I²R at TREC 2006 Genomics Track

Nie Yu, Yang Lingpeng, Zhang Jie, Su Jian, Ji Donghong
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
{ynie,lpyang,zhangjie,sujian,dhji }@i2r.a-star.edu.sg

Abstract

This paper describes the method we used for the Genomics Track of TREC 2006. BM25 model is implemented to retrieve relevant documents. We also tried to re-ranking documents based on the initial retrieval before passage retrieval. Passages are retrieved based on the concepts defining in topics and concept coverage. Results of submitted runs are listed and discussed.

1 Introduction

The enormous amount of biological literature makes the strong expectation of efficient retrieval ways for biological information. This motivated various research on information retrieval from large scale of information or corpus. The Text Retrieval Conference (TREC) provides a platform for testing and experiments of retrieval methods. In this year, the genomics track of TREC developed a new single task that focuses on retrieval of passages with linkage to the source document.

2 The Passage Retrieval Task

The document collection for the TREC 2006 Genomics Track consists of full-text HTML documents from 49 journals, containing 162,259 documents. There are 28 official topics, with seven topics from each generic topic template (GTT) of Genomics Track 2005. Following is an example for a GTT and an instance of it:

GTT: *Find articles describing the role of a gene involved in a given disease.*

Instance: *Find articles describing the role of Interferon-beta involved in Multiple Sclerosis.*

The target of this task is to submit up to 1000 passages per topic that are predicted to be relevant to answering the topic question. A passage is identified by the document ID(PMID), the start offset into the text file in characters, and the length of the passage in characters. Submitted passages from all track attendants are pooled together. Then the expert judges will be presented with the text of the maximum-length legal span containing each pooled passage. They evaluate and identify the portion of presented text that contains an answer, which is used to measure performance of all submitted runs. There are three levels of retrieval performance measuring: passage retrieval, aspect retrieval, and document retrieval. According to how to generate queries from topics, runs are grouped into “automatic”, “manual” and “interactive”. We submitted three automatic runs for the task of this year.

3 Methods

We expand queries via pseudo relevance feedback. Okapi BM25 [1][2] is implemented to retrieve relevant documents. Single words are used as features with BM25 method. We also tried to exploit document re-ranking in the retrieval process. Figure 1 describes the framework of our system.

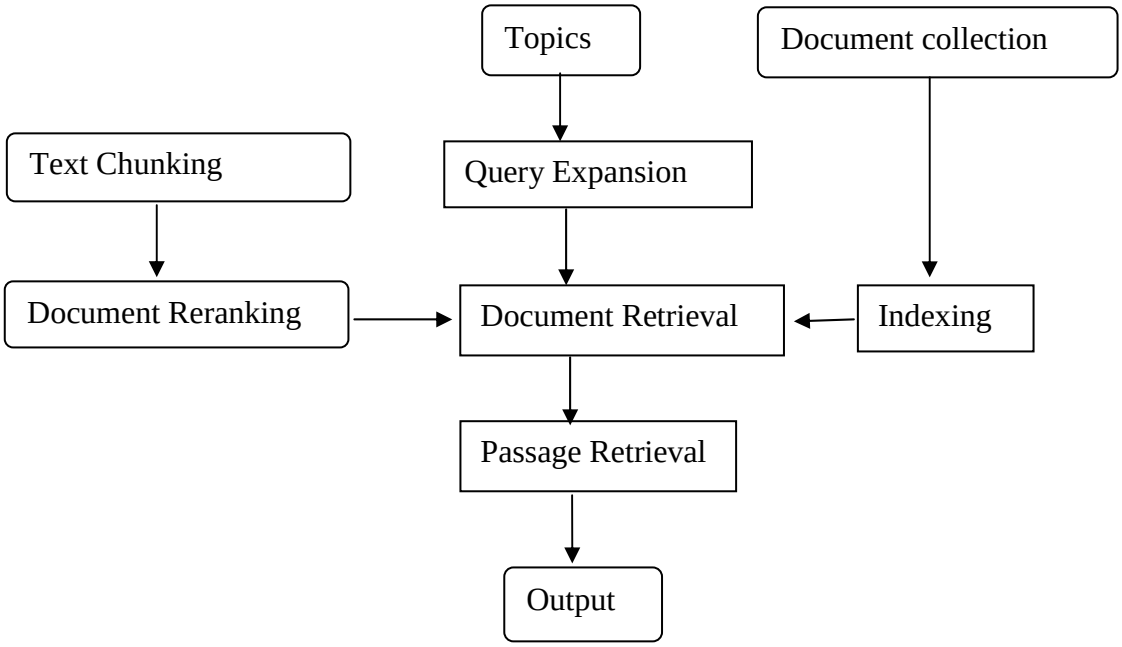

Figure 1: Framework of our system

3.1 Indexing

Documents are indexed before experiments. All HTML files are parsed into plain text files first to remove tags and other format characters. Indexing is made upon the plain text files for all words, including stop words and any continuous letter combination between two delimiter characters, eg. space character. With indexing of single words, indexing of any phrase could be gotten if needed. Since the offset position and length of retrieved answer for topics in the HTML files must be denoted in submitted results, answers retrieved from plain text files are reversed back into position and length in original HTML files after retrieval.

3.2 Document Retrieval

Okapi BM25 [1][2] is implemented to retrieve the top 1000 documents for each query, where a score of each document is calculated as following formula and ranked.

$$\sum_{T \in Q} w^{(1)} \frac{(k_1 + 1)tf}{K + tf} \frac{(k_3 + 1)qtf}{k_3 + qtf} \quad (1)$$

Here $w^{(1)}$ is the Robertson/Spark Jones weight of T in Q:

$$\log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)} \quad (2)$$

Rocchio feedback [3] for BM25 is adopted.

