Patent Retrieval Based on Multiple Information Resources

Kan Xu[1], Hongfei Lin[1(✉)], Yuan Lin[2], Bo Xu[1], Liang Yang[1], and Shaowu Zhang[1]

[1] School of Computer Science and Technology, Dalian University of Technology, Dalian, China
{xukan, hflin, zhangsw}@dlut.edu.cn, {xubo2011, yangliang}@mail.dlut.edu.cn
[2] WISE Lab, Dalian University of Technology, Dalian, China
zhlin@dlut.edu.cn

Abstract. Query expansion methods have been proven to be effective to improve the average performance of patent retrieval, and most of query expansion methods use single source of information for query expansion term selection. In this paper, we propose a method which exploits external resources for improving patent retrieval. Google search engine and Derwent World Patents Index were used as external resources to enhance the performance of query expansion methods. LambdaRank was employed to improve patent retrieval performance by combining different query expansion methods with different text fields weighting strategies of different resources. Experiments on TREC data sets showed that our combination of multiple information sources for query formulation was more effective than using any single source to improve patent retrieval performance.

Keywords: Information Retrieval · Query expansion · Learning to rank · Patent retrieval

1 Introduction

The amount of patent information is growing rapidly with an abundant production of digital collection of documents. It is a real challenge to accessing to useful information among this large size dataset. Although patent search engine like Derwent World Patents Index and Google patent search engine have large databases, the search results are not satisfactory. People got very different results when they use different search engines with the same keywords, and they cannot determine which result is more relevant to the keywords. So it is necessary to integrate multiple patent data sources and search methods to improve the performance of patent retrieval.

Automatic query expansion technologies have been widely used in information retrieval (IR) [1–3] In particular, the pseudo-relevance feedback (PRF) which uses query expansion has been proven to be effective [4, 5]. The process of query expansion modified the original keyword query submitted by the user and it would be better represented the underlying intent of the query. The formulated query is then used as an

S. Ma et al. (Eds.): AIRS 2016, LNCS 9994, pp. 125–137, 2016.
DOI: 10.1007/978-3-319-48051-0_10

input to the search engine's ranking algorithm. Thus, the primary goal of query formulation is to improve the overall quality of the ranking presented to the user in response to the query. However, the general query expansion method cannot be introduced directly to special tasks, such as patent retrieval. The patent documents, which are constructed by several special text fields, are different from Web page documents. These fields describe different aspects of patent and have different importance. The traditional expansion methods select candidate terms from the whole document without considering the information from fields which are not suitable for patent retrieval. The existing work [6–8] did not pay enough attention to it. In previous work [9, 10], we proposed a query expansion method, which used patent text fields as the resource of expansion terms, the performance was improved by introducing the field information to query expansion. However, we only use the pseudo-relevance feedback documents for expansion terms. There are still some external information resources which can be used to improve the retrieval performance. It is highly effective to query expansion by using external information resources [11–13].

Learning to rank [14] has become an important research issue for information retrieval. It is an effective approach to improve the ranking performance. The basic premise for learning to rank method is that there are three types of input spaces, they are pointwise, pairwise, and listwise samples. In this paper, we will apply the learning to rank approach to optimize the combination of information sources to improve the performance of patent retrieval.

The remainder of our paper is organized as follows. Section 2 reviews some related work. Section 2 explores the impact of different information resources for patent retrieval. Section 3 proposes the learning to rank based query expansion approach on Derwent World Patents Index and Google search engine for patent retrieval. In Sect. 4, we report the experimental results. Finally, we conclude the paper and discuss future work in Sect. 5.

