# DUTIR at TREC 2007 Enterprise Track

Jianmei Chen, Hui Ren, Linhong Xu,Hongfei Lin, Zhihao Yang
Department of Computer Science and Technology, Dalian University of Technology
No 2 LingGong Road Shahekou District, Dalian 116023, China.
cindynjust@126.com, renhuei18@163.com, qingniao1203@163.com,
{hflin, yangzh}@dlut.edu.cn

## Abstract

This paper describes our experiments on the two tasks of the TREC 2007 Enterprise track. In data preprocessing stage we stripped the non-letter character from documents and query. For the Document Search, we built the index by indri and lemur, handled the query topic and then retrieved relevant documents by indri and lemur. For the Expert Search, we recognized candidates from collection, established correlative document pool, built the index by indri and lemur, and then got expert list and supporting documents.

## 1. Introduction

We participated in both the Document Search task and Expert Search Task at the Enterprise Track of Trec 2007. The Document search task is to search for messages regarding to a topic. Retrieved documents should be those that help the science communicator create an overview page in the given topic area. The expert search is to look for a person or multiple people who are experts on a subject and supporting documents which can explain why the person is an expert in a subject.

## 2. Data Preprocess

Different from the past, the interest of Enterprise Track this year is CSIRO (Commonwealth Scientific and Industrial Research Organization). The document collection is a crawl of the publicly available web pages from the "*.csiro.au" domain, known as the CSIRO Enterprise Research Collection (CERC or .CSIRO). The collection contains a mixture of document types including web pages, news/email archives, directory services, and document archives. Each type has particular characteristics, including inter and intra document structure.

The documents provided by TREC are full-text articles in HTML format. To improve performance, we cleaned CSIRO collection by removing most HTML tags except <AUTHER>, as we would use the <AUTHER> tag in correlative document pool generation. Furthermore, when removing the HTML tags, we reserved the URLs in them, since we thought the URLs implied the importance of a document and maybe useful for ranking [1]. Some other sections which we thought were noises such as all texts within the HTML tags "< DOCHDR >" and "</ DOCHDR >" were also deleted.

Based on this collection, we cleaned the collection further including removing the special character, such as "–", "/," etc [2]. This could be superior to matching the "if-else" and the "if else". Moreover, considering the encoding of the text, we parsed documents in ISO-8859-1, which could promise some non-English, such as "é", to be identified correctly.

# 3. Document Search Task

This task is to search some pages which contain a discussion about the topic, and messages could have pro or con point about the given topic.

## 3.1 Overview

Firstly, we preprocessed the cleaned CSIRO Enterprise Research Collection corpus, based on which an index was built by indri and lemur [3]. Then we handled the query topic in the similar way of cleaning the documents, i.e. stripping the special character and stopping word. At last relevant documents were retrieved by indri and lemur. Figure 1 depicts the overview of our retrieval system.

## 3.2 Query Analysis

A discussion search topic contains a query, a narrative, and a page. Our system makes use of the query and narrative fields of the topic to generate one or more abstract query representations from which specific search engine queries can be generated. Aside from standard stopword removal, we performed appending the field of narrative for query. Through experiments, we found the following facts. Firstly, stripping the stop-words from the field of query directly was superior to composing the query by bigram method. Secondly, appending the field of narrative for query helps to improve the precise of retrieval results.

## 3.3 Document retrieval

As discussed previously, our document retrieval component uses one or more query generator and search engine pairs. In our TREC system, we used indri and lemur to build the index. We developed query generation components for each of these search engines in an attempt to generate search queries that best leverage the expressiveness of the query languages of the underlying search engine. Finally, when we ranked the documents, we found that BM25 was superior to other ranking methods.

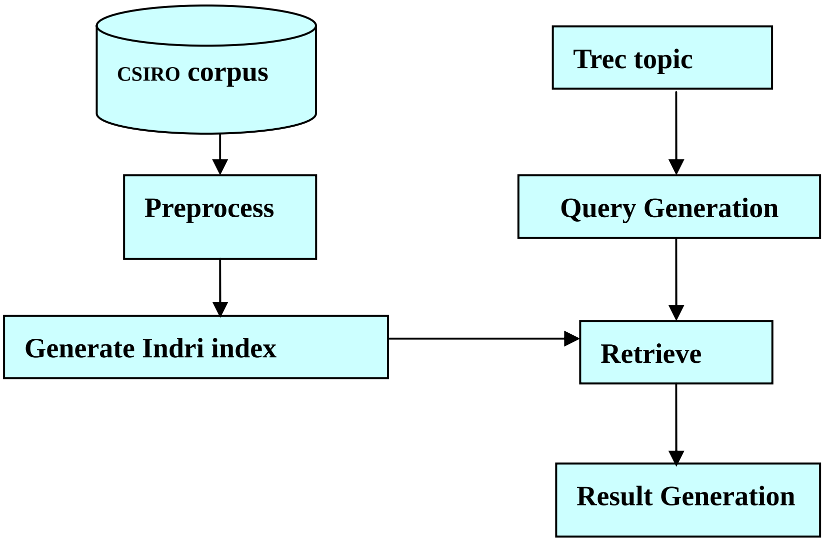

Figure 1: Framework of DS track IR system.

