for manual classification.
4. Categorization speed was not a problem, but training the systems for taxonomies containing over 100 categories sometimes took over a week on a desktop PC.
5. The analysis of confusion matrices, showing where documents were erroneously classified, indicated that manual and automated procedures typically made similar mistakes. This probably results from overlapping categories (for example “organic chemistry” and “pharmaceuticals”) and is to be expected when purely statistical tools are applied to the IPC.

In Figure 5, the precision obtained with the Inxight categorizer is displayed as a function of the number of categories, shown in a logarithmic scale. A linear trend of decreasing accuracy is then obtained. It should be noted that these results are given at 100% recall, which may be an exaggerated requirement in a production setting. By lowering the required recall, the precision shown in Figure 5 rises.

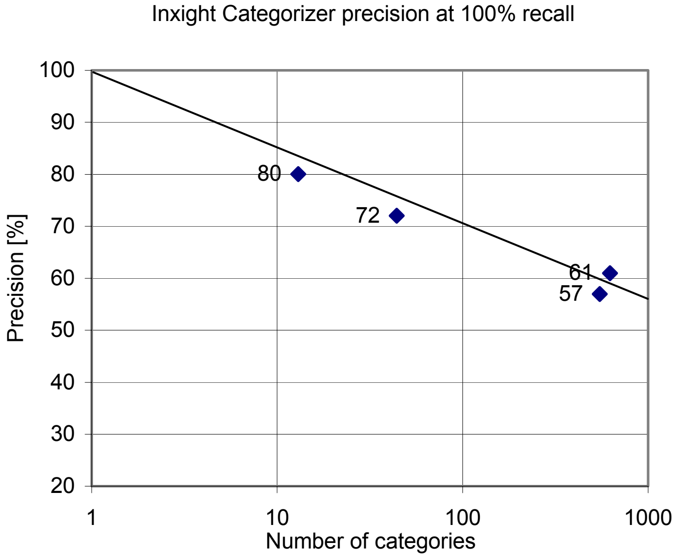


Figure 5 : Patent full-text automated categorization results with Inxight, as a function of the number of categories, redrawn using data from [Krier02]. The solid line is a guide to the eye.

Detailed results have been published by one of the participants [Koster01], who compared a Rocchio algorithm with Winnow, a learning algorithm that refines its discrimination between relevant and non-relevant documents as the training set is presented [Sebastiani02]. As the number of training documents increases, the precision and recall of the classification improved continuously for categorization by abstract, with Winnow outperforming Rocchio for larger training sets. This highlights the importance of providing large training sets. Over all tests imposed by the EPO, and compared to all commercial classifiers, the Winnow algorithm performed best [Koster01].

Experiments with word stemming have shown that it increases precision but lowers recall for full-text patent categorization, but it lowers precision and increases recall for patent abstract categorization [Koster01]. There is thus a difference between long and short versions of patents, probably related to the different vocabulary employed. Indeed, the vocabulary is much more diverse in the full text of patents than in the abstracts [Koster02].

An investigation into the best technique for selecting terms from the full vocabulary for discriminating between patent categories has also been published [Koster02]. Although small differences are found, the best techniques are related to the information gain related to each term, particularly when only small numbers of discriminating terms are retained (under 50 terms per class). These sophisticated measures are derived from information theory [Sebastiani02].

Assigning IPC codes to patents may benefit from extracting the IPC codes of earlier patents cited in the patent application, which should often prove to be similar. Tests in this direction are ongoing at the EPO. An 80% success rate at extracting patent citations from the full text of patent applications has already been reported [Krier02].

Overall, most categorization packages tested by the EPO had similar accuracies [Koster01]. Following the testing phase, an automated patent pre-classifier has been installed at the EPO, based on Inxight software.

3.2.2 Patent categorization research in the US

The USPTO patent classification system consists of around 400 classes and over 135,000 subclasses arranged in a hierarchy of varying depth. Each subclass can contain up to 2,000 patents, after which new subclasses are added to subdivide the set further. Most subclasses only contain about 20 patents [Larkey98]. US patent classification principles are different from those applied to the IPC, vary throughout the classification system, and are explained in the literature [Falasco02]. Some simple tools for assessing the consistency of a patent subclass by examining the distribution of reference patents have been developed as part of a student project [Eshler01].

A hierarchical patent classification system, restricted to 12 US subclasses organized in three levels and concerned with Communication, Electricity, and Electronics, was developed by researchers at IBM [Chakrabarti97, Chakrabarti98]. They used Bayesian algorithms and auto-generated small sets of discriminating words at each node of the hierarchical taxonomy. The best number of discriminating words depended strongly on the category and varied between 160 and 9130 terms [Chakrabarti98]. In a small-scale test involving 500 training patents and 300 testing patents per subclass, the average recall of the hierarchical classifier was 66%, which was higher than that of a flat classifier used for comparison. It is unclear whether the full-text of the patents was employed or not. By comparing document models that account for the reoccurrence of words in the text (Bernoulli model) with those that do not (Binary model), it was established that accounting for reoccurrences improved classification accuracy by about 8%.

Another interesting conclusion of this research is that the classification of patents may be significantly harder than that of standard Reuters news articles. Indeed, the categorization tool developed in [Chakrabarti97] achieved 87% accuracy when classifying news articles in 30 categories, but only 66% accuracy with 12 categories of US patents.

In [Chakrabarti98b], the authors attempt to improve their results by including patent hypertext information, *i.e.* basing the categorization of a patent not only on its textual content, but also on information in patents that it cites and in patents citing it. This approach is implemented when training and testing the system by considering the full set of linked documents as a whole rather than by presenting each document separately to the categorizer. In this respect, the classification tool developed is unique.

If citing and cited documents are taken into account by indexing their words and terms in conjunction with those of the primary patent, results are found to be disappointing and the categorization effectiveness is reduced. This is caused by a lack of category specificity in the words contained in all citing and cited patents. However, if the categories of the citing and cited documents are used as indexing information for the primary patent, the classification accuracy is much improved, sometimes halving the error rate [Chakrabarti98b]. The system performs a primary classification based only the textual content of the patent, and then refines its estimates from the hyperlink information. It should be pointed out that this technique requires sophisticated processing that has yet to be implemented in any commercial product the authors of this report are aware of.

Leah Larkey, at the University of Massachusetts, has developed a system to classify patents according to the US patent classification system [Larkey98, Larkey99]. The approach chosen makes use of stopword removal, stemming, and combines a k-NN scheme with a Bayesian algorithm. The inclusion of phrases in the terms representing the patents was not found to be beneficial [Larkey98]. The accuracy of hierarchical classifiers, when tested in a small portion of the full US database concerned with speech-related patents, was not found to be better than that of a flat classifier. The system makes use of the Inquery engine [Callan92], an academic effort that has since ceased being developed.

A full patent classification assistance prototype on a subclass level is reported to have been built [Larkey99]. The on-line user interface includes a natural-language query entry box with an option for adding similar co-occurring phrases and compound terms automatically. These additional terms have been built by processing patent documents automatically. Additional phrases can be added at the user's discretion, by manually selecting relevant ones from an automatically-compiled list. The system thus represents a first step towards developing a patent ontology.

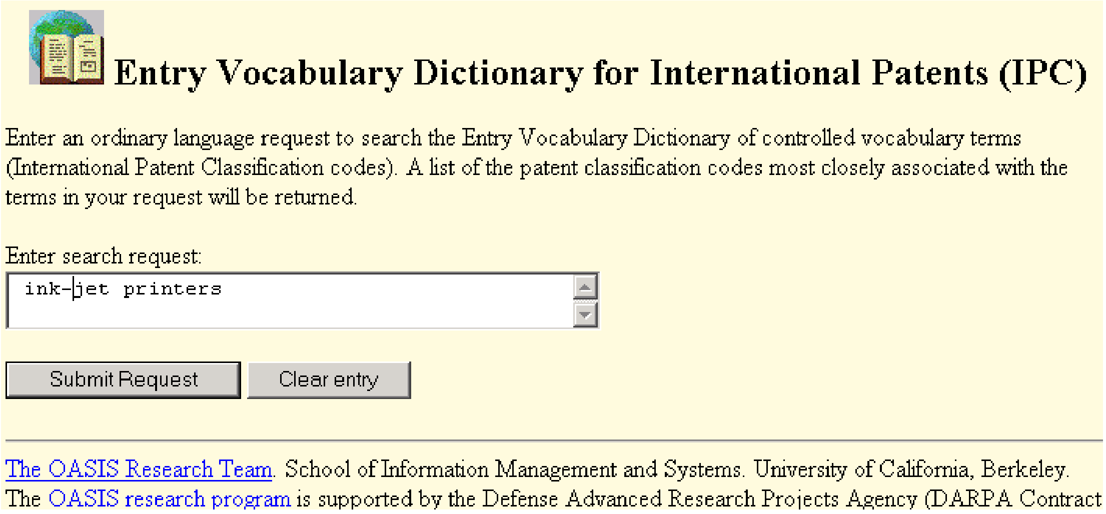
The full text of the patents is not indexed, but only some sections or portions of sections are examined, with weights reflecting their importance. Best performance is reported using the title, the abstract, the first twenty lines of the background summary, and the claims section to represent each patent [Larkey99].

The classification system uses a k-NN algorithm, which requires comparing the search string with a set of the most similar training patents. Because the database of patents is huge, containing over 5 million patents divided into 400 topically-organized document collections [Larkey00], the patents are retrieved in a two-step procedure. First the search string is compared to a set of 400 virtual documents each describing a document collection, and then the 10 best document collections are searched for similar patents. Tests were performed to determine whether the document collections should be separated by date or by topic [Larkey00]. Topical organisation proved better at retrieving similar patents from the whole corpus. Because of the computational power needed, the system was built around a four-processor Sun workstation in 1999 [Larkey99]. The subclasses of similar patents are ranked, and a final subclass is proposed to the user.

The system is not currently made available to the public and has not been tested by the authors of this report. No details about the system's precision and recall at subclass level have been published. It is believed this system is not currently in use at the USPTO [Karetka02].

The University of California OASIS research team has built a different automated system suggesting a variety of patent classification codes. It is available online at metaphor.sims.berkeley.edu/oasis/patents.html, for US patent classification, and at metaphor.sims.berkeley.edu/oasis/ipc.html, for IPC classification codes, as shown in Figure 6 [Gey99]. A natural-language search facility is provided, where the user can formulate a query. This string is automatically augmented with a set of associated terms derived from a controlled vocabulary. Options for searching by phrase or by word are provided in the user interface. The result is output as a set of suggested patent classification codes most closely matching the search query. For the IPC, subgroup codes are suggested. Links are then provided to USPTO and WIPO patent databases to retrieve patents from the chosen categories.

In practice, the system is difficult to use, as 10 subgroups are always returned. It is also unclear what the ranking of the proposals is, as no confidence levels are provided.



Entry Vocabulary Dictionary for International Patents (IPC)

Enter an ordinary language request to search the Entry Vocabulary Dictionary of controlled vocabulary terms (International Patent Classification codes). A list of the patent classification codes most closely associated with the terms in your request will be returned.

Enter search request:

ink-jet printers

The OASIS Research Team. School of Information Management and Systems. University of California, Berkeley. The OASIS research program is supported by the Defense Advanced Research Projects Agency (DARPA Contract

Figure 6: OASIS IPC classification tool, from metaphor.sims.berkeley.edu/oasis/ipc.html

At metaphor.sims.berkeley.edu/oasis/ipc2uspc.html, an interface attempting to map IPC codes to US codes, but not vice-versa, is also provided by the formerly-named OASIS group, now know as the Metadata research Program. A list of US patent codes relating to an IPC code is returned from the search, but no details about how the match is constructed have been found.

In a recent workshop [Lewis01], a presentation was made about patent categorization at the Ford Motor Company, where competitor monitoring is the primary purpose. Commercial tools and research software were evaluated to classify patents in 4,000 categories. Tests indicated that support vector machines outperform k-NN nearest-neighbour approaches, both of which give better results that a Naïve Bayes algorithm. It was found that by collapsing sparsely-populated categories together, better results were achieved. No further details have been published about this project's results.

3.2.3 Patent categorization research in Japan

A patent categorization test at the JPO has been reported in [Mase98]. Unfortunately, it has not been possible to locate a copy of this report. The website of the journal in question further indicates that a majority of its contributions are in Japanese. Therefore, we reprint here only the research abstract, which is available on the Internet:

> "This paper presents keywords-based patents categorization and discusses its simulation study. […] We propose a classification knowledge generation method, which extracts keywords that characterize the particular category from a lot of patent documents […]. We also propose keyword extraction and ranking method based on the structure of patent documents and the syntactics of the sentence. We did experimental simulation using maximum 310,000 training patent documents. The maximum classification accuracy was 96.0% (38 categories) and 82.8% (2,815 subcategories) when three categories are assigned to each document. We also evaluated how much training data is necessary from the viewpoint of classification knowledge maintenance. The results show that approximately 1,000 patent documents per each subcategory were necessary to classify most correctly and effectively. The results of this simulation strongly encouraged us to develop a patent classification system to support the category assignment work." (Reprinted from [Mase98]).

This work has again highlighted the importance of providing large training sets, and apparently achieves excellent accuracy. It would thus be interesting to contact the authors for further details. From the information currently available, it is unclear whether the system supports patents in English or Japanese.

The Industrial Property Cooperation Center, an affiliated organization of the JPO, also presented the OWAKE project, a primary automatic classification system, at WIPO in March 2000 [Owake00]. The system is reported to perform patent preclassification on the basis of morphological, syntactic, and semantic analyses. It attributes IPC categories and one of around 3,000 themes. The accuracy of the system is said to need improvement to perform reliable categorization at the IPC main group level [Owake00]. Difficulties are attributed to the treatment of proper nouns, accounting for the placement of words in the title or the abstract, and accounting for word frequencies. The authors feel that including synonyms and other relations between words might improve the performance of the system. This presentation may have been reporting the similar work as that detailed in the preceding paragraph.

Researchers at KDD Laboratories have developed a patent retrieval system [Inoue00]. While not designed expressly for document categorization, it is capable of locating patents similar to a given search string—or a given search patent—based on Bayesian techniques exploiting term similarity in the documents. Presumably, by examining the IPC codes of the pre-classified patents that are found, a guess at the optimal classification of the search patent can be determined, although the authors did not consider this application. Because of the need to compare a search patent with a large number of pre-classified patents to locate similarities, the authors use a cluster-based hierarchical search system to speed-up the exhaustive search. Performance and quality comparisons with exhaustive searching are unfortunately not reported. At typical levels of thresholding, the system is reported to retrieve around 80% of patents similar to a given search patent, based on the comparison with similar patents manually located by a panel of experts.

Another patent retrieval system based on the syntactic analysis of Japanese patent documents is presented in [Hyoudo98]. The accuracy of patent retrieval is improved by analysing sentence semantics when compared to simple term proximity testing in the searched patents of the search string. Reported precision and recall are high, at 92% and 96% respectively, although only a small-scale test of 10 queries was performed and the number of patents searched is not reported. The application of patent retrieval to categorization was not considered in this work.

