World Patent Information 31 (2009) 89–96

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Semantic enrichment and added metadata – Examples of efficient usage in an industrial environment ☆

Peter Atzmüller *, Gerald Landl

voestalpine Stahl GmbH, Intellectual Property Rights & Standards & Literature, voestalpine-strasse 3, AT-4020 Linz, Austria

ARTICLE INFO

Keywords:
Text mining
Semantic enrichment
Analysis and visualisation
Patent information
Competitive intelligence
Steel technology

ABSTRACT

The need for more comprehensive and "easy to use" search tools especially for patent documents still exists. In this paper, the developments of the ongoing project of voestalpine Stahl GmbH together with their partners regarding the implementation of a semantic search tool and an appropriate visualisation are revealed.

The main focus is on the different methods of analysis and examples of applications of the software. The features are on the one hand, a semantic representation of the content of the patent documents, in combination with the user's feedback and the use of landscape visualisation during the retrieval process. On the other hand a satisfactory quality is ensured by using among other features a document corpus which is finite and focused on the technology areas of voestalpine.

1. Introduction

For several years the need for more comprehensive (e.g. semantic) search tools for patent data has been increasing. Mainly, but not only, due to the fact that the amount of published patent documents which have to be considered is constantly rising.

Also, many different visualisation tools for the search results have been on the market for a fairly long time and there are several thorough overviews available as well [1].

The aim of this paper is to show several examples where a semantic search system is efficiently used in an industrial environment. In this article a newly developed software system which improves the search results by using semantic enrichment will be presented. Finally, an overview of further developments and limitations will be given.

2. "Traditional approach for searchers and their limitations/ drawbacks

There are several methods which are used for searching patent information. All of these approaches have different advantages and disadvantages. Following now is a listing of those approaches (without any claim on completeness).

2.1. "Classic" metadata

For many years this kind of metadata has been used for patent information retrieval. Examples of this "given" metadata are classification symbols (e.g. IPC), keywords in claims, description, title, names of inventor, assignee and/or dates (publication, application) and many more.

But this approach cannot deliver comprehensive results because there are mistakes in names or keywords on the one hand due to OCR errors or on the other hand, even due to misspelling.

Also different descriptions of the same concepts and different languages are proving such deficits. Finally, if taken alone, this method generates much "noise" that means low recall and/or precision of the results which can only be overcome by investing much time and effort.

2.2. Extracted metadata

Extracted metadata describes all kind of metadata which is generated solely from the document(s).

One very common example of extracted metadata are citations implemented by the applicants themselves which are not considered in the search reports of the examiners. Also the "consolidation" of assignee names or inventor names and their group affiliation (e.g. VAE is a member of the voestalpine group) are state of the art (e.g. described in [2]). Most recently for example extraction and clustering of regional information of the address of the inventor and/or assignee has been done.

Also this kind of metadata has its weaknesses. The effort which has to be made in order to get useful results is very high and data

☆ Some subject matter of this article were presented by the author, Peter Atzmüller, at the International Patent Information Conference and Exposition, IPI ConfEx, in March 2008, in Seville, Spain and by Gerald Landl at the Information Retrieval Facility Symposium, IRFS, in November 2007, in Vienna, Austria.

* Corresponding author. Tel.: +43 50304 15 9114.
E-mail address: Peter.Atzmueller@voestalpine.com (P. Atzmüller).

0172-2190/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2008.10.005

has to be always up-to-date (e.g. group affiliations). Still, one is faced with the mistakes (systematic/random) of names or ambiguous/wrong citations.

2.3. Added metadata

This kind of metadata describes every type of information which is added by intellectual work from engineers/scientists. For instance, these could be comments on patent documents, relevance feedback, internal keywords, reference to projects, translations or additional documents.

One of the greatest features, but at the same time, the greatest deficit might be the subjectivity of this kind of information, especially the comments and relevance feedback. Additionally, this kind of information very much depends on the time of submission (state of technology) and the position of the author (his/her viewpoint).

2.4. Landscapes

Use of landscapes to visualise patent information is also nothing new. It is used in several software products on the market, like STN® AnaVist™ [3], OmniViz or Aureka® etc. [4,5]. Paper [5] describes the analysis of search results using Aureka® (after the prior step of transcription of the results with e.g. Microsoft® Access or Lotus Notes®). The final conclusion of the Aureka® analysis is that it takes "... 6–8 h of the client's time to develop a thorough understanding of how the map can be interpreted."

Landscapes visualise the "similarity" – measured by different methods – of the documents in a given document pool; the big advantages of the software systems are the quick "reading and comparing" abilities of a software.

Nevertheless, there are also limitations for landscapes like the problem when dealing with a document pool which contains different languages or completely heterogeneous document pools (no peaks).

Additionally, there are problems with the repeatability of the results. Moreover, one needs a "suitable" amount of documents in order to get good results. Too few will generate an "empty" landscape where every document is its own hill; too many will lead very quickly to unbearable calculation needs (exponential growth of complexity).

3. Comprehensive approach to overcome those drawbacks

So how can one overcome the limitations of the presented kinds of metadata and visualisation?

