Developing an IR and NLP-Based Algorithm for Biomedical Literature Searching

Yung-Tai Yen, Chien-Yeh Hsu*

Graduate Institute of Medical Informatics, Taipei Medical University, Taipei, Taiwan
*Corresponding author

Abstract

MEDLINE is the largest biomedical journal database in the world. When researchers search for literatures by using Entrez system, they usually retrieve thousands of articles and the query results can only be sorted by PubMed ID, author, journal and publication date. However, users usually need to spend lots of time to review these articles to find out which they are really interested in.

The purpose of this study is to apply Information Retrieval (IR) and Natural Language Processing (NLP) methods to develop an algorithm that can refine the search results of Entrez system. We have built a simple query system, the Biomedical Literature Searching System (BLSS) based on our algorithm. The system outperformed the Entrez for retrieving more relevant documents. In the future, we will build an application to help MEDLINE users retrieve articles that are most likely related to their queries.

Keywords:

MEDLINE; Information Retrieval; Natural Language Processing; Text Processing

1. Introduction

Biomedical information exists in both research literatures and various structured databases, but it is difficult to get knowledge from diverse information sources. Natural Library in Medicine (NLM) provides an integrated journal database, MEDLINE, to help researchers get information from literature search. However, its search engine, Entrez system, doesn't have efficient information retrieval applications. For example, it is not easy to locate most relevant information that the users really want from lots of query results.

In our research, we have used several disease MeSH terms as the examples, such as lymphoma related biomedical literatures, and applied Information Retrieval (IR) and Natural Language Processing (NLP) methods which have been used in computational linguistics to refine the search results of Entrez system.

The specific aims of our research are summarized as follows:

1. To establish a database for the linguistic terms (dictionary and lexicon collection) from MEDLINE articles.
2. To develop a ranking algorithm based on IR methods to refine the query results of Entrez system.
3. To apply NLP models for semantic analysis of MEDLINE articles.
4. To develop a language model and generate context rules for advanced search functions.

2. Architecture

Figure 1 demonstrates the architecture of our research and the following illustrates it more detailed.

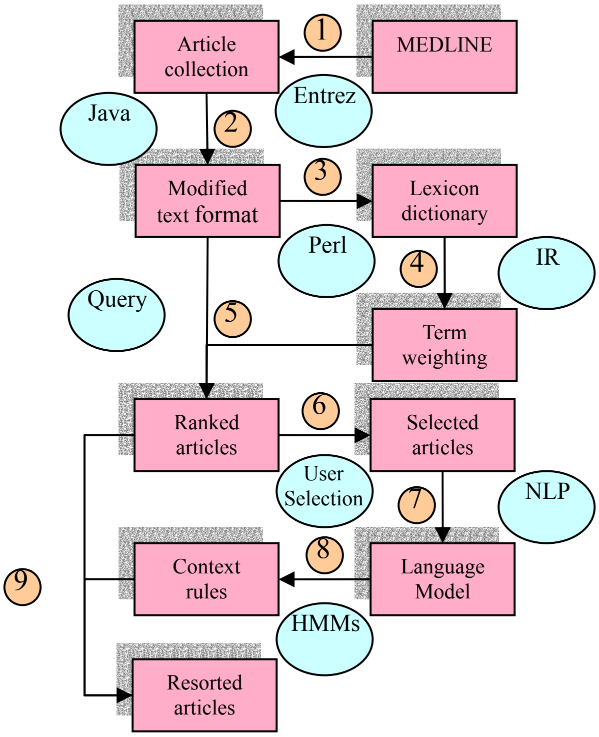

Figure 1: The research architecture.

1. We collect specific topic abstracts from MEDLINE by using Entrez system.
2. Using a Java program to modify the text format of the collection.
3. A lexicon dictionary is built with a Perl program.
4. We also use Perl to implement IR model and then calculate a term–weighting score for each index term.
5. Then we make some queries in our system to find related articles and get ranked articles by these term-weighting scores.
6. Some articles with high term-weighting score are selected.
7. NLP methods are applied to do semantic analysis for those selected articles.
8. A language model will be built by the results of NLP methods and the Hidden Markov Models (HMMs)[3][4].
9. Finally we will generate context rules based on the language model and apply the rules to resort the articles that are ranked by IR model.