3.2.4 Patent maps

An issue related to the categorization of documents in general, and patents in particular, is the graphical representation of sets of documents. So-called Kohonen maps are projections on a two-dimensional plane of a set of documents, in which related documents are placed nearby [Lagus98, Kohonen00]. Document similarity is established by comparing word or term distributions. To produce such maps, it is therefore necessary to perform a document analysis involving word frequencies that is similar to that commonly used for document categorization.

An interesting application of concept mapping to the field of patents has been published recently [Kohonen00]. Over 6,000,000 US, Japanese, and European patent abstracts in all sections of the IPC have been represented on a single map, allowing a user to browse patent concepts and drill down to a list of similar patents. To establish this map required several weeks of computation on a multiprocessor supercomputer. The resulting patent-browsing application is not made public at this time, although a similar one relating to newsgroups is online at websom.hut.fi/websom.

In [Kohonen00], an example of a patent map is shown. The map is labelled automatically with concept words. The user can zoom in to two level of detail by clicking on the map. At the third level of detail, the individual patent abstracts are displayed. A search for the term “color display“ is shown and relevant areas are automatically located on the map. In one of these, patents from IPC group G02F 1/1335 are in evidence. If one was therefore trying to place a new patent in an existing IPC group, one might try to search the map taking some patent keywords as a search string and to examine the groups of neighbouring pre-classified patents.

In the course of their research, the authors of [Kohonen00] tested the categorization of patent abstracts. They obtained an accuracy of 60.6% when classifying patent abstracts in the 21 subsections of the IPC. Because of the enormous workload required to process millions of documents, they explored techniques for speeding up patent categorization by randomly reducing the number of terms used in the description of each document. In essence, this corresponds to a projection on a randomly-oriented hyper-plane in the space of vectors characterizing each patent abstract. They found that such projections did not dramatically reduce the accuracy of the categorization, which was then around 55-59%, but provide enormous computational advantages.

Another application of patent mapping has recently been performed, but on a much-reduced set of patents related to engine oil technology [Lamirel01]. In this research, published in French, a technique for creating a set of maps providing different viewpoints of the same collection of patents is developed. Each map groups patents according to different criteria, such as the uses, the advantages, the titles, or the authors of the patents. Techniques for cross-map browsing are also developed.

3.3 Patent-specific software

Below, we list software products that are (or will be) developed specifically for patent-related retrieval and classification tasks.

3.3.1 USPTO PLUS

The USPTO currently makes use of the PLUS system (Patents Linguistic Utility Service) [Smith02], which operates on a query by example basis. Word lists are extracted from patent application titles, abstracts, and portions of the full-text. Stopwords are then removed and the lists are compared with those extracted from pre-classified documents to retrieve a ranked list of similar patents.

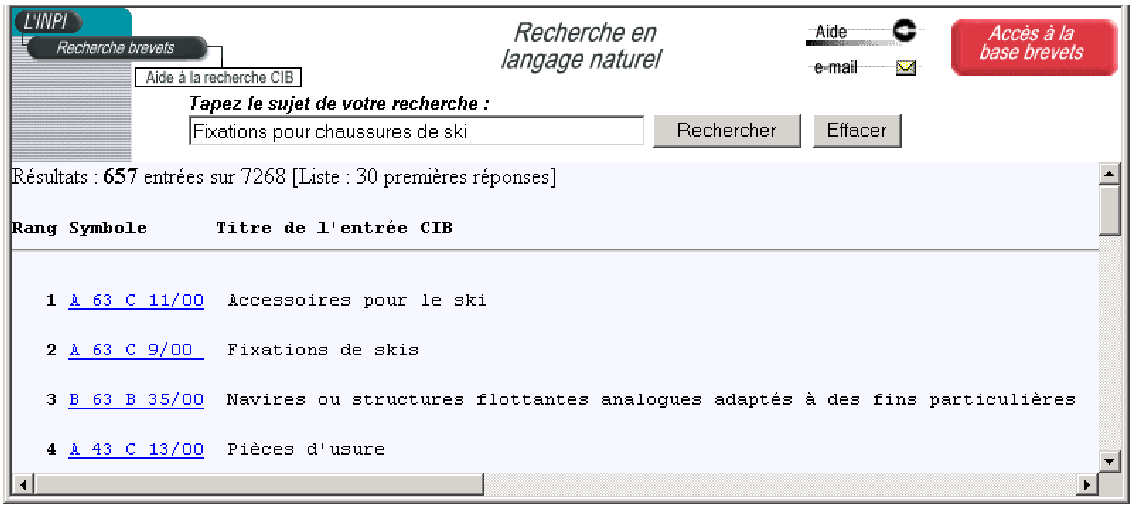
Although the system is designed primarily to aid with prior art searching, USPTO staff use of the codes of the ranked similar patents to guide the classification of new patent applications.

3.3.2 Lingway TACSY – INPI CIB-LN

Lingway (www.lingway.com), a French company formed by collaborators of Lexiquest, previously known as Erli, has developed a Taxonomy Access and Coding System (TACSY) allowing natural language access to a taxonomy category. This has been applied to the IPC categories. The user can formulate a natural-language query in French, consisting of a sentence describing the field of interest, and the software suggests the best predictions of the corresponding IPC subgroups, with a set of confidence levels. Category indexing is first performed through a linguistic indexation of the category descriptions, without relying on any pre-classified documents. To query the category index, a linguistic analysis of the search phrase is performed, polysemic words are contextually disambiguated, and equivalent words are additionally included. This allows all relevant categories to then be located.

The same software, or a very similar version of it, is known as CIB-LN (an abbreviation of the French equivalent of “Natural Language IPC access”), and is available online for free at the INPI website: www.inpi.fr/Cibln/. An overview of the system has been published in [Lyon99] and a technical description is also available

[Leclercq99]. Testing by the EPO with 350 queries resulted in an average of 79% category recall with 55% of the correct classifications appearing in the top 20 answers given by the system [Lyon99]. Parts of the IPC have been omitted, notably in the Chemistry section (IPC section C). An example of the system use is shown in Figure 7.



Tapez le sujet de votre recherche : Fixations pour chaussures de ski

```
Résultats : 657 entrées sur 7268 [Liste : 30 premières réponses]

Rang Symbole      Titre de l'entrée CIB

  1 A 63 C 11/00  Accessoires pour le ski

  2 A 63 C 9/00   Fixations de skis

  3 B 63 B 35/00  Navires ou structures flottantes analogues adaptés à des fins particulières

  4 A 43 C 13/00  Pièces d'usure
```

Figure 7: Using INPI CIB-LN with a French version of the IPC, from www.inpi.fr/Cibln

The linguistic analysis is performed with a tool from Lexiquest, which generates Boolean search strings sent to a Verity 97 search engine [Leclercq99]. The system is based on a 50,000-word dictionary including 35,000 concepts [Normier02]. Approximately 7,300 categories are searched in CIB-LN, corresponding to a search down to main group level of the IPC. A hyperlinked result set is auto-generated allowing subgroups to be examined manually. It currently handles 30,000 queries a month [Normier02].

Problems associated with this system derive from the fact that no use has been made of patent documents in the construction of the tool. Furthermore, it is unclear how references present in the IPC have been handled.

3.3.3 e-Patent

The e-Patent project (www.eu-projects.com/epatent/) aims to extend the French INPI system to English, Spanish, and German, and provide better ranking and translation aids [Normier02]. It is developed by a collaboration between INPI, the UK, Spanish, and German patent offices, Lingway and Jouve, a French software development company active in information management. The project started in January 2002 and will run for 2 years. It benefits from a €2.4-million investment.

Based on multilingual semantic networks, the project aims to interpret natural language queries to retrieve patents. When a user enters a natural-language question, the system will expand the query using its ontology, built from the full-text of patent collections, and perform a search on both the IPC definitions and the patent database, which will have been converted to a syntactically-tagged XML format in advance [Normier02]. Merging the results of both searches will improve the final accuracy of the tool.

A system of automated patent translation making use of the same multilingual ontology is also foreseen [Normier02].

Currently under development and beta testing, the full e-Patent system should be made available to the public by the end of 2002 in English, with translations to follow in 2003.

3.3.4 Derwent

Derwent runs a commercial patent classification system (www.derwent.com). This classification is of a much simpler nature than the IPC, with around 1,000 categories, and is marketed as a replacement or an alternative to the IPC.

Derwent's domain experts re-categorize all patents according to their system. Patents are divided into 21 broad subject areas or Sections. These are designated A-M (Chemical); P-Q (Engineering); and S-X (Electronic and Electrical). These Sections are then further subdivided into Classes. Each Class consists of the Section letter, followed by two digits. A subset example of the classification system is shown in Table 3.

Derwent taxonomy sample
A Polymers and Plastics
A1 Addition and Natural Polymers
A2 Condensation Polymers
A3 Processing: General Additives and Applications
A31 Preliminary processes.
A32 Polymer fabrication – such as moulding, extrusion, forming, laminating, spinning.
A35 Other processing and general – including vulcanisation, welding of plastics and adhesive processes. Testing.
A41 Monomers and Condensants
A60 Additives and Compounding Agents
A8/9Applications
B Pharmaceuticals
C Agricultural Chemicals
D Food, Detergents, Water Treatment and Biotechnology
E General Chemicals
F Textiles and Paper-Making
G Printing, Coating, Photographic
H Petroleum
J Chemical Engineering
K Nucleonics, Explosives and Protection
L Refractories, Ceramics, Cement and Electro(in)organics
M Metallurgy

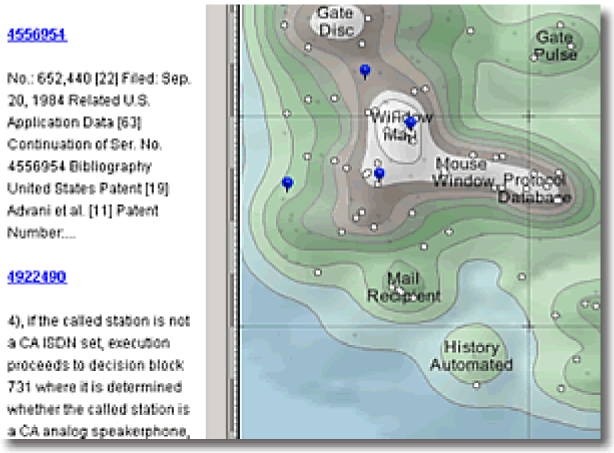
Table 3 : Subset of the Derwent patent classification system

A natural-language online patent search facility is available, containing patent information from 40 different sources, including national patents from most major industrial countries. Patents are summarized and the titles are rewritten to make them more meaningful. IPC codes are listed in published abstracts and are sometimes corrected if patents are misclassified by US patent examiners [Stembridge98]. Comparisons between IPC classifications performed by various national patent offices (other than the USPTO), which are all listed in the Derwent databases, have shown that inconsistencies in IPC subclass or group level classification sometimes happen [Stembridge98]. This illustrates that human categorization is not without error.

There is no online tool for automated Derwent code categorization that the authors of this report are aware of.

3.3.5 Aurigin Aureka

The client/server version of Aureka, marketed by Aurigin (www.aurigin.com), a division of MicroPatent USA (www.micropat.com), consists of an extensive set of patent databases along with a selection of sophisticated analytical tools. A subscription is required to access the system and receive updated of patent content. The Aureka system provides enhanced, structured access to US, EPO, and Japanese patent data, as well as EPO, Japanese, and PCT patent application data.



4556954

No.: 652,440 [22] Filed: Sep. 20, 1984 Related U.S. Application Data [63] Continuation of Ser. No. 4556954 Bibliography United States Patent [19] Advani et al. [11] Patent Number:...

4922490

4), if the called station is not a CA ISDN set, execution proceeds to decision block 731 where it is determined whether the called station is a CA analog speakerphone,

Figure 8: Aurigin themescape demonstration, from www.aurigin.com/aureka.html

Aurigin provides a so-called themescape, shown in Figure 8, similar to the Kohonen maps presented in paragraph 3.2.4, where patents are located according to their textual similarities. The system can analyse patent abstracts, the claims section, the title, or the full-text of patents to find similarities. Details about the algorithms used for finding similarities are not publicly available.

A citation tree analysis tool allows referenced patents to be visualized in a hyperbolic viewer, thereby allowing branches interesting to the user to be explored more deeply.

3.3.6 VxInsight

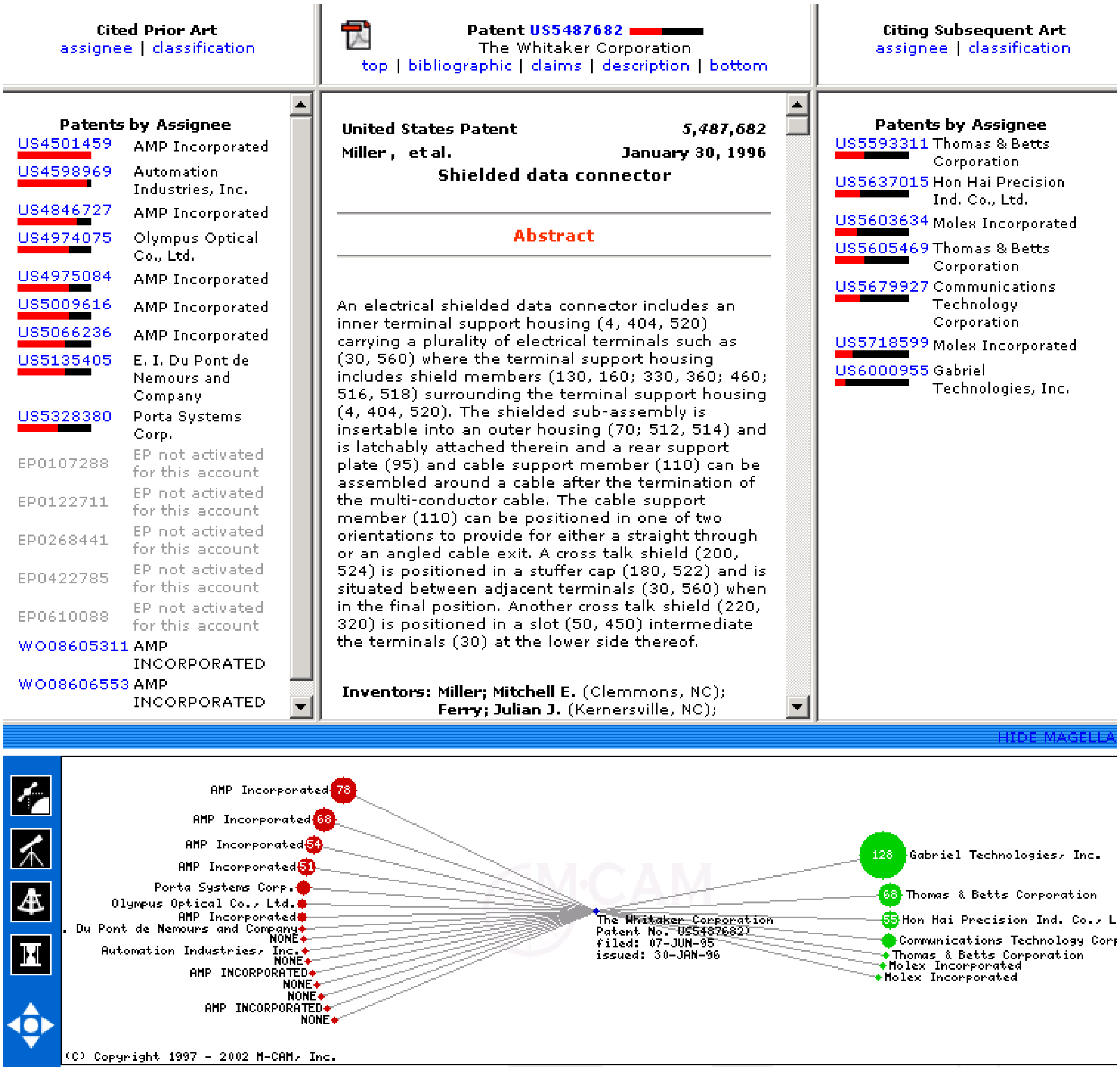
VxInsight is a generic tool for discovering relationships within very large databases. It was developed at the US Sandia National Research Laboratories, and can be licensed for private use. There may be licensing requirements for the export to or use by citizens of certain countries. VxInsight runs on both NT-based PCs and on low-end SGI 02 desktop workstations. It works with SQL-based relational databases.