2 Related Work

2.1 Patent Retrieval

In recent years, researchers show growing interests in patent retrieval. Their research mainly focused on exploring methods on query formulation for topics. Keywords was extracted to form new queries in the early work [15, 16]. Full patent texts were used as the query to reduce the burden on patent examiners which was advocated by Xue and Croft [17]. Text mining, bibliographic coupling and citation analysis were also used in patent retrieval [18, 19]. Chen and Chiu [20] developed an IPC-based vector space model for patent retrieval and achieved a higher accuracy than normal patent search engine. Rusinol et al. [21] presented a flowchart recognition method for patent image retrieval. Recent work showed that the best retrieval results were obtained when using terms from all the fields of the queried patents [22]. It seems that field information is very effective to improve the patent retrieval. However, there are still few works on exploring the text fields to improve query expansion. This paper will use the patent text field information to select candidate terms and improve the results of patent retrieval.

We also investigate the capability of text field of patent in improving the performance of retrieval as promising information for query expansion.

2.2 Query Expansion and External Sources

Pseudo-relevance feedback (PRF) is an effective automatic query expansion method by reformulating the original query using expansion terms from pseudo-relevant documents. Traditional PRF has been implemented in several retrieval models, such as vector space model [23], probabilistic model [24], relevance model [25], mixture model [26], and so on. Meanwhile, there are many research work which focus on improving traditional PRF in different ways. For example, using passages instead of documents [27], using a local context analysis method [1], using a query-regularized estimation method [4], using latent concepts [3], and using a clustered-based re-sampling method for generating pseudo-relevant documents [5]. These methods follow the basic assumption that the top-ranked documents from an initial search contain useful terms that can help discriminate relevant documents from irrelevant ones.

Two external information sources will be employed in our experiment, Google Search Engines and Derwent World Patents Index. Google is one of best search engines in the world, which can provide the accurate information for the users according to the their queries, so we also want to use Google to provide the relevant web pages to expand the query terms for patents. The Derwent World Patents Index (or DWPI) is a database containing patent applications and grants from 44 of the world's patent issuing authorities. Compiled in English by editorial staff, the database provides a short abstract detailing the nature and use of the invention described in a patent and is indexed into alphanumeric technology categories to allow retrieval of relevant patent documents by users. Each record in the database defines a patent family, the grouping of patent documentation recorded at the various patent offices as protection of an invention is sought around the world. Each patent family is grouped around a Basic patent, which is usually the first published example of the invention. All subsequent filings are referred back to the Basic patent as Equivalent patents. The database has some 20 million "inventions", corresponding to ten millions of patents, with almost a million new inventions added each year. Since Derwent database is so effective to the patent research, we will use it as another external information resource to patent query expansion.

2.3 Learning to Rank

Learning to rank approaches can be divided into three categorizations, the pointwise approach, the pairwise approach, and the listwise approach. Different approaches model the process of learning to rank in different ways. They define different input and output samples, using different hypotheses and employ different loss functions. This paper will focus on the construction of samples of listwise approach for further analysis. The listwise approach addresses the ranking problem in a natural way. It takes ranking lists as samples in both learning and prediction. The structure of ranking is

maintained and ranking measures is incorporated directly into the loss functions. More specifically, the listwise approach takes the labeled query-document list as one instance. LambdaMART [28] is the boosted tree version of listwise approach of learning to rank, which is based on RankNet. Boosting and LambdaMART have been shown as the best performing learning methods on public data sets. LambdaMART rankers won Track 1 of the 2010 Yahoo Learning To Rank Challenge. It has been proven to be an effective ranking method for merging the ranking features to improve the performance of retrieval. In this paper, we will use this approach to improve the ranking performance of patent retrieval based on multiple query expansion methods and text fields.

3 Query Expansion Using External Information Resources

3.1 Query Expansion Model

In this section, we introduce our method for patent query expansion. Our query expansion model includes two Rocchio models, one is the original Rocchio model [23], and the other is modified Rocchio model [9].

