UM-D at TREC 2005: Genomics Track

LiPing Huang, ZhiHang Chen, and Yi Lu Murphey,
Department of Electrical and Computer Engineering
The University of Michigan-Dearborn
Dearborn, Mi 48128-1491
yilu@umich.edu, 313-593-5028

1 Introduction

The University of Michigan-Dearborn team participated in the ad hoc task and submitted two runs in TREC 2005. The Genomics track is different from others since it focuses on document retrieval in genomics domain as opposed to general retrieval tasks such as question-answering, multi-lingual IR, etc. Since we were not familiar with the knowledge in biomedical field, we utilized the database publicly available online to obtain alias and variations of names for genes/proteins. We generated a term list based on each topic description and their alias and variations. The terms were further transformed into a logical expression in which terms were connected by "AND" and "OR". The documents satisfying the logical expression are retrieved and their similarity scores are calculated based on the weighted terms using the method of Okapi BM25 proposed by Robertson et al[RWJ94][RWB98] [BCC04].

2 Ad hoc Retrieval Task

2.1 Overview

The genomics track of TREC 2005 consisted of two tasks: ad hoc retrieval task and categorization task. The ad hoc task is a conventional searching task, which is designed to retrieve documents that are relevant with respect to certain topics in a subset of medical publications. The document collection for this task is a 10-year subset of the MEDLINE bibliographic database of the biomedical literature, which consists of a total of **4,591,008** documents. There is no training data for this task. However, sample topics and relevance judgment are provided.

50 topics collected from real biologists were provided to the participants. These topics are structured in 5 templates. The topics in template 1 contain statements describing standard methods or protocols for doing some sort of experiment or procedure, topics in template 2, 3, 4, and 5 are names of genes and processes and diseases. This characteristic of lacking of details motivated us to use logic expressions rather than term lists.

The following briefly describes the methods and experiments conducted that lead to our two final submitted runs.

2.2 Architecture

The architecture of our system is shown in Figure 1. The detail of each section is described in the following.

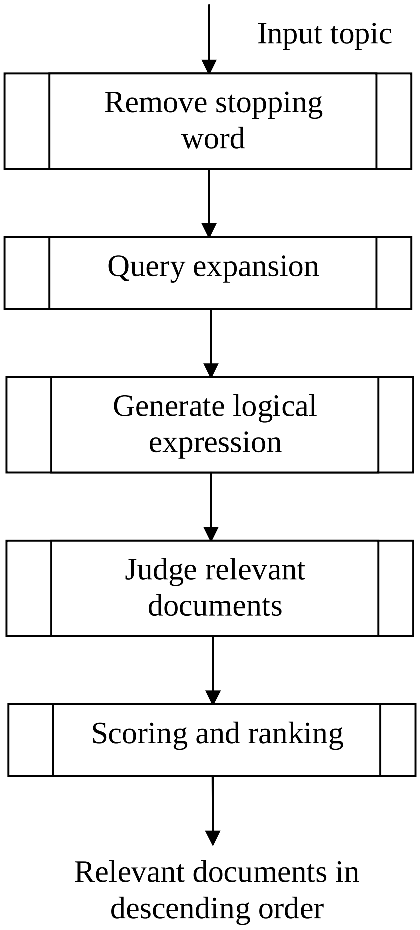

Figure 1. Architecture of System for Ad Hoc Retrieval Task

2.2.1 Query expansion

Given a topic, stopping terms such as a, the, that, etc. were first removed from the query. The system then extracted a list of important keywords and noun phrases from the description of topic. This is a semi-manual process. All fields except ID were used in this phase. In order to expand original query, this list was used as search keys on the medical databases described below to get synonyms or aliases of keywords and noun phrases. There are mainly three types of synonyms that can appear in topics of this year:

- acronyms for standard method or protocol

  For example, IP represents immuno precipitations

- name of gene/protein

  For example, the name of gene APC is adenomatous polyposis coli, and GFP represents green fluorescent protein.

- name of disease

  For example, CAA represents Cerebral Amyloid Angiopathy

Acronyms can be dealt with in a straightforward fashion by adding the respective long form to the topic whenever a known acronym is encountered.

