York University at TREC 2010: Chemical Track

Jiashu Zhao[1], Xiangji Huang[1], Zheng Ye[1,2]
[1] Information Retrieval and Knowledge Managment Lab, York University, Toronto, Canada
[2]Information Retrieval Lab, Dalian University of Technology, Dalian, China
jessie@cse.yorku.ca, {yezheng,jhuang}@yorku.ca

Abstract

We participated two chemical information retrieval tasks, Technology Survey (TS) task and Prior Art (PA) task in TREC 2010 chemical track. We aim to discover extra relevant chemical compounds for a given query. We investigate various basic retrieval models as well as corresponding Pseudo Relevance Feedback(PRF) models in chemical experiments. In order to eliminate low quality feedback documents, QRocDFR, an early work of York University, is introduced to consider the quality of each feedback chemical document. Further analysis could be made when we were able to reach the evaluation results.

Keywords

Chemical Information Retrieval, Patent, BM25, DFR, Language Model, QRocDFR, RM3

1 Introduction

This is the second year that TREC chemical track has been carried out. This paper describes the work done by members at York University in Canada for the TREC 2010 chemical track, which continues our work in the TREC 2009 chemical track [1]. We participated in both the Technology Survey (TS) and the Prior Art (PA) retrieval tasks. 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 TREC 2010 chemical track data collection is very similar to the one of 2009, but larger. Same as in 2009, the test corpus used in this year's chemical track also consists of two types of documents, chemical patents and chemical articles. There are 1.3 million patents covering patents in the chemical field until 2009, and 181,076 scientific articles from: The Royal Society of Chemistry, PubMed Central, Hindawi Publishing, International Union of Crystallography, Oxford Publishing and Molecular Diversity Preservation International. In addition to last year's plain XML format, TREC 2010 chemical track contains images and chemical structure information (in the form of CDX or MOL files). Our work focuses on retrieval using plain XML chemical documents.

The TREC 2010 Chemical Track contains two ad-hoc retrieval tasks: Technology Survey and Prior Art. The TS task contains 30 short topics, which are generated with the help of human experts. 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. There is also a subtask using title and claim only and a smaller set of topics including 100 patents.

The remainder of this paper is organized as follows. In Section 2, we describe three most popular basic retrieval models. In Section 3, we introduce Pseudo Relevance Feedback (PRF) models:

QRocDFR based on DFR, and RM3 based on KL-divergance retrieval model. In Section 4, we test the introduced models on both TS and PA task in chemical track.

2 Basic Retrieval 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 three well-known basic retrieval models, BM25 [2], DFR [3], and KL-divergence retrieval model [4] in this year's chemical track.

2.1 BM25

In BM25, search term is assigned weight based on its within-document term frequency and query term frequency [2]. The corresponding weighting function is as follows.

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

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

Divergence from Randomness (DFR) is a componential framework that measures the relevance of documents following the probabilistic paradigm [3]. In the DFR framework, the weight of a document d for a given query term t is given by:

$$\omega(d,t) = qtw(t) * IG * (-log_2 Prob(tf)) \tag{2}$$

where IG is the information gain, which is given by a conditional probability of success of encountering a further token of a given word in a given document on the basis of the statistics on the retrieved set. $Prob(tf)$ is the probability of observing the document d given tf occurrences of the query term t. $-log2Prob(tf)$ measures the amount of information that term t carries in d. qtw is the query term weight component, which measures the importance of individual query terms. In the DFR framework, the query term weight is given by:

$$qtw(t) = \frac{qtf(t)}{qtf_{max}} \tag{3}$$

where $qtf(t)$ is the query term frequency of t, namely the number of occurrences of t in the query. qtf_{max} is the maximum query term frequency in the query.

2.3 KL-divergence Retrieval Model

In KL-divergence retrieval model, the query and document are represented as language models [4]. The Kullback-Leibler divergence between the query language model θ_Q and the document language model θ_D is defined as

$$\begin{aligned} D(\theta_Q|\theta_D) &= \Sigma_w P(w|\theta_Q) log \frac{P(w|\theta_Q)}{P(w|\theta_D)} \\ &= -\Sigma_w P(w|\theta_Q) log P(w|\theta_D) + cons(q) \end{aligned} \tag{4}$$

where $cons(q)$ is a document-independent constant that can be dropped, since it does not affect the ranking of documents.

3 Relevance Feedback

In Chemical IR, it shows that specialized relevance feedback models are needed. Chemical compounds are complicated and consisting more than one chemical elements. Chemical compounds can be molecular compounds held together by covalent bonds, salts held together by ionic bonds, inter-metallic compounds held together by metallic bonds, or complexes held together by coordinate covalent bonds. Either way, a chemical compound is related to a serious of chemical reactions, chemical elements, and other chemical compounds. Due to the large amount of chemical terms, it is impossible for human beings to gather all the related terms together for any given chemical compound. Therefore, we introduce two relevance feedback weighting models in this section.

