A Model for Extracting Keywords of Document Using Term Frequency and Distribution

Jae-Woo Lee and Doo-Kwon Baik

Software System Lab., Dept. of Computer Science & Engineering,
Korea University,
1, 5-ka, Anam-dong, SungBuk-ku, 136-701, Seoul, Korea
It21c@Korea.ac.kr,Baik@Swsys2.korea.ac.kr

Abstract. In information retrieval systems, it is very important that indexing is defined very well by appropriate terms about documents. In this paper, we propose a simple retrieval model based on terms distribution characteristics besides term frequency in documents. We define the keywords distribution characteristics using a statistics, standard deviation. We can extract document keywords that term frequency is great and standard deviation is great. And if term frequency is great and standard deviation is small, the terms can be defined as paragraph keywords. Applying our proposed retrieval model we can search many documents or knowledge using the document keywords and paragraph keywords.

1. Introduction

Information retrieval is one of the most important technologies at present. We can always get many information in the Internet or distributed computing systems using various information retrieval models. For searching proper information that we need, it is necessary to extract keywords of documents helping many web clients' requests. These information retrieval models specify how representations of text documents and information needs should be compared in order to estimate the likelihood that a document will be judged relevant. The estimates of the relevance of documents to a given query are the basis for the document rankings that are now a familiar part of information retrieval systems. Many models, including the probabilistic or Bayes classifier, have been proposed and are being used [1,2,3].

In information retrieval systems, it is very important that indexing is defined very well by appropriate terms about documents. In this paper, we propose a simple retrieval model based on terms distribution characteristics besides term frequency in documents. We define the keywords distribution characteristics using a statistics, standard deviation. By the standard deviation we define meaningful terms as document keywords or paragraph keywords, and the terms are selected by using stemming, filtering stop-lists, synonym for search meaningful terms in a document including TF-IDF(Term Frequency - Inverse Document Frequency). And then we can search many documents or knowledge using the keywords [2,4,5,6,7].