York University at TREC 2009: Chemical Track

Jiashu Zhao[1], Xiangji Huang[1], Zheng Ye[1,2], Jianhan Zhu[3]
[1] Information Retrieval and Knowledge Managment Lab, York University, Toronto, Canada
[2]Information Retrieval Lab, Dalian University of Technology, Dalian, China
[3]University College London, London, UK
jessie@cse.yorku.ca, {yezheng,jhuang}@yorku.ca, jianhan.zhu@ucl.ac.uk

Abstract

Our chemical experiments mainly focus on addressing three major problems in two chemical information retrieval tasks, Technology Survey (TS) task and Prior Art (PA) task. The three problems are: (1) how to deal with chemical terminology synonyms? (2) how to deal with chemical terminology abbreviation? (3) how to deal with long queries in Prior Art (PA) task? In particular, we propose a query expansion algorithm for TS task and a keyword-selection algorithm for PA task. The Mean Average Precision (MAP) for our TS task run "york09ca07" using *Algorithm* 1 was 0.2519 and for our PA task run "york09caPA01" using *Algorithm* 2 was 0.0566. The evaluation results show that both algorithms are effective for improving retrieval performance.

Keywords

Chemical Information Retrieval, Mean Average Precision, Patent, BM25, DFR

1 Introduction

In this paper, we describe the work done by members at York University in Canada and University College London in UK for the TREC 2009 Chemical track. This is the first year that the chemical track is carried out. We participated in both the Technology Survey (TS) and the Prior Art (PA) retrieval tasks of the Chemical track. Our goal of participating in this year's TREC Chemical track is to evaluate Information Retrieval (IR) models and their term weighting functions in the chemical domain, and to address the challenges in searching large-scale chemical and patent documents.

The test corpus used in this year's chemical track consists of two types of documents, chemical patents and chemical articles. There are 2,648,160 different patents with a total size of 112GB, and 59,000 different articles with the size of 3GB. Two retrieval tasks are provided this year, which are listed below.

1.1 Technology Survey (TS)

The TS task contains 18 short topics, which are generated with the help of the IP experts. These topics have been generously provided by chemical patent experts based on their own experience. The retrieval results from participants to these 18 topics have been evaluated manually by both graduate students specialized in chemistry and the domain experts who provide these topics. The purpose of this task is to understand the weak points of the participating systems and specific areas where effectiveness can be improved. The task aims to find patents or articles describing standard methods about organic, high molecular weight, pharmaceuticals, and inorganic.

1.2 Prior Art (PA)

The PA task contains 1000 long automatically generated topics, each of which is a full patent. The aim of this task is to find relevant patents with respect to a set of 1,000 existing patents. The results were assessed based on existing citations from the 1,000 patents and their family members.

The remainder of this paper is organized as follows. In Section 2, we describe basic retrieval models utilized in this paper. In section 3, we present the algorithm and experimental results for TS task. In section 4, we propose an approach for BM25 based keyword selection for PA task and its experiment results. In section 5, we conclude the paper with a discussion of our findings and future work.

2 Weighting Models

The retrieval documents are ranked in the order of their probabilities of relevance to the query. Search term is assigned weight based on its within-document term frequency and query term frequency. We used two well-known weighting models, BM25 [1] and DFR [2] in this year chemical track.

2.1 BM25

In BM25, search term is assigned weight based on its within-document term frequency and query term frequency [3]. The corresponding weighting function is as follows.

$$w = \frac{(k_1+1)*tf}{K+tf} * \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} * \frac{(k_3+1)*qtf}{k_3+qtf} \\ \oplus \quad k_2 * nq * \frac{(avdl-dl)}{(avdl+dl)} \tag{1}$$

where 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 a specific 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, qtf is within-query term frequency, dl is the length of the document, $avdl$ is the average document length, nq is the number of query terms, the k_is 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. In our experiments, the values of k_1, k_2, k_3 and b are set to be 1.2, 0, 8 and 0.75 respectively.

2.2 DFR

DFR is a generalisation of one of the very first models of Information Retrieval, Harter's 2-Poisson indexing-model [4]. Its weighting function has the following form.

$$\omega = TF * qtf * NORM * log_e(\frac{N+1}{n\_exp}) \tag{2}$$

where w, qtf, N, tf, dl and $avdl$ have the same meaning as in BM25, and

$$TF = tf * log_2(1 + avdl/dl)$$
$$NORM = (tf+1)/(df*(TF+1))$$
$$n\_exp = idf * (1 - e^{-f})$$
$$f = qtf/df$$

where idf is the term's inverse document frequency weight.