3.1 Relevance Feedback under DFR framework: QRocDFR

QRocDFR [5] was an earlier work of York University, it updates the query term weight component of the DFR framework by considering an expansion terms importance in the pseudo relevance set, has the follows steps:

```
QRocDFR:
  (1) For a given query, performs the first-pass retrieval and considers the R highest ranked
      documents as the pseudo feedback set $D_R$
  (2) Assigns an expansion weight $w(t, D_R) = \Sigma_{d_R \in D_R} w(t, d_R)$ to each term t in $D_R$
  (3) The T most weighted expansion terms are added to the query, using updating formula
      $Q_1 = \alpha * Q_0 + \beta * \Sigma_{r \in R} \frac{r*q(d_r)}{R}$
```

Figure 1: QRocDFR

In step (2), the expansion weight $w(t, D_R)$ is the mean of the expansion weights in each individual feedback document d_R. Document ranking scores from the first pass retrieval is applied to compute the expansion weight $w(t, D_R)$.

In step (3), the quality of each feedback document are considered. It is shown that the retrieval performance will be degraded by the low-quality feedback documents, when the feedback document set size is large [5]. QRocDFR, a quality-biased pseudo relevance feedback (PRF) method, promotes expansion terms in the high-quality documents, and penalizes those in the low-quality documents. $Q_1 = \alpha * Q_0 + \beta * \Sigma_{r \in R} \frac{r*q(d_r)}{R}$ is a quality-baised factor, where $q(d_r)$ is the quality score of feedback document d_r in the R highest ranked documents in the first-pass retrieval. The document quality score is given by the sum of the expansion weight of the original query as $q(d_r) = \sum_{t \in Q} w(t, d_r)$.

3.2 Relevance Feedback under LM framework: Relevance Model

The essential issue in the KL-divergance retrieval model is to estimate θ_Q and θ_D. In general, a feedback language model θ_F is derived to smooth θ_Q [6]. The updated query language model is as follows:

$$\theta_{Q'} = (1 - \alpha) * \theta_Q + \alpha * \theta_F \tag{5}$$

Relevance model is a representative and state-of-the-art approach for estimating query language models within language modeling framework [7]. Relevance models do not explicitly model the relevant or pseudo-relevant document. Instead,they model a more generalized notion of relevance R. The formula of RM1 is:

$$p(w|R) \propto \Sigma_D p(w|\theta_D)p(\theta_D)p(Q|\theta_D) \tag{6}$$

The relevance model $p(w|R)$ is often used to estimate the feedback model θ_F , and then interpolated with the original query model $\theta_Q$in order to improve its estimation. The interpolated version of relevance model is called RM3.

4 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 eight automatic runs in total for the 2010 TREC Chemical track, including seven TS task runs, and one PA task run. For TS task, we did experiments on the basic retrieval weighting models and also their relevance feedback versions. Due to the limitation of number of submissions, we only submitted runs with empirical optimal parameters. The overview of the experimental settings of our submitted runs are shown in Table 1.

Run	Model	Parameter Settings	Decription
york09ca02	BM25	b=0.3, k_1=0.12, k_3=1000	Title
york09ca03	LM	$\mu = 500$	Title
york09ca05	DFR	Parameter Free	Title
york09ca06	QRocDFR	Doc No.=20, Expansion Term No.=30, Parameter= 0.7	Title
york09ca07	BM25	b=0.3, k_1=0.12, k_3=1000	Title & Chemical
york09ca08	LM	$\mu = 500$	Title & Chemical
york09ca09	QRocDFR	Doc No.=20, Expansion Term No.=30, Parameter= 0.7	Title & Chemical

Table 1: Information of Submitted TS runs

Due to the reason that the organizers have not finished evaluation of TREC 2010 Chemical Track TS task, we are not able to show or analyze the results currently. Besides the official runs, we also conducted more experiment, we can further analyze the details after the evaluation is done.

For PA task, the keywords within a query and extracted the top keywords for further retrieval were ranked [1]. We use a part of BM25 as weighting function, in order to consider both the frequency of the keyword in the query and its effect on the whole collection.

$$\hat{w} = \log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)} * \frac{(k_3 + 1) * qtf}{k_3 + qtf} \tag{7}$$

The parameters are the same as described in Function 1. There are two parts of Function 7, 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. The top keywords are therefore extracted for retrieval. The retrieval models and their performance is shown in Table 2. We can see that using both title and claim can achieve better performance than using title only. Using PRF model, QEAdap, can further improve the performance.

Model	Parameter Settings	Description	MAP
DLM	$\mu = 1000$	Title	0.0136
DLM	$\mu = 1000$	Title & Claim (top 40 words for query)	0.0269
DPH	Parameter Free	Title & Claim (top 50)	0.0320
DPH & QEAdap	Parameter Free	Title & Claim (top 50)	0.0339

Table 2: MAP for PA runs

5 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] J. Zhao, X, Huang, Z. Ye, Z., J. Zhu, York University at TREC 2009: Chemical Track, 2009

[2] 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.

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

[4] J. Lfferty, C. Zhai: Document language models, query models, and risk minimization for information retrieval. In: SIGIR 01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2001) 111119

[5] Ye, Z. and He, B. and Huang, X. and Lin, H.(2010), Revisiting Rocchios Relevance Feedback Algorithm for Probabilistic Models. *AIRS*

[6] Zhai, C., La?erty, J.: Model-based feedback in the language modeling approach to information retrieval. In: CIKM 01: Proceedings of the tenth international conference on Information and knowledge management, ACM (2001) 403410

[7] Lavrenko, V., Croft, W.B.: Relevance-based language models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.* pp. 120-127, 2001

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

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

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

[11] 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.