Traditionally, one would have to invest much time and effort to either read and analyse the high amount of results and/or to improve the search strategy in order to improve precision without dropping the recall.

3.1. Our comprehensive approach

Our approach with the new system is to combine all these kinds of metadata and visualisation methods, adding semantic information and ontologies, and consecutively, using all of them during the search and evaluation process.

Firstly, free patent information from primary sources are imported to the software X/Pat® from the German company, G.e.I. Kramer&Hofmann mbH [6]. This is mainly used for basic retrieval and adding of metadata by our internal users (R&D engineers and IP department). Secondly, the software system being developed by the Austrian company, m2n – consulting and development gmbh [7], and the Know-Center Graz [8], which is attuned to the needs of searching patent information together with voestalpine Stahl GmbH. The System is based upon the m2n Intelligence Management Framework of m2n gmbh and the KnowMiner developed by the Know-Center in 2006 [9,10] and uses Semantic Technologies, Rules Based Engineering and Knowledge Discovery to realise the specific knowledge extraction and distribution processes.

The function of this tool is to provide the landscape visualisation of the results and to ensure the semantic enrichment of the data from X/Pat®.

Fig. 1 shows the relationship between these two systems in detail.

3.2. Preprocessing and data representation

One unique, homogenous index from the patent document information – like publication number, inventor names, abstract, claims, specification etc. and the non patent literature (NPL) information (in the future) will be generated. This index is represented in a semantic manner as *Resource Description Framework* (RDF)

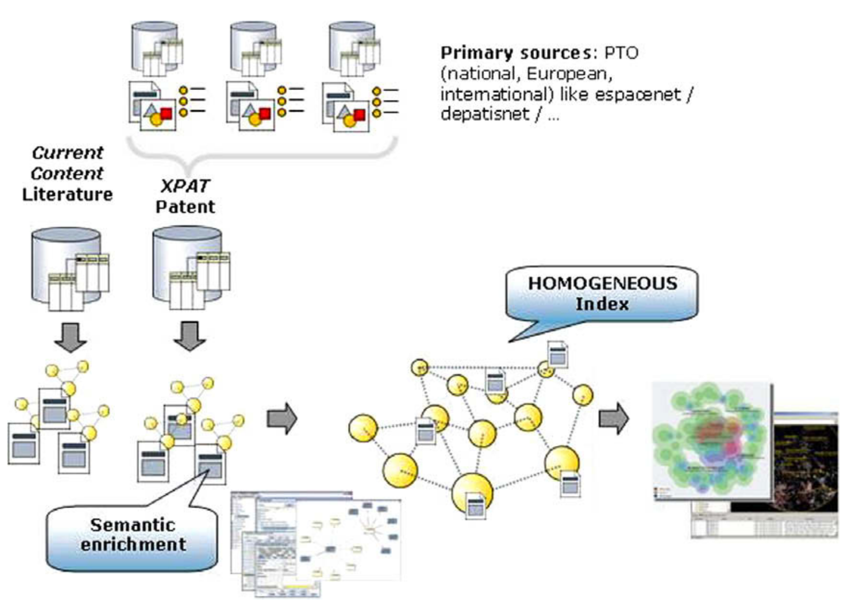

Fig. 1. Relationship between the primary sources, the software X/Pat and the new software system.

graph (for example "airbag is part of car" or "airbag has part steel pipe", quite similar to [11]).

Patent content will be broken down into paragraphs, sentences, phrases and tokens. By utilizing Part-of-Speech tagging linguistic information is added identifying nouns, verbs, proper nouns etc.

Stop words are filtered and removed. Information extraction techniques relying on the annotated NLP information are used for populating the RDF network with metadata like persons, places and organizations.

Currently, additional metadata categories like chemical compounds are planned. Besides populating known concepts, co-occurrence analysis of important terms as well as of different metadata categories like authors, IPC codes or similar official classification schemes and so on allows the system to derive significant relationships, subsequently used for similarity calculation and retrieval.

Although, very different as described in [11] Section 3, currently no Upper Ontologies or Web Ontology Language Description Logic (OWL-DL) is used. This OWL-DL defines the inference and reasoning capabilities; for example if a steam locomotive is a railcar and a railcar is a vehicle, then the steam locomotive is a vehicle. Mainly due to scalability and performance aspects, the system relies on statistical measurements for estimating similarities among patents. However, our system utilizes OWL-DL for workflow and application control. This allows the easy management and adaptation of workflows, making the system flexible for different and/or changing uses.

As outlined above, relationships among entities and terms are automatically extracted and utilized for similarity calculation in the retrieval and clustering process outlined below. Similarly the integration of a-priori knowledge by considering structural information of the e.g. IPC hierarchy information for similarity calculation is currently under development. Thus, considering structured knowledge in the similarity process allows for implicitly integrating a-priori knowledge in all algorithms relying on similarity estimates among documents.

Also, the software described in paper [11] is restricted to two technology fields and a limited amount of documents – about one order of magnitude lower than our document pool.

3.3. Concept extraction and relationship identification

After pre-processing, clustering techniques allow us on the one hand to derive the latent semantic concepts of the patents and on the other hand to prepare the landscape visualisation of the data set.

