2010

Automatic Learning of A Supervised Classifier for Patent Prior Art Retrieval

Hung-Chen Chen
National Taiwan University, JesseHCChen@gmail.com

Yu-Kai Lin
National Tsing Hua University, yklin@mx.nthu.edu.tw

Chih-Ping Wei
National Taiwan University, cpwei@im.ntu.edu.tw

Chin-Sheng Yang
Yuan-Ze University, csyang@saturn.yzu.edu.tw

AUTOMATIC LEARNING OF A SUPERVISED CLASSIFIER FOR PATENT PRIOR ART RETRIEVAL

Hung-Chen Chen, Department of Information Management, National Taiwan University, Taipei, R.O.C., JesseHCChen@gmail.com

Yu-Kai Lin, Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan, R.O.C., yklin@mx.nthu.edu.tw

Chih-Ping Wei, Department of Information Management, National Taiwan University, Taipei, Taiwan, R.O.C., cpwei@im.ntu.edu.tw

Chin-Sheng Yang, Department of Information Management, Yuan Ze University, Taoyuan, Taiwan, R.O.C., csyang@saturn.yzu.edu.tw

Abstract

Prior art retrieval is the process of determining a set of possibly relevant prior arts for a specific patent or patent application. Such process is essential for various patent practices, e.g. patentability search, validity search, and infringement search. To support the automatic retrieval of prior arts, existing studies generally adopt the traditional information retrieval (IR) approach or extend the IR approach by incorporating additional information such as citations, classes of patents. Those approaches only exploit partial information of patents and thus may limit the performance of prior art retrieval. In response, we propose a novel approach which employs comprehensive information of patents and performs a supervised approach for prior art retrieval. Unlike traditional supervised learning approach which requires manual preparation of a set of positive and negative training examples, the proposed supervised technique includes a simple but effective mechanism for automatic generation of training examples. Our empirical evaluation on a large dataset consisted of 52,311 semiconductor-related patents indicates that the proposed supervised technique significantly outperforms the traditional full-text-based IR approach.

Keywords: Prior Art Retrieval, Supervised Learning, Patent Search, Patent Management

1 INTRODUCTION

The rapid advancement of technology and explosion of knowledge-based economy over the last decades have intensified the use of patents as a source of competitive advantage. According to the patent statistics reports[1] conducted by the United States Patent and Trademark Office (USPTO), there are more than 180 thousand new patent grants in 2008 which double the number in 1988. These statistics indicate the increasing importance of patents in business management and administration. Companies in high-tech industries, such as information and communication technology, electronics, biotechnology, and pharmacy, generally spend a lot of R&D budget to obtain their core competence. Patents provide a powerful tool for those companies to prevent others from making, using, selling, or distributing the granted inventions without permission

With the growing importance and usage of patents, there is a pressing and urgent need to develop diverse services for supporting effective patent management. One of the most important patent-related services is patent search. Foglia (2007) categorizes patent searches into four types: informative search, patentability search, validity search, and infringement search. Except informative search, the rest three types of patent searches are mainly supported by prior art retrieval.

Current practice of prior art retrieval still highly relies on manual investigation by domain experts, and thereby is costly and inefficient. Due to the pressing need of supporting automatic prior art retrieval, many studies concentrate on proposing novel and effective techniques. Most of them take the information retrieval (IR) perspective or extend IR approach by incorporating additional information, such as citations and classes of patents. However, they only employ partial information, such as textual contents, citations, and IPC classes, of patents to conduct automatic prior art retrieval. Consequently, the effectiveness of prior art retrieval is limited. In response, we propose a supervised approach which learns a classifier for prior art retrieval on the basis of comprehensive information of patents. Specifically, ten variables are designed to fully exploit useful information of patents. Moreover, unlike traditional supervised learning approach which requires manual preparation of a set of positive and negative training examples, the proposed supervised technique incorporates a simple but effective mechanism for automatic generation of training examples. According to our empirical evaluation results, the proposed supervised technique significantly outperforms its benchmark technique, i.e., a traditional full-text-based prior art retrieval, and the designed mechanism for automatic generation of training examples is effective in supporting the learning of the supervised classifier for prior art retrieval.

