User Interface for Customizing Patents Search: An Exploratory Study

Arthi M. Krishna(✉), Brian Feldman, Joseph Wolf, Greg Gabel, Scott Beliveau, and Thomas Beach

United States Patent and Trademark Office, Alexandria, VA, USA
{arthi.krishna, brian.feldman, joseph.wolf, greg.gabel, scott.beliveau, thomas.beach}@uspto.gov

Abstract. Prior art searching is a critical and knowledge-intensive step in the examination process of a patent application. Historically, the approach to automated prior art searching is to determine a few keywords from the patent application and, based on simple text frequency matching of these keywords, retrieve published applications and patents. Several emerging techniques show promise to increase the accuracy of automated searching, including analysis of: named entity extraction, explanations of how patents are classified, relationships between references cited by the examiner, weighing words found in some sections of the patent application differently than others, and lastly using the examiners' domain knowledge such as synonyms. These techniques are explored in this study. Our approach is firstly, to design a user interface that leverages the above-mentioned processing techniques for the user and secondly, to provide visual cues that can guide examiner to fine tune search algorithms. The user interface displays a number of controls that affect the behavior of the underlying search algorithm—a tag cloud of the top keywords used to retrieve patents, sliders for weights on the different sections of a patent application (e.g., abstract, claims, title or specification), and a list of synonyms and stop-words. Users are provided with visual icons that give quick indication of the quality of the results, such as whether the results share a feature with the patent-at-issue, such as both citing to the same reference or having a common classification. This exploratory study shows results of seven variations of the search algorithm on a test corpus of 100500 patent documents.

Keywords: Patent similarity · Prior art search · User interface design

1 Introduction

Searching for prior art is one of the most critical, time-intensive aspects of patent examination. Searching is like navigating through ocean waters, using beacons to avoid rough shorelines and to reach the correct destinations. There is an ocean of prior art available today in the form of patents, published applications and non-patent literature. Examiners use beacons in the form of search tools and databases to navigate the prior art ocean. And in doing so, they aim to avoid irrelevant references - i.e. rocky shorelines–and apply the relevant ones - i.e. to reach the correct destinations.

C. Stephanidis (Ed.): HCII 2016 Posters, Part I, CCIS 617, pp. 264–269, 2016.
DOI: 10.1007/978-3-319-40548-3_44

Automatic prior art retrieval algorithms, if accurate, can assist expert examiners by identifying literature that would otherwise take substantial research to uncover. Examiners today have access to an automated search tool, PLUS - the Patent Linguistic Utility Services, which is limited in its usability. PLUS uses a typical approach to automating prior art search, namely determining a select few keywords from the input patent application and retrieving published patent documents that exhibit a high level of textual similarity to these keywords. However, simple keyword searches have limited utility in the patent prosecution context because of the high prevalence of uncommon language patterns [8] and intentional creation by patent applications of 'abstract vocabulary' specific to their claimed invention. Thus, there has been extensive research on how keyword matching can be augmented by various techniques such as using classification systems and citations within references [1, 2, 4].

Fig. 1. Interface shows patents results similar to US11061715

Our new approach to building a reliable automated prior art search system is two-fold. Firstly, we build a search system that not only performs the basic keyword searches, but also has several layers of augmented processing techniques that can be controlled and modified as the search progresses. By designing a user interface (Fig. 1) that allows the expert to alter the behavior of search algorithm, for instance defining the relative weights of different sections of the patent (e.g., title, claims, specification and abstract), experts can create strategies of patent retrieval algorithms best suited to examining a particular application. Secondly, the user interface has quick visual cues that provide immediate feedback regarding the quality of the patent search results. For example, visual indicators show whether the input patent belongs to the same "patent family" as the result, thereby helping the user judge the quality of the patents search.

2 Implementation

We use Apache Solr Lucene [6] open source search system with cloud capability (version 5.1) to implement the core of the patent search system tool, which hereafter we refer to as Sigma. The web-based user interface for Sigma is built using AngularJS.

The patent text available in public corpus (http://patents.reedtech.com/) is ingested into Solr indices with the following Solr fields – id, title, abstract, claims, specification and full_text. Lucene uses term frequency – inverse document frequency (tf-idf) [7] calculations for text relevancy and retrieval. Among Solr's various built in query parsers, such as Simple and Field query parsers, the MoreLikeThis query parser, which can take an entire document as the input and retrieve other documents similar to it, is the one we choose for this implementation. The MoreLikeThis parser finds the top unique terms in the input document, and uses these terms to retrieve related documents. A number of customizations are allowed by MoreLikeThis at runtime, such as the number of terms to be searched and the "boost" or importance of Solr fields (e.g. abstract - 10). We heavily exploit the customization features which will be described in the next sections.