As base clustering algorithm, a bisecting k-Means [14] is used supporting splitting and merging of concepts as well as incremental updates. Thus, starting from the whole corpus the collection is split several times in the same way according to the defined similarity measure, assigning the documents to the most suitable cluster, for example the cluster with the highest similarity to a patent.

As outlined above, the integration of hierarchical information like IPC codes is currently under development. One important aspect in the use of a-priori knowledge is that expert users may explore different aspects of the dataset. Therefore, IPC codes can be combined with other orthogonal aspects of the data set like for example authors, locations etc. Orthogonal in this case means that these different aspects are evaluated independently of each other and correlated afterwards. As a result you can see e.g. which company and which place correlates to each other.

One goal is that users are given the freedom and flexibility for developing different viewpoints on the data sets and explore those viewpoints by visual means. So for example patents can not only be arranged hierarchically by content based similarity, but also by similarity among patent holders or companies – valuable information to analyse the patent landscape.

Due to scalability reasons, we focused on clustering for extracting the latent semantic concepts as opposed to the approach of latent semantic indexing, as described in paper [12] Section 4. Currently, the system scales up to 100,000+ patents with around 10,000 patents per minute, whereas first experiments show potential for an increase in data volume by using different clustering strategies.

While clustering allows segmentation of the data set on a document level, our co-occurrence analysis derived above allows the generation of a network of related concepts on a sub document level. It allows for identification of related or *synonymously* used

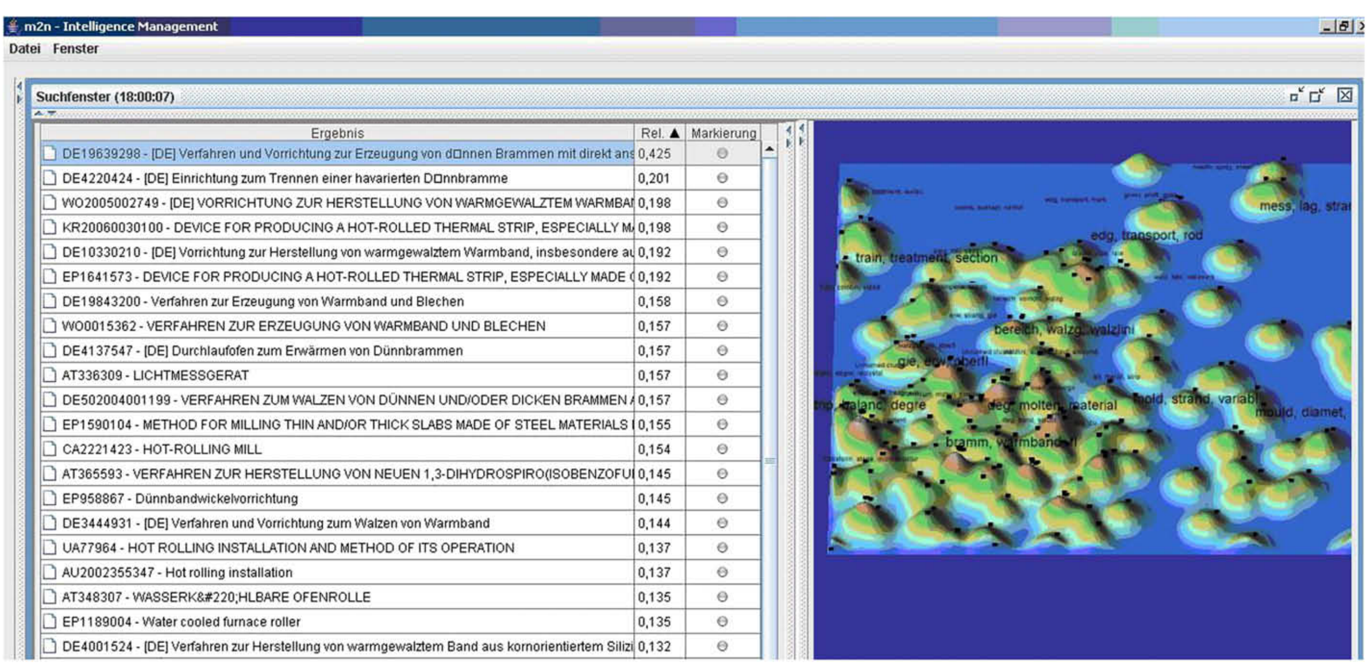

Fig. 2. Overview of a result of the prior art search with landscape visualisation.

terms for retrieval. Thus, data set specific relationships can be taken into account for analysing the data. Similarly to clustering, co-occurrence analysis scales up to 100,000+ documents, but with a significantly longer runtime.

3.4. Visualisation and human computer interaction (HCI) components

One major part of the system is the utilization of visualisation techniques for supporting explorative analysis of patent spaces. On top of the extracted concept clusters, projection algorithms transform the high dimensional relationships among patents onto a 3D surface, explorable by the user.

