Developing Semantic Search for the Patent Domain

Daniel Eisinger
Technische Universität
Dresden
BIOTEC
Tatzberg 47/49
01307 Dresden, Germany
daniel.eisinger@biotec.tu-dresden.de

Jan Mönnich
Technische Universität
Dresden
BIOTEC
Tatzberg 47/49
01307 Dresden, Germany
jan.moennich@biotec.tu-dresden.de

Michael Schroeder
Technische Universität
Dresden
BIOTEC
Tatzberg 47/49
01307 Dresden, Germany
ms@biotec.tu-dresden.de

ABSTRACT

The patent domain is a very important source of scientific information that is currently not used to its full potential. Issues such as high numbers of patents, complicated language style and inconsistently used vocabulary make the task of searching for relevant patents extremely complex. While this is already a problem for patent professionals who have to invest a lot of time and effort into their search, it is even more problematic for academic scientists with little experience in this domain.

Semantic search functionality has been demonstrated to provide large advantages for document search in other domains. As an example, the search engine GoPubMed offers advanced search functionality for the biomedical domain based on annotating documents with relevant concepts from various ontologies. In this paper, we report on our efforts to provide comparable advances for the patent domain. We introduce the patent search prototype GoPatents, and we describe the experiments that we performed during its development in the areas of term extraction, term and IPC class co-occurrence analysis, automated patent categorization, and automated annotation with ontology concepts.

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 [2]. Despite that potential, most academic researchers are to our knowledge not using patents, presumably due to the high complexity of the domain.

The number of patent applications continues to rise, reaching 2.35 million worldwide in 2012 alone [11] - only one year after surpassing two million for the first time ever in 2011. [10]. The number of patent grants is also at an all-time high, exceeding the one million mark for the first time in 2012 [11]. Additionally, the 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. 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" [1].

As a result of these problems, professional patent searches usually don't rely exclusively on keywords. The most important way to improve pure keyword searches is through the use of the classification information that is provided by the patent offices. 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 system that provides an easier option for scientists to perform high-quality patent searches and assists patent professionals in completing and refining their initial queries. In order to provide such as-

Published at Ceur-ws.org
Proceedings of the First International Workshop on Patent Mining and Its Applications (IPAMIN) 2014. Hildesheim. Oct. 7th. 2014.
At KONVENS'14, October 8–10, 2014, Hildesheim, Germany.

sistance, 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). Since the benefit of using existing annotations for semantic search has already been demonstrated in the biomedical domain, we use the controlled vocabulary "Medical Subject Headings" (MeSH) that is used to annotate all document abstracts on the biomedical literature database PubMed as a point of comparison. Following this analysis, we give a detailed description of multiple approaches we are proposing to improve patent search, and we introduce the patent retrieval prototype GoPatents that incorporates some of these proposals.

2. COMPARATIVE ANALYSIS OF MESH AND IPC

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 a large patent corpus as well as the annotations to all PubMed documents published by early 2011. Table 1 summarizes some core results of our analysis.

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
occurrence of class labels in text	frequent	very rare
average number of annotations per document	9	2
proportion of documents with multiple annotations	86%	53%
proportion of documents with related annotations (i.e., same hierarchy tree)	81%	46%

Table 1: 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.

The number of unique MeSH entries is considerably smaller than the number for IPC, but since the hierarchy tree of MeSH allows for the same heading to appear more than once, the sizes are comparable, as are the hierarchies (cf. Figure 1).

The comparison of the terms on the other hand shows some major differences. IPC is focused on alphanumeric 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. As Figure 2 shows, there are also large differences between the numbers of MeSH annotations per document and the numbers of IPC annotations per patent: While most patents have less than five assigned classes, PubMed documents have around nine on average and often even considerably more. Additionally, we were able to show that the annotation sets for patents are much less diverse than for PubMed, leading us to question the completeness of the existing assignments.

