EuroVis Workshop on Visual Analytics (2016)
N. Andrienko and M. Sedlmair (Editors)

Patent Retrieval: A Multi-Modal Visual Analytics Approach

Daniel Seebacher, Manuel Stein, Halldór Janetzko, and Daniel A. Keim

University of Konstanz, Germany

Abstract

Claiming intellectual property for an invention by patents is a common way to protect ideas and technological advancements. However, patents allow only the protection of new ideas. Assessing the novelty of filed patent applications is a very time-consuming, yet crucial manual task. Current patent retrieval systems do not make use of all available data and do not explain the similarity between patents. We support patent officials by an enhanced Visual Analytics multi-modal patent retrieval system. Including various similarity measurements and incorporating user feedback, we are able to achieve significantly better query results than state-of-the-art methods.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Display algorithms H.3.3 [Information Systems]: Information Search and Retrieval—Relevance feedback

1. Introduction

Intellectual property and technical inventions are usually protected by patent claims. Patents contain a set of exclusive rights for an invention granting patent holders the right to exclusively manufacture, sell, or use the patented invention. As a result, patenting an invention provides a competitive advantage and in some cases additional income in the form of licensing fees. Consequently, companies invest a substantial amount of money in research and development leading to an increasing number of patent applications. This trend can be observed in Figure 1 visualizing patent filings of the *European Patent Office*. The linear trend results in a quadratic increase of the overall number of patents stored in databases.

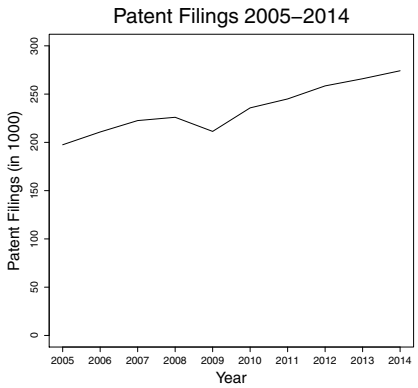

Figure 1: Number of patent filings from 2005 to 2014 per year at the European Patent Office. [epo16]

The increasing number of patent applications proves to be challenging for patent experts whose task is, among other things, to validate the patentability of an invention. The workflow of an official patent investigator is depicted in Figure 2 showing the general European patent application process beginning with filing an invention ending with the final grant. The patent application will be published when arriving at the patent office and simultaneously the validation process begins with a preliminary assessment. If the preliminary assessment yields a positive result, the patent application will be substantively examined and prior art will be searched extensively. A successful patent application must satisfy several criteria (for example, novelty, inventive step, and industrial applicability). The prior art search is conducted to ensure the novelty of submitted inventions which means that no similar or identical inventions are already part of the state of the art. If all criteria are met, the patent will be granted and published.

In our work, we want to focus on the most time-consuming step of this workflow, namely the prior art search (highlighted in blue in Figure 2). In this step, the official has to consider all the publicly available information such as articles in journals, interviews, and all existing 9.5 million patents [Wor15]. Existing information retrieval systems designed to help with this task contain common and severe drawbacks. They do not make use of all available data and, additionally, do not explain why one result is preferred over another. With the vast amounts of information existing today, finding relevant and helpful prior art proves to be a difficult task. The goal of Visual Analytics is to build a bridge between human and

DOI: 10.2312/eurova.20161118

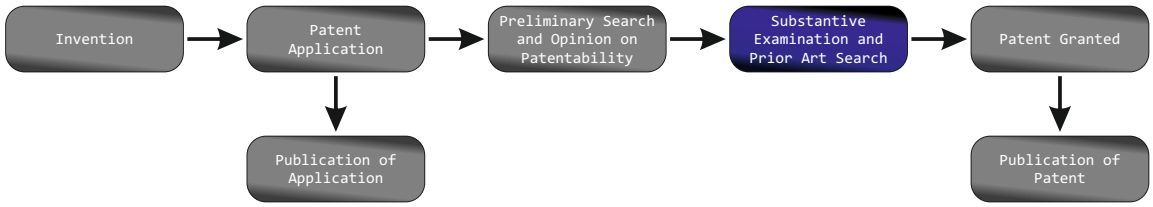

Figure 2: European patent application process. We support the domain experts during the substantive examination and prior art search.

computer, making the knowledge discovery process as effective as possible. We believe that Visual Analytics is helpful to get an overview of the resulting patents and simultaneously enable interactive user feedback to the data analysis process.