Such landscape visualisations are suitable for analysing relationships in large data sets and provide the basis for more sophisticated, quantitative correlation analysis. Colour coding allows for the identification of orthogonal aspects in the data set; for example different companies in the context of the concepts identified by automatic means like regional information.

Interaction methods allow different combinations of selecting/deselecting relevant patents so that users are able to narrow down their information requirements. Exploration using this landscape visualisation can be either done on the complete data set of up to 100,000 patents or interactively on selected respectively retrieved patents. Multiple synchronized views allow the combination of multiple landscapes. So if a set of relevant documents is selected in one landscape, the same patents are also selected in another landscape.

Overall, the visualisation components allow different ways of integrating the various explorative analysis methods at different stages of the workflow. An extension for analysing time based relationships such as concept drift or thematic changes are currently being planned. Considering changing relationships over time should reveal trends and time based patterns that are hardly observable by standard technology.

4. Different examples of efficient usage of the new system

There are several different tasks and needs which can be fulfilled by the system. These will be presented below:

4.1. Searching prior art for opposition/invalidation

The task is to search for prior art relevant to one given document – a very common application.

In the system the document under investigation will be analysed as described as in Section 3. The text of this document (title, abstract, claims, description, etc.) is evaluated, concepts retrieved and a search for "similar" documents is conducted. The similarity

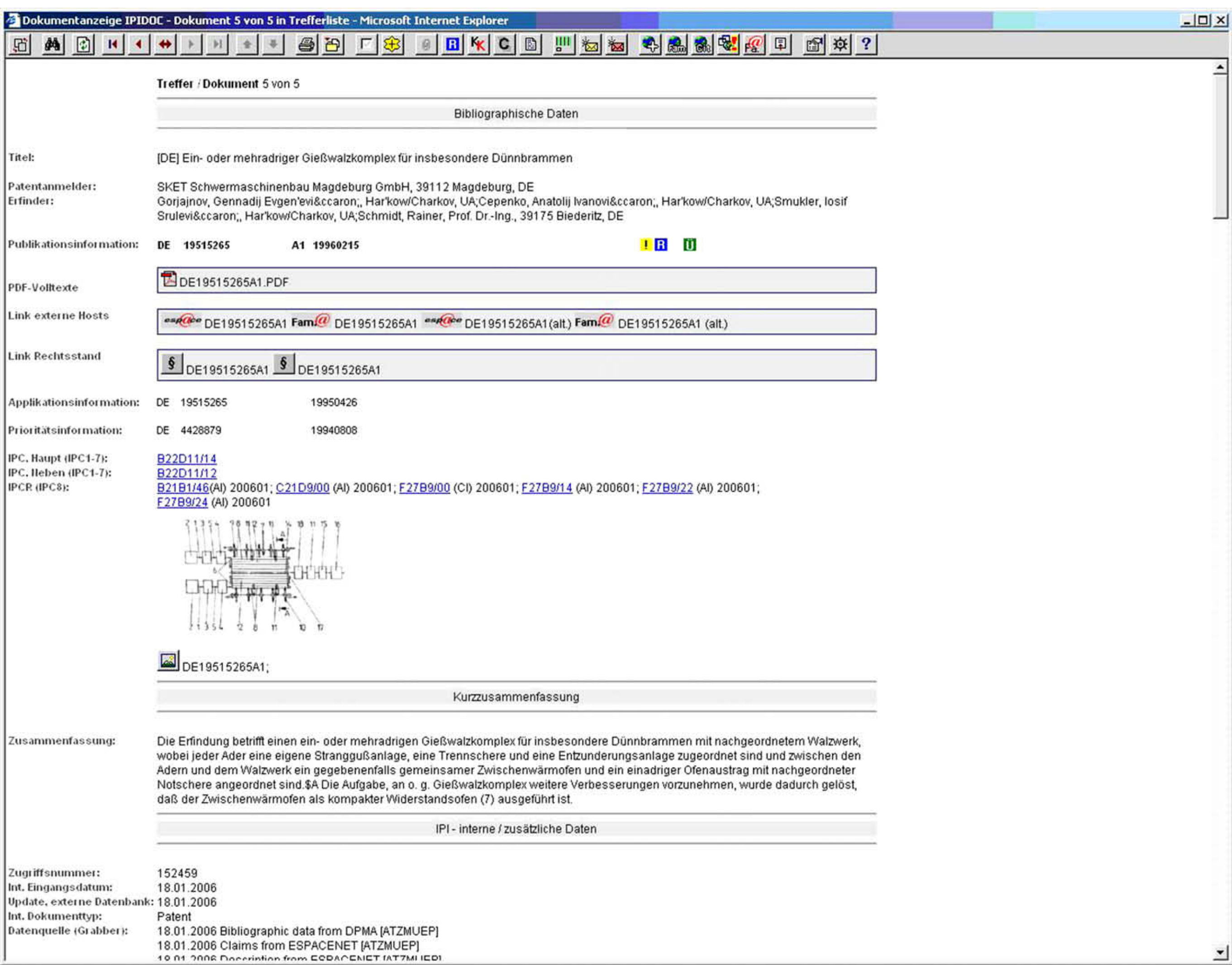

Fig. 3. Detail view of one document in X/Pat®.

is calculated by the number of the same concepts and term occurrences or co-term occurrences. As a result, a list of dozens to a few hundred documents sorted by relevance (= similarity) is obtained and a landscape is produced (see Fig. 2).

