Leibniz-Gemeinschaft

Abschließender Sachbericht

Trend Mining für die Wissenschaft

Leibniz-Einrichtung: FIZ Karlsruhe
Aktenzeichen: SAW-2013-FIZ_KA-2
Projektlaufzeit: 01.07.2013 – 30.06.2016
Ansprechpartner: Dr. Michael Schwantner

Projektname	**Trend Mining für die Wissenschaft**
Projektakronym	T4P
Aktenzeichen	SAW-2013-FIZ KA-2
Projektzeitraum	01.07.2013 – 30.06.2016
Förderlinie	2 – Besonders innovative und risikoreiche Vorhaben
Antragssteller	FIZ Karlsruhe Leibniz-Institut für Informationsinfrastruktur Hermann-von Helmholtz-Platz 1 76344 Eggenstein-Leopoldshafen
Projektpartner	Institut für Informationswissenschaft und Sprachtechnologie Universität Hildesheim Universitätsplatz 1 31141 Hildesheim
Publikation	Struß, Julia Maria; Mandl, Thomas; Schwantner, Michael; Womser-Hacker, Christa (2014): "Understanding Trends in the Patent Domain. User Perceptions on Trends and Trend Related Concepts." In: *Proc. of the First International Workshop on Patent Mining and its Applications.* KONVENS'14. IPaMin. Hildesheim, Oct. 7th. 2014.
Pressemitteilungen etc.	Erwähnung des Projekts in folgenden Artikeln: „Das große Graben" Leibniz-Journal 1/2014, S.23. „Auf der Jagd nach der Millionenprognose" manager magazin online, URL: http://www.manager-magazin.de/unternehmen/it/trendmining-mit-big-data-in-der-cloud-prognosen-erstellen-a-943926.html, 16. Januar 2014. "Schatzsuche in Patentdatenbanken" Stuttgarter Zeitung, Nr. 6, S. 18, 9.1.2014.

Dr. Michael Schwantner
FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Prof. Dr. Christa Womser-Hacker
Universität Hildesheim

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Stiftung Universität Hildesheim 2003

Abstract

As a result of the globalisation and the growing competitive pressure, the timely recognition of new trends and new developments in research is becoming increasingly important for the individual scientist as well as for institutes and research labs, be they academic or industrial. At the same time, however, this is becoming more and more difficult due to an increasingly fragmented research landscape and the still rapidly growing number of scientific publications. Since patents combine the characteristics of scientific articles, technical descriptions and legal texts, they represent a universal field of experimentation. Trend mining in patents is a new field of research, since until now trend mining methods had mainly been applied in more general areas and to documents with a simpler linguistic structure.

To understand the information needs and working environments of the target audience, and most importantly their understanding of trends and needs regarding the functionality of a trend mining system, a requirements analysis has been conducted in the form of a qualitative survey with scientists, who are working with patents, and information professionals from the patent domain.

Within the scope of this project, a prototype has been developed and evaluated which allows for (semi-automatic) tracking of trends on the basis of patents by methods independent from subject areas. For this purpose a corpus consisting of European 1,088,898 patents has been created. With the prototype we evaluated, adapted and developed semantic and statistical procedures that could be used to identify trends and their chronological development. Methods of topic detection and tracking were used which include linguistic approaches as well as similarity measures and clustering methods.

The evaluation of the prototype in the T4P project has been conducted at several steps throughout the development process. On the one hand, the performance and quality of the selected algorithms was evaluated by the use of a gold standard. On the other hand, a user-centred evaluation was conducted. For both types of evaluations, we gave a priority to the trend evaluations, as this was the main focus of the project.

Table of Content

1. Objectives of the Project

As a result of the globalisation and the growing competitive pressure, the timely recognition of new trends and new developments in research is becoming increasingly important for the individual scientist as well as for institutes and research labs, be they academic or industrial. At the same time, however, this is becoming more and more difficult due to an increasingly fragmented research landscape and the still rapidly growing number of scientific publications.

Within the scope of this project, a prototype is to be developed and evaluated which allows for (semi-automatic) tracking of trends on the basis of patents by methods independent from subject areas. Patents combine the characteristics of scientific articles, technical descriptions and legal texts, and thus represent a universal field of experimentation. Trend mining in patents is a new field of research, since until now trend mining methods had mainly been applied in more general areas and to documents with a simpler linguistic structure. So we will evaluate, adapt and develop semantic and statistical procedures that can be used to (semi-) automatically identify trends and their chronological development. For this purpose, we have to examine the development and changes of topic areas within a given set of documents. Methods of topic detection and tracking (TDT) will be used which include linguistic approaches, especially morphological analysis, as well as similarity measures and clustering methods.

2. Related Work

There are several papers that address the technical aspects of trend mining in the patent domain. Most of them (e.g., [1], [2] and [3]) focus on identifying technology trends retrospectively. Others consider related areas like the identification of patents with high novelty or deal with technology monitoring in patents (e.g., [4] and [5]).

Most publications address the problem of trend mining by means of machine learning techniques (e.g., [3], [6], [7], and [8]), particularly by using clustering techniques (e.g., [9], [10]) and network analysis (e.g., [3], [8], [9], [11], and [12]). Most works use different visualisation techniques to present the results (e.g., [13]) and leave it to the user to decide whether there really is a trend.

A wide range of features has been examined in these papers, for example terms that were selected based on their frequency, to mention the most common one (e.g., [10], [14]), adjective-noun pairs for potential technology features and verb-noun pairs for potential technology functions (e.g., [8]), noun and verb phrases (e.g., [2]), or subjective-action-object-relations ([3], [4]). Most of these works, however, do not present a sound evaluation of their approaches or only offer evaluations of selected steps from the complete process, due to missing evaluation resources. Most of the time case studies are performed instead. To our knowledge, there has not been any study on the understanding of trends and the informational background of the potential users of such a system so far.