The rest of this paper is organized as follows. In Section 2, existing techniques for automatic prior art retrieval are reviewed. In Section 3, the overall framework and detailed design of the proposed supervised technique for prior art retrieval is presented. Design of the empirical evaluation and important evaluation results are illustrated in Section 4. Finally, we conclude in Section 5 with a summary of this study as well as some future research directions.

2 RELATED WORK

In this section, we review several existing techniques related to prior art retrieval, including text-based prior art retrieval, citation-based prior art retrieval, and patent-class-based prior art retrieval.

2.1 Text-based Prior Art Retrieval

Text-based prior art retrieval is the most straightforward approach which represents a patent as a vector described by the features, generally words or phrases, occur in its textual contents (e.g., title,

[1] http://www.uspto.gov/web/offices/ac/ido/oeip/taf/reports.htm

abstract, claim, and description). The similarity between the vector of a specific query patent and the vector of a candidate relevant prior art are estimated by using traditional IR ranking functions, such as BM25 (Robertson et al., 1996) and SMART (Buckley et al., 1995). Text-based prior art retrieval approach is simple and may achieve acceptable effectiveness on determining a set of possibly relevant prior arts from the corpus. However, text-based approach only employs partial information of patents, specifically features in the textual contents, to evaluate the relevance of a query patent and the candidate prior arts and thus may limit the performance of prior art retrieval. In response, some studies extend text-based prior art retrieval approach by incorporating additional information, such as citations or IPC classes, of patents to conduct an either integrated or multi-staged solution for better retrieval effectiveness.

2.2 Citation-based Prior Art Retrieval

A patent is required to cite a list of prior arts and may be cited by other patents. Through the citation linkages, the relationship among patents can be identified and may improve the ranking effectiveness of prior art retrieval. Since the relationships among patents are similar to those among web pages or academic articles, the studies related to link analysis algorithms, such as PageRank (Brin & Page, 1998) and HITS (Kleinberg, 1999), have been employed to extract useful information for improving the effectiveness of prior art retrieval (Fujii, 2007; Tiwana & Horowitz, 2009).

Fuji (2007) proposed a two-stage method by incorporating the citation information into the traditional text-based prior art retrieval technique to enhance the retrieval effectiveness. The first stage adopts the BM25 method to retrieve top-*n* patents with highest content-based similarities to a query patent. In the second stage, a PageRank-like analysis is applied to estimate the citation score for each patent of the *n* patents retrieved in the first stage. Finally, a hybrid score of each patent to a specific query patent is estimated by considering both the content-based similarity and the citation score. Unlike our objective of determining a set of possibly relevant prior arts for a specific patent, Tiwana and Horowitz (2009) proposed a parametric algorithm, namely FindCite, to find prior arts by some query keywords. Specifically, FindCite first collects a set of patents that contain the query keywords and then utilizes both backward citations and forward citations to identify the prior arts that satisfy the given citation criterion.

2.3 Patent-class-based Prior Art Retrieval

Patent classification systems are designed to facilitate patent management and retrieval practice in patent offices. For example, in the USPTO, each patent is manually assigned an original class and a set of cross-reference classes by its patent examiner(s). Since the assignments are based on either application aspect, functionality aspect, or both (Adams, 2001), the patent class information can be employed to determine the possibly prior arts by estimating the patent class similarity between two patents. Takeuchi et al. (2004) utilized the Generalized Cosine Similarity Measure (GCSM) (Ganesan et al., 2003) to measure IPC (i.e., International Patent Classification) similarities of two patents. Finally, the similarity between a patent and a specific query patent is estimated by s weighted sum equation. Their experimental results demonstrate that the hierarchical structure information of IPCs can help to improve the effectiveness of prior art retrieval.

Both citation-based and patent-class-based prior art retrieval approach have the ability to improve the effectiveness of prior art retrieval. However, it is difficult for users to set the appropriate parametric constant to control the weight of the citation score or IPC score without domain knowledge of concerned patent database. Moreover, only partial information, such as textual contents, citations, or IPC classes, of patents is employed. The effectiveness of prior art retrieval may be limited. In response, we propose a supervised approach which automatically learns a classifier for prior art retrieval without the need of human involvement. In addition, comprehensive information of patents is adopted in our proposed technique.

