Knowledge-Based Systems 24 (2011) 393–405

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Word AdHoc Network: Using Google Core Distance to extract the most relevant information

Ping-I Chen *, Shi-Jen Lin

Department of Information Management, National Central University, Chung-Li 320, Taiwan, ROC

ARTICLE INFO

Article history:
Received 4 March 2010
Received in revised form 23 November 2010
Accepted 23 November 2010
Available online 29 November 2010

Keywords:
Similarity distance
Search engines
Information retrieval
Keyword sequence
n-gram

ABSTRACT

In recent years, finding the most relevant documents or search results in a search engine has become an important issue. Most previous research has focused on expanding the keyword into a more meaningful sequence or using a higher concept to form the semantic search. All of those methods need predictive models, which are based on the training data or Web log of the users' browsing behaviors. In this way, they can only be used in a single knowledge domain, not only because of the complexity of the model construction but also because the keyword extraction methods are limited to certain areas. In this paper, we describe a new algorithm called "Word AdHoc Network" (WANET) and use it to extract the most important sequences of keywords to provide the most relevant search results to the user. Our method needs no pre-processing, and all the executions are real-time. Thus, we can use this system to extract any keyword sequence from various knowledge domains. Our experiments show that the extracted sequence of the documents can achieve high accuracy and can find the most relevant information in the top 1 search results, in most cases. This new system can increase users' effectiveness in finding useful information for the articles or research papers they are reading or writing.

1. Introduction

The search engine has become an indispensable tool in people's daily lives. Individuals can search for any kind of knowledge and can find all of the newest information in the world. The only way to use the search engine is to enter some keywords that represent what the user wants to know. Jansen et al. [21] found that most users only enter 2.35 terms in the search engine, usually because the users lack sufficient domain knowledge about entering precise keywords that describe their thoughts.

In the past few years, the Google search engine has offered keyword expansion, a feature that can provide useful next or additional keywords to help the user find the most relevant and accurate search results. But the possible combinations of keywords are so numerous that the Google recommendations will not always work. The search engine can only provide the most frequently entered keyword sets, which are determined by other users. This method is called "collaborative recommendation". Both methods use tools such as semantic nets, ontology [48], and Markov chains [5] to model users' behavior and to identify their interests. Thus, the system can find people who share interests in order to form communities and to use their search results as the best potential keyword sequences.

However, each time users want to search for information, they may want to find different kinds of information in different knowledge domains. A user might enter the keyword "Apple" in order to learn about the Apple company's newest product. But in another search, the same user might use the same keyword to search for McDonald's Apple pie. Using the traditional methods, the training model only can be constructed in a single domain. When modeling several domains of knowledge at the same time, the model will be so huge that it will take an extremely long time to search for the potential keywords in all those domains. In addition, keyword extraction methods, like TF-IDF, always rely on term frequency (TF) to find the most important keywords. The TF-IDF methods need pre-collected documents or Webpage sets to calculate the inverse document frequency (IDF) values. But in Web information retrieval, all the execution should be on-line and real-time.

Additionally, users' browsing behaviors are widely varied. Therefore, it is impossible to determine the exact IDF values in order to evaluate the importance of the keywords except by observing users and collecting information about their search behaviors. To use this system on a mobile handset device, the repository of that information must be very small; hence, it is impossible to save the pre-defined model or dataset on such devices. Thus, there is a need for a new way to enhance the keyword expansion methods,

* Corresponding author.
E-mail addresses: chenpingi@gmail.com, 974403002@cc.ncu.edu.tw (P.-I. Chen), sjlin@mgt.ncu.edu.tw (S.-J. Lin).

0950-7051/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.11.006

turning it into a real-time execution system, and to minimize the system in order to save space in the repository.

In our previous work, we used the Google similarity distance algorithm to calculate and find the potential keywords in articles that users viewed [6]. The users can acquire information about each keyword so that all the articles or Web pages that the users are reading can become a Wikipedia-like system. In other words, this system can automatically provide information about which keywords in an article are most important to that user. This is easier than marking the keywords and searching in Google to discover their meanings. By using an NGD algorithm, the system can achieve the goal of on-line real-time execution without using any repository. However, this system only can provide information about each keyword. To find the most relevant information about the article, the keywords must be expanded into a longer sequential set. For example, if a user reads an article about the Google similarity distance, the user may want to know whether any research relates to that article or whether other papers reference it. A set of keywords must be extracted as the "term vector" to represent this article; that vector can then be used to search Google or a database to find the most relevant research articles [22].

