Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed

Dissertation

zur Erlangung des akademischen Grades
Doktor rerum naturalium (Dr. rer. nat.)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

eingereicht von

Dipl.-Math. Daniel Eisinger
geboren am 31. März 1982 in Augsburg

Betreuender Hochschullehrer:	Prof. Dr. Michael Schroeder Technische Universität Dresden
Gutachter:	Prof. Dr. Christa Womser-Hacker Universität Hildesheim
Tag der Verteidigung:	7. Oktober 2013

Dresden, August 2013

Abstract

The patent domain is a very important source of scientific information that is currently not used to its full potential. Searching for relevant patents is a complex task because the number of existing patents is very high and grows quickly, patent text is extremely complicated, and standard vocabulary is not used consistently or doesn't even exist. As a consequence, pure keyword searches often fail to return satisfying results in the patent domain. Major companies employ patent professionals who are able to search patents effectively, but even they have to invest a lot of time and effort into their search. Academic scientists on the other hand do not have access to such resources and therefore often do not search patents at all, but they risk missing up-to-date information that will not be published in scientific publications until much later, if it is published at all.

Document search on PubMed, the pre-eminent database for biomedical literature, relies on the annotation of its documents with relevant terms from the Medical Subject Headings ontology (MeSH) for improving recall through query expansion. Similarly, professional patent searches expand beyond keywords by including class codes from various patent classification systems. However, classification-based searches can only be performed effectively if the user has very detailed knowledge of the system, which is usually not the case for academic scientists. Consequently, we investigated methods to automatically identify relevant classes that can then be suggested to the user to expand their query. Since every patent is assigned at least one class code, it should be possible for these assignments to be used in a similar way as the MeSH annotations in PubMed.

In order to develop a system for this task, it is necessary to have a good understanding of the properties of both classification systems. In order to gain such knowledge, we perform an in-depth comparative analysis of MeSH and the main patent classification system, the International Patent Classification (IPC). We investigate the hierarchical structures as well as the properties of the terms/classes respectively, and we compare the assignment of IPC codes to patents with the annotation of PubMed documents with MeSH terms. Our analysis shows that the hierarchies are structurally similar, but terms and annotations differ significantly. The most important differences concern the considerably higher complexity of the IPC class definitions compared to MeSH terms and the far lower number of class assignments to the average patent compared to the number of MeSH terms assigned to PubMed documents.

As a result of these differences, problems are caused both for unexperienced patent searchers and professionals. On the one hand, the complex term system makes it very difficult for members of the former group to find any IPC classes that are relevant for their search task. On the other hand, the low number of IPC classes per patent points to incomplete class assignments by the patent office, therefore limiting the recall of the classification-based searches that are frequently performed by the latter group. We approach these problems from two directions: First, by automatically assigning additional patent classes to make up for the missing assignments, and second, by automatically retrieving relevant keywords and classes that are proposed to the user so they can expand their initial search.

For the automated assignment of additional patent classes, we adapt an approach to the patent domain that was successfully used for the assignment of MeSH terms to PubMed abstracts. Each document is assigned a set of IPC classes by a large set of binary Maximum-Entropy classifiers. Our evaluation shows good performance by individual classifiers (precision/recall between 0.84 and 0.90), making the retrieval of additional relevant documents for specific IPC classes feasible. The assignment of additional classes

to specific documents is more problematic, since the precision of our classifiers is not high enough to avoid false positives. However, we propose filtering methods that can help solve this problem.

For the guided patent search, we demonstrate various methods to expand a user's initial query. Our methods use both keywords and class codes that the user enters to retrieve additional relevant keywords and classes that are then suggested to the user. These additional query components are extracted from different sources such as patent text, IPC definitions, external vocabularies and co-occurrence data. The suggested expansions can help unexperienced users refine their queries with relevant IPC classes, and professionals can compose their complete query faster and more easily. We also present *GoPatents*, a patent retrieval prototype that incorporates some of our proposals and makes faceted browsing of a patent corpus possible.

Acknowledgments

The work presented in this thesis was carried out in equal parts at Roche Diagnostics GmbH in Penzberg and at the Biotechnology Center (BIOTEC) in Dresden. It would not have been possible without all the help I received from many people in both places. First and foremost, I want to thank both my academic supervisor Michael Schroeder and my Roche supervisor Ulrich Wieneke for the scientific and financial support I received as well as the positive work environment and the academic freedom I was afforded.

I am also very grateful to Markus Bundschus for sharing both his scientific knowledge and his own experiences as a PhD candidate. Claudia Albrecht, Victor Denk and Rainer Volk introduced me to the patent search process and generously agreed - together with Markus - to invest the time necessary for evaluating the quality of different term extraction results. Martin Baron shared his experience with different text mining tools and use cases. During my time in Dresden, I benefited from many interesting and insightful discussions with Thomas Wächter and George Tsatsaronis which ultimately resulted in joint publications. I also want to thank Jan Mönning for introducing me to the GoPubMed framework during our joint development of GoPatents, and Götz Fabian for very helpful discussions about various programming-related topics.

Both at Roche Penzberg and in Group Schroeder, I received a very warm welcome after my arrival. I consider myself extremely lucky to have worked with not one, but two groups of extraordinarily friendly colleagues. In Penzberg, their number was too high to mention everyone, so I will restrict myself to Traudi Sieber and Sabine Schreindl as representatives for all members of the always-helpful library staff as well as the aforementioned Markus and Viktor for the Scientific Information Services. Special thanks go to Bibiane Büring and Claudia for making it their personal mission to help me settle in quickly in my early days there.

Among the current and former members of Group Schroeder that I haven't mentioned yet, I want to thank the driving force behind my improved "Kicker" abilities, Matthias Reimann, my frequent football discussion partner Rainer Winnenburg (despite his misguided team choice), our great system administrator and professional group photographer Alex Mestiashvili, my tennis nemesis Christoph Baldow, my Berlin flatmate Sebastian Salentin, the always friendly and helpful Zerrin Isik, movie friends Anne Tuukkanen, Alicja Kozikowska, Maria Kissa and Alina Petrova and the most funky among my friends, Joachim Haupt and Sainitin Donakonda.

Most of all, I want to thank Simone Daminelli for many very enjoyable Neustadt and TV show evenings, Janine Roy for constantly trying out new sports and inviting me to join her, and Norhan Mahfouz for being the best Arabic teacher and friend I could have wished for.

Last but certainly not least, many thanks to my parents Helmut and Ingrid and my brothers Christopher and Johannes for always being there for me.

Contents

1 Introduction

As evidenced by a growing number of reports about various high-profile patent trials in recent years, having the necessary information about all relevant competitor patents can be vital to a company's interests. At the same time, patents can also be a valuable source for academic research, since current research results are often first published in a patent and only afterwards (or never) in a journal. Experts have estimated that only 10-15% of the patent content is also described in other publications, and that 80-90% of all scientific knowledge is contained in patents [1]. Despite that potential, most academic researchers are to our knowledge not using patents, presumably due to the high complexity of the domain.

As we will show in Section 2.1, this complexity is in part due to the high and fast-growing number of existing patent documents (about 80 million [2]). Additionally, these documents are not always available in English, which makes finding all relevant documents extremely difficult. But even for the documents with English-language versions, there are some unique challenges that separate the patent domain from most other document types. While it is not unusual to rely mainly on keywords for searching most other document corpora, this approach does not return satisfactory results for many patent search tasks. Not all information is contained in text, it is often also necessary to consider the images in the patent, e.g., for chemical formulas. Section 2.3 describes some problems with the patent text itself: Many patents are very long, and so are individual sentences of the patent. Different sections of the patent text are written in completely different styles, patent authors don't always use standard terminology (or it may not even exist), and many patents are written in very unspecific language. The problem has been summarized by the European Patent Office (EPO) in the following way, using the term "patentese" for the unconventional language style that is typically only used in patents: "Newcomers to intellectual property are often surprised or even shocked at the way words or phrases familiar in everyday language are used very differently in the world of patents. Grammatical constructions that would be unthinkable in everyday speech or writing are used routinely in patentese. Patentese has words which do not even exist in ordinary languages. Furthermore patentese exists in every conceivable natural language version" [3].

As a result of these problems, professional patent searches usually don't rely exclusively on keywords. Various metadata (e.g., the inventors or assignees of the patents) can be used to filter

or expand the results of an initial search, and can therefore be considered valuable resources that set patents apart from other document types. However, the most important way to improve pure keyword searches is through the use of classification information. All patents are assigned classes from expansive classification systems in order to categorize them according to the type of invention they represent (cf. Section 2.4). This information can also be used to filter or expand search results, but in order to make the most of these possibilities, the searcher must have detailed knowledge about the classification system. Unfortunately, this is not the case for many academic researchers. Even for professional patent searchers, the process of constructing and refining patent queries is quite complicated and time-consuming.

Consequently, it is desirable to offer a) an easier option for scientists to formulate high-quality patent queries and b) a system that assists patent professionals in completing their initial queries. A system for this task would ideally be able to both complete the user's initial keyword query and propose relevant expansions based on classification information. In order to provide such assistance, it is important to have a clear understanding of the properties of patent classification systems. We therefore carry out an in-depth investigation of the most common patent classification system, the International Patent Classification (IPC). We examine both its internal properties (namely its classes and the hierarchic structure) and the way its classes are assigned to patent documents. As a point of comparison, we perform the equivalent analyses on the controlled vocabulary "Medical Subject Headings" (MeSH) that is used to annotate all document abstracts on the biomedical literature database PubMed. As Section 2.5 will demonstrate, MeSH is already used to improve search results on PubMed by automatically expanding user queries and offering additional search functionality.

As a solution to problems we discovered through our analysis, we propose two approaches: a system for the automated assignment of additional classes to patent documents and a guided patent search system that assists the user by offering query expansion suggestions. Suggestions are given in the form of both additional keywords and additional IPC classes, and they are retrieved from patent texts, IPC class definitions and class co-occurrence data or using existing knowledge from external sources.

1.1 Thesis outline

Following this introduction, this thesis is organized as follows:

- Chapter 2 gives a more detailed overview of the patent domain and describes the problems that are addressed by this thesis. These problems include the high number of patent documents, the special properties of the text that make keyword search problematic and the complicated classification systems that are used to improve patent search. Additionally, we give a short introduction to existing (non-patent) document search engines that are already using classification systems to improve the search results. We conclude the chapter with an overview of previously published work that is relevant to this thesis.

- Chapter 3 describes our comparison of two annotation/classification systems for different document types: IPC classes assigned to patents and MeSH terms assigned to PubMed documents. We compare the hierarchies as well as the terms of both systems, and we then investigate differences in the way that documents are annotated. As a consequence of our analysis, we identify potentially problematic aspects for patent search. After that, the two following chapters propose solutions to these problems.

- In Chapter 4, we present our effort to assign additional classes to existing patents using a large set of binary Maximum-Entropy classifiers. We describe our method in detail, and we evaluate our results based on existing classification data. Positive and negative aspects of our categorization results are identified, and we discuss ways to improve the results by filtering them appropriately for different applications of our system.

- Chapter 5 describes different proposals for assisting both professional and unexperienced patent searchers. We present methods to suggest additional keywords as well as classes based on initial user queries. The possible sources of the suggestions include patent text, class definitions, external knowledge sources and co-occurrence information. We also introduce our prototype system *GoPatents* that incorporates some of our proposed methods.

- Chapter 6 describes the methods used in the three preceding chapters (i.e., for the comparison MeSH-IPC, automated patent categorization and guided patent search) in more technical detail.

- The thesis is concluded in Chapter 7 with summaries of the existing problems and our approaches to solving them.

2 Background

Summary

The annual number of patent applications has been rising for decades, reaching a new record in 2011 with more than two million applications worldwide. In total, the number of existing patent documents has surpassed 80 million, with almost 8 million patents in force by 2011; all these numbers are expected to rise further.
In addition to the high number of documents, patents have some unique characteristics that cause problems for patent search. Patent text is often extremely complicated, text styles differ between sections and authors, and standard vocabulary is avoided by some authors or may not even exist yet. These problems necessitate the use of classification information in addition to keywords. Many different classification systems can be used and combined for this purpose, with most systems being based on the International Patent Classification.
Other document annotation systems have been used to improve document search in other domains. The controlled vocabulary "Medical Subject Headings" serves this purpose for multiple efforts to offer different or better search functionality for the biomedical literature database PubMed, making it a logical point of comparison for our goal of improving patent search with the help of classification information.
Among the related work, efforts to automatically categorize patents have with one exception not yet attempted to use the complete hierarchy down to the subgroup level. Patent retrieval systems often include the option to use classification information for search, but in most cases it is not implemented in a way to make its use convenient for the searcher, and additional relevant classes or keywords are not suggested.

Parts of this chapter were previously published in:

- Daniel Eisinger, Thomas Wächter, Markus Bundschus, Ulrich Wieneke, Michael Schroeder. Analysis of MeSH and IPC as a Prerequisite for Guided Patent Search. Bio-Ontologies 2012. http://bio-ontologies.knowledgeblog.org/346

- Daniel Eisinger, George Tsatsaronis, Markus Bundschus, Ulrich Wieneke, Michael Schroeder. Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed. Journal of Biomedical Semantics 2013, 4:S1. http://www.jbiomedsem.com/content/4/S1/S3

As has been mentioned in Chapter 1, there are numerous specific complications for patent search compared with other more conventional sources. This chapter will investigate these issues in more detail. The first section gives information on the number of existing patent documents and the growth of the domain, and the following sections examine the special properties of patent text and patent classification systems. After presenting the basic ideas behind the use of document annotations for search improvement, the chapter is concluded with a presentation of related work.

2.1 Patent statistics

The number of patent applications continues to rise, for the first time surpassing two million worldwide in 2011 alone [4]. As Figure 1 shows, this number has grown each year since 1995, with two notable exceptions in 2002 and 2009 due to economical crises. After growth had been slowing down since 2005, the growth rate increased greatly after 2009, surpassing 7% in both 2010 and 2011. These growth rates represent the strongest two consecutive years since at least 1995, which is especially remarkable considering the continued effects of the global financial crisis.

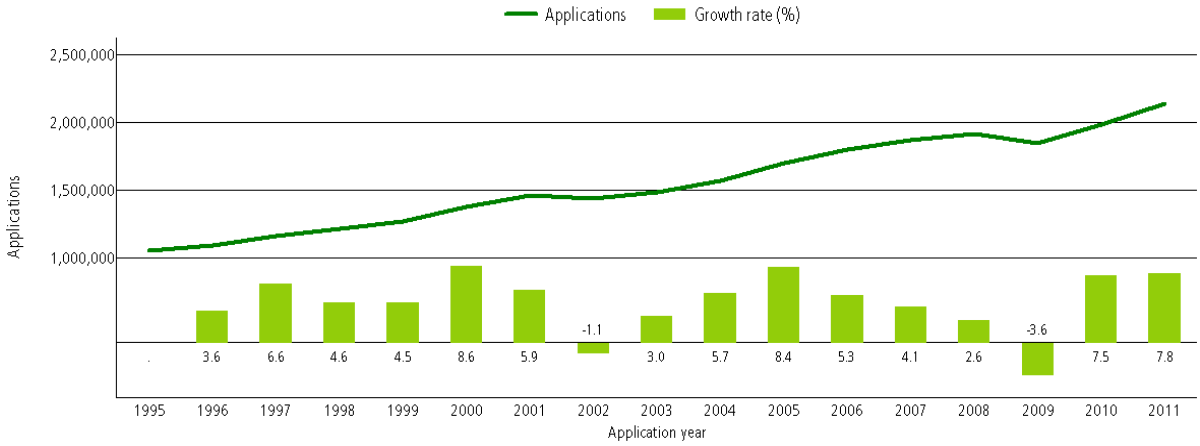


Figure 1: Number of patent applications worldwide 1995-2011 (adapted from [4]). With two exceptions, the number of applications has grown each year.

Figure 2 shows the long-term trend for the five top patent offices according to their 2011 total application numbers. It reveals that while extreme increases in China play a large role in the overall increase, all other top offices except Japan have also been contributing. China surpassed the United States in 2011, becoming the largest patent office in the world according to the number of received applications. Japan on the other hand reached its peak around 2000 already and has since then

been surpassed by both the US and China. The European and Korean patent offices are far below the top three, but have also for the most part shown moderate but continuous growth during the last three decades.

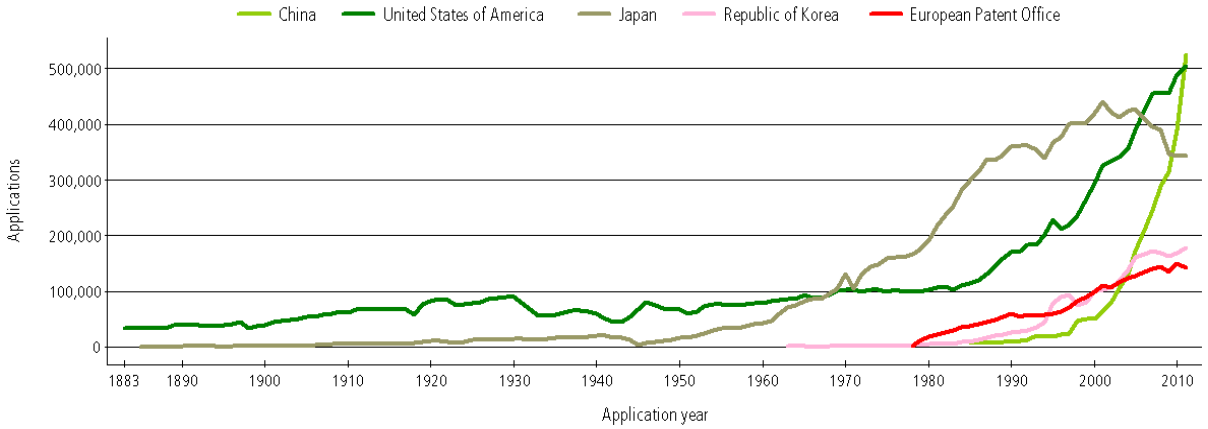


Figure 2: Trend in patent applications for top five offices (adapted from [4]). Application numbers have been growing extremely fast in China for the last decade, and all offices except Japan show strong growth over the last two decades.

While the number of patent applications is a good indicator for trends in the patent world, it is also important to look at the number of granted patents. As shown in Figure 3, the basic trend in the number of granted patents is similar to the situation for applications shown in Figure 1. While the succession of weak and strong years isn't completely identical, both data sets show strong growth in 2010 and 2011 after a number of weaker years. In 2011, a new record was set with almost one million grants.

There is no reliable data on the total number of existing patent documents, but it was estimated to be about 50 million by 2006 already [1], and the European patent database Espacenet contained almost 80 million documents as of late 2012 [2]. However, many of these documents correspond to patents that expired (either due to the patent protection ending or the patent owner failing to pay the renewal fees) or to applications that have been rejected or are still under consideration. The number of patents that were in force worldwide in 2011 was estimated to be 7.88 million, up from 6.88 million three years previously [4]. The ten offices with the largest number of patents in force are presented in Figure 4 with the respective numbers of patents. The US are the clear leader in this category with more than 2.1 million active patents, and Japan with more than 1.5

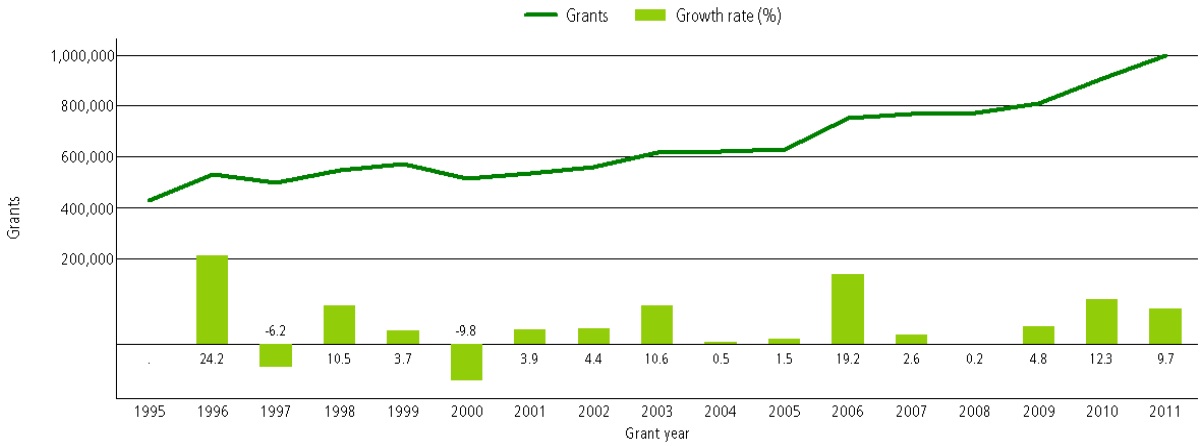


Figure 3: Number of patents granted worldwide 1995-2011 (adapted from [4]). With the exception of the years 1997 and 2000, the number of patent grants has grown each year.

million is still far ahead of China despite its decline in applications over the last decade (cf. Figure 2). Figure 4 also ranks the major European countries according to their active patents, showing Germany ahead of the United Kingdom and France as the top three. Note that the fact that a patent is not in force does not have to mean that it is not relevant. If an application is pending, it may still become a valid patent, and the content of expired patents can be very important for determining prior art. The number of documents in a comprehensive patent corpus is therefore extremely large.

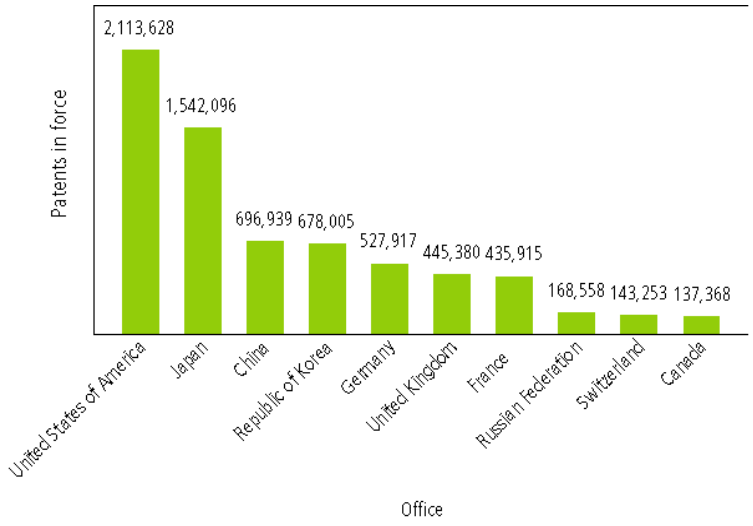


Figure 4: Number of patents in force worldwide 2011 for top ten offices (adapted from [4]).

All statistical figures in this section were adapted from the 2012 edition of the World Intellectual Property Indicators report [4] that was published by the World Intellectual Property Organization (WIPO)[1]. The represented data are taken from the WIPO Statistics Database that is based on information supplied by national and regional patent offices from around the world. Missing data (usually from small offices) are estimated for the world totals, potentially leading to small errors. However, since the available data covers 98% of all data, estimation errors should not have a major effect.

2.2 Patent search types

Patent search is an extremely complicated and time-intensive search task. According to a survey among 81 patent professionals [5,6], the average amount of time needed to complete a single search task is 12 hours, with complicated tasks taking up to 40 hours. There are a number of vastly different motivations for performing patent search, and also a number of slightly different ways to define the resulting different search types (cf. [3, 7–12]). It is very important that patent search is conducted appropriate to the search type, since it has a large influence on the required recall and precision of the search, and the search strategy must therefore be tailored to the objective of the search [12–14]. The following list gives a brief overview of different objectives and the consequences for the search process.

- The "technology survey" / "state of the art search" / "patent landscape search" is supposed to give a broad overview over a specific field before a company invests resources into research projects or enters into licensing agreements. Other applications for this search type include competitor analysis, technology trend watch and the compilation of country or enterprise statistics [9]. Since this search type is not required to deliver comprehensive results, technology surveys are supposed to be the least complex and time-consuming patent searches. Precision is therefore more important than recall.

- The "infringement/clearance/freedom to operate search" is supposed to evaluate legal risks caused by the introduction of a new product. Its purpose is the search for patents that the new product may infringe upon. Since a single missed patent could cause very serious legal and financial problems later, this search type has a strong need for high recall. Therefore, it

[1] http://www.wipo.int/portal/index.html.en

often necessitates a very elaborate and time-consuming search strategy.

- The "novelty/patentability search" looks for previous patents that may be in conflict with a new patent application. It is usually carried out both by the company before applying for a patent and by the patent examiner after that. On the one hand, prior art found by the company can be used to phrase the claims in a way that will distinguish the patent from existing ones and therefore raise the likelihood of its acceptance. On the other hand, the patent examiner uses the prior art that resulted from their search to decide whether the patent application fulfills the novelty requirement that is necessary for the patent to be granted. It is very important to find any existing prior art, since missing anything might result in an erroneously accepted patent and could therefore lead to legal problems for the company later.

- The "validity/invalidity search" can be considered a variation of the patentability search. The main difference concerns the source document for the search, which is an already granted patent in this case; the objective is either to judge the value of a patent or to find a way to invalidate a competitor's patent. If relevant elements of a patent had been patented before its application, it is possible to get the granted patent protection removed by legal means.

The term "prior art search" is also frequently used to describe a patent search type, especially for the search that is performed by the patent examiner in order to determine whether the patent should be granted. However, it is not always made clear if it is used in this way (as a synonym for a patentability search) or as a generic term that also describes infringement and invalidity searches.

It should be noted that for most of these patent searches, other types of text are also relevant. According to Article 54(2) of the European Patent Convention (EPC), "prior art" comprises "everything made available to the public, in writing, by public use, or otherwise, before the filing date of a patent application" [15]. The patent laws of the USA and of Japan also have corresponding articles. This means if the content of a new patent application was previously published in a scientific journal, the patent will not be granted - regardless of whether the author of the publication is an employee of the company applying for the patent or some third party with the same idea. Despite the relevance of other types of text, this thesis is concentrated on improving patent search since there are numerous well-known options for searching most other document sources. The following section investigates special properties of the text used in patents.

2.3 Patent text

In addition to the high number of patents, the text of individual patents is also often fairly long. An average length of almost 4,000 words per full-text patent was calculated by Iwayama *et al.* [16] for a test collection of Japanese patents, and the standard variance of over 3,000 words showed that there is also a high amount of variation in patent lengths. The longest patent consisted of over 250,000 words, and the highest number of unique words in a patent was over 29,000.

Patent text has some special characteristics separating it from most texts from other sources. One key point is the language used in different fields of the patent: While the abstract is usually written in a fairly clear and natural way, the claims have a very rigid structure and are written in a very technical language style, as evidenced by the patent excerpt provided in Figure 5.

Additionally, each claim is required to be defined in a single sentence, leading to long and complicated grammatical constructions that chain multiple sentences and cause problems for grammatical parsers. Parapatics *et al.* [17] used the well-known Stanford Natural Language Parser on the claims of a small test corpus and reported large differences between the results for independent claims (e.g., Claim 1 in Figure 5) and those for less complex dependent claims that refer to and extend other claims (e.g., Claims 2 and 3 in Figure 5). While dependent claims contained less than 35 words on average and were almost always parsed correctly, independent claims contained more than 120 words, and the proportion of successful parses dropped by 20-50% depending on the heap size. In other words, claims are often quite hard to understand - for human readers as well as computers. Unfortunately, the claims section is arguably the most important part of the patent [8], since it is the legal basis of all patent protection [18].

The different text styles can lead to significant differences of the vocabulary used in different sections of the patent [19]. Additionally, authors from completely different backgrounds are involved in writing patents, and the technical fields covered by the patents also vary considerably [18]. The consistency of vocabularies between different patents is therefore also very low, leading to further issues for natural language processing (NLP) tasks [20].

A further complication is caused by the different author intentions compared to research papers: While scientists usually have the goal to inform as many people as possible about their findings, many patent authors have the opposite goal. A patent assignee is granted timed exclusive rights to use and license the patented invention, but in exchange for that privilege they have to publish

What is claimed is:

1. A stringed musical instrument comprising an instrument body having front and rear surfaces, sound producing means extending over a portion of said front surface, and a device mounted onto said rear surface for positioning said instrument body at an angular orientation to a player's body, said device including attachment means movable between an inoperative position overlying said rear surface and an operative position at an angle to said rear surface, a pair of spaced-apart mounting blocks attached to said rear surface and support means coupled to said mounting blocks for rotationally supporting therebetween said attachment means, said attachment means engaging said player's body when in said operative position for maintaining said instrument body in said angular orientation and disengaging from said player's body when in said inoperative position for maintaining said instrument body in other than said angular orientation.

2. The stringed musical instrument of claim 1 wherein said support means comprises a rod extending between said mounting blocks.

3. The stringed musical instrument of claim 2 wherein said attachment means is movably mounted on said rod for rotational movement between said operative and inoperative positions and for lateral movement between a locked and unlocked position.

Figure 5: Beginning of claims section from US patent 4656917, "Musical instrument support".

detailed information about it. Competitors may be able to use that information to find out more about the technological processes used by the assignee, and possibly even to "design around" the patent - reaching the same goal without violating the patent. Therefore, obvious search keywords are often intentionally omitted from patents and replaced by very specialized vocabulary, making the retrieval of relevant patents difficult [21]. On the other hand, many companies also use very unspecific vocabulary in order to make the scope of their patents as broad as possible. As an example, the images and text description of US patent 4656917 make it clear that the patented invention is intended for guitars. Despite that, the claims text of the patent (cf. Figure 5) only mentions "stringed musical instruments"; this prevents potential competitors from applying the same principle to violins or other string instruments. Lastly, since patents describe new inventions, it can't be avoided that new terminology has to be introduced in some cases. Since different inventors are often working on similar new technologies simultaneously, different terms will be

introduced for the same subject before the technology becomes mature enough to have an accepted common vocabulary. In many cases, multiple different terms remain in use by different authors [22]. As a result of all these factors, finding most or all relevant keywords for a particular patent search will at the very least require a large investment of time and effort.

Consequently, professional patent searches usually don't rely exclusively on keywords. While much information is contained in the text parts (mainly title, abstract, claims and description), it is often also necessary to pay attention to the included drawings and various metadata (e.g., inventor, assignee). Most importantly however, the use of classification information can improve pure keyword searches considerably and is therefore an integral part of most professional patent searches [23]. The following section will therefore provide an overview of the most important patent classification systems as well as further details about the benefits (and potential problems) their use can bring to patent search.

2.4 Patent classification systems

As a consequence of the problems with keywords, professional searchers have long been using classification information [24, 25] - either for expanding keyword-based searches or for filtering overly large result sets. Patents are classified into hierarchical systems of categories by patent offices according to the type of invention they represent. This work will concentrate on the International Patent Classification (IPC)[2] since it is the best-known and most widely spread classification system [26]; it is used by over 100 patent-issuing bodies worldwide [27] and contains around 70,000 classes. For more details on the IPC, see Section 2.4.1. In addition to that, most other important systems are directly based on it; they take the IPC hierarchy and add additional subclasses.