3 SUPERVISED CLASSIFIER LEARNING FOR PRIOR ART RETRIEVAL

To improve the effectiveness of existing techniques, we propose a supervised approach which employs comprehensive information of patents to learn a classifier for prior art retrieval. Unlike traditional supervised learning approach which requires the manual preparation of a set of positive and negative examples as training data, the proposed technique incorporates an intelligent mechanism for automatic generation of training examples. Each training example is then represented using a comprehensive set of variables. On the basis of the set of represented training examples, a supervised learning algorithm is applied to automatically learn a classifier for prior art retrieval. The overall process of the supervised classifier learning is illustrated in Figure 1.

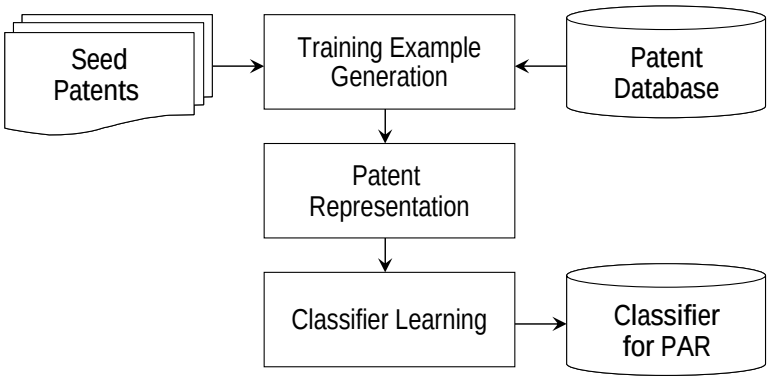

Figure 1. Overall process of the supervised classifier learning for prior art retrieval

3.1 Training Example Generation

Given some granted patents as seed patents, this phase aims at automatically generating positive examples (i.e., prior arts) and negative examples (i.e., non-prior arts) for classifier learning. Each seed patent s_i is adopted to generate a set of prior arts PA_i and a set of non-prior arts NPA_i as positive examples and negative examples, respectively. The selection of PA_i is simple and naïve. Since each seed patent s_i is already granted, the list of cited prior arts in s_i is directly employed as PA_i. On the other hand, those patents highly similar to s_i in content but exactly not prior arts of s_i are employed as NPA_i. Specifically, we first estimate the content-based similarities between the seed patent s_i and all patents in the database. Subsequently, those patents top-ranked on content-based similarities but not cited by s_i are selected as negative training examples. In this study, we select identical numbers of positive and negative training examples (i.e., $|PA_i| = |NPA_i|$) for each seed patent s_i. The union of positive and negative training examples generated by all seed patents are applied for subsequent classifier learning.

3.2 Patent Representation

In this phase, each training patent is represented using a comprehensive set of variables for supervised learning. We design ten variables, i.e., assignee (AS), inventor (IN), original USPC (OR-USPC), cross-reference USPC (CR-USPC), original IPC (OR-IPC), cross-reference IPC (CR-IPC), abstract (ABS), claim (CLM), description (DES), and full text (FT), to fully exploit useful information in patent documents for prior art retrieval. Definitions of the ten variables adopted are discussed in the following.

- **Assignee (*AS*):** *AS* measures whether a patent p_j has the identical assignee to that of its seed patent s_i. Specifically, $AS_j = 1$ if the assignees of p_j and s_i are identical; otherwise, $AS_j = 0$.
- **Inventor (*IN*):** *IN* measures the similarity of inventors between a patent p_j and its seed patent s_i. Since a patent may have more than one inventor, we estimate *IN* by the degree of overlapping

between p_j's inventors and s_i's inventors. That is, $IN_j = \frac{|Inventor_j \cap Inventor_i|}{|Inventor_j|}$, where $Inventor_j$ (or $Inventor_i$) is the set of inventors of patent p_j (or seed patent s_i).

