THUIR at TREC 2004: Genomics Track*

Jiao Li, Xian Zhang, Min Zhang, Xiaoyan Zhu
State Key Lab of Intelligent Tech. & Sys., CST Dept, Tsinghua University, Beijing 100084, China
jiao-li04@mails.tsingha.edu.cn, vivian@tsinghua.org.cn, {z-m, zxy-dcs}@tsinghua.edu.cn

1. Introduction

This is the first time that THUIR participates in TREC Genomics Track. We took part in both Ad hoc retrieval task and Categorization task.

Based on our retrieval system TMiner, our research in the Ad hoc retrieval task focuses on: (1) Category of organism retrieval strategy; (2) Primary Feature Model; (3) Query Expansion (QE) technology; (4) Result fusion method.

Five official runs have been submitted at triage task in the Categorization task. Unigrams are used as features in Vector Space Model, and the high dimension feature vectors are trained and classified by SVM classifier with RBFs as the kernel function. Three ways are taken to improve the classifier: (1) Perform feature selection to reduce the dimension of feature vectors; (2) Weight the important features; (3) Balance between the positive dataset and the negative dataset.

2. Ad Hoc Retrieval Task

This task is a conventional searching task to retrieve Medline citations that contains the description given in the query of interest. Last year's genomics track data is used as training set.

A Medline record consists of many fields, and there are totally about 62 fields in the all corpus. However, some fields are not found in every record, and other fields may occur multiple times in one record. We call the title (<TI>) and abstract (<AB>) fields as text field, while others as non-text field. For the purpose of exacting enough information from the corpus, we analyze the non-text field. Finally, our retrieval fields include <TI> <AB> <MH> <RN> <TT><GS>.

Besides, the query of this task is a group of biologist's information need statements, including <TITLE>, <NEED> and <CONTEXT> field. As query quality is also an important factor to system performance, and our research on novelty task indicates that long query perform well in the retrieval, we use all the information in the need statements and weight the field of <TITLE>.

2.1. Primary Feature Model

As known, terms appear in the title, heading or emphasized fields in the text are more generally important for retrieval than the other body text. This is also adapted to the retrieval based on Medline record, as its format follow the field style like web page. We presented Primary Feature Model proposed before [1], which takes special information fields as Primary Feature Fields (PFF) and performs DF-related term weighting on them. In genomics track of this year, we use the model and take the fields of <TI> and <MH> as PFFs.

In the training set, namely TREC 2003 Primary task dataset, and the experimental results turn out that with a broad rate BM2500/PFS scale, we get improvement(7.53%) using PFS weighting scheme combined with BM2500 weighting.

2.2. Category of Organism Retrieval Strategy

In the <MH> field of Medline, there is information about organism, e.g. MH - Mice. Similarly, in

* Supported by by Chinese National Key Foundation Research & Development Plan (973) (Grant No.2004CB318108) and Chinese Natural Science Foundation (Grants No. 60223004, 60321002, 60303005).

the query, it contains the description of organism, e.g. Carcinogenesis and hairless mice. Therefore, the idea of category retrieval is inspired naturally. After categorizing the document and query according the organism, the retrieval relationship between them is shown as follow Figure 1.

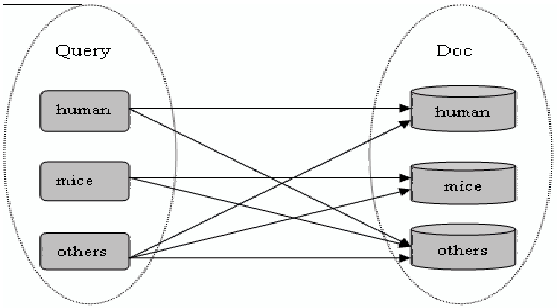

Figure 1. retrieval relation between query and documents

2.3. Query Expansion (QE) Technology

The technology of QE includes two parts: one is using a public online dictionary of abbreviations from MEDLINE (http://abbreviation.stanford.edu/); and the other is Pseudo relevance feedback technology which feedback top M terms in top ranked N documents to expand the query. Where, the first approach only performs on the <TITLE> field of each query, and the second one which had been already described in our last year's novelty track report performs all the long queries. The experimental results indicate that QE technology works well on the performance of recall while little improvement on mean average precision (MAP).

2.4. Result Fusion Method

AS Primary Feature Model works well on MAP with QE method on recall, the result fusion idea is generated. We work on the similarity score lists generated by PF model and QE technology. There are some basic fusion methods, such as Fox [2], linear merge [3], inverted ranking [4], as well as filtering and enhancing [4]. Our research focuses on the last two methods to get our fusion algorithm, which is shown as below:

For each $D_i \in L_1$ {
    if $D_i \in L_2$ then $s_i = \lambda / r_{i2}$
    else $s_i = 1/r_{i1}$
}

Where r_{i1} and r_{i2} are the ranks of document D_i in base list L_1 and enhance list L_2 respectively. The parameter λ is called enhance factor, which could control the effect of the enhance list to document rank. We calculate the document score with document rank instead of document similarity, as the rank of the relevant document in the retrieval document list is more significant for a retriever.

In this year's Genomics task, we use two kinds of fusion strategies: (1) PF score list is used as base list, and QE score list is used as enhance list; (2) QE score list is used as base list, and PS score list is used as enhance list, and the enhance factor λ is set to 0.2 in both cases.

2.5. Submitted Official Runs

Two official runs results and one unofficial one are listed in the follow table 1:

Runs		Run Type	P @ 10	P @ 100	MAP
1	THUIRgen01	manual	0.5820	0.3924	0.3435
2	THUIRgen02	automatic	0.5940	0.3944	0.3434
3	Unofficial run	automatic	0.5920	0.3864	0.3520

1. Select both of the text fields and non-text fields, long query with <TITLE> field weighting, divide

the documents and queries into three subsets according to the categorization rule, use BM2500+PFS weighting, use relevance feedback technology and expand the queries based on an online dictionary manually, use the first result fusion strategy mentioned above.

