MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Applying Clinical Ontology for Biomedical Information Retrieval

Yung-Tai Yen[a], B. Chen[b], Ching-Lih Shyu[c], Yu-Chuan Li[d], Chien-Yeh Hsu[c]

[a] *Graduate Institute of Medical Sciences, Taipei Medical University, Taiwan*
[b] *Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan*
[c] *Graduate Institute of Medical Informatics, Taipei Medical University, Taiwan*
[d] *Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan*

Abstract

Information retrieval (IR) is a technology to help people find information, however it is hard to formulate information need into just a few words. In order to solve this problem, we adopted knowledge ontology in the IR process to design a query expansion. We implemented three IR models, Vector Space Model (VSM), Hidden Markov Model (HMM), and Topical Mixture Model (TMM) and evaluated these models with and without ontology-based query expansion. In conclusion, TMM has the best performance and people who are not familiar with specific domain knowledge, the ontology-based query expansion can help them generate queries and obtain relevant information.

Keywords:

ontology, information retrieval, Vector Space Model, Hidden Markov Model, Topical Mixture Model

Introduction

Due to the advancement of computer technologies, tremendous volumes of biomedical data have been digitized. In order to retrieve useful information from these data, the development of intelligent information retrieval (IR) methods becomes more important. In most IR systems, term matching is a common approach; however no searching result can be obtained if query terms are not found in the system lexicon. For biomedical professionals, they know what keywords can be used to find correct information; for people who don't have enough knowledge, it is difficult to obtain accurate information by using inappropriate keywords. Thus, we developed an ontology-based query expansion adopted in three IR models, Vector Space Model (VSM), Hidden Markov Model (HMM), and Topical Mixture Model (TMM) and evaluated the performance of these IR models with and without ontology-based query expansion.

Methods

Information Retrieval Model

Vector Space Model

Each document, D_i, is represented by a feature vector, $\overrightarrow{D_i}$; each component, $g(t)$, in the vector is associated with the statistics of a specific index term, t,

$$g(t) = \left(1 + \ln\left(c(t)\right)\right)\ln\left(N / N_t\right) \tag{1}$$

where $c(t)$ denotes the occurrence count of term t within document D_i, and the logarithmic operation is used to compress its distribution. The term weighting scheme, $1+ln(c(t))$, is a variation of conventional schemes. N_t is the total number of documents in the collection in which the specific indexing term t appears, and N is the total number of documents in the collection. A query, Q, is also represented by a vector, $\vec{Q}$, in which every component is a word in the query. The cosine measure is used to estimate the relevance between query Q and document D:

$$R(Q,D) = \cos(\vec{Q},\vec{D}) = \left(\vec{Q} \bullet \vec{D}\right) \Big/ \left(\left\|\vec{Q}\right\| \cdot \left\|\vec{D}\right\|\right) \tag{2}$$

HMM/N-gram Model

In our implementation of the HMM/N-gram model, each document is composed of a mixture of N-gram distributions, which can be represented as a special case of HMM. The N-gram distributions are estimated on the basis of frequency of words or word pairs occurring in documents and are smoothed using linear interpolation with background unigram or bigram language models estimated from a large outside text corpus. In this study, we only implemented HMM/Unigram model. For example, given a query Q and a document D, the relevance measure for the HMM/Unigram model can be expressed as:

$$P(Q|D) = \prod_{n=1}^{N}\left[m_1 P(q_n|D) + m_2 P(q_n|Corpus)\right] \tag{3}$$

where $P(q_n|D)$ is the unigram probability of a index term q_n within the document D and $P(q_n|Corpus)$ is the probability of an index term q_n within the large corpus. Weighting parameters, m_1 and m_2 are summed to one and tied among the entire document collection. They are opti-

mized using the expectation-maximization (EM) algorithm.

Topical Mixture Model

Each document D_i is interpreted as a mixture model and a set of k latent topical distributions characterized with unigram language modeling are used to predict the input query words and each of the latent topics is associated with a document-specific weighting parameter.

$$P(Q|D_i) \approx \prod_{n=1}^{N}\sum_{k=1}^{K} P(q_n|T_k)P(T_k|D_i) \tag{4}$$

where $P(q_n|T_k)$ is the probability of the query word q_n occurring in a specific latent topic T_k, and $P(T_k|D_i)$ is the posterior probability of the topic T_k conditioned on the document D_i, with the constraint $\sum_{k=1}^{K} P(T_k|D_i)=1$. The key idea of TMM is that the relevance measure of a query word q_n and a document D_i is not calculated directly based on the frequency of q_n occurring in D_i, but instead based on the frequency of q_n in the latent topic T_k as well as the likelihood that D_i generates the respective topic T_k.