In this work, we contribute an effective semi-automatic patent retrieval Visual Analytics system allowing the patent official to integrate his domain knowledge. Moreover, we show that by incorporating the expert's knowledge our proposed methods achieve significant better results than standard solutions. Eventually, we provide a glyph design based on our requirement analysis improving the user's understanding of the results as well as the result ordering.

2. State of the Art in Patent Retrieval

We learned from interviews with a domain expert of the European Patent Office that searching for prior art is a challenging task. Each of the over 4.000 employed patent officials has own methods and approaches to look for prior art. For example, in the area of mechanics some patent examiners first create a subset of patents using keywords and filters for patent classes. Afterwards, they manually inspect up to several thousand patents assessing whether a filed invention is already covered by the state of the art. Therefore, a system presenting similar prior art to a query patent is very desirable. Such a system would drastically reduce the duration of the manual exploration and verification process. Furthermore, it is considered crucial to let the domain experts interactively explore the search space and enhance the query results. At present patent examiners have two tools to assist them in their prior art search: EPOQUE [Jon90], a boolean search engine and a text search engine based on ElasticSearch [Ela16] that mainly works with keywords. The drawback of both of these commonly used systems is that they do not incorporate images, illustrate why some results are preferred over others or give the user the possibility to integrate their expert knowledge. Instead, they mostly perform boolean queries based on metadata and similarity queries based on the text contained in patent documents.

In current research, images have already been used in the field of patent retrieval [MSEMSZ03, YQHE06, SVK10], but only seldom combined with text and/or metadata [CPVP08]. Evaluations, however, show that a combination of all modalities results in a better performance [CRJ11]. Moreover, current research patent retrieval systems seldom take advantage of the user and his expert knowledge, for instance by relevance feedback, which refers to techniques that use query results and relevance information to iteratively improve the performance of retrieval systems [SB97]. Relevance information can either be provided by the user explicitly or implicitly, or can be generated automatically. Another approach incorporating multi-modal patent retrieval as well as user interaction is PatViz [KBGE09]. The main difference between PatViz and our system is the way how users can interact with the system. In PatViz the user is given the possibility to refine query using a powerful visual query builder. In our system, by contrast, we algorithmically determine weights for modalities based on user feedback.

3. System

A patent can be modeled as an object of various modalities, with text, images, and metadata being the most descriptive ones. The text content of a patent, for example, can be further split into its structural elements like title, description, or claims. We decided to use all available data for the retrieval of patents because evaluations as conducted by Belkin et al. [BKFS95] show, that employing all available modalities results in a better performance than identifying and using only the single best one. Consequently, we designed a pipeline as depicted in Figure 3 to support multi-modal patent retrieval

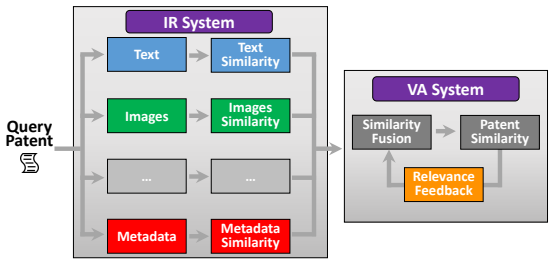

Figure 3: Overall Visual Analytics workflow. Combines the process pipeline for multi-modal patent retrieval with user interaction possibilities to iteratively improve the performance.