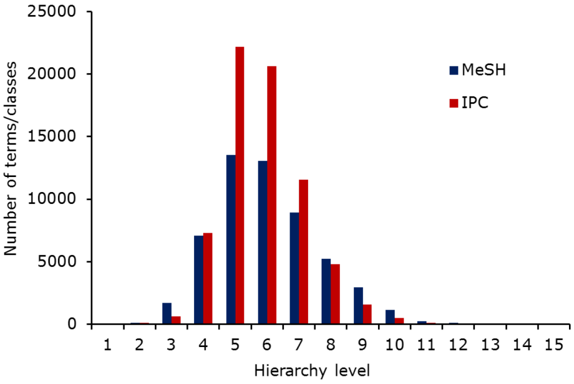


Figure 1: IPC vs. MeSH - Terms/classes per hierarchy level. Both hierarchies expand in similar ways.

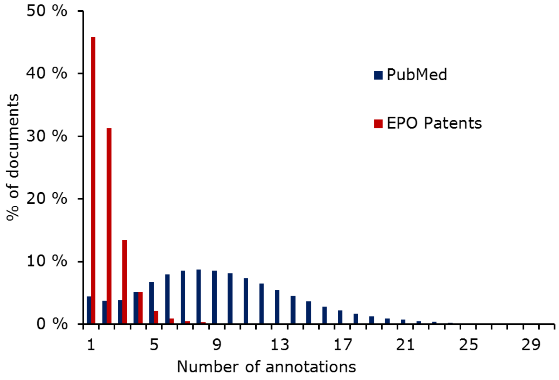


Figure 2: Percentage of documents with number of annotations. The average number of MeSH annotations per PubMed document is much higher than the number of IPC classes per patent.

We therefore believe that the use of IPC for patent search comes with two serious disadvantages: First, the complexity of the system causes significant problems for non-professional patent searchers since it is very difficult to find the complete set of IPC classes that are relevant for the search task at hand. Second, the low number of class assignments may lead to unexpectedly low recall for classification-based patent searches that are often performed by professional searchers in order to overcome the problems of keyword search.

3. SEMANTIC SEARCH FOR THE PATENT DOMAIN

This section describes our attempts to solve the problems caused for patent search by incomplete class assignments and complex patent text. We automatically assign additional classes, expand initial queries, and we annotate patent documents to make faceted search functionality possible.

3.1 Patent Categorization

The most straightforward way of dealing with the problem of incomplete class assignments 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.

Previous approaches to automated patent categorization were usually restricted to higher levels of the hierarchy (e.g., [7, 9, 6]). The only prior effort to classify patents down to the lowest level of the IPC involved a complicated three-phase algorithm that is not well suited for application on a large corpus; in addition, it already removes large parts of the hierarchy in the first step, which we believe makes it is too restricting for our goal of finding new relevant but potentially very different classes that were not previously assigned [3]. We therefore based our system on an approach that has been used successfully for the automated assignment of MeSH terms to PubMed documents by Tsatsaronis *et al.* [8]. It 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.

In order to evaluate the results of our categorization efforts, we constructed two training corpora from the EPO dataset that was also the basis of our previous analysis. The first corpus (C_{73}) follows strict quality requirements and contains 73 classes while the second one (C_{1205}) has more relaxed requirements and is therefore much larger with 1205 classes. 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}. 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.

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

Table 2: Evaluation results for confidence threshold 0.5. The precision values are identical for both corpora, but recall is considerably higher for the smaller corpus.

Table 2 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 results are for the most part encouraging, with most values approaching 0.9. For the purpose of our first task, this means that we can 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, a precision score of ≈ 0.9 leads to a high number of incorrect assignments. While using higher values for the confidence threshold has a positive effect on precision, it is accompanied by a severe drop in recall and therefore leads to a significantly lower F_1-measure. This problem is caused by slower precision growth for the individual classifiers compared to the situation for PubMed/MeSH, making additional steps necessary. We propose two filtering options: Since most patent queries also include a keyword component, many of the incorrect assignments are filtered out automatically since they don't include the required keywords. Additionally, 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) to much less so (allow pairs of classes if their respective ancestors of a certain hierarchy level have been co-assigned). For a small set of example patents, this filter had the desired effect of filtering out unrelated classes while accepting related ones.