An application of the tool to the analysis of US patents has been published [Boyack00]. While the tool produces similar maps to those of Aurigin Aureka, it computes similarities between patents not by relying on textual content, but only by analysing citations between patents.

An example shown in [Boyack00] displays 10’805 patents issued by the US Patent Office in January 2000. The height of the mountain peaks gives some indication of the number of patents within them. The patents are seen to be mostly grouped together in isolated peaks, each of which corresponds to a patent class in the classification system. Where peaks merge together, the corresponding classes can be expected to cover similar material, which automated categorization tools may have more difficulty separating.

3.3.7 M-CAM Doors

M-CAM, headquartered in Virginia (www.m-cam.com), produces the M-CAM Doors software, which is a powerful searchable database of the world’s patents, available through online web access, and run on a commercial basis. A typical screenshot from the demonstration package is shown in Figure 9. The full text of a patent is displayed, along with lists and graphs of citing and cited reference patents.



Cited Prior Art
assignee | classification

Patents by Assignee

US4501459	AMP Incorporated
US4598969	Automation Industries, Inc.
US4846727	AMP Incorporated
US4974075	Olympus Optical Co., Ltd.
US4975084	AMP Incorporated
US5009616	AMP Incorporated
US5066236	AMP Incorporated
US5135405	E. I. Du Pont de Nemours and Company
US5328380	Porta Systems Corp.
EP0107288	EP not activated for this account
EP0122711	EP not activated for this account
EP0268441	EP not activated for this account
EP0422785	EP not activated for this account
EP0610088	EP not activated for this account
WO08605311	AMP INCORPORATED
WO08606553	AMP INCORPORATED

Patent US5487682
The Whitaker Corporation
top | bibliographic | claims | description | bottom

United States Patent **5,487,682**
Miller , et al. **January 30, 1996**
Shielded data connector

Abstract

An electrical shielded data connector includes an inner terminal support housing (4, 404, 520) carrying a plurality of electrical terminals such as (30, 560) where the terminal support housing includes shield members (130, 160; 330, 360; 460; 516, 518) surrounding the terminal support housing (4, 404, 520). The shielded sub-assembly is insertable into an outer housing (70; 512, 514) and is latchably attached therein and a rear support plate (95) and cable support member (110) can be assembled around a cable after the termination of the multi-conductor cable. The cable support member (110) can be positioned in one of two orientations to provide for either a straight through or an angled cable exit. A cross talk shield (200, 524) is positioned in a stuffer cap (180, 522) and is situated between adjacent terminals (30, 560) when in the final position. Another cross talk shield (220, 320) is positioned in a slot (50, 450) intermediate the terminals (30) at the lower side thereof.

Inventors: Miller; Mitchell E. (Clemmons, NC);
Ferry; Julian J. (Kernersville, NC);

Citing Subsequent Art
assignee | classification

Patents by Assignee

US5593311	Thomas & Betts Corporation
US5637015	Hon Hai Precision Ind. Co., Ltd.
US5603634	Molex Incorporated
US5605469	Thomas & Betts Corporation
US5679927	Communications Technology Corporation
US5718599	Molex Incorporated
US6000955	Gabriel Technologies, Inc.

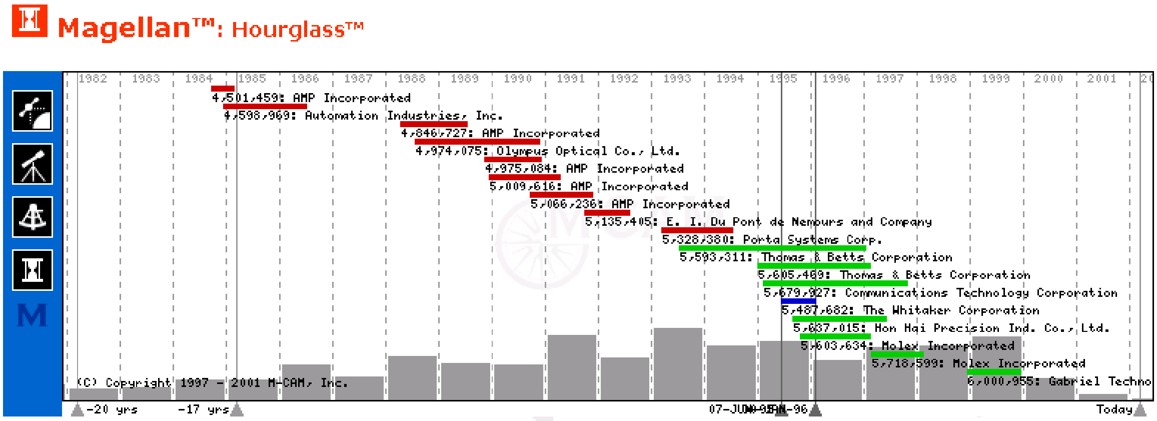
HIDE MAGELLA

(C) Copyright 1997 - 2002 M-CAM, Inc.

Figure 9: M-CAM Doors sample page, from doors3.m-cam.com

Features are available for patent searching, for creating graphs of related patents, for finding similar patents, for historical analyses, and for concept querying. In Figure 10, a demonstration of one of these features is shown.

Tools for concept querying by semantic indexing, for identifying claims uniqueness, and for analysing patent groups are available on subscription but cannot be tested freely.



4,501,459: AMP Incorporated
4,598,969: Automation Industries, Inc.
4,846,727: AMP Incorporated
4,974,075: Olympus Optical Co., Ltd.
4,975,084: AMP Incorporated
5,009,616: AMP Incorporated
5,066,236: AMP Incorporated
5,135,405: E. I. Du Pont de Nemours and Company
5,328,380: Porta Systems Corp.
5,593,311: Thomas & Betts Corporation
5,605,469: Thomas & Betts Corporation
5,679,927: Communications Technology Corporation
5,487,682: The Whitaker Corporation
5,637,015: Hon Hai Precision Ind. Co., Ltd.
5,603,634: Molex Incorporated
5,718,599: Molex Incorporated
6,000,955: Gabriel Techno

(C) Copyright 1997 - 2001 M-CAM, Inc.

Figure 10: Temporal distribution of patents in a given US classification subclass, together with a list of patent assignees, from doors3.m-cam.com

3.3.8 Delphion

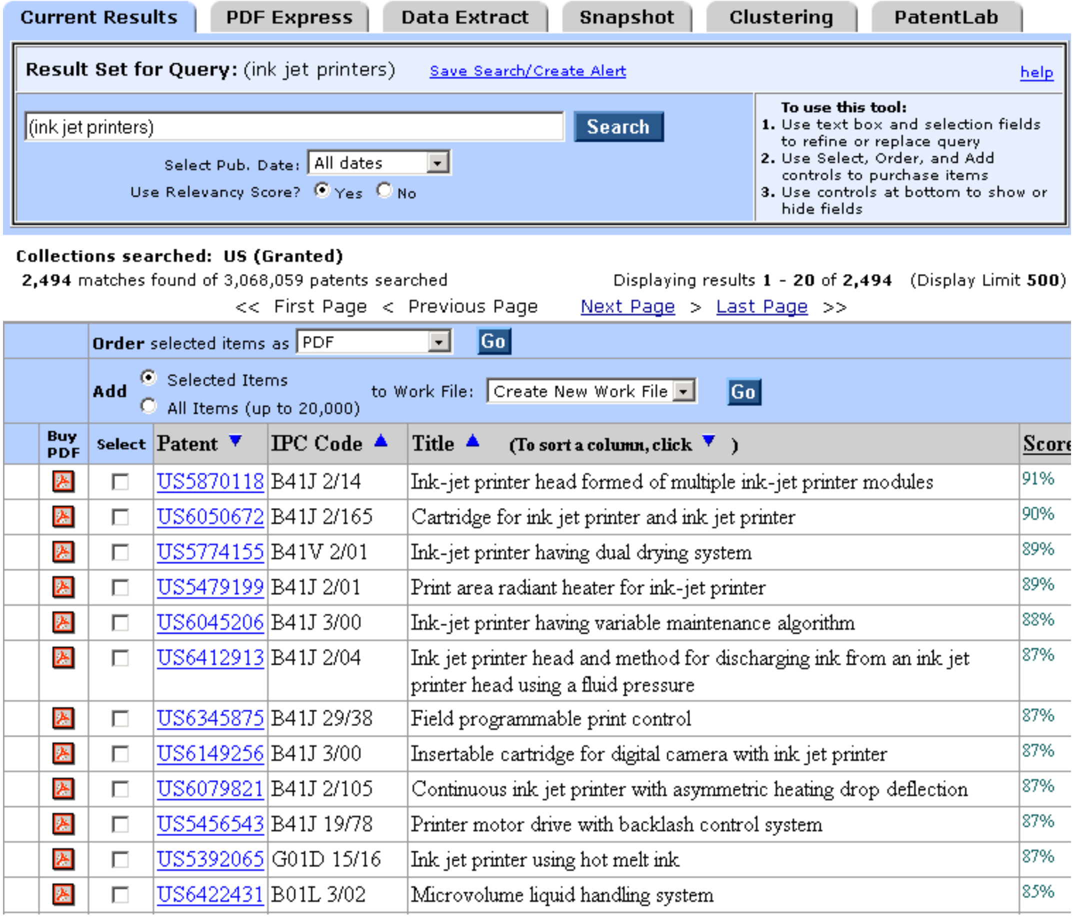
Delphion (www.delphion.com) is an IBM spin-off company now running the old IBM patent database. After subscription, facilities are available for querying various databases, retrieving PDF versions of patents, graphing citation lists, automated alerting, and patent clustering. Data sources consist of: US granted patents and patent applications, European granted patents and patent applications, WIPO PCT publications, JPO abstracts, and INPADOC files.

In Figure 11, a list of US patents related to "ink-jet printers" has been displayed, together with their IPC codes. Facilities for clustering the results, based on linguistic analysis of the textual content are available, as shown in Figure 12, when a subscription to the service is granted.

During clustering, documents are assigned uniquely to one defined cluster. Clusters of similar documents are displayed along with the extracted keywords that characterize each cluster. Results can be visualized graphically with a map that provides an overview of the clusters and an indication of the relationship among them. The most relevant clusters can be focused on and a drill-down procedure into any cluster allows individual documents to be viewed.

Frequent access to patent clustering requires a $200 per month unlimited subscription. Corporate discounts are available for 2, 10, 100, and 1000 users from the same company.

Wisdomain (www.wisdomain.com) distributes PatentLabII freely, a simple tool for analysing patent data downloaded from the Delphion website. This free product does not support topic analysis or categorization. However, Wisdomain's more recent commercial Focust product has a module for clustering patents systematically by similarity.



Result Set for Query: (ink jet printers) Save Search/Create Alert help

(ink jet printers) Search

Select Pub. Date: All dates

Use Relevancy Score? ● Yes ○ No

To use this tool:
1. Use text box and selection fields to refine or replace query
2. Use Select, Order, and Add controls to purchase items
3. Use controls at bottom to show or hide fields

Collections searched: US (Granted)

2,494 matches found of 3,068,059 patents searched

Displaying results **1 - 20** of **2,494** (Display Limit **500**)

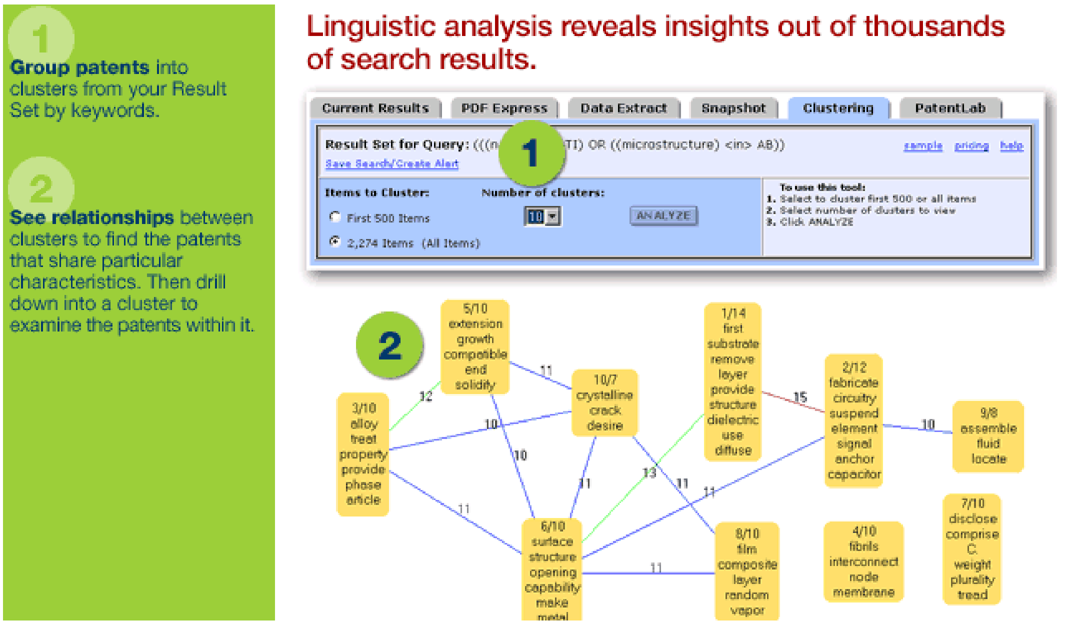
<< First Page < Previous Page Next Page > Last Page >>

Order selected items as PDF Go

Add ● Selected Items ○ All Items (up to 20,000) to Work File: Create New Work File Go

Buy PDF	Select	Patent ▼	IPC Code ▲	Title ▲ (To sort a column, click ▼)	Score
	□	US5870118	B41J 2/14	Ink-jet printer head formed of multiple ink-jet printer modules	91%
	□	US6050672	B41J 2/165	Cartridge for ink jet printer and ink jet printer	90%
	□	US5774155	B41V 2/01	Ink-jet printer having dual drying system	89%
	□	US5479199	B41J 2/01	Print area radiant heater for ink-jet printer	89%
	□	US6045206	B41J 3/00	Ink-jet printer having variable maintenance algorithm	88%
	□	US6412913	B41J 2/04	Ink jet printer head and method for discharging ink from an ink jet printer head using a fluid pressure	87%
	□	US6345875	B41J 29/38	Field programmable print control	87%
	□	US6149256	B41J 3/00	Insertable cartridge for digital camera with ink jet printer	87%
	□	US6079821	B41J 2/105	Continuous ink jet printer with asymmetric heating drop deflection	87%
	□	US5456543	B41J 19/78	Printer motor drive with backlash control system	87%
	□	US5392065	G01D 15/16	Ink jet printer using hot melt ink	87%
	□	US6422431	B01L 3/02	Microvolume liquid handling system	85%

Figure 11: Delphion search results showing lists of IPC codes, from www.delphion.com/research



1 **Group patents** into clusters from your Result Set by keywords.

