Document Re-ordering Based on Key Terms in Top Retrieved Documents

Yang Lingpeng, Ji Donghong, Nie Yu, Zhou Guodong

Institute for Infocomm Research
21, Heng Mui Keng Terrace
Singapore 119613
{lpyang, dhji, ynie, zhougd}@i2r.a-star.edu.sg

Abstract. In this paper, we propose a method to improve the precision of top retrieved documents by re-ordering the retrieved documents in the initial retrieval. To re-order the documents, we first automatically extract key terms from top N (N<=30) retrieved documents, then we collect key terms that occur in query and their document frequencies in top N retrieved documents, finally we use these collected terms to re-order the initially retrieved documents. Each collected term is assigned a weight by its length and its document frequency in top N retrieved documents. Each document is re-ranked by the sum of weights of collected terms it contains. In our experiments on 42 query topics in NTCIR3 Cross Lingual Information Retrieval (CLIR) dataset, an average 17.8%-27.5% improvement can be made for top 10 documents and an average 6.6%-12% improvement can be made for top 100 documents at relax/rigid relevance judgment and different parameter setting.

1 Introduction

For Chinese Information Retrieval where query is a short description by natural language, many retrieval models, indexing strategies, query expansion strategies and document re-ordering methods have been proposed. Chinese Character, bi-gram, n-gram (n>2) and word are the most widely used indexing units. The effectiveness of single Chinese Characters as indexing units has been reported in [7]. The comparison between the three kinds of indexing units (single Characters, bi-grams and short-words) is given in [5]. It shows that single character indexing is good but not sufficiently competitive, while bi-gram indexing works surprisingly well and it’s as good as short-word indexing in precision. [9] suggests that word indexing and bi-gram indexing can achieve comparable performance but if we consider the time and space factors, it is preferable to use words (and characters) as indexes. It also suggests that a combination of the longest-matching algorithm with single characters is a good method for Chinese IR and if there is a module for unknown word detection, the performance can be further improved. Some other researches give similar conclusions. Bi-gram and word are considered as the top two indexing units in Chinese IR and they are also used in many reported Chinese IR systems.