3. Requirement Analysis

In order to provide a system that supports the target audience mentioned above in planning their research strategies through (semi-)automated trend detection, one needs to understand the information needs and working environments of these user groups, and most importantly their understanding of trends and requirements regarding the functionality of a trend mining system. The findings of a qualitative survey on this subject with both scientists, who are working with patents, and information professionals from the patent domain, are presented in this chapter.

3.1. Method

We are interested in gaining deeper insights in the users' understanding of trends as well as their requirements towards a trend mining system. Therefore, and due to the lack of prior studies in this area, we chose a qualitative approach and conducted semi-structured interviews.

In order to get a better idea of the working environment and the specific needs of information professionals in the patent domain, two pre-interviews were conducted with domain experts from a major information infrastructure institute working with patents and offering software products for information professionals in the patent domain. Due to these pre-interviews, the area of interest was narrowed down to the engineering sciences as patent documents in chemistry-related domains would have added the additional challenge of handling chemical notations.

Seven interviews were conducted subsequently. Three interview partners were scientists and four interview partners were information professionals, who either have a background as professional patent searchers, work in the IP management or work in a company offering different patent services to clients. The questions asked during the interviews were adapted to the respective target audience (scientists and information professionals) and the order of the questions could change during the interviews, depending on the individual development of the interview. The interviews were audio recorded and transcribed afterwards[1], before they were analysed.

3.2. Results

The analysis of the interviews shows numerous differences in the understanding of trends or the characteristics which make a trend interesting to the target audience, although the interview partners mostly had a rather homogeneous background in engineering. Essentially, two types of trends that are interesting to the target audience could be identified: Trends at the top level of an entire research area or domain and subject-specific or technical developments within a specific area of interest. The results also show that the time spans encompassing a trend can be quite different according to the content granularity of interest and the domain of interest. Additionally, not only emerging trends are of interest to the target

[1] One interview partner did not allow to audio record the interview, therefore the interview notes were used for further analysis.

audience, but also trends which have reached their height or even those that are declining, as this denotes that a technology has reached a stage where it can be used and licensed by other organisations to incorporate them in their own products. The interest on trends at this stage is mainly ascribed to SMEs.

The study also shows that research is needed with regard to the question of which content related sections of a patent are best applicable for trend mining, due to the fact that almost every content related section has been named by at least one interview partner. The findings show as well that at least for some of the patent searchers, it is important to integrate their customers and clients in the trend mining process. Therefore, a system with such a target audience should also incorporate visualisation techniques that allow for exploring analysis results together with clients and make it easy for a non-patent specialist to understand the results shown by the trend mining system.

For further results of the requirement analysis, please refer to [15].

4. Provisioning / preparation of the patent corpus

As part of its STN service, FIZ Karlsruhe offers the database EPFULL which contains all patent applications filed with the European Patent Office (EPO). These perfectly suit the purposes of our project since they encompass many different fields, for example chemistry, physics and mechanical engineering. Since patents use syntactically and semantically complex language patterns – the style of the detailed description section is similar to that of scientific publications, whereas the claims are heavily influenced by legalese – we can assume that methods which produce good results with a collection of patents will also prove successful when applied to other document types, in particular scientific publications.

At the time the corpus was constructed, EPFULL had 6,500,751 records from 1978 to 2014. However, the last two years, 2013 and 2014, are not complete, as the patent offices publish most patent applications only 18 months after they have been filed [16]. Besides bibliographic metadata such as application date, applicant name or inventor, the records comprise titles, descriptions, and claims in English, French or German (not all these parts exist for every record, though). FIZ Karlsruhe has converted the data into a proprietary XML format used for all databases offered in STN. After converting this XML format into a JSON format, we loaded the data into the Apache Lucene based ElasticSearch [17] engine.

Due to so-called patent families it is quite normal that patent databases contain records that are partial duplicates of each other. The very first application for a certain invention is at the same time the first member of a new family. The family grows if a patent is slightly reformulated during the patenting process, and the application and the granted patent become two separate, though almost identical, documents. The family also grows when the patent is applied for in more than one country. For trend mining purposes these duplicates are undesired since they distort the needed statistical data.

The concept of patent families is rather complicated, and there are several definitions. We used the definition for simple patent families [18] which considers all patents with identical

priorities (Priority Number (PRN) and Priority Date (PRD)) as members of the same patent family. In each family only the member with the largest set of English claims or the longest English description is then used for trend mining. During the preparation of the corpus these documents get special *family flags*. To avoid multiple variants during the granting process (so-called domestic families), we only used patents with kind code A1 or A2 (the original applications with or without search report).

The final corpus consists of 1,088,898 patents.

5. A prototype for trend mining on patent documents

In developing the prototype we followed the methods of user-centred design. The insights gained from the described user surveys (cf. Section 3) have been used to build a prototype which allows for trend mining on a preselected subset of the EPFULL database. It consists of state-of-the-art open source components and tools some of which were adapted to suit the peculiarities of patent documents. The following two subsections describe the components and the functionalities provided for user interaction.

5.1. Components of the prototype

5.1.1. Client-Server-Architecture

The prototypical application is implemented in a client-server architecture with a Java client for the user interaction. Data and business logic has been implemented as a Java Maven project by using Eclipse under the SUSE Linux Enterprise Server. The code was adapted to the MapReduce requirements in order to be runnable on a Hadoop cluster.

5.1.2. Retrieval

Usually, trend mining will not be done on the entire EPFULL database but on a defined subset which represents the subdomain the user is interested in and where he wants to look for trends. For this purpose, a search and retrieval component accepts a user query which will be expanded automatically to ensure that duplicates are filtered out. That is, the system modifies the query in such a way that it only retrieves English documents of kind code A1 or A2 and documents with family flags (cf. Section 4).