- **Original USPC (*OR-USPC*):** *OR-USPC* measures whether the major U.S. patent class of a patent p_j is identical to that of its seed patent s_i. Each patent is generally assigned with several USPC codes for describing the related areas of this patent. The original U.S. patent class of a patent is its first class code and the remaining class codes are considered as cross-reference USPC. Accordingly, $OR\text{-}USPC_j$ = 1if the major USPCs of p_j and s_i are identical; otherwise, $OR\text{-}USPC_j = 0$.
- **Cross-reference USPC (*CR-USPC*):** *CR-USPC* measures the similarity of cross-reference USPCs between a patent p_j and its seed patent s_i. Since a patent generally has several cross-reference USPC codes which are organized in hierarchical structure, we adopt the generalized cosine similarity measure (GCSM), adopted in Takeuchi et al.'s (2004) study, to estimate *CR-USPC*. Specifically, $CR\text{-}USPC_j = \frac{G(\vec{CR}_i, \vec{CR}_i)}{\sqrt{G(\vec{CR}_j, \vec{CR}_j)}\sqrt{G(\vec{CR}_i, \vec{CR}_i)}}$, where CR_j (or CR_i) is the set of cross-reference USPC codes of p_j (or s_i). $G(\vec{CR}_j, \vec{CR}_i) = \sum_{k=1}^{|CR_j|}\sum_{l=1}^{|CR_i|} G(\vec{c}_k, \vec{c}_l)$, where c_k (or c_l) a cross-reference USPC code in CR_j (or CR_i). $G(\vec{c}_k, \vec{c}_l) = \frac{2 \times depth(LCA(\vec{c}_k, \vec{c}_l))}{depth(\vec{c}_k) + depth(\vec{c}_l)}$, where $LCA(\vec{c}_k, \vec{c}_l))$ is the least common ancestor of c_k and c_l.
- **Original IPC (*OR-IPC*):** In addition to the USPC system, international patent classification (IPC) system is also well-known and commonly applied to classify patents. Similar to USPC codes, a patent is generally assigned with several IPC codes and its first class code and the remaining class codes are considered as original IPC and crsss-reference IPC respectively. We employ the identical method in measuring *OR-USPC* to estimate the $OR\text{-}IPC_j$ of p_j.
- **Cross-reference IPC (*CR-IPC*):** The $CR\text{-}IPC_j$ of a p_j is measured identically to that of $CR\text{-}USPC_j$.
- **Abstract (*ABS*):** *ABS* measures the content-based similarity of abstracts between a patent p_j and its seed patent s_i. Specifically, $ABS_j = \frac{\vec{p}_j \cdot \vec{s}_i}{|\vec{p}_j| \times |\vec{s}_i|}$, where $\vec{p}_j$ (or $\vec{s}_i$) is the feature vector of p_j (or s_i) in the abstract section. In this study, a POS tagger (Tsuruoka & Tsujii, 2005) is adopted to syntactically tag each word in a patent. According to our preliminary experimental results, nouns are representative features and thus are applied to represent a patent.
- **Claim (*CLM*):** *CLM* measures the content-based similarity of claims between a patent p_j and its seed patent s_i. CLM_j is estimated identically to that of ABS_j.
- **Description (*DES*):** *DES* measures the content-based similarity of descriptions between a patent p_j and its seed patent s_i. DES_j is estimated identically to that of ABS_j.
- **Full Text (*FT*):** *FT* measures the content-based similarity of full texts (i.e., abstracts, claims, and descriptions) between a patent p_j and its seed patent s_i. FT_j is estimated identically to that of ABS_j.

For each training example p_j, we represent it with its corresponding values of the ten variables. The represented examples are adopted as the training data for subsequent classifier learning.

3.3 Classifier Learning

A supervised learning algorithm, specifically the Naïve Bayes classifier in this study, is employed to learn the desired classifier for prior art retrieval. Given h variables $f_1, f_2, \ldots, f_h$, the probability that a patent p_j belongs to the set of prior arts A_{si} of a patent s_i is computed via Bayes rule as:

$$p(p_j \in A_{si} \mid f_1, f_2, \ldots, f_h) = p(f_1, f_2, \ldots, f_h \mid p_j \in A_{si}) \frac{p(p_j \in A_{si})}{p(f_1, f_2, \ldots, f_h)}.$$

Assuming statistical independence of the features, the statistics is transformed into:

$$p(p_j \in A_q \mid f_1, f_2, \ldots, f_h) = \frac{\prod_{m=1}^{|h|} p(f_m \mid p_j \in A_{si}) p(p_j \in A_{si})}{\prod_{m=1}^{|h|} p(f_m)},$$

where $p(p_j \in A_{si})$ is a constant, and $p(f_m \mid p_j \in A_{si})$ and $p(f_m)$ can be estimated from the training dataset.