That way, the "Deutsche Feinklassifikation" (DEKLA) adds nearly 40,000 subdivisions to the German version of the IPC [28] (but there is no official English translation), the European Classification (ECLA) basically doubles the size of the IPC to more than 132,000 classes, and the Japanese File Index (FI) contains about 170,000 classes [1]. Despite their large size, there are additional refinements for both the European and Japanese systems. The little-known "In Computer Only" (ICO) index terms were developed at the EPO and can be used for information of secondary importance or to index additional aspects of an invention [26, 29]. They were mainly intended for internal use by patent officers, but have been included in a number of patent databases offered by

[2] http://www.wipo.int/classifications/ipc/en

commercial providers like Derwent, Questel or STN. The number of codes had already surpassed 106,000 in 2010, making their size comparable to the complete ECLA itself. The ICO term hierarchy mirrors ECLA to a certain degree: Each ECLA section (A to H) has an corresponding ICO section (K to T), but not all classes from the main system have equivalent ICO classes. Similarly, the so-called F-terms[3] are used to index aspects of Japanese patents that are not completely covered by the assigned FI codes. As explained by Iwayama *et al.*, technological fields ("themes") are defined as sets of FI codes with individual collections of viewpoints relevant for the theme, and there is a list of possible elements for each viewpoint [30]. For example, the theme "Dental Preparations" (4C089) is relevant for the FI code A61K 6/00 ("Preparations for dentistry") and all its subcodes until A61K 6/10. The seven viewpoints for theme 4C089 include aspects like "Purpose of use", "Metal component" and "Inorganic component". Each of these viewpoints has between eight and 20 elements, e.g., "containing silicon" and "containing phosphorus" for the viewpoint "Inorganic component". This additional data is supposed to give a more complete overview of the contents of the classified patent than it would be possible when only using the classification system by itself. A more detailed example for the connection between File Index and F-terms is presented by Wolter [26].

An important system that is entirely separate from the IPC is the United States Patent Classification (USPC), containing more than 150,000 classes[4]. The US Patent and Trademark Office (USPTO) offers a US-to-IPC conversion tool, but it is only intended for giving hints about potentially relevant classes; the systems are too different to make a direct concordance possible. The USPC is arguably the most complicated system, with different classes being based on different classification principles and very unintuitive class codes that do not reflect the hierarchical position of classes for the most part [31,32]. Even experienced patent searchers confess to having problems with finding the right classes for their search and complain about a decline in quality in recent versions of the system [26]. Possibly partially in response to such concerns, the USPTO and the European Patent Office (EPO) reached an agreement in 2010 to create a new system based on ECLA and integrating elements from both ICO and USPC. This new system is called Cooperative Patent Classification (CPC)[5] and contains about 250,000 entries in its first version that was for-

[3] http://www5.ipdl.inpit.go.jp/pmgs1/pmgs1/!frame_E?hs=1&gb=2&dep=1
[4] http://www.uspto.gov/patents/resources/classification/overview.pdf
[5] http://www.cooperativepatentclassification.org/about.html

mally launched in January 2013. It has now replaced both ECLA and USPC and is used by more than 45 patent offices worldwide according to the USPTO[6]. The State Intellectual Property Office of the People's Republic of China (SIPO) was announced as an additional cooperation partner in June 2013[7], and the first updated CPC scheme (2013.07) was released in July[8]. However, only a minority of detailed class definitions has been made publicly available as of early August 2013. It is not yet clear if the introduction of CPC will make former systems ECLA and USPC redundant, or if they will still be used for searching older patent corpora. It has also been pointed out that the new system can only improve the previous situation if the examiners from both patent offices can be trained to use it in the same way [26].

The Derwent World Patent Index (DWPI) Classification[9] is a commercial product that is intended to improve the performance of patent retrieval tools. The DWPI Classification isn't directly based on the IPC, but some sections are closely aligned to it. Relevant classes are added manually to patents by experts in their fields. The number of classes is considerably smaller than for the major official systems, but we were unable to find the exact number, and we could not calculate it ourselves since the system is not freely available. Derwent's approach of having experts assign classes from its private system is supposed to compensate for a problem that was already recognized decades ago [33]: different interpretations of the scope of IPC classes or different classification philosophies can lead to inconsistent and missing class assignments [9], especially between international patent offices. The following paragraph presents more detailed analyses of this problem, and Table 1 gives an overview of all classification systems we introduced.

As mentioned before, classification information is the most important element of patent searches apart from keywords. Becks *et al.* report drastic recall improvements for patent retrieval from the inclusion of classification information [34]. Parisi *et al.* recommend the use of classification information in searches, but they warn that the assigned class codes may not always be the most appropriate ones, going so far as to say that existing assignments may be "subjective, incomplete or inconsistent and sometimes even random" due to overworked patent examiners [35]. In a practice-oriented article on chemical patent search, Annies investigates the coverage of different classification systems for patents about chemical formulations [23]. He points out that the inclusion

[6]http://www.uspto.gov/news/pr/2013/13-01.jsp
[7]http://www.epo.org/news-issues/news/2013/20130604.html
[8]http://www.cooperativepatentclassification.org/cpcSchemeAndDefinitions.html
[9]http://ip-science.thomsonreuters.com/support/patents/dwpiref/reftools/classification/

System name	Approx. # of entries	Main region	Type of system	based on IPC
IPC	70,000	international	main classification	=
ECLA	130,000	Europe	main classification	directly
ICO	110,000	Europe	additional indexing	indirectly
File Index	170,000	Japan	main classification	directly
F-Terms	350,000	Japan	additional indexing	indirectly
DEKLA	110,000	Germany	main classification	directly
USPC	150,000	United States	main classification	no
CPC	250,000	United States / Europe	main classification	directly
DWPI	?	international	(separate) main cl.	partially

Table 1: Major patent classification systems. Most systems contain between 110,000 and 170,000 classes and are directly or indirectly based on the IPC.

of classification information is crucial when searching for formulation types and dosage forms due to the non-standardized wording describing the inventions in this domain. However, just using one classification system may not be enough to retrieve a comprehensive result set since the coverage of different systems varies greatly: For a set of patent documents about chemical formulations and dosage forms indexed either with Derwent Manual Codes or IPC, only a minor proportion of the documents was indexed with the respective relevant classes from both systems (36% for pharmaceutical formulations and just 19% for agrochemical formulations). He concludes that the inconsistent application of the different classification systems necessitates the addition of as many codes as possible for improving search results. In a detailed overview of the practical implications of using the different systems we explained above for patent search, Wolter makes the same argument [26]. He argues that the advantages of combining multiple classification systems make up for the high difficulty and large necessary time investment of finding the relevant classes in multiple systems. These advantages include the option to find patents from more countries (especially Japan), the potential identification of additional subject matter that might not be included in the searchable text and the possibility to search for different aspects of inventions that may be covered in different ways by the various systems. In a case study searching for prior art on the anti-ulcer drug pantoprazole, Emmerich compares the results of full-text patent searches with value-added patent information such as DWPI and reports that the use of this additional information is essential for all searches with high commercial relevance [36]. Other case studies have recommended the use of classification information for searches concerning antibodies [37] and biopharmaceuticals [38]. Despite the described advantages of using multiple systems, the IPC is still described as "one very

important arrow in a patent searcher's quiver" with its general use remaining "uncontested" [26]. We will give a short introduction to the IPC in the following section.

2.4.1 International Patent Classification

The first version of the IPC was published in 1968; it was initially updated roughly in five year intervals. After a major reform effort, the eighth IPC version entered into force in 2006. After this, the frequency of updates became significantly higher; the currently valid version, IPC-2013.01, is already the seventeenth version. However, the majority of the changes after IPC8 did not alter the structure of the hierarchy significantly, and a lot of existing patent data still carries IPC codes according to that version. For that reason, the majority of our work is based on IPC8. It contains 69487 entries in up to 14 hierarchy levels; that number has risen to 71644 entries in the 2013 version.

The IPC hierarchy for all versions is divided into eight sections that correspond to very general categories such as "Human necessities" (Section A) or "Chemistry; metallurgy" (Section C). Table 2 shows an overview of all eight sections of the IPC. Each section is made up of numerous classes (e.g., A61), and each class contains multiple subclasses such as A61K. Each subclass is again divided into main groups (e.g., A61K 38/00), and for most main groups there are additional subgroups such as A61K 38/17. The definitions for these hierarchy entries are listed in Table 3; a more detailed description of the hierarchy can be found in the WIPO-published "Guide to the IPC"[10]. Despite its specific meaning in IPC terminology, we will use the term "classes" as an umbrella term for all individual entries of the IPC instead of the correct term for the respective level of the hierarchy, in order to improve readability.

Section	Definition
A	Human necessities
B	Performing operations; transporting
C	Chemistry; metallurgy
D	Textiles; paper
E	Fixed constructions
F	Mechanical engineering; lighting; heating; weapons; blasting
G	Physics
H	Electricity

Table 2: IPC sections with definitions. The sections correspond to a very broad categorization of patents.

[10] http://www.wipo.int/classifications/ipc/en/guide/guide_ipc.pdf

As the class examples in the previous paragraph show, individual entries of the IPC hierarchy are mainly represented by alphanumeric codes. The corresponding definitions are complicated and often depend on each other. Table 3 shows the complete definition tree for subgroup A61K 38/17 from the example above. The class hierarchy and the definitions will be investigated in more detail in Sections 3.1 and 3.2.

Class code	Definition
A	Human necessities
A61	Medical or veterinary science; hygiene
A61K	Preparations for medical, dental or toilet purposes
A61K 38/00	Medicinal preparations containing peptides
A61K 38/16	Peptides having more than 20 amino acids; gastrins; somatostatins; melanotropins; derivatives thereof
A61K 38/17	from animals; from humans

Table 3: IPC definition tree for class A61K 38/17. Definitions can be long and complicated and depend on each other.

This example illustrates the necessity to also consider the superordinate code definitions in order to understand what kind of invention is represented by the given code. The definition "from animals; from humans" is not useful by itself, but combining it with the superordinate code definitions results in a clear description of the category represented by the code. It also shows that the code alone does not accurately represent the hierarchy in all cases: While class 38/17 is directly subordinate to 38/16, there is no direct hierarchical connection between classes 38/16 and 38/15 (definition: "Depsipeptides; derivatives thereof"). Finding the most relevant classification codes to be used for search therefore constitutes a significant challenge, especially for users with little experience.

As a consequence of the high complexity of searching patents, major pharma companies employ patent professionals to relieve their scientists of this difficult and time-consuming task. Their patent searches often combine keywords and classification codes (and possibly additional metadata) in a single query. Many researchers without access to such resources ignore patents in favor of more accessible scientific literature. However, as we mentioned above, they thereby risk missing a lot of current research results. This study intends to investigate ways for

1. assisting patent professionals in finding the relevant elements (keywords and classification codes) for their search queries more efficiently and

2. enabling non-professionals to start using patents as an important additional information

source.

Due to the problems for keyword search caused by the complexities of patent text (cf. Section 2.3), we focused on the use of classification information for these tasks. The idea of using existing annotations to improve document search has already been successfully implemented for other text domains, as the short overview in the following section shows.

2.5 Improving document search using term annotations

There are numerous examples for term annotation systems - controlled vocabularies, taxonomies or ontologies - that are used in the biomedical domain for assisting document search. Many text mining and document retrieval applications rely on text annotations with terms from existing term systems - either by using existing manual annotations or by automatically assigning relevant terms. These document annotations can be used to directly improve document search: If the user query contains one or more annotation terms, documents that were assigned these terms can be considered relevant for the query even if the terms are not contained in the text. Section 2.5.1 shows an example of this functionality in the biomedical literature database PubMed.

Other systems make use of document annotations by offering *faceted search* (also called *faceted browsing*) functionality to users. This means that the resulting documents can be filtered according to their annotation terms, allowing the user to quickly and easily reach a result set with very high relevance. This is especially useful if the annotation systems are hierarchically organized, since this adds the possibility of choosing more specific or more general filter terms in reaction to the results of the search. The search engine *GoPubMed*[11] offers an alternative entry to the documents available via PubMed and offers faceted browsing of search results using the terms from Medical Subject Headings (MeSH) and Gene Ontology (GO) as well as a protein database. Users can immediately see which of the terms from these three systems were assigned to the result documents most frequently and choose a combination of relevant terms with a small number of clicks [39, 40]. Figure 6 shows a small part of the term annotations for an example search result. The "M" and "G" letters to the left of the terms tell the user whether the term is from MeSH or Gene Ontology, and the numbers to the right give the number of documents among the result set that were annotated with the respective term. In the figure, the user has decided to restrict the search to documents about coronary vessels and filter out documents with a connection to stem cell development. The

[11] http://gopubmed.com/web/gopubmed/

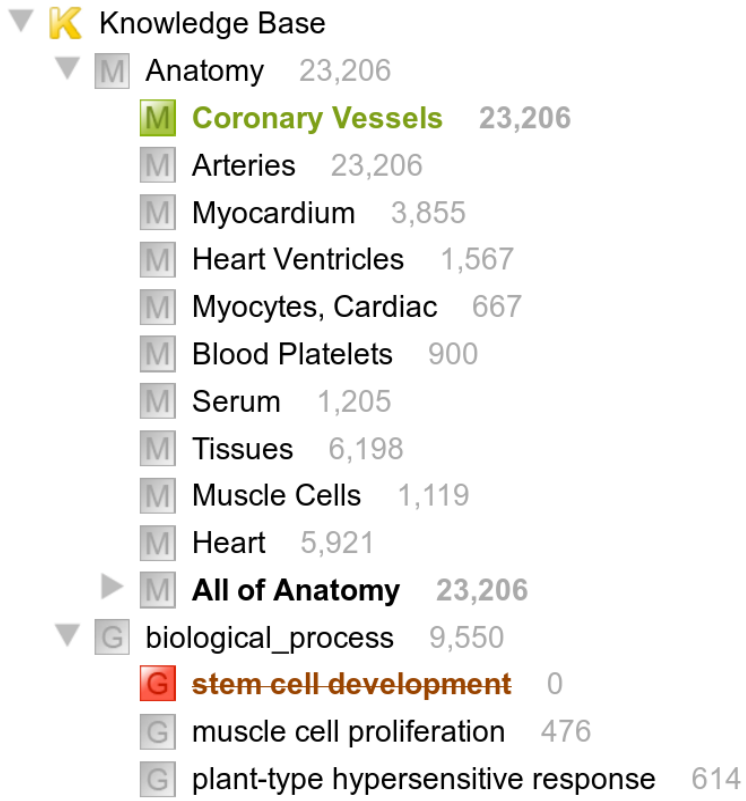


Figure 6: Faceted search with GoPubMed. Excerpt from result page showing the option to filter the result set by including or excluding MeSH or GO terms.

system *EBIMed*[12] also accepts PubMed queries and annotates the resulting documents with GO terms and protein names, but its goal is the identification of related pairs of terms rather than a set of relevant documents [41]. Its sister application *Whatizit*[13] asks the user to enter a query in the form of a text passage and offers sets of modules for different text analysis tasks such as the annotation with relevant terms from different collections, the identification of relations between terms and the retrieval of related PubMed abstracts [42]. All these systems are mainly or exclusively intended for use with PubMed documents; there is no comparable system for patents.

In order to test the validity of our approach of using patent classification information for these tasks, we analyzed the code assignment in the patent domain and compared it with the assignment of Medical Subject Headings to documents in the biomedical literature database PubMed. Although PubMed has a considerably more narrow focus than the patents do, we consider this comparison a

[12] www.ebi.ac.uk/Rebholz-srv/ebimed
[13] http://www.ebi.ac.uk/webservices/whatizit

useful approach for the following reasons: First, PubMed represents (to our knowledge) the largest freely accessible collection of scientific documents (or more precisely, abstracts) indexed with a controlled vocabulary, making it a natural target for our comparison of document annotations. Second, as we will describe in the next section in more detail, the assigned terms are already used for improving PubMed searches, mirroring our plan for the IPC codes. And third, although our patent corpus contains patents from many different fields, it is our main objective to improve patent search for the biomedical domain.

2.5.1 Medical Subject Headings

The Medical Subject Headings (MeSH) are a controlled vocabulary thesaurus of biomedical terms curated by the National Library of Medicine (NLM). MeSH was first published in 1960, at the time containing 4300 descriptors. It is now updated annually, and the newest version of the hierarchy, MeSH 2013, contains 26853 headings. However, the hierarchy tree of MeSH allows for the same heading to appear more than once, and therefore the tree contains 54095 entries in total. The hierarchy starts with 16 relatively broad categories such as "Anatomy" or "Organisms" and gets much more specific (e.g., "Olfactory Receptor Neurons" or "Locusta migratoria") in deeper hierarchical levels. However, due to the considerably smaller focus of MeSH compared to IPC, the main trees are already much more specific than the IPC section definitions (cf. Table 2). Table 4 shows the root terms of the MeSH main trees.

The MeSH terms are used in the biomedical literature database PubMed as a document indexing system, i.e., for annotating documents with relevant terms that describe their content. The PubMed database contains more than 22 million citations, 90% of which have MeSH annotations[14]. PubMed users can therefore restrict their search to documents that have been annotated with some very specific terms in which they are interested. On top of that, MeSH terms are used to automatically improve the recall of PubMed searches through query expansion: By mapping keywords from a search query to MeSH terms, relevant documents are included in the search results even if they only contain synonyms or hyponyms of the original keyword. As an example, Figure 7 shows how the system expands the simple query "heart attack". Searching with the expanded query returned 194,209 hits on August 6, 2013, of which only 1.5% contained the exact phrase "heart attack". That means that the search results for the initial user query would have been very lacking without

[14] http://nnlm.gov/training/resources/meshtri.pdf

Tree	Root term
A	Anatomy
B	Organisms
C	Diseases
D	Chemicals and Drugs
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment
F	Psychiatry and Psychology
G	Phenomena and Processes
H	Disciplines and Occupations
I	Anthropology, Education, Sociology and Social Phenomena
J	Technology, Industry, Agriculture
K	Humanities
L	Information Science
M	Named Groups
N	Health Care
V	Publication Characteristics
Z	Geographicals

Table 4: MeSH main trees. The root terms are more specific than the IPC section definitions.

the automated inclusion of MeSH information. As a result, even PubMed users who are completely unfamiliar with MeSH can benefit from the search improvements it makes possible.

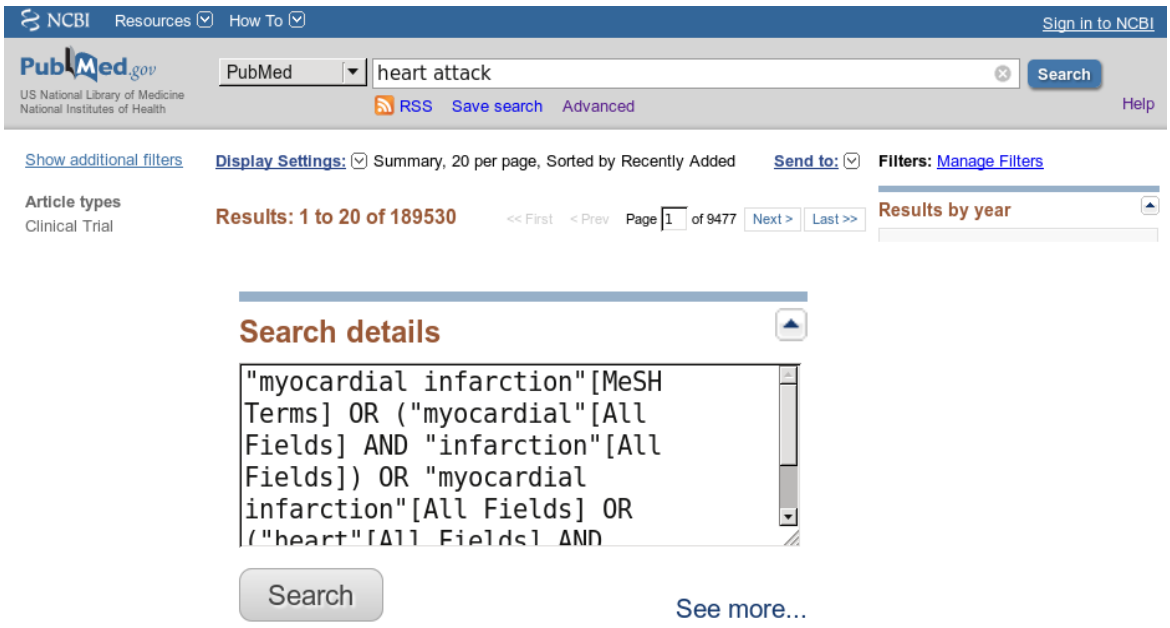


Figure 7: Automated query expansion on PubMed website. The unspecific search query "heart attack" is automatically expanded into the corresponding more comprehensive query.

Although the patent domain also lists manually assigned and therefore presumably highly accurate categories for each document in the form of IPC classes, a similar use of the system has not previously been investigated. We believe that there are two possible benefits to such an approach: First, since most scientists are unfamiliar with the intricacies of the patent domain, it is desirable to offer them an easier option to formulate patent queries that include classification information - as we described in Section 2.3, relying on keyword queries leads to insufficient results in most cases. Second, professional patent searchers have to invest a lot of time and effort into composing complete queries, and would therefore benefit greatly from a system that helps them with this complicated task. In order to provide such assistance, it is important to have a clear understanding of the properties of both classification systems. In this thesis, we therefore investigate differences between the IPC and the established MeSH hierarchy and their implications for patent search. As a solution to problems we discovered through our analysis, we propose two approaches: a system for the automated assignment of additional classes to patent documents and a guided patent search system that assists the user by offering query expansion suggestions derived from class co-occurrence data or using existing knowledge from external sources.

2.6 Related work

The importance of MeSH for the biomedical field has led to extensive research. Apart from the efforts to use the MeSH annotations to improve PubMed search functionality [39, 41, 42] that we described in Section 2.5, there are mature approaches for automatically assigning MeSH terms to documents, and MeSH terms are successfully used for query expansion. Trieschnigg *et al.* investigate multiple MeSH classification systems and report that a K-Nearest Neighbor approach clearly outperforms the other systems, but its classification speed is relatively low (around a second per abstract) and it is not well-suited for assigning rarely used MeSH terms [43]. For the same task, Tsatsaronis *et al.* present a fast system based on Maximum-Entropy classification that offers very high accuracy [44]; we will describe their system in more detail in Sections 4.1 and 4.2. MeSH has also been used in combination with patents, e.g., for tagging diseases [45].

IPC-related research is much more limited than for MeSH, but scientific interest has been growing over the last decade. As we described in Sections 2.3 and 2.4, there are a number of publications about the professional approach to patent search and classification information, many of which identify problems and suggest solutions [11,14,23,26,35–38]. However, there is no published

in-depth analysis of either hierarchy and no systematic comparison of both hierarchies, although there are some papers dealing with MeSH [46, 47] or IPC [25, 48–50] in a more general way.

2.6.1 Patent categorization

Giving the correct classification codes to patent applications is a very important task for the examiners working at patent offices. It is needed both internally and externally: Internally, it represents a prerequisite for assigning the application to an expert in its field. Krier *et al.* describe the process that was previously in place at the EPO for this task [51]: Three well-trained pre-classifiers (technical staff) were able to pre-classify all applications before more highly-qualified specialists made the exact classification decision. Increasing technological complexity has made it impossible for such a small group to accurately decide upon the pre-classification of today's wide variety of applications. To make matters worse, errors in this process are consuming a lot of time from highly-qualified patent examiners and are therefore very expensive for the patent office. Apart from these internal needs of patent offices, the correct classification of patent documents is also extremely important for external patent searchers. As described in Section 2.3, the addition of classification information to patent queries is often necessary for finding all relevant results. Of course, the effectiveness of using classification information for search relies on the correctness of the classification data. Since it is also a complicated and time-consuming task, the need for automation in the patent categorization process has been recognized since the late 1950s [52], and the recent growth in patent numbers (cf. Section 2.1) has made that need even more urgent.

Since patent offices suffer most from these problems, it is not surprising that most early research had direct connections to patent offices. Larkey *et al.* describe a tool developed for patent examiners at the USPTO at the end of the last century [53]. It was able to propose appropriate classes for a patent using k-nearest neighbor classification (i.e., based on the assigned classes of existing patents containing similar terminology). At the EPO, Krier *et al.* evaluated multiple (unnamed) commercial systems for patent categorization without giving any details about the methods that were used by those systems [51]. Their tests were carried out in the late 1990s and early 2000s and suggested that automatic pre-classification to a very high hierarchical level was already possible, but results for deeper levels were still far from satisfactory. At WIPO, the problems of patent categorization were investigated in the early 2000s as part of the *CLAIMS* project [54]. Fall *et al.* [55] tried classifying the patent documents from the *WIPO-alpha* test collection using multiple

different machine learning methods. The *WIPO-alpha* collection contains about 75,000 English-language patents in 114 classes and 451 subclasses. They reported categorization precision of up to 55% for the top prediction on the class level and up to 41% on the subclass level and concluded that support vector machines outperformed other algorithms such as Naive Bayes and k-nearest neighbor. For the same project, Fall *et al.* investigated the applicability of their results to German patents [56]. For a small corpus of almost 80,000 German-language documents and a set of 115 classes and 367 subclasses, they reported similar results as for the English corpus. Neither study evaluated classification precision below the subclass level.

In later years, different workshops such as the Japanese NTCIR [30] and more recently the CLEF-IP evaluation track [57] added patent categorization tasks. While the workshops NTCIR-5 [30] and NTCIR-6 [58] both had patent classification tasks concentrating on the Japanese F-term system (for an explanation see Section 2.4), later NTCIR workshops switched to IPC [59]. Li *et al.* participated in the F-term classification subtask [60], using support vector machines with varying parameters. They reported a surprising finding: Including information about the F-term hierarchy made the results considerably worse, although hierarchical information had led to improvements on other hierarchical classification tasks [61, 62]. The results from the published approaches vary depending on the hierarchical level that was used: Trappey *et al.* [63] report precision values slightly above 0.9 for a small subset of IPC subclasses and main groups, Tikk *et al.* [64] correctly identify up to 37% of main groups, and Verberne *et al.* [65] reach an F_1-score of 0.7 for the subclass level in their best run. To our knowledge, there is only one prior effort to classify patents down to the lowest level of the IPC: Chen *et al.* [66] use an elaborate three-phase approach to reach 36% accuracy for that difficult task. We will describe their method in detail in Chapter 4 where we introduce the method we used for the same task.

In addition to these efforts for reproducing existing class assignments, there are some approaches to patent categorization that were not based on existing systems: Lai *et al.* describe a method for the automatic construction of a classification system for patents from a very specific field (e.g., the semiconductor foundry industry) [67], and Loh *et al.* investigate the automatic assignment of patents to a new classification system based on the innovation methodology TRIZ [68].

2.6.2 Patent retrieval

Although it is our goal to assist patent searchers in finding the patents they are looking for, we are more focused on helping them formulate the queries than on the actual patent retrieval. However, since this is an active research topic that is closely connected to our goal, we will give a basic overview of a broad variety of existing approaches for this task.

While early systems for retrieving patent documents were already developed in the 1970s [69], the number of publications related to patent retrieval has grown considerably in recent years. The majority of the existing research into patent retrieval systems has concentrated on prior art search. The proposed systems often expect as input an existing patent or patent application and produce a ranked list of similar patents as output. This task was also a part of several patent-related workshops such as the patent retrieval task at NTCIR [70], the prior art candidates search task at CLEF-IP [71] and the prior art task at TREC-CHEM [72]. The systems differ greatly in the input they are expecting (e.g., complete patents, only the text fields, only the claims section, or even non-patent texts) and therefore also in the elements that are used in their retrieval methods (e.g., text only, citation information, classification information, information about inventors/assignees).

In an early approach to solving the prior art task, Osborn *et al.* use the SMART information retrieval system [73, 74] with additional shallow NLP techniques to find similar patents to a query patent [75]. A more advanced system is presented by Takaki *et al.*: Documents are divided into "subtopics" (i.e., thematically coherent text fragments), similar documents are retrieved for each subtopic, and the results are combined into one ranked list of documents [76]. Graf *et al.* generate prior art queries by contrasting the distributions of shared terms in related as well as unrelated pairs of documents [77], while Xue *et al.* evaluate a number of different features (e.g., patent field and weighting method) for extracting useful query words [78, 79]. Magdy and Jones examine three different strategies for automatic query expansion, but report no general improvement in the quality of the resulting ranking [80]. The same can be said for experiments on extending the commonly used "bag of words" approach by taking dependency triples into account [81, 82], and Becks *et al.* even report significant decreases in both precision and recall as a result of using linguistic phrases instead of terms [83]. Cluster-based language models were also shown to lead to marginal improvements at best [84].

Mase *et al.* only use the claims section of the query patent to extract query terms; the resulting

ranking is then combined from two different searches, a broad recall-oriented search on the full texts and a second narrower precision-oriented search just on the claims sections [20]. The same method was also reported to have a small positive effect on retrieval from a large set of patents from the chemistry domain by Gobeill *et al.*, but the same experiment revealed a larger improvement from analyzing the citations [85]. Citation analysis is also used by Fujii *et al.* in an approach that combines a text-based retrieval method with a citation score into a combined result ranking [86], and Lopez and Romary take this approach a step further by also including the citation texts [87,88]. The latter approach performed best among all competitors of the CLEF-IP 2010 Prior Art Patent Search Task, followed by a relatively simple system by Magdy and Jones that also combined citation analysis with standard information retrieval techniques [89, 90].

Szarvas *et al.* combine standard text-based document retrieval with IPC assignment information [91], and Mahdabi *et al.* investigate different approaches to creating a "query model" representing the patent in question, using different patent fields as well as IPC information [21]. The IPC assignments are also the basis of a new vector space model for patent retrieval proposed by Chen *et al.* [92, 93]. Instead of using document-term vectors as in most traditional approaches, document-category vectors are calculated based on the common vocabulary of patents with the same assigned class.

Bashir and Rauber argue that a high number of relevant patents are not retrieved with current systems [94]. To improve the retrievability of these patents, they propose expanding automatically generated queries using pseudo relevance feedback. The problem of differences in retrievability of patents was also examined by Bache [95], and it is shown that retrieval models taking term frequencies into account (in this case tf-idf and a version of Okapi-BM25) are much better than Boolean search in this context. However, Boolean search is still the basis of most patent search engines, and professional users consider Boolean operators the most important search functionality of a patent retrieval system [5, 6].

This chapter has introduced the patent domain by giving statistics on patent numbers and describing different search types and their requirements. The special properties of the patent texts as well as the different classification systems were explained with regard to their effect on

patent search. Additionally, we examined previous approaches that use document annotations to implement new and improved search functionality; many approaches concentrate on PubMed documents and the corresponding MeSH annotations. Therefore, we will compare MeSH and IPC and their respective annotations to PubMed abstracts and EPO patents in the following chapter.

3 Comparative analysis of MeSH and IPC

Summary

Both MeSH and IPC are hierarchical systems used for the annotation/classification of documents. The hierarchies are fairly similar in size and structure, although there are some differences in the relationship between terms and hierarchy entries.
The differences between the terms themselves are larger: IPC is focused on class codes while MeSH emphasizes terms, IPC definitions are longer, more complicated and less self-contained than MeSH headings, and are therefore much less likely to appear in the text.

The main differences between both systems concern their assignment to documents: The set of MeSH terms assigned to a single PubMed document is usually much larger and much more diverse than the set of IPC classes assigned to a patent. Consequently, there are limits to the current usage of classification information for patent search: The complexity of the system makes finding the relevant classes difficult, and the sparseness of class assignments means that relevant documents may be missed in the search due to missing class assignments.

Parts of this chapter were previously published in:

- Daniel Eisinger, Thomas Wächter, Markus Bundschus, Ulrich Wieneke, Michael Schroeder. Analysis of MeSH and IPC as a Prerequisite for Guided Patent Search. Bio-Ontologies 2012. http://bio-ontologies.knowledgeblog.org/346
- Daniel Eisinger, George Tsatsaronis, Markus Bundschus, Ulrich Wieneke, Michael Schroeder. Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed. Journal of Biomedical Semantics 2013, 4:S1. http://www.jbiomedsem.com/content/4/S1/S3

Our analysis of MeSH and IPC can be divided into three parts: The first two parts concern the respective hierarchies and terms of the systems themselves, while the third part examines their usage for document classification. We analyzed the latter by collecting classification information from all patent applications to the EPO between 1982 and 2005 (over one million) as well as the annotations to all PubMed documents published by early 2011 (over 20 million). Our analysis has the goal of assisting patent search; we are therefore less interested in the reasons for any discrepancies than in their implications for search. Table 5 summarizes some core results of our analysis, and the following sections give more detailed reports.