Each preprocessed modality is used to calculate a similarity between a query patent and each patent in the database. Since this can be done independently for each modality, it is possible to use many different similarity measurements. We use the BM25 similarity [The15a] of Lucene [The15b] to compute

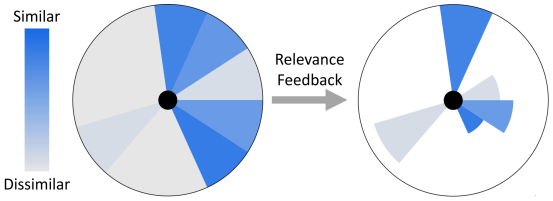

Figure 4: Each sector of the glyph visualization corresponds to one modality. The similarity of each modality is mapped to the color of the sector. After at least one relevance feedback iteration we can determine weights for each modality and map them on the size of the sectors.

scores for the text modalities, like many state-of-the-art patent retrieval systems [BMWH10, GPTR10, MKG*11, PIRF10]. Sidiropoulos et al. have shown in an evaluation that their Adaptive Hierarchical Density Histogram method is currently the best method for patent images [SVK10]. For the metadata, we use the Jaccard similarity coefficient [Jac01], like Lopez and Romary [LR09]. Finally, the similarity measurements for each modality are combined into a single similarity score allowing an ordering of similar patents. We use the linear combination method [BCB94] which allows setting a unique weight for the similarities of each modality.

For the exploration of the query results, we designed a glyph visualization as shown in Figure 4. This glyph visualization allows us to not only present the similarities of each modality but also the corresponding weight in the linear combination method, which reflects their influence on the combined patent similarity. Each sector of the glyph visualization corresponds to one modality. The weight of the modalities is mapped linearly to the radius of the sector. The similarity in each modality is mapped to the color of the sector, with grey for dissimilar and blue for similar patent aspects. This glyph visualization was inspired by star glyphs [SFGF72], which are often used and known to work in a small-multiple setting. It should be possible to considerably reduce the size of glyph visualization, allowing to use them in a small multiple, while still making the important modalities recognizable.

In order to improve the query results, it is important to incorporate expert knowledge and to offer the opportunity to interactively specify relevance of query results. Consequently, the ordering of results and the similarity score should be adapted to fit the user's expectation and requirements. This is why we provided the user with interaction possibilities by relevance feedback. For each result the user has the possibility to give binary feedback, i.e., one result is relevant or not. We employ the linear combination method with the trained weights by multiple regression (*LCR*) of Wu [Wu12], who showed that *LCR* outperforms other methods. Using *LCR*, we calculate a weight for each modality to determine its influence in accordance to the analyst's feedback. The resulting weights will correspond to the user judgment. Additionally, we are providing several drill-down capabilities to help the user gain further insights.

Having all system requirements in mind, our designed prototype can be seen in Figure 5. In addition to showing the most similar patent applications, the system's purpose is to explain why the system rated a patent as similar. The system consists of two parts: An overview over the query patent is provided on the left hand side, and on the right hand side the results of the query are shown. To give an overview over the query patent we show, among other things, word clouds to visualize the most prominent terms in text modalities. Word clouds are a familiar and easily comprehensible visualization for most people. Users may tend to prefer them over more sophisticated visualization, like noticed in [KBGE09]. The term frequency is mapped to the color and size of the term in the word cloud. The aforementioned glyphs are used to explain how a patent has been ranked. Furthermore, we offer different comparison views for the various modalities, to quickly compare the query patent with a result patent. One example would be the polar word clouds of the Kumo library [Cas16], as depicted in the foreground in Figure 5, which allow to quickly compare the most prominent terms of the query and a result patent. The term frequency is again mapped to the size and color of the terms in the polar word cloud.