The learned naïve Bayes classifier can then be applied to predict the relevant prior arts of a query patent q. Specifically, for each patent p_j in the patent database, we calculate its values for the ten variables using the method mentioned in Section 3.2 and then represent it accordingly. The represented patents are subsequently submitted to the naïve Bayes classifier to estimate the probabilities of being prior arts of the query patent q. The top-n patents with highest Bayes probability values are returned as the relevant prior arts for the query patent q.

4 EXPERIMENTAL SETTINGS AND RESULTS

In this Section, we describe the experimental settings and discuss important evaluation results of the proposed technique. We first introduce the evaluation dataset and the evaluation criteria in Section 4.1. Subsequently, a tuning experiment of the effect of number of features on our performance benchmark technique, i.e., a traditional text-based approach, is described in Section 4.2. In Section 4.3, the comparative performance of the proposed technique and its benchmark technique is discussed. We also examine the effect of different methods for non-prior art selection in Section 4.4.

4.1 Evaluation Dataset and Evaluation Criteria

To evaluate the effectiveness of the proposed approach, we collect patents, categorized into the USPC class 438 (i.e., Semiconductor Device Manufacturing: Process), from the USPTO patent database. The issued dates of the collected patents range from 1976 to 2005. We randomly select 150 patents issued in 2004 and have at least 10 prior arts as the seed patents of the proposed technique. The average, minimum, and maximum numbers of prior arts of the seed patents are 26, 10, and 279 respectively. In addition, 300 patents issued in 2005 are randomly selected as the testing patents for both the proposed technique and its benchmark technique. The average, minimum, and maximum numbers of prior arts of the testing patents are 15, 1, and 287 respectively. Because the selected seed patents and testing patents may cite prior arts that do not belong to class 438, we also collect and include those patents into our patent dataset. Consequently, the total number of patents in our patent dataset is 52,311.

We use top-n recall rate (Fujii, 2007) as the evaluation criteria to measure the effectiveness of each technique under investigation. Assume that the total number of prior arts of a testing patent q is y and there are x exact prior arts of q among the n retrieved patents. The top-n recall rate of q is calculated as follows: $Recall(q) = \frac{x}{y}$. We then average the top-n recall rates of the 300 testing patents as the final performance.

4.2 Parameter Tuning

In this study, the traditional full-text-based approach for prior art retrieval is adopted as our benchmark technique. Previous studies have shown that feature selection is important for system efficiency and effectiveness. Thus, we conduct a tuning experiment to determine the appropriate number of features for our benchmark technique. On the basis of the BM25 values of features in the patent database, we select top k features with the highest scores to represent the patents. Specifically, we set the number of features (k) as 50, 100, 200, 300, and all and then investigate their effects on the recall rates. The full-text-based approach achieves the best performance when k is 100. Therefore, we

adopted *k* as 100 for the full-text-based approach. Moreover, the identical number of features is selected for the four variables, i.e., *ABS*, *CLM*, *DES*, and *FT*, of the proposed supervised approach.

4.3 Comparative Evaluation Results

Figure 2 illustrates the comparative evaluation results of the proposed supervised approach and its benchmark technique. The supervised approach outperforms the full-text-based approach among all the top-*n* recall rates examined. Among different *n* examined, the average improvement of the proposed supervised approach is 3.23%. Moreover, a statistical testing on the top-*n* recall rates of the supervised and full-text-based techniques indicates that the performance difference is significant with $p < 0.01$. This evaluation result suggests that the inclusion of comprehensive information of patents for prior art retrieval improves the retrieval effectiveness compared with only partial information of patent is adopted. Moreover, although rarely employed in existing studies, supervised learning can be applied to retrieve prior arts and obtain promising results.