## 3.4 Summaries of runs and results

The following runs were submitted and the detail info of runs was displayed in Table 1

| Run identifier | Index type | query | Ranking method | remark |
|---|---|---|---|---|
| DUTDS 1 | Indri | Query, Narrative | BM25 | Manual |
| DUTDS 2 | Indri | Query, Narrative | BM25 | Auto |
| DUTDS 3 | Indri | Query, Narrative, Page | Indri | Auto |
| DUTDS 4 | Indri | Query | Indri | Auto |

**Table 1: the description of four runs**

The results of four runs were displayed in Table 2.

| Run ID | MAP | R-prec | Bpref | Reciprocal rank | p@10 |
|---|---|---|---|---|---|
| DUTDS 1 | 0.4015 | 0.4167 | 0.4062 | 0.8202 | 0.6100 |
| DUTDS 2 | 0.3316 | 0.3482 | 0.3469 | 0.7432 | 0.5100 |
| DUTDS 3 | 0.3487 | 0.3596 | 0.3903 | 0.7125 | 0.5380 |
| DUTDS 4 | 0.3364 | 0.3711 | 0.3751 | 0.4990 | 0.5120 |

**Table 2: Results for Discussion Search**

In Table 2, the first column displays the run identifier, the second reports the mean average precision (MAP), other columns display other important factors. In terms of the MAP measure, DUTDS 2 whose query text was taken from query field, narrative field is lowest. DUTDS1 which used Query and Narrative fields by manual and ranked by BM25 increased the MAP about 6.5% over DUTDS4 which used Query fields by manual and ranked by indri.

# 4. Expert Search

Expert search is one of the two tasks of 2007 Enterprise Track. In this task participants should retrieve a list of candidate experts on a subject.

## 4.1 Name recognition

We recognize candidate through three phases. Firstly, emails which include “csiro” are collected from collection. And then all names in collection are tagged by named entity recognition system which is finished by the Board of Trustees of The Leland Stanford Junior University. This package provides a high-performance machine learning, including facilities to train models from supervised training data and pre-trained models for English, which includes serialized model is a 3 class NER tagger that can label: PERSON, ORGANIZATION, and LOCATION entities. And we use the label of PERSON. Finally, some rules are made to match PERSON names tagged and emails as follow:

If the name of the label PERSON contains full name and matches the certain email, the person is considered the candidate in CSIRO.

If the name of the label PERSON contains first name or last name and matches the certain email, full name is searched in this document. And then if full name is found and could match the email, so the person is considered the expert in CSIRO. Moreover, we replaced all names and emails in the handled corpus into candidate identities. The number of the candidates is 3460.

## 4.2 Correlative document pool generation

Though the collection was preprocessed, we thought not all the texts contributed to expert finding. If a candidate id appeared in < AUTHER > field, we got the whole document as the correlative document pool. Otherwise we singled out one hundred words around the candidate identity to form the correlative document pool. We built two kinds of correlative document pool. One was the compositive relative document pool; the other was the dispersive relative document pool.

In the compositive relative document pool, every candidate had a profile. Since we had 3460 candidates, the documents in the collection were divided into 3460 relative document pools.

Upon request, every retrieved expert should be provided with corresponding supporting documents which can explain why the candidate is an expert in this subject. Accordingly, we dealt with the compositive correlative document pool. We took the "candidate ID- document ID" as the supporting document ID, in this way the compositive correlative document pool of a candidate was divided into some dispersive relative document pools [4].