5.1.3. Pre-processing

From the answer set the features have to be extracted which can be used to identify the topics. Therefore, each document undergoes a sequence of linguistic and statistical processes. First, the document is split into sentences which are then split into tokens, i.e. single words. Based on a part-of-speech tagging, the noun phrases are extracted. Terms are stemmed and stop words are removed. The latter task was based on a standard list ([19], [20]) that had been extended with some frequent, patent specific expressions. The amount of terms is increased by extracting all possible n-grams, where a n-gram is a sequence of n consecutive words with n=1,2,3. Finally, all terms are filtered by their frequency in the answer set: terms with an absolute frequency less than 3 or a relative frequency higher than 0.3 are disregarded. We chose these parameters after previous tests with various values and the resulting clusters (cf. Section 6.1.2). We used the well-known OpenNLP [21] tools for sentence detection, PoS-

Tagging, and noun phrase extraction. For stemming the PlingStemmer [22] was applied and for tokenisation the built-in tool of Apache Lucene [23], the basis of ElasticSearch.

The next step is the computation of the document vectors. We use the standard tf-idf-measure [24] as vector elements for a term t in document d:

$$tfidf(t,d) = tf(t,d) * \log\frac{|D|}{1+h(t,D)}$$

$tf(t,d)$ is the frequency of t in d; $h(t,D)$ is the number of documents in D in which t occurs, and D denotes the complete document set.

5.1.4. Concept generation

The first step to identify trends is to identify all topics discussed in the individual patents. To do this, the prototype supports two approaches: topic modelling and document clustering.

For topic modelling, we used the Latent Dirichlet Allocation algorithm (LDA, [25]) to identify the *concepts* which could be the names of technologies, products, methods, materials, or services. The LDA model assigns to each concept a probability of occurrence in the document [26]. This means that the concepts do not divide the document set into disjunctive subsets since a document may be represented by two or more concepts. Only those concepts which score highest are used in the subsequent operations.

For document *clustering*, we investigated several approved methods: overlap clustering, hierarchical clustering, suffix tree clustering, x-means, and k-means [27] with cosine similarity metric. The latter proved to be the most suitable: it is more selective than the others and it is performant even for large datasets. To help the user to better conceive the meaning of each cluster, titles are generated for them by extracting the bigrams of the abstracts of all documents belonging to the cluster. After removing stop words, the bigram with the highest frequency is used as a cluster title.

A third type of concepts is represented by the *IPCs*. In the course of the application process the patent authorities assign one or more classification codes of the International Patent Classification (IPC) to each patent. The most descriptive IPC is marked as the main IPC and can be interpreted as the main topic of the patent.

5.1.5. Trend calculation

Each trend is based on a topic. In this project we followed an approach where a topic could be represented by either a concept derived from the LDA, a cluster, or an IPC. Concepts, clusters, or IPCs in turn are represented by the sets of documents in which they occur. The first step in trend analysis is to plot the number of documents for a topic along a timeline. A moving average calculated with five data points is added to the resulting graph. For better judgement whether a graph represents an upward or downward trend, a strong one or only an insignificant one, a statistical analysis is necessary. The prototype provides two functions for that purpose. The first one approximates the last n data points of the graph with a linear function $ax + b$. A second function adapts an exponential function ae^{bx} to the whole graph. In both cases the parameters a and b are calculated with the method of least squares. For the

linear function, the computation of a and b starts with $n = 5$ and n is incremented as long as the coefficient of determination R^2 is greater than 0.9. This approach is based on the requirement analysis (cf. Section 3.2), where experts stated that especially the last years are the most interesting. The exponential function is, in contrast, adapted to the whole data set. Here, only those topics which yield a positive trend ($b > 0$) are considered. For both types of functions, the absolute value of parameter b is an indicator of the strength of the trend – the greater the value, the stronger the trend.

5.2. User interaction

A typical user's workflow to identify trends can be subdivided into three steps: provisioning of the document set, getting an overview over the document set's topics, and studying the trends.

5.2.1. Search and retrieval

Only in the minority of cases will the user want to conduct a trend analysis on the complete document corpus which consists of all patent applications in Europe and covers subject areas from agriculture to x-rays. Normally, the user is interested in the trends and developments in a defined subject area. As a consequence, he first of all has to select the document set which covers the area of his interest as accurately as possible, e.g., via a state-of-the-art search. Our prototype provides the user with a retrieval component where he can enter a Boolean query according to the Apache Lucene Query Parser Syntax [28]. In addition to the operators AND, OR, AND NOT, he can narrow down his search on all necessary fields like

- title, abstract, description, or claims,
- inventor, applicants,
- time period,
- IPC.

With the exemplary query

```
AC:JP AND (TIEN:laser OR CLMEN:“laser beam“)
```

the user searches for all patents which were applied for in Japan and bear the word *laser* in the title or the phrase *laser beam* in the claims field. Since the retrieval functionality was not in the focus of this project, the prototype does not provide more sophisticated functions like ranking or query expansion. After submitting the query, the system retrieves the matching documents, analyses them, and makes the concepts and clusters (cf. Sections 5.1.3 and 5.1.4) available.

5.2.2. Examining the answer set

The user has now several functions at hand to survey the system's response. Figure 1 shows the user interface after the user has searched for the main IPC B64 which represents the class *Performing Operations; Transporting - Aircraft; Aviation; Cosmonautics*. The search field is in the upper left corner and the middle part lists number, title, assignee, and publication date of the patents found. If the user is interested in more details, he can have a look at a basic machine-generated summary (just the most significant phrases from the document) or at the

full text itself. In a more general way, he can create pie charts which will show him the distribution of inventors, assignees, countries, or IPCs.

The user can perform these tasks not only at this point of the workflow but also at any subsequent time, e.g., when he has had a look at the proposed trends (cf. Section 5.2.4).