3.3 Document Re-ranking

To re-rank retrieved documents, we use the NP and VP chunks in the documents, and suppose that these chunks will contribute to the re-ranking. Here, we only focus on the chunks which also occur in the queries. So, the chunks can also be referred to as query chunks. To weigh a query chunk, we consider the following three factors.

i) Relative distribution: the ratio of document frequency of a chunk in the top K retrieved documents against the document frequency of the chunk in the whole document collection.
ii) Chunk length: the number of words a chunk contains.
iii) Document ranking position: the serial number of a document in top K documents.

Given top K retrieved documents to be re-ranked and query chunk t, the weight assigned to t is given by the following formula[4].

$$\sqrt{\frac{(\sum_{i=1}^{K} df(t, d_i) \times (1 + 1/i))/K}{DF(t, C)/R}} \times \sqrt{|t|} \quad (3)$$

$$df(t, d_i) = \begin{cases} 1 & t \in d_i \\ 0 & t \notin d_i \end{cases} \quad (4)$$

where d_i is the *i-th* ($i=1, \ldots, K$) document, R is the total number of documents in the whole collection C, $DF(t,C)$ is the number of documents which contain t in C, $|t|$ is the length of term t.

After weighting each query chunk, we can re-order top K retrieved documents by chunks t_j in q and their weightings:

Step 1 For each document d_i in top K retrieved documents, calculate its re-ordered similarity value S_i by its initial similarity value R_i in the initial retrieval;

$$w = \sum_{t_j \in q, d_i} W(t_j) \quad (5)$$

$$S_i = \begin{cases} w \times R_i & (w>0) \\ R_i & (w=0) \end{cases} \quad (6)$$

Step 2: Re-order top K retrieved documents by their new re-ordered similarity values $S=\{S_1, S_2, \ldots, S_i, \ldots, S_K\}$.

3.4 Passage Retrieval

Considering a question as a collection of concepts, we believe that the good answer to a question should be covering all main concepts of the question. So since the target of this year's task is to locate answers for topics accurately, we define the problem of answer

retrieving as finding texts covering main concepts in topics within a limited range in documents. According to the topic templates used for this year's task, we noticed that there are two main concepts for each topic: Gene & Disease, or Gene & Biological Process, or Gene Mutation & Biological Impact. We treated them as two main concepts and other words in the topic as the third topic. For example, given the topic "What is the role of MMS2 in cancer?", the two main topics are "MMS2" and "cancer", and "role" belongs to the third concept.

To retrieve the answer of a topic from documents, we try to find text span no longer than a limited length that covers concepts as many as possible but at least covers the two main topics.

To weigh and rank eligible candidate text spans, we define the text span weight w_s as:

$$w_s = w_d w_t w_l \tag{7}$$

$$w_d = \frac{k_1 + 1/\sqrt{r}}{k_1 + 1} \tag{8}$$

$$w_t = \sum_{concept_i} c_i (k_2 + (1 - k_2)\frac{f_i}{t_i}) \tag{9}$$

$$w_l = \sqrt{\frac{1}{l}} \tag{10}$$

Here k_1 and k_2 are constant parameters, r is the ranking of document containing this text span after the previous document retrieving. Given a concept, c is the weight pre-assigned for the concept, f is the frequency of all concept terms within the text span, t is the number of concept terms. l is the length of the text span. According to formulas mentioned above, w_t is the weight for the text span according to it's concept coverage, w_l makes the shorter text span tend to get higher weight, w_d reflects the influence of ranking of containing documents. We rank text spans by their weight w_s, and discard text spans with weight value smaller than a threshold value.

4 Results and Discussions

We submitted three runs named 'i2rg061', 'i2rg062' and 'i2rg063'. In our experiments the maximum length for eligible candidate text spans was set to 250 bytes. Parameters (k_1, k_2, c_1, c_2, c_3) were set to (2, 0.5, 0.3, 0.3, 0.4). Table 1 lists the MAP of 3 runs evaluated by TREC organizer.

	MAP_document	MAP_passage	MAP_aspect
i2rg061	0.215	0.047	0.081
i2rg062	0.222	0.045	0.076
i2rg063	0.214	0.045	0.080

Table 1: MAP of our submitted runs

To make the answer as short as possible, we only take text spans between two concept terms. The first run and second run both fetch text span begins with a term of a main concept and ends with a term of another main concept. Meanwhile the third run fetch text span begins and ends with any concept term, but still terms from 2 main concepts must be contained. Document re-ranking was exploited in the second run i2rg062, but not used in

the first and third run. From the evaluation results we found that there was little difference between runs.

For the task of this year, we focused on accurate text span retrieval for given topics. Our method is based on the concepts defining in the topic and concept coverage of text spans. We also believe that the ranking of documents could contribute to rank candidate text spans. Because of the lack of training set, we set the parameters directly. The evaluation data set generated from the track of this year would be very helpful to further investigate the efficiency of our passage retrieval method and the influence coming from document retrieval in the process.

5 References

[1] Robertson, S.E. and Walker S. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In Proceedings of the 1994 ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 232-241.

[2] Robertson, S.E., Walker S., Jones S., Hancock-Beaulieu, M.M. and Gatford, M. 1995. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference(TREC-3), NIST Special Publication 500-225, Washington D.C., 109-126.

[3] Rocchio, J.J. 1971. Relevance feedback in information retrieval, In The SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton ed. Prentice-Hall, Englewood Cliffs, NJ, 313-323.

[4] Yang Ling Peng and Ji Dong Hong. Improving Retrieval Effectiveness by Using Key Terms in Top Retrieved Documents. Proceedings of the 27th European Conference on Information Retrieval, ECIR 2005, LNCS 3408, pp.169-184.