The original Rocchio model is defined as follows:

$$Q_1 = \lambda * Q_0 + (1 - \lambda) \sum_{r \in R} \frac{r}{|R|} \tag{1}$$

where Q_1 is the expansion query, Q_0 is the original query. R is the pseudo relevance document collection, r is the relevant document. The modified Rocchio model is based on patent fields. In this paper, the model is defined as follows:

$$Q_2 = \lambda * Q_0 + (1 - \lambda) \sum_{r \in R} \frac{\sum_{f \in F} r_f * q_{r_f}}{|R|} \tag{2}$$

where Q_2 is the expansion query, Q_0 is the original query. R is the pseudo relevance document collection, r_f is the field f of the relevant document r. q_{rf} is the weight of r_f. We expand the original queries by this formula.

3.2 Information Resources for Patent Retrieval

The common information resource for pseudo-relevance feedback is the top ranked documents from the corpus with a given query. Relevance feedback takes the results that are initially returned from a given query to perform a new query. The content of the assessed documents is used to adjust the weights of terms in the original query and/or to add words to the query. So the first resource is the TREC data for patent. A patent document is composed of several fields of information, in particular the title, the abstract, the description and the claims. We use these content text fields as research objects to improve the quality of expansion terms. The title field contains the title

of patent. The abstract field contains the text of summary or main idea of a patent. The description field consists of the some sentences about different aspects of a patent content. The claims are the boundary associated with a patent, which is assumed to describe its limits. All the information from the fields may be related to the relevance, and the terms appear in the different fields have different degrees of relevance. So we try to apply the fields to weight the terms for query expansions.

A common web data source from Google for query expansion of patent retrieval is very effective. When the query is submitted to the search engine, the answer is returned in the form of title and abstract texts. The texts and real user search queries are very similar because most title and abstract texts are succinct descriptions of the destination page. The relevant documents for the given query are the second resource of query expansion. The fields we use to query expansion from Google are title and abstract.

The third resource is based on Derwent World Patents Index. The initial set of candidates associated with a query is restricted by considering only those anchor texts that point to a short set of top ranked patents from a larger set of top-ranked patents. These patents can provide more effective information for query expansion. The patent is represented by title and abstract texts. The fields we use to query expansion are title and abstract.

3.3 Term Selection for Query Expansion

For query expansion, there are two steps: select the pseudo relevance document collection R and evaluate the weight of q_f.

In this paper, the pseudo relevance documents come from three information resource: TREC patent data set, Google and Derwent World Patents Index. For TREC patent data set, the first step is the pseudo feedback document selection, which applies three ranking methods for top-k documents: TF*IDF, BM25, BM25F.

TF*IDF [29] contains two variables, term frequency and inverse document frequency. There are various ways to determine the exact values of both variables. For term frequency, the simplest choice is to use the raw frequency of a term in a document, i.e. the number of times that term occurs in a document.

$$w_{t,d} = tf_{t,d} * \log(\frac{N}{n_t}) \tag{3}$$

where $tf_{t,d}$ is the number of times that term t occurs in document d. n_t is the number of the documents which contain the term t. N is the number of documents in the collection.

BM25 [24] is based on the probability model. The retrieved documents are ranked in the order of their probabilities of relevance to the query. A query term is assigned a weight based on its within-document term frequency and within-query frequency. The weighting function used in our experiments is BM25, shown as follows:

$$\omega = \frac{(k_1+1)*tf}{K+tf} * w^{(1)} * \frac{(k_3+1)*qtf}{k_3+qtf} \oplus k_2 * nq * \frac{(avdl-dl)}{(avdl+dl)} \tag{4}$$

$$w^{(1)} = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{5}$$

w is the weight of a query term, N is the number of indexed documents in the collection, n is the number of documents containing the term, R is the number of documents known to be relevant to a specific topic, r is the number of relevant documents containing the term, tf is within-document term frequency, q_{tf} is within-query term frequency, dl is the length of the document, $avdl$ is the average document length, n_q is the number of query terms, the k_i s are tuning constants (which depend on the database and possibly on the nature of the queries and are empirically determined), K equals to $k_1 * ((1-b)+b*dl/avdl)$, and $\oplus$ indicates that its following component is added only once per document, rather than for each term.