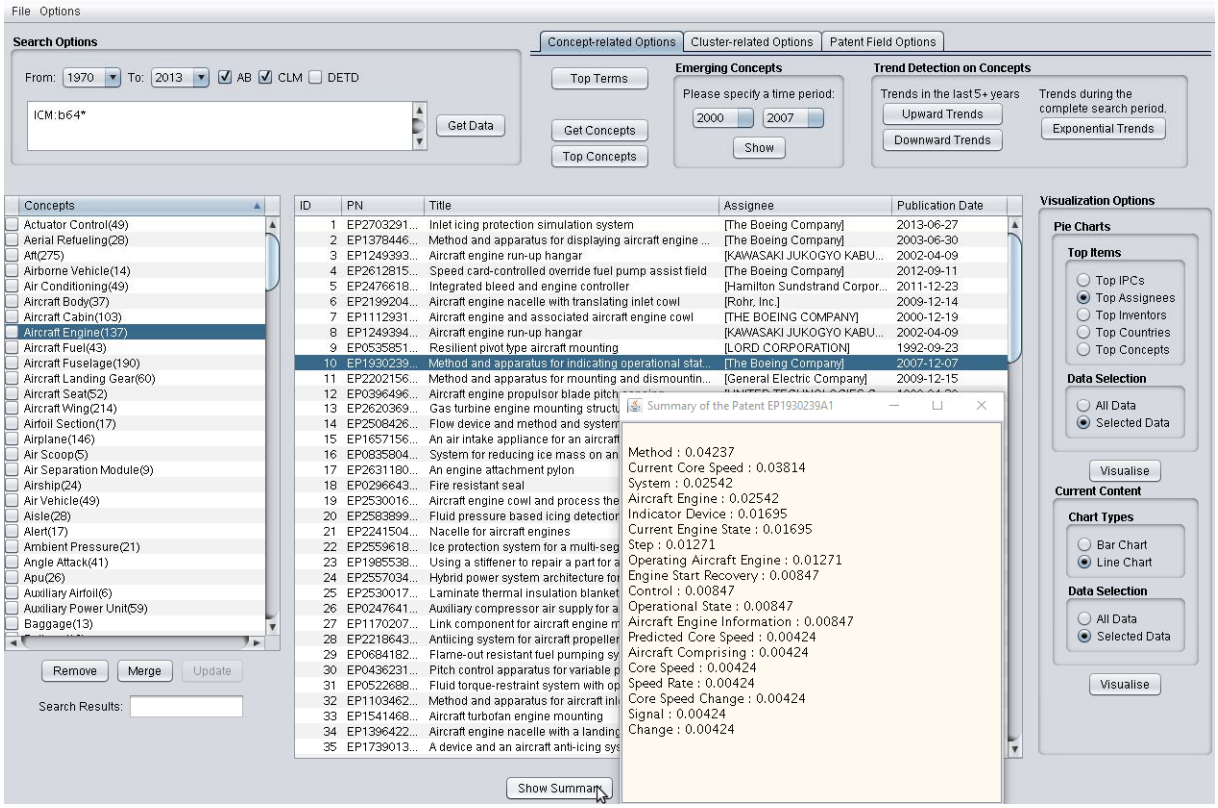

Figure 1: User interface with display of the answer set

5.2.3. Examining the topics

Independent of the type of topic – concept, cluster, or IPC – the user has always a direct access to the documents of the selected topic. This allows him to evaluate whether the system has built the topic consistently.

As for the concepts, pie charts allow a quick glance at the frequency distribution of the identified topics. The user can thus easily assess those which dominate the answer set or are worth examining in detail. In some cases the concepts proposed by the system may not be distinctive enough. Two or more concepts may be similar or, depending on the actual context, closely related. The user can correct this by arbitrarily merging them. The newly created concepts are then treated as normal concepts in the subsequent operations.

As a patent strives to specify novelties, a lack of proper words or phrases to describe this novelty will often be inevitable and this deficiency will lead to neologisms. It is therefore interesting to investigate these novel, *emerging* concepts. So there is a function which allows the user to select all concepts that had their first occurrence in a patent filed within a user-defined period of time.

The functionality for clusters is slightly different. In addition to the cluster title the most frequent terms are given to illustrate the cluster's content. If such a term or a whole cluster is

selected, the corresponding patents are shown together with their respective similarity score with the cluster's centroid.

5.2.4. Trend analysis

Basically, the user can display the timeline for each topic. In many cases he will see immediately whether the topic represents an interesting trend or not. The drawback is that this way he will have to check every topic – all concepts, clusters, and IPCs. The prototype alleviates this task by identifying those topics with a strong trend characteristic (cf. Section 5.1.5). On demand, the user gets a list of upward or downward trends ranked according to their strength. To assess these trend types, only the last years of the timeline are considered. This approach is based on the requirement analysis (cf. Section 3.2) where experts stated that especially the last years are the most interesting. Another trend type, the exponential trend, enables the user to identify trends that were strong in the past but in the meantime have declined and would therefore not be recognised if only the last years were examined.

Together with the graph the system displays the estimated parameters for the trend function and the coefficient of determination R^2. This allows for a better judgement on how well the calculated function replicates the data.

Figures 2 to 5 show the different steps of trend analysis a user can take. Figure 2 shows the timeline for the basic set, i.e. all patents with class B64 assigned. There is a slight growth until 1988 followed by a lateral movement until 2005 when a sudden increase can be observed. (The drop at the end of all the timelines is an artefact resulting from the incompleteness of the data, cf. Section 4).

Figure 3 shows the trend line for the concept "rotor blade". Obviously, the curve of this trend progresses quite differently from that of the basic set. Until 2006, only very few patent applications were registered, whereas the number of filings almost exploded in the years that followed. This is probably due to the literal rise of the drones, for which rotor blades are essential.

Figure 4 shows a downward trend for the cluster with the title *spacecraft body* and the terms *system spacecraft*, *momentum*, *thruster*, and *orbit*. Again, there is a striking difference to the basic set – and explicable perhaps by the NASA's shrinking budget.