IPC code	Class definition (abbrev.)	Features 1 to 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
G01N 33/50	Chemical analysis of biol. materials	sample, test, cell, specimen, light

Table 3: Most influential positive classifier features. Features were extracted from binary Maximum-Entropy classifiers trained on IPC classes with biomedical significance. The positive features for the classifiers in the list are useful for identifying patents that belong to the class.

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 3 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, there is very little overlap in the most influential features. As an example, the five top features are completely disjunct for class A61B 17/00 about surgical instruments and its descendant A61B 17/70 about spinal positioners.

3.2 Guided Patent Search

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. Since professional patent search queries are a combination of keywords and class codes in most cases, we investigated ways to expand both of these components. 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.

We demonstrated that additional relevant keywords can be extracted from a variety of sources including IPC class definitions and external resources such as MeSH. Most importantly, we extract keywords from existing patents using established natural language processing techniques after an initial evaluation showed the validity of this approach. Our method is based on analyzing patents from an IPC class that has been identified as relevant by the user. Since significant numbers of documents are available for most patent classes, this approach is able to deliver large numbers of keyword suggestions that are characteristic for the respective class. In a way, extracting relevant words from class patents is an expansion of our categorization efforts. Table 3 shows that this approach is able to discover useful keywords for search. Since we are also interested in relevant multi-word terms, we performed a more in-depth examination of different ranking algorithms for such extracted term candidates. Additionally, we investigated the influence of the background corpus on the result quality. The evaluation of the resulting term rankings was performed manually by four information professionals from the Scientific & Business Information Services department of Roche Diagnostics Penzberg. Interestingly, the experts disagreed often about the relevance of a term, indicating the high complexity of the problem.

We evaluated the established statistical term extraction measure *tf-idf* as well as previously published measures *wf-idf* and *Log-Likelihood Ratio (LLR)*, and we introduced two new variants of *tf-idf* and *wf-idf*. In order to judge the quality of the resulting term lists based on the scores given by our experts, we calculated different quality measures such as the average "discounted cumulative gain" (DCG) of the different rankings. Figure 3 shows clear differences between the ranking methods we investigated: The frequently used *tf-idf* measure was clearly the worst option for the task, 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 were also considerably better than *tf-idf*.

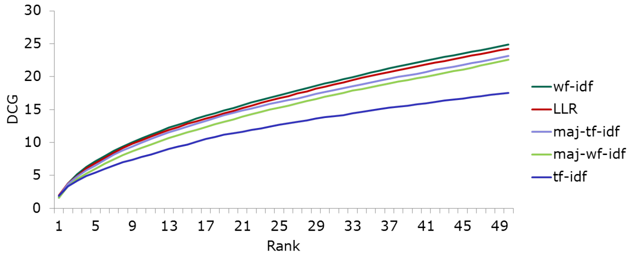


Figure 3: Influence of different ranking measures on the DCG value of extracted terms. Measure *wf-idf* performs best, followed by *LLR* and our proposed measures *majority-tf-idf* and *majority-wf-idf*. The DCG value is the lowest by far for *tf-idf*.

We experimented with background corpora that were either closely ("diagnostics") or distantly ("pharma") related to the class that we extracted the terms from, as well as a general corpus with no direct relation. Figure 4 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 are highest for the diagnostics background corpus, followed by the pharma corpus and the general corpus.

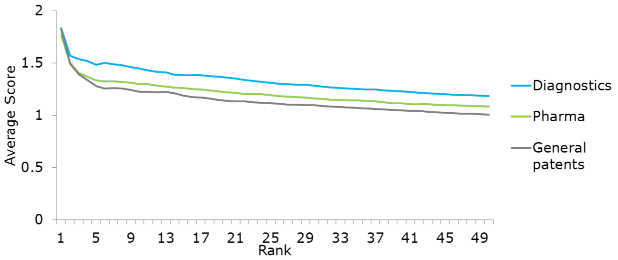


Figure 4: Influence of different background corpora on the average scores of extracted terms. On average, the extracted terms score highest with the closely related corpus (diagnostics) and lowest with the most distant corpus (general patents).