Using the Apache Unstructured Information Management Applications (UIMA) framework and Apache OpenNLP library, we extract different parts of speech from the patent text and introduce them as Solr fields to the index. Other fields introduced include noun n-grams, nouns and chemicals.

3 Customizing Search

The objective of the user interface is to expose the underlying search algorithm and patent related processing in a manner that is simple and easy for a patent expert to understand and manipulate. The various elements of the user interface that help control the search algorithm are detailed in this section.

3.1 Weights for Patent Sections

A considerable amount of research has been done on how the words in the different sections of a patent [3] such as claims and abstract, affect the relevancy of the results. By allowing the users to select which sections of the patent they would like to use for matching and how much relative weight to assign to the words in these sections, we are able to allow users to optimize their search.

As shown in Fig. 2, we expose the different sections of the patents—abstract, claims, title and description—as options that can be checked or unchecked, and the sliders next to them control the weights. The option of choosing the patent sections to compare is a powerful tool, as for example the claim-to-claim comparison of patents can provide insights into double patenting, a phenomenon where the same invention is attempted to be patented more than once by the applicant. We can offer this flexibility to the user by indexing the various patent sections as different Solr fields and assigning different boost values to these fields through the payload to the search call.

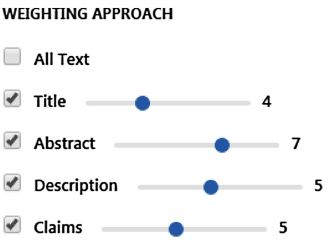

Fig. 2. Sections of patents are chosen with different weights assigned through sliders

3.2 Natural Language Processing

Gerunds in claims have been studied to be an indicator of the strength of the claims. Other parts of speech, such as noun phrases [10], are shown correlated to the relevancy of patent similarity results. However, the benefits to emphasizing on different parts of speech are not uniform across the different patents subject matters; for example, the business methods patents tend to have a large portion of the invention described in verbs instead of nouns. The user interface exposes the different part of speech and chemical matching as options that the user can control as required by the subject area of interest.

3.3 Input Words

Select words from the patent application are used for matching with the corpus of patents available. As we have seen in the previous sections the weights influence the ranking of these words. The words chosen and their corresponding weights are displayed as a word cloud at the left margin of the interface as shown in Fig. 3. In this initial effort, we are able to let the user choose the number of top words to consider, and we limit it to the 25–500 range in order to avoid overloading the system. Users can explicitly exclude any of the words in the tag cloud in order to restrict the results. Also, a limited list of synonyms can be inputted and WordNet [9] synonym support maybe turned on if desired.

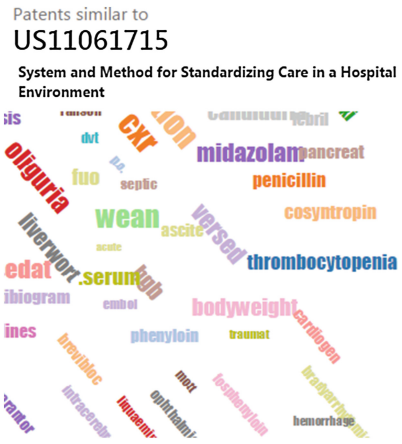

Fig. 3. Tag cloud of interesting terms in the patent US11061715

4 Validation Indicators

To follow up on our approach of letting patent experts to fine tune the patent search algorithm, we expose visual cues that make it easy for the experts to judge the quality of the results. Each result has a number of indicators which show signs of how the result is related to the patent application. If a patent is listed as being in the same family as the patent application, indicator is colored in. Patent documents within a family tends to be the most related to each other with a high degree of similarity. The CPC (Cooperative Patent Classification) and USPC (United States Patent Classification) indicators are shown active if the patent application shares at least one CPC or USPC class respectively with the patent application. Any overlap in the patents cited by the result and the patents cited by the patent application results in the reference indicator being on. The shared art unit indicator reflects if the input patent application and the result have been assigned to examiners in the same art unit.

5 Preliminary Results

The United States Patent Office created a test corpus of 100500 patent documents with 60300 granted patents and 40200 pre-grant publications. 8 sample searches across the different technical domains has been conducted by specialists in the fields, and the handpicked best references has been made available for testing.