Property	MeSH	IPC
number of hierarchy entries	54095	69487
number of unique entries	26581	69487
number of main trees	16	8
number of hierarchy levels	13	14
average string length main labels/class definitions	18	50
string length longest main label/class definition	104	596
string length shortest main label/class definition	2	3
average number of synonyms	8	0
occurrence of class labels in text	frequent	very rare
average number of annotations per document	9	2
number of unique annotations	25646	56599
proportion of documents with multiple annotations	86%	53%
proportion of documents with related annotations (i.e., same hierarchy tree)	81%	46%

Table 5: Comparative analysis MeSH vs. IPC. The hierarchical structures are similar, but MeSH terms are shorter and more likely to occur in text. The number of MeSH annotations per document far surpasses the number of classes per patent.

3.1 Hierarchies

As noted in Sections 2.4.1 and 2.5.1 and shown in Table 5, the number of unique MeSH entries is considerably smaller than the number for IPC; it's less than half of the IPC's number of entries. However, the hierarchy tree of MeSH allows for the same heading to appear more than once, and therefore the tree contains 54095 entries in total (slightly above $\frac{3}{4}$ of the IPC number). Since this analysis is concerned with the hierarchical relation between entries, we decided to use the "tree view" of MeSH. In MeSH terminology, this corresponds to using the unique tree numbers of

the separate entries instead of the MeSH IDs that are identical for different tree positions of the same heading. This distinction between the number of hierarchy entries and unique entries already illustrates the main difference between the hierarchies: While each IPC class can only be directly subordinate to exactly one other entry, MeSH allows its entries to have more than one parent. In one instance, this has arguably caused an inconsistency in MeSH: In tree branch F ("Psychiatry and Psychology"), the heading "Ethics" (MeSH-ID D004989) is subordinate to "Morals" (D009014); however, in branch K ("Humanities"), their roles are reversed. This apparent contradiction may be caused by the ambiguous definition for "Ethics" that MeSH uses, since it describes both "The philosophy or code pertaining to what is ideal in human character and conduct" and "the field of study dealing with the principles of morality". "Morals" on the other hand is more clearly defined as "Standards of conduct that distinguish right from wrong".

In addition to allowing multiple parent nodes, MeSH has a second distinctive feature separating it from the IPC: Not all hierarchy entries can be mapped to MeSH IDs and vice versa. More precisely, there are MeSH headings with IDs that aren't a part of the hierarchy as well as hierarchy entries that do not have a corresponding MeSH ID. According to the MeSH Browser[15], the headings "male" (MeSH-ID D008297) and "female" (D005260) - not to be confused with headings "men" (D008571) and "women" (D014930) - are meant as tags for male or female organs, diseases, processes, etc. They have MeSH IDs, but no tree numbers - in other words, they do not have any parent or child nodes. On the other hand, the most general entries in the hierarchy, the "roots" of the different trees in MeSH, do not have MeSH IDs and can therefore not be found in the MeSH Browser. They are however listed as tree tops in the MeSH tree structure navigation and were therefore included in the analysis. In this regard, the IPC hierarchy is simpler than MeSH, since each IPC entry has exactly one fixed position inside the hierarchy.

Contrary to these conceptual differences, the structural comparison of the hierarchies did not reveal any significant differences. As Table 5 shows, their sizes are in the same range (about 70,000 IPC classes and 54,000 entries in the MeSH tree) and they have almost the same depth (14 levels for IPC, 13 for MeSH). Figure 8 shows that the distributions of nodes over different levels of both hierarchies are also similar, with the larger IPC having a higher concentration of nodes in levels 4 to 7 of the hierarchy and MeSH having a slightly more even distribution of nodes among the different hierarchy levels.

[15] http://www.nlm.nih.gov/mesh/2012/mesh_browser/MBrowser.html

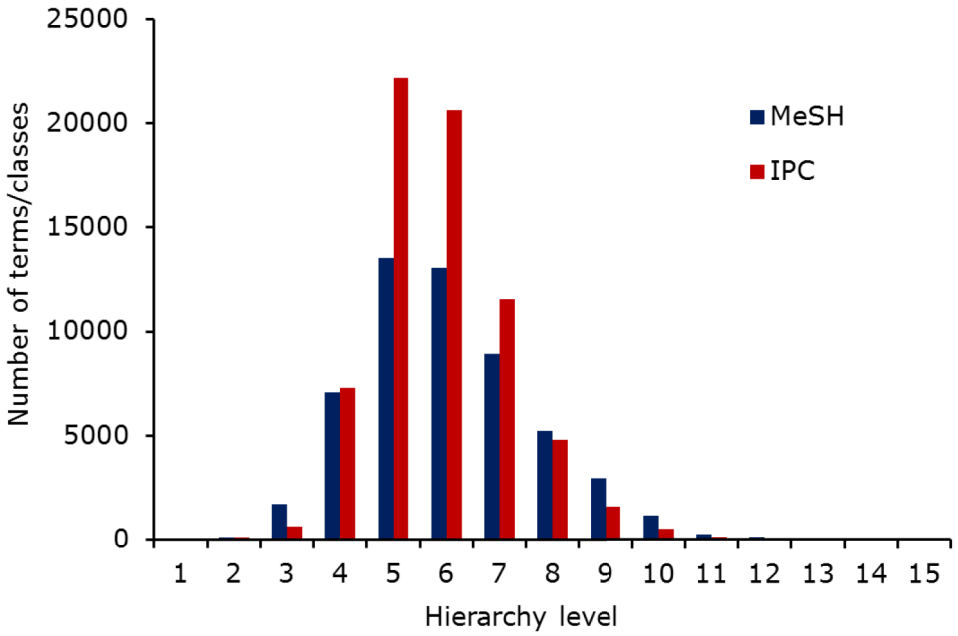


Figure 8: IPC vs. MeSH - Terms/classes per hierarchy level. Both hierarchies expand in similar ways.

Closely related to the findings depicted in Figure 8, Figure 9 shows the average number of child nodes per hierarchy level and illustrates some slight differences in the node distributions: While IPC and MeSH grow mostly in parallel after the fourth hierarchy level, there are large differences in the first four levels. The IPC expands especially fast from the first and third hierarchy levels where the average IPC class has 16.1 and 11.5 child nodes respectively while the growth of MeSH is more modest with 6.5 and 4.2 child nodes per heading. In the second hierarchy level however, the expansion of MeSH with 14.6 children on average is much more significant than that of the IPC with just 5.0. By level 4, the growth of both hierarchies has slowed down considerably, but IPC is again growing significantly faster than MeSH. On the other hand, it can also be seen from Figures 8 and 9 that both IPC and MeSH expand between levels 1 and 5 and start contracting after that.

As mentioned before, the hierarchical structures are remarkably similar overall apart from these minor differences. The next section investigates the terms of both systems.

3.2 Terms

Unlike the very similar structures, the comparison of the terms shows some major differences. The first difference between MeSH and IPC is their emphasis on terms/concepts versus identifiers: While

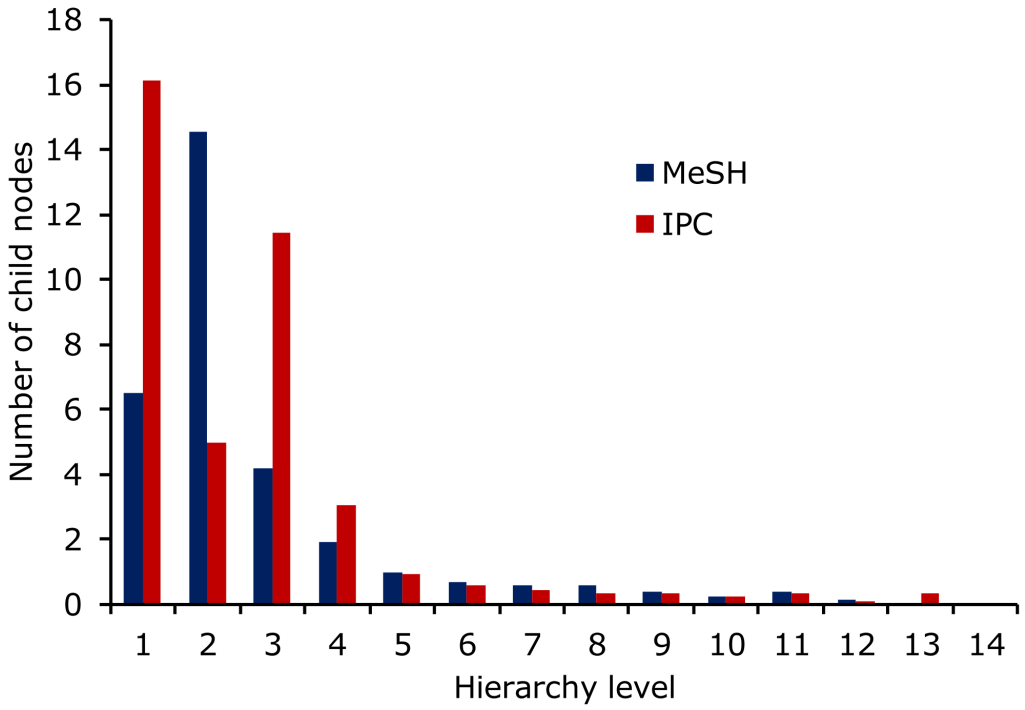


Figure 9: IPC vs. MeSH - Nodes and children per hierarchy level. The expansion of both systems differs considerably in the first three levels.

each heading has its own MeSH ID, the emphasis is clearly on the term itself - as is evidenced by the commonly used phrase "MeSH terms". The IPC on the other hand is first and foremost a collection of alphanumeric codes which are signifying their place in the hierarchy instead of their semantic meaning. Unlike MeSH terms, these codes do not give an uninformed user any useful information about the patents that should be assigned to this class. This information is instead contained in additional class definitions that are more akin to MeSH's scope notes. The class definitions are therefore necessary to understand what kind of invention is assigned a specific code. As an example, looking up the MeSH headings for a document about insects on PubMed will lead the user to the term "Insects", not its identifier "D007313". However, a patent about an immunoassay is assigned to class "G01N 33/53" (or one of its subclasses), and the definition of this class has to be checked separately if the user does not know it.

The examples of codes and corresponding definitions in Section 2.4.1 show an additional difference: While each MeSH entry constitutes a self-contained phrase (often containing their hierarchical predecessor in part or even in whole), IPC definitions below the subclass level can often only be

understood by considering their hierarchical ancestors. For example, a typical branch of the MeSH tree contains the descending entry sequence "Neoplasms" - "Neoplasms by site" - "Breast Neoplasms" - "Breast Neoplasms, Male"; in the IPC however, dependent sequences like "Wet end of machines for making continuous webs of paper" - "Wire-cloths" - "Seams thereof" - "sewn" are much more common.

The described differences in the style of formulating class definitions might lead to the assumption that MeSH definitions will on average be longer than IPC definitions; after all, a MeSH heading must stand on its own while IPC class definitions may depend on their ancestor nodes. As Table 5 shows however, the opposite is true: the average length of an IPC definition exceeds 50 characters, while the average MeSH heading contains less than 18 characters. As an example, the longest MeSH term is "3-Pyridinecarboxylic acid, 1,4-dihydro-2,6-dimethyl-5-nitro-4-(2-(trifluoromethyl)phenyl)-, Methyl ester", while the longest definition of an IPC class is shown in Figure 10. This example shows that IPC definitions can include figures (of chemical structures usually) to clarify the content of their classes; MeSH terms however are text-only. As an aside, the shortest IPC definition ("Tin" as a subclass of a class about biocides containing metal atoms) is also longer than the shortest MeSH term ("Id", i.e., "The part of the personality structure which harbors the unconscious instinctive desires and strivings of the individual").

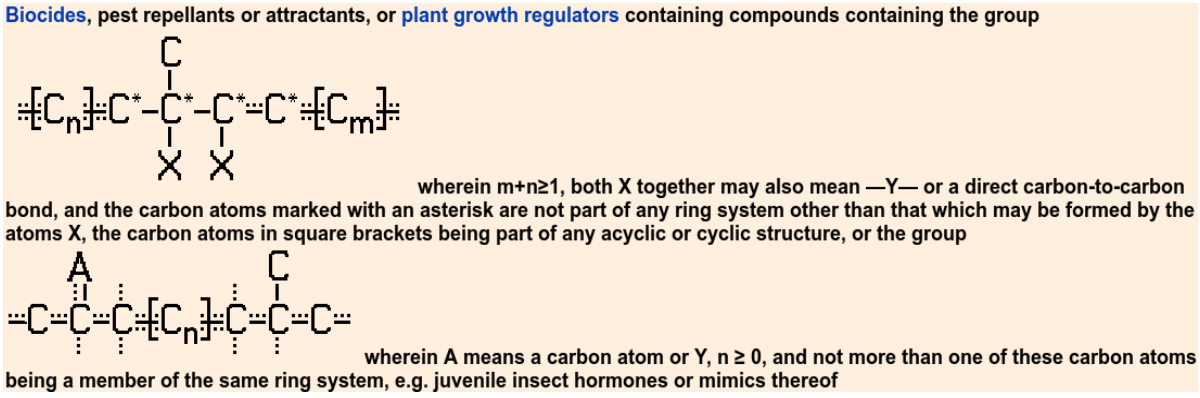


Figure 10: Longest IPC class definition (for class A01N 49/00)

In addition to being longer and less self-contained than MeSH terms, most of the IPC definitions are also considerably more abstract and complicated. Unlike MeSH, the IPC also does not include any synonyms for the class definitions. All of these differences contribute to a very low probability of occurrence of class labels in text, while MeSH terms occur much more frequently. Shah *et al.*

report for a small corpus of 104 *Nature Genetics* articles that of the terms that were attached to the documents by the indexers at the National Library of Medicine (NLM), 72% were found in the article on average [96]. However, this result is based solely on single-word MeSH terms and might therefore be considered overly optimistic. Schuemie *et al.* on the other hand search a larger corpus for all assigned MeSH terms and report finding 62% on average after including children of the headings [97]. In order to quantify how rarely IPC codes occur in text, we searched the complete texts of all 14,600 documents from our test corpus C_{73} (for a definition see Section 4.3) for their class definitions. Less than 2% of the documents contained their respective definition, and most of these hits were for class names that weren't informative on their own (e.g., class C07K 14/47, "from mammals").

As a consequence, IPC classes cannot be assigned to patents by simply extracting them from text. This is one of the main reasons for the much more extensive use of automated (pre-)annotation of PubMed documents compared to patents. One possible approach to solving this problem is the assignment of classes using machine learning methods, i.e., training a classifier on existing classification data to predict assignments for new data. We will investigate this possibility in Chapter 4.

3.3 Usage for document classification

IPC and MeSH are both used as classification/annotation systems for documents: all EPO patent applications are assigned at least one IPC code, and all PubMed articles from participating journals are annotated with appropriate MeSH terms. In order to analyze the assignment of hierarchy entries to documents, we collected classification information from the two corpora we introduced at the beginning of this chapter. These corpora contain about 1.06 million EPO patent applications and over 20 million PubMed documents in total. Figure 11 shows that there are large differences between the numbers of MeSH annotations per document and the numbers of IPC annotations per patent.

For the EPO patents, 45.8% were assigned only one class, 31.3% were assigned two, 13.4% were assigned three and 5.1% were assigned four. This means that more than 95% of the patents had four class assignments or less, leading to an average of just below two classes per patent. The situation is much different for the PubMed documents. The proportion of documents with one annotation is 4.3%, and 3.5%, 3.4% and 4.3% respectively for two, three and four annotations.

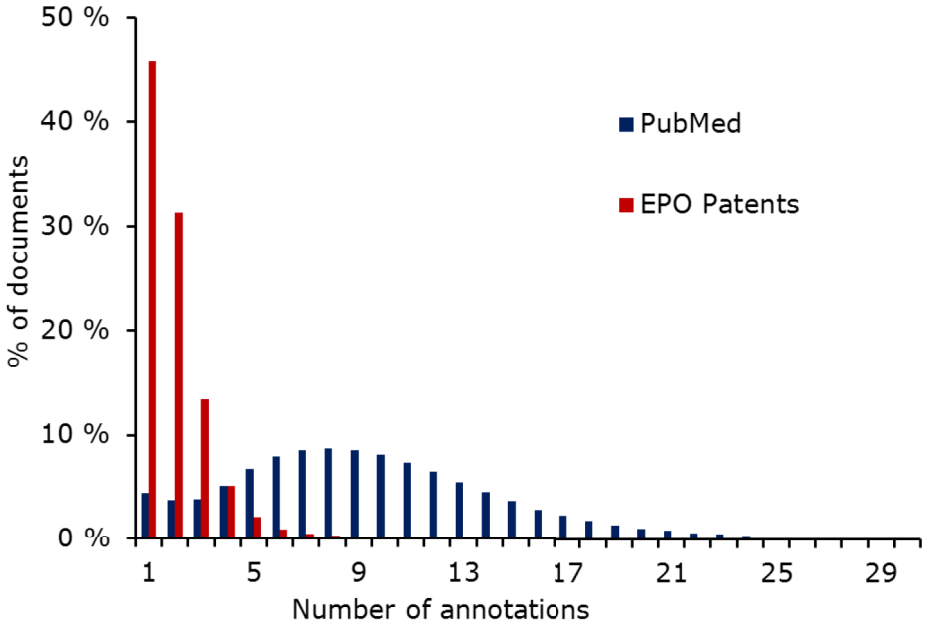


Figure 11: Percentage of documents with number of annotations. The average number of MeSH annotations per PubMed document is much higher than the number of IPC classes per patent.

Adding up these percentages reveals that less than 16% of the PubMed documents have four annotations or less, which is a huge difference compared to the 95% we reported for the patents. While the set of patents with one assigned class was the largest by far, the PubMed documents are much more evenly spread. The largest set is the one with nine annotations in this case with 7.6%, but the sets with seven, eight and ten annotations have very similar sizes with 7.2%, 7.5% and 7.4% respectively. Averaging over the whole corpus then results in less than two assigned IPC classes for the patents in our corpus, while PubMed documents have almost nine MeSH terms (cf. also Table 5).

The following tables show one example respectively for the set of annotations to a patent and a PubMed document. The annotations presented in Table 6 are from the document with PubMed identifier 14771478 and the title "Little known tropical diseases", and Table 7 shows the IPC classes assigned to EPO patent 1481970, "Novel Herbicides". Since IPC class codes do not directly explain the meaning of the class, Table 7 also gives abbreviated definitions of the superordinate classes. Both documents represent the highest number of annotations that were assigned to any document in the respective corpus.

Abdomen	Acne Vulgaris	Acrodermatitis	Actinomycosis
Alcaligenes faecalis	Alopecia	Amebiasis	Anal Canal
Ants	Appendicitis	Areca	Arthritis
Aspergillosis	Axilla	Blastomycosis	Bronchi
Buttocks	Candidiasis	Colitis	Colon, Sigmoid
Cough	Cryptococcosis	Culicidae	Cysticercosis
Darier Disease	Deficiency Diseases	Dermatitis	Dermatitis, Seborrheic
Diarrhea	Dysentery, Bacillary	Ear	Edema
Emetine	Erythema	Extremities	Eyelids
Face	Fever	Fibroma	Furunculosis
Gingiva	Hair	Halitosis	Hand
Hookworm Infections	Insects	Intertrigo	Intestines
Keloid	Keratosis	Keratosis, Seborrheic	Lactose
Leg	Lichens	Lip	Lipodystrophy
Lipoma	Liver	Malaria	Microsporum
Mite Infestations	Mycetoma	Mycoses	Myiasis
Nails	Nose	Oculomotor Muscles	Palatine Tonsil
Pemphigus	Phlebitis	Phlebotomus Fever	Pigmentation
Pityriasis	Pruritus	Pyoderma	Salmonella Infections
Scalp	Skin	Skin Diseases	Spermatic Cord
Spirochaetales Infections	Splenomegaly	Starvation	Stomatitis
Sweat	Sweat Glands	Syphilis	Tattooing
Tinea	Toes	Tongue	Toxoplasmosis
Tropical Medicine	Ulcer	Ultraviolet Rays	Umbilicus
Urethra	Urine	Vagina	Vibration
Wind	Wounds and Injuries	Xanthium	

Table 6: MeSH term annotations for PubMed document 14771478, "Little known tropical diseases".

A comparison of the sets of annotation terms shown in Tables 6 and 7 shows that the PubMed annotations for the example document are extremely diverse, with covered MeSH trees including "Anatomy" (e.g., "Hand"), "Diseases" (e.g., "Malaria"), "Organisms" (e.g., "Ants"), "Phenomena and Processes" ("Wind") and others. On the other hand, the patent's IPC classes shown in Table 7 cover only two IPC sections (Section A, "Human necessities", and Section C, "Chemistry; metallurgy"). Even among these sections, they are limited to only three subclasses in total, namely $A01N$, $C07C$ and $C07D$ - as mentioned above, the table also contains abbreviated definitions of these subclasses to show the close relations between these assignments compared to the situation for the PubMed document.

These observations led us to investigate the diversity of the different assignments to a single document. This property does not have a straightforward measure, but the distances between

A01N 35/06 A01N 43/40	A01N 43/16 A01N 43/42	A01N 43/18 A01N 43/72	**Biocides; pest repellants**
C07C 49/747	C07C 49/753	C07C 69/708	**Ketones/esters of carboxylic acids**
C07D 207/38 C07D 231/32 C07D 237/04 C07D 303/32 C07D 309/32 C07D 333/32 C07D 403/12 C07D 417/12	C07D 209/96 C07D 231/34 C07D 265/02 C07D 307/60 C07D 309/38 C07D 333/50 C07D 409/12	C07D 211/86 C07D 231/36 C07D 279/06 C07D 307/94 C07D 311/96 C07D 335/02 C07D 413/12	**Heterocyclic compounds**

Table 7: IPC classes assigned to EPO patent 1481970, "Novel Herbicides".

simultaneous annotations in the respective hierarchies can give an indication of the differences between both corpora. We used all documents (PubMed and patents) with more than one annotation to calculate maximum and minimum distances for each annotation set.

More formally: Given a hierarchy H (in our case either MeSH or IPC) and two entries a and b of the hierarchy, we define the distance between a and b as the length of the shortest path between them in H. For a subset A of H consisting of all annotations to a single document, we then define the maximum (minimum) annotation distance as the maximum (minimum) over the pairwise distances of elements of A. Since both MeSH and IPC are organized as forests (i.e., unions of trees) instead of singular trees, we inserted one artificial root node into each hierarchy in order to represent them as trees.

Figure 12 shows one example each for a PubMed document and a patent. Among other terms, the PubMed document is annotated with "Cyanides" from the "Chemicals and Drugs" tree as well as with "Risk Factors" from the "Health Care" tree. These terms show multiple aspects of the document, and their distance in MeSH is 14. On the other hand, the most distant IPC classes assigned to the patent have the definitions "Chemical Analysis" and "[Immunoassay] using isolate of tissue [...]". These classes are directly related, and the shortest path connecting them in the IPC hierarchy has length 3. This means that the annotations to this patent only cover one aspect or at best a very narrow range of aspects. Additionally, some of the assignments may be considered redundant since they are direct ancestors of another assigned class and therefore implicitly covered.

As these examples show, the distance between terms hints at whether they belong to the same main tree of the hierarchy; this is an important property of the annotation terms that we will

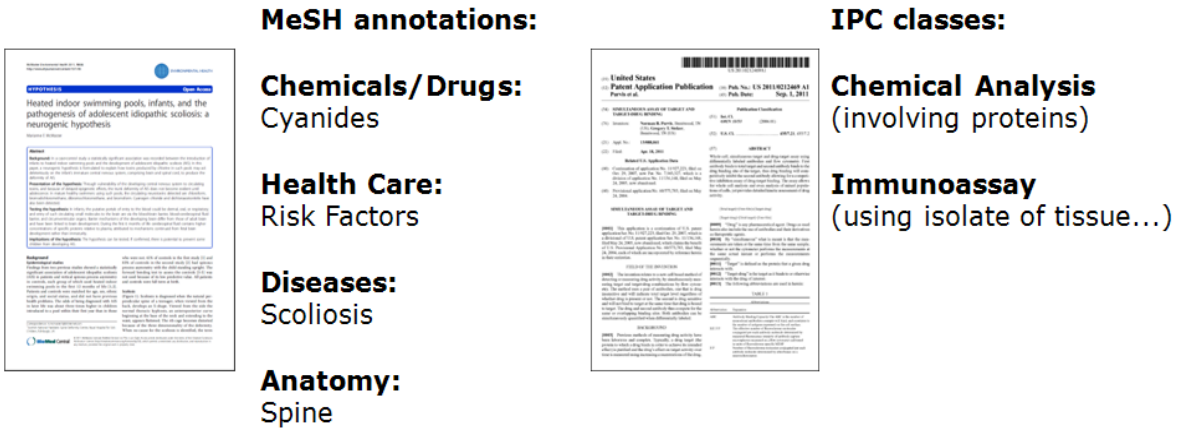


Figure 12: Annotation sets for example documents (left: PubMed, right: patent). The MeSH terms for the PubMed document are very diverse and cover multiple aspects, while the IPC classes are closely related.

examine in more detail below. However, the distance gives more information than that since it also allows us to estimate the diversity of annotations from the same tree. Figures 13 and 14 show the minimum and maximum differences for PubMed as well as our patent corpus.

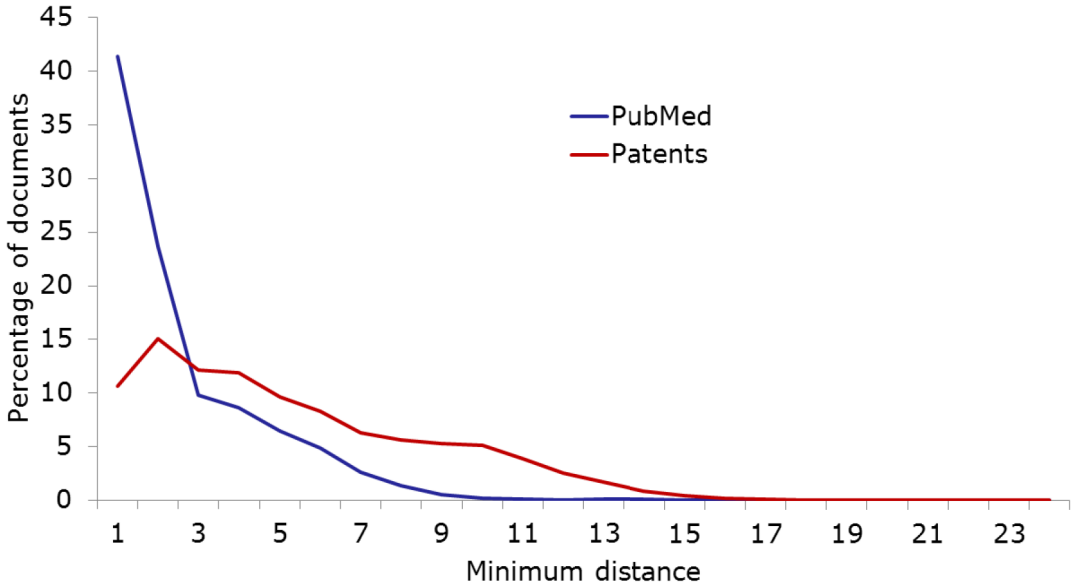


Figure 13: Minimum hierarchical distances of multiple annotations assigned to the same document. PubMed documents have more very closely related annotations than patents.

As Figure 13 shows, many PubMed documents are annotated with very similar MeSH headings, in many cases (around 41.4%) even pairing a term with its direct parent (i.e., annotations have

distance 1). Arguably, this means that there is more redundancy contained in the MeSH annotations, since the parent term is implicitly assigned together with the child term. This situation is less common in our patent corpus (only 10.7%), although there are also many patents that are annotated with closely related classes. A more detailed analysis of this property as well as some other relationships between annotations follows in Figure 15.

The analysis of maximum distances (Figure 14) has the opposite result: While the maximum distances for patents do not differ too much from the minimum distances, the maximum distances for PubMed documents are much larger. The higher number of PubMed annotations is certain to play an important role in these differences. However, the extent of the differences between both corpora is surprisingly large. This result indicates that PubMed annotations cover a considerably broader spectrum of aspects than the assigned patent classes. This means that in addition to having a fairly small number of class assignments per patent, these assignments are also much more closely related than the PubMed/MeSH ones.

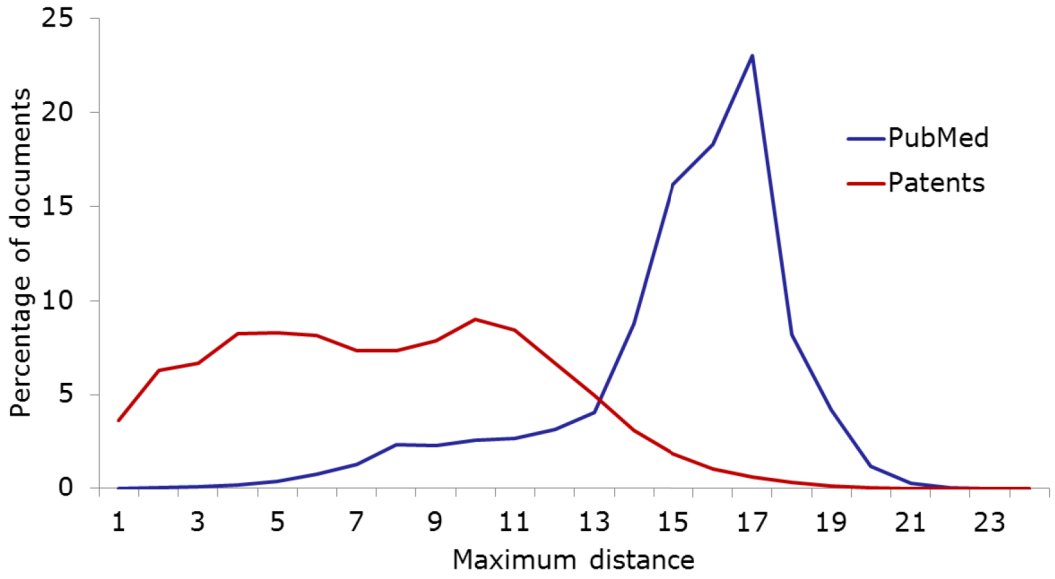


Figure 14: Maximum hierarchical distances of multiple annotations assigned to the same document. The maximum distance is considerably larger for PubMed documents than for patents, as is the difference between maximum and minimum distances.

In addition to the path lengths between annotations, we examined the relations between annotations more directly by checking how often terms were co-assigned with closely related terms. We again used the annotation sets of all documents with multiple annotations, and tested them

for the inclusion of at least one pair of nodes that belong to the same hierarchical tree or share an ancestor/descendant, parent/child or sibling relationship. (Documents with parent/child annotations are of course a subset of those with ancestor/descendant annotations.) Figure 15 shows the percentage of documents (among those with multiple annotations) that were assigned pairs of annotations that share any of these relationships. For both PubMed and the patent corpus, a high percentage of the documents with multiple assignments contains pairs of annotations from the same hierarchy tree. While this is not a problem for PubMed due to the high overall number of annotations, the much lower number for patents may be a cause for concern: Including patents with just one annotation, over 83% of all patents are classified into only one of the eight main sections of IPC. Since the main trees correspond to extremely general domains such as "Human necessities", we believe that some aspects of many patents are not covered by the currently assigned classes.