2 **See relationships** between clusters to find the patents that share particular characteristics. Then drill down into a cluster to examine the patents within it.

Linguistic analysis reveals insights out of thousands of search results.

Result Set for Query: (((n TI) OR ((microstructure) <in> AB)) sample pricing help

Save Search/Create Alert

Items to Cluster: ○ First 500 Items ● 2,274 Items (All Items)

Number of clusters: 10 ANALYZE

To use this tool:
1. Select to cluster first 500 or all items
2. Select number of clusters to view
3. Click ANALYZE

5/10 extension growth compatible end solidify
10/7 crystalline crack desire
1/14 first substrate remove layer provide structure dielectric use diffuse
2/12 fabricate circuitry suspend element signal anchor capacitor
9/8 assemble fluid locate
3/10 alloy treat property provide phase article
6/10 surface structure opening capability make metal
8/10 film composite layer random vapor
4/10 fibrils interconnect node membrane
7/10 disclose comprise C weight plurality tread
11
12
10
10
15
10
13
11
11
11
11
11

Figure 12: Delphion features facilities for grouping patents by concepts, from www.delphion.com/research

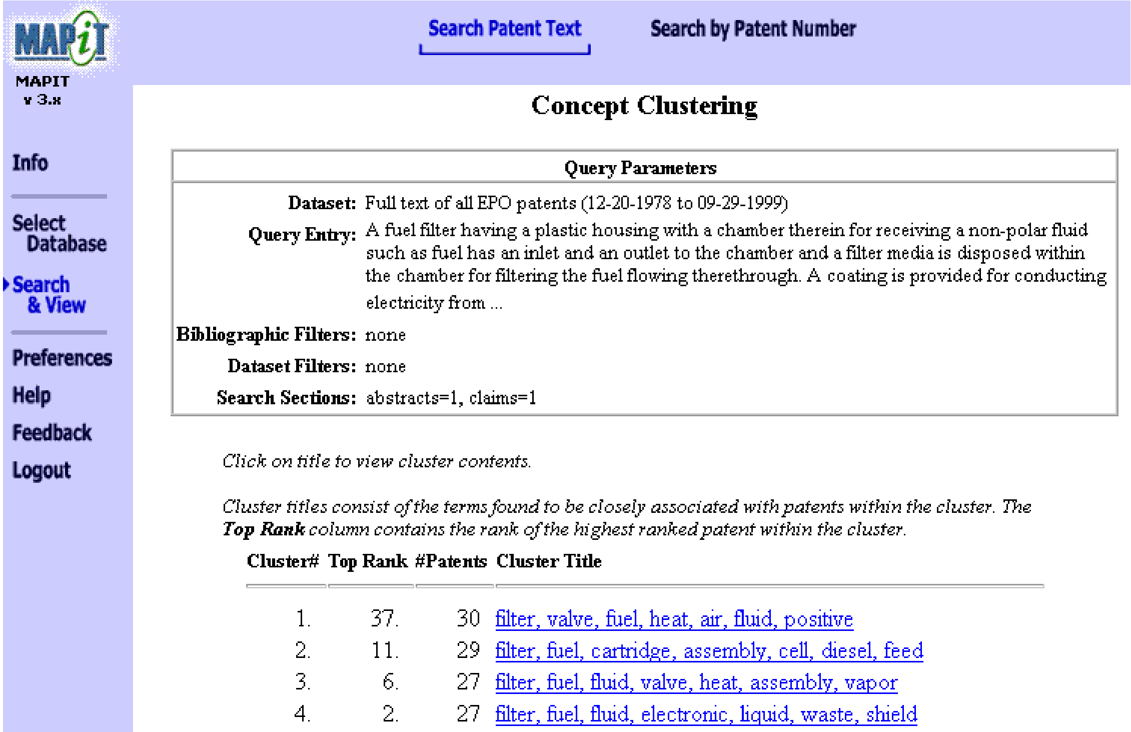
3.3.9 MapIt

MapIt is a tool produced in 1999 by Manning & Napier Information Services—a US technology incubator in New York state. It is designed to perform in-depth analysis on patent information.

Available patent information includes files from the USPTO in full text, PCT data including bibliographic information and abstracts, documents from the EPO, and bibliographic information and abstracts from the Patent Abstracts of Japan database. MapIt also provides the capability to build personal datasets from previously saved results.

MapIt provides options for results sorting and visualization, including a hyperbolic browser of patent citation information. A concept query facility allows one to search a topic or technology of interest, using natural English phrases and sentences. In Figure 13, the concept clustering result window is shown.

No information about the clustering algorithms is publicly available. However, it is based on a linguistic analysis of the abstract of the patent. The top 500 patents in a results set are analysed for clustering and a maximum of 30 defined clusters can be displayed.



Concept Clustering

Query Parameters

Dataset: Full text of all EPO patents (12-20-1978 to 09-29-1999)

Query Entry: A fuel filter having a plastic housing with a chamber therein for receiving a non-polar fluid such as fuel has an inlet and an outlet to the chamber and a filter media is disposed within the chamber for filtering the fuel flowing therethrough. A coating is provided for conducting electricity from ...

Bibliographic Filters: none

Dataset Filters: none

Search Sections: abstracts=1, claims=1

Click on title to view cluster contents.

Cluster titles consist of the terms found to be closely associated with patents within the cluster. The ***Top Rank*** *column contains the rank of the highest ranked patent within the cluster.*

Cluster#	Top Rank	#Patents	Cluster Title
1.	37.	30	filter, valve, fuel, heat, air, fluid, positive
2.	11.	29	filter, fuel, cartridge, assembly, cell, diesel, feed
3.	6.	27	filter, fuel, fluid, valve, heat, assembly, vapor
4.	2.	27	filter, fuel, fluid, electronic, liquid, waste, shield

Figure 13: MapIt concept clustering of patents, from www.mnis.net/mapitdemo

3.4 Related projects

We list below projects not directly related to patent categorization, but of interest for various reasons, as explained in each case.

3.4.1 Markush Structures

A system for processing natural language descriptions of chemical patents and extracting a formal syntactic description of chemical reactions has been described in the literature [Chowdhury92]. This research produced a voluminous academic literature about automated patent interpretation from 1981 to 1993. A full list of references is available online: www.bci.gb.com/about/bibliography.htm.

GENSAL, a formal language for the unambiguous description of generic chemical structures, has been developed and applied to patent documents. It is designed to be intelligible to a chemist or patent agent, yet sufficiently formalised to be amenable to computer analysis. The system is reported to process satisfactorily 86% of 545 chemical patents containing information about Markush structures, a highly-variable and modular set of chemical compounds [Austin01, Chowdhury92].

While not a direct categorization system, this project has shown that specialized knowledge in the chemical field may be required for a correct interpretation of patents related to chemical compounds. However, even in such a technical domain, natural language processing may also be automated by a purpose-built tool.

Some of the authors of this research now run Barnard Chemical Information Ltd, a company that offers a range of off-the-shelf software packages for analysis of large files of chemical structures (www.bci.gb.com). They offer tools for chemical clustering, on the basis of similarities between their structures.

Tools for searching patent databases for Markush structures have been developed by various organizations and companies, and are listed in [Austin01]. They allow chemical structures to be drawn on screen, translated into a search-specific form, and relevant patents to be retrieved.

The Merged Markush Service (MMS, www.inpi.fr/inpi/mms) is an extensive searchable patent information service for the pharmaceutical and chemical communities. INPI and Derwent maintain the Merged Markush Service. The MMS structure file, which is the core of the Merged Markush Service, provides a comprehensive coverage of patent chemical structures. Its main features are that all chemical areas are included, it is extremely current, and over 12 years of data are covered. As explained in [MMS], the principle of operation is that a user formulates a query, in the form of a chemical structure, which is translated into a format suitable for database searching. Patent documents relating to similar compounds are then retrieved.

Where patents are classified in the IPC by the compounds they describe, such a tool could be useful for determining the correct IPC code of a new patent application. By listing the IPC codes of patents retrieved from the MMS database, a good guess about the IPC code of the new structure could be obtained.

3.4.2 Patent Trend Discovery

Researchers at IBM built a system in 1997 to discover trends in patent databases [Lent97]. The tool was named PatentMiner, but should not be confused with www.patentminer.com, a commercial patent retrieval service that is now discontinued.

The system analysed word phrase distributions in the full-text of US granted patent documents. The user was requested to specify a topic trend, such as a sudden rise in interest in a topic over a two-year period, using a graphical interface or a specific shape definition language. Relevant patent topics were then displayed, as exemplified in [Lent97]. For example, the tool allowed current popular patent topics to be conveniently retrieved.

3.4.3 Desire II

Desire II was a European research programme run until 2000 whose aims included the automatic classification of web pages [Koch00]. The authors focused on categorizing engineering information in a hierarchical taxonomy of 800 categories, known as EELS (Engineering E-Library, Sweden). An example of this taxonomy is given below.

EELS taxonomy sample
400 Civil Engineering
500 Mining Engineering
600 Mechanical Engineering
700 Electrical Engineering
800 Chemical Engineering
820 Agricultural Engineering and Food Technology
821 Agricultural Equipment and Methods
821.1 Agricultural Machinery and Equipment
821.2 Agricultural Chemicals
821.3 Agricultural Methods
821.4 Agricultural Product
821.5 Agricultural Wastes
821.6 Farm Buildings and Other Structures
900 Engineering, General

Table 4: A subset of the Swedish Engineering E-Library taxonomy

he simple classification algorithm consisted of matching pre-built keywords describing the categories with the full text of the web pages, and weighting the matches according to their frequency and context (i.e. whether the keywords appeared in the title, the header, or the body of the page). There is thus no machine learning or training sets involved. By selecting the top matching categories and thresholding adequately, an estimated 57%-66% of 1,000 pages were classified correctly.

3.4.4 SIC classification

The Standard Industrial Classification (SIC) Code was developed by the US Federal Government to aid in gathering economic statistics and to help compare information from different government agencies. SIC codes define industries according to their economic structure and classify businesses by their primary activities.

In Table 5, a subset of the SIC classification codes is listed. The SIC classification is hierarchical, with 11 divisions divided into 83 major industry groups that are further subdivided into 416 industry groups and 1005 industries.

SIC taxonomy sample
0000-0999 Agriculture
0111-0191 Agricultural Production—Crops
0211-0291 Agricultural Production—Livestock
0711-0783 Agricultural Services
0811-0851 Forestry
0912-0971 Fishing, Hunting, and Trapping
0912 Finfish
0913 Shellfish
0919 Miscellaneous marine products
0921 Fish hatcheries and preserves
0971 Hunting, trapping, game propagation
1000-1499 Mining
1500-1999 Construction

Table 5 : Example of SIC classification codes

The categorization of company descriptions into the correct SIC category has been tested with several categorization products [Dolin99]. The authors trained a commercial categorization engine (Verity Search'97 Developer's Kit) and two academic products to recognise descriptions of activities submitted by companies to the Federal government. The training sets were extremely small, between 1 and 6 documents per category, and only a SIC subset of 206 categories was studied. The authors generated a ranked list of proposed categories for each company with the software tools. The procedure involved determining a single representation document for each category and matching it with the test documents. Technical difficulties with the software were initially found when classifying large documents. Taken as black boxes, the two academic packages were more successful than Verity, the best yielding the correct SIC category in 45% of cases. In a further 20% of cases, the correct SIC code was ranked second in the list of categories automatically generated.

A natural-language search tool built by the OASIS research group for finding SIC classification codes down to industry level is available online at metaphor.sims.berkeley.edu/oasis/sic.html [Gey99]. It is based on the same controlled vocabulary system as the patent search system detailed in 3.2.2.

3.4.5 KerMIT

KerMIT (Kernel Methods for Image and Text classification, clustering, ranking and filtering) is an ongoing European project concerned with the development of algorithms and software for the classification, clustering, ranking, and filtering of digital documents (www.eurokermit.org). In particular, the focus of the project is on investigating the use of kernel methods (support vector machines and others) in processing multilingual (French, English, Czech, Hungarian, German) and multimedia documents (containing text and images). In particular, new algorithms have been developed for category ranking [Crammer02], but full applications have still to be presented.

The KerMIT research project is a joint project between a consortium of two industrial partners-Reuters Group Plc and Xerox Research Centre Europe-and four academic partners-the Hebrew University of Jerusalem, Royal Holloway and Bedford New College at the University of London, *Università degli Studi di Milano*, and *Università di Genova*.

The Kermit project started in March 2001 and runs for 3 years.

3.4.6 WISPER

The European WISPER project-Worldwide Intelligent Semantic Patent Extraction & Retrieval (www.bmtproject.net/wisper/)-aims to demonstrate that flexible tools to permit automated patent mark-up can be developed to allow directed patent search and retrieval.

WISPER will be a new multilingual portal for intelligent access to massive patent databases. The semantic content of patents will be encoded and visualised by using advanced text mining, intelligent image analysis and user profiling to allow ontology development and thereby encourage small and medium-sized enterprise participation in patenting. Users will be able to search both text and images for the first time. Results will be shown in detailed graphical interfaces.

The WISPER project is partially funded by the European Commission's Information Society Technologies Programme. Partners in the project include British Maritime Technology (UK), Hillgate Patent Services (UK), TEMIS (F), Synthema (I), TXT e-solutions (I), *Consorzio Milano Ricerche* (I) and Haptica (Ire).

The WISPER project started in June 2002 and runs for 3 years.

4 Categorization software

4.1 Commercial products

In the following paragraphs, we present some information about general-purpose commercial categorization engines. We do not include solutions that only provide document clustering. A full list of products is found here: www.kdnuggets.com/software/classification.html. Most enterprise-scale document management solutions include a module for text-categorization, but details about the accuracies and algorithms employed are business-sensitive information that is often not published. However, automated document classification is good enough for commercial entities to be able to sell products that include it.

News about commercial electronic document management software can be found online at www.gilbane.com. A free electronic newsletter is in particular available.