3. Method

The procedures of our research are described as follows:

Literature collection and data preparation:

We collected text information about lymphoma for the MEDLINE citation database of the NLM. Those articles are constrained to be the subset of MeSH tree number (C04.557.386). Figure 2 is the tree structure of this subset.

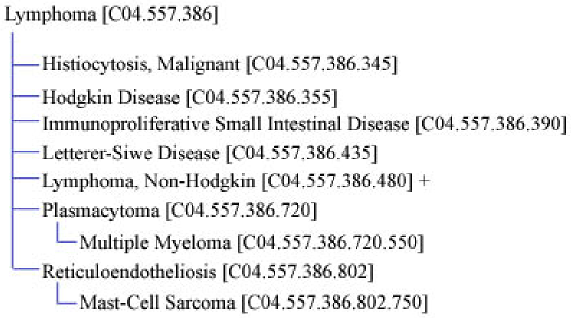
Lymphoma [C04.557.386]
Histiocytosis, Malignant [C04.557.386.345]
Hodgkin Disease [C04.557.386.355]
Immunoproliferative Small Intestinal Disease [C04.557.386.390]
Letterer-Siwe Disease [C04.557.386.435]
Lymphoma, Non-Hodgkin [C04.557.386.480] +
Plasmacytoma [C04.557.386.720]
Multiple Myeloma [C04.557.386.720.550]
Reticuloendotheliosis [C04.557.386.802]
Mast-Cell Sarcoma [C04.557.386.802.750]

Figure 2: The tree structure of the subset C04.557.386

Term-weighting scores of collected articles:

In order to decide which article is most likely related to users' query keywords, we applied methods such as Vector Space Model[2][3] to calculate a term-weighting score for each word excludes stop words and built a ranking algorithm for sorting the articles.

Some terms of Vector Space Model are defined as:

N: the number of total documents.

$W_{i,j}$: the weighting of index term k_i in document d_j.

$\vec{dj}$: the document vector ($w_{1,j}, w_{2,j},\ldots,w_{t,j}$).

In the Vector Space Model, $W_{i,j}$ is influenced by two factors:

1. The "*tf*" factor: the frequency of index term k_i in document d_j (intra-document).
2. The "*idf*" factor (inverse document frequency): the frequency of index term k_i between documents (inter-documents). Thus, $w_{i,j} = tf_{i,j} \times idf_i$.

If n_i is the number of index term k_i in documents and $freq_{i,j}$ is the raw frequency of k_i in document d_j ,the normalized frequency of k_i document d_j is

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}}.$$

If idf_i is the inverse document frequency of k_i, we have $idf_i = \log\frac{N}{n_i}$.

The term-weighting of k_i is,

$$w_{i,j} = f_{i,j} \times \log\frac{N}{n_i}.$$

The result of this model will be a procedure that can easily select the articles that are highly related to users' query.

Semantic analysis and language model:

The N-Gram Model is used for contextual analysis to calculate N-Gram probabilities of each sentence contained users' query keywords in the selected articles. To implement N-Gram Model, we used the Unigram and Bigram Model to estimate the N-Gram Model.

For the Unigram Model, we calculated the unigram probability of each word:

$$P(W_i) = \frac{N(W_i)}{\sum_{j=1}^{|V|} N(W_j)}. \quad \forall W_i \in V, i \in [1,\ldots,|V|]$$

For the Bigram Model, we calculated the bigram probability of each word:

$$P(W_n | W_{n-1}).$$

In order to estimate the N-gram Model, we combined the bigram sequence, which is:

$$P(W_1^n) \approx \prod_{k=1}^{n} P(W_k | W_{k-1}).$$

Then we will apply HMMs to build a language model [5] by the result of the N-Gram Model. Figure 3 is the flowchart of the language model.

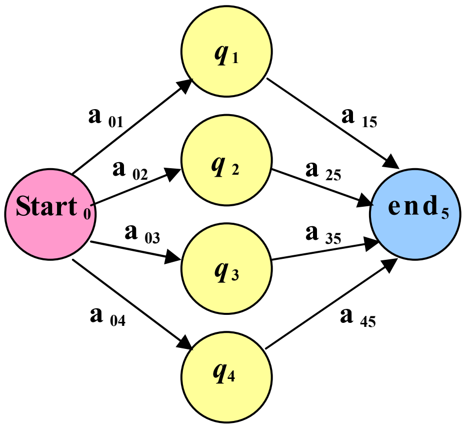

Figure 3 : The language model based on HMMs.