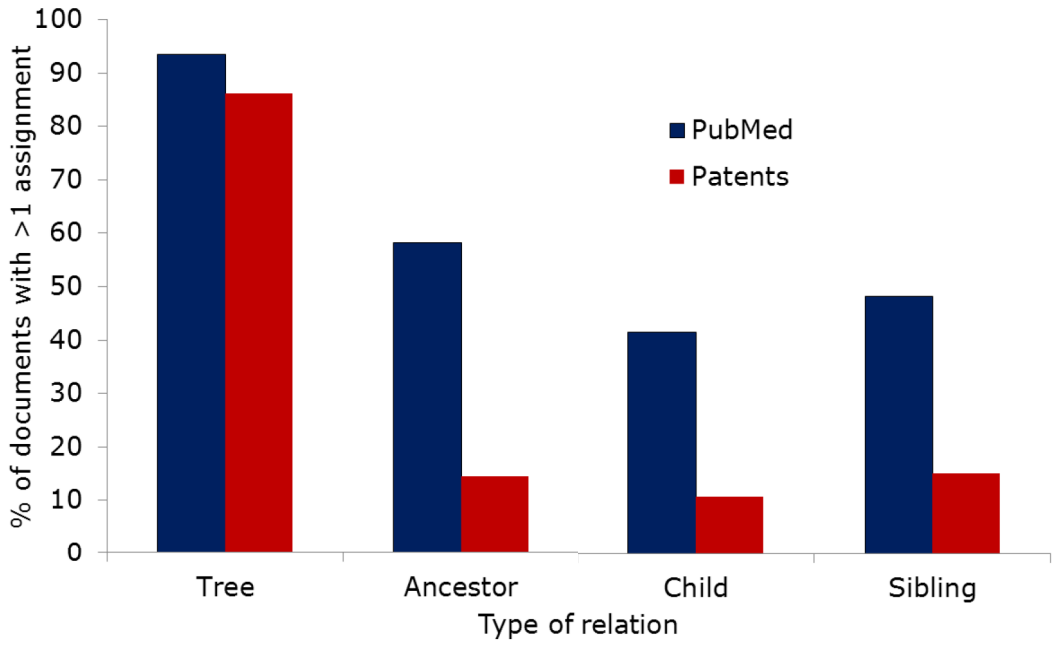


Figure 15: Hierarchical relationships of terms assigned to the same document. PubMed documents have considerably more closely related annotations, but the percentage of documents with annotations from the same main tree is almost equal for patents.

On the other hand, the proportion of documents with directly related annotations (parents/ancestors or siblings) is much lower in the patent corpus than in PubMed. It makes sense that direct ancestors usually aren't assigned to the same document since the general annotations are already implicitly contained in the more specific ones. If the search engine in use features the

possibility to automatically expand general queries by including all subordinate terms of query terms, it is not necessary to include any annotation except the most specific one in order to avoid redundant annotations. More formally: Let A be the set of terms that was assigned to a specific document, and let B be a set of terms that corresponds to the nodes in a descending path in the hierarchy, i.e., a path in which the distance between the root of the tree and the current node grows from node to node. If A represents a subset of B, A can be reduced to its most specific element, i.e., $\{a^* \in A | dist(root, a^*) \geq dist(root, a) \forall a \in A\}$. Provided that the search engine uses the query expansion strategy described above, this does not cause any loss in recall: If a document was found by searching the original set of annotations (i.e., A), that means that the search query contains at least one of the terms in A. Let a' be that term. The definition of a^* implies that $dist(root, a^*) \geq dist(root, a')$. Since $A \subseteq B$, a^* is a direct descendant of a', and therefore the search engine will also retrieve the document in question if all other annotations are removed.

However, unlike the missing parent/child and ancestor/descendant pairs among the patent annotations, the low number of IPC sibling annotations (both in absolute numbers and in comparison to MeSH) may point to a potential problem: If there are more relevant sibling classes than there are actual class assignments for many patents, the search for exact class codes will have missing results. The same problem applies if the low number of sibling assignments is due to them being replaced by their common parent. It is therefore advisable for a patent search engine to allow for easy addition of relevant sibling classes.

3.4 Problems for IPC-based search

It could be argued that some of the described discrepancies are likely caused by differences in classification guidelines between patent offices and PubMed and may therefore be intended. However, that does not change the fact that both IPC and MeSH are used to improve search results on document corpora, and we believe that the use of IPC for that purpose comes with two serious disadvantages:

- Complexity of the classification system:

  The complexity of the IPC terms causes significant problems for non-professional patent searchers. It is very difficult to find the complete set of IPC classes that are relevant for the search task at hand, and it is even more difficult to combine them in a way that max-

imizes recall while still limiting the corresponding drop in precision to a manageable level. It is therefore not surprising that classification-based patent search is mostly performed by experienced professionals.

On the other hand, while the exclusive use of keywords for searching patent databases is more intuitive for less experienced users, it often leads to bad results due to the complicated language used in patents.

- Sparse class assignments:

  The low number of class assignments indicates that relevant aspects of many patents are not covered by the existing class assignments, and the problem is made worse by the relatively close relation of many co-assigned classes. Since the classification information is often used to filter the patent search results, the recall of these patent searches may be lower than expected.

Given these disadvantages, patent search engines should offer additional functionality for helping the user find the required results. Since the class definitions are needed to understand the meaning of the class codes, the system must include easy access to them. Additionally, since many definitions depend on their ancestor classes, the engine should give the user an easy overview over the relevant parts of the hierarchy. Unfortunately, many popular existing engines such as Espacenet[16], Google Patents[17] and FreePatentsOnline[18] do not display this basic information on the same page as the patent. In addition to that, patent search engines generally don't include any functionality to alert the user to additional relevant search terms and classes, and in many cases don't even offer auto-completion suggestions to the user when they start typing their query.

In this chapter, we analyzed the IPC hierarchy and compared it to MeSH, the prime example for a large-scale annotation system that is used to improve document search. While the hierarchies were discovered to be fairly similar, there were some differences in the terms. One of the main differences we found in the way that documents are annotated concerns the number of annotations per document: While PubMed documents were assigned almost nine MeSH terms on average,

[16] http://worldwide.espacenet.com/
[17] http://www.google.com/?tbm=pts
[18] http://www.freepatentsonline.com/

patents had less than two classes. Since this sparseness of class assignments may lead to low recall in patent searches, we will investigate a method to automatically assign additional classes in the following chapter.

4 Assigning additional classes to patents

Summary

In an effort to combat the problem of low class assignment numbers for patents, we investigate a method to automatically assign additional classes. Existing approaches for this task were with one exception not intended for or evaluated on the whole classification hierarchy. The single existing approach for this task has only been used on a smaller scale and is believed to be too computationally expensive for large corpora. Our work therefore represents the first time that patent categorization has been evaluated on a large corpus for all hierarchy levels.
Our method is based on training a series of Maximum Entropy-classifiers (one for each class) on existing class assignments and applying them to each document that is supposed to get additional class assignments. It has been used successfully for assigning MeSH terms to PubMed documents.

We first evaluate our method's ability to recreate existing class assignments. With precision and recall values close to 90%, the performance of individual classifiers is satisfactory, making our approach feasible for the task of finding additional documents for specific classes. However, since we intend to use a large number of classifiers on each document, these values still lead to large numbers of incorrect assignments. We propose the combination of our categorization results with keyword search as well as the use of class co-occurrence information for filtering search results to overcome this problem.
Our evaluation of the classifier features with the highest and lowest weights (i.e., the words that are most important for the classifier suggestion) shows that the chosen positive features are useful for making classification decisions while the negative features may be too unspecific.

Parts of this chapter were previously published in:

- Daniel Eisinger, Thomas Wächter, Markus Bundschus, Ulrich Wieneke, Michael Schroeder. Analysis of MeSH and IPC as a Prerequisite for Guided Patent Search. Bio-Ontologies 2012. http://bio-ontologies.knowledgeblog.org/346
- Daniel Eisinger, George Tsatsaronis, Markus Bundschus, Ulrich Wieneke, Michael Schroeder. Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed. Journal of Biomedical Semantics 2013, 4:S1. http://www.jbiomedsem.com/content/4/S1/S3

In Chapter 3, we analyzed the IPC classes assigned to our corpus of EPO patents. We showed that the average number of class assignments is very low compared to PubMed, and that many of the existing assignments are closely related in addition to that. We hypothesized that the low number of class assignments combined with the close relations between annotations results in recall problems for patent search, especially since classification-based search is the most important way to make up for the problems of keyword search on patents (cf. Sections 2.3 and 2.4). The most straightforward way of dealing with this problem would be the assignment of additional classes, but due to the high number of patents as well as the high complexity of the classification system, this can only be done automatically. Depending on the accuracy of the automatic assignment of relevant classes, the method can be useful for two related but different ways of dealing with the low number of assigned patent classes:

1. Given a class, find documents for this class.

   If the user knows that a particular class is highly relevant for their search, the automatic class assignments can be used to discover additional patents that should have been assigned to the class. The recall of the search can therefore be improved considerably.

2. Given a document, find classes for this document.

   If the user has already collected a small set of relevant documents, the automatically assigned classes for these documents can help them find the classes that are related to these documents, even if there is no classification data available or if there are missing assignments. These additional classes again enable them to refine their initial search query.

We will give an overview of previous approaches to the problem of automated patent class assignment in the following section before detailing the method we used and discussing our results in the rest of this chapter.

4.1 Previous approaches

As we described in Section 2.6.1, there have been multiple published approaches for the assignment of IPC codes to patents. These approaches are usually restricted to higher levels of the hierarchy such as the class or subclass level [63,65] or the main group level [63,64]. The WIPO has also made a patent categorization tool available, offering users the possibility to have documents categorized

to any of these levels[19]. To our knowledge, there is only one prior effort to classify patents down to the lowest level of the IPC [66]. This approach by Chen and Chang is based on a three-stage method for categorization that was designed to find the most fitting subgroups by reducing the complexity of this task stepwise: First, the best 11 subclasses are identified using a pre-learned Support Vector Machine (SVM). This step is intended to reduce the complexity drastically by excluding large parts of the hierarchy without strongly impacting the accuracy. In the second phase, a new SVM is learned from the documents below the subclasses chosen in the first phase. This new classifier is then used to predict the 37 most appropriate subgroups for the document that is supposed to be categorized. (The best values for the different parameters of the method were chosen empirically.) The chosen subgroups are then further subdivided in the third phase using the K-means clustering algorithm, and the final decision for one subgroup is made using the K-Nearest Neighbor algorithm on the document and the subgroup clusters. As a result of this rather complicated process, 36% of patent documents were categorized into their original subgroup. This is compared to 20% for the same method without the last stage, and 30% for the *HITEC* algorithm that also takes the hierarchy into account and had previously only been used on the main group level [64].

While all previous approaches were solely focused on automatically recreating the existing assignments, it is our goal to find additional relevant classes that the patent was not originally assigned to. We therefore believe that the method proposed by Chen and Chang is not ideal for our purpose, since it already removes large parts of the hierarchy in the first step. While this may make sense for retrieving the one main class that was previously assigned to the patent, it is too restricting for our goal of finding new relevant but potentially very different classes that were not previously assigned. Additionally, the need to train large numbers of new SVM classifiers in the second phase and to apply a clustering algorithm as well as another round of classification in the third phase, makes the method unsuitable for our intended application on a large corpus of documents.

We therefore based our system on the approach that has been used successfully for the automated assignment of MeSH terms to PubMed documents by Tsatsaronis *et al.* [44]. The Maximum Entropy approach is used to train one binary classifier for each MeSH term based on four feature types:

[19] https://www3.wipo.int/ipccat/

- lexical tokens from the document title
- lexical tokens from the abstract
- the name of the journal the document was published in
- the year of publication.

After the classifiers have been trained, each learned model is applied to all documents that are supposed to be categorized. For each classifier that puts the document into the positive category with high confidence, the corresponding MeSH term is added to the document's annotations. After all classifiers have been applied to all documents, a set of MeSH term annotations is retrieved for each document. Tsatsaronis *et al.* evaluate the method on a MeSH subset consisting of four of the 16 main trees and report very good results with an F_1-measure of 92.4%. The results are compared to the Naive Bayes method as well as the use of Decision Trees, both of which turn out to be clearly worse. We introduce the method in more detail in the following section.

4.2 Maximum Entropy

The Maximum Entropy approach estimates a probability distribution from existing data, based on the assumption that the distribution should be "as uniform as possible" if no external knowledge is available. This principle also gave the approach its name: *Entropy* measures the uncertainty of the outcome of a random variable, and its value is *maximized* if the random variable is uniformly distributed. Intuitively, this can be seen through the example of a coin toss: Its uncertainty is largest for a fair coin, since the probability of guessing the outcome of the next toss correctly is always 50%. However, if the coin is known to have a higher probability to show heads, the probability of guessing the next toss correctly (by always guessing heads) increases. Maximum Entropy has been used for various tasks in Natural Language Processing (e.g., language modeling [98] and part-of-speech tagging [99]) since the mid-nineties and was first proposed for text classification in 1999 by Nigam *et al.* [100]. It is closely related to the class of Expectation-Maximization (EM) algorithms that was named and formally introduced in 1977 by Dempster *et al.* [101] after the approach had already been used earlier for special cases (e.g., [102, 103]). The EM algorithm is introduced as "a general approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data", and its name stems from the division of each

iteration into an expectation step and a maximization step that are mutually influencing each other. In the expectation step, the expected value of the likelihood function is calculated using the current parameter value estimations, and in the maximization step, these parameter values are updated in order to maximize the likelihood function from the first step.

For the purpose of text classification, the existing data is represented by documents that have been labeled with certain categories (the *training set*), and the probability distribution that is estimated by the approach is used to assign classes to new documents (the *test set*). In order to do that, features are extracted from the training set. A feature is a measurable property of the documents, e.g., the number of occurrences of a certain word in the text or the year in which the document was published. For estimating the probability distribution, each feature f_i is assigned a parameter λ_i with initial value 0. Based on the relationship between feature values and class assignments in the training documents, these parameters are then updated iteratively until they converge. The result is a probability distribution based on the chosen features weighted by the corresponding parameters. Classes can then be assigned to a new document according to the classification scores that are calculated by combining the document's feature values with these parameters.

While Tsatsaronis *et al.* reported performance improvements from the use of different feature types, two of the feature types that were used for the PubMed documents are more problematic in the patent world: There is no directly equivalent entity to the journal of publication in the patent domain, and the year of publication can be ambiguous due to the existence of priority patents and different versions of the same patent. Consequently, we decided to restrict ourselves to the feature type that we consider most reliable for our patent corpora, the lexical features.

MaxEnt can be used for binary classification (i.e., one of two classes is assigned) as well as multi-class classification (one of multiple classes). Since our goal is multi-label classification (i.e., the relevant subset of all classes should be assigned), we trained one binary classifier for each class and applied all classifiers to each document.

4.3 Training corpora

In order to evaluate the results of our categorization efforts, we constructed training corpora from the EPO dataset that was also the basis of our previous analysis. We used three parameters to choose the sets of classes and documents that we based these corpora on:

- number of patents

  Since the Maximum Entropy method improves with a growing number of training documents [44], it is reasonable to restrict the categorization task to classes that were used to annotate some minimum number of patents. We therefore excluded classes that were assigned to fewer patents.

- text length

  While the EPO dataset contains the bibliographic data to all European patents, it doesn't always include the complete texts. Since the classifiers rely on the text, we only used documents that surpassed a certain minimum text length.

- only primary classification/also secondary classification

  Of the classes assigned to a patent, one is always emphasized as the primary one; i.e., the one that is supposed to correspond to the central aspect of the patent. When choosing training documents for a class, it might therefore be advisable to concentrate on patents with the primary classification.

We constructed one corpus with strict requirements (only widely used classes, long patents and primary classification) and another with more relaxed requirements (also less widely used classes, shorter patents and secondary classification). The details are presented in Section 6.2. As a result of applying these requirements to our set of patents, the first corpus contains 73 classes while the second one is much larger with 1205 classes. In order to enhance readability, we will refer to the first corpus as C_{73} and the second one as C_{1205} for the remaining parts of this thesis. This size difference in connection with the expected higher quality of the documents due to the constraints we mentioned above should lead to better categorization results for C_{73} than for C_{1205}.

4.4 Evaluation

With our initial evaluation, we tested our method's ability to retrieve the classes that were actually assigned to the patents. Therefore, all of these classes were considered correct while everything else was considered wrong. While this approach can not evaluate our method's suitability for our objective of assigning new classes, it is nevertheless valuable for determining the quality of the classifiers by comparing their results with the categorization decisions made by the experts at the patent offices. Table 8 shows the macro-average scores (precision, recall and F_1-measure) of all

classifiers using 10-fold cross-validation for the confidence threshold 0.5; the use of other values is investigated below.

Corpus	Precision	Recall	F_1-measure
C_{73}	0.88	0.90	0.89
C_{1205}	0.88	0.84	0.86

Table 8: Evaluation results for confidence threshold 0.5. The precision values are identical for both corpora, but recall is considerably higher for the smaller corpus.

As Table 8 shows, the results are for the most part encouraging, with most values approaching 0.9. The recall value is 6% higher for the smaller corpus. Applying t-test to the recall values from the common classes of both corpora (i.e., the 73 classes from C_{73}) confirmed that this difference is statistically significant with very high confidence ($\alpha < 0.001$). However, despite the size differences between both corpora, the precision values are equal. This may suggest that the C_{73} results are about as good as can be expected from the use of the Maximum-Entropy method on patent texts. These F_1 values are still reasonably close to those reported by Tsatsaronis *et al.* for MeSH, especially considering the known problems caused by the complex patent text (cf. Section 2.3) as well as the fact that we restricted ourselves to using lexical features only. However, there are clear differences in precision and recall: In our case, both values are almost even, with recall even having a higher value than precision in C_{73}. In contrast to that, Tsatsaronis *et al.* report a drastically higher value for precision compared to recall ($> 99\%$ compared to $< 87\%$). As we will discuss in more detail below, this difference causes problems for our goal of assigning additional classes to patents.

The quality of the trained classifiers can also intuitively be judged by looking at the features that make the largest difference in categorizing documents. Table 9 shows the five most influential positive features from binary Maximum-Entropy classifiers for a subset of IPC classes with biomedical significance, i.e., the features that were assigned the highest positive values by the Maximum Entropy method. The occurrence of these words in a document that is supposed to be classified increases the likelihood of positive classification; in other words, the document is more likely to be assigned the category represented by the classifier. Almost all features listed in the table appear to be well-suited to making this distinction, since they are representative of their respective class. Although some of the class definitions are closely related (i.e., "Chemical analysis of biological materials" and "Chemical analysis involving proteins"), there is very little overlap in the

IPC code	Class definition (abbrev.)	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
A61B 5/00	Measurement for diag. purposes	light	sensor	blood	patient	tissue
A61B 17/00	Surgical instruments	tissue	suture	end	surgical	closure
A61B 17/70	Spinal positioners	rod	bone	portion	member	screw
A61F 13/15	Absorbent pads	absorbent	material	napkin	web	diaper
A61M 25/00	Catheter	catheter	distal	end	tube	lumen
B01L 3/00	Laboratory glassware	sample	fluid	channel	chamber	surface
G01N 33/50	Chemical analysis of biol. materials	sample	test	cell	specimen	light
G01N 33/543	Immunoassay	binding	analyte	sample	surface	antibody
G01N 33/68	Chemical analysis involving proteins	protein	peptide	antibody	detection	disease

Table 9: Most influential positive classifier features. Features were extracted from binary Maximum-Entropy classifiers trained on IPC classes with biomedical significance, leaving out stop words and words with three or less characters. The occurrence of positive feature words makes a document more likely to be assigned to the class. The positive features for the classifiers in the list are useful for identifying patents that belong to the class.

most influential features. For the two related classes about "chemical analysis" from the example (i.e., G01N 33/50 and G01N 33/68), the five top features are completely disjunct. The same is true for class A61B 17/00 about surgical instruments and its descendant A61B 17/70 about spinal positioners.

The situation is different for the most influential negative features for the same IPC classes shown in Table 10. The features are for the most part not specific enough to separate the class from other, even distantly related classes; most of them seem to be suited only for excluding documents from very distant classes. Additionally, some features are repeated in completely unrelated classes (i.e., "cell", "signal" and "antibody"). From our previous examples, classes A61B 17/70 about spinal positioners and G01N 33/50 about chemical analysis of biological materials are not closely related, but the negative feature sets for both classes contain the words "ink" and "signal". These differences are consequences of our choice to randomly select documents from other classes for the set of negative examples. This appears to have caused an overly strong random influence on the features selected by the classifier based on which documents were chosen for the negative

sample. While this method was positively evaluated by Tsatsaronis *et al.* [44] for assigning MeSH annotations, there may be more promising methods for selecting the negative sample. In Section 5.1, we evaluate in a different context what influence the choice of a background corpus has on the quality of terms extracted from patent documents.

IPC code	class definition (abbrev.)	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
A61B 5/00	Measurement for diag. purposes	cell	recording	antibody	network	sequence
A61B 17/00	Surgical instruments	absorbent	connector	electrode	unit	shell
A61B 17/70	Spinal positioners	acid	cell	ink	network	signal
A61F 13/15	Absorbent pads	tag	radio	insulin	nucleic	air
A61M 25/00	Catheter	acid	control	ink	information	signal
B01L 3/00	Laboratory glassware	user	terminal	ink	value	protein
G01N 33/50	Chemical analysis of biol. materials	plug	DNA	ink	signal	sequence
G01N 33/543	Immunoassay	cell	collagen	channel	section	particle
G01N 33/68	Chemical analysis involving proteins	antibody	Elafin	data	system	mobile

Table 10: Most influential negative classifier features. The occurrence of negative feature words makes a document less likely to be assigned to the class. The negative features for the listed classifiers are useful for excluding documents from very distant classes, but too unspecific for distinguishing between closely related classes.

While precision and recall values around 0.9 are generally acceptable when a single classifier is used, these values are problematic in our situation - especially when applied to the corpus with many classes. As we mentioned at the start of this chapter, it is important to distinguish between two different learning tasks: First, the retrieval of additional documents for a given relevant class. For the purpose of this task, our results are very promising with precision, recall and F_1-measure close to 0.9. We can therefore retrieve additional documents with high confidence. The second task, finding additional classes for a given document, is more problematic however. Since we apply all classification models to all documents, most documents are assigned more than one hundred classes in the case of C_{1205}. While this may appear to contradict our claim of the good performance of the individual classifiers, it does not: Even if every classifier makes the right decision in nine

out of ten cases, applying more than 1000 such classifiers will still lead to many wrong decisions. This means that although the performance of the individual classifiers is satisfactory, we have to take additional steps in order to make its use for our intended application feasible. In order to reduce the number of class suggestions, we tried various higher values for the confidence threshold. In the PubMed/MeSH experiments detailed in [44], the highest F_1-measure was reached for the confidence threshold 0.6. Unfortunately, our patent classifiers react less positively to raising the threshold, as can be seen in Figure 16: While raising the value from 0.5 to 0.6 clearly has a positive effect on precision, the corresponding drop in recall is much more severe and leads to a significantly lower F_1-measure. Raising the value further only has negligible effects on the classification quality, leading to very slight precision increases and recall decreases.

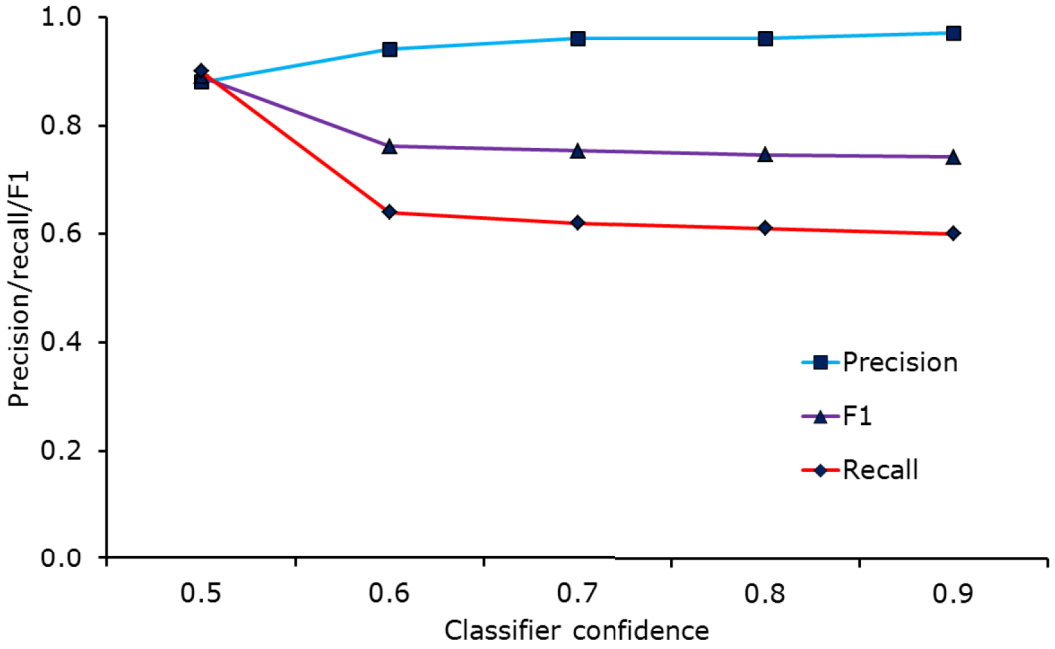


Figure 16: Classification results for corpus C_{73} depending on the confidence threshold. The F_1 measure is highest for the value 0.5 due to a rapidly decreasing recall. Increasing the threshold further after the value 0.6 only leads to small changes.

Due to the high number of patent classes, our method's precision would have to be very close to 1 in order to make it feasible for the task of assigning additional classes to all documents. Unfortunately, since raising the confidence threshold only leads to moderate increases in precision, we cannot reach a value high enough for practical application of the method by itself. Still, since most queries also include a keyword component, it is possible to use the described approach to

improve recall for such combined searches.

Despite that, we also tried to filter the assignments of our approach in order to make it useful by itself: As before, we applied every classifier to every model. However, instead of setting a confidence threshold for the classification score, we decided in advance how many classes were supposed to be assigned to each document. After calculating all classification scores, we only retained the predetermined number of highest-ranking classes. Figure 17 shows the recall of the method depending on the number of assigned classes, both for the exact class (i.e., the subgroup) and the more general main group. We calculated the main group recall by considering all subgroups below the closest main group as correct; in terms of our example from Section 2.4.1, for a patent from class A61K 38/17, also classes such as A61K 38/00 and A61K 38/16 are accepted.

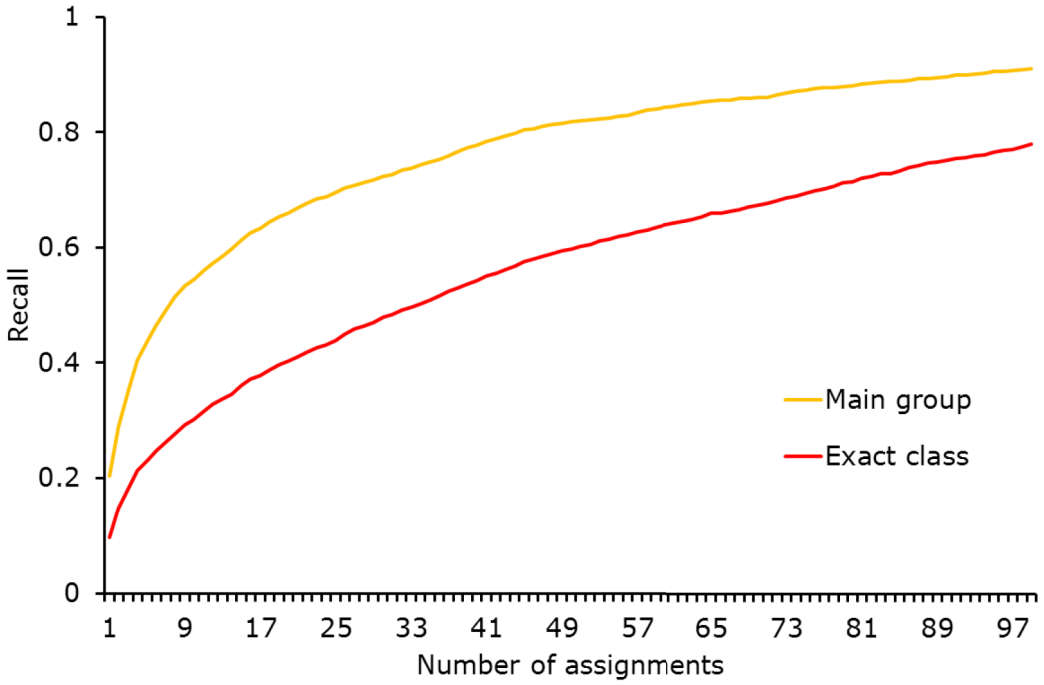


Figure 17: Recall for corpus C_{1205} depending on the number of assigned classes. The value grows rapidly until around ten classes, then continues growing at a slower pace.

We chose to have the method assign ten classes in order to strike a balance between recall and precision. A small-scale manual evaluation of the results revealed that this method is able to recreate some assignments and to add relevant classes that were not assigned. As an example, patent $EP1286824$ about an "apparatus for clamping and releasing contact lens molds" was correctly assigned to class B29D 11/00 about the production of optical elements, and it was also assigned to

the relevant class G02C 7/02 about lens systems - this class was not among the original assignments. However, even among the ten assigned classes for each patent, there were usually at least five completely irrelevant ones. The example patent about contact lens production was also assigned to class A61K 31/485 which is about "medicinal preparations involving morphinan derivatives" as well as class B29D 30/06 which is about "pneumatic tyres or parts thereof". These results make the practical application of the method without any other filters doubtful. They are however not unexpected considering the slow precision growth that we pointed out in Figure 16: Our method effectively increases the confidence threshold further, reaching different values for each classifier. But since precision remains almost constant (albeit at a high level), this is still not enough to remove all irrelevant assignments.

However, we have identified another possibility to filter out incorrect class assignments. As we will explain in more detail in Section 5.5, we performed an analysis on pairs of classes that were frequently assigned to the same patent document. This data can also be used for this task, since it is likely that classes that were never assigned to the same document do not cover similar subjects. Consequently, we implemented a filter that accepts additional class assignments only if there is an existing patent that was assigned a similar combination of classes. The filter has multiple possible settings, from very restrictive (only allow classes that have previously co-occurred directly) too much less so (allow pairs of classes if their respective ancestors of a certain hierarchy level have been co-assigned). An initial test had the result we were hoping for: For the example patent about contact lens production from the previous paragraph, our approach was able to filter out the incorrectly assigned classes about morphine and pneumatic tyres, since they had never been assigned to the same patent as class B29D 11/00 that had been assigned by the patent office. On the other hand, the newly assigned relevant class about lens systems had previously co-occurred with class B29D 11/00 and was therefore not filtered out.

In order to give a more general evaluation of the classifier features, we compared the feature sets of frequently co-occurring classes in contrast to randomly chosen classes. If a pair of patent classes is frequently assigned to the same patent document, it is reasonable to assume that these classes share a certain connection. Consequently, the corresponding classifiers should share at least some of their most important features. We chose the 100 most frequently co-occurring pairs of classes as well as 100 additional random classes. We then calculated the number of common features between different classes among the 100 top features for each classifier. Figure 18 shows these numbers for

the co-occurring pairs compared to the average for all 100 random classes. For all co-occurring class pairs, the overlap is considerably higher than for the randomly chosen classes. On average, the co-occurring classes have 35 common features, while the random classes only have nine. These results show that there is a clear connection between the co-occurrence of classes and the feature overlap of the corresponding classifiers. We therefore conclude that the features chosen by the classifiers are useful representations of the class patents' content.