4.1.1 Applied Semantics

Applied Semantics is dedicated to developing software in the field of knowledge management and produces the Auto-Categorizer product (www.appliedsemantics.com/as_solutions_autocat.shtml). The company's description of its application in the categorization area is as follows:

> "Auto-Categorizer works in real-time and can organize documents into an industry standard or custom taxonomy. Unlike technologies that require the time-consuming process of creating a training set of documents or designing a set of rules to make the categorization system function, Applied Semantics' Auto-Categorizer utilizes a simple mapping process. Using our Taxonomy Administrator, a user can quickly define category names and associate categories with the concepts in the Applied Semantics ontology that best describe the category. The mapping process is much simpler and more direct than developing extensive training sets. Instead of finding dozens or even hundreds of documents related to a category, you find the two or three concepts that represent it. Once set up, the system remains easy and inexpensive to modify and maintain."

Its technology is explained in a white paper [AppliedSemantics], claims to make use of ontologies, and is summarized below:

> "Applied Semantics' Conceptual Information Retrieval and Communication Architecture technology is composed of two principal elements, the Applied Semantics Ontology and the linguistic processing engine.
>
> The Applied Semantics Ontology consists of meanings, or concepts, and relationships between those meanings. An ontology would recognize the multiplicity of relations that a word has with other words: Java, for example, is an alternate name for coffee but is also the name for an Indonesian island and a computer language.
>
> […] Prior to understanding the overall meaning of any content, the linguistic processing engine must disambiguate each word or phrase in the document. The linguistic processing engine performs the following steps:
>
> - The tokenizer segments raw content into its individual tokens and recognizes and marks sentences.
> - The Part of Speech Tagger analyses the tokens in a sentence and assigns a syntactic category tag to each token.

- The Named Entity Recognition and Regular Pattern Identification system identifies any series of tokens that should potentially be treated as a unit, suggesting what type of unit it may be.
- The Term Segmenter maps single tokens or sequences of tokens to the terms represented in the ontological database. Competing terms – terms that overlap on one or more tokens – are each given a probability with respect to their competitors."

This approach seems to be rather simplistic and no test results are available to assess its actual efficiency. It appears to rely not on a real ontology, but rather on a classified list of concepts with few descriptors and relationships. Although it might be useful to test Applied Semantics' Auto-Categorizer in the IPC context, we do not believe that such a tool would perform better than tools previously tested at the EPO.

4.1.2 Arisem

Arisem (www.arisem.com/fr) is a French knowledge management company whose chief technology officer, Alain Garnier, comes from Lexiquest. Arisem's technology is described in [Arisem]. It is based on the combination of a classification tree and an information extractor. The classification tree represents a kind of primitive ontology. Here is how Arisem summarizes its tools:

> "The Arisem systems automate the classification of large amounts of information according to the business priorities of your company.
>
> Unlike search engines, end user to have a good grasp of how to formulate search queries, Arisem offers access to information via "classification trees" or a series of categories that reflect the vision of your enterprise:
>
> The system allows the simple and manual creation of categories and sub-categories upon which semantic filters are implemented. The information gathered will then be able to be organized and classified automatically, logically and correctly. […]
>
> The classification can offer statistics at any given moment, thus enabling the user to follow the activity of categorization. According to the mode of usage, the users will have personalization options that keep track of earlier consultations.
>
> The date, relevance, source, type of information and the relevant extract relating to the theme of the classification node are given prior to the decision to consult a text."

The tests described in their white paper [Arisem] were performed on the Reuters corpus. The number of categories used is unclear, but may be 775. After extensive manual tuning, recall results are not very impressive (72%) and precision results are poor (19%). It could be useful to perform a test with Arisem's technology on the IPC corpus to check how these tools behave in the specific context of patents.

4.1.3 Autonomy

Autonomy (www.autonomy.com) is a general-purpose concept search engine that does not rely on words but on so-called concepts for pattern matching. Autonomy can index a variety of external content sources: shared file systems, web content, or databases, for example. A specific Autonomy module, the Autonomy Classification Server, allows documents to be classified in pre-defined categories, based on the similarity of the new document with previous documents already manually classified. The algorithms employed are of Bayesian nature and are language independent, since such they make no use of linguistic analyses or semantic networks [GGautonomy].

The rich set of categorization features also include automatic clustering, spectrograph viewing of cluster information, 2D cluster maps and automatic taxonomy generation [Autonomy]. Autonomy categorizer supports the following features:

- Dynamic and consistent categorization
- Create, activate, modify, delete categories
- Training categories by example
- Multiple category matching
- Flexible categorization actions (list categorized documents to a specified file, move categorized documents to a directory, make a copy of the categorized documents, email categorized documents to an email address, import the categorized document, index the categorized document)

The Gartner Group suggests that setting up and training a full-blown Autonomy portal—including categorization and other knowledge management functions—can take months for effective use by over 1,000 users [GGautonomy].

4.1.4 Documentum

Documentum, an enterprise-level document management system, provides an add-on module called Content Intelligence Services (CIS), which provides the following elements:

- A server, which is the core engine for CIS and which performs information extraction and conceptual categorization to enable auto-tagging and auto-classification of content.
- A Verity full-text indexer that is used by CIS when processing documents and when displaying documents via the resource viewer feature.
- An XML-based taxonomy import utility that enables importing of pre-built or existing taxonomies and automatically sets up the Documentum repository folder structure to reflect the imported taxonomy nodes.
- An administration tool suite that includes a web based control panel, a web based administrator, a domain map editor and scripts to generate reports and perform administrative functions.

The Documentum classifier incorporates a semantic analysis engine. It supports multiclassification tasks by making use of thresholds. Categories are assigned on the basis of a weighted list of concepts that describes a category (as in the Rocchio approach) [Documentum].

CIS runs on Windows NT and 2000 and is certified with Oracle and MS SQL Server databases. The list price for CIS is $50k per CPU, on top of the standard Documentum licence.

4.1.5 IBM

IBM Intelligent Text Miner (www-3.ibm.com/software/data/iminer/fortext/) is a text analysis tool that integrates with IBM's DB2 universal database and provides:

- Language identification to discover the language of a document
- Clustering to group related documents by contents
- Categorization to assign documents to a set of pre-defined categories

- Summarization of documents
- Feature extraction to identify key elements of free-text
- A text search engine to search for textual information and to uncover related concepts with Java-based samples for GUI application development

The categorization system in this product is a centroid approach, similar to the Rocchio algorithm, in which the features are vocabulary items. The categories are represented by vectors consisting of the categories' most salient features (one vector per category). In the centroid approach, the comparison is essentially a vector-space comparison between a document feature vector and the category vectors, as in the Rocchio algorithm [Mack01].

A full-product, limited-time trial version is available at no charge. A one-processor licence costs $30k.

Researchers at IBM used this technology internally, to monitor and classify US patents, at least until 1998 [Hehenberger98]. These activities have subsequently been spun off to Synthema (www.synthema.it/english), an Italian consultancy and are incorporated in the Temis suit of products (www.temis-group.com). In [Hehenberger98], IBM researchers show a cluster analysis of all Korean patents issued in 1991.

IBM also sells the IBM Text Analyzer Business Component (www-3.ibm.com/software/webservers/components/textanalyzer.html), which integrates with the IBM WebSphere web application server. The IBM web site claims that:

- Text Analyzer is the most accurate categorization engine in the market.
- It is simple to use: It has a simple set of APIs and is easy to put into production because document categories can be set up easily. Once started, categories can be added easily without requiring expert assistance.
- Text Analyzer is unique in the number of languages it can process, as it can categorize double-byte character languages, such as Chinese, Japanese, and Arabic.
- It can return multiple categories with different confidence levels that enables a richer set of routing choices.

The technology employed relies on decision rules, and does therefore not resemble statistical engines such as those based on k-NN, Rocchio, or support vector machine algorithms. Instead, a set of categorization decision rules are produced automatically during training [Johnson02]. These can be tweaked or updated later by users.

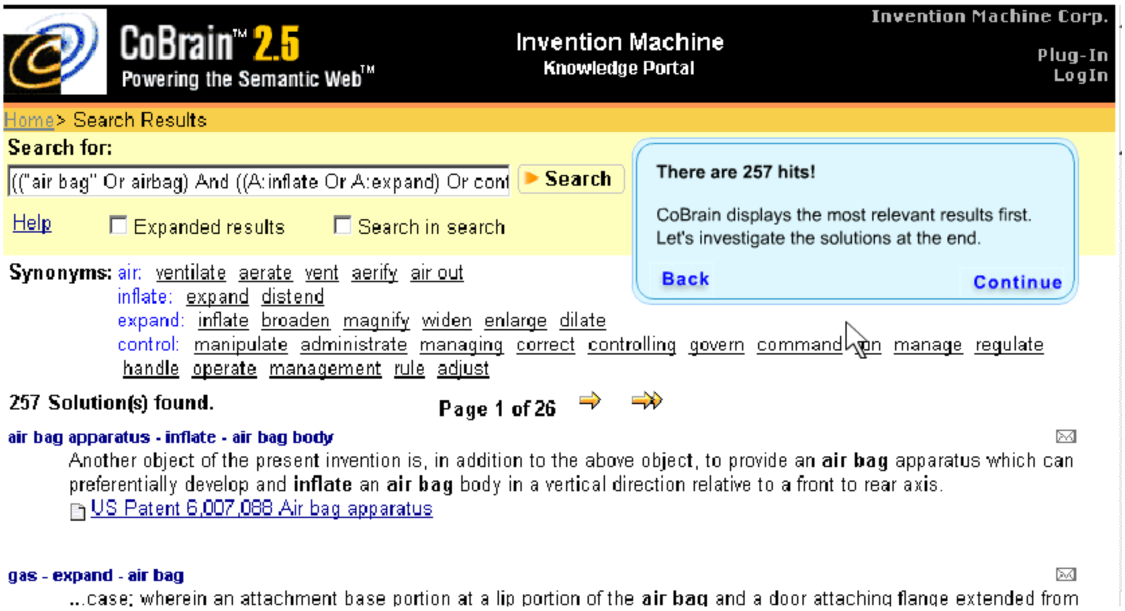
This component achieved more than 87% precision and 81% recall against the industry-standard Reuters 21578 benchmark when applied to 93 categories [IBMWebsphere, Johnson02].

4.1.6 Invention Machine CoBrain

The linguistic experts at Invention Machine have developed a technology that understands the meaning of individual sentences, defined as the ability to recognize the main structural sentence elements—namely the subject, action, and object of each sentence. The computer can then identify and extract meaningful concepts, as well as determine the semantic relationships between these concepts.

Invention Machine's technology performs semantic analysis of unstructured documents, which includes syntactic and semantic parsing of sentences. Invention Machine's linguistic knowledge base, an extensive compendium of lexical and grammatical resources, as well as advanced pattern-recognition rules, supports these algorithms.

A demo is available online at www.invention-machine.co.uk, where a limited set of content sources can be searched. In Figure 14, a set of results related to «air bags» has been retrieved from a US patent database. A set of terms lexically similar to the search string is displayed from a topic map, and several relevant US patents have been found.



(("air bag" Or airbag) And ((A:inflate Or A:expand) Or cont

There are 257 hits!

CoBrain displays the most relevant results first. Let's investigate the solutions at the end.

Synonyms: air: ventilate aerate vent aerify air out
inflate: expand distend
expand: inflate broaden magnify widen enlarge dilate
control: manipulate administrate managing correct controlling govern command un manage regulate handle operate management rule adjust

257 Solution(s) found. Page 1 of 26

air bag apparatus - inflate - air bag body

Another object of the present invention is, in addition to the above object, to provide an **air bag** apparatus which can preferentially develop and **inflate** an **air bag** body in a vertical direction relative to a front to rear axis.

US Patent 6,007,088 Air bag apparatus

gas - expand - air bag

...case; wherein an attachment base portion at a lip portion of the **air bag** and a door attaching flange extended from

Figure 14: CoBrain demonstration web page, from www.invention-machine.co.uk

This technology does not appear to have been incorporated into a stand-alone categorization product.

4.1.7 Inxight

Developed originally by Xerox Research, Inxight Categorizer (www.inxight.com/products/categorizer) is a robust enterprise application that classifies documents for fast, accurate delivery. The categorizer is said to be highly scalable and to manage thousands of categories and millions of documents. This categorizer was tested and adopted by the EPO as a tool for pre-classifying patents.

Its technology is based on [Inxight]:

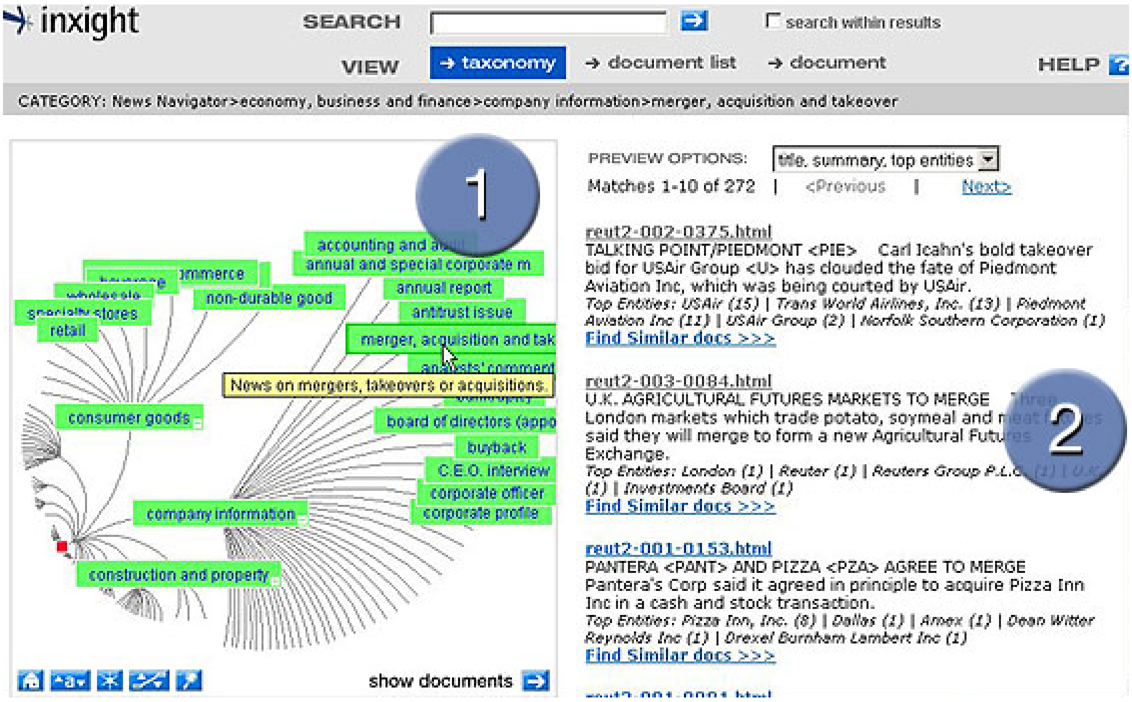
- **Natural Language Processing:** Inxight's patented linguistic technology is said to understand the context of documents in multiple languages. Facilities for stemming and tokenising documents in various languages are provided, including German and Finnish.