The term vector model uses index terms as vectors of identifiers to represent documents. The Apache Lucene text search engine uses this method to calculate documents' relevancy rankings. The document similarities theory is then used to compare the deviations of angles between the documents' vectors and the query keywords. Many researches explain how to find a meaningful keyword sequence in order to enhance the accuracy of the search results. Still, the model must be trained in advance so that the system usability is always restricted.

In this paper, we adapt the NGD algorithm and try to extract a meaningful keyword sequence based on the Webpage or article that is viewed. The NGD algorithm has been used to conduct some systems. We proved that using the number of search results to calculate the relations among the keywords is a workable method. However, we find a significant problem; the number of keyword search results will be varied. Thus, the relations of keywords will become unstable because the NGD algorithm can only be used for search results. The NGD algorithm should be used to extract the sequence of documents as index terms and to use this sequence to cluster those relative documents. When using this method, two similar data that were collected at different times are likely to be totally different because the relations of the keywords have been changed so greatly.

We propose a new algorithm, called "Google Core Distance" (GCD), to improve the stability of the relations. In this process, the GCD algorithm will become a distance-based method rather than a probability method. Therefore, it will be impossible to combine this method with the traditional LSI algorithm to find a *n-gram* sequence of keywords as the term vector. We used the famous PageRank algorithm and combined it with the BB's graph-based clustering algorithm to find the sequence. This idea is based on the sensor network's routing algorithm, and we have named it "Word AdHoc Network" (WANET). Our method can be used to find a most important keyword, the most important two or three keywords, and so on. By using a Hop-by-Hop Routing algorithm, we can extract the word-by-word sequences from the documents or Web pages to represent the term vectors. Thus, to find the best keyword sequences, our system will focus on the co-occurrence of each two keywords, the connectivity of those relative keywords (including the relations of keyword to keyword and the relations of a keyword's relative keywords), and the best routing path.

We used the previously mentioned algorithms to conduct the system, and we used the Elsevier Webpage to choose four different knowledge domains. We randomly selected 10 journals in each domain as subjects for the experiment. In each journal, we chose the top 25 most-downloaded papers as our data set, and we used our system to extract the most important keyword sequences. Next, we used those sequences to search in Google and to evaluate the strength of the sequences' ability to find the original papers in the Top-k search results. The experiment's results showed that the 4-gram sequence of the determined keywords can identify the most relevant search results.

We believe that using our system can help users, especially researchers, to find the most relevant information in a more efficient way. When a user is writing an introduction to a research paper, if the system can immediately calculate and find related previous researches, then it will be much easier for the user to define the paper topic and to determine whether any previous work has examined the same problem [27]. When the manager of a company is reading a newspaper on a mobile device, if the system can find other related articles and provide a summary, then the manager can make informed decisions—anytime, anywhere. Information from around the world is collected on the Internet, but users must know how to find the precise information and then to use it properly. By using this kind of system, users can control the information and can understand current events, thereby enhancing their knowledge.

The rest of this article is organized as follows. In Section 2, we will introduce some relative research articles and compare them with the system which we proposed in this article. In Section 3, we will introduce our proposed methods in detail to emerge the original thinking of the system's design. In Section 4, we evaluate the NGD algorithm and our proposed GCD algorithm by using the spearman's footrule measurement. Then, we conduct the experiment which uses the research articles from different knowledge domain to evaluate whether our system can achieve the goal of either high accuracy or cross domain or not.

2. Related works

Most users do not know how to enter precise keywords to represent what they want to find in the search engine, especially when looking for knowledge in unfamiliar domains [1]. Two main problems need to be solved: (a) extracting the potential keyword; (b) finding a meaningful keyword sequence [7,13,20,34,40]. In the past few years, many different solutions to this problem have been developed. A sequence of keywords can be extracted to represent the documents, and searching for that sequence in the search engine will offer the most relevant information to the user. In this section, these methods will be explained and compared with our system in detail.

2.1. Semantic similarity

Semantic similarity is a concept that has been used to measure the similarity of documents or terms. Some algorithms use human-defined ontologies to measure the distance between words. Others use the vector-based model to represent their correlations.

2.1.1. Vector-based model

In 1990, Deerwester et al. proposed the latent semantic analysis (LSA), which is a technique used for natural language processing and for providing the relationship measurements of word–word and word–passage. This method analyzes the potential relationships among a set of documents and terms by using a set of concepts related to those documents and terms. Thus, a document becomes a column vector, and the query that is entered by the user also becomes a vector. Finally, the two vectors can be compared to measure their similarity. The problem with the LSA is that it does