3 TS Task

3.1 Query Expansion Algorithm for TS Task

TS task's queries are one or two sentences long, which show research demanding of companies or experts. Some abbreviations of chemical terminologies are also appeared in the queries. Therefore, an important issue for TS retrieval task is how to expand the queries in the chemical realm so that information retrieval systems can get more information while searching. In the "narratives" of the queries, the chemical experts recommended that "There are other names see ChemID plus or PubChem", which are two well known websites that provide chemical termonology explanation. Therefore, we integrated the professional chemical information from the suggested website ChemID plus [5] and PubChem [6] in our *Algorithm* 1.

Input: a query from chemical experts
Output: a list of possible variants for the chemical terminologies in the query

Method:
(1) Extract keywords from the query by eliminating stop words
(2) Search the extracted keywords on ChemID plus and PubChem, and add their variants in the list
(3) Remove redundant keywords from the list

Figure 1: *Algorithm* 1 for TS task

3.2 Experimental Results

Our experiments were conducted on a double-processor server which has 2 Intel(R) Quad 2.66GHz CPU and 4G memory. York University submitted nine automatic runs in total for the 2009 TREC Chemical track, including seven TS task runs, and two PA task runs.

The Average Precision (AP) of our run "york09ca07" and the average AP over all the participants on the 18 TS topics are shown in Figure 2. We achieved better performance in our run "york09ca07" than the average AP.

We also compared the Mean Average Precision (MAP) of our run "york09ca07" with the average MAP over all participants and the best run in this year in Table 1 . The MAP of "york09ca07" is much higher than the average MAP, and very close to the best run's MAP.

Run	MAP
max MAP among all participants	0.3014
average MAP over all participants	0.1620
york09ca07	0.2909

Table 1: MAP comparison of "york09ca07" and the best and average of all participants

Moreover, the performance of some other official runs of York University is shown in Table 2. In this table, comparisons are given for different approaches. Firstly, the BM25 model achieves better performance than the DFR model under given parameter settings. Secondly, using *Algorithm* 1

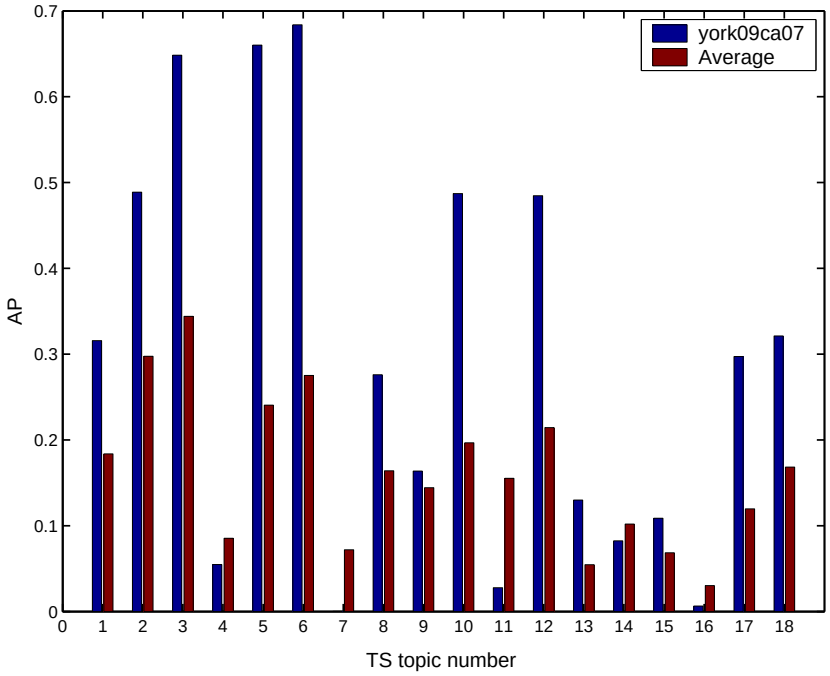

Figure 2: AP comparison on each topic of "york09ca07" and the average over all participants

with extended keywords can improve both MAP for Chemical retrieval models. Therefore, our best run is "york09ca07" which uses BM25 as well as *Algorithm* 1. Most importantly, the performance of all the runs in Table 2 are all significantly better than that of the average's.