- **Statistical algorithms:** Self-tuned categorization using the k-NN method that learns by example as more documents and categories are added.

Inxight Categorizer includes a Java and C API with XML output. Windows, Linux, and Solaris are supported.

Inxight also provides a tool to produce a hyperbolic view of web links or document categories, as illustrated in Figure 15. There is a demo available online at www.inxight.com/map/. A free demonstration application can also be downloaded to produce custom maps limited to 300 links. The authors of this report have found it easy to use this tool and add links by cutting and pasting from Microsoft Internet Explorer.

Inxight also sells the just-released SmartDiscovery portal solution, shown in Figure 15, an integrated product for document management, summarization, categorization, concept searches, and taxonomy management.



CATEGORY: News Navigator>economy, business and finance>company information>merger, acquisition and takeover

PREVIEW OPTIONS: title, summary, top entities

Matches 1-10 of 272 | <Previous | Next>

reut2-002-0375.html
TALKING POINT/PIEDMONT <PIE> Carl Icahn's bold takeover bid for USAir Group <U> has clouded the fate of Piedmont Aviation Inc, which was being courted by USAir.
Top Entities: USAir (15) | Trans World Airlines, Inc. (13) | Piedmont Aviation Inc (11) | USAir Group (2) | Norfolk Southern Corporation (1)
Find Similar docs >>>

reut2-003-0084.html
U.K. AGRICULTURAL FUTURES MARKETS TO MERGE Three London markets which trade potato, soymeal and meat futures said they will merge to form a new Agricultural Futures Exchange.
Top Entities: London (1) | Reuter (1) | Reuters Group P.L.C. (1) | U.K. (1) | Investments Board (1)
Find Similar docs >>>

reut2-001-0153.html
PANTERA <PANT> AND PIZZA <PZA> AGREE TO MERGE Pantera's Corp said it agreed in principle to acquire Pizza Inn Inc in a cash and stock transaction.
Top Entities: Pizza Inn, Inc. (8) | Dallas (1) | Amex (1) | Dean Witter Reynolds Inc (1) | Drexel Burnham Lambert Inc (1)
Find Similar docs >>>

reut2-001-0001.html

Figure 15: Inxight SmartDiscovery screenshot. Part (1) shows a hyperbolic view of a taxonomy, while part (2) shows the corresponding documents. From www.inxight.com.

4.1.8 Lexiquest

SPSS INC, Chicago, a software company active in data mining and automated decision-making now markets Lexiquest.

Lexiquest Categorize (www.spss.com/spssbi/lexiquest/categorize.htm) is a linguistic-based general categorization tool that features [Lexiquest]:

- **Linguistic features:** Using natural-language processing technology and semantic networks, LexiQuest Categorize is able to recognize and extract compound words, phrases and idioms that would typically be treated as individual words by other products. This has a dramatic effect on the overall accuracy of the system. Dictionaries are available in English, French, and German.
- **Term Extractors:** LexiQuest Categorize employs the same technology as the text-mining tool from the same company, LexiQuest Mine, and as such has the ability to extract specific types or categories of information from text. This enables the categorization to be independent of the domain being processed and accurate regardless of the industry.
- **Volume/Speed:** LexiQuest Categorize is suited to cataloguing extremely large volumes of data very quickly (250,000 pages of text per hour).

In its datasheet, Lexiquest claims to provide over 70% accuracy in its first two category proposals, when employing a 600-category taxonomy. No details about the specifics of the test set are made available.

4.1.9 Oracle

Oracle Text uses standard SQL to index, search, and analyse text and documents stored in an Oracle database, files and on the Web. Oracle Text can analyse document themes and summaries; search text using a variety of strategies, including full-text Boolean, exact phrase, proximity, section searching, misspellings, stemming, wildcard, thesaurus, word equivalence, scoring, and thematic; search HTML and XML sections and tag values; render search results in various formats including unformatted text, HTML with automatic keyword highlighting, and original document format; analyse and index most document formats with over 150 document filters; supports 39 languages; bulk load documents in Oracle8i/9i with SQL*Loader [Oracle].

The system comes with a pre-built taxonomy of 425,000 groups arranged in 2,000 main classes, but custom taxonomies can also be employed. The algorithm employed creates a theme vector to represent each document, by stemming and selecting relevant keywords. Uniquely, it also weights the terms according to their position in a sentence, with leading terms receiving higher weights. Categorization is performed by comparing the document theme vector with category theme vectors, as in the Rocchio algorithm. Documents can be classified in more than one category [Alpha01].

4.1.10 Recommind

Recommind (www.recommind.com) markets the Mindserver platform, which analyses and indexes information for use in a variety of applications. The MindServer Categorizaton module takes structured and unstructured information and automatically maps content into an existing information structure (taxonomy, ontology, and subject heading classification structure).

The core technology powering Recommind’s MindServer Product Suite is based on patented, proprietary machine learning techniques including the Probabilistic Latent Semantic Indexing algorithms developed by Recommind’s Chief Scientist, Professor Thomas Hofmann (Brown University, USA). This algorithm is a term-reduction technique that only selects from the corpus vocabulary words that are best at discriminating between categories. Recommind claims to support topic maps and provide word-sense disambiguation.

The MindServer platform is said to automatically identify the concepts that describe a document, regardless of the language or subject. As a result, MindServer is able to automatically understand that words can have multiple meanings and that there are multiple ways to express the same concept or query. Features include the ability to:

- Automatically or semi-automatically incorporate documents into a taxonomy- or multiple taxonomies
- Accurately assign predefined descriptors, thesaurus terms, or metadata to a document collection
- Accurately map documents into multiple categories; Automatically generate metadata, tag documents, and map into XML
- Train the system in an automated, semi-automated, or manual fashion

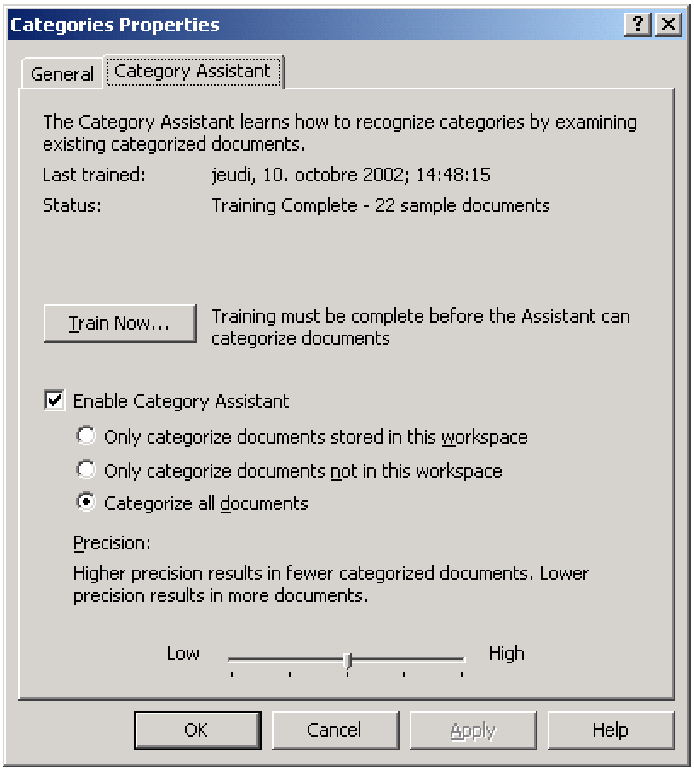
MindServer requires 1Gb RAM with a greater than 500Mhz processor as well as over 5Gb of disk space. Additional requirements are determined by the data collection used. Versions of MindServer are currently available for Windows NT, Windows 2000, Solaris 2.5 and higher, and Linux 7.1 or higher. It has an API and supports integration with Java and C++ applications.

4.1.11 SharePoint

SharePoint, a web-based enterprise document-management system (www.microsoft.com/sharepoint), supports document classification in predefined categories containing pre-indexed documents. Documents to categorize must first be up loaded into the web-based document repository. A document can then be automatically categorized in several categories.

The technology used in SharePoint's text classification scheme is a in-house Microsoft development. The principal Microsoft researchers who worked on it are Susan Dumais and John Platt [Johnston01]. The algorithms employed are based on Support-Vector-Machine classification [Hearst98], a highly performing algorithm in most comparative categorization tests.

In Figure 16, a screen shot of SharePoint categorization assistant is shown. One can see that the automated categorization is designed to provide a hassle-free user experience with very few parameters to set. A single parameter adjusts the system to request high recall or high precision. While advantageous from an end-user perspective, it allows little flexibility when testing the system. Although SharePoint has an API for custom developments, it does not provide support for tuning the automated categorization.



The Category Assistant learns how to recognize categories by examining existing categorized documents.

Last trained: jeudi, 10. octobre 2002; 14:48:15

Status: Training Complete - 22 sample documents

Training must be complete before the Assistant can categorize documents

☑ Enable Category Assistant

○ Only categorize documents stored in this workspace

○ Only categorize documents not in this workspace

● Categorize all documents

Precision:

Higher precision results in fewer categorized documents. Lower precision results in more documents.

Figure 16: SharePoint Categorization Assistant screenshot

The SharePoint interface has been translated into German, French, Spanish, Italian, and Japanese. In addition, SharePoint provides noise word files and thesaurus files for the following languages: Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Spanish, Swedish, and Thai [GGsharepoint].

4.1.12 Semio

SemioTagger (www.semio.com) is said to efficiently organize and expose unstructured text contained within an enterprise. Uniquely built for large data volumes and exceptional granularity, SemioTagger offers enterprise scalability and easy integration facilities.

SemioTagger uses patented linguistic analysis to identify the key concepts in documents and organize documents into categories, using categorization rules one can see and modify. In this respect, it appears to function using a ruled-based technique rather than a statistical method such as a Rocchio algorithm. With this approach, complete control over how documents are categorized and presented to users is obtained. SemioTagger's automated categorization claims to achieve an accuracy of 95% of manual efforts.

4.1.13 Verity

Verity Intelligent Classifier provides administrators with the tools to quickly organize enterprise information into business-specific categories. Information from all popular document formats and repositories are automatically classified using business rules that can be automatically built by Verity Intelligent Classifier without the aid of subject experts. Intelligent Classifier lets the system administrator fine tune the rules to suit specific objectives. It thus appears that the classification algorithm is based on decision rules [Verity].

Verity Intelligent Classifier is a stand-alone Windows Application, with no API. It currently supports servers running Windows NT 4.0 with service packs 3 and 4, 64 MB RAM is required, as well as 100Mbytes of hard drive space and at least a 200 Mhz Pentium-level processor.

No product cost information is available on the Verity web site. They must be contacted separately at: www.verity.com/contact/index.html

Verity also sells a more versatile developer platform, known as Verity K2 Developer, to allow commercial independent software vendors to add Verity's indexing, basic and advanced search, content organization, and social network capabilities to e-business applications, portal frameworks and software infrastructure. Further details are online at www.verity.com.

Verity is used as a publishing and search portal for the database of all Swiss patents, available online at ch.espacenet.com [VerityCH].

4.1.14 Others

Other vendors marketing categorization software include Gammasite (www.gammasite.com), Inktomi (www.inktomi.com), Liquent (www.liquent.com), Temis (www.temis-group.com), and Yellowbrix (www.yellowbrix.com, whose core technology relies on support vector machines).

4.2 Freeware packages

A list of freeware categorization tools can be found on www.kdnuggets.com/software/classification.html, although most tools listed on this site are not specific to text categorization. Most of these packages are available for research and educational activities only. In particular, the following packages collectively implement a rich set of different algorithms:

- **SVMlight**: svmlight.joachims.org. SVMlight is implementation of the support vector machine algorithm in C. Extensively developed and revised, version 5.0 is now available. It solves classification and clustering tasks and has a large number of options. Various optimisations have been performed for speed. It can handle over 100,000 training examples and many thousand categories. Source code, tutorials, manuals, and executables for Unix and Windows are available. This is a versatile tool often used in academic categorization algorithm comparisons, and is usually a top performer. The code provided implements the core part of the classification. Supporting software for indexing documents to SVMlight's input format must be developed separately.

- **Bow/Rainbow:** www-2.cs.cmu.edu/~mccallum/bow. Bow is a library of C code for text analysis, including facilities for importing text files, tokenising documents, and finding word vector weights and word probabilities. Only the source code is provided, and documentation is sparse. Rainbow is a comprehensive command-line text classification tool that implements Naïve Bayes, Rocchio, support vector machines, and k-NN algorithms.

- **BoosTexter**: www.research.att.com/~schapire/BoosTexter. BoosTexter is a text categorization programme implementing a decision rules algorithm [Schapire00]. It can handle large data sets and multiclassification, but appears to have difficulty handling large numbers of categories. Only the binary code is available, for Unix and Windows.

- **SnoW**: l2r.cs.uiuc.edu/~danr/snow.html. SnoW, or Sparse Network of Winnows, is a multi-class categorizer implementing Perceptron, Naïve Bayes, and Winnow-type algorithms. It scales linearly with the number of features. User guides and source codes in C++ are available. Various support tools for linguistic analysis are also available.

- **CBA**: www.comp.nus.edu.sg/~dm2/. The Classification Based on Associations (CBA) text categorizer implements a decision-rules algorithm for single-class classification. Executables are available freely for academic research.

- **SIMPL**: www.cse.iitb.ac.in/~soumen/main/download.html. Soumen Chakrabarti indicates on his homepage that he will soon be releasing a freeware version of his new SIMPL (Simple Iterated Multiple Projection on Lines) algorithm, which combines the accuracy of support vector machines with the simplicity of Bayesian classifiers.

Various categorization algorithms can be tested online by classifying a two-dimensional set of points. Online applets found here:

- www.cs.technion.ac.il/~rani/LocBoost/index.html: Comprehensive implementation of a variety of categorization algorithms applied to a two-category taxonomy.

- page.inf.fu-berlin.de/~tapia/svm/SvmApplet2.html: Categorization into two categories, using various support vector machine variations.

- www.csie.ntu.edu.tw/~cjlin/libsvm/: Categorization into three categories, using various support vector machine variations.

Freeware packages that support additional linguistic tasks include:

- **Porter Stemmer**: www.tartarus.org/~martin/PorterStemmer. A popular English word stemming algorithm developed by Martin Porter, with source code available in Perl, Python, Lisp, Java, C, and C#.

- **Snowball**: snowball.tartarus.org. Multilingual stemming algorithms by Martin Porter in C and Java, covering Romance, Scandinavian, Germanic languages, Russian, and Finnish.

- **IGLU-Java**: iglu-java.sourceforge.net. IGLU is a general-purpose Java class library implementing various data mining functions, such as the Porter stemming algorithm, word frequency calculations, and the creation of document indexing vectors, using the popular TFIDF normalization [Sebastiani02].

- **WordNet**: www.cogsci.princeton.edu/~wn. WordNet is a large ontology that is freely available, including for commercial work. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. Overall, WordNet includes 168,000 words grouped in 91,600 synsets.

