Evaluating Document-to-document Relevance based on Document Language Model: Modeling, Implementation and Performance Evaluation[*]

Ge Yu, Xiaoguang Li, Yubin Bao, Daling Wang

School of Information Science and Engineering, Northeastern University
Shenyang 110004, P.R.China
{xgli7312@mail.163.com, yuge@mail.neu.edu.cn}

Abstract To evaluate document-to-document relevance is very important to many advanced applications such as IR, text mining and natural language processing. Since it is very hard to define document relevance in a mathematic way on account of users' uncertainty, the concept of topical relevance is widely accepted by most of research fields. It suggests that a document relevance model should explain whether the document representation describes its topical contents and the matching method reveals the topical differences among the documents. However, the current document-to-document relevance models, such as vector space model, string distance, don't put explicitly emphasis on the perspective of topical relevance. This paper exploits a document language model to represent the document topical content and explains why it can reveal the document topics and then establishes two distributional similarity measure based on the document language model to evaluate document-to-document relevance. The experiment on the TREC testing collection is made to compare it with the vector space model, and the results show that the Kullback-Leibler divergence measure with Jelinek-Mercer smoothing outperforms the vector space model significantly.

1 Introduction

The evaluation of document relevance is a very important task for many advanced applications such as information retrieval (IR), text mining, natural language processing, and has been extensively studied until now. In general, document relevance can be divided into two categories: query-to-document relevance and document-to-document relevance [1]. The former focuses on document relevance to a user's query, while the latter aims at an entire document in contrast to a small number of words in a specific user query. On account of the inherent redundancies and ambiguities in textual descriptions and high dimension problem, document-to-document relevance is more complex.

[*] Supported by the National Natural Science Foundation of China under Grant No.60173051 and the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institution of the Ministry of Education, China