In addition to PubMed (a service of the National Library of Medicine, which can provide access to MEDLINE as well as a dozen other databases), there are additional genomics resources online which can provide rich annotation and linkage:

- The AcroMed database of biomedical acronyms
- The Eukaryotic Genes database at the University of Indiana
- Saccharomyces Genome Database

From the retrieved documents of search engine, appropriate synonyms/aliases were selected to add in the list for further expansion. In addition, some variations of gene or protein names were generated and added into the list in order to tolerate minor typos or variations of naming. Space or other symbols were added or deleted between character and number in the name of genes/proteins. For example, Cop 1 and Cop-1 both indicate Copolymer 1. HPV16 and HPV 16 can both represent human papillomavirus type 16. Expanded queries were generated based on final version of the list.

2.2.2 Logical connectives

The next step is to combine terms and noun phrases in the expanded query with logic connectives AND or OR. The logical connection between multiple terms was defined as AND if all of them are required to appear in the actual query or in a document. The "and" that existed in the original query was transformed into AND in logical expression. Other relations such as the role of a gene in a specific biological process was also transformed into logical connective of AND. On the other hand, one concept may have multiple variations or synonyms, OR was defined as a logical connective between concept and its synonym. The "or" that existed in the original query or in the description of GTT was transformed into OR in logical expression. In some scenarios, for example, one or more mutations of a given gene were required to be included in the retrieved documents. After the original query was expanded with gene's mutations, OR was added between the names of mutations.

2.2.3 Relevant document

The rule for judging relevance of document in our system is to satisfy logical expression and term matching as well. The system first judges if each document from a collection satisfies the logical expression of the query. If this document contains all terms (i.e. term matching) that occurred in the chain connected by AND, and at the same time it matches any chain of terms that connected by OR, the document is regarded as relevant. As a result of the logic expression and term matching, a collection of documents matching the topic is retrieved. These documents are further ranked using the similarity score proposed by S.Robertson et al. The documents are sorted in descending order of similarity scores. The top 1000 documents (if there are) are returned as relevant ones for the topic.

Several experiments were conducted to refine the logic expressions for the 50 topics. For each topic, we ran the list of logic expressions of keywords and phrases derived using the method above on the test document collection. If the return has too many documents, we tried to remove the less specific keywords from the expression, change logical operators to restrict the search to more specific meanings, and limit the keywords that are allowed for partial matching. If too few documents were retrieved we look for new key words or variations of keywords to add to the logic expression. This process is also assisted by the method used in the second run to estimate the size of the return documents for each topic.

Finally we used the sample data set to evaluate our two runs. The result indicated the first run has generated better results.

2.2.4 Similarity ranking

Once a document is retrieved as relevant to a specific topic, we use the following Okapi BM25 method to assign a similarity score to the document [RWJ94][RWB98] [BCC04]. For every topic description, we generate a term list including the alias and variations of the words.
Every term T in the term list is assigned a term weight

$$w_T = \ln(\frac{|D| - |D_T| + 0.5}{|D_T| + 0.5})$$

where D is the document corpus provided by the NIST Genomic Track, and D_T is the set of documents containing the term T. For a given logic expression Q = $\{T_1,\ldots,T_n\}$, the score of a document D_i is computed using the formula

$$\sum_{T \in Q} w_T \cdot q_T \cdot \frac{d_T \cdot (1 + k_1)}{d_T + k_1 \cdot ((1-b) + b \cdot \frac{lenD_i}{lenAVE}}$$

where d_T is the number of occurrences of the term T in the document D_i, $lenD_i$ is the length of the term list for D_i, lenAVE is the average document length in the corpus, and q_T is the query-specific relative weight of the term T. Usually, q_T equals the number of occurrences of T in the input query. We set $k_1 = 1.2$ and $b = 0.75$.

2.2.5 Second run

From each topic we generate a list of keywords and phrases and their variations as described above. In this run we use the keyword and phrase list to match the documents directly in the 2004_TREC_ASCII_MEDLINE. The matching of the topic query and a document is measured and scored by the Okapi BM25 method described above. The top 1000 documents are returned as relevant ones for each topic.

2.3 Experiment

The performance of our two runs for 50 topics is shown in the table below:

	First run	Second run
MAP	0.1221	0.0544
p@5 average	0.3918	0.1959
p@10 average	0.3224	0.1755
p@100 average	0.1473	0.0843

Table 1. Ad hoc retrieval result for 50 topics

REFERENCE

[RWJ94] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference (TREC 1994), November 1994.

[RWB98] S. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7. In Proceedings of the Seventh Text REtrieval Conference (TREC 1998), November 1998.

[BCC04] Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval. Presented at the *2004 Text REtrieval Conference (TREC 2004)*, Gaithersburg, Maryland, November 2004.