International Journal of Innovative Computing, Information and Control
Volume **6**, Number **4**, April **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 1843–1864

A NOVEL TOPIC SELECTION ALGORITHM BASED ON WORD DISTRIBUTION

Chun-Wei Tsai[1], Ko-Wei Huang[2], Heng-Yao Hsu[2], Ming-Chao Chiang[1] and Chu-Sing Yang[3]

[1]Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung 80424, Taiwan
cwtsai87@gmail.com; mcchiang@cse.nsysu.edu.tw

[2]Department of Computer and Communication Engineering
[3]Department of Electrical Engineering
National Cheng Kung University
Tainan 70101, Taiwan
elone530@gmail.com; q3695122@mail.ncku.edu.tw; csyang@ee.ncku.edu.tw

Received October 2008; revised February 2009

Abstract. *Over the past decade, more and more users of the Internet rely on the search engines to help them find the information they need. However, the information they find depends, to a large extent, on the ranking mechanism of the search engines they use. Not surprisingly, it generally consists of a large amount of information that is completely irrelevant. To help users of the Internet find the information they are looking for quickly, an efficient algorithm for building the summaries of a collection of documents found by a search engine in response to a user query, called DiSco (Distribution Scoring), is proposed. The main idea in the design of DiSco is to balance the rate of coverage and overlap in the selection of topic words for document summarization, especially when the datasets are large. Moreover, several measure methods such as coverage, overlap, and the computation time are employed in evaluating the performance of the proposed algorithm. All our simulation results indicate that the proposed algorithm outperforms all the state-of-the-art algorithms evaluated in terms of not only the quality of the summarizations but also the computation time.*

Keywords: Document summarization, Search engine, Topic

1. **Introduction.** Finding out the information that users need from a large amount of data has always been a challenge of information retrieval. Search engine is certainly a useful tool for helping users of the Internet find the information they need quickly. Unfortunately, it, in general, consists of a great amount of information that is totally irrelevant, as is typical of search engines and even clustering search engines. In other words, with all the search engines and even clustering search engines are associated two major problems:

- *The returned information is not centralized:* This problem is due to the fact that the useful information returned by a search engine tends to spread over a large number of similar documents instead of being centralized in a single document, thus making it extremely difficult to identify and retrieve.
- *The classified information is overlapped:* This problem is due to the fact that the web pages returned by a search engine are similar to each other even after grouping.

In general, the first problem shows up because ranked lists used by a search engine are not summarized in terms of the topics and thus are not suitable for browsing task