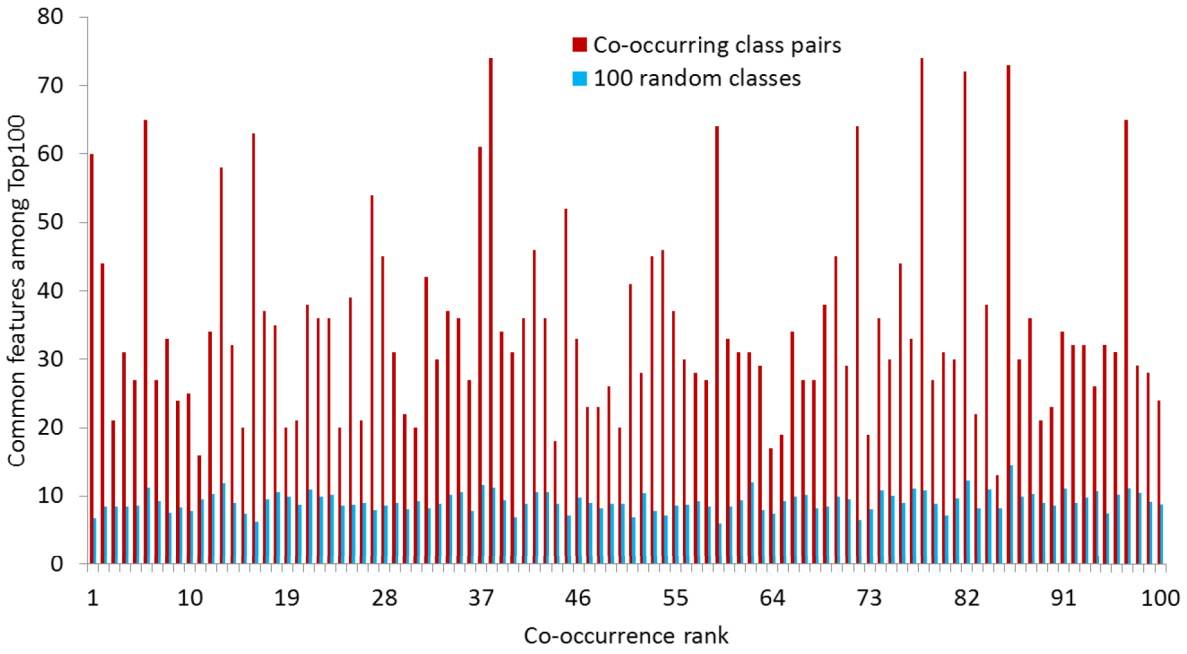


Figure 18: Classifier feature overlap among the Top 100 features for frequently co-occurring and random classes. The overlap is generally much higher for co-occurring classes, showing the significance of co-occurrence information.

This chapter examined the first part of our approach to address the problem of low class assignment numbers that we identified in Chapter 3, the automated assignment of additional classes. In the following chapter, we will investigate the second part, the expansion of the search query with additional relevant keywords and IPC classes.

5 Guided patent search

Summary

The second part of our approach to address the problem of low numbers of patent class assignments and simplify patent search combines multiple systems intended to guide the user towards quickly and easily formulating patent queries that are as complete as possible. An initial user query is used to determine additional relevant query components.
Additional class codes are suggested based on their co-occurrence with already entered classes, and we propose possibilities for retrieving additional keywords from different sources such as class definitions and external resources.
Most importantly, additional keywords are extracted from existing patents following the lessons learned from an in-depth evaluation of extraction methods. This evaluation was performed manually by four information professionals and showed that the *wf-idf* and *LLR* measures outperform the frequently used *tf-idf* measure by a wide margin, and that closely related patents should be chosen for the background corpus. The discovered terms and classes are recommended to the user so they can decide which of the proposals should be included in the final query.
As a proof of concept, we developed the patent retrieval prototype *GoPatents* that incorporates our proposals. Its implementation is based on the semantic search engine GoPubMed. Our system contains the documents from our EPO corpus and is mainly intended for searching patents with biomedical relevance.
The user has continuous access to an overview of the IPC hierarchy complete with definition trees, and they can directly see how result documents are distributed over the IPC hierarchy. This information also enables the user to filter or expand search results with additional classes. All resulting patent documents are also annotated with relevant terms from MeSH as well as the Gene Ontology and a protein database, making faceted browsing based on completely different aspects of the documents possible.
Result statistics are calculated automatically, giving the user an overview of the main classes, terms and applicants of the result patents as well as the temporal trend of the relevant patents. This gives the user a better intuition for whether their query is retrieving the intended documents.

Parts of this chapter were previously published in:

- Daniel Eisinger, Thomas Wächter, Markus Bundschus, Ulrich Wieneke, Michael Schroeder. Analysis of MeSH and IPC as a Prerequisite for Guided Patent Search. Bio-Ontologies 2012. http://bio-ontologies.knowledgeblog.org/346
- Daniel Eisinger, George Tsatsaronis, Markus Bundschus, Ulrich Wieneke, Michael Schroeder. Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed. Journal of Biomedical Semantics 2013, 4:S1. http://www.jbiomedsem.com/content/4/S1/S3

An additional possibility for tackling the problem of low class assignment numbers is the expansion of user queries to make up for the "missing" assignments. Since professional patent search queries are a combination of keywords and class codes in most cases, we investigate ways to expand both of these components in the following sections.

5.1 Extracting keywords from class patents

If a user query contains a class code, it can be assumed that the user is confident of the relevance of that class. If the system is able to identify and suggest keywords that are characteristic for this class, this enables the user to further refine their search. In order to find these relevant keywords, NLP techniques can be used to extract keywords from the patent texts that belong to the query classes. Since significant numbers of documents are available for most patent classes, this approach is able to deliver large numbers of keyword suggestions. In a way, extracting relevant words from class patents is what we already did for our categorization efforts: Each classifier gives weight parameters to the words contained in patent documents, with high values corresponding to words that are typical for the class. Table 9 in Section 4.4 shows the word features with the highest values for a selection of IPC classes with biomedical relevance, demonstrating that this approach is able to discover useful search terms. However, as Table 10 in the same section shows, the most important negative classifier features are not particularly suitable for distinguishing between closely related classes. We believe that this is due to our choice of negative example documents, i.e., a set of randomly chosen documents from all classes. While Tsatsaronis *et al.* reported good results for the same choice of negative examples for PubMed documents, we decided to investigate the influence that the choice of background corpus has on the quality of the terms that are extracted automatically from documents using statistical methods. Additionally, we investigated which ranking method for the extracted terms works best for patent documents. In order to judge the quality of the resulting term lists, we relied on the manual evaluation performed by four domain experts from the Scientific & Business Information Services department of Roche Diagnostics Penzberg.

Our categorization method is based purely on the individual words of the patent text. However, the extraction of multi-word terms from the patent text can result in even more valuable suggestions for query expansion. In order to reach that goal, we started with some preliminary term extraction experiments for individual documents that we describe in Section 5.1.1. Based on these results, we then expanded our approach to complete patent classes and performed a more in-depth evaluation

including expert feedback (Section 5.1.2).

5.1.1 Preliminary experiments

As a preliminary step, we carried out a basic investigation into whether statistical term extraction systems are able to retrieve terms with high relevance from patents. We used the web version[20] of the term extraction system that is also included in DOG4DAG [104]. It uses a variant of the well-known *tf-idf* measure to extract terms that are over-represented in a document compared to a pre-processed background corpus. The *tf-idf* method is based on counting the occurrences of each term in the given document as well as in the corpus documents and assigning two values to each term. The first value is called term frequency (*tf*) and corresponds to the number of times the given term occurs in the given document, usually normalized to account for different document lengths. The second value is called inverse document frequency (*idf*) and is calculated by dividing the total number of documents in the corpus by the number of documents containing the given term. The *tf-idf*-measure for a term is then calculated by multiplying its term frequency with the logarithm of its inverse document frequency. That way, a term's measure (and therefore its perceived importance for the given document) is high when it occurs often in the given document while rarely (or never) occurring in the corpus documents. Terms can then be ranked according to their *tf-idf* measure, with high-ranking terms assumed to be important for the document, while terms with lower ranks are considered to be less relevant.

The quality of this ranking depends to some degree on the corpus that was used for calculating the document frequencies. Ideally, its documents should meet two requirements: First, it is important that they cover a spectrum of topics that is not overly narrow, and that they aren't focused on exactly the same domain as the document the terms are supposed to be extracted from. If they are, it is likely that the document's important terms are also contained in many corpus documents and will therefore have a low *idf* value, leading to a low *tf-idf* rank. However, so far it was not clear how distant the relationship between the corpus and the document should be; we will investigate that question in Section 5.1.2. Second, they should have a similar type of text as the document itself. Otherwise, there is a risk that high-ranking terms are specific to the language of that type of text and not the domain the text is about. For example, if a patent document's terms are ranked by comparing them to a corpus of purely scientific texts, typical terms like "inventor", "invention",

[20] previously available at http://projects.biotec.tu-dresden.de/IdavollPlatform/; no longer functional

"claim", "description" and "embodiment" are likely to get very high ranks since they don't occur in most scientific texts. However, that doesn't make these terms relevant for the domain the patent belongs to - they're just a part of the typical language that is shared by almost all patents.

5.1.1.1 Validity of term extraction from patents

For our initial test, we compiled a small test collection of nine patents concerning different aspects of pipettes. We manually searched these patents for relevant terms and compared our results with the terms that were retrieved by the DOG4DAG Term Generation Platform [104]. The patent texts were tokenized and POS-tagged by the platform, sentences were identified and noun phrases following the pattern $[adj|verb]*[fill]\{2\}[noun]+$ were extracted as term candidates. The tag "fill" in the noun phrase pattern represents unspecific words that are often part of noun phrases, e.g., "of", "the", "for",... DOG4DAG then ranks these term candidates using an approximation of the *tf-idf* measure, with the set of all PubMed abstracts used as a background corpus. The domain covered by PubMed abstracts should be large enough to meet the first requirement for the corpus mentioned above, but the same can't be said about the second requirement: Since PubMed doesn't include any patents, it had to be expected that patent-specific terms would be ranked highly due to the large text differences between patents and scientific publications (cf. Section 2.3). We therefore took the occurrence of patent terminology into account during our evaluation of the quality of the extracted terms. For that purpose, we devised a five-point scoring scheme:

- Score 0: term has no relation to the domain and doesn't include patent terminology; e.g., terms like "end", "one", "user", "then",...
- Score 1: term includes patent terminology and has either no relation to the domain or is an extension of a domain term *that has also been found*, e.g., patent terms like "invention", "claim", "embodiment", unusual adverbs like "therethrough", "therewith", "therefrom" or extended terms like "pipette as claimed in claim", "said pipette tip", "aforementioned piston ring",...
- Score 2: term is either part of a domain term or includes one, e.g., "glass" instead of "glass pipette", "tip" instead of "pipette tip", "vacuum" instead of "vacuum source/vacuum pump/vacuum regulator",...

- Score 3: domain term that has also been found by the manual search, e.g., “pipette tip”, “shucking device”, “vacuum pump”,...
- Score 4: domain term that has been overlooked during manual search.

The results of our initial experiment were largely positive. Figure 19 shows the distribution of scores over the top 500 terms retrieved by the system (the maximum number of results) for one example patent from our test collection. Many of the “useful” terms (i.e., terms with scores between 2 and 4) are ranked highly, while terms without domain relevance (score 0) are usually pushed to lower ranks. It is therefore possible to retrieve a significant portion of the relevant terms with decent precision, although due to the appearance of additional useful terms on lower ranks, the recall may not be as high as would be desirable. On the other hand, if high precision is not essential, most relevant terms are among the retrieved terms. In the evaluated patents, about 73% of the manually extracted terms were found by the term generation system, and for an additional 13%, at least a part of the term was. Only 14% of terms were missed completely, in most cases because they were phrased in a complicated way that did not fit the pattern the system was looking for.

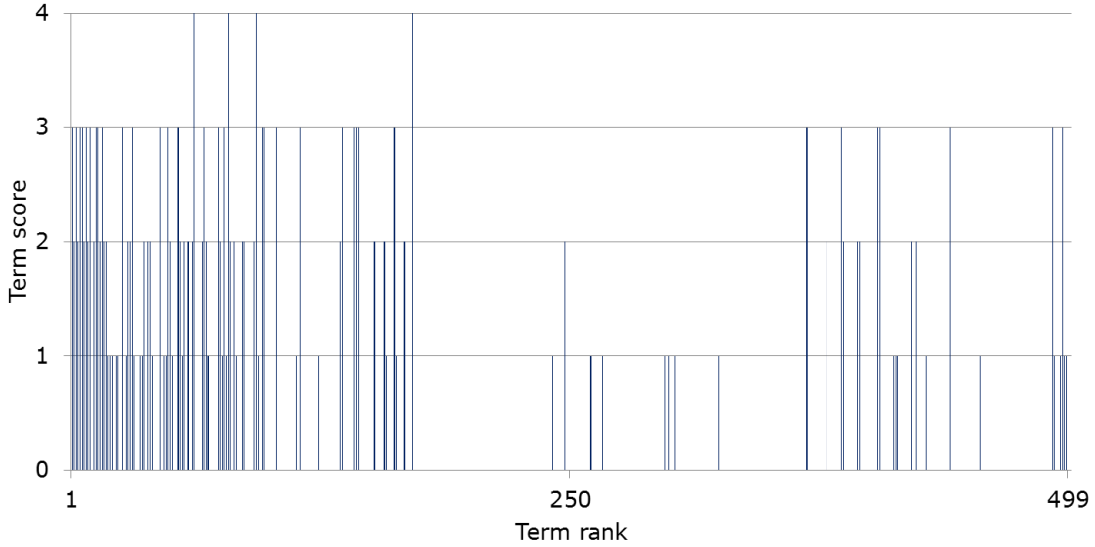


Figure 19: Distribution of term scores over the top 500 terms extracted using the DOG4DAG term generation platform. The concentration of “useful” terms (scores between 2 and 4) is highest among the top-ranked terms, but there is still a number of additional good terms on lower ranks.

5.1.1.2 Influence of the background corpus

While the high concentrations of relevant terms (scores 2 to 4) on high ranks and irrelevant terms (score 0) on low ranks as shown in Figure 19 is a positive result, the picture is less clear for terms with score 1. The fact that these terms are distributed widely means that the method is not effective for filtering out terms with patent terminology. As we explained previously, this result was not unexpected due to the choice of PubMed abstracts as a background corpus. Since most of the patent-specific terms occur rarely or never in these abstracts, they get a high *idf* value and therefore a rank that is higher than it should be. This problem can be solved by using different documents for calculating the *idf* value, documents that also contains these patent terms. A preliminary test of this possibility on the nouns of the pipette patents showed promising results. Table 11 shows the ten highest-ranking nouns of the example patent. When using all PubMed abstracts as a corpus (left column of Table 11), patent-specific nouns such as "FIG" (frequently used in patents for referencing a figure), "embodiment" and "invention" are among the top ten. However, when the patent corpus is used instead, all these nouns lose their high ranks and are replaced by more relevant ones (e.g., "fluid" and "adherence").

Term rank	Background corpus	
	PubMed	EPO patents
1	pipette	pipette
2	member	tip
3	tip	member
4	FIG	body
5	portion	passage
6	body	air
7	air	fluid
8	passage	adherence
9	embodiment	cords
10	invention	portion

Table 11: Comparison of highest-ranking nouns extracted from a patent when using either PubMed abstracts or patent texts as a background corpus. The use of patents leads to lower ranks for nouns that are connected to the patent domain, but not to the specific patent.

An alternative way of avoiding high-ranked terms with typical patent language but no domain relevance would be the creation of a stop word list that removes these terms completely. While this approach would also successfully remove the patent terms shown in Table 11, it may also have negative consequences for other terms that contain patent terminology but are still relevant for the

patent class. We therefore believe that using patent documents for the background corpus is the preferable approach for removing patent-specific vocabulary.

5.1.2 Term extraction evaluation

While our initial term extraction experiments (cf. Section 5.1.1) were restricted to extracting terms from individual patents, our goal of proposing query terms for specific patent classes makes term extraction from a complete document corpus (i.e., the patents of the respective class) necessary. Since our initial experiments resulted in promising terms and showed improved performance from using a patent corpus for the background statistics, we decided to also investigate further which patents should be included in the background corpus. We experimented with background corpora that were either closely or distantly related to the class that we extracted the terms from, as well as a general corpus with no direct relation. Apart from that, we also wanted to find out whether different statistical methods would be able to improve upon the performance of *tf-idf*.

5.1.2.1 Evaluated statistical term extraction measures

Apart from *tf-idf*, we used the previously published measures *wf-idf* and *Log-Likelihood Ratio (LLR)* for the term extraction, and we introduced two new variants of *tf-idf* and *wf-idf*.

Since *tf-idf* tends to over-represent the term frequency to the detriment of the document frequency [105, 106], we also used the variant that is introduced with the name *wf-idf* by Manning *et al.* [107]. The measure *wf-idf* is almost identical to *tf-idf* as defined in Section 5.1.1, but it applies a logarithm to the *tf* value before multiplying it with the logarithmized *idf* value. This is intended to reduce the boost that terms get from appearing very frequently, which makes sense in the patent domain where some terms are repeated very often. Exact definitions of all measures are given in Section 6.3.2.

Both *tf-idf* and *wf-idf* combine all documents from the domain corpus (i.e., the patent documents from the respective patent class) into one very large document that is used to calculate the term frequencies. We are also introducing new variations of both measures that we named *majority-tf-idf* and *majority-wf-idf* because they are based on taking the term rankings from different individual documents and combining them into a "majority vote". More precisely, terms are extracted and ranked for each document individually according to the respective measure. For each of the documents, terms on high ranks are awarded fixed point scores that are then added

up to retrieve a combined term ranking for the whole corpus. These measures are intended to give higher ranks to very specific terms that are extremely important for a small number of documents while rarely occurring in the others.

We also used an existing measure that preserves the corpus structure instead of relying on the simplification of treating the domain corpus like one big document, the *LLR* measure. It uses only the document frequencies from both corpora and doesn't take term frequency into account at all. *LLR* was proposed by Ted Dunning mainly for measuring the collocation strength of bigrams [108] - in other words, the Log-Likelihood Ratio was used as a measure of "unithood" (i.e., the collocation strength of terms) rather than "termhood" (i.e., the degree to which the term is representative for the domain). It was however also previously applied to measuring termhood [109,110]. When used for this purpose, *LLR* represents a statistical test of whether a term is more or less likely to appear in the domain corpus compared to the background corpus. Only terms more likely to appear in the domain corpus were included in the resulting term lists. For more details on all term extraction measures as well as additional parameters of our term extraction and ranking algorithms we refer again to Section 6.3.2.

5.1.2.2 Evaluation method

There are two main options for evaluating the quality of extracted terms:

1. Comparison with gold standard

   If a complete vocabulary of relevant domain terms is available, retrieved terms can be matched against this vocabulary in order to judge their relevance.

2. Expert evaluation

   Domain experts can manually go through the retrieved terms and judge their relevance based on their knowledge.

Both methods have advantages and disadvantages. In the first case, large-scale experiments can be carried out with relatively little effort since the evaluation can be automated for the most part. However, this is only possible if a complete gold standard for the respective domain is available, which is usually not the case for individual IPC classes. The second option on the other hand relies on the availability of experts for the respective domain - a prerequisite that is often even harder to meet, especially for specialized domains. Since every domain expert has a very specific point of

view on what is and isn't relevant for their domain, it is important that there is more than one expert. Even if a team of domain experts agrees to judge the retrieved terms, they usually do not have the time to go through thousands of terms, which means that the evaluation scheme as well as the terms that are to be evaluated have to be chosen carefully in order to minimize the required time investment while at the same maximizing the value of the evaluation. However, if these prerequisites can be met, the resulting evaluation has the potential to be more nuanced and informative than purely automatic procedures. In our case, four information professionals from the Scientific & Business Information Services department of Roche Diagnostics Penzberg agreed to judge the rankings resulting from the investigated combinations of parameters. We devised a simple rating scheme with scores between 0 and 2, with the following intended meanings:

- Score 2: This is a relevant term for the patent class.
- Score 1: This term is somewhat relevant, but not essential for the class; or:
  This string is not a relevant term by itself, but it is part of a relevant term, or it contains one.
- Score 0: Either this string cannot be considered a term, or it is a term without domain relevance.

Table 12 shows some examples for extracted terms and the scores that were assigned by the four experts. The example terms were extracted from patents of IPC class G01N 33/66 about the chemical analysis of biological material involving blood sugar. The detailed definition of the class is presented in Table 13 in Section 5.1.2.3. The evaluation procedure as well as the criteria we used for comparing the results are described in Section 6.3.4.

The examples in Table 12 show that while there were some unanimous decisions, the experts had some differences of opinion for most terms. In many cases, this meant that neighboring scores were assigned to a term (i.e., scores 2 and 1 or 1 and 0), but there were even cases where one expert judged a term to be very relevant and another one considered it to be completely irrelevant. Some of these divisive terms are included in Table 12, e.g., "sialic acid", "lectin" and "hemoglobin". Surprisingly, this was true for almost ten percent of all terms that were evaluated by the experts. This shows that even among domain experts, the perceived quality of a term depends strongly on the perspective of the user, again indicating the high complexity of the problem.

Term	Score 1	Score 2	Score 3	Score 4	Avg. score
glucose concentration	2	2	2	2	2.0
glucose measurement	2	2	2	2	2.0
glucose dehydrogenase	2	2	2	2	2.0
glucose monitor	2	2	2	2	2.0
blood sugar level	2	2	2	2	2.0
fructose	2	2	2	2	2.0
diabetes	2	2	2	2	2.0
insulin	2	2	2	2	2.0
carbohydrate	2	1	2	2	1.75
saccharide	2	1	2	2	1.75
oligosaccharide	2	1	2	2	1.75
polysaccharide	2	1	2	2	1.75
galactose	2	1	2	2	1.75
sugar chain	2	1	2	2	1.75
glucose meter	2	1	2	2	1.75
test strip	2	2	1	2	1.75
glycosylation endproducts	1	2	1	2	1.5
carbohydrate electrophoresis	1	2	1	2	1.5
glycolysis	1	2	1	2	1.5
reagent test pad	2	1	2	1	1.5
sialic acid	2	1	2	0	1.25
lectin	2	1	2	0	1.25
inositol content	1	1	1	2	1.25
biological fluid	1	1	1	2	1.25
blood sample	1	1	1	1	1.0
boronic acid	2	0	1	1	1.0
capillary channel	1	1	1	1	1.0
hemoglobin	1	1	2	0	1.0
abundance ratio	1	0	2	0	0.75
tetrazolium	1	0	1	1	0.75
fluid sample	1	0	1	1	0.75
mediator	1	0	1	0	0.5
interstitial fluid	1	0	1	0	0.5
hematocrit	1	0	1	0	0.5
albumin	1	0	0	0	0.25
protein	1	0	0	0	0.25
device	0	0	0	1	0.25
present invention	0	0	0	0	0.0
time	0	0	0	0	0.0
substance	0	0	0	0	0.0

Table 12: Examples for extracted terms, evaluation scores that were manually assigned by four experts, and resulting score averages. For most terms, the experts' opinions were divided.

5.1.2.3 Corpora

Each of the measures that we used (cf. Section 5.1.2.1) compares the term statistics in the domain corpus with those in a separate background corpus. Since it was one of our objectives to examine the influence of the background corpus on the quality of the ranking, we used three different background corpora with a varying degree of relatedness to the domain for each of the measures. All corpora consisted solely of patent texts and were carefully assembled with the help of an information professional based on IPC main groups and subgroups. We only considered patents with a publication date between 1990 and 2010 for all corpora, and we chose a subset of about 100.000 patents for each of the background corpora. While this number is relatively small for a background corpus, we believe that it represents a good balance between accuracy and running time of the extraction algorithms due to the considerable length of most patents. In order to avoid including potentially high numbers of very similar documents, we only included the most recent patent from each INPADOC patent family.

For the domain corpus of our evaluation, we chose an IPC class that all our experts were familiar with, subgroup G01N 33/66 about the chemical analysis of biological materials involving blood sugars. The exact definition tree of class G01N 33/66 is shown in Table 13.

Class code	Definition
G	Physics
G01	Measuring; Testing
G01N	Investigating or analysing materials by determining their chemical or physical properties
G01N 33/00	Investigating or analysing materials by specific methods not covered by groups G01N 1/00-G01N 31/00
G01N 33/48	Biological material, e.g. blood, urine
G01N 33/50	Chemical analysis of biological material, e.g. blood, urine; Testing involving biospecific ligand binding methods; Immunological testing
G01N 33/66	involving blood sugars, e.g. galactose

Table 13: IPC definition tree for class G01N 33/66.

For the background corpus most closely related to the domain, we collected patents from the diagnostics domain, but excluded patents belonging to G01N 33/66. The second corpus consisted of patents from the pharmacological domain which is still related to the class, but much less closely than the first background corpus. We again excluded patents also belonging to G01N 33/66 or to

the classes that formed the first background corpus. For the most general corpus we didn't restrict ourselves to any IPC groups and simply included patents from all fields except those covered by the other corpora. A graphical representation of the relationships between our individual patent corpora in the form of a Venn diagram is shown in Figure 20.

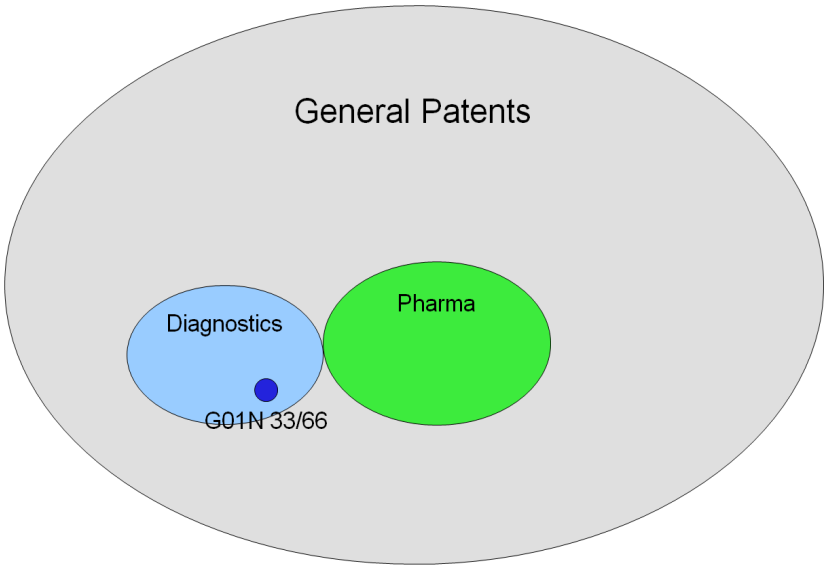


Figure 20: Relation between different background corpora. The diagnostics corpus contains the term extraction class, the pharma corpus is more distantly related to it, and the general corpus covers all classes.

5.1.2.4 Results

In order to judge the quality of the resulting term lists based on the scores given by our experts, we measured the average evaluation scores of the terms, and we calculated the "discounted cumulative gain" (DCG) of the different rankings. Section 6.3.4 explains in more detail how we calculated these measures.

Figure 21 shows the average expert scores for the top n terms ($n \leq 50$) for the separate ranking measures. According to the scoring system we detailed in Section 5.1.2.2, the best possible score average is 2 (i.e., all experts agree that a term is relevant) and the worst possible score average is 0 (i.e., all experts agree that a term does not have any relevance). As the figure shows, the best results are achieved using either *wf-idf* or *LLR*, followed by *majority-tf-idf* and *majority-wf-idf*. The frequently used measure *tf-idf* clearly results in the worst-scoring ranking. Interestingly, while this order of the individual ranking scores stays the same over most of the ranks, it appears that

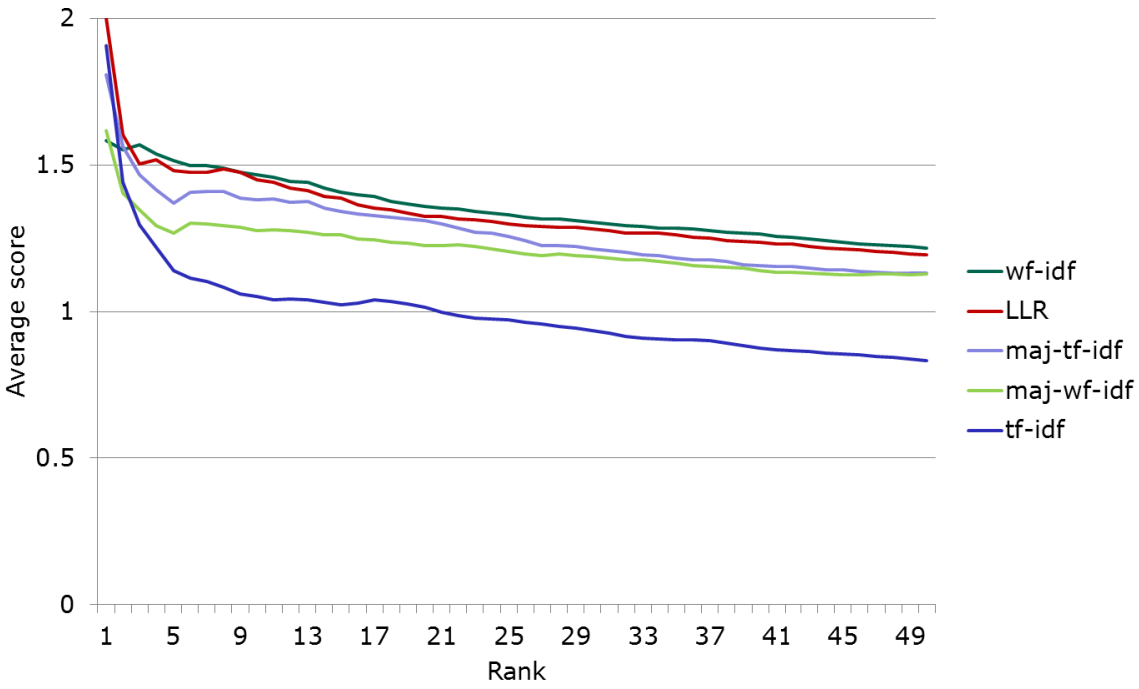


Figure 21: Influence of different ranking measures on the average score of extracted terms. Measures *wf-idf* and *LLR* perform best, followed by our proposed measures *majority-tf-idf* and *majority-wf-idf*. The frequently used *tf-idf* measure results in the lowest scores.

both *wf-idf* and *majority-wf-idf* have some problems with assigning the very highest ranks: While *wf-idf* becomes the ranking with the highest scores after a small number of ranks, it is considerably behind most of the others at first. On the other hand, *tf-idf* starts out close to the top before quickly becoming the worst measure by far. Similarly, *majority-wf-idf* starts out much worse than *majority-tf-idf*, but catches up to it on the lower ranks.

