PRIS at TREC 2011 Legal Track

Discovery Based on Relevant Feedback

Jiayue Zhang, Wenyi Yang, Xi Wang, Lihua Wu, Yongtian Zhang,
Weiran Xu, Guang Chen, Jun Guo
School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications
jyz0706@gmail.com, yeeyang19@gmail.com

Abstract

In order to finish the task of TREC 2011 Legal Track, this paper puts forward an experiment method, which combines indri and relevant feedback to evaluate the probability of relevance of every document in a collection.

1 Introduction

The goal of the Legal Track at the Text Retrieval Conference (TREC) is to assess the ability of information retrieval techniques to meet the needs of the legal profession for tools and methods capable of helping with the retrieval of electronic business records, principally for use as evidence in civil litigation. In the USA, this problem is referred to as "e-discovery." The 2011 Task uses exactly the same dataset as the 2010 Learning Task, which is the set of documents that results from message-level du-duping of the Enron emails. The collection is consisted of 685,592 e-mail messages and attachments (approximately 4GB of text) from 159 mailbox directories. The Learning Task of the Legal Track investigated the effectiveness of e-Discovery search techniques at learning from examples to estimate the likelihood of responsiveness as a probability of relevance of every document in a collection. In addition, the 2011 Task uses three new requests for production, which is named Topic 401, Topic 402 and Topic 403, respectively. Each topic included a one-sentence request for documents to produce for each topic.

2 Indexing

In the Learning Task, we are provided thousands of documents for each topic. To index the collection, we processed the files as follows:

We converted the message text in the “edrmv2txt-v2” file to an XML form. For each message (including attachments), we added a “<DOC>” tag before the message,

followed by the document id inside “<DOCNO>…</DOCNO>” tags, and a closing “</DOC>” tag at the end of the message. In order to distinguish the different types of the document, we used the “<TYPE>…</TYPE>” to sigsssssn the message as a mail or an attachment. If we got an attachment from “edrmv2nativeattach”, we used the content between “<CATEGORY>…</CATEGORY>” to show the attribute of this attachment, for instance, xls and doc,etc. Between the tags “<LENGTH>” and “</LENGTH>”, the bytes of the message were demonstrated, and the details of each document were followed by the tag “<TEXT>”, ending with </TEXT>. In particular, the “Date:”, “From:”, “To:” and “Subject:” lines were just treated as plain text like any line of the body of the email. The reason for converting the collection to this XML format was that we could then index it with building the indri.

3 Query

We used the method of relevant feedback to query experiment results with the combination of indri theory, which could reflect the probability of relevance of every document quite well.

First of all, we constructed query conditions according to the content of three requests of the production. We extracted key words showed as follows:

Topic 401: <text>#scoreif(enrononline #combine(#1(on line service) #combine(design development operation marketing) #combine(purchase sale trading exchange) #1(financial instruments) #1(financial products) #combine(#1 (derivative instruments) commodities futures swaps #1(common stock) equity dollars)))</text>

Topic 402: <text>#scoreif(#band({rule regulation law standard proscription act} {Enron Enrononline} {testimony evidence witness exhibit}) #combine(legal illegal permitted prohibited act testimony evidence exhibit witness proof commission {purchase sale trading invest exchange transaction})) </text>

Topic 403: <text>#scoreif(environment #combine(#1(environmental impact) #combine(#5(take measure) governing) #combine (emmissions spills pollution noise #3(animal habitats) leaking discharge) Enron #band(environmental {standards rule regulation law proscription}) #combine(#1(conform to) #1(comply with) avoid circumvent influence)))</text>

We chose top 100 in the experiment results as the most relevant ones to submit, and then we got responsiveness determinations from the TA.

Secondly, in order to finish the experiment more effective, we design the query sentences with the theory of relevant feedback. We applied VSM in inquires and reconstruction of documents based on the former determinations from the TA. We thought there were similarities between VSM of the different labeled relevant documents, and the diversities of the VSM existing in the labeled irrelevant ones. Therefore, the kernel idea of the method was the inquire values of the reconstruction should be closed to the VSM of the relevant documents as much as possible. Supposed the collection of the relevant documents as C_r, in this situation, the best

inquire vector value to distinguish the relevant and irrelevant documents expressed as:

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\forall \vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\forall \vec{d}_j \notin C_r} \vec{d}_j$$

Finally, we got the $\vec{q}_{opt}$ values of the words form each of the documents which could be positive or negative. We chose the word with the biggest positive value as the query key words, and deny the query results relating to the words with the smallest negative value. With this method, we submitted our experiment results with the web interface.

References

[1] Jun Guo. Web Search. Higher Education Press, Beijing, 2009

[2] Stephen Tomlinson. Learning Task Experiments in the TREC 2010 Legal Track. Proceedings of TREC2010.

[3] Text Retrieval Conference (TREC) Legal Track Home Page. http://trec-legal.umiacs.umd.edu/