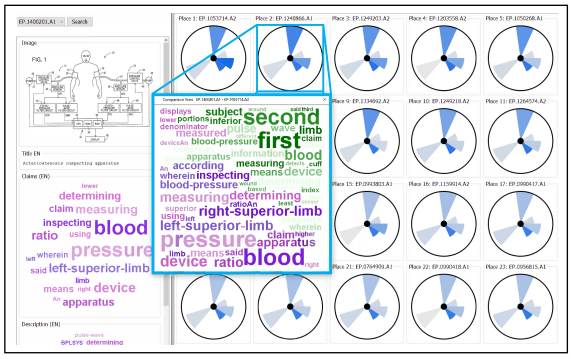

Figure 5: On the left-hand side information about the query patent are shown. In the background on the right-hand side the results and their corresponding glyph visualizations are displayed. In the foreground a comparison view of the English patent claims of the query and a result patent is shown.

4. Discussion

According to our experience, we believe our system is of great use for patent officials. The simultaneous display of query and result objects allows a quick and easy comparison. Our glyph visualization enables a better understanding which modalities of a result are similar to the query object using a grey to blue color scale. We chose a visual design that is easy to understand, thus increasing the acceptance by patent officials, although there might be many more sophisticated design alternatives. Additionally, we can show the impact of the modality on the similarity score. We include domain knowledge and expert feedback by interactions rating query results.

We motivated our research with prior art search being a manual, tedious, and time-consuming process. We propose a Visual Analytics system explaining why one result is rated more similar than others. Thus, analysts have the opportunity to make fully informed feedback decisions to the system.

In our evaluation, we formulated two hypotheses assessing the consequence of user feedback. Firstly, we wanted to show that using all available data results in better performance than identifying and using only the single best modality. Our second hypothesis is that using the user feedback results in an even better performance. For our evaluation, we used the publicly available dataset for the *"Image Prior Art Candidate Task"* of the *CLEF-IP 2011* [CLE16]. The dataset contains approximately 48.000 patents. Additionally, there is a query relevance file containing the information, which patents in the dataset are relevant for a given query patent. For each patent in the dataset the title, description, and claims of the patent are available in one or several of the following languages: English, German, and French. Handling multiple languages is an additional challenge which we did not handle in this work, but which may prove useful in further work.

In detail, our evaluation consisted of three experiments. Firstly, we tested the performance of our system if only single modality was used. An overview over the different modalities we used in this experiment and their respective results can be seen in Table 1. Description and claims are worse in comparison to title and metadata, since they are mostly available in a single language only and cannot be compared with other languages. Only using images performed worse than other modalities, but this is expected, as other evaluations show that images discriminate worse than text or metadata [CRJ11].

In our second experiment, we used the *CombSUM* method of Fox and Shaw [FS94] combining the similarity measures of all modalities into a single measurement. Using this fused similarity, we achieve a *Mean Average Precision (MAP)* of 0.089 which is already nearly a third better than the results for single modalities. Using the *Paired Bootstrap Test*, as described by Sakai [Sak14], we calculated a p–value of 0.045 which allows us to reject the null hypothesis that the results are equal. Therefore, we are able to confirm our first hypothesis, since we can show that using all modalities results in a statistically significant better performance of our retrieval system.

Finally, we tested the influence of user feedback on the performance of our system. We used the query relevance file to simulate a user providing relevance feedback. We conducted an initial query, asked our simulated user for relevance feedback for the first 20 results and used this information to determine weights for the individual modalities using the *LCR* method. After a single iteration, our *MAP* could already be improved from 0.089 to 0.128. After four iterations we reached a plateau at 0.18, nearly doubling the *MAP* and confirming our hypothesis that user interaction improves the performance.