The identified terms that are relevant for certain classes can 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. 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.

Apart from mapping keywords to classes and vice versa as shown in the previous paragraphs, 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; 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 that many searches for one of the classes will also have additional relevant results in the second class. Figure 5 shows one example of such a pair of classes, including their definition hierarchy. Although the left class is clearly more application-oriented than the right one, we argue that many searchers interested in patents from one class will also find relevant patents in the other one. 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.

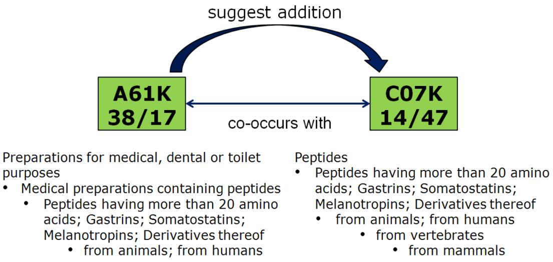

suggest addition
A61K 38/17
C07K 14/47
co-occurs with
Preparations for medical, dental or toilet purposes
• Medical preparations containing peptides
• Peptides having more than 20 amino acids; Gastrins; Somatostatins; Melanotropins; Derivatives thereof
• from animals; from humans
Peptides
• Peptides having more than 20 amino acids; Gastrins; Somatostatins; Melanotropins; Derivatives thereof
• from animals; from humans
• from vertebrates
• from mammals

Figure 5: Example for semantically related IPC classes without any hierarchical relation, detected using co-assignment information.

3.3 Annotation of Patent Documents with Gene/Protein Names

The biomedical search engine *GoPubMed*[1] offers its users faceted browsing of search results using the terms from Medical Subject Headings (MeSH) and Gene Ontology (GO) as well as a protein database. 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.

In order to provide patent searchers with similar functionality, we need a system that can annotate patent documents with the relevant concepts from the ontological resources we intend to use. The protein/gene annotator that is used for GoPubMed provides excellent performance for the types of text it was developed for, namely biomedical abstracts. Its quality has been demonstrated at the BioCreative workshop, where it was the best-performing system for the task of gene normalization [5]. However, due to the special properties of patent text it is by no means trivial to transfer existing text mining systems to the patent domain. We therefore developed a new version of the annotator for patent text, based on the original pipeline described in [4].

In order to help us test the performance of our new annotator, professional patent searchers collected a small set of patents related to neoplasms and made it available to us. The set consisted of 50 patents in total, including a large number of USPTO patents and smaller numbers of WIPO and EPO patents. A team of master students with expertise in the field manually listed all genes and proteins mentioned in the text. Our gold standard was then created in two further steps in a semi-automated fashion, by first matching these lists to the patent text automatically and then manually curating the result of this process.

In order to evaluate our new gene annotator for patent text, we used it to assign gene names to this manually annotated test corpus of neoplasm patents. The results showed a very large variation between individual patents, as had to be expected from the equally large variation of text styles and structures of the patents. On average, we reached a somewhat satisfactory precision of 0.75, while the recall still shows a lot of room for improvement at 0.39. These values correspond to an F_1 measure of 0.51. Although these results aren't nearly as good as the ones achieved by the original BioCreative annotator, we believe that they represent a promising starting point given the inherent complexity of the patent domain. We hope that an analysis of common annotation errors will help us further adapt the system to these special requirements, leading to clear improvements especially concerning the recall of the method. Further analysis of patents with particularly good or particularly bad annotation results may also help in this process. The current version of the annotator is however already able to provide clear improvements for patent search. In preparation for the patent search prototype GoPatents, it has been applied to an EPO corpus of 1.8 million patents, to which it assigned 157 million annotations. The complex and long texts also result in high processing requirements; assigning the annotations to the aforementioned EPO corpus took approximately 6000 CPU hours.