BM25F [30] is an extension of the BM25 function to a document description over multiple fields. A key property of this function is that it is nonlinear. Since BM25F reduces to BM25 when calculated over a single field, we will refer to both functions as BM25F, where F is a specification of the fields contained in the document description. In this paper, we use BM25F as the initial retrieval method for feedback documents, which considers multiple fields. BM25F is computed as follows for document d, with a document description over fields F, and query q:

$$S = \sum_{t \in q} TF_t * I_t \tag{6}$$

The sum is over all terms t in query q. It is the Robertson-Sparck-Jones form of inverse document.

We apply the BM25F approach as the initial retrieval method, and select the documents ranking on top-k positions as the candidate collection for the second step. TF*IDF and BM25 are used as baselines for comparison, which rank the documents for top-k pseudo feedback documents without field information, i.e. taking the whole document as a field.

The second step is to decompose every pseudo relevant document generated from the first step into several pieces according to the fields of patent, while each field is regarded as an independent short document. We use the BM25 approach to calculate the relevance between the query and the field document. The relevance score can be seen as the importance of field, which we used to weight the fields of the patent. We also evaluate the importance of each term in the short field document by the query expansion methods, such as TF, TF*IDF, BO1 and BO2 [31]. This analogy suggests us to use the other urn model for IR to obtain alternative methods of expansion for the query, which is the Bose-Einstein statistics. Note that one possible approximation of the Bose-Einstein statistics is given by the geometric distribution G. The probability P

generating the geometric distribution has the same parameter λ = N as the Poisson process. P defined as follows:

$$P = \frac{1}{1+\lambda} \tag{7}$$

The urn model based on BE can be thus used for measuring the information content of terms in the query expansion process giving us:

$$Weight(t) = -log_2(\frac{1}{1+\lambda_{E_q}}) - F_{E_q} * log_2(\frac{\lambda_{E_q}}{1+\lambda_{E_q}}) \tag{8}$$

where F_{Eq} is the frequency of the term and λ_{Eq} is defined by:

$$\lambda_{E_q} = \begin{cases} \frac{F_{E_q}}{N} & [BO1] \\ TotFr_{E_q}.\frac{F_{Eq}}{TotFr_D} & [BO2] \end{cases} \tag{9}$$

where $TotFr_D$ is the total number of term tokens in the collection D. We use these expansion methods to evaluate the relevant importance of a term in the patent fields, which combine the weights of fields to obtain the final weight of the term in the patent document. The finally expanded queries will be used to improve the ranking accuracy.

In this paper, we also take ranking methods and weight evaluation methods as parameters for the patent retrieval method. If there are *M* optional parameter settings for a method, *N* ranking methods and *K* weight evaluation methods, and *L* information resources, the number of features is *M*N*K*L*. The experiments focus on the effectiveness of different forms of patent retrieval methods on learning a ranking model.

3.4 LambdaMart

The performance of patent retrieval system is also evaluated by IR measures such as MAP and NDCG. Learning to rank approaches can define the ranking loss function such as cross entropy loss according to the relevance judgments. By minimizing the loss, it can learn a ranking model to improve ranking performance directly. The aim of query expansion is also to improve the performance of ranking. Therefore, learning to rank can be used to learn a model for query expansion approaches.

LambdaMART combines MART and LambdaRank. MART is a boosted tree model, a linear combination of the outputs of a set of regression trees. LambdaMART utilizes gradient boosting to optimize its loss function defined in the same way as LambdaRank. Gradient Boosting produces an ensemble of weak learner to form a strong one. LambdaRank constructs its loss function based on RankNet, whose loss function is a differentiable function of the model parameters based on cross entropy objective function. The λ for a given document in the ranking list gets contributions from all other documents under the same query with different labels. The λ can also be interpreted as a force, which indicates whether the document should move up or move