Query expansion based on the ontological structure

The SARS ontology, which was developed by referring to biomedical literatures, medical textbooks, and MeSH was used to designed an ontology-based query expansion: First, synonyms of query words are extracted from a synonym database as additional query words. Then if a query word is matched in the ontological term, its precedent parent node terms, its sibling node terms, and its descendant child node terms are also extracted as additional query words. In addition, the synonym extraction procedure is carried out again to find synonyms of the extracted ontological terms. Finally, the original query is expanded to a vector that includes SARS related words and corresponding weightings.

$$\{(q_1,W_1),(q_2,W_2),...,(q_k,W_k),...(q_n,W_n)\} \tag{8}$$

Results

We defined two retrieval modes, ontology-based mode and keyword-based mode for evaluation of IR model with and without ontology-based query expansion. The values of precision for the three IR models are shown in Table 1. In both retrieval modes, TMM has the best performance than the other IR models. In general, the values of precision of ontology-based mode are always lower than the keyword-based mode.

Table 1 - Values of precision of three IR models

IR Model	VSM	HMM	TMM
Keyword-based	0.5047	0.4896	0.7346
Ontology-based	0.4798	0.3631	0.6544

The interpolated recall-precision curves of three IR models in the ontology-based mode are shown in Figure 1. Except for recall level 0, the precisions of TMM are much higher than those of VSM and HMM. The patterns of the recall-precision curves of these models are different: TMM curve is concave downward and VSM and HMM curves are concave upward. This means the acceleration of decrease in precision with increase in recall of TMM is negative and those of the other two models are positive. Usually, the models are operated in the range of low recall values; therefore, TMM is the best model we should choose for the reason that it's curve slope in that range is the lowest among the three models.

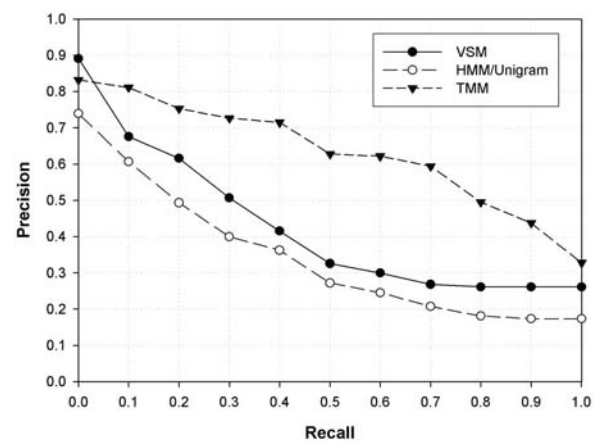

Figure 1 – Interpolated recall-precision curves of VSM, HMM/Unigram model, and TMM with ontology-based mode

Discussion

The value of precision of TMM in both keyword-based and ontology-based retrieval mode is much higher than other IR models. In Figure 1, the interpolated recall-precision curve of TMM is more stable than others. Thus the performance of the ontology-based query expansion is the best when it is applied to TMM. Relevant information is required to train TMM, however it is usually unavailable at beginning of building an IR system. An alternative way is that using VSM or HMM to develop the IR system and logging user's queries and reviewed documents. Then we can establish relevant information according to searching logs for training and updating model parameters. In addition, people usually review only top 10 or 20 records from searching results. For VSM and HMM, precision at recall level 0 is greater than 0.7 in the ontology-based mode. Therefore the performances of VSM and HMM are good enough when people only review preceding retrieval results.

Conclusion

We used ontological structure to design query expansion and applied it to three IR models. The evaluation results showed that the performance of TMM is better than other two IR models in both keyword-based and ontology-based retrieval mode. In practical use, we conclude that for users who are familiar with specific knowledge, using keyword-based method to obtain more precise information; otherwise, using ontology-based query expansion to generate queries and find relevant information.