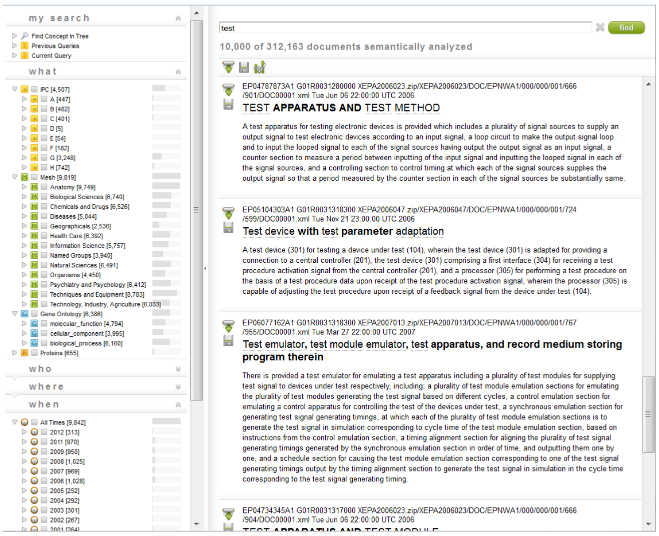
While our corpus cannot be considered a representative sample, our analysis of its documents led to some interesting observations. With the publication years of our patents spread between 2001 and 2011, we were able to observe a significant growth in the average number of annotations per patent beginning in 2006. The highest number of annotations to a single patent surpassed $2,500$ gene names. We hypothesize that the development and more wide-spread application of high-throughput techniques is at least partially responsible for this increase. We also kept track of which part of the patents individual annotations were assigned to. Unsurprisingly, the Description section was responsible for the largest number of annotations. However, a very large number of annotations is also contained in tables, which can cause problems for some automated extraction methods.

3.4 GoPatents - A Semantic Patent Search Prototype

In order to give a demonstration of some of our proposals, we implemented the patent retrieval prototype *GoPatents*

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

that enables the user to filter the resulting patent documents using terms from MeSH, Gene Ontology and a protein database. This functionality is brought over from GoPubMed, but we added the possibility of using 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 6, showing the following main components of the system:



test

10,000 of 312,163 documents semantically analyzed

EP04787873A1 G01R0031280000 XEPA2006023.zip/XEPA2006023/DOC/EPNWA1/000/000/001/666 /901/DOC00001.xml Tue Jun 06 22:00:00 UTC 2006

TEST **APPARATUS AND** TEST METHOD

A test apparatus for testing electronic devices is provided which includes a plurality of signal sources to supply an output signal to test electronic devices according to an input signal, a loop circuit to make the output signal loop and to input the looped signal to each of the signal sources having output the output signal as an input signal, a counter section to measure a period between inputting of the input signal and inputting the looped signal in each of the signal sources, and a controlling section to control timing at which each of the signal sources supplies the output signal so that a period measured by the counter section in each of the signal sources be substantially same.

EP05104303A1 G01R0031318300 XEPA2006047.zip/XEPA2006047/DOC/EPNWA1/000/000/001/724 /599/DOC00001.xml Tue Nov 21 23:00:00 UTC 2006

Test device **with** test **parameter** adaptation

A test device (301) for testing a device under test (104), wherein the test device (301) is adapted for providing a connection to a central controller (201), the test device (301) comprising a first interface (304) for receiving a test procedure activation signal from the central controller (201), and a processor (305) for performing a test procedure on the basis of a test procedure data upon receipt of the test procedure activation signal, wherein the processor (305) is capable of adjusting the test procedure upon receipt of a feedback signal from the device under test (104).