In this example the German Patent Publication with the publication number DE 19515265 A1, referred to later as "P1" with the title "Single- or multi-line casting-rolling installation partic. for thin strip" has been chosen. The results of the similarity search is also revealed in Fig. 2.

The similarity of the other documents with respect to P1 is presented in the column "Rel." – Relevance. The higher the number the more similar it is; up to a maximum of 1.0 – family members of the P1 would generate 1.0 (or almost this number – depending if the claims, abstract, and so on are identical).

The patent documents with the highest similarity within our document corpus of almost 100,000 documents are DE 19639298 C2 (P2) with 0.425 relevance, DE 4220424 C2 with 0.201, WO 2005002749 A2 with 0.198, KR 20060030100 A with 0.198, DE 10330210 A1 with 0.192, EP 1641573 A2 with 0.192, DE 19843200 C1 with 0.157, etc.

All those documents are related to the same technical area and – especially the most relevant ones – highly related to the topic of P1. P2 is classified with almost all the same IPC codes as P1, although there was no consideration of this information in the software (we want to implement this feature in the future, but always allowing it to be switchable or adjustable in order to avoid "self-restriction" and controlling of the results).

The user can choose one peak of the landscape and go into details back in X/Pat® (Fig. 3 shows a detailed view of one document in X/Pat®). Usually the peak with the most promising keywords, which are extracted from the abstracts of the documents, will be chosen.

The advantage of the view in X/Pat® is the comprehensiveness of all relevant data – so all images, the entire patent document (PDF) and the complete patent family are available there, as are internet links to the legal status database DPInfo, etc.

Benefits include a less time-consuming method to get relevant documents and additionally an "easier" approach for the user who does not need to have know-how about in-depth search strategies like IPC classes, usage of Boolean operators, etc. Finally, in this example missing documents (= NOT 100% comprehensiveness) is not as dangerous as it would be in Freedom-to-operate (FTO) searches.

4.2. *Creation of selective dissemination of information (SDI-profiles and alerts)*

The creation of SDI profiles for the different R&D engineers in order to build effective monitoring of newly published patent documents in a defined technological field is usually very time consuming.

When using Boolean operators one has to build a highly complex query in order to get reasonable results (good precision and – even more important – excellent recall).

A typical example for such a SDI search query is shown in Fig. 4.

The new approach now is to start with a very simple search for first results. This result list should contain also "bad" examples and can be fairly broad. Support Vector Machines (SVM) [15] which are methods to build a classifier with minimized training set error and test set error by splitting the good examples and the bad examples and trying to find rules which best separate those groups. These SVM are used as one part of the machine learning process, these "negative" examples are very important to improve the quality of the classifier.

This result list will be provided to the R&D specialist who can process it quite quickly and assign colours, representing relevance levels, to those documents:

– Red means this document should not be a result of the query (negative);
– Yellow means undecided/unclear but may be relevant (extended/broader positive);

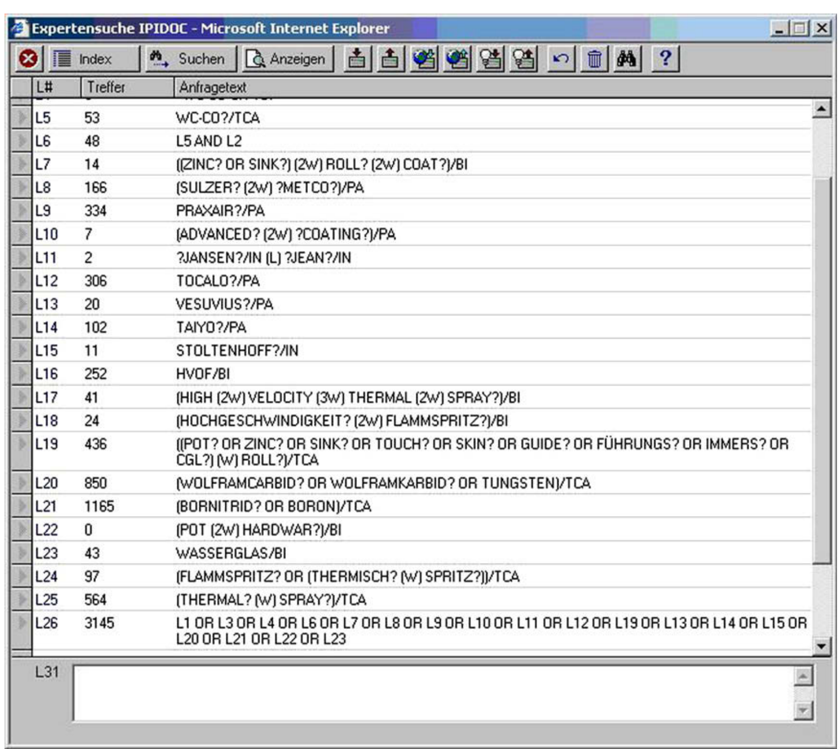

Fig. 4. Traditional SDI profile – built with the use of Boolean operators.