The differences between the different measures become even clearer in Figure 22 where the DCG values for all measures are compared. DCG is also based on the scores that were assigned by our experts, but it incorporates real-valued weights that give the terms on high ranks a stronger influence on the final score. This leads to a slightly clearer distinction between the different measures, although their order remains unchanged. In general, the results for the different ranking methods were remarkably consistent over all evaluation measures we calculated: The frequently used *tf-idf* measure was without exception the worst of the five measures that we investigated, and *wf-idf* as well as *LLR* were consistently the best. The two new measures we proposed, *majority-tf-idf* and *majority-wf-idf*, were unable to reach the scores that were achieved by *wf-idf* and *LLR*, but they

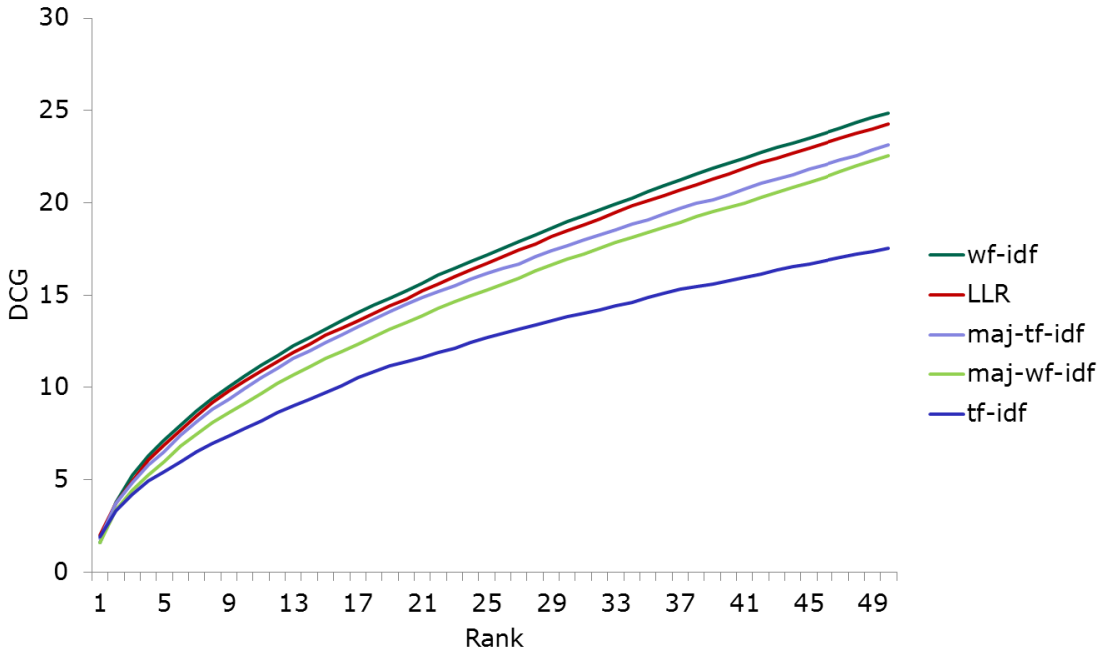


Figure 22: Influence of different ranking measures on the DCG value of extracted terms. Similar to Figure 21, measure *wf-idf* performs best, followed by *LLR* and our proposed measures *majority-tf-idf* and *majority-wf-idf*. The DCG value is the lowest by far for *tf-idf*.

were also considerably better than *tf-idf*. However, unlike the original measures where *tf-idf* was clearly worse than *wf-idf*, our measures reversed that situation to a certain degree. While the differences were much smaller, *majority-tf-idf* generally outperformed *majority-wf-idf*. This lends further evidence to the observation that *tf-idf* does not work well for overly long texts such as our artificial "domain document" that consists of multiple individual documents. Since the majority variants of both measures rank the extracted terms for each document individually, this disadvantage does not exist anymore, making *majority-tf-idf* the superior measure.

The second main aspect of our term extraction experiments was the evaluation of the influence of the background corpus on the quality of the extracted terms. As we explained in Section 5.1.2.3, we created three different corpora: One corpus that contains the class the terms were extracted from ("diagnostics"), one that is relatively closely related to it ("pharma") and one with no close relation ("general patents"). Figure 23 shows the average term scores for the first 50 term ranks, demonstrating that for our purpose of extracting relevant terms for a very specific domain, there is a clear benefit from choosing a background corpus that is closely related to the domain. The scores

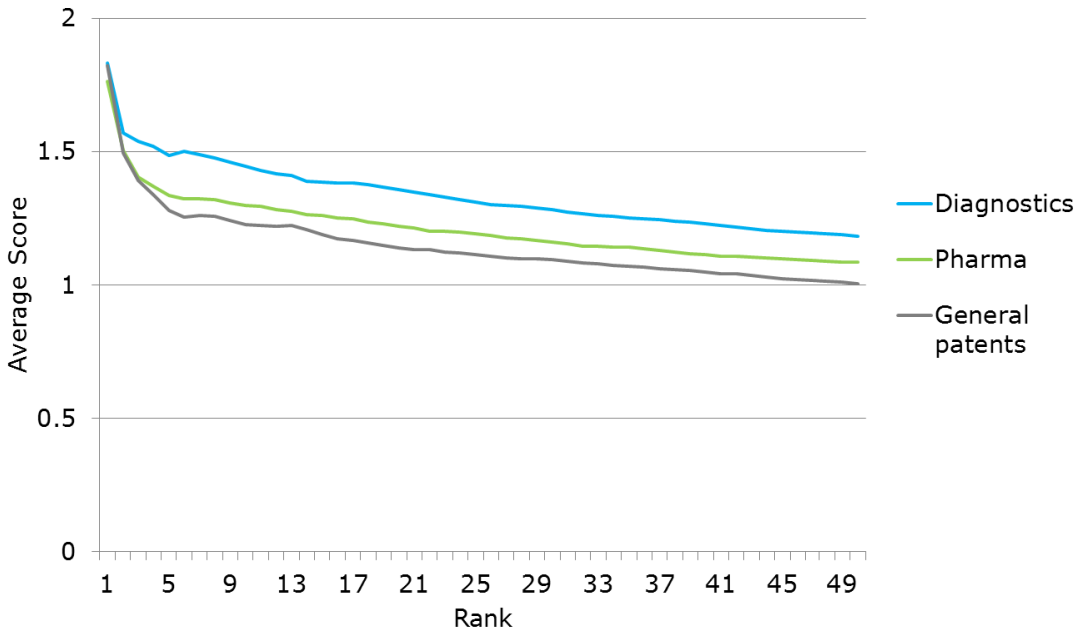


Figure 23: Influence of different background corpora on the average scores of extracted terms. On average, the extracted terms score highest with the closely related corpus (diagnostics) and lowest with the most distant corpus (general patents).

are highest for the diagnostics background corpus, followed by the pharma corpus and the general corpus with almost constant distances between the average scores. While the score differences between the different corpora are not quite as large as for some of the different ranking measures (cf. Figure 21), they still make it very clear that the background corpus should be taken into consideration when terms need to be extracted from documents.

Figure 24 takes a closer look at the differences between the results for different background corpora by comparing two measures for the precision of the term rankings. The upper three lines correspond to a fairly loose evaluation of the ranking where terms with an average score above 1.0 are considered good while the three lower lines have the much stricter requirement of a perfect 2.0 score average (i.e., all experts have to agree that the term is valuable). As the figure shows, the difference caused by the corpora is relatively small in the first case, but considerably larger in the second. This means that the result improvement from using the closely related corpus is mainly caused by a higher number of terms with very high relevance rather than a higher number of terms with medium relevance.

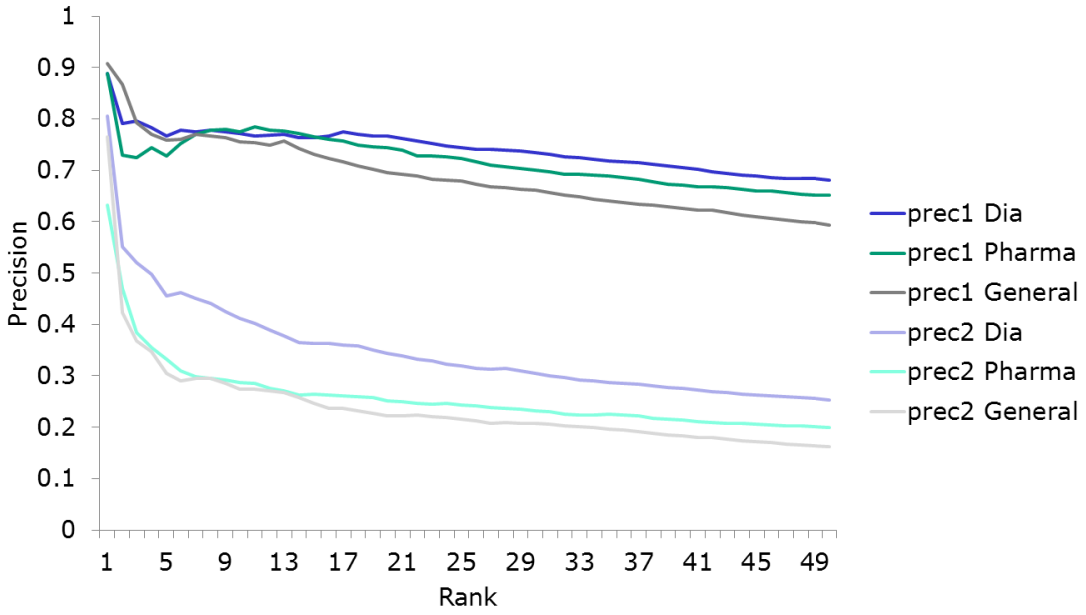


Figure 24: Influence of different background corpora on different precision values of extracted terms. The score advantages of the closely related corpus (diagnostics) are mainly due to very high-scoring terms (i.e., precision2).

5.1.3 Conclusion

In conclusion, our term extraction experiments showed that *tf-idf* in its original form is not a good choice for the extraction of relevant terms from a domain corpus such as the documents assigned to a certain patent class. However, switching to the similar *wf-idf* measure results in much better term rankings, and the Log-Likelihood Ratio is almost as good. While our proposed measures were unable to match these performances, their performance was still clearly ahead of *tf-idf*. Additionally, the different ranking approach led to a lower overlap with the top terms from other rankings and can therefore be considered a good complimentary approach for finding additional relevant terms.

Regarding the background corpus used for comparing the term statistics, our initial term extraction experiments indicated that a patent corpus should be used when the terms are extracted from patents. In our in-depth evaluation of different corpora, we were also able to demonstrate that a corpus with a close relationship to the domain documents clearly results in better term rankings than a more distantly related corpus or a mostly unrelated one. It is therefore advisable to choose different background corpora according to the domain of the documents that are the target of the

term extraction instead of using one fixed corpus for all term extraction tasks.

The quality improvement of retrieved terms that is achieved by implementing these proposals can be quite dramatic. Table 14 shows two different term rankings that were retrieved from the same set of documents, simply by changing the parameters of the extraction algorithm. Terms with high average scores (≥ 1.5) were marked blue, while terms with low scores (≤ 0.5) were marked red. This example again demonstrates that using the *wf-idf* measure and a specific background corpus (Ranking 2) is clearly preferable to the *tf-idf* measure and a general background corpus (Ranking 1). While these were the main results of our term extraction analysis, we also examined minor additional parameters of the algorithm. We present the corresponding results and the exact parameters used for the rankings in Table 14 in Section 6.3.4.

Rank	Ranking 1		Ranking 2	
	Term	Score	Term	Score
1	glucose	2.0	glucose	2.0
2	sample	0.75	carbohydrate	1.75
3	blood	1.0	blood sugar	2.0
4	concentration	0.5	glucose concentration	2.0
5	method	0.5	sugar level	2.0
6	reagent	1.0	blood glucose	2.0
7	protein	0.25	saccharide	1.75
8	diabetes	2.0	sugar	2.0
9	patient	0.25	blood sugar level	2.0
10	invention	0.0	glucose meter	1.75
11	presence	0.25	glucose measurement	2.0
12	sensor	1.0	test strip	1.75
13	present invention	0.0	glucose monitor	2.0
14	amount	0.25	glucose dehydrogenase	2.0
15	enzyme	1.0	glucose level	2.0

Table 14: Highest-ranking extracted terms for two different sets of parameters with average assigned evaluation scores. Plural terms that were also included in singular form were removed. The quality of Ranking 2 is clearly far superior to Ranking 1 despite being retrieved from the same set of documents.

5.2 Extracting keywords from class definitions

As Section 5.1 showed, extracting terms from class patents results in good keyword suggestions, but the extraction process is relatively complicated, especially if different background corpora have to be constructed to fit the respective patent class. A more straightforward way of turning class codes

into keyword suggestions is by considering the corresponding IPC definitions. Keywords and terms can again be extracted directly from the definition text by using established NLP techniques such as part-of-speech tagging, grammatical parsing and noun-phrase chunking. This simple approach can lead to high-quality term suggestions for refining an initial query. If the subordinate definitions of the class code are taken into account, this can especially result in very useful filter terms for the keyword search. For example, if a searcher enters the class code G01N 21/00 ("Investigating or analyzing materials by the use of optical means, i.e. using infra-red, visible, or ultra-violet light"), the system can extract closely related terms and phrases such as "cuvette", "refractivity", "specular reflectivity" and many others from the subordinate definitions.

In addition to this direct approach, the morphosyntactic structure of some definitions can be exploited to get additional related keywords that do not appear in the IPC definition hierarchy. For definitions containing lists of related terms, we use the system described by Fabian *et al.* [111] to find additional terms with the same relation and suggest the top-ranking ones to the user. This system was originally intended to assist ontology engineers in improving and extending their ontologies, and has therefore been integrated into the two most important ontology editors, *OBO-Edit* [112] and *Protégé*[21], as part of the ontology generation plug-in *DOG4DAG* [104]. It suggests additional co-hyponyms to existing ontology entries, i.e., terms that share a sibling relationship with a set of existing ontology terms that have a common parent. Its term suggestions are extracted from web pages using the pages' HTML structure and different textual patterns, making use of the information contained in the ontology. However, while ontological information helps improve the quality of the generated siblings, the system does not rely completely on the availability of an ontology. Sibling terms can also be recovered just by entering a small number of seed terms. This functionality can be used for any IPC definition that contains a list of sibling terms; in many cases, such terms appear in the form of examples for the inventions covered by the class. As an example, for the IPC definition "Orthopaedic devices [...] such as splints, casts or braces" (class A61F 5/01), the system proposes the relevant sibling terms "slings", "collars", and "crutches". For a baseline Boolean keyword query simply connecting the terms with "OR", the result set almost doubles in size after the inclusion of the generated sibling terms. Our system detected 3053 IPC classes ($\approx 4\%$) that contain enumerations and can therefore in principle be used in this way for query expansion.

[21] http://protege.stanford.edu/

5.3 Extracting keywords from external ontologies

Existing ontologies are another possible source for additional keywords. If an ontology term can be matched to an IPC class definition, any additional information contained in the ontology about the term (e.g., its synonyms) can be used to add suggestions for the user. As a proof of concept, we again used the GoPubMed annotation pipeline [39] to map MeSH terms to an IPC subset with biomedical relevance. For that purpose, we selected all subclasses of the IPC class "A61K" with the definition "Preparation for medical, dental or toilet purposes" (981 subclasses). The annotation results provided at least one MeSH term for 865 of these classes (88%), and three or more terms for 466 classes (48%). Many IPC classes were matched with very relevant MeSH terms, e.g., class A61K 48/00 ("medicinal preparations containing genetic material which is inserted into cells of the living body to treat genetic diseases; gene therapy") with MeSH terms including "Genes", "Cells" and "Gene Therapy". On the other hand, there were also incorrect matches, often due to shortened MeSH synonyms. For the example class, the MeSH term "Containment of Biohazards" was considered a match because the word "containing" in the class definition was mapped to "Containment" which MeSH lists as a synonym. Since our system proposes expansion terms to the user instead of automatically adding them, this high level of coverage represents a valuable addition despite the inclusion of some false positive annotations. Additionally, as we will demonstrate in Section 5.6, mapping ontology terms to the patent texts that are available from a patent retrieval system makes faceted browsing of patent search results possible.

The availability of a domain ontology also offers the possibility of enhanced sibling generation: If an IPC definition contains a MeSH term as well as one of its child terms in the form of an example, it is reasonable to assume that all other child terms are also relevant. Following this intuition, IPC definitions of this form (e.g., "Sulfonylureas, e.g. glibenclamide, tolbutamide, chlorpropamide") lead to term suggestions with very high precision (for the example: "Carbutamide", "Acetohexamide", etc.). Of the biomedical IPC subset, this was possible for 72 classes (7%).

5.4 Repurposing class-keyword mappings for class suggestion

In the previous sections, we have presented multiple ways of identifying additional terms and keywords for a given classification query:

- by applying NLP methods to patents from the class for term extraction and using statistical

measures for ranking them

- by reusing the features that were identified as most important by the respective classifier
- by extracting noun phrases from class definitions
- by retrieving additional terms such as synonyms from existing ontologies.

Each of these possibilities corresponds to a mapping between terms and patent classes and can therefore also be used in the opposite direction, for proposing classification components to add to keyword queries. If the user enters a keyword that has been mapped to an IPC class, this class can be suggested to the user for expanding their query. Since the size of the classification system prevents users from knowing all class definitions, this information has to be displayed with the suggested class code. Consequently, even users unfamiliar with the IPC can profit from classification information without investing too much effort into getting to know the classification system. This is especially true for the biomedical domain, since the availability of detailed domain ontologies leads to very precise class suggestions. The WIPO website used to offer similar functionality: The system *TACSY* expected a small number of keywords as input and suggested related IPC classes from different hierarchy levels. However, it was not made clear what the system's class proposals were based on and the service was stopped in November of 2012 without further explanation[22].

5.5 Using class and term co-occurrences for query expansion proposals

Apart from mapping keywords to classes and vice versa as shown in the previous sections, it is also possible to use the co-occurrence of either to retrieve more relevant components of the same type for the query. For keywords, we have already presented various possible sources for co-occurrence statistics: Common features of classifiers, terms extracted from the same class documents or from closely related class definitions and especially ontology terms with a known relationship. For patent classes, the existing patent data represents a more direct source. In order to find closely related classes to suggest to the user, we analyzed the class co-assignments in our patent corpus. We collected all pairs of classes that were assigned to the same patent and ranked them both on the absolute number of co-assignments and the relative number in the form of their Jaccard-Index. We hypothesize that pairs of classes with high ranks in either ranking are related closely enough

[22] http://www.wipo.int/tacsy

that many searches for one of the classes will also have additional relevant results in the second class. We therefore propose to suggest these frequently co-occurring classes to the user for query expansion.

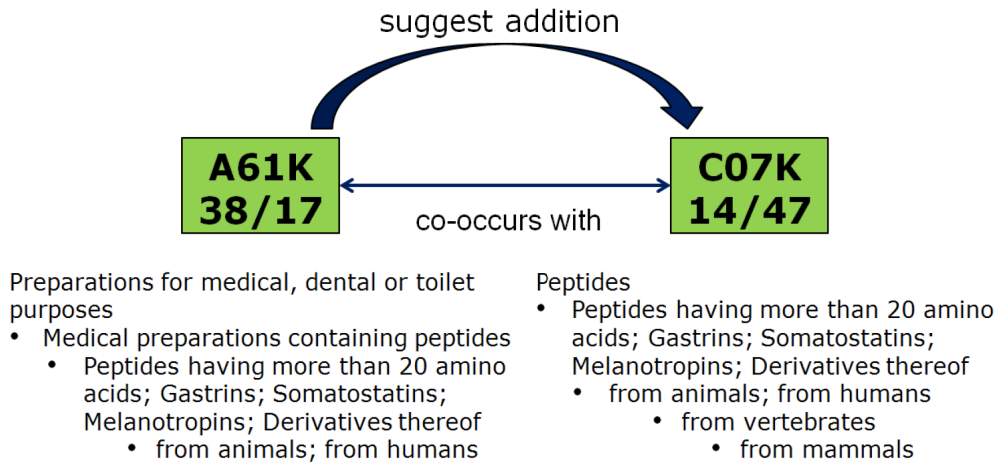


Figure 25: Example for semantically related IPC classes without any hierarchical relation, detected using co-assignment information.

Many resulting class suggestions are from the same hierarchical tree but not directly related, i.e., they cover patents with similar aspects to the ones searched for by the user. Additionally, the rankings include pairs of classes from completely separate parts of the hierarchy that are also highly related; in many cases, they represent different points of view. Figure 25 shows one example of such a pair of classes, including their definition hierarchy. The left class is clearly more application-oriented than the right one, since it deals with "medical preparations containing peptides" while the right class concentrates on the peptides themselves. However, we argue that many searchers interested in patents from one class will also find relevant patents in the other one. We used the professional patent search tool Thomson Innovation[23] to find out how recall is affected when only one class is used for search. For these example classes, searching for only the first class leads to over 50% missed possible results, and searching only for the second still leads to 25% missed results. The situation is similar for the pair of classes shown in Figure 26, also detected using co-assignment information.

As these examples show, it is possible to find additional relevant classes to expand user queries based on class co-occurrence. The relevance of this co-occurrence information was also confirmed

[23] http://info.thomsoninnovation.com/

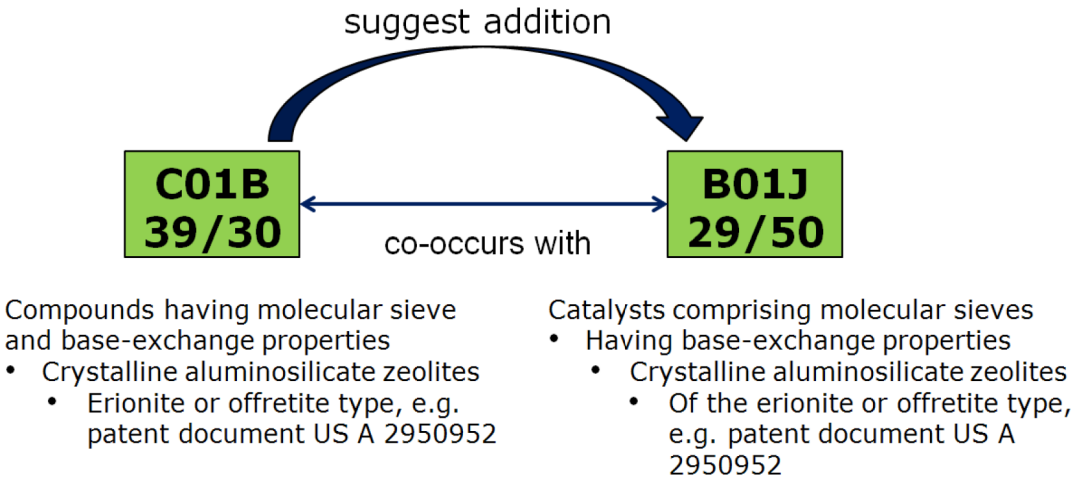


Figure 26: Second example for semantically related IPC classes without any hierarchical relation, detected using co-assignment information.

by the larger overlaps in the top features of the corresponding classifiers compared to random class pairs that we presented in Figure 18 in Section 4.4. In order to ensure the quality of our suggestions, we based the co-assignment statistics solely on the existing EPO assignments.

5.6 Patent retrieval system prototype *GoPatents*

In order to give a demonstration of our proposals, we implemented a patent retrieval prototype in cooperation with Jan Mönning. Since our implementation is based on the semantic search engine *GoPubMed*, we decided to name our system *GoPatents*. As described in Section 2.5, GoPubMed enables the user to filter the resulting PubMed abstracts using terms from MeSH, Gene Ontology and a protein database. This functionality is brought over to GoPatents, with the added benefit of also allowing the use of IPC classes for the same purpose. The user interface is divided into two columns, a main window on the right and a side column on the left; an overview is given in Figure 28, showing the following main components of the system:

- The term hierarchies (left column, second from top)

  GoPatents enables the user to refine their search using relevant concepts from different sources. All patent documents have been automatically annotated with MeSH terms, GO concepts and proteins from a protein database. Additionally, the assigned IPC classes have been extracted from the XML patent files. The complete hierarchies of all these annotation systems are shown continuously, with the numbers in square brackets as well as the gray bars next to the

respective term indicating how many of the retrieved documents were annotated with it. The arrows next to each term enable the user to expand lower levels of the hierarchies for more precise information about the contents of the search results. Since the IPC class codes are not informative for users without patent search experience, hovering the mouse over a code opens a pop-up window with the complete definition hierarchy of the term. An example of this functionality is shown in Figure 27.

Each concept from each of the included hierarchies can be used to refine the search in different ways. Clicking on a term opens a small pop-up window that offers options to restrict the search to documents containing the term or conversely, excluding all these documents from the search. Both of these filtering options can be combined for multiple terms from different hierarchies, thereby allowing the user to quickly compose a very focused query.

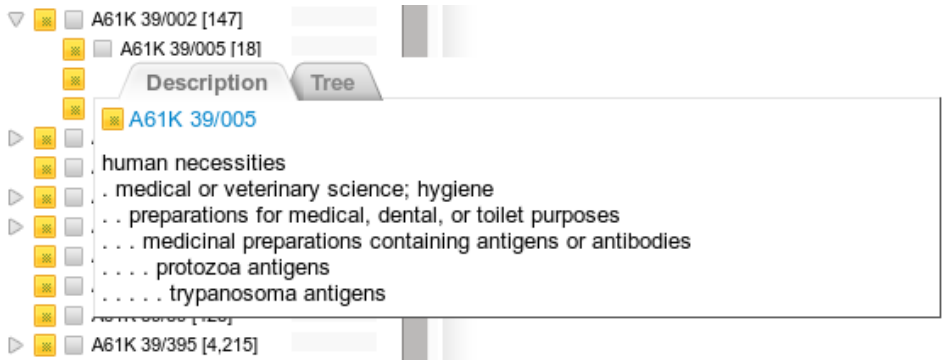


Figure 27: IPC class definition tree displayed by GoPatents patent retrieval system prototype. Hovering the mouse over any IPC code opens a pop-up window with this information.

- The additional filtering options (left column, third to fifth from top)

  Besides automatically annotated terms and IPC classes, GoPatents offers additional possibilities for faceted browsing. Applicant information is extracted from the XML patent files along with publication dates, and search queries can be refined further by applying this data from the "who" and "when" components in the left column in the same way that we described above.

- The search field for entering queries (main window, top)

  Queries can consist of keywords, IPC classes, terms from the different included hierarchies as well as the previously described additional filtering options.

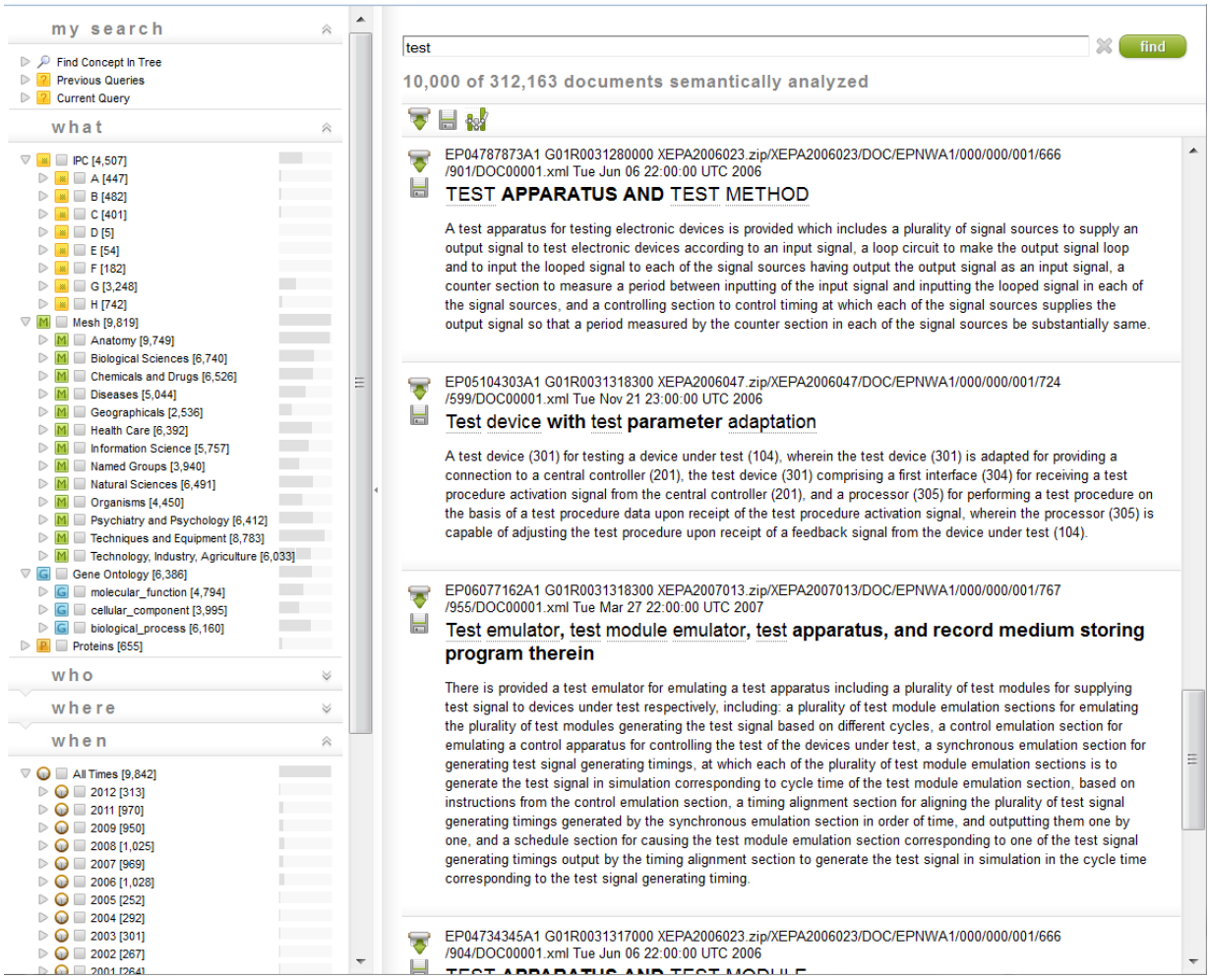


Figure 28: Overview of GoPatents patent retrieval system prototype. The query is entered in the box on top, result documents are shown below, and the faceted browsing functionality is available in the left column.

- The search results (main window, bottom)

  Patents that fit the initial query as well as any additional requirements made by including or excluding other facets are displayed in the main part of the window. This result overview shows the patent titles along with a text snippet that is supposed to help the user judge whether the patent is of interest to him. Clicking on a result opens the complete patent in a separate browser tab.

- The search history/concept finder (left column, top)

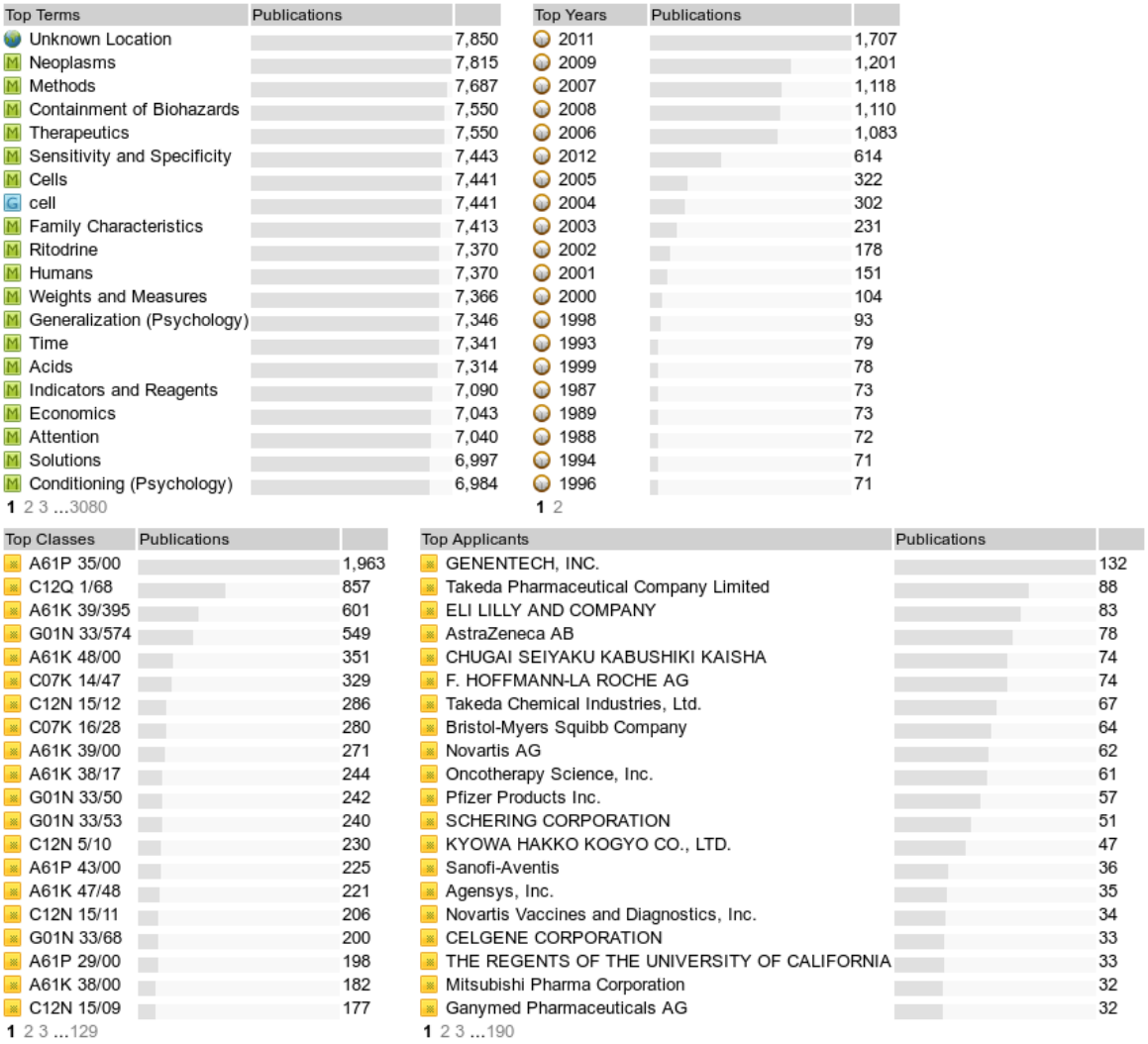
  Since it is very easy to modify queries considerably with few mouse clicks, it is useful to

give the user access to their search history. If a query modification proves unsuccessful, this enables them to go back to a previous result. The concept finder can be used to search the hierarchies for relevant concepts that can be added to a query.