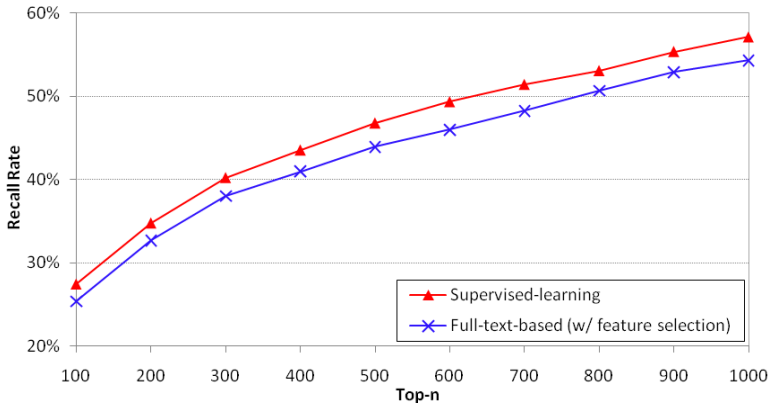

Figure 2. Comparative evaluation results of the two techniques investigated

4.4 Effects of Different Non-prior Art Selection Methods

When selecting non-prior arts as negative training examples for the supervised training, we choose those patents with highest content-based similarities but are not cited by the seed patents. Except the *most similar* method adopted in the proposed technique, we design two additional methods, namely *closest neighbor* and *random*, to examine the effects of non-prior art selection on the effectiveness of prior art retrieval. The closest neighbor method also depends on the content-based similarities to select non-prior arts. However, instead of choosing those non-prior arts most similar to the seed patents, the closest neighbor method chooses those patents whose content-based similarities are ranked right after the exact prior arts but not cited by the seed patents as non-prior arts. On the other hand, the random method randomly selects patents not cited by the seeds patents as non-prior arts. As shown in Table 1, both the closest neighbor method and the random method lead to a very poor performance. The most similar method for non-prior art selection can generate representative negative training examples for our supervised technique to learn an effective classifier.

Method / Recall	Most Similar	Closest Neighbor	Random
R@100	27.41%	0.05%	0.00%
R@200	34.76%	0.18%	0.01%
R@500	46.78%	1.84%	0.03%
R@1000	57.11%	3.18%	0.14%

Table 1. Effects of different non-prior selection methods on the proposed supervised technique

5 CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this study, we propose a supervised approach which learns a classifier for prior art retrieval on the basis of comprehensive information of patents. A simple but effective mechanism is designed to automatically generate a set of positive and negative examples for supervised classifier learning. Moreover, we design ten variables to capture and represent useful information presents in patents. Our empirical evaluation results suggest that the proposed supervised prior art retrieval technique significantly outperforms its benchmark technique. Our evaluation results also demonstrate the effectiveness of the mechanism for automatic generation of positive and negative training examples.

Some ongoing and future research directions are briefly discussed as follows. First, we currently limit the prior arts of patents to be patents issued in USPTO. However, other publication (i.e., foreign patents, scientific literatures, oral lectures, etc.) available to the public qualifies as prior arts. It is essential to include those diverse forms of prior arts into the proposed technique. Second, ten variables are designed to learn the supervised classifier for prior art retrieval. It is desired to incorporate additional variables to extend the effectiveness of the supervised classifier. Last, the proposed technique is only evaluated in a dataset covers the USPC class 438. Examining the effectiveness of the proposed technique using additional datasets from the USPTO database or other patent databases, e.g., European Patent Office and Japan Patent Office, is also one of our future research directions.

Acknowledgement

This work was supported by Boost Program of National Tsing Hua University under Contract Number 99N2534E1.

References

Adams, S. (2001). Comparing the IPC and the US classification systems for the patent searcher. World Patent Information, 23(1), pp.15-23.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7), pp. 107-117.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1995). New retrieval approaches using SMART. In Proceedings of 3rd Text REtrieval Conference (TREC-3), pp. 25-48.

Foglia, P. (2007). Patentability search strategies and the reformed IPC: A patent office perspective. World Patent Information, 29(1), pp. 33-53.

Fujii, A. (2007). Enhancing patent retrieval by citation analysis. In Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, pp. 793-794.

Ganesan, P., Garcia-Molina, H., & Widom, J. (2003). Exploiting hierarchical domain structure to compute similarity. ACM Transactions on Information Systems (TOIS), 21(1), pp. 64-93.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5), pp. 604-632.

Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., & Payne, A. (1996). Okapi at TREC-4. In Proceedings of the 4th Text REtrieval Conference (TREC-4), pp. 73-96.

Takeuchi, H., Uramoto, N., & Takeda, K. (2004). Experiments on patent retrieval at NTCIR-4 Workshop. In Working Notes of the 4th NTCIR Workshop Meeting, pp. 271-276.

Tiwana, S., & Horowitz, E. (2009). FindCite: automatically finding prior art patents. In Proceedings of the 2nd international workshop on Patent information retrieval, pp. 37-40.

Tsuruoka, Y., & Tsujii, J. (2005). Bidirectional inference with the easiest-first strategy for tagging sequence data. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 467-474.