Run	Description	MAP	Percentage better than average
york09ca03	BM25	0.2143	32.28%
york09ca04	DFR	0.1709	5.46%
york09ca07	BM25 and *Algorithm* 1	**0.2909**	**79.57** %
york09ca08	DFR and *Algorithm* 1	0.2541	56.85%

Table 2: More results on the 2009 Chemical TS task

4 PA Task

4.1 Keyword Selection Algorithm for PA task

2009 Chemical track also includes another task, PA task, which contains 1000 patents as topics. The uniqueness of this task is that a patent is normally very long, ranges from pages to hundreds of pages. In the mean time, it is not possible for an IR system to process a long topic containing tens of thousands words. Even if a system can handle the query, the retrieved documents will not be highly relevant to the query, due to too much less important information is retrieved. Therefore, an essential question in PA task is to preprocess the topics properly.

In our experiment, we firstly ranked the keywords within a query and extracted the top keywords for further retrieval. A good ranking algorithm should consider both the frequency of the keyword

in the query and its effect on the whole collection [7]. Therefore, we consider to use a part of BM25 as weighting function.

$$\hat{w} = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} * \frac{(k_3+1)*qtf}{k_3+qtf} \quad (3)$$

The parameters are the same as described in Function 1. There are two parts of Function 3, where the first part is calculated based on the term frequency on the whole collection and the second part is calculated based on the term frequency on the query. We aim to lift the keywords that balance these two aspects at the same time. *Algorithm* 2 for PA task is described as following.

Input: a query(patent), the collection, number of output keywords n_k for a query
Output: a list of $n_k$possible keywords

Method:
(1) Extract all the keywords from the query, and put them in a list
(2) Go through the collection, and calculate the first part of Function 3 for each keyword
(3) Calculate the second part of function 3, and therefore the weight $\hat{w}$
(4) Rank the keywords by weight $\hat{w}$
(5) Reserve the top n_k keywords and remove the others

Figure 3: *Algorithm* 2 for PA task

4.2 Experimental Results

Our experimental results of PA task are shown in Table 3. We conducted five runs with different n_k, since it is hard to estimate how many keywords should be included in a query. With the n_k increasing, the MAP decreases, which indicates that more keywords may not be better than fewer keywords in the PA task in 2009 Chemical track.

n_k	20	30	50	70	100
MAP	0.0575	0.0566	0.0516	0.0450	0.0343

Table 3: Performance of our PA task runs with respect to the parameter

5 Conclusions and Future Work

The contributions of our work are as follows. First, we have designed and implemented two algorithms for 2009 Chemical track. *Algorithm* 1 uses website for chemical terminology expansion in TS task, and *Algorithm* 2 introduces the collection's information in PA task query keyword selection. We find that both algorithms are effective for improving retrieval performance in chemical domain. Second, we demonstrate that BM25 model is more effective than DFR model in chemical retrieval domain. Third, we show that less keywords even performs better in Chemical track PA task.

There is still future work to do in chemical and patent IR. One future work is to select more suitable IR model parameters for Chemical track. We only tried a group of default parameters in the adopted weighting models. So parameter selection of BM25 and DFR will be further investigated for improving retrieval performance. Another future work is to make better use of chemical domain knowledge, for a detail example, the structure of chemical compositions. Finally, the term selection

problem in Chemical track could be further refined. We only considered a term's information both on the query and on the whole collection in this paper. Our further work will utilize IR feedback approaches to select the terms on some top ranked documents.

6 Acknowledgements

This research is supported in part by the research grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada. We would also like to thank Michael Siu at Vice-President Research Office of York University for his support and advice.

References

[1] M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker and P. Williams (1996), Okapi at TREC-5. *Proceedings of 5th Text REtrieval Conference*, pp. 143-166, 1996.

[2] G. Amati. Probabilistic models for information retrieval based on divergence from randomness. *PhD thesis, Department of Computing Science, University of Glasgow*, 2003.

[3] Mladen Kovacevic, Xiangji Huang: York University at TREC 2008: Blog Track. *Proceedings of the 17th Text Retrieval Conference*, 2008.

[4] S. P. Harter, A Probabilistic Approach to Automatic Keyword Indexing, *Journal of the American Society for Information Science*, 1975.

[5] ChemID plus. URL address: http://chem.sis.nlm.nih.gov/chemidplus/

[6] PubChem. http://pubchem.ncbi.nlm.nih.gov/

[7] X. Huang, Y. R. Huang, M. Wen, A. An, Y. Liu, J. Poon: Applying Data Mining to Pseudo-Relevance Feedback for High Performance Text Retrieval. *ICDM*, 2006.