– Green means that the document is a positive result of the query (positive).

The user does not have to go through the complete list, but only through parts of it.

Additionally, the user can also assign colours (= relevance feedback) to complete peaks of the landscape.

Fig. 5 shows one example of assignment of colours – in this case a positive feedback (green) to one peak[1].

After the evaluation a new iteration based on the feedback will be automatically conducted and the results presented. The user gives feedback to the newly added documents and starts the process again. During that cycle the amount of green documents is constantly rising and that of red ones going towards zero.

On reaching a certain threshold (e.g. 50% of the results have to be green) this process is successfully finished and the query (that is the trained classifier) can be saved.

If there are new documents imported to the document pool matching this query this document(s) will be delivered to the relevant user(s) "work basket" automatically.

Advantages of this process are similar to those of example (a), namely, time saving and easy to handle search tasks. Also the quality of results will be improved and "surprising" results can be expected. This means there will be some hits which seem to be mistakes on the first glance but they are just well camouflaged documents matching the query.

4.3. White pattern analysis

Another application for the software is the detection of free fields within certain technologies.

The starting point is a search for patent documents of a certain corporation or several groups as assignee or a general overview about a technology (e.g. via an IPC code).

The second step will be the clustering of the documents regarding the technology topics in order to get an overview of the different technologies patented. After that a visualisation of the results as landscape is carried out. Within this landscape a certain assignee or group can be marked (coloured).

Finally, the white patterns within a certain technology respectively strong and weak fields of certain companies can be extracted and will be shown in the landscape. Obviously, one cannot get exact information about the content of the white space, therefore you will have a look at the documents on the border of the white fields to examine the technology of the free areas. Alternatively, one can choose a broader selection as underlying map and select with different parameters like inventor name, assignee name, etc. those documents of interest. This will enable obtaining information about the technology field not represented by one of the chosen parameter.

Fig. 6 shows this process. This screenshot of a landscape overview reflects the situation of two different companies (= assignees) in a certain technology field. As first step, a landscape has been generated in order to build up the "basis" overview of the underlying technology field. In order to separate visually the two assignees; one assignee is marked "green" and the other one "yellow" and the landscape visualization is limited to those two parties[2]. As result the empty space within reveals the technology in which neither assignee is represented.

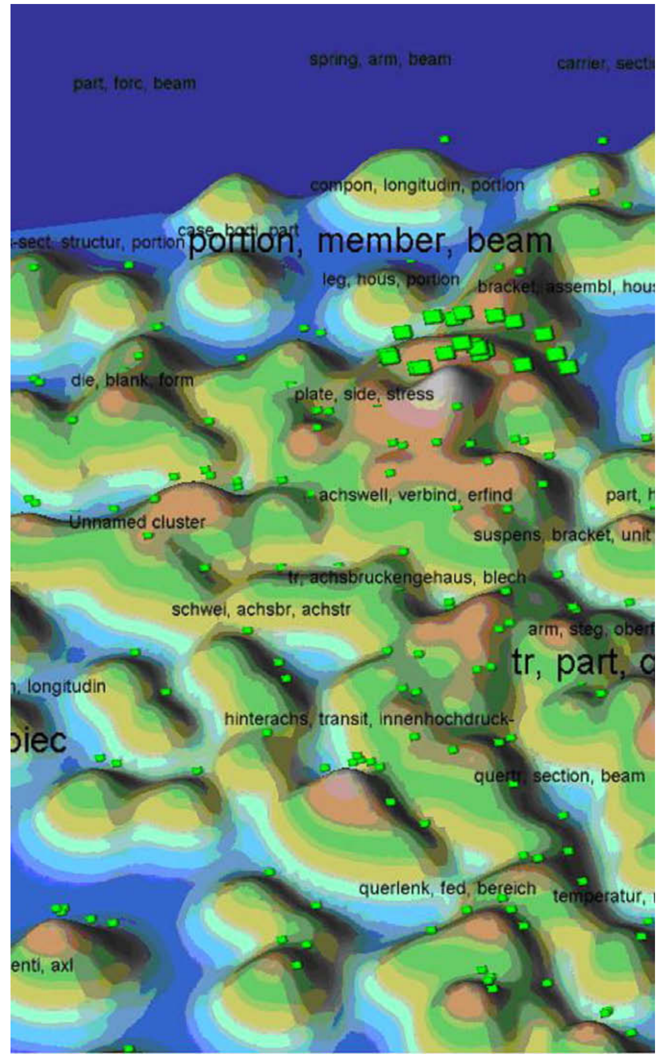

Fig. 5. Example for assignation of relevance feedback to documents within the LS.