The timeline in Figure 5 is for the IPC subgroup B64D 37/32 which is used to classify patents relating to safety measures, like preventing explosive conditions, in the context of aircrafts. While the linear analysis did not recognise a trend here, the moving average indicates a growing trend which is confirmed by the exponential trend line. Not very surprisingly, the growth for this topic started soon after 2001.

Figure 2

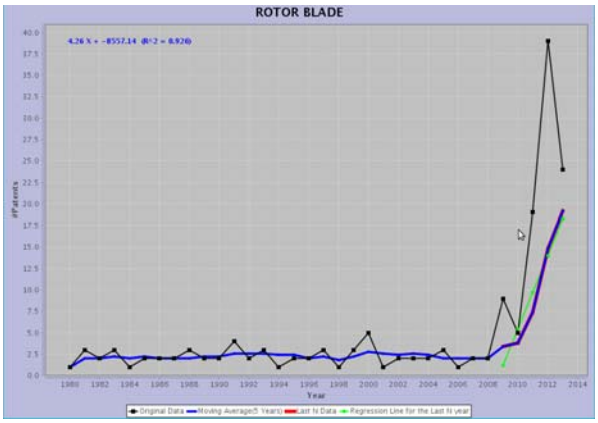

Figure 3

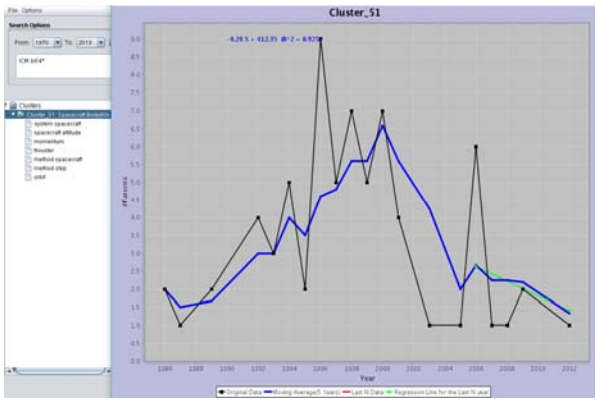

Figure 4

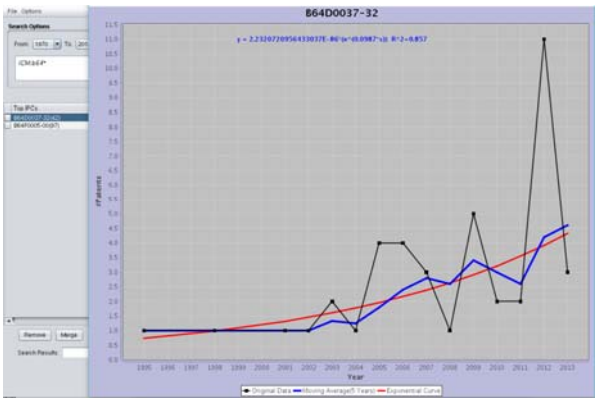

Figure 5

6. Concluding Evaluation

The evaluation of the prototype in the T4P project is conducted at several steps throughout the development process. On the one hand, the performance and quality of the selected algorithms is evaluated by the use of a gold standard. On the other hand, user-centred evaluation is conducted. For both types of evaluations, we give a priority to the trend evaluations, as this is the main focus of the project.

6.1. Gold Standard Evaluation

A common way to evaluate the quality of a system's output is to compare it with a given gold standard. Within this project, several intermediate steps will be evaluated which means that the quality of each part of the prototype can be observed. The prototype builds clusters in order to find themes or topics within a given data set. These clusters will then be analysed with regard to the time-related information given by the patents within each cluster in order to determine trends within the identified clusters.

A first step in the gold standard evaluation is therefore the evaluation of the clusters returned by the system (see Section 6.1.2). Only if the topics identified in this step match the topics given in the gold standard, a further analysis of the trends within these clusters is sensible. The details of the latter are explained in Section 6.1.3.

In order to conduct the evaluations mentioned above, a suitable gold standard is needed. To the best of our knowledge, no gold standard for the given problem of trend identification is freely available and therefore it is necessary to build it within the project.

6.1.1. Gold Standard

A gold standard consists of the artefacts (e.g., documents and human judgments on them) with respect to a specific problem. We therefore use publicly available reports generated by patent experts containing trend information. These reports are published on the website of the *World Intellectual Property Organization* (WIPO) [29]. The reports have been prepared to "provide a snap-shot of the patent situation of a specific technology, either within a given country or region, or globally. They can inform policy discussions, strategic research planning or technology transfer." [29]. They usually subdivide the selected technology and this information can be used to evaluate the clustering. Moreover, many – but not all – of the reports make statements about trends for the area in question. When using the reports, it has to be taken into account that these are based on another, larger data base, namely the applications at WIPO. We selected five reports (on 3D printing, contact lenses, robotic arms, fibre optic sensors, and slot machines) which are provided by *Gridlogics Technologies Pvt. Ltd*. and were partially generated by the use of *Patent iNSIGHT Pro*. The reports contain the queries that were used by the experts in order to filter the document set for further analysis; some of these reports also contain the queries for subtopics of the selected topic (e.g., the category or subtopic of *automobiles* for the topic *3D printing*) that were used for the categorisation of the topic. In order to build the gold standard, the queries were used to build a data set based on the EPFULL database. Therefore, the queries needed to be translated into JSON, the format used by ElasticSearch.

The reports further contain information about the number of patents within different categories of each criteria, such as *type* (e.g., type of contact lenses: soft contact lenses), *use scenarios, materials,* etc., which can be used to evaluate the clustering. The selected reports also provide information about the number of patents filed per year for the main topic and some of the subtopics.

6.1.2. Cluster Evaluation

The cluster evaluation is based on the data set generated by the queries in the reports. The clusters found by the system are then compared to the given subtopics (e.g., in the criterion of types and materials) of the reports. At this point it should be mentioned that this is hard for the k-means clustering, since the reports' subtopics overlap, whereas the k-means algorithms assort every patent only to one cluster. Having the users' requirements in mind and for the sake of a realistic evaluation scenario we still stick to this evaluation method.