Table 1. Example of learning features of TREC-CHEM

ID	(N/M)	Ranking methods	Weight evaluation
1	10/50	BM25	BO1
2	20/100	TF	IDF
3	20/150	BM25F	BO2

down in this round of optimization and also the distance it will move. The λ for a document is the sum of λ_{ij} computed by using the formula as below.

$$\lambda_{ij} = \frac{\partial C(s_i - s_j)}{\partial s_i} = \frac{-\sigma}{1 + e^{\sigma(s_i - s_j)}} |\Delta NDCG| \tag{10}$$

Loss function C has the same form as RankNet based on a probability function combining the score of each document. LambdaRank modifies the gradient with the variation of NDCG through swapping the rank positions of the two documents. LambdaMART uses λ as the gradient of loss function and use boosted regression tree as its model to decrease ranking loss in iterations as MART does. In this paper, we mainly utilize the multiple query expansion methods to extract features for ranking model. We expect that it is effective to improve the ranking accuracies of patent retrieval.

Feature space is constructed by different parameter settings, different ranking methods, and different weight evaluation methods. Overall, there are 18 features, which can be directly used in learning algorithms. The ranking methods include TF*IDF, BM25, and BM25F, and the weight evaluation methods include BO1 and BO2. The example of feature set is shown in Table 1. Table 1 gives some details of implementation of these features, and for the parameter settings *N*/*M* means to extract M expansion terms from *N* documents. *N* can be set to be10, 20, and *M* can be set to be 50, 100, and 150. Ranking methods include BM25, BM25F and TF*IDF. BO1 and BO2 are used as weight evaluation methods.

4 Experiments and Results

In this section, we show the experimental results of query expansion based on patent fields. The TREC-CHEM collection is the experimental data set. We adopt all the topics from TS (Technology Survey) task from TREC-CHEM2010 and TREC-CHEM2011 as our query set. Our research is based on data set of the subtask technology survey. This set contains TS-topics, which is manually created by human experts. Each topic has a description as a natural language expression of information need based on data described in a patent document. The systems should return a set of documents that answer this information need as good as possible. These topics are created to be interesting, so their main priority will be as similar as possible to a genuine information need of an expert searcher. We only use the patent documents in this collection. A patent document is composed of several fields, including title, abstract, description, and claims. These special text fields are used to improve the

quality of expansion terms. For the information resources from Google and Derwent, we select expansion query terms from the title and abstract fields. The 6-fold cross validation is used to obtain the average results. The results are evaluated by mean average precision (MAP) and P@n.

4.1 Effectiveness of Query Expansion Based on Patent Fields

In this section, we conduct the experiment based on TREC data patent fields. We compare the method based on text field for expansion terms (short for TFET) with retrieval methods without query expansion (Original) and the oracle method (use the best feature to rank the documents of test topic of every fold). Table 2 lists the results of these methods.

Table 2. Performance comparison of ranking methods (TFET, Original, and Oracle)

Methods	P@5	P@15	P@20	MAP
Original	0.3333	0.1944	0.1708	0.2173
TFET	0.3833	0.2000	0.1750	0.2342
Oracle	0.3833	0.2278	0.1875	0.2608

From Table 2, we can see that TFET method achieves better performance than original method. Especially for MAP and P@5, the ranking performance of TFET method is much better than Original method, and is similar to the performance of Oracle method in terms of P@5. Results show that query expansion approach based on field information is indeed effective in improving the patent retrieval results. However, TFET is not as good as Oracle method in terms of other evaluation methods. The results of Oracle method come from the best ranking feature of test set of every fold. Therefore, it is feasible to develop a method considering the impact of different ranking features other than using a single ranking feature. Based on these results, the optimization of the query expansion based ranking methods for queries could be expected to further improve the retrieval performance. Now our goal is to develop an effective method to construct a ranking model based on different ranking features.

4.2 Effectiveness of Learning to Rank Model

In order to take full advantage of all the ranking methods, we introduce a learning to rank model: LambdaMART to learn a ranking model from the ranking features. In this section the TFET and Original methods serve as baseline approaches. We will examine the effectiveness of LambdaMart model whose features are extracted from TREC data sets. Table 3 lists the results of the ranking methods.