The Sigma tool has been stood up on one Amazon Web Services (AWS) m3 server with 2 CPU, 3.75 GB RAM and 32 GB attached storage. We used 6 different settings of the search algorithm to do preliminary analysis of the performance. The table shows the PRES [5] results across the different technology areas.

Though on average algorithm 3 seems to do the best, as shown in Fig. 4, no one algorithm has shown to be superior across all the various technology areas (areas C & H do well on algorithm 4). These are preliminary results were obtained on a limited dataset, and a small subset of the algorithms available; future work is needed to verify.

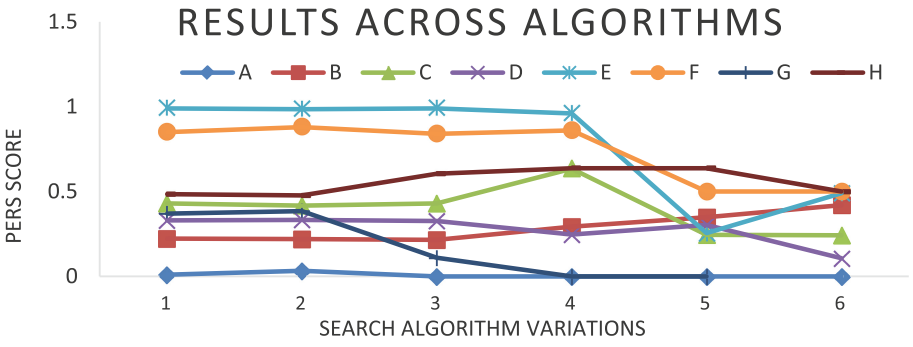

Fig. 4. The chart shows the scores of technology areas A through H on 6 variations of the search algorithms.

6 Conclusions and Future Work

Automated prior art search systems are useful to patent examiners when they are highly accurate. By augmenting simple keyword search algorithms with features that the experts can manipulate, our early results show promise that we can improve one of the most critical, time-intensive aspects of patent examination.

Our preliminary findings are based on a limited corpus of only patent applications. Given the importance of navigating the entire ocean of prior art, we will soon be expanding this corpus to include non-patent literature such as scientific research papers and foreign documents. We will investigate ways to leverage established linkages between prior art references and patent applications to improve the search algorithm results. These results also suggest that the customized algorithms have different optimizations for different settings. Future research will involve exploring how examiners alter the behavior of the search algorithm and ways to validate how examiners use the results. By capturing these settings, we will explore the development of a feedback loop to optimize the behavior of the search algorithm using a "committee of experts" approach and to facilitate knowledge transfer between examiners.

Acknowledgements. We would like to thank David Chiles, Joseph Bailey, Jamie Kucab, Aaron Pepe, Zoin Amir, Arva Adams, Jamie Simpson, Brigit Baron, Terrel Morris and Britt Hanley for their support.

References

1. Bashir, S., Rauber, A.: Improving retrievability of patents in prior art search. In: European Conference on Information Retrieval, pp. 457–470 (2010)
2. Fujii, A.: Enhancing patent retrieval by citation analysis. In: Proceedings of SIGIR 2007 (2007)
3. D'hondt, E., Verberne, S.: Conference and Labs of the Evaluation Forum and Intellectual Property (CLEF-IP) 2010: Prior Art Retrieval using the different sections in patent documents. In: Braschler, M., et al. (2010)
4. Herbert, B., Szarvas, G., Gurevych, I.: Prior art search using international patent classification codes and all-claims-queries. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 452–459. Springer, Heidelberg (2010)
5. Magdy, W., Jones, G.J.: PRES: a score metric for evaluating recall-oriented information retrieval applications. In: Proceedings of the 33rd International ACM (2010)
6. Smiley, D., Pugh, E., Parisa, K., Mitchell, M.: Apache Solr 4 Enterprise Search Server, 1st edn. Packt Publishing, Birmingham (2014)
7. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inf. Syst. **26** (2008)
8. Verberne, S., D'hondt, E., Oostdijk, N., Koster, C: Quantifying the challenges in parsing patent claims. In: Proceedings AsPIRe (2010)
9. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
10. Andersson, L., Mahdabi, P., Hanbury, A., Rauber, A.: Exploring patent passage retrieval using nouns phrases. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 676–679. Springer, Heidelberg (2013)