Modality	Mean Average Precision
Title (E)	0.069
Description (E)	0.004
Claims (E)	0.032
Title (G)	0.044
Description (G)	0.003
Claims (G)	0.014
Title (F)	0.055
Description (F)	0.001
Claims (F)	0.011
Metadata	0.044
Images	0.006

Table 1: An overview over the different modalities we used and the *Mean Average Precision (MAP)* we achieved using only individual modalities for the *"Image Prior Art Candidate Task"* [CLE16]

We believe that our system is generalizable and applicable in domains other than patent retrieval. Many objects consist of multiple modalities, an obvious example being newspaper articles, which also contain text, images and metadata. The system and especially our glyph visualization could also be adapted to be used for plagiarism detection. Here, not only finding plagiarisms, but also visualizing why a given result was identified as such is of high interest. Then of course, other similarity measurements have to be considered, for example citation-based similarities as proposed by Bela Gipp [Gip14].

Additional future work includes switching from the current *"late fusion"* to an *"early fusion"* approach [SWS05]. Instead of fusing the similarities at the end we could already fuse them in an earlier step, for example, by combining the feature vectors we calculated for each modality. This would give us the opportunity to use the relevance feedback of the user to directly influence the query, for example, using the Rocchio algorithm [Roc71]. Furthermore, we plan to make the system available to domain experts of the European Patent Office for an exhaustive qualitative evaluation.

5. Conclusion

We presented a Visual Analytics workflow to support patent examiners during their prior art search. We use multi-modal patent retrieval to improve the retrieval performance of our system, applied a glyph visualization to help the user gain insight on why some results are ranked better than others and provide user interaction possibilities by relevance feedback and drill-down capabilities. Our evaluation shows that we could improve the performance compared to classical retrieval systems, which only use a single modality and we discussed possible promising future work opportunities.

References

[BCB94] BARTELL B. T., COTTRELL G. W., BELEW R. K.: Automatic Combination of Multiple Ranked Retrieval Systems. In *Proc. 17th Int. Conf. on Research and Development in Information Retrieval (SIGIR)* (1994), pp. 173–181. 3

[BKFS95] BELKIN N. J., KANTOR P., FOX E. A., SHAW J. A.: Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management 31*, 3 (1995), 431–448. 2

[BMWH10] BECKS D., MANDL T., WOMSER-HACKER C.: Phrases or Terms? The Impact of Different Query Types. In *CLEF (notebook Papers/LABs/Workshops)* (2010). 3

[Cas16] CASON K.: Kumo. URL: https://github.com/kennycason/kumo, Last accessed on: 16.02.2016. 3

[CLE16] CLEF-IP 2011, UNIVERSITY OF TECHNOLOGY VIENNA: Data for the Image PAC Tasks, Last accessed on: 02.08.2016. URL: http://www.ifs.tuwien.ac.at/~clef-ip/download/2011/index.shtml. 4

[CPVP08] CODINA J., PIANTA E., VROCHIDIS S., PAPADOPOULOS S.: Integration of Semantic, Metadata and Image Search Engines with a Text Search Engine for Patent Retrieval. In *SemSearch* (2008), pp. 14–28. 2

[CRJ11] CSURKA G., RENDERS J.-M., JACQUET G.: XRCE's Participation at Patent Image Classification and Image-based Patent Retrieval Tasks of the Clef-IP 2011. In *CLEF (Notebook Papers/Labs/Workshop)* (2011). 2, 4

[Ela16] ELASTIC.CO: Elasticsearch, Last accessed on: 11.02.2016. URL: https://www.elastic.co/products/elasticsearch. 2

[epo16] EPO.ORG: European patent filings 2005–2014 per country of residence of the applicant, Last accessed on: 04.02.2016. URL: https://www.epo.org/about-us/annual-reports-statistics/statistics/filings.html. 1

[FS94] FOX E. A., SHAW J. A.: Combination of Multiple Searches. *NIST SPECIAL PUBLICATION SP* (1994), 243–243. 4

[Gip14] GIPP B.: *Citation-based plagiarism detection: detecting disguised and cross-language plagiarism using citation pattern analysis*. Springer, 2014. 4