While all the described functionality is active and working in GoPatents, it is still a prototype that is not yet fit for professional patent search. The system has not yet been optimized and is therefore due to the size of the corpus not fast enough when the results for keyword queries are retrieved. This can lead to very long delays (almost five minutes) after submitting an initial query and limits the number of documents that are semantically analyzed. However, once the results have been retrieved, adding and removing different facets works much faster. Additionally, our proposed query expansion methods have yet to be fully integrated into the system. On the other hand, result statistics are calculated automatically and can be accessed instantly by the user as soon as the result set has been retrieved. As Figure 29 shows, these statistics cover multiple aspects of the result set:

- The most frequently assigned terms from the different hierarchies (MeSH, GO and proteins) are listed. This gives the user an intuitive overview of the main topics covered by the retrieved patents, and is therefore a good indicator of whether the search retrieves the documents that the user intended to find.
- The most frequent patent classes in the result set are collected, informing the user about the parts of the IPC hierarchy that are most closely connected to their query. This may help the user discover additional aspects of the search results, and therefore allow them to refine their query by adding or excluding additional classes.
- The top publication years of the retrieved patents are listed in order to display the temporal trend of patent publications relevant to the query.
- The applicants with the highest numbers of patents among the results are retrieved. This is especially useful for professional patent searchers looking for the main competitors in a given area.

In this chapter, we proposed various methods for retrieving additional keywords and patent classes to expand and refine patent search. Additionally, we introduced the patent retrieval prototype GoPatents that incorporates some of our proposals. The following chapter will describe the



Top Terms	Publications
Unknown Location	7,850
Neoplasms	7,815
Methods	7,687
Containment of Biohazards	7,550
Therapeutics	7,550
Sensitivity and Specificity	7,443
Cells	7,441
cell	7,441
Family Characteristics	7,413
Ritodrine	7,370
Humans	7,370
Weights and Measures	7,366
Generalization (Psychology)	7,346
Time	7,341
Acids	7,314
Indicators and Reagents	7,090
Economics	7,043
Attention	7,040
Solutions	6,997
Conditioning (Psychology)	6,984

1 2 3 ...3080

Top Years	Publications
2011	1,707
2009	1,201
2007	1,118
2008	1,110
2006	1,083
2012	614
2005	322
2004	302
2003	231
2002	178
2001	151
2000	104
1998	93
1993	79
1999	78
1987	73
1989	73
1988	72
1994	71
1996	71

1 2

Top Classes	Publications
A61P 35/00	1,963
C12Q 1/68	857
A61K 39/395	601
G01N 33/574	549
A61K 48/00	351
C07K 14/47	329
C12N 15/12	286
C07K 16/28	280
A61K 39/00	271
A61K 38/17	244
G01N 33/50	242
G01N 33/53	240
C12N 5/10	230
A61P 43/00	225
A61K 47/48	221
C12N 15/11	206
G01N 33/68	200
A61P 29/00	198
A61K 38/00	182
C12N 15/09	177

1 2 3 ...129

Top Applicants	Publications
GENENTECH, INC.	132
Takeda Pharmaceutical Company Limited	88
ELI LILLY AND COMPANY	83
AstraZeneca AB	78
CHUGAI SEIYAKU KABUSHIKI KAISHA	74
F. HOFFMANN-LA ROCHE AG	74
Takeda Chemical Industries, Ltd.	67
Bristol-Myers Squibb Company	64
Novartis AG	62
Oncotherapy Science, Inc.	61
Pfizer Products Inc.	57
SCHERING CORPORATION	51
KYOWA HAKKO KOGYO CO., LTD.	47
Sanofi-Aventis	36
Agensys, Inc.	35
Novartis Vaccines and Diagnostics, Inc.	34
CELGENE CORPORATION	33
THE REGENTS OF THE UNIVERSITY OF CALIFORNIA	33
Mitsubishi Pharma Corporation	32
Ganymed Pharmaceuticals AG	32

1 2 3 ...190

Figure 29: Result statistics automatically generated by GoPatents prototype. The most frequently assigned terms from the different hierarchies are listed, as are the best-represented IPC classes, years and applicants.

methods we used in previous chapters (i.e., for the analysis of MeSH and IPC as well as the patent categorization) as well as in this chapter (with additional minor results for the term extraction) in more technical detail.

6 Methods

This chapter describes the methods we used for our experiments in more detail. The first section explains our analyses of MeSH and IPC, the second section concerns our experiments with patent categorization, and the third section gives additional information about the term extraction from patent documents.

6.1 Analysis of MeSH and IPC

Both analyses were carried out in two steps: We first retrieved and analyzed the terms and their hierarchical relationships, and then the annotations to our document corpora.

6.1.1 MeSH

For our analysis of the MeSH hierarchy, we used the XML version of MeSH 2012 retrieved from the MeSH homepage[24]. We extracted all MeSH terms with their MeSH IDs as well as the tree numbers from the file. The tree numbers were then used for reconstructing the hierarchy. We implemented graph-based methods for calculating different hierarchical properties such as the number of nodes per hierarchy level. For the PubMed/MeSH annotation analysis, we used the complete Medline dataset with MeSH annotations, downloaded from PubMed on September 22, 2011. After extracting the necessary information about documents and annotations, we analyzed it using a custom implementation, calculating different characteristics of the data such as the average number of annotations per document.

6.1.2 IPC

We reconstructed the IPC hierarchy using HTML files available from the WIPO homepage[25]. After manually entering the eight sections of the IPC with their definitions as top nodes of the hierarchy, our implementation automatically extended the hierarchy step by step: Each section file (e.g., "A.htm") contained all main classes of the section, allowing us both to add them to our representation of the hierarchy and to retrieve the corresponding HTML files. Then the subclasses were extracted from the main class files (e.g., "A01.htm"), and the main groups and subgroups from the subclass files (e.g., "A01B.htm"). Since the class codes do not correctly reflect the parent/child

[24] http://www.nlm.nih.gov/mesh/filelist.html

[25] http://www.wipo.int/ipc/itos4ipc/ITSupport_and_download_area/20060101/subclass/advanced/en/20060101_en_al_xml.zip

relationship between entries at the subgroup level (cf. Section 2.4.1), we used the dot representation in the files to ensure the accuracy of our representation of the hierarchy. Class definitions were also extracted from the files in string form; images contained in a number of chemistry-related class definitions as well as references to related classes were removed. The analysis of the hierarchy was carried out using a slightly modified version of our MeSH implementation, leading to directly comparable results.

For the patent annotation analysis, we used XML files published by the EPO via their subscription-based "Product 14.12"[26]. We used the complete set of patent applications from the years 1981 to 2005, and we extracted document numbers as well as all classification information from the files. The reason for our exclusion of more recent patents was a change in EPO publication policies and data formats. Since there have been multiple updates to the IPC that are not reflected in the EPO's files, we decided to use the 2006 version of IPC in order to minimize the number of class assignments that could not be matched. The document numbers were used to make sure that different versions of the same patent were not counted multiple times. For the classification information, we extracted both primary and secondary classification codes and combined them into one set of codes per patent. We again used a modification of the PubMed implementation to perform our analyses of the data.

6.2 Automated patent categorization

As we described in Section 4.1, our approach to assigning additional classes to patents was based on an approach by Tsatsaronis *et al.* for assigning MeSH terms to documents using Maximum Entropy (MaxEnt) classification. For each MeSH term that was to be used for document annotation, a binary MaxEnt model was learned from already annotated documents and applied to new ones [44]. We applied the same principle to patent classification, learning IPC classifiers from existing patent documents with classes manually assigned by professional patent examiners. We will now describe the details of our implementation. Our corpus was again a subset of the EPO dataset we used for the IPC analysis. For the classification, we used patents published after 2005 and before July 2012. As we described in Section 4.3, we constructed two corpora by applying three different criteria: the number of patents, the text length and the optional restriction to primary classification. We included all classes that had the required number of patents fulfilling the text length requirement.

[26] http://www.epo.org/searching/subscription/raw/product-14-12.html

In the first corpus, only patents that had the class as primary classification were counted. We then collected the required number of patents for each class by randomly choosing from the complete set. Table 15 shows the values we chose for the parameters as well as the number of classes that fulfilled the requirements and the resulting number of patents per corpus.

training corpus	number of patents/class	minimum text length	restricted to primary classification	number of classes	total number of patents
C_{73}	200	8000 characters	yes	73	14600
C_{1205}	100	2000 characters	no	1205	120500

Table 15: Training corpora for patent categorization. C_{73} has more patents per class with longer text and only primary classification.

We used the Java API of the open-source machine-learning toolkit Mallet (version 2.0.7) [113] for our classification efforts. The pre-processing was done in two steps: For each of the patent documents that were chosen from the EPO corpus for inclusion in C_{73} or C_{1205}, we first created a text file that contained all text fields from the corresponding XML file. We then created a feature vector from each text file by using existing and custom implementations of Mallet's *Pipe* interface in sequence. The classifiers were trained by executing the *train* method from the Mallet class *MaxEntTrainer*.

The training sets for each classifier were constructed as follows: For the positive set, all patents that the corpus contained for the respective class were included. For the negative set, a few different approaches were investigated by Tsatsaronis *et al.* [44]. Since the differences were very small, we decided to use the most simple option: We randomly chose the same total number of patents as in the positive set from the set of all other classes. In order to avoid the over-representation of individual classes, we shuffled all these classes and randomly selected one document from each of them in turn. Despite taking that step, the negative features seem to be overly concentrated on separating very distant technological fields and less useful for detecting subtle differences between classes (cf. Table 9). We plan to investigate different possibilities for constructing the negative set, e.g., increasing the number of documents from fairly similar classes. However, while this may help fine-tune the negative features, it is possible that the currently high quality of the positive features will suffer.

We used 10-fold cross-validation and calculated the macro-average scores (cf. Table 8). Since the cross-validation methods that are included in Mallet do not conserve the ratio of positive and

negative training documents, we implemented a custom method for this task as well as for the evaluation of the categorization results.

Since both our approach and our objective for patent categorization differ considerably from the previous approaches we mentioned in Section 2.6.1, comparing the results directly is not possible: Almost all existing approaches are restricted to higher levels of the hierarchy, and all of them are used for assigning one single class instead of sets of classes.

6.3 Term extraction from patents

Automatic term extraction from any kind of document is usually performed in two steps: First, term candidates are identified by applying certain natural language processing techniques to the text. Second, statistical methods are used to rank the term candidates by how well they represent the document; this step usually involves the comparison of term statistics with one or more existing corpora. The following subsections describe in detail how both steps were performed during our term extraction process.

6.3.1 Text processing

The patent texts from both the domain and the background corpora were pre-processed using the LingPipe API [114]:

- Texts were split into sentences using the *LingPipe* MedlineSentenceModel which is designed for scientific abstracts from the biomedical domain. While patent texts don't fit that description exactly, the sentence splitter worked well enough for our purposes. The only correction that proved necessary was the insertion of an artificial sentence boundary at the end of the patent titles (in the form of a period symbol) to prevent the sentence splitter from disregarding the final word of the title.

- Sentences were tokenized using LingPipe's IndoEuropeanTokenizerFactory.

- The resulting tokens were assigned part-of-speech tags according to the *LingPipe* MedPost HiddenMarkovModel [115]. As for the sentence splitter, this model was not originally intended to be used for patent texts, but the quality of the tag assignment was for the most part satisfactory.

- Based on the assigned tags, noun phrase candidates were extracted from the patent texts. In order to maximize precision, we restricted phrase extraction to the very reliable pattern *adjective* noun+*. MedPost uses separate tags for proper and plural nouns as well as comparative and superlative adjectives (Smith *et al.* give a complete list of tags in [115]); our term extraction algorithm treated proper nouns like regular nouns, while plural nouns were only allowed as the last word of a phrase candidate. Adjectives in comparative or superlative form were treated like regular adjectives. Since MedPost tends to assign the "noun" tag to single letters, single-word candidates containing less than three and multi-word candidates containing less than five characters were discarded, as were candidates containing punctuation.

All extracted noun phrases were collected in an index containing the total number of occurrences of the phrase in the corpus (term frequency) as well as the number of corpus documents containing the term (document frequency). The following section describes how the different ranking methods that we examined use this information to calculate the rankings.

6.3.2 Ranking methods

As we described in Section 5.1, we evaluated three of the most common statistical methods for ranking automatically extracted terms as well as two new variants that we are proposing. Our goal was to find the best method for ranking terms according to their "termhood", i.e. the degree to which "a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts" [116]. Usually, the termhood is measured by comparing term frequencies in a domain corpus to those in at least one more general background corpus.

The measure *tf-idf* is commonly used in information retrieval. Search results are intended to be improved by giving high ranks to documents that don't just contain the search terms, but are best represented by them compared to the other documents in the available collection. In order to determine the terms that represent a document, every term is assigned a weight that is proportional to its number of occurrences in the document, and inversely proportional to the number of background documents containing it. This simple idea forms the basis of a number of slightly different definitions of tf-idf; we follow the definition given by Manning *et al.* [107].

The *tf-idf* weight is calculated by multiplying two scores, the term frequency (tf) and the inverse document frequency (idf). The term frequency is defined simply as the number of occurrences of a term in the document. It can be normalized to a value between 0 and 1 by dividing the number

of occurrences by the total number of terms in the document. The document frequency is then defined as the number of documents that contain the term, and the inverse document frequency of the term t is given by

$$idf_t = \log \frac{N}{df_t},$$

where N represents the total number of documents.
The *tf-idf* weight of a term t in a document d is then defined as

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t.$$

In order to avoid division by zero for terms that don't occur in any corpus document, the document frequency is often increased by one for all terms.

It is not our objective to extract terms that represent a single document; we want to find terms that are representative of the whole domain. We therefore apply the *tf-idf* principle to our situation by treating the domain corpus like one (very large) document and the background corpus like the rest of the document collection. That means that in our case term frequency is calculated by adding up term frequencies from all domain documents, and document frequency corresponds to the number of background documents containing the term. For calculating the document frequencies, we consider the "domain document" a part of the background corpus; otherwise, domain terms that aren't included in any background document would lead to division by zero during the *idf* calculation.

Since *tf-idf* tends to over-represent *tf* to the detriment of *idf* (cf. [105, 106]), we also used the variant that Manning *et al.* call *wf-idf* in the same way [107]. It is calculated by replacing *tf* with

$$wf = \begin{cases} 1 + \log tf & tf > 0 \\ 0 & otherwise \end{cases},$$

and then multiplying by *idf* as before.

We also propose a new way of using *tf-idf* for extracting terms from a corpus instead of a single document, the measure we introduced in Section 5.1.2 under the name *majority tf-idf*. Our system applies the *tf-idf* method as defined above to each domain document, while still using the whole background corpus as a point of comparison. A pre-defined number of top-ranking terms in each document are assigned a number of points depending on their rank. These points are added up over all domain documents for all terms that have a high rank in any of them, resulting in

a new term ranking that favors terms that are very important in a comparably small number of documents over others that are somewhat important in many documents. Intuitively, this could be expected to lead to a more specialized ranking. There are a lot of possible choices for the allocation of points to top-ranking terms that may result in more or less specific terms. For our extraction experiments, we decided to concentrate just on the most important terms for each document. We therefore used the following simple point scheme: The first three terms in each ranking are assigned 3, 2 and 1 point(s) respectively. For *majority wf-idf*, we did the same calculation using *wf-idf* on the individual documents.

The third existing measure we used was the Log-Likelihood Ratio (LLR) that also preserves the corpus structure. As a consequence, LLR relies on the document frequencies in both corpora and doesn't take term frequency into account. It is the basis of a statistical test that is used to measure how well experimental data fits different hypotheses about the model the data originated from. Applied to our requirements, the document frequencies in the domain and background corpus of a specific term are the data, and the hypotheses are the following:

- H_0: Term t is not domain-specific; it is as likely to occur in the domain corpus as it is in the background corpus.
- H_1: Term t is domain-specific; it is *more or less* likely to occur in the domain corpus than in the background corpus.

As defined by Dunning, the value of LLR is then the logarithmic ratio of the maximum value of the likelihood function assuming H_0 and the maximum value of the likelihood function assuming H_1. We are using a binomial model for calculating the likelihood function with parameters p_1 and p_2 that represent the probability of a term occurring in the domain corpus and the background corpus respectively. If n_1, n_2 are the numbers of documents in the domain corpus and the background corpus, and k_1, k_2 are the corresponding numbers of documents containing the term, the likelihood function for a specific term can then be calculated as follows:

$$L(p_1, p_2) = {p_1}^{k_1}(1-p_1)^{n_1-k_1}\binom{n_1}{k_1}{p_2}^{k_2}(1-p_2)^{n_2-k_2}\binom{n_2}{k_2}$$

This likelihood function is maximized for the values $p_1 = \frac{k_1}{n_1}$ and $p_2 = \frac{k_2}{n_2}$. Note that H_0 assumes that the term is as likely to occur in the domain corpus as it is in the background corpus,

which means that the same value must be chosen for p_1 and p_2. The likelihood function is then maximized for the value $p_1 = p_2 = \frac{k_1+k_2}{n_1+n_2}$, and the ratio of both maxima is

$$\lambda = \frac{(\frac{k_1+k_2}{n_1+n_2})^{k_1+k_2}(1-\frac{k_1+k_2}{n_1+n_2})^{n_1+n_2-k_1-k_2}}{(\frac{k_1}{n_1})^{k_1}(1-\frac{k_1}{n_1})^{n_1-k_1}(\frac{k_2}{n_2})^{k_2}(1-\frac{k_2}{n_2})^{n_2-k_2}}$$

The logarithm is used to simplify the calculation, and terms are then sorted by their resulting *LLR* score to make the final ranking. However, hypothesis H_1 poses a problem: Since it does not distinguish between the corpora, terms that are more likely to occur in the domain corpus than in the background corpus can get the same score as terms for which the opposite is true. We are therefore following the solution that was proposed by Gelbukh *et al.* for this problem, by requiring term candidates to fulfill an additional condition [110]. In our case, only terms that satisfy the inequality

$$\frac{k_1}{n_1} > \frac{k_2}{n_2}$$

with k_i and n_i defined as above (k_1 and n_1 are the variables that belong to the domain corpus) are included in the ranking; all others are simply discarded. This means that even though hypothesis H_1 only states that the probabilities differ between both corpora, the terms that make up the final ranking are only the ones that are more common in the domain corpus.

We also considered using LLR for the "majority method" we defined above. But since this measure uses document frequencies instead of term frequencies for the domain corpus, applying it to just one document can not be expected to lead to good results - the document frequency would simply have the value one for every term candidate in the document, making the least common term that occurs in the document automatically its top term. Our expectations were confirmed by the resulting term lists; since their rankings were quite obviously much worse than all other results, we excluded them from the evaluation.

6.3.3 Additional parameters

As we described in detail in Sections 5.1.2.1, 5.1.2.3 and 6.3.2, we were mainly interested in examining the influence of different ranking measures and background corpora; we already discussed the results for these aspects in Section 5.1.2.4. However, we also investigated additional minor parameters that influence the ranking quality of our algorithm. This section describes these parameters, and Section 6.3.4 compares the results. We were planning to answer the following questions:

- Which fields of the patent are most useful for phrase extraction?
- Should the statistics also be updated for substrings of candidate terms?
- Which cutoff value should be used for the document frequency of candidate terms?

As we described in Section 2.3, the writing styles and therefore the vocabularies of most patents differ significantly between individual fields of the same patent. Consequently, it is not clear which fields of the patent should be used for the extraction of additional search terms. On the one hand, it seems likely that the relatively natural writing style of the abstract and description sections would be more suitable for this task than the often deliberately obfuscating language of the claims section. On the other hand, since the claims are legally the most important part of the patent, they should also not be disregarded. We examined the use of each of these fields individually, and we additionally tried using the patent titles or the complete patent text including all these fields (i.e., title, abstract, description and claims).

The next aspect that we investigated concerns the substrings of candidate terms. The tag pattern we used for the extraction leads to a high number of phrase candidates that contain other, shorter candidates. There are two straightforward ways of dealing with these cases:

1. Extract the longest possible sequence of words as well as all subsequences that fit the pattern.
2. Extract only the longest possible word sequence and ignore all subsequences.

Choosing one of these approaches can have fairly substantial consequences on the collected phrase statistics - and therefore on the resulting ranking. Very common and important domain phrases may be ranked much higher by the first method than by the second one if they occur mostly as substrings of more specific phrases. On the other hand, it can be argued that the second method gives more weight to the complete phrases included in the domain texts and lessens the risk of extracting overly general phrases.

The third additional parameter that we examined was the minimum document frequency a term was required to have in the corpus in order to be included in the final ranking. Since our ranking methods require a positive value for the document frequency, terms that occurred in the domain corpus but never in the background corpus were not included in the final rankings. This makes sense, because extremely rare terms often originate from typos. We evaluated three different

values (2, 5 and 10) that the document frequency of a term was required to surpass for a term to be included.

As we described in Section 6.3.1, we collected term frequencies and document frequencies for all extracted noun phrases in an index in order to calculate the rankings. Both of these statistics were collected for all individual combinations of the parameters we examined:

- background corpora: diagnostics, pharma, general patents
- patent fields: the complete patent texts as well as the individual fields title, abstract, description and claims
- candidate substrings: only count the longest candidate/also count its substrings
- minimum document frequency: values 2, 5 and 10.

This resulted in 78 different sets of *tf* and *df* statistics. (Since patent titles are usually fairly short, we only used the value 2 for the minimum document frequency.) After applying each of the five ranking methods - *tf-idf*, *wf-idf*, *LLR*, *majority-tf-idf* and *majority-wf-idf* - to each set of statistics, we retrieved 294 different term rankings (78 each for the existing measures, 30 each for our proposed measures that do not take the minimum document frequency into account). The complete set of term rankings is illustrated in Figure 30, where each square represents five term lists corresponding to the five ranking measures.

The colors in Figure 30 represent the different ranking measures we evaluated. Each square stands for the three best average scores that were retrieved for the respective set of parameters, with the top left triangle displaying the color of the best measure and the bottom right triangle representing the second and third. As an example, the square on the top left represents the term rankings that were retrieved from the patent titles with minimum document frequency 2, using the diagnostics corpus for the background and including substrings of term candidates into the statistics. For this choice of parameters, the best result was achieved from using the *LLR* measure, followed by *wf-idf* and *majority-wf-idf*. The figure gives some insight into how the combination of parameters affects the results. Interestingly, although the *wf-idf* measure was the best measure overall (cf. Section 5.1.2.4), its advantage comes mainly from the more general background corpora; it is in most cases surpassed by the *LLR* measure on the more specific diagnostics corpus. On the other hand, *LLR* is surprisingly weak when combined with the general patent corpus. For most

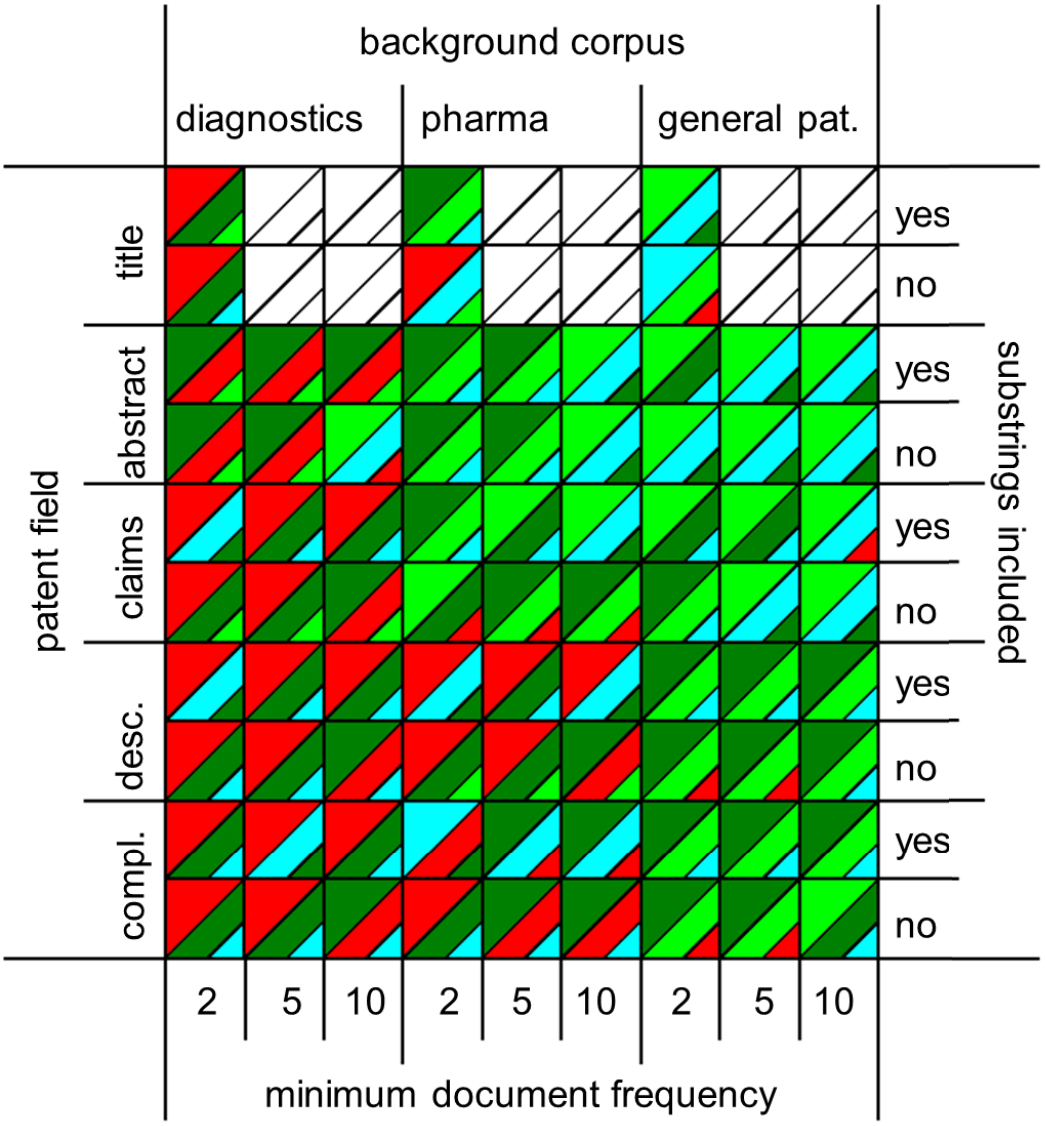


Figure 30: Overview of term extraction rankings. Each square corresponds to five different ranking lists, with the top three represented by color (red=LLR, dark/light green=wf-idf/majority-wf-idf, light blue = majority-tf-idf).

combinations of the three remaining parameters, it does not even manage to be among the best three of our five measures. The opposite is true for our two newly proposed measures, *majority-tf-idf* and *majority-wf-idf*. When the diagnostics corpus is used for the background, both measures are topped by *wf-idf* and *LLR* in almost all cases. However, the measures (especially *majority-wf-idf*) seem to be much better-suited for more general background corpora where they mostly outperform *LLR* and are often even ahead of *wf-idf*, at least for the patent fields title, abstract and claims.

One result that we already discussed in Section 5.1.2.4 is confirmed by the overview in Figure 30: the very weak overall performance of the regular *tf-idf* measure. There is not a single combination of parameters for which the *tf-idf* ranking even made it into the top three of the five measures.

An example for the term rankings resulting from different parameter combinations was given in Table 14 in Section 5.1.2.4. Both of the rankings in the table were using the abstract field of the patents, and the left ranking was retrieved using *tf-idf* with the general background corpus excluding substrings of term candidates and using the value 2 for the minimum document frequency, while the right term ranking resulted from using *wf-idf* with the diagnostics corpus including substrings and using the value 5. The following section describes our evaluation of these term rankings.

6.3.4 Evaluation

As we mentioned in Section 5.1.2.2, it was important to minimize the time investment required from the domain experts. We therefore restricted the evaluation to the top 50 terms of each list. Since there was a lot of overlap in the different ranking lists, we combined all 294 lists into one, thereby reducing the number of terms from almost 15000 partly identical terms on individual lists to slightly less than 600 terms on one list. This meant that the evaluation was a manageable task for the experts while still delivering meaningful information. In order to evaluate the individual parameters of the ranking algorithm that we wanted to evaluate, we separated the complete set of rankings into corresponding clusters and calculated the averages of the scores for the different ranks. Additionally, we calculated the "discounted cumulative gain" (DCG) of the individual rankings. DCG is frequently used in information retrieval to judge the quality of the ranked result sets of search engines. It is based on relevance scores for the results that are weighted according to their ranks. Formally, the DCG at rank $r \in \mathbb{N}$ is defined as follows:

$$DCG_r = score_1 + \sum_{i=2}^{r} \frac{score_i}{log_2(i)},$$

where $score_i$ is the average relevance score for the term at position i. Since our experts evaluated the first 50 terms from each list, we can calculate the DCG up to position 50.

In Section 5.1.2.4, we presented the results of our evaluation for the different ranking methods as well as the different background corpora. Figure 31 shows the same analysis for the different text fields of the patents. It shows that despite the different writing styles and vocabularies used in different parts of the patent, the term extraction still works best when the entire patent text

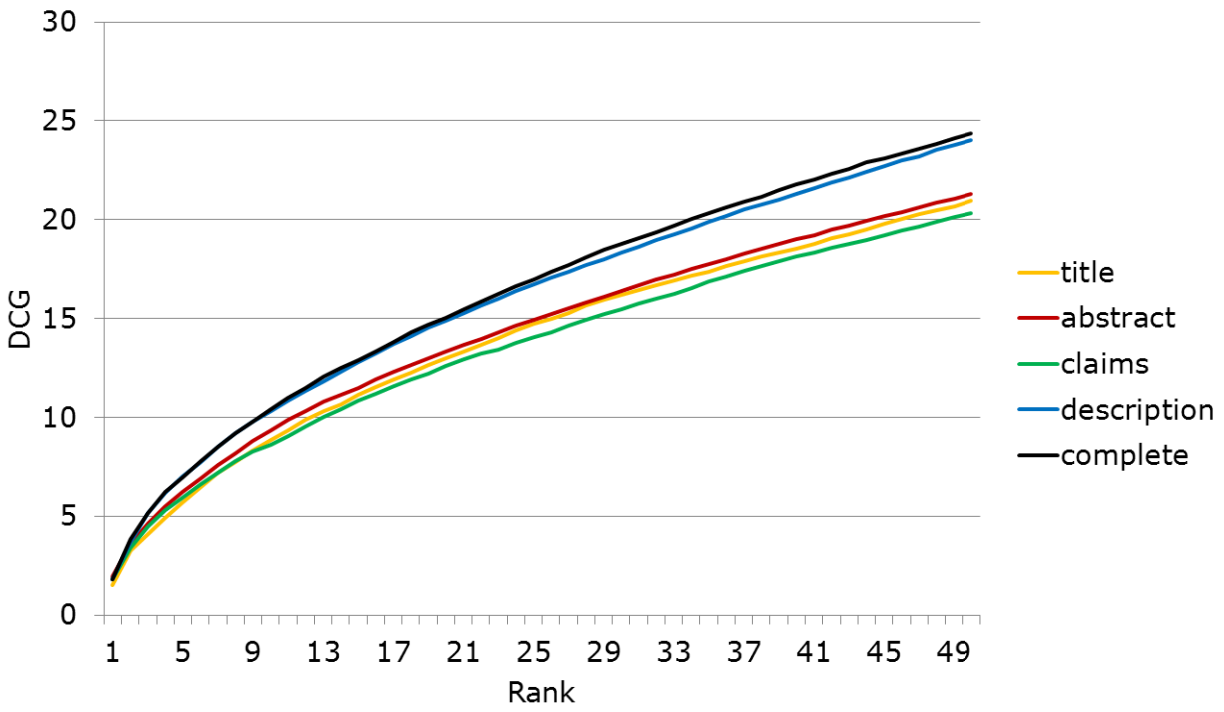


Figure 31: DCG values of term rankings extracted from different patent fields according to expert scores. The complete patent gives the best results, while the claims section leads to the worst term ranking.

is used. The description section of the patent is a close second, which is not overly surprising since it usually contains the largest part of the patent text. All other fields are doing considerably worse, with the abstract retrieving better results than the titles and the claims being the least useful source of terms. Since titles tend to be very short and often relatively vague, it has to be considered a surprise that the results outperform the much longer claims, however narrowly. On the other hand, this confirms our suspicion that the language of the claims section is not well suited for the extraction of terms.