There are several evaluation measures for granulated clusters like k-means clusters such as Dunn, purity, Mutual Information (MI), F-measure, Rand Index (RI) and Jaccard, although they are not necessarily in line with the purpose of overlapping classes. For this evaluation the focus is on Normalised Mutual Information (NMI). It allows comparing different clusterings as it represents a fixed interval (between 0 and 1) for each cluster and unlike the other measures, it is not influenced by the number or size of clusters. For more information about the clustering measure formulations see [30].

6.1.3. Trend Evaluation

The evaluation of the trends will also be based on the gold standard described above. The data sets generated by the queries in the reports are used as the starting point for the system to find trends. The suggested trends are then compared to the trend information given for subtopics within the WIPO reports. This kind of evaluation can be compared to classic information retrieval and therefore recall, precision, and f-measure are proposed as evaluation measures. One drawback of this approach, however, is that it applies only to reports providing explicit trend information.

As mentioned before, the results of the prototype are obtained through different methods. In this section, we depict the results of *contact lenses* at each section and analyse them. The contact lenses topic includes 267 patents, containing 17 (non-empty) categories of the *types* criterion, 22 categories for the *materials* criterion and 6 categories for the *uses* criterion. The time series of these categories are compared with the gold standard trend chart (in the report) and the quality of their potential trend are labelled as upward trend or downward trend. The labels are then ready to be compared to the prototype's output.

Concept-based Approach

In the concept-based approach, the prototype reveals two downward trends categorised as *Radius Curvature* and *Spherical (Contact Lenses)*. *Radius of Curvature* is a measure used to point out the quality of the convex surface of lenses. This optical specification is utilised throughout the manufacturing process of the product to control its quality according to standards. This is not a category regarding any of the available criterions while, logically, it can be considered a quality measure which represents the quality of spherical contact lenses. The *Spherical (Contact Lenses)* are a type of contact lenses that have been gaining attention from 2006 to 2009 according to EPFULL, but from 2009 the applications of patents in this field have dropped gradually. There is a drop from 2009 to 2012 in the EPFULL spherical contact lenses data, however, it still appears above the gold standard uptrend until the second half of 2011; this indicates that a relatively large amount of the entire applications from 2004 to 2012 have been submitted in recent years, though they are decreasing moderately. According to the contact lenses chart of the WIPO report, this downward move of the EPFULL curve cannot be assumed as a downtrend as a large part of it is situated above the curve in the gold standard chart and at some points the slope is better and at some points worse than it. From this, we cannot conclude whether the trend is generally a downtrend or an uptrend; therefore, both found downward trends by the prototype do not satisfy the gold standard. Results are reflected in Table 2.

The Prototype did not detect any uptrend in the EPFULL while the gold standard approves that there is actually no uptrend under the topic contact lenses. Under these circumstances, there is a positive null response which results in a precision and recall equal to 1. Out of 45 (non-empty) categories in three criteria, there was no uptrend and, despite a variety of categories, the prototype could successfully detect that none of them is actually an uptrend. This shows the ability of the prototype in identifying and ruling out the non-upward trends, however with this data; we cannot conclude anything about detecting the actual upward trends. These results are shown in Table 1.

Cluster-based Approach

The downtrend cluster detected by the prototype is completely covered by the following categories of the *materials* criteria*: Hydroxyethylmethacrylate, Lotrafilcon A, N Vinyl Pyrolidone, Phosphorylcholine, Polystyrene, Senofilcon A*. None of these categories are downward trends according to the gold standard. In the *types* criterion, more than half of this cluster overlaps *Hybrid Contact Lens* (61%) and *Soft Contact Lens* (77%). Both of these categories of the types are downward trends in the gold standard. In the *uses* criterion, the maximum coverage which is *Hyperopia* with 100% overlap, is not known as any type of trend in the gold standard. The other two categories with more than half of overlap are *Astigmatism* and *Myopia*. *Myopia* is a downtrend in the gold standard. Considering the subtopics with the most coverage by the detected downtrend cluster as its representatives, we observe the prototype detects downtrends relatively good for the *types* and *uses* criterion though it fails to detect downtrends in the material criterion (see Table 2). The prototype could also rule out the existence of any uptrend for the selected topic, the f-measure=1 reflecting the positive null response of uptrend detection is shown in the corresponding column of cluster-based approach in Table 1.

IPC-based Approach

The prototype detects two upward trends based on the IPCs of the patents. Both of these fall in the category "no trend" as they both are ascending and descending at different intervals while falling below and rising above the gold standard uptrend in the report. Moreover, the number of the patent applications is decreasing in the recent years, and certainly this cannot be a sign of an uptrend. This approach failed completely to detect the downtrends, while it also found two uptrends under a topic without any uptrend (see Table 1 and Table 2) which is a sign of poor performance in that regard.

Criterion	**Concept**	**Cluster**	**IPC**
Types	1	1	0
Materials	1	1	0
Uses	1	1	0

Table 1: F-measure of different criterion of the topic contact lenses for the prototype's uptrend detection (here the f-measures=1 reflects the positive null response)

Criterion	**Concept**	**Cluster**	**IPC**
Types	0	0.57	0
Materials	0	0	0
Uses	0	0.67	0

Table 2: F-measure of different criterion of the topic contact lenses for the prototype's downtrend detection

6.2. User-Centred Evaluation

The user-centred evaluation mainly focuses on the results the prototype delivers, meaning the automatically suggested trends and the corresponding judgements by patent experts. This evaluation can be regarded as a classical Interactive Information Retrieval (IIR) evaluation, where a user interacts with a system and judges whether the presented results are relevant or not. Relevance has multiple facets in this context:

- A presented trend must be relevant to the topic the user is interested in.

- The identified trends are correct, meaning they really present a trend.
- The type of trend (increasing, decreasing, …) is correct.