From Table 3, we can see that the LambdaMart ranking model based on TREC data is superior to TFET method in all of the evaluation methods. Moreover, the relative improvement of LambdaMart is even over that of Oracle method for P@5. And in

Table 3. Performance comparison of ranking methods (TFET, Original, TREC, and Oracle)

Methods	P@5	P@15	P@20	MAP
Original	0.3333	0.1944	0.1708	0.2173
TFET	0.3833	0.2000	0.1750	0.2342
TREC	0.4000	0.2167	0.1875	0.2469
Oracle	0.3833	0.2278	0.1875	0.2608

terms of P@20, it also achieves the same results as the Oracle method. As the information of test set is unknown in the training process and the ranking model is learned from training set as well as the feature selection of TFET, it seems that it is effective to take into account the impact of all the ranking features based on text fields for patent retrieval. It also reveals that the query expansion method based on learning to rank model can improve the ranking performance of patent retrieval.

4.3 Effectiveness of External Information Resources

On above experiments, we only use the TREC data sets for query expansion to extract the features for learning to rank approach. In this section, we also apply the Google and Derwent information resources for query expansion in order to obtain the features for the ranking model. From Table 4, we can see that the LambdaMart ranking model based on TREC data is superior to TFET method in all of terms of evaluation methods. It is also effective to improve the ranking performance by using Google and Derwent information resources. Especially when we use all the features from TREC, Google and Derwent information resources, the ranking model learned from that can achieve the best performance. It seems that it is effective to take the impact of all the information resources based on text fields into account for patent retrieval. It also reveals that the query expansion method based on learning to rank model using multiple information resource can improve the ranking performance of patent retrieval.

Table 4. Performance comparison of ranking methods

Methods	P@5	P@15	P@20	MAP
Original	0.3333	0.1944	0.1708	0.2173
TFET	0.3833	0.2000	0.1750	0.2342
TREC	0.4000	0.2167	0.1875	0.2469
Google	0.4333	0.2722	0.1875	0.2375
Derwent	0.3833	0.2555	0.1875	0.2166
All	**0.4833**	**0.3000**	**0.2541**	**0.2727**

5 Conclusion

In this paper, we explored the multiple information resources for query expansion. For TREC topics, we measure the importance of expansion terms on the retrieval performance. Our experiments show that the query expansion method is an effective

approach for patent retrieval. Furthermore, we investigate the effectiveness of learning to rank model based on the query expansion ranking features. The experimental results demonstrate that, the ranking model which is based on multiple information resources, can effectively cope with the patent ranking problem. In future work, for the pseudo relevant selection method, we will try other retrieval methods to obtain more relevant documents. For the term ranking model, we plan to explore more term ranking methods for further accuracy of patent retrieval.

There are several important differences between our work and previous work on improving query expansion: (1) we examine the effectiveness of different information resources for the patent query expansion; (2) we cast the combination of information sources as an optimization problem that can be solved under a learning to rank framework; (3) we take different query expansion approaches by different resources as features for learning; (4) we apply learning to rank approach with the ranking features to improve the performance of patent retrieval.

Acknowledgement. This work is partially supported by grant from the Natural Science Foundation of China (No. 61272370, 61402075, 61572102, 61572098, 61272373), Natural Science Foundation of Liaoning Province, China (No. 201202031, 2014020003), State Education Ministry and The Research Fund for the Doctoral Program of Higher Education (No. 20090041110002), the Fundamental Research Funds for the Central Universities.