There are four states and eight transition probabilities in this model:

1. q_1 : The unigram probability of word (W_i) in the document (d_j).
2. q_2 : The bigram probability of word sequence (W_1^n) in the document (d_j).
3. q_3 : The unigram probability of word (W_i) in all documents (D).
4. q_4 : The bigram probability of word sequence (W_1^n) in all documents (D).

Context rules:

Incorporating with the language model and the sentences contained keywords, we will generate the context rules as follow:

$$P(W_1^n | LM) = \alpha P(W_i | d_j) + \beta P(W_1^n | d_j) + \gamma P(W_i | D) + \delta P(W_1^n | D).$$

Therefore, we will calculate the probabilities of context rules based on the ranked result generated by IR methods. According to these probabilities, the ranked result will be resorted and the article with higher probability will be expected to be more likely related to user's query.

4. Results

We have collected 37810 abstracts under the MeSH tree number (C04.557.386) constraint. From this collection, we have calculated a lexicon by eliminating the selected stop words and it contains 4315869 entries. The lexicon can be used as the index terms for the Vector Space Model.

We have implemented the Vector Space Model by Perl and calculated a term-weighting score of each index term. Now we have built a simple query system, the Biomedical Literature Searching System (BLSS), which is based on the algorithm and is ready for the further research.

5. Conclusion

For the purpose of evaluating the performance of BLSS, we have used some gene symbol and immunoglobulin keywords including “bcl-2”, “CD4”, and “IgH” to search for related articles by using Entrez system in MEDLINE, and then the search results were queried again by BLSS corresponding to the term-weighting scores of articles. The 20 preceding outputs of BLSS and Entrez were compared and reviewed by experts in biomedicine. We calculated interpolated precisions at standard recall levels of both systems as in Figure 4. Further more we also calculated Mean Average Precision for BLSS (0.80) and Entrez (0.66). Thus we found that BLSS can retrieve more relevant documents than Entrez system does.

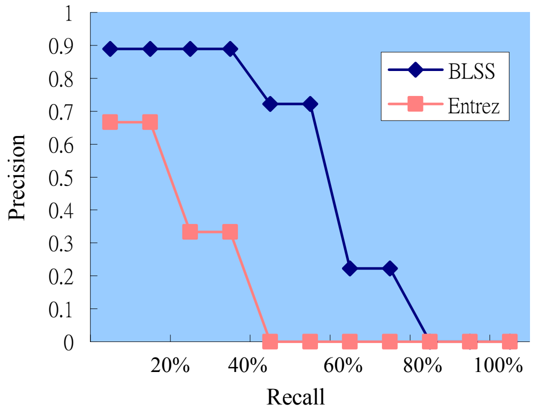

Figure 4 : Interpolated precisions of two systems.

At the mean time, we are evaluating our system by making more complicated queries and use these queries to search for articles in Entrez system. The result will be the baseline answers of BLSS and we can evaluate and tune the system by some indicators such as recall and precision.

After that we will calculate the unigram and bigram probabilities and build a language model to generate context rules. The ranked result will be resorted by the probabilities of specific rules.

6. Reference

[1] D. Jurafsky and J. H. Martin, “Speech and Language Processing”, *Prentice-Hall*, 2000.

[2] Kenney Ng, “Survey of Approaches to Information Retrieval of Speech Messages”, *Draft, SLSG, MIT,* 1996.

[3] R. Baeza-Yates and B. Ribeiro-Neto, “Modern Information Retrieval”, *Addison Wesley Longman*, 1999.

[4] Lawrence R. Rabiner, “A Tutorial on Hidden Markov models and Selected Applications in Speech Recognition”, *Proceedings of The IEEE* Vol. 77, No2, February, 1989.

[5] Djoerd Hiemstra and Franciska de Jong, “Statistical Language Models and Information Retrieval: natural language processing really meets retrieval”, *Glot International* 5(8), pages 288-294, 2001.