[GPTR10] GOBEILL J., PASCHE E., TEODORO D., RUCH P.: Simple Pre and Post Processing Strategies for Patent Searching in CLEF Intellectual Property Track 2009. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, vol. 6241 of *Lecture Notes in Computer Science*. Springer, 2010, pp. 444–451. 3

[Jac01] JACCARD P.: *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901. 3

[Jon90] JONCKHEERE C.: EPOQUE (EPO QUEry service) the Inhouse Host Computer of the European Patent Office. *World Patent Information 12*, 3 (1990), 155–157. 2

[KBGE09] KOCH S., BOSCH H., GIERETH M., ERTL T.: Iterative integration of visual insights during patent search and analysis. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (2009), IEEE, pp. 203–210. 2, 3

[LR09] LOPEZ P., ROMARY L.: Multiple Retrieval Models and Regression Models for Prior Art Search. *CoRR* (2009). 3

[MKG*11] MAHDABI P., KEIKHA M., GERANI S., LANDONI M., CRESTANI F.: *Building Queries for Prior-Art Search*. Springer, 2011. 3

[MSEMSZ03] MAHMOUDI F., SHANBEHZADEH J., EFTEKHARI-MOGHADAM A.-M., SOLTANIAN-ZADEH H.: Image Retrieval based on Shape Similarity by Edge Orientation Autocorrelogram. *Pattern Recognition 36*, 8 (2003), 1725–1736. 2

[PIRF10] PÉREZ-IGLESIAS J., RODRIGO A., FRESNO V.: Using BM25F and KLD for Patent Retrieval. In *Cross-Language Evaluation Forum (notebook Papers/LABs/Workshops)* (2010). 3

[Roc71] ROCCHIO J. J.: Relevance Feedback in Information Retrieval. In *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, 1971. 4

[Sak14] SAKAI T.: Metrics, Statistics, Tests. In *Bridging Between Information Retrieval and Databases*. Springer, 2014, pp. 116–163. 4

[SB97] SALTON G., BUCKLEY C.: Improving Retrieval Performance by Relevance Feedback. *Readings in Information Retrieval 24*, 5 (1997), 355–363. 2

[SFGF72] SIEGEL J. H., FARRELL E. J., GOLDWYN R. M., FRIEDMAN H. P.: The surgical implications of physiologic patterns in myocardial infarction shock. *Surgery 72*, 1 (1972), 126–141. 3

[SVK10] SIDIROPOULOS P., VROCHIDIS S., KOMPATSIARIS I.: Adaptive hierarchical density histogram for complex binary image retrieval. In *International Workshop on Content-Based Multimedia Indexing (CBMI)* (2010), pp. 1–6. 2, 3

[SWS05] SNOEK C. G. M., WORRING M., SMEULDERS A. W. M.: Early versus Late Fusion in Semantic Video Analysis. In *Proc. 13th Annual Int. Conf. on Multimedia* (2005), pp. 399–402. 4

[The15a] THE APACHE SOFTWARE FOUNDATION: BM25Similarity. URL: https://lucene.apache.org/core/5_3_0/core/org/apache/lucene/search/similarities/BM25Similarity.html, Last accessed on: 04.11.2015. 2

[The15b] THE APACHE SOFTWARE FOUNDATION: Lucene. URL: http://lucene.apache.org/, Last accessed on: 20.08.2015. 2

[Wor15] WORLD INTELLECTUAL PROPERTY ORGANIZATION: IP Rights in Force. URL: http://ipstats.wipo.int/ipstatv2/keysearch.htm?keyId=205, Abgerufen am: 17.08.2015. 1

[Wu12] WU S.: Linear Combination of Component Results in Information Retrieval. *Data & Knowledge Engineering 71*, 1 (2012), 114–126. 3

[YQHE06] YANG M., QIU G., HUANG J., ELLIMAN D.: Near-Duplicate Image Recognition and Content-based Image Retrieval using Adaptive Hierarchical Geometric Centroids. *Pattern Recognition, International Conference on 2* (2006), 958–961. 2