In order to be able to evaluate the correctness of a trend presented by the system, the study participants need to be experts in the corresponding domain and also in the patent field, as a trend shown in patents could likely show a different development (starting time, time-span etc.) in other publication areas. In order to find study participants with this kind of expertise, the recruitment is performed at a conference in the patent domain (PatInfo 2016 – the 38th Colloquium of the Ilmenau University of Technology on Patent Information). The participants of such a conference are experts in the patent domain and are usually specialised on one or two research areas (see interviews conducted at the beginning of the project).

To be able to evaluate the outcome of the prototype in an interactive user test, the key requirement is that the response time of the prototype is as quick as possible (if possible real-time response). Since the processing of topics with a lot of patent documents takes some time, the users are asked to name a topic in their area of expertise during the recruitment. The topic is then processed prior to the actual user test and can be loaded to a local computer during the test. That way the almost real-time response can be achieved.

6.2.1. Structure of the user test

At the beginning of the user test, the users are asked to answer some questions regarding their background, like their field of expertise, how long they are working in the patent domain, how often they are presented with trend searches during their everyday work life, what kind of steps they usually take for identifying trends in their domain etc. Afterwards a short presentation of the prototype is given, so the user can familiarise himself/herself with the functions and the user interface. The user is also allowed to ask questions during the prototype presentation and give comments about the presented functionality and the prototype itself.

In the third step, the study participant is asked to formulate an information need and conduct a trend search with the prototype. The participant is allowed to ask questions about the usage of the prototype, since the short intro in the beginning might not be sufficient to be familiar enough with the system to use it on his own.

Once the participant receives the trend suggestions by the system, he is asked to give detailed relevance judgements about each of the presented trends, meaning that he is asked to answer questions about each of the above listed facets regarding the relevance of a trend.

6.2.2. Results

Three participants could be recruited at the PatInfo, two coming from the industry and one working at the EPO. The topics covered by these patent experts are very different from each other and show quite some differences in the document collection size of the topic (around 500 up to almost 4,000 patents). These topics can therefore help to judge the performance of the system in different scientific areas which naturally have different volumes of publications. The first participant is an expert in the area of fire-extinguishing systems (IPC: A62C*), the second on lighting (IPC: F21*) and the third on Nano-technology (IPC: B82Y*). An overview of the results can be found in Table 3.

	Trend type	Cluster-based Trends	Concept-based Trends	IPC-based Trends
Topic 1	Upward	2 trends suggested, both correct	3 trends suggested - 1 correct - 1 too general - 1 not relevant to the topic	3 trends suggested - 2 correct - 1 relevant to the topic, but no trend
	Downward	none found	none found	none found
Topic 2	Upward	6 trends suggested, - 2 trends correct - 1 not relevant anymore/ downward - 1 cluster build around a common term, but is no trend - 2 clusters too general	27 trends suggested - 1 correct - 1 is a relevant topic, but no trend - other concepts were too general or irrelevant	not interesting for the interview partner
	Downward	none found	none found	
Topic 3	Upward	2 upward trends suggested, both not homogeneous	2 suggested - 1 is a correct trend - 1 too general and ambiguous	6 trends suggested - 3 are relevant to the topic and present a trend - 1 cannot be judged by the expert - some expected IPCs are missing
	Downward	1 suggested and likely correct	none found	none judged

Table 3: Details on the judgement of the trends suggested by the system

If, interpreting these results very strictly, only those trends designated as "correct" are regarded as "true positives" and all the other as “false positives”, the following results can be derived from the above table for the precision.

	Trend type	Cluster-based Trends	Concept-based Trends	IPC-based Trends
Topic 1	Upward	2 / 2 = 1.00	1 / 3 = 0.33	2 / 3 = 0.67
	Downward			
Topic 2	Upward	2 / 6 = 0.33	1 / 27 = 0.04	
	Downward			
Topic 3	Upward	0 / 2 = 0.00	1 / 2 = 0.50	3 / 6 = 0.50
	Downward	1 / 1 = 1.00		
All Topics, all types		5 / 11 = 0.45	3 / 32 = 0.09	5 / 9 = 0.56

Table 4: Precision for three topics

Trends calculated on the basis of the IPCs performed best. This is not surprising since the IPCs are assigned manually and thus an error source is excluded. That the results are not better could be explained by an insufficient selectivity of the IPCs. The disappointing values for the concept-based trends are mainly caused by too broad concepts, especially with topic 2. The results for clusters are heterogeneous. After all, only with clusters downward trends were found.

Statements about the recall cannot be made since this would have meant to manually determine beforehand for each topic all existing trends. This would have required a much greater effort on the part of the experts.

7. Conclusions and Outlook

Particular problems encountered during the evaluation were the lack of a real gold standard and the low number of available experts. The WIPO reports could only serve as a makeshift solution, since in most cases the description of trends had not been their central objective and

only a minority provided useful statements about trends at all. As a result, only five topics were available for evaluation, a number too small to be reliable. Moreover, the WIPO reports are based on the WIPO data which are more extensive than the EPO data and have a broader geographical distribution. Therefore, one has to expect that these two data sets may result in different trends for the same topic. Since in the field of intellectual property management the experts are often highly specialised in their subject area and in the nature of their activities, the number of experts and interviews was too small to be truly representative.

Nevertheless, the evaluation provided many valuable indications for improvements. To begin with, the extracted concepts are often very general and not suitable as a topic for a trend. The LDA algorithm used should at least be supplemented with more elaborate filtering methods.

One drawback of the k-means clustering is the fixed number of non-overlapping clusters. However, it is often the case that a document belongs to several topics and therefore to several trends. In addition, some clusters often contained only a small amount of patents. Using a more complex feature vector reflecting more detailed document properties as input to the clustering algorithm might be worth studying. More fine-tuning is also needed for the naming of the clusters. Too often they are nondescript.