EP06077162A1 G01R0031318300 XEPA2007013.zip/XEPA2007013/DOC/EPNWA1/000/000/001/767 /955/DOC00001.xml Tue Mar 27 22:00:00 UTC 2007

Test emulator, test module emulator, test **apparatus, and record medium storing program therein**

There is provided a test emulator for emulating a test apparatus including a plurality of test modules for supplying test signal to devices under test respectively, including: a plurality of test module emulation sections for emulating the plurality of test modules generating the test signal based on different cycles, a control emulation section for emulating a control apparatus for controlling the test of the devices under test, a synchronous emulation section for generating test signal generating timings, at which each of the plurality of test module emulation sections is to generate the test signal in simulation corresponding to cycle time of the test module emulation section, based on instructions from the control emulation section, a timing alignment section for aligning the plurality of test signal generating timings generated by the synchronous emulation section in order of time, and outputting them one by one, and a schedule section for causing the test module emulation section corresponding to one of the test signal generating timings output by the timing alignment section to generate the test signal in simulation in the cycle time corresponding to the test signal generating timing.

EP04734345A1 G01R0031317000 XEPA2006023.zip/XEPA2006023/DOC/EPNWA1/000/000/001/666 /904/DOC00001.xml Tue Jun 06 22:00:00 UTC 2006

TEST **APPARATUS AND** TEST MODULE

Figure 6: 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 term hierarchies (left column, second from top)
  GoPatents enables the user to refine their search using relevant concepts from different sources. The complete hierarchies of all annotation systems we used are shown continuously with an indication of how many of the retrieved documents were annotated with it. The user can expand lower levels of the hierarchies for more precise information. 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 class.
- The additional filtering options (left column, third to fifth from top)
  GoPatents offers additional possibilities for faceted browsing: Search queries can be refined further to filter for specific applicants or publication dates.
- 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.
- The search results (main window, bottom)
  Snippets of the 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, providing links to the full patents.

In addition to the described functionality, the user's search history is made available, and the hierarchies can be searched for relevant concepts. Result statistics are calculated automatically and can be accessed instantly by the user as soon as the result set has been retrieved. These statistics cover multiple aspects of the result set, including the most frequently assigned terms from the different hierarchies (MeSH, GO and proteins), the most frequent patent classes and the top applicants.

4. CONCLUSION

We presented our approaches to some of the problems that have to be faced by patent searchers, e.g., complicated text, inconsistent vocabulary and incomplete class assignments. Our suggestions include the use of automated categorization for adding assignments and improving recall, different guided patent search strategies that help the user refine their queries, and the use of automated annotators to make faceted browsing possible in the patent domain. Our prototype GoPatents demonstrates some of the potential that semantic search can bring to the patent domain.

5. REFERENCES

[1] K. H. Atkinson. Toward a more rational patent search paradigm. In *Proceedings of the 1st ACM workshop on Patent information retrieval*, PaIR '08, pages 37–40. ACM, 2008.

[2] S. Brügmann. PATEXPERT project deliverable 8.1 - state of the art in patent processing, 2006.

[3] Y.-L. Chen and Y.-C. Chang. A three-phase method for patent classification. *Information Processing & Management*, 48(6):1017–1030, 2012.

[4] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez. Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, 24(16), 2008.

[5] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, et al. Overview of BioCreative II gene normalization. *Genome biology*, 9(Suppl 2):S3, 2008.

[6] D. Tikk, G. Biró, and A. Törcsvári. A hierarchical online classifier for patent categorization. *Emerging Technologies of Text Mining: Techniques and Applications. IGI Global*, 2008.

[7] A. Trappey, F. Hsu, C. Trappey, and C. Lin. Development of a patent document classification and search platform using a back-propagation network. *Expert Systems with Applications*, 31(4):755–765, 2006.

[8] G. Tsatsaronis, N. Macari, S. Torge, H. Dietze, and M. Schroeder. A maximum-entropy approach for accurate document annotation in the biomedical domain. *Journal of Biomedical Semantics*, 3:S2, 2012.

[9] S. Verberne, M. Vogel, and E. D'hondt. Patent classification experiments with the linguistic classification system LCS. In *Proceedings of CLEF 2010, CLEF-IP Workshop*, 2010.

[10] World Intellectual Property Organization. World intellectual property indicators - 2012 edition, 2012.

[11] World Intellectual Property Organization. World intellectual property indicators - 2013 edition, 2013.