5 Conclusions

In operational contexts, reliable automated categorization is often achieved by a combination of modifying category definitions, by providing more training data, by considering better learning algorithms, and by manual feature engineering to get to an acceptable level of effectiveness [Lewis02].

In the case of automated IPC patent classification, the taxonomy is a given complex constraint. To achieve the requested accuracy, it will therefore be necessary to consider all other possibilities for improving effectiveness carefully. This study has surveyed the current state-of-the-art in text categorization algorithms, applications, and results.

Training sets

All published literature highlights the importance of training sets in developing accurate automated categorization software. Such tools cannot produce classification knowledge automatically if they lack the basic information on which to develop their understanding, no matter how sophisticated the algorithm. In particular, large training sets that are evenly distributed across the taxonomy categories are always preferred.

Academic researchers often use Reuters, newsgroups, or medical abstract collections for training and testing new categorization techniques. This results partly from the desire to compare new work with previously-published literature, and partly from the lack of other comprehensive online resources.

Patent databases form an excellent source of training data because the classification has been performed manually with extreme care by subject experts, because the taxonomy is complex and hierarchical, and because documents in several languages has been classified in the same taxonomy. Nevertheless, the research community has not tapped this source extensively due to difficulties in accessing data sets.

Algorithms

In the field of text classification algorithms, progress is continuously being made in academic research and commercial products. In particular, support vector machine algorithms, implemented in the commercial SharePoint portal and the academic Rainbow and SVMlight freeware packages, provide excellent results and have not been tested yet on an extensive patent corpus. The last of these tools is particularly well adapted to development spikes as it often serves as a benchmark classifier.

Classifying committees also promise to bring improved accuracies to patent classification. When successful, they combine different algorithms' best abilities to deal with high or low numbers of training samples in each category.

Research into hierarchical classifiers is still at an early stage, and past implementations have relied on custom solutions specifically tailored to the problem at hand. Results are encouraging, but do not seem to have yet been implemented in commercial products.

Patent classification tests

At the request of the European Patent Office (EPO), a systematic comparative study of patent IPC categorization tasks has been performed. Such studies are precious because extrapolating accuracies from other work is difficult in view of possible differences in document structure, in document length variation, in vocabularies, in the number of categories, in the sizes of training sets, and in the nature of the classification task (whether single-label classification or multiclassification). Furthermore, the evaluation of categorization algorithms is complicated by the various measures reported, whether precision, recall, or other average and peak measures. Categorization tasks can also be biased towards providing high category precision or high recall by adequate thresholding.

The EPO tests showed that patent classification is more accurate when the full texts of patents—rather than patent abstracts and bibliographic information—are employed. In line with other studies, the Inxight k-NN categorizer provided sufficient accuracy for EPO patent pre-classification needs, scaling well with the size of the corpus and the number of categories. Around 72% precision at 100% recall was achieved with around 50 categories. Sacrificing some recall and flagging documents for manual categorization improved the precision of these results.

The application in view for the WIPO CLAIMS project differs from the EPO tasks in that a single category only was to be attributed by the EPO system. If an IPC classification assistance tool is developed at WIPO, several categories will probably be suggested, thus improving the chances of a good match and raising the effectiveness found in the EPO tests.

The difficulty of developing accurate IPC classification tools is highlighted by the results achieved by two online categorizers (CIB-LN and the OASIS research tool), who ambitiously attempt to associate natural language queries with IPC subgroups. Neither of these tools fulfils user expectations. Currently, the European e-Patent programme is attempting to improve on results with a 6-institution combined effort. This development scale illustrates the necessary complexity.

Outlook

Little academic research has been published about automated classification systems applied to taxonomies consisting of several hundred categories. This suggests that effective classification to IPC subgroup level, comprising around 69,000 categories, is beyond the current state-of-the-art. Instead, an automated system supporting subclass classification, into approximately 600 categories, seems a realistic target today, as performances are continuously improving in the field. Reports of excellent patent classification accuracies from Japan are most encouraging in this respect, although details about these tools are still lacking.

A number of promising avenues to explore rely on the development of a customized classification tool, rather than the implementation of a generic commercial product. In particular, little use appears to have been made up to now in applying the catchword indexes to IPC classification. The use of references in the IPC category definitions could probably also be exploited better when several classifications seem relevant at first sight. Such developments may need to rely on the use of linguistic techniques and ontologies to disambiguate polysemic words.

The implementation of an IPC categorization tool requires competencies in a number of fields. Mathematical issues related to document term selection as well as algorithm choice and implementation are important. Computational issues related to large database management and system response time must be considered. Linguistic aspects related to ontologies and multilingual vocabularies should be mastered. Specific knowledge about the current use of the IPC taxonomy itself is crucial. Furthermore, information about user and application interface requirements must be available. All these facets of a classification tool must be explored in detail to build an effective system.

Further surveying work

An exhaustive survey of patent categorization results would benefit from documents obtained from a number of sources not available to the authors of this document. In particular, peer-reviewed academic papers listed in INSPEC—the leading English-language bibliographic information service of technical literature in science—and published in academic journals should be surveyed. Internal documents relating to tests at the JPO and the Japanese OWAKE categorizer would be of great interest. Further details about current USPTO categorization activities are also needed.

Appendix A: Research methodology

The documents collected for this study have been in large part procured from free public sources. Heavy use has been made of www.google.com for locating primary online resources.

Research papers and conference proceedings have been found and retrieved using citeseer.nj.nec.com/cs. Some published peer-reviewed research papers have also be located using www.ingenta.com. The homepages of academic text-categorization researchers sometimes provide copies of research papers that are otherwise only accessible from journals requiring subscriptions. Papers from the World Patent Information journal can only be downloaded with a subscription.

Software datasheets and white papers have been procured from vendor web sites. Papers from Gartner Research have also been consulted at www.gartner.com, where a subscription is required for downloading documents.

In Table 6, we present a selected list of interesting web sites relating to document and patent categorization.

URL	Description
www.wipo.int	WIPO's comprehensive portal.
www.european-patent-office.org	The homepage of the European Patent Office.
ep.espacenet.com	Europe's network of patent databases, where full-text patents can be searched for and downloaded.
www.ipmenu.com	A global guide to intellectual property resources on the Internet.
www.piug.org	The International Society for Patent Information.
www.elsevier.com/locate/issn/01722190	Web site of the World Patent Information journal. The full text of this publication is online, but a subscription is required to download the full text.
www.european-patent-office.org/epidos/conf/patlibal.htm	PATLIB is a network of patent information centres throughout Europe.
liinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html	Fabrizio Sebastiani's Bibliography on Automated Text Categorization, a comprehensive guide to published academic literature.
www.cs.helsinki.fi/group/doremi/categorization/bibliography.html	Automated Text Processing Related Short Bibliography, mainly focused on academic research.
dewey.yonsei.ac.kr/memexlee/links/categorization.htm	Another text categorization bibliography, maintained by Jae Yun Lee.
www.kdnuggets.com/software/classification.html	List of software packages for document categorization, both commercial and freeware sources are listed, sorted by algorithms.
ftp://ftp.sas.com/pub/neural/FAQ.html	Frequently asked questions of the comp.ai.neural-nets newsgroup, which contains some relevant information on machine learning with neural networks for classification tasks.

URL	Description
www.ai.mit.edu/projects/jmlr/	The online Journal of Machine Learning Research, providing free access to research papers, some of which focus of document categorization.
www.mayallj.freeserve.co.uk/	A commercial web site listing a variety of Intellectual Property resources on the web, including patent news and software.

Table 6: Selected list of web site relating to document and patent categorization

In Table 7, we list important actors external to WIPO who are relevant to this study. In particular, we consider major academic researchers in document categorization and staff of international organizations involved in patent categorization. Papers and research by these authors have been particularly useful in this study.

Name	Email	Role
Michael Buckland	buckland@sims.berkeley.edu	Professor in charge of former OASIS team, who developed a simple IPC categorizer.
Soumen Chakrabarti	soumen@cse.iitb.ac.in	Assistant Professor at the Indian Institute of Technology Bombay, previously with IBM, developed a hierarchical patent classifier.
Susan Dumais	sdumais@microsoft.com	Major contributor to the categorisation software in Microsoft SharePoint Portal Server
Thorsten Joachims	tj@cs.cornell.edu	Academic author and editor of a large number of text categorization articles.
Kees Koster	kees@cs.kun.nl	Author of several papers on automated classification of European patents.
Marc Krier	mkrier@epo.org	Author of the major EPO study in patent categorization.
Leah Larkey	larkey@cs.umass.edu	Author of several papers on automated classification of US patents.
Dave Lewis	ddlewis2@worldnet.att.net	Independent text categorization consultant who runs the DDLBETA mailing list for discussion of text categorization.
Fabrizio Sebastiani	fabrizio@iei.pi.cnr.it	Research scientist in categorization techniques, author of the definitive review of machine learning techniques for text classification [Sebastiani02].
Yiming Yang	yiming@cs.cmu.edu	Author of several articles evaluating and comparing categorization algorithms.
Francesco Zaccà	fzacca@epo.org	Author of the major EPO study in patent categorization.

Table 7: Who’s who in patent and document categorization

Appendix B: Bibliography

[Alpha01] S. Alpha, P. Dixon, C. Liao, C. Yang, Oracle at TREC 10: Filtering and Question-Answering, proceedings of 10th Text Retrieval Conference (TREC 10) 2001, 423-433, trec.nist.gov/pubs/trec10/papers/orcltrec10.pdf.

[Adams00] S. Adams, Using the International patent Classification in an online environment, World Patent Information 22, 2000, 291-300, www.sciencedirect.com/science/journal/01722190.

[Applied Semantics] Ontology Usage and Applications, Applied Semantics White Paper, 2001, www.appliedsemantics.com/as_support_white_paper.shtml.

[Arcanum] Report on IPC research by Arcanum Development, Hungary, www.arcanum.com

[Arisem] Reuters Corpus: A Benchmark for Automatic Classification Systems, Arisem Group White Paper, 2001, www.arisem.com/en/downloads/index.html.

[Austin01] R. Austin, The Complete Markush Structure Search: Mission Impossible?, proceedings of PIUG North East Workshop, 2001, www.stn-international.de/training_center/chemistry/piug1.pdf.

[Autonomy] Autonomy Classification Server, Technical Brief, Autonomy, 2001, www.autonomy.com/Extranet/Technical/Products/ACI%20Servers/TB%20Autonomy%20Classification%20Server.pdf.

[Banik01] E. Banik, A. Bracy, Text Classification Algorithms & Morphological Analysis on Hungarian Newswire Articles, www.ling.upenn.edu/~ebanik/research.html

[Bennett02] P. B. Bennett, S. T. Dumais, E. Horvitz, Probabilistic Combination of Text Classifiers Using Reliability Indicators: Models and Results, proceedings of the SIGIR'02 conference, 2002, ftp://ftp.research.microsoft.com/pub/ejh/class_combine.pdf.

[Biskri02] I. Biskri, S. Delisle, Text Classification and Multilingualism: Getting at Words via N-grams of Characters, proceedings of CSI 2002, 2002, citeseer.nj.nec.com/510485.html.

[Boyack00] K. W. Boyack, B. N. Wylie, G. S. Davidson, D. K. Johnson, Analysis of Patent Databases Using VxInsight, proceedings of 9th Int. Conf. On Information and Knowledge Management, 2000, www.cs.sandia.gov/projects/VxInsight/pubs/npivm00.pdf.

[Callan92] J. P. Callan, W. B. Croft, S. M. Harding, The INQUERY Retrieval System, proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications, 1992, 78-83, citeseer.nj.nec.com/26307.html.

[Calvert01] J. Calvert, M. Makarov, The reform of the IPC, World Patent Information 23, 2001, 133-136, www.sciencedirect.com/science/journal/01722190.

[Chakrabarti97] S. Chakrabarti, B. Dom, R. Agrawal, P. Raghavan, Using taxonomy, discriminants, and signatures for navigating in text databases, proceedings of 23rd VLDB conference, 1997, citeseer.nj.nec.com/chakrabarti97using.html.

[Chakrabarti98] S. Chakrabarti, B. Dom, R. Agrawal, P. Raghavan, Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies, VLDB journal 7, 1998,163-178, citeseer.nj.nec.com/chakrabarti98scalable.html.

[Chakrabarti98b] S. Chakrabarti, B. Dom, P. Indyk, Enhanced hypertext categorization using hyperlinks, Proceedings of SIGMOD98, ACM International Conference on Management of Data, ACM Press, New York, 1998, 307-318, citeseer.nj.nec.com/chakrabarti98enhanced.html.

[Chowdhury92] G. G. Chowdhury, M. F. Lynch, Automatic Interpretation of the Texts of Chemical Patent Abstracts I & II, J. Chem. Inf. Comput. Sci 32, 1992, 463-473.

[Crammer02] K. Crammer, Y. Singer, A New Family of Online Algorithms for Category Ranking, proceedings of SIGIR'02 conference, 2002, www.cs.huji.ac.il/~singer/papers/mcml.ps.gz.

[Dalessio00] S. D'Alessio, K. Merray, R. Schiaffino, A. Kerschenbaum, The Effect of Using Hierarchical Classifiers in Text Categorization, Proceeding of RIAO-00, 6th International Conference on *Recherche d'Information Assistée par Ordinateur*, Paris, 2000, 302-313, citeseer.nj.nec.com/410559.html.

[Ding01] Y. Ding, D. Fensel, Ontology Library Systems: The key to successful Ontology Re-use, proceedings of International Semantic Web Working Symposium 2001, paper 58, www.semanticweb.org/SWWS/program/full/paper58.pdf.

[Dolin99] R. Dolin, J. Pierre, M. Butler, R. Avedon, Practical Evaluation of IR within

Automated Classification Systems, proceedings of 8th international conference on Information and Knowledge Management (CIKM), 1999, citeseer.nj.nec.com/241926.html.

[Documentum] Content Intelligence Services, Structuring Unstructured Content, Documentum Technical White paper, 2002, www.documentum.com/products/collateral/platform/wp_tech_cis.pdf.

[Dumais00] S. Dumais, H. Chen, Hierarchical Classification of Web Content, Proceedings of SIGIR'00, 23rd ACM International Conference on Research and Development in Information Retrieval, 2000, 256-263, citeseer.nj.nec.com/dumais00hierarchical.html.

[EKAW00] Knowledge Engineering and Knowledge Management: Methods, Models, and Tools, edited by Dieng and Corby, Springer-Verlag, Berlin-Heidelberg, 2000.

[EKAW02] Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, edited by Gómez-Pérez and Benjamins, Springer-Verlag, Berlin-Heidelberg, 2002.

[Eshler01] J. Eshler, J. Messina, Z. Rina Paz, J. Ricketts, Improving Patent Organization with an Automated Classification Assessment Tool, 2001 Systems Engineering Capstone Conference Proceedings, 97-102, www.sys.virginia.edu/capstone/past/cap2001/2001-24.doc.