The results based on the IPCs were not as good as expected, but during the interviews one expert proposed to base trends on the frequency of those IPC *pairs* which occur for the first time within a given time span. This would be analogous to the emerging trends the prototype calculates. Obviously, an IPC pair represents a more specific topic than a single IPC and one drawback of IPC topics was their lack of specificity.

Finally, the ranking of the trends could be enhanced by taking into account the difference between the trend found and the base set: If the timeline of a topic only slightly differs from that of its base set, the trend is not really remarkable.

A doctoral study started at the Dept. of Information Science and Natural Language Processing of the Hildesheim University addresses some of the above mentioned problems. This is all the more worthwhile, as FIZ Karlsruhe’s product management considers the functionality realised with the prototype very promising.

8. References

All URLs have been checked in December 2016.

[1] B. Yoon and Y. Park, "A Systematic Approach for Identifying Technology Opportunities: Keyword-based Morphology Analysis", *Technological Forecasting and Social Change 72 (2),* pp. 145–160, 2005.

[2] Y. Kim, Y. Tian, Y. Jeong, R. Jihee and S.-H. Myaeng, "Automatic Discovery of Technology Trends from Patent Text", *Proceedings of the 2009 ACM Symposium on Applied Computing (SAC), Honolulu, Hawaii, USA,* pp. 1480-1487, 2009.

[3] S. Choi, J. Yoon, K. Kim, J. Y. Lee and C.-H. Kim, "SAO Network Analysis of Patents for Technology Trends Identification: a Case Study of Polymer Electrolyte Membrane Technology in Proton Exchange Membrane Fuel Cells", *Scientometrics, Vol. 88(3),* pp. 863-883, 2011.

[4] J. M. Gerken, "PatMining – Wege zur Erschließung textueller Patentinformationen für das Technologie-Monitoring", Dissertation, Universität Bremen, Bremen. Available online at http://elib.suub.uni-bremen.de/edocs/00, 2012.

[5] P. Hu, M. Huang, P. Xu, W. Li, A. K. Usadi and X. Zhu, "Finding Nuggets in IP Portfolios. Core Patent Mining Through Textual Temporal Analysis.", *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM, Maui, Hawaii, USA,* pp. 1819–1823, 2012.

[6] W. M. Pottenger and T.-H. Yang, "Detecting Emerging Concepts in Textual Data Mining", *Computational Information Retrieval,* pp. 89–105, 2001.

[7] M.-J. Shih, D.-R. Liu and M.-L. Hsu, "Mining Changes in Patent Trends for Competitive Intelligence", *Advances in Knowledge Discovery and Data Mining. 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan. Lecture Notes in Computer Science, 5012,* p. 999–1005, 2008.

[8] J. Yoon, S. Choi and K. Kim, "Invention Property-function Network Analysis of Patents: a Case of Silicon-based Thin Film Solar Cells", *Scientometrics 86 (3),* pp. 687–703, 2011.

[9] P.-L. Chang, C.-C. Wu and H.-J. Leu, "Using Patent Analyses to Monitor the Technological Trends in an Emerging Field of Technology: a Case of Carbon Nanotube Field Emission Display", *Scientometrics 82 (1),* pp. 5-19, 2010.

[10] B. Yoon and Y. Park, "A Text-mining-based Patent Network: Analytical Tool for High-technology Trend", *The Journal of High Technology Management Research 15 (1),* pp. 37–50, 2004.

[11] H. Park, K. Kim, S. Choi and J. Yoon, "A Patent Intelligence System for Strategic Technology Planning", *Expert Systems with Applications 40 (7),* pp. 2373–2390, 2013.

[12] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Thao and X. Yan, "PatentMiner. Topic-driven Patent Analysis and Mining.", *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and DataMining. ACM SIGKDD. Beijing, China,* 2012.

[13] C. Lee, J. Jeon and Y. Park, "Monitoring Trends of Technological Changes Based on the Dynamic Patent Lattice: A Modified Formal Concept Analysis Approach", *Technological Forecasting and Social Change 78 (4),* pp. 690–702, 2011.

[14] M.-Y. Wang, D.-S. Chang and C.-H. Kao, "Identifying Technology Trends for R&D Planning Using TRIZ and Text Mining", *R&D Management 40 (5),* pp. 491–509, 2010.

[15] J. Struß, T. Mandl, M. Schwantner and C. Womser-Hacker, "Understanding Trends in the Patent Domain. User Perceptions on Trends and Trend Related Concepts.", *Proceedings of the First International Workshop on Patent Mining and Its Applications,* KONVENS'14. IPaMin. Hildesheim, Oct. 7th. 2014.

[16] https://www.epo.org/applying/basics.html.

[17] https://www.elastic.co.

[18] http://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families/definitions.html.

[19] https://code.google.com/p/stop-words/.

[20] A. Blanchard, "Understanding and customizing stopword lists for enhanced patent mapping", *World Patent Information,* Vol 29 (4), December 2007, pp. 308 -316, 2007.

[21] https://opennlp.apache.org/documentation/manual/opennlp.html.

[22] http://resources.mpi-inf.mpg.de/yago-naga/javatools/doc/javatools/parsers/PlingStemmer.html.

[23] https://lucene.apache.org/core/4_9_0/core/org/apache/lucene/analysis/Tokenizer.html.

[24] G. Salton and M. McGill, *Introduction to modern information retrieval*, McGraw Hill, 1983.

[25] D. Blei, A. Ng and M. Jordan, "Latent dirichlet allocation", *Journal of Machine Learning Research,* Vol. 3, 2003, pp. 993-1022, 2003.

[26] https://mahout.apache.org/users/clustering/latent-dirichlet-allocation.html.

[27] https://mahout.apache.org/users/clustering/k-means-clustering.html.

[28] https://lucene.apache.org/core/2_9_4/queryparsersyntax.html.

[29] http://www.wipo.int/patentscope/en/programs/patent_landscapes.

[30] C. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval, Ch. 16.3 Evaluation of Clustering,* Cambridge University Presss, 2008.