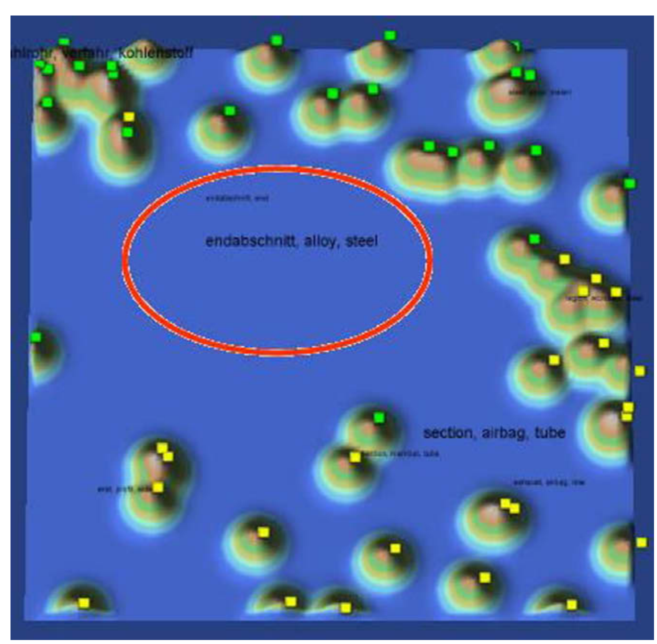

Fig. 6. Example for white pattern analysis.

[1] For interpretation of color in Fig. 5, the reader is referred to the web version of this article.

[2] For interpretation of color in Fig. 6, the reader is referred to the web version of this article.

But there are some third party patent documents present in that certain field, otherwise there would be no named cluster within the "empty space".

In conclusion, this application generates basic material for strategic decisions and is able to point to promising technological fields wherein one is in a weaker or stronger position than the competitors.

4.4. Explorative search (to be implemented)

This function serves as an easy-to-use feature for R&D engineers having a creative approach where they just have a look at some technologies without any "determined goal" when starting their search.

Starting with a huge landscape, showing the overall content of the database in several clusters (for instance IPC sub-classes as filter) the engineers can go deeper into one of these clusters and further separations within those clusters. The final goal is to find interesting documents without "searching" in the usual sense.

We expect users to adopt this function as a helpful tool for brainstorming.

5. Conclusion/further developments/limitations

From our point of view the approach discussed in this paper results in several benefits. Firstly, the unique usage of landscape visualisation during the search process and not only as a tool for presenting the results. Secondly, the semantic enrichment of the data (addition of the context of keyword → meaning of the content) leading to hits which can't be retrieved via traditional Boolean searches.

Finally, the basic is prefiltered to documents which are most likely to be relevant for certain technology fields. This leads automatically to more reliable, faster and improved results by avoiding ambiguity of keywords arising from completely different technology fields.

Planned developments of the current system:

- The data content and the association index (see Fig. 7) where the thickness of the arrows and the respective distance of the concepts is showing the dependencies and relevancies (but also avoidance of overlapping is considered!) will be tailored to the needs of our business and technology fields (steel producing and steel processing);

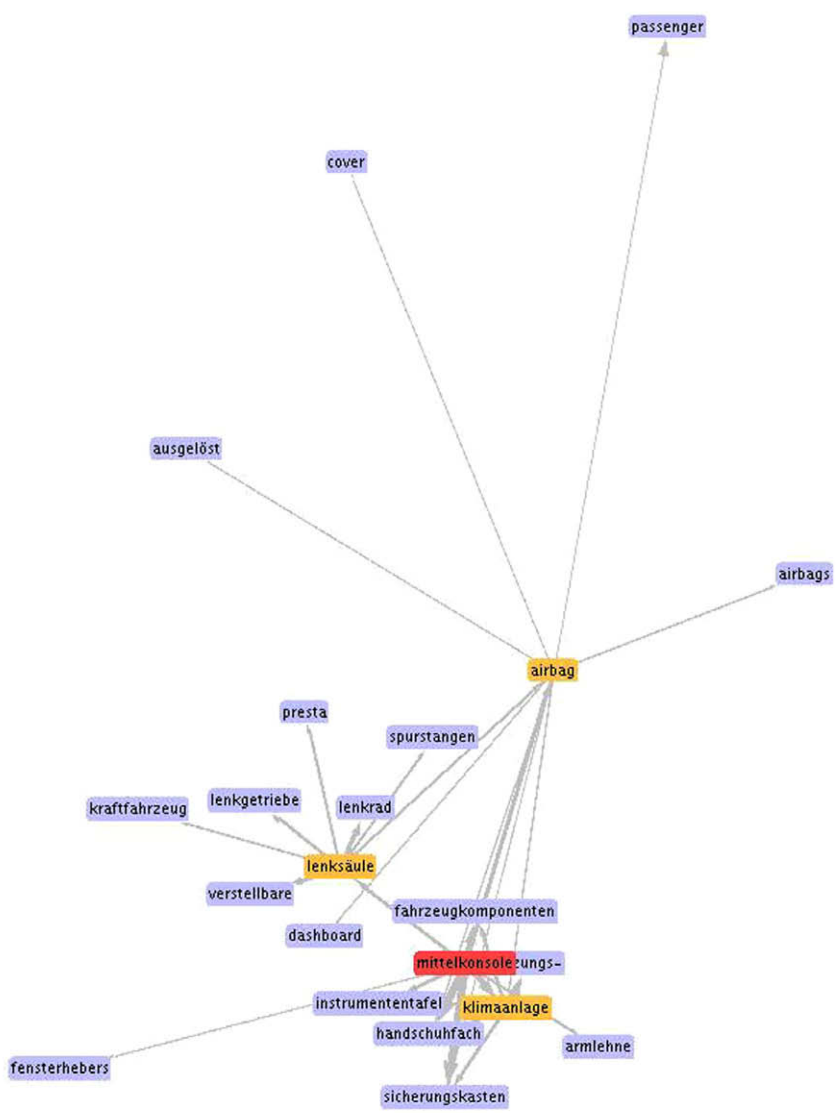

Fig. 7. Example for an association index – in this case an airbag module.