## 4.3 Expert list and supporting document generation

The expert search task requires a list of support documents provided for each expert. There are two ways to achieve this purpose. The first way is finding the experts firstly, and then finding the corresponding supporting documents. The second way is that the supporting documents are found before getting the experts themselves, which is one of the natural ways for expert search. We used the first way in DUTEXP1, DUTEXP2, and DUTEXP3, and the second way in DUTEXP4. The detail information of our four submitted runs is displayed in Table 3.

| Run ID | Task | Correlative document pool | Index type | Query | Ranking method |
|---|---|---|---|---|---|
| **DUTEXP1** | **Expert** | compositive | Indri | query | BM25 |
| | **Support Document** | dispersive | | candidate identities + query | Indri |
| **DUTEXP2** | **Expert** | compositive | | query+ narr | BM25 |
| | **Support Document** | dispersive | | candidate identities + query + narr | Indri |
| **DUTEXP3** | **Expert** | compositive | | query + narr + candidate identities in page | BM25 |
| | **Support Document** | dispersive | | candidate identities + query + narr + candidate identities in page | Indri |
| **DUTEXP4** | **Support** | dispersive | | query + narr | Indri |

| | | | | | |
|---|---|---|---|---|---|
| | **Document** | | | | |
| | **Expert** | Calculate the average score of each candidate’s dispersive relative document pools. Rank the average scores and get the expert list. | | | |

**Table 3: Detail information of four runs**

In DUTEXP1, DUTEXP2, and DUTEXP3, an index was built based on the correlative pool firstly. We attempted to compose the query in several ways for each topic. The expert list was gained through the retrieved BM25 [5] score. Then we added the candidate identities to the original query and utilized indri to gain the supporting documents of the expert. In need of special note are: in DUTEXP3 the query included candidate identities in page which means the page within the tags <page> and </page> given in topic.

In DUTEXP4, an index was built by Indri based on the dispersive relative document pool. We use <query> and <narr> fields in topics to form queries and got support documents by indir. Then calculated the average score of each candidate’s support documents, ranked the average scores and got the expert list.

## 4.4 Results

| Run ID | MAP | R-prec | MRR | p@1 | p@5 | p@10 |
|---|---|---|---|---|---|---|
| DUTEXP1 | 0.2630 | 0.2252 | 0.4288 | 0.3600 | 0.5000 | 0.5800 |
| DUTEXP2 | 0.3324 | 0.3334 | 0.5362 | 0.4600 | 0.6200 | 0.6400 |
| DUTEXP3 | 0.3404 | 0.3232 | 0.5348 | 0.4600 | 0.6000 | 0.6800 |
| DUTEXP4 | 0.1876 | 0.1720 | 0.2929 | 0.2200 | 0.3600 | 0.4600 |

**Table 4: Results for Expert Search**

Table 4 shows the results for Expert Search. DUTEX4 is unsatisfactory, which is because we calculated the average score of each candidate’s support documents, ranked the average scores and got the expert list. This method is not delicate. In terms of the MAP measure, DUTEX3 is better than DUTEX2. The only difference between them is the query. We can see that it is better when query included candidate identities in the <page> field given in topic. DUTEX2 and DUTEX3 gains the preferable results since we modified its queries by manual. So we can conclude that it is effective to apply manual interfere in the process.

# 5. Conclusions

In this paper we describe our experiments on the TREC 2007 Enterprise Track. We found that structured information, such as thread structure, was not useful in the discussion search, and data preprocess that special characters were removed from W3C collection increased the MAP by about 3%. But in the mass our performance on training queries was not consistent with test queries, so different topics influenced the results to a certain extent. Expert search task is different from the traditional search problem. To resolve this problem, a new method which we called it correlative document pool, was applied and the result indicates the effectiveness of this methodology. There are so many pronouns in the document and it is very important to identify the expert in our method. Therefore, we will try to apply the technology involved in anaphora resolution in the future.

# References

[1] Jianhan Zhu, Dawei Song, Stefan Ruger, Marc Eisenstadt, Enrico Motta. The Open University at TREC 2006 Enterprise Track Expert Search Task. Proceedings of the 15th Text Retrieval Conference, 2006.
[2] Zhihao Yang, Hongfei Lin, Yanpeng Li, Linhong Xu, Yu Pan and Baoyan Liu. DUTIR at TREC 2006: Genomics and Enterprise Tracks. Proceedings of the 15th Text Retrieval Conference, 2006.
[3] http://www.lemurproject.org
[4] Conglei Yao, Bo Peng, Jing He, Zhifeng Yang. CNDS Expert Finding System for TREC2005. Proceedings of the 14th Text Retrieval Conference, 2005.
[5] J. Allan, J. Callan, K. Collins-Thompson, B. Croft, F. Feng, D. Fisher, J. Lafferty, L. Larkey, T. N. Truong, P. Ogilvie, L. Si, T. Strohman, H. Turtle, and C. Zhai. The lemur toolkit for language modeling and information retrieval. http://www.cs.cmu.edu/~lemur/, 2003.