The evaluation of the different approaches to collecting the substring statistics also shows a preference for one of the two approaches we examined. As the average scores presented in Figure 32 demonstrate, the inclusion of substrings of the extracted terms led to better results. This is presumably due to important domain phrases being ranked up when their frequent appearances in the patent texts as substrings of longer, more specific domain phrases are also taken into account when the frequency statistics are generated.

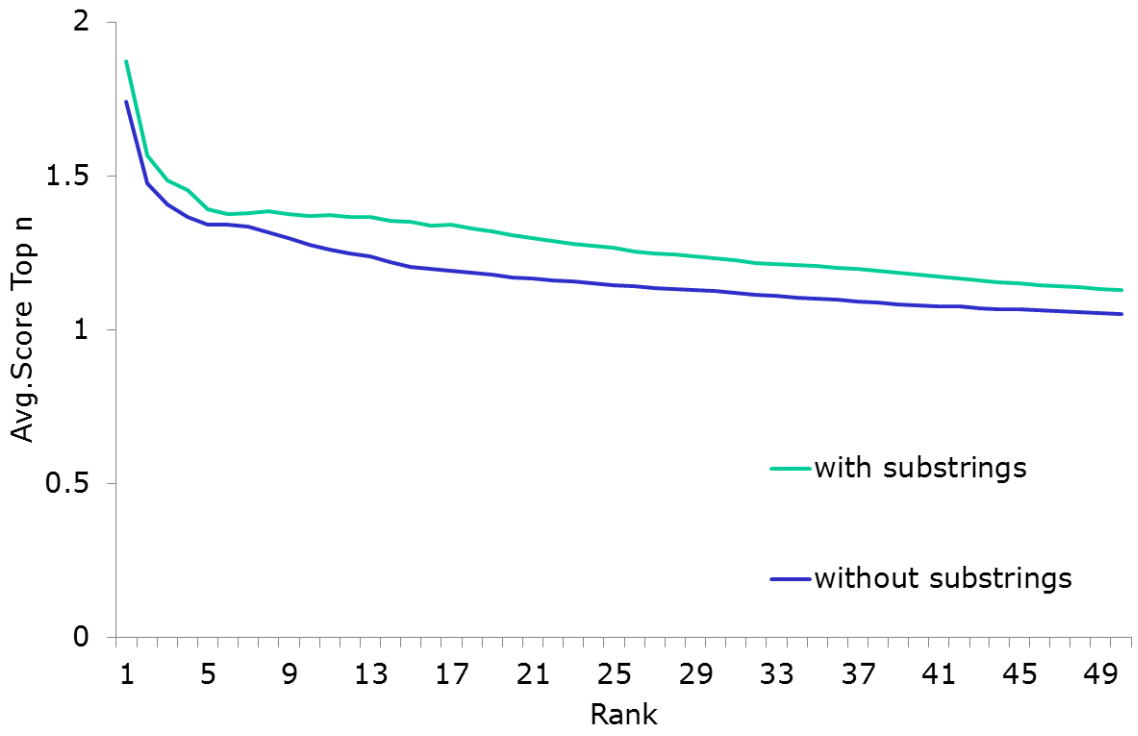


Figure 32: Average scores of term rankings using different methods for computing term statistics according to expert scores. The average scores are considerably higher when all substrings of term candidates also influence the frequency statistics than when they are discarded.

The last aspect of the algorithm on the other hand, the minimum required document frequency, did not show a clear preference for any of the values we examined. The average scores of all resulting term rankings were almost identical.

This chapter explained the technical details of our approaches to solving the different problems we investigated. In the following concluding chapter, we summarize these problems as well as our solutions.

7 Conclusions

This chapter summarizes our findings from the comparative analysis of MeSH and IPC in Chapter 3 and our scientific contributions to patent categorization and guided patent search that we developed in response to the problems that were revealed by our analysis.

7.1 Comparison of MeSH and IPC

The main goal of this work was finding ways to improve patent search for both unexperienced patent searchers and professional users. The large hierarchical systems that are used for patent classification have a lot of potential to be used for this task, since both groups of users can benefit: users unfamiliar with the classification systems can improve their search with automatically suggested classes relevant for their text query, and professional users can save a lot of time and effort that they would otherwise invest into collecting all relevant classes. Motivated by the advanced search functionality that is offered to PubMed users based on MeSH annotations, we conducted a detailed comparative analysis of MeSH and the IPC, which is the most widely used patent classification system internationally. Such an in-depth analysis has not been published previously, and it resulted in some important information that should be considered for the development of patent retrieval systems incorporating classification-based search. Our comparison concerned three different aspects of both systems:

1. the hierarchies:

   The hierarchies of both systems are relatively similar. The size of the IPC hierarchy surpasses that of MeSH, but they are in a similar range and the node distributions in both hierarchies are also comparable.

2. the terms:

   There are considerable differences with respect to the individual entries of both hierarchies. While MeSH emphasizes natural language terms, the IPC uses alphanumeric codes that are then explained by additional definitions. While MeSH terms are self-contained and often fairly short, these definitions are often quite long and rely on each other. Since many definitions are also fairly abstract and complicated, their literal occurrence in text is far less frequent than it is for MeSH terms, making the automatic annotation of IPC classes to patents more challenging.

3. the annotation of documents:

   Since the PubMed search improvements that this work was inspired by are made possible by using the MeSH annotations that were assigned to PubMed documents, this was the most important aspect of our analysis. We found that the number of IPC classes assigned to patents is considerably lower than the number of MeSH annotations to PubMed documents. Additionally, we showed that the IPC classes assigned to the same patent are often closely related, in many cases only covering one of the eight very general main sections of the classification system.

We concluded from these findings that patent search is severely limited. Unexperienced patent users usually lack the necessary knowledge to find the relevant classes, and the high complexity of the classification system forms a high barrier of entry. Additionally, as we explained in Sections 2.3 and 2.4, the exclusive use of keywords for search suffers from multiple problems: the language used in patents is very complex, the writing style changes both between different patents and between individual fields of the same patent, and standard vocabulary does not exist or is used inconsistently. Professional patent searchers know about these problems and include classification information in their search, but this approach also has some issues. The low average number of assigned IPC classes and their frequent close relatedness led us to believe that the class assignments performed at the patent office are not complete. This observation was also confirmed by multiple recent publications by professional patent searchers (e.g., [23, 26, 35]). As a consequence, the classification-based result filters that are often used in professional searches may be more restrictive than the users intend them to be, which limits the recall of these searches.

7.2 Automated patent categorization

We approached the problems that we discovered during our analysis of MeSH and IPC from two directions, the first one being the assignment of additional IPC classes to patents that already have a set of annotations from the patent office. For this task, we followed an approach that was used successfully by Tsatsaronis *et al.* for the assignment of MeSH terms to PubMed abstracts. We trained a large set of binary Maximum-Entropy classifiers (one for each class) and used the complete set of classifiers on each patent document in order to retrieve a set of classes that the document belongs to. Our experiments represent the largest-scale categorization effort yet that includes the lowest level of the IPC hierarchy.

Our results were for the most part promising, with precision, recall and F_1-measure values between 0.84 and 0.90 for individual classifiers and different test corpora. These values enable us to improve the recall for classification-based searches by including additional patents that were automatically assigned the query class. Since almost all patent queries also include keyword components, it is possible to limit the loss of precision that accompanies the gain in recall. We also demonstrated that the classifiers are able to choose word features that are relevant for the respective class and have a higher overlap with related classes than with randomly chosen classes. On the other hand, our precision values are not sufficient on their own for retrieving all relevant classes for a particular patent, since our method leads to a high number of incorrect assignments in this case. However, we proposed to solve this problem by using class co-occurrence data, and we showed that this approach can filter out many of these incorrect assignments without removing the correct ones.

7.3 Guided patent search

Apart from adding additional classes to patents, we approached the problem of lacking class assignments from a second direction: We investigated ways to expand the initial search query entered by the user. Since professional patent queries are usually composed of classes and keywords, we developed approaches to expand both components. We presented ways to extract additional keywords from three different sources: patent texts, IPC definitions and external knowledge sources.

In order to retrieve keywords and terms from class patents, we used existing term extraction techniques and investigated their effectiveness on patent corpora. Initial tests confirmed that the methods were applicable, although they also showed that there was still room for improvement, e.g., by using a patent corpus for the frequency comparison during the creation of the term statistics. Consequently, we performed in-depth experiments into how the parameters of the algorithm should be chosen to warrant the extraction of useful terms. The term lists that resulted from our large-scale comparison of five different aspects of the algorithm were manually evaluated by four experts, leading to results that clearly show strong advantages from choosing certain parameters correctly.

The two major insights we gained from these experiments concern the ranking measures and the background corpora. While the *tf-idf* measure is commonly used for the task of extracting relevant terms from text, our evaluation showed that its results are far from ideal. Both the closely related *wf-idf* measure and the well-known *LLR* measure resulted in considerably better term lists, and even two relatively simple new variants that we proposed clearly surpassed the *tf-idf* measure.

In order to find out how the background corpus should be chosen, we examined three different ones with varying degrees of relatedness to the source texts. The resulting term lists show a clear benefit from having related documents in the background corpus, as the results were clearly best for the most closely related corpus and worst for the most distant, general patent corpus. An interesting minor result of our expert evaluation was the fact that the description section of patents is most useful for extracting terms - only very narrowly surpassed by the complete patent - while the claims section is least useful.

All connections that are discovered between terms and classes can be repurposed for class suggestions as well, giving unexperienced users the option of simply entering relevant keywords and directly having the system suggest related patent classes to them. Additionally, keywords and terms that were found to be closely related through any of the above-mentioned sources can be suggested when one of the terms is entered. In order to also find connections between related classes, we retrieved class co-occurrence data from our patent corpus and showed that this information can uncover valuable information about related classes.

Lastly, we introduced the patent retrieval prototype GoPatents that we developed in cooperation with Jan Mönning to demonstrate some of our proposals. It is based on the semantic search engine GoPubMed and enables users to search a large EPO patent corpus using faceted browsing with IPC class codes as well as with terms from vocabularies such as MeSH and GeneOntology. While it has not yet reached the point at which it can be used productively, it is already able to demonstrate the benefits of incorporating classification information together with such term annotation systems to allow a new form of patent search.

List of Figures

List of Tables

References

1. Brügmann S: **PATEXPERT Project Deliverable 8.1 - State of the Art in Patent Processing** 2006.

2. Montecchi T, Russo D, Liu Y: **Searching in Cooperative Patent Classification: Comparison Between Keyword and Concept-based Search**. *Advanced Engineering Informatics* 2013.

3. Atkinson KH: **Toward a More Rational Patent Search Paradigm**. In *Proceedings of the 1st ACM workshop on Patent information retrieval*, PaIR '08, ACM 2008:37–40.

4. World Intellectual Property Organization: **World Intellectual Property Indicators - 2012 Edition** 2012.

5. Joho H, Azzopardi LA, Vanderbauwhede W: **A Survey of Patent Users: An Analysis of Tasks, Behavior, Search Functionality and System Requirements**. In *Proceedings of the third symposium on Information interaction in context*, IIiX '10, ACM 2010:13–24.

6. Azzopardi L, Joho H, Vanderbauwhede W: **A Survey on Patent Users: Search Behavior, Search Functionality and System Requirements**. *IRF Report* 2010, **1**:2010.

7. Yoo H, Ramanathan C, Barcelon-Yang C: **Intellectual Property Management of Biosequence Information from a Patent Searching Perspective**. *World Patent Information* 2005, **27**(3):203–211.

8. Iwayama M, Fujii A, Kando N, Marukawa Y: **Evaluating Patent Retrieval in the Third NTCIR Workshop**. *Information Processing & Management* 2006, **42**:207–221.

9. Foglia P: **Patentability Search Strategies and the Reformed IPC: A Patent Office Perspective**. *World Patent Information* 2007, **29**:33–53.

10. Goebel MW: **Mirror, Mirror on the Wall, Squishy and Soggy, 2 Nanos Tall: Strategies, Methods and Tools for Searching Homogeneous Catalysts – An EPO Perspective (Part 1. Introduction and Patents)**. *World Patent Information* 2010, **32**:39–52.

11. Paranjpe PP: **Patent Information and Search**. *DESIDOC Journal of Library & Information Technology* 2012, **32**(3).

12. Burhan M, Jain SK: **Tools for Search, Analysis and Management of Patent Portfolios**. *DESIDOC Journal of Library & Information Technology* 2012, **32**(3).

13. Fujii A, Iwayama M, Kando N: **Introduction to the Special Issue on Patent Processing**. *Information Processing & Management* 2007, **43**(5):1149–1153.

14. Alberts D, Yang C, Fobare-DePonio D, Koubek K, Robins S, Rodgers M, Simmons E, DeMarco D: **Introduction to Patent Searching**. *Current Challenges in Patent Information Retrieval* 2011, **29**:3–43.

15. van Staveren M: **Prior Art Searching on the Internet: Further Insights**. *World Patent Information* 2009, **31**:54–56.

16. Iwayama M, Fujii A, Kando N, Marukawa Y: **An Empirical Study on Retrieval Models for Different Document Genres: Patents and Newspaper Articles**. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '03, ACM 2003:251–258.

17. Parapatics P, Dittenbach M: **Patent Claim Decomposition for Improved Information Extraction**. In *Proceeding of the 2nd international workshop on Patent information retrieval*, PaIR '09, ACM 2009:33–36.

18. Shinmori A, Okumura M, Marukawa Y, Iwayama M: **Patent Claim Processing for Readability: Structure Analysis and Term Explanation**. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing - Volume 20* 2003:56–65.

19. Konishi K, Kitauchi K, Takaki T: **Invalidity Patent Search System of NTT DATA**. In *Proceedings of NTCIR-4 Workshop Meeting* 2004.

20. Mase H, Matsubayashi T, Ogawa Y, Iwayama M, Oshio T: **Proposal of Two-stage Patent Retrieval Method Considering the Claim Structure**. *ACM Transactions on Asian Language Information Processing* 2005, **4**(2):190–206.

21. Mahdabi P, Keikha M, Gerani S, Landoni M, Crestani F: **Building Queries for Prior-Art Search**. *Multidisciplinary Information Retrieval* 2011, :3–15.

22. Hu P, Huang M, Xu P, Li W, Usadi AK, Zhu X: **Finding Nuggets in IP Portfolios: Core Patent Mining through Textual Temporal Analysis**. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, ACM 2012:1819–1823.

23. Annies M: **Best Practice in Search and Analysis of Chemical Formulations: From Chemical Recipes to Complex Formulation Types and Dosage Forms**. *World Patent Information* 2012, **34**(3):206–212.

24. Vijvers W: **The International Patent Classification as a Search Tool**. *World Patent Information* 1990, **12**:26–30.

25. Adams S: **Using the International Patent Classification in an Online Environment**. *World Patent Information* 2000, **22**(4):291–300.

26. Wolter B: **It Takes All Kinds to Make a World – Some Thoughts on the Use of Classification in Patent Searching**. *World Patent Information* 2012, **34**:8–18.

27. Harris CG, Arens R, Srinivasan P: **Comparison of IPC and USPC Classification Systems in Patent Prior Art Searches**. In *Proceedings of the 3rd international workshop on Patent Information Retrieval*, ACM 2010:27–32.

28. **DEPATISnet - Hilfe - IPC und DEKLA**. *http://depatisnet.dpma.de/ipc/help.do*.

29. Goebel MW, Hintermaier FS: **Searcher's Little Helper - ICO Index Terms**. *World Patent Information* 2011, **33**(3):260–268.

30. Iwayama M, Fujii A, Kando N: **Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task**. In *Proceedings of NTCIR-5 Workshop Meeting* 2005.

31. Falasco L: **Bases of the United States Patent Classification**. *World Patent Information* 2002, **24**:31–33.

32. Falasco L: **United States Patent Classification: System Organization**. *World Patent Information* 2002, **24**(2):111–117.

33. Blinnikov V, Belov V, Makarov M: **Some Problems in the Use of the International Patent Classification**. *World Patent Information* 1984, **6**(2):63–68.

34. Becks D, Womser-Hacker C, Mandl T, Kölle R: **Patent Retrieval Experiments in the Context of the CLEF IP Track 2009**. In *Multilingual Information Access Evaluation I – Text Retrieval Experiments, Proceedings of CLEF 2009* 2010:491–496.

35. Parisi C, Rodríguez-Cerezo E, Thangaraj H: **Analysing Patent Landscapes in Plant Biotechnology and New Plant Breeding Techniques**. *Transgenic Research* 2013, **22**:15–29.

36. Emmerich C: **Comparing First Level Patent Data with Value-added Patent Information: A Case Study in the Pharmaceutical Field**. *World Patent Information* 2009, **31**(2):117–122.

37. Falciola L: **Searching Biotechnology Information: A Case Study**. *World Patent Information* 2009, **31**:36–47.

38. Dirnberger D: **A Guide to Efficient Keyword, Sequence and Classification Search Strategies for Biopharmaceutical Drug-centric Patent Landscape Searches - A Human Recombinant Insulin Patent Landscape Case Study**. *World Patent Information* 2011, **33**(2):128–143.

39. Doms A, Schroeder M: **GoPubMed: Exploring PubMed with the Gene Ontology**. *Nucleic Acids Research* 2005, **33**:783–786.

40. Doms A: **GoPubMed: Ontology-based Literature Search for the Life Sciences**. *PhD thesis*, PhD thesis, Technical University of Dresden 2009.

41. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P: **EBIMed—Text Crunching to Gather Facts for Proteins from Medline**. *Bioinformatics* 2007, **23**(2):237–244.

42. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A: **Text Processing through Web Services: Calling Whatizit**. *Bioinformatics* 2008, **24**(2):296–298.

43. Trieschnigg D, Pezik P, Lee V, de Jong F, Kraaij W, Rebholz-Schuhmann D: **MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval**. *Bioinformatics* 2009, **25**(11):1412 –1418.

44. Tsatsaronis G, Macari N, Torge S, Dietze H, Schroeder M: **A Maximum-Entropy Approach for Accurate Document Annotation in the Biomedical Domain**. *Journal of Biomedical Semantics* 2012, **3**:S2.

45. Gurulingappa H, Müller B, Klinger R, Mevissen H, Hofmann-Apitius M, Fluck J, Friedrich C: **Patent Retrieval in Chemistry based on Semantically Tagged Named Entities**. In *The Eighteenth Text RETrieval Conference (TREC 2009) Proceedings* 2009.

46. Coletti MH, Bleich HL: **Medical Subject Headings Used to Search the Biomedical Literature**. *Journal of the American Medical Informatics Association* 2001, **8**(4):317–323.

47. Darmoni SJ, Soualmia LF, Letord C, Jaulent MC, Griffon N, Thirion B, Névéol A: **Improving Information Retrieval using Medical Subject Headings Concepts: a Test Case on Rare and Chronic Diseases**. *Journal of the Medical Library Association : JMLA* 2012, **100**(3):176–183.

48. Makarov M: **The Process of Reforming the International Patent Classification**. *World Patent Information* 2004, **26**(2):137–141.

49. Adams S: **Comparing the IPC and the US Classification Systems for the Patent Searcher**. *World Patent Information* 2001, **23**:15–23.

50. Wongel H, Farassopoulos A: **Changes to the IPC Effective from January 2011**. *World Patent Information* 2012, **34**:4–7.

51. Krier M, Zacca F: **Automatic Categorisation Applications at the European Patent Office**. *World Patent Information* 2002, **24**(3):187–196.

52. Mooers CN: **The Next Twenty Years in Information Retrieval: Some Goals and Predictions**. In *Papers presented at the the March 3-5, 1959, western joint computer conference*, IRE-AIEE-ACM '59 (Western), New York, NY, USA: ACM 1959:81–86.

53. Larkey LS: **A Patent Search and Classification System**. In *Proceedings of the fourth ACM conference on Digital libraries* 1999:179–187.

54. Fall CJ, Benzineb K: **Literature Survey: Issues to be Considered in the Automatic Classification of Patents**. *World Intellectual Property Organization* 2002, **29**.

55. Fall CJ, Törcsvári A, Benzineb K, Karetka G: **Automated Categorization in the International Patent Classification**. *SIGIR Forum* 2003, **37**:10–25.

56. Fall C, Törcsvári A, Fiévet P, Karetka G: **Automated Categorization of German-language Patent Documents**. *Expert Systems with Applications* 2004, **26**(2):269–277.

57. Piroi F, Tait J: **CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain**. In *Proceedings of CLEF 2010* 2010.

58. Kando N: **Overview of the Sixth NTCIR Workshop**. In *Proceedings of 6th NTCIR Evaluation Workshop* 2007:1–9.

59. Kando N: **Overview of the Seventh NTCIR Workshop**. In *Proceedings of 7th NTCIR Evaluation Workshop* 2008:1–9.

60. Li Y, Bontcheva K, Cunningham H: **SVM Based Learning System for F-term Patent Classification**. In *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Crosslingual Information Access* 2007.

61. Rousu J, Saunders C, Szedmak S, Shawe-Taylor J: **Learning Hierarchical Multi-category Text Classification Models**. In *Proceedings of the 22nd international conference on Machine learning* 2005:744–751.

62. Cesa-Bianchi N, Gentile C, Zaniboni L: **Incremental Algorithms for Hierarchical Classification**. *The Journal of Machine Learning Research* 2006, **7**:31–54.

63. Trappey A, Hsu F, Trappey C, Lin C: **Development of a Patent Document Classification and Search Platform using a Back-Propagation Network**. *Expert Systems with Applications* 2006, **31**(4):755–765.

64. Tikk D, Biró G, Törcsvári A: **A Hierarchical Online Classifier for Patent Categorization**. *Emerging Technologies of Text Mining: Techniques and Applications. IGI Global* 2008.

65. Verberne S, Vogel M, D'hondt E: **Patent Classification Experiments with the Linguistic Classification System LCS**. In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2010), CLEF-IP Workshop* 2010.

66. Chen YL, Chang YC: **A Three-Phase Method for Patent Classification**. *Information Processing & Management* 2012, **48**(6):1017–1030.

67. Lai K, Wu S: **Using the Patent Co-citation Approach to Establish a New Patent Classification System**. *Information processing & management* 2005, **41**(2):313–330.

68. Loh HT, He C, Shen L: **Automatic Classification of Patent Documents for TRIZ Users**. *World Patent Information* 2006, **28**:6–13.

69. Attar R, Fraenkel AS: **Local Feedback in Full-Text Retrieval Systems**. *J. ACM* 1977, **24**(3):397–417.

70. Fujii A, Iwayama M, Kando N: **Overview of Patent Retrieval Task at NTCIR-4**. In *Working otes of the 4th TCIR Workshop* 2004:225–232.

71. Roda G, Tait J, Piroi F, Zenz V: **CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain**. *Multilingual Information Access Evaluation I. Text Retrieval Experiments* 2010, :385–409.

72. Lupu M, Hanbury A, Rauber A: **4th International Workshop on Patent Information Retrieval (PaIR'11)**. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, New York, NY, USA: ACM 2011:2623–2624.

73. Salton G: **A Document Retrieval System for Man-machine Interaction**. In *Proceedings of the 1964 19th ACM national conference*, ACM '64, New York, NY, USA: ACM 1964:122.301–122.3020.

74. Buckley C, Singhal A, Mitra M: **New Retrieval Approaches using SMART: TREC 4**. In *Text REtrieval Conference* 1996:25–48.

75. Osborn M, Strzalkowski T, Marinescu M: **Evaluating Document Retrieval in Patent Database: A Preliminary Report**. In *Proceedings of the sixth international conference on Information and knowledge management* 1997:216–221.

76. Takaki T, Fujii A, Ishikawa T: **Associative Document Retrieval by Query Subtopic Analysis and its Application to Invalidity Patent Search**. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* 2004:399–405.

77. Graf E, Azzopardi L, Rijsbergen K: **Automatically Generating Queries for Prior Art Search**. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, *Volume 6241*. Edited by Peters C, Nunzio GM, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G, Springer Berlin Heidelberg 2010:480–490.

78. Xue X, Croft WB: **Transforming Patents into Prior-art Queries**. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, ACM 2009:808–809.

79. Xue X, Croft WB: **Automatic Query Generation for Patent Search**. In *Proceeding of the 18th ACM conference on Information and knowledge management* 2009:2037–2040.

80. Magdy W, Jones G: **A Study on Query Expansion Methods for Patent Retrieval**. In *Proceeding of the 18th ACM conference on Information and knowledge management* 2011.

81. D'hondt E, Verberne S, Oostdijk N, Boves L: **Re-ranking based on Syntactic Dependencies in Prior-Art Retrieval**. In *Proceedings of the Dutch-Belgium Information Retrieval Workshop*, *Volume 2010* 2010.

82. D'hondt E, Verberne S, Alink W, Cornacchia R: **Combining Document Representations for Prior-art Retrieval** 2011.

83. Becks D, Eibl M, Jürgens J, Kürsten J, Wilhelm T, Womser-Hacker C: **Does Patent IR profit from Linguistics or Maximum Query Length?** 2011.

84. Kang IS, Na SH, Kim J, Lee JH: **Cluster-based Patent Retrieval**. *Information Processing & Management* 2007, **43**(5):1173–1182.

85. Gobeill J, Teodoro D, Pasche E, Ruch P: **Report on the TREC 2009 Experiments: Chemical IR Track**. In *the Eighteenth Text REtrieval Conference (TREC-18)* 2009.

86. Fujii A: **Enhancing Patent Retrieval by Citation Analysis**. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* 2007:793–794.

87. Lopez P, Romary L: **Multiple Retrieval Models and Regression Models for Prior Art Search**. *Arxiv preprint arXiv:0908.4413* 2009.

88. Lopez P, Romary L: **Experiments with Citation Mining and Key-term Extraction for Prior Art Search** 2010.

89. Magdy W, Jones G: **Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task** 2010.

90. Magdy W, Lopez P, Jones G: **Simple vs. Sophisticated Approaches for Patent Prior-art Search**. *Advances in Information Retrieval* 2011, :725–728.

91. Szarvas G, Herbert B, Gurevych I: **Prior Art Search using International Patent Classification Codes and All-Claims-Queries**. In *Working Notes of the 10th Workshop of the Cross Language Evaluation Forum* 2009.

92. Chen YL, Chiu YT: **An IPC-based Vector Space Model for Patent Retrieval**. *Information Processing & Management* 2011, **47**(3):309–322.

93. Chen YL, Chiu YT: **Vector Space Model for Patent Documents with Hierarchical Class Labels**. *Journal of Information Science* 2012.

94. Bashir S, Rauber A: **Improving Retrievability of Patents in Prior-art Search**. *Advances in Information Retrieval* 2010, :457–470.

95. Bache R: **Patent Retrieval - A Question of Access**. *World Patent Information* 2011.

96. Shah PK, Perez-Iratxeta C, Bork P, Andrade MA: **Information Extraction from Full Text Scientific Articles: Where are the Keywords?** *BMC Bioinformatics* 2003, **4**:20.

97. Schuemie M, Weeber M, Schijvenaars B, Van Mulligen E, Van Der Eijk C, Jelier R, Mons B, Kors J: **Distribution of Information in Biomedical Abstracts and Full-Text Publications**. *Bioinformatics* 2004, **20**(16):2597–2604.

98. Rosenfeld R: **A Maximum Entropy Approach to Adaptive Statistical Language Modelling**. *Computer Speech and Language* 1996, **10**(3):187–228.

99. Ratnaparkhi A, et al.: **A Maximum Entropy Model for Part-of-Speech Tagging**. In *Proceedings of the conference on empirical methods in natural language processing* 1996:133–142.

100. Nigam K, Lafferty J, McCallum A: **Using Maximum Entropy for Text Classification**. In *IJCAI-99 workshop on machine learning for information filtering* 1999:61–67.

101. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1977, **39**:1–38.

102. Hartley HO: **Maximum Likelihood Estimation from Incomplete Data**. *Biometrics* 1958, **14**(2):174–194.

103. Sundberg R: **Maximum Likelihood Theory for Incomplete Data from an Exponential Family**. *Scandinavian Journal of Statistics* 1974, :49–58.

104. Wächter T, Schroeder M: **Semi-automated Ontology Generation within OBO-Edit**. *Bioinformatics* 2010, **26**(12):i88.

105. Hisamitsu T, Tsujii J: **Measuring Term Representativeness**. *Extraction in the Web Era* 2003, :45–76.

106. Hisamitsu T, Niwa Y, Nishioka S, Sakurai H, Imaichi O, Iwayama M, Takano A: **Term Extraction Using a New Measure of Term Representativeness**. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition* 1999:475–484.

107. Manning CD, Raghavan P, Schütze H: *Introduction to Information Retrieval*. Cambridge University Press 2008.

108. Dunning T: **Accurate Methods for the Statistics of Surprise and Coincidence**. *Computational linguistics* 1993, **19**:61–74.

109. He T, Zhang X, Xinghuo Y: **An Approach to Automatically Constructing Domain Ontology**. In *Computational linguistics* 2006.

110. Gelbukh A, Sidorov G, Lavin-Villa E, Chanona-Hernandez L: **Automatic Term Extraction Using Log-likelihood based Comparison with General Reference Corpus**. *Natural Language Processing and Information Systems* 2010, :248–255.

111. Fabian G, Wächter T, Schroeder M: **Extending Ontologies by Finding Siblings Using Set Expansion Techniques**. *Bioinformatics* 2012, **28**(12):i292–i300.

112. Day-Richter J, Harris MA, Haendel M, Lewis S: **OBO-Edit — An Ontology Editor for Biologists**. *Bioinformatics* 2007, **23**(16):2198–2200.

113. McCallum AK: **MALLET: A Machine Learning for Language Toolkit** 2002. [http://mallet.cs.umass.edu].

114. Alias-i: **LingPipe 4.0.1**. *http://alias-i.com/lingpipe* 2008, [http://alias-i.com/lingpipe].

115. Smith L, Rindflesch T, Wilbur WJ: **MedPost: a Part-of-speech Tagger for BioMedical Text**. *Bioinformatics* 2004, **20**(14):2320–2321.

116. Kageura K, Umino B: **Methods of Automatic Term Recognition: A Review**. *Terminology* 1996, **3**(2):259–289.