– Integration of improved consideration of e.g. IPC codes and similar official schemes for the similarity measurement;
– Feedback cycle for semantic recognition of keywords should be implemented (user can give feedback to the occurrence of some keywords);
– Non-patent-literature (NPL) will be implemented and considered, which leads then to anticipation of patent publications;
– Improvement of the stop word list in order to enhance the quality of the clustering and landscape visualisation [13];
– Implementation of a broader document set in order to have a possibility to find unexpected hits.

There is still plenty of room for improvements to the process of retrieval and analysis and we are convinced this system will lead to a new and more efficient workflow.

The work in this project is still on going and a lot of research and evaluation work has to be done, but nevertheless a prototype has already been implemented within voestalpine and the mentioned improvements of the software which lead to an operational version which is expected to be available at the beginning of 2009. Full functionality will be reached around the beginning of 2010.

Acknowledgements

Thanks to Mr. Andreas Hofmann, Ms. Doris Ipsmiller and Dr. Michael Granitzer and their team members for their support and developments for the realisation of the software.

References

[1] Eldridge J. Data visualisation tools – a perspective from the pharmaceutical industry. World Pat Inform 2006;28:43–9.
[2] van der Ligt G. Patent landscaping – process, text mining and statistical patent valuation. IPIConfEx presentation, Seville, Spain; 2008.
[3] Fischer G, Lalyre N. Analysis and visualisation with host-based software – The features of STN® AnaVist™. Pat Inform 2006;28:312–8.
[4] Yang Y et al. Text mining and visualisation tools – impressions of emerging capabilities. World Pat Inform 2008. doi:10.1016/j.wpi.2008.01.007.
[5] Badger E. Techniques for analyzing literature search results. World Pat Inform 2008. doi:10.1016/j.wpi.2008.01.008.
[6] <http://www.interhost.de/>.
[7] <http://www.m2n.at/>.
[8] <http://www.know-center.tugraz.at/>.
[9] Granitzer M. KnowMiner: Konzeption und Entwicklung eines Generischen Wissenserschließungsframeworks dissertation. Austria: Technical University of Graz; 2006.
[10] Klieber W et al. KnowMiner: Ein service orientiertes knowledge discovery framework. In: Proceedings XML-Tage, Berlin, Germany; 2006.
[11] Wanner L et al. Towards content-oriented patent document processing. World Pat Inform 2008;30:21–33.
[12] Ryley JF et al. Advanced document retrieval techniques for patent research. World Pat Inform 2008;30:238–43.
[13] Blanchard A. Understanding and customizing stop word lists for enhanced patent mapping. World Pat Inform 2007;29:308–16.
[14] Bradley PS, Fayyad UM. Refining initial points for K-means clustering. In: Shavlik JW, editor. Proceedings of the fifteenth international conference on machine learning. San Francisco, CA: Morgan Kaufmann Publishers; 1998. p. 91–9.
[15] Vapnik V et al. Support vector method for function approximation, regression estimation, and signal processing. Advances in neural information processing systems 9, San Mateo, CA; 1997. p. 281–7.

Peter Atzmüller is patent information specialist at the R&D department of voestalpine Stahl GmbH. His diploma thesis was "Evaluation and integration of a patent management system" in 2005-02 (published as book in June 2008/ISBN: 3639036921 respect. 978-3639036923). He was certified as "Patentrechercheur LGA" from the German Landesgewerbeanstalt Nürnberg in 2008-03 and has been involved in posting activities in several patent information forums (e.g. PIUG). Additionally, he was invited to act as chairman at several EPO conferences. His main projects are Implementation of the patent management system in voestalpine Stahl GmbH 2005–2006 development of an integrated patent management workflow therein since 2006 and project manager of the described "Semantic patent analysis" project since 2007.

Gerald Landl is head of the IP-rights – Standards – Literature department at voestalpine Stahl GmbH. In 2001 he finished his degree in technical chemistry at Johannes Kepler University, Linz with a diploma thesis in chemical engineering. During his university years he also developed databases for safety relevant data providing quick access for managers within the safety–health–environment department at DSM Fine Chemicals Austria exploring the needs and obstacles of company wide information systems. After one year in R&D on conductive polymers for heating purposes and another year R&D on high temperature protection coatings at voestalpine Stahl GmbH being constantly involved with patent search, he implemented a corporate wide information system for standards and became head of the department IP-rights & Standards department in 2005.