[Esteban98] A. D. Esteban, M. de B. Rodríguez, L. A. U. López, M. G. Vega, Integrating Linguistic Resources in an Uniform Way for Text Classification Tasks, proceedings of 1st International Conference on Language Resources and Evaluation, Granada, 1998, www.esi.uem.es/laboratorios/sinai/postscripts/lrec98.ps.

[Falasco02] L. Falasco, Bases of the United States Patent Classification, World Patent Information 24, 2002, 31-33, www.sciencedirect.com/science/journal/01722190.

[Fellbaum98] Christiane Fellbaum (Editor), WordNet: An Electronic Lexical Database and Some of its Applications, MIT Press, Cambridge, 1998.

[Gaustad01] T. Gaustad, G. Bouma, Accurate Stemming of Dutch for Text Classification, Computational Linguistics in the Netherlands 2001, odur.let.rug.nl/~tanja/publis/clin01.pdf.

[Gey99] F. Gey, Y. Kim, A. Chen, B. Lam, J. Purat, R. Larson, Advanced search technologies for Unfamiliar Metadata, proceedings of Metadata'99 Third IEEE Metadata conference, 1999, citeseer.nj.nec.com/gey99advanced.html.

[GGautonomy] S. Hayward, A. Linden, Autonomy: Delivering Infrastructure for Information Access, Gartner Group Research Note C-11-0611, June 2000, www.gartner.com.

[Gghypecycle] A. Linden, Innovative Approaches for Improving Information Supply, Gartner Group Research Note M-14-3517, September 2001, www.gartner.com.

[GGKM] B. Rosser, Knowledge Mapping — Automated or Manual?, Gartner Group Research Note DF-07-8831, May 1999, www.gartner.com.

[Ggontologies] J. Jacobs, A. Linden, Semantic Web Technologies Take Middleware to Next Level, Gartener Group Technology Report T-17-5338, 2002, www.gartner.com.

[GGsharepoint] K. Shegda, Microsoft SharePoint Portal Server, Gartner Group Product Report DPRO-99297, July 2001, www.gartner.com.

[Ggtaxonomy] D. Logan, Understanding and Using Taxonomies, Gartner Group Research Note TU-13-5274, May 2001, www.gartner.com.

[Gruber93] T. R. Gruber, A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition 5, 1993, 199-220, gicl.mcs.drexel.edu/people/regli/Classes/KBA/Readings/KSL-92-71.pdf.

[Hearst98] M. A. Hearst (Ed.), Trends & controversies: Support vector machines. IEEE Intelligent Systems 13, 1998, 18-28, www.computer.org/intelligent/ex1998/pdf/x4018.pdf.

[Hehenberger98] M. Hehenberger, P. Coupet, Text Mining applied to Patent Analysis, proceedings of Annual Meeting of American Intellectual Property Law Association 1998, www-1.ibm.com/support/docview.wss?uid=swg27002175

[Hull01] D. Hull, S. Aït-Mokhtar, M. Chuat, A. Eisele, E. Gaussier, G. Grefenstette, P. Isabelle, C. Samuelsson, F. Segond, Language technologies and patent search and classification, World Patent Information 23, 2001, 265-268, www.sciencedirect.com/science/journal/01722190.

[Huffman95] S. Huffman, Acquaintance: Language-independent Document Categorization by N-grams, proceedings of TREC-4 conference, 1995, citeseer.nj.nec.com/huffman95acquaintance.html.

[Hyoudo98] Y. Hyoudo, K. Niimi, T. Ikeda, Comparison between Proximity Operation and dependency Operation in Japanese Full-text Retrieval, proceedings of SIGIR'98 conference, Melbourne, 1998, ikd.info.gifu-u.ac.jp/paper/1998/sigir98.ps.gz.

[IBMWebsphere] IBM WebSphere Business Components V1.2 datasheet, IBM, 2001, www-3.ibm.com/software/webservers/components/pdf/wsbcv12.pdf.

[Inoue00] N. Inoue, K. Matsumoto, K. Hoashi, K. Hashimoto, Patent Retrieval System Using Document Filtering Techniques, proceedings of ACM SIGIR 2000 Workshop on Patent Retrieval, 2000, citeseer.nj.nec.com/454960.html.

[InvMachine] Precision in Knowledge Retrieval White Paper, Invention Machine, 2002, www.invention-machine.com/prodserv/whitepaper.cfm.

[Inxight] Inxight Categorizer datasheet, Inxight, 2002, www.inxight.com/pdfs/datasheets/categorizer_ds.pdf.

[InxightWP] Inxight Categorizer Versus Naïve Bayes Classification, Inxight White Paper, Inxight, 2002, www.inxight.com/pdfs/white_papers/categorizer_knn_vs_bayes.pdf.

[IPC7] International Patent Classification, Seventh edition, Volume 10, WIPO, 1999, www.wipo.int/classifications/fulltext/new_ipc/.

[Ittner95] D. J. Ittner, D. D. Lewis, D. D. Ahn, Text categorization of low quality images, proceedings of 4th annual symposium on document analysis and information retrieval (SDAIR95), 1995, 301-315, www.cs.rochester.edu/u/davidahn/papers/sdair.ps.gz.

[Johnson02] D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz, A decision-tree-based symbolic rule induction system for text categorization, IBM systems journal 41, No. 3, 2002, 428-437, researchweb.watson.ibm.com/journal/sj/413/johnson.pdf.

[Johnston01] S. J. Johnston, Microsoft Research: Who Benefits?, Informationweek.com news, March 12, 2001, www.informationweek.com/828/microsoft.htm.

[Kando00] N. Kando, What Shall We Evaluate?—Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in Patent Searching and Patent Attorneys, proceedings of ACM SIGIR 2000 workshop on patent retrieval, 2000, research.nii.ac.jp/ntcir/sigir2000ws/sigirprws-kando.pdf.

[Karetka02] G. Karetka, personal communication, WIPO, 2002.

[Kingston01] J. Kingston, High Performance Knowledge Bases: Four approaches to Knowledge Acquisition, Representation and Reasoning for Workaround Planning, Expert Systems with Applications 21, 2001, 181-190, www.informatics.ed.ac.uk/publications/online/0050.pdf.

[Koch00] T. Koch, A. Ardö, Automatic classification, Desire II D3.6a, Overview of results, www.lub.lu.se/desire/DESIRE36a-overview.html.

[Koller97] D. Koller, M. Sahami, Hierarchically Classifying Documents Using Very Few Words, proceedings of the Fourteenth International Conference on Machine Learning ICML-97, 1997, 170-178, citeseer.nj.nec.com/koller97hierarchically.html.

[Koster01] C. H. A. Koster, M. Seutter, J. Beney, Classifying Patent Applications with Winnow, proceedings of Benelearn 2001 conference, Antwerpen, 2001, www.cs.kun.nl/peking/benelearn.ps.gz.

[Koster02] C. Peters, C. H. A. Koster, Uncertainty-based Noise reduction and Term Selection in Text categorization, proceedings of 24th European Colloquium on IR Research (ECIR 2002), 2002, www.cs.kun.nl/~kees/home/papers/trec9.ps.gz.

[Kohonen00] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive document collection, IEEE transactions on neural networks 11, 2000, citeseer.nj.nec.com/378852.html.

[Krier02] M. Krier, F. Zaccà, Automatic categorization applications at the European patent office, World Patent Information 24, 2002, 187-196, www.sciencedirect.com/science/journal/01722190.

[Lagus98] K. Lagus, Generalizability of the WEBSOM method to document collections of various types, proceedings of 6th European Congress on Intelligent Techniques & Soft Computing (EUFIT'98) 1, 1998, 210-214, citeseer.nj.nec.com/lagus98generalizability.html.

[Lamirel01] J.-C. Lamirel, Y. Toussaint, J. Ducloy, C. Czysz, C. François, *Réseaux Neuronaux Avancés pour la Cartographie de la Science et de la Technique : Application à l'Analyse des Brevets*, proceedings of VSST 2001, 2001, 215-229, http ://www.inist.fr/uri/pdf/3_vsst21.pdf.

[Larkey00] L. S. Larkey. M. E. Connell, J. Callan, Collection Selection and Results Merging with Topically Organized U.S. Patents and TREC Data, CIKM 2000 proceedings, 2000, 282-289, citeseer.nj.nec.com/larkey00collection.html.

[Larkey99] L. S. Larkey, A Patent Search and Classification System, Proceedings of DL-99, 4th ACM Conference on Digital Libraries, 1999, 179-187, citeseer.nj.nec.com/larkey99patent.html.

[Larkey98] L. S. Larkey, Some Issues in the Automatic Classification of U.S. patents, AAAi-98 working notes, 1998, ciir.cs.umass.edu/pubfiles/ir-135.ps.

[Leclercq99] I. Leclercq, *L'INPI, Internet, et le Commerce Electronique*, proceedings of the PATLIB99 conference, 1999, www.european-patent-office.org/patlib/patlib99/presentations/leclercq2.pdf.

[Lent97] B. Lent, R. Agrawal, R. Srikant, Discovering Trends In Text Databases, proceedings of 3rd Int .Conf. On Knowledge Discovery and Data Mining, 1997, citeseer.nj.nec.com/lent97discovering.html.

[Lewis01] D. Lewis, F. Sebastiani, Report on the Workshop on Operational Text Classification Systems OTC-01, 2001, www.acm.org/sigir/forum/F2001/textClassification.pdf.

[Lewis02] D. Lewis, DDLBETA mailing list communication, 2002.

[Lexiquest] Lexiquest Categorize Executive Report and Data Sheet, SPSS, 2002, www.spss.com/spssbi/iberica/productos/lexiquest/lexiquest_categorize.pdf.

[Lingway] Lingway's TACSY product datasheet, Lingway, 2002, www.lingway.com/commonfiles/TACSY.pdf.

[Lyon99] M- Lyon, Language related problems in the IPC and search systems using natural language, World Patent Information 21, 1999, 89-95, www.sciencedirect.com/science/journal/01722190.

[Mack01] R. Mack, Y. Ravin, R. J. Byrd, Knowledge portals and the emerging digital knowledge workplace, IBM Systems Journal 40, 2001, 925-955, researchweb.watson.ibm.com/journal/sj/404/mack.pdf.

[Mase98] H. Mase, H. Tsuji, H. Kinukawa, M. Ishihara, Automatic Patents Categorization and Its Evaluation, IPSJ Journal 39, 1998, www.ipsj.or.jp/members/Journal/Eng/3907/article018.html.

[Mase98b] H. Mase, Experiments on Automatic Web Page Categorization for IR systems, technical report, Stanford University, 1998, citeseer.nj.nec.com/164846.html.

[MMS] P. Borne, M. P. O'Hara, C. Roesch, R. W. Skippon, Merged Markush Service User's Manual, INPI, 2002, www.inpi.fr/inpi/mms/usermanual.htm.

[Normier02] B. Normier, Multilingual access to European patent databases, seminar on "Protection and access to innovation in the Net: An environment for technological progress. Patents, trade marks, design", Madrid, 2002, www.wipo.org/scit/en/meeting/7/pdf/e_patent.pdf.

[Oracle] Oracle Text Technical White Paper, Oracle, 2002, technet.oracle.com/products/text/pdf/text_techwp.pdf.

[Owake00] Delegation of Japan: OWAKE system – primary automatic classification. Section 59 of WIPO report IPC/CE/29/11, 29th session of the committee of experts of the IPC union, 13-17 March 2000, www.wipo.org/classifications/en/ipc/ipc_ce/29/11.htm.

[Peters02] C. Peters, C. H. A. Koster, Uncertainty-based Noise Reduction and Term Selection in Text Categorization, proceedings of 24th European Colloquium on Information Retrieval Research (ECIR 2002), 2002, www.cs.kun.nl/~kees/home/papers/trec9.ps.gz.

[Porter02] M. Porter, Snowball stemmer homepage, snowball.tartarus.org

[Rodríguez97] M. de B. Rodríguez, J. M. Gómez-Hidalgo, B. Díaz-Agudo, Using WordNet to Complement Training Information in Text Categorization, Proceedings of the International Conference on Recent Advances in Natural Language Processing, 1997, xxx.unizar.es/ps/cmp-lg/9709007.

[Ruiz02] M. E. Ruiz, P. Srinivasan, Hierarchical Text categorization Using Neural Networks, Information Retrieval 5, 2002, 87-118, citeseer.nj.nec.com/ruiz02hierarchical.html.

[Schapire00] R. E. Schapire, Y. Singer, BoosTexter: A Boosting-based System for Text Categorization, Machine Learning 39, 2000, 135-168, citeseer.nj.nec.com/388632.html.

[Scott98] S. Scott, S. Matwin, Text Classification Using WordNet Hypernyms, proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language

Processing Systems, Association for Computational Linguistics, Montreal, 1998, 38-44, citeseer.nj.nec.com/sam98text.html.

[Scott99] S. Scott, S. Matwin, Feature Engineering for Text Classification, proceedings of ICML-99, 16th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, 1999, 379-388, citeseer.nj.nec.com/330021.html.

[Sebastiani02] F. Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys 34, March 2002, 1-47, faure.iei.pi.cnr.it/~fabrizio/Publications/ACMCS02.pdf.

[Smith02] H. Smith, Automation of patent classification, to appear in World Patents Information (preprint provided by M. Makarov).

[Stembridge98] B. Stembridge, International Patent Classification in Derwent databases, proceedings of the advanced seminar on the International Patent Classification, WIPO, IPC/SEM/98/4, 1998, www.wipo.org/classifications/en/ipc/ipc_ce/sem_98/4.pdf.

[Sowa01] J. F. Sowa, Building, Sharing, and Merging Ontologies, 2001, www.jfsowa.com/ontology/ontoshar.htm.

[Verity] Verity Intelligent Classifier datasheet, 2000, www.verity.com/products/pdf/MK0340_Int_Classifer.pdf.

[VerityCH] Verity in Action: Swiss Federal Institute of Intellectual Property, Verity, 2002, www.verity.com/customers/verticals/research/pdf/MK0430_VIA_SwissFed.pdf.

[Wermter02] S. Wermter, C. Hung, Selforganizing classification on the Reuters news corpus, Proceedings of the International Conference on Computational Linguistics, Taipeh, Taiwan, 2002, www.his.sunderland.ac.uk/ps/coling-232.pdf.

[Wermter99] S. Wermter, G. Arevian, C. Panchev, Recurrent Neural Network Learning for Text Routing, Proceedings of the International Conference on Artificial Neural Networks (ICANN99), 1999, 898-903, www.his.sunderland.ac.uk/ps/icann99.pdf.

[Yang99] Yiming Yang, An Evaluation of Statistical Approaches to Text Categorization, Information Retrieval 1, 1999, 69-90, citeseer.nj.nec.com/yang97evaluation.html.

[Yang99b] Yiming Yang, Xin Liu, A re-examination of text categorization methods, proceedings of SIGIR'99 conference, 1999, citeseer.nj.nec.com/yang99reexamination.html.

[Yamazaki97] T. Yamazaki, I. Dagan, Mistake-driven Learning with Thesaurus for Text Categorization, Proceedings of NLPRS-97, the Natural Language Processing Pacific Rim Symposium, 1997, xxx.lanl.gov/abs/cmp-lg/9706006.