2. There are two points of differences from the last run: all the processes in this run are done automatically; On the other hand, this run uses the second result fusion strategy.
3. Select both of the text fields and non-text fields, long query with <TITLE> field weighting, use BM2500 only.

3. Categorization Task

In the triage subtask of the categorization task, unigrams are used as features in Vector Space Model, and the high dimension feature vectors are trained and classified by SVM classifier with RBFs as the kernel function. Three ways are taken to improve the classifier:

(1) Perform feature selection to reduce the dimension of feature vectors based on Term Frequency (TF), Document Frequency (DF) and Chi^2.
(2) Give larger weight to the features that have more information based on the biological knowledge.
(3) Balance between the positive dataset and the negative dataset by the help of K-means clustering.

3.1. Feature Selection

Feature selection is an important in Text Categorization because the original feature space consists of all the unigrams that occur in all the documents which can be hundreds of thousands of terms. So we use TF, DF and Chi^2 to select more informative terms. Chi^2 is defined to be [5]:

$$Chi(t_k, c_i) = \frac{n\left[P(t_k, c_i) \times P(\overline{t_k}, \overline{c_i}) - P(t_k, \overline{c_i}) \times P(\overline{t_k}, c_i)\right]^2}{P_d(t_k) \times P_d(c_i) \times P_d(\overline{t_k}) \times P_d(\overline{c_i})}$$

Where t_k is term k, c_i is category i, n is the number of all the documents.

TF can be used to roughly remove the stopwords which occur too many times in documents. Chi^2 is a normalized statistical value which could tell how a term distributes over category i and the rest categories. But it is not reliable for low-frequency terms, so TF and DF are used to filter the low frequency items.

The number of all the unigrams that occur in the triage task is about 240900. After the feature selection by the help of TF, DF and Chi^2, it is reduced to 1000~3000, which is a reasonable for a normal classifier.

3.2. Biological Feature Weighting

The triage task is not just a traditional Text Categorization work. They are about proteins and genes of mouse. Therefore, it is natural to turn to existing bioinformatical knowledge for help. In the feature selection stage, the words which occur in our existing Gene Ontology library or MGI library are selected by a lower threshold. Then 200-500 more terms, which are obviously more informative, are selected as the features of the final vector space.

3.3. K-means Clustering

In training set, the number of the positive documents is 375, while the number of the negative documents is 5462. Obviously it is so unbalanced that it is hard to build a good classifier. There are two reasons: First, we must give appropriate weights to positive and negative set in training, which could be easy to overfit the training data. Second, the negative data points distribute in almost all the vector space, so it is hard to find common features in the negative data.

By clustering the large negative dataset, we can avoid these disadvantages to some extent. The original negative dataset is divided into several subsets. In this way, the number of documents in each

subset is comparable to the size of positive set. The distribution of data points in each subset will be tighter than in the original negative set. This solves the second shortcoming above.

We finally divided the negative data into 7 subsets, between which and the positive dataset a classifier is built. When predicting, the test vector will be input into all the 7 classifiers and then a vote will be taken among the 7 results to get the final output. The evaluation results of our five submitted official runs are listed in the next section.

3.4. Submitted Official Runs

Five official runs results are listed in the follow table:

Runs		Precision	Recall	F-score	Normalized Utility
1	THIRcat01	0.1021	0.6024	0.1746	0.3375
2	THIRcat02	0.1033	0.5571	0.1743	0.3154
3	THIRcat03	0.0914	0.5500	0.1567	0.2765
4	THIRcat04	0.0908	0.7881	0.1628	0.3935
5	THIRcat05	0.1082	0.4167	0.1718	0.2450

4. Future Work

In the Ad hoc retrieval task, we take use of our TMiner system and some approaches ever used in other TREC tasks. Moreover, some parameter settings mainly accord to the experimental results of the primary task in TREC 2003 Genomics Track, but either the query expression or judgment method is different from the last year. Therefore, mass work on result analysis and method introduction should begin. Otherwise, there are a great many open sources about BioNLP on the website, which might help our information retrieval, except the online dictionary mentioned in section 4.2.3. Besides, we will also concentrate on the fast growth of biological texts, which ask for efficient retrieval sooner or later, though the measurements of time cost haven't been involved in the last two years.

In the Categorization task, we complete the first subtask and submit five official runs because of the time limited. In this year's task, our research focuses on using machine learning methods to solve the problems of triage task. We will continue our research on the other two annotation subtasks in the categorization task, while analyzing the existing results and methods of the first subtask. Otherwise, we will attempt to find new features to categorize the biological documents, as feature selection is so important in Text Categorization.

Reference

[1] Min Zhang, Ruihua Song, and Shaoping Ma, DF or IDF? On the Use of HTML Primary Feature Fields for Web IR, the 12th World Wide Web conference, (www2003), poster, Hungarian, 2003.

[2] E. A. Fox and Joseph A. Shaw. Combination of multiple searches. In Proceedings of the 2nd Text REtrieval Conference (TREC2), 1993, pages 243-252.

[3] Thompson, P. A combination of expert opinion approach to probabilistic information retrieval. Part I: The concept model. Information Processing and Management, 2693) 1990, pp371-382.

[4] Min Zhang, Study on Web Text Information Retrieval [D], Dissertation Submitted to Tsinghua University in partial fulfillment of the requirement for the degree of Doctor of Engineering, 2003.6.

[5] Yiming Yang, Xin Liu: A Re-Examination of Text Categorization Methods. SIGIR 1999: 42-49