References

1. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 4–11. ACM (1996)
2. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A framework for selective query expansion. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 236–237. ACM (2004)
3. Metzler, D., Croft, W.B.: Latent concept expansion using Markov random fields. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 311–318. ACM (2007)
4. Tao, T., Zhai, C.: Regularized estimation of mixture models for robust pseudo-relevance feedback. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 162–169. ACM (2006)
5. Lee, K.S., Croft, W.B., Allan, J.: A cluster-based resampling method for pseudo-relevance feedback. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 235–242. ACM (2008)
6. Magdy, W., Jones, G.J.: An efficient method for using machine translation technologies in cross-language patent search. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1925–1928. ACM (2011)
7. Ganguly, D., Leveling, J., Magdy, W., et al.: Patent query reduction using pseudo relevance feedback. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1953–1956. ACM (2011)

8. Leveling, J., Magdy, W., Jones, G.J.: An investigation of decompounding for cross-language patent search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1169–1170. ACM (2011)
9. Xu, K., Liu, W., Lin, H., et al.: Patent query expansion using text fields. J. Comput. Inf. Syst. **8**(13), 5607–5614 (2012)
10. Xu, K., Lin, H., Liu, W., et al.: Learning to rank based query expansion for patent retrieval. J. Comput. Inf. Syst. **9**(13), 5387–5394 (2013)
11. Diaz, F., Metzler, D.: Improving the estimation of relevance models using large external corpora. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 154–161. ACM (2006)
12. Lin, Y., Lin, H., Jin, S., et al.: Social annotation in query expansion: a machine learning approach. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 405–414. ACM (2011)
13. Xu, Y., Jones, G.J., Wang, B.: Query dependent pseudo-relevance feedback based on wikipedia. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 59–66. ACM (2009)
14. Liu, T.-Y.: Learning to rank for information retrieval. Found. Trends Inf. Retrieval **3**(3), 225–331 (2009)
15. Konishi, K.: Query terms extraction from patent document for invalidity search
16. Itoh, H., Mano, H., Ogawa, Y.: Term distillation in patent retrieval. In: Proceedings of the ACL-2003 Workshop on Patent Corpus Processing, vol. 20, pp. 41–45. Association for Computational Linguistics (2003)
17. Xue, X., Croft, W.B.: Transforming patents into prior-art queries. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 808–809. ACM (2009)
18. Liu, S.-H., Liao, H.-L., Pi, S.-M., et al.: Development of a patent retrieval and analysis platform–a hybrid approach. Expert Syst. Appl. **38**(6), 7864–7868 (2011)
19. Mahdabi, P., Crestani, F.: The effect of citation analysis on query expansion for patent retrieval. Inf. Retrieval **17**(5–6), 412–429 (2014)
20. Chen, Y.-L., Chiu, Y.-T.: An IPC-based vector space model for patent retrieval. Inf. Process. Manag. **47**(3), 309–322 (2011)
21. Rusiñol, M., de las Heras, L.-P., Terrades, O.R.: Flowchart recognition for non-textual information retrieval in patent search. Inf. Retrieval **17**(5–6), 545–562 (2014)
22. Wanagiri, M.Z., Adriani, M.: Prior art retrieval using various patent document fields contents. In: CLEF (Notebook Papers/LABs/Workshops) (2010)
23. Rocchio, J.J.: Relevance feedback in information retrieval. In: Proceedings of the Smart Retrieval System, pp. 313–323 (1971)
24. Robertson, S.E., Walker, S., Beaulieu, M., et al.: Okapi at TREC-4. In: Proceedings of the Fourth Text Retrieval Conference, pp. 73–97. NIST Special Publication (1996)
25. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 120–127. ACM (2001)
26. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 403–410. ACM (2001)
27. Yeung, D.L., Clarke, C.L., Cormack, G.V., et al.: Task-specific query expansion (MultiText Experiments for TREC 2003). In: TREC, pp. 810–819 (2003)
28. Wu, Q., Burges, C.J., Svore, K.M., et al.: Adapting boosting for information retrieval measures. Inf. Retrieval **13**(3), 254–270 (2010)

29. Salton, G., Wong, A., Yang, C.-S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
30. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 42–49. ACM (2004)
31. Amati, G.: Probability Models for Information Retrieval Based on Divergence from Randomness. University of Glasgow (2003)