*International Journal of Innovative Computing, Information and Control*
Volume **6**, Number **1**, January **2010**

ICIC International ©2010 ISSN 1349-4198
pp. **399–408**

# ENHANCING TOPIC TRACKING FOR CHINESE NEWS WEB PAGES WITH TEMPORAL INFORMATION AND KEY WEB CONTEXTS

Jing Qiu, Lejian Liao and Peng Li

Beijing Laboratory of Intelligent Information Technology
School of Computer Science
Beijing Institute of Technology
5 South Zhongguancun Street, Beijing, 100081, P. R. China
{ qiuhao; liaolj; wowolee }@bit.edu.cn

Received December 2008; revised April 2009

Abstract. *With the continuous growth in the number of news Web sites available and the diversity in their presentation of contents, there is an increasing need for mining the news correlation on the Web to keep track of successive development of specific event. In this paper a new approach to topic tracking of Chinese news Web pages is presented. We have developed several relatively simple NLP methods for extracting temporal information from news texts and key Web contexts from HTML, with the goal of improving performance of TDT tasks. Temporal information is used to increase the similarity between the news and topics which happened at the same time because of the fact that the same type of events happened on the same time. Because key Web contexts are used to get weighted model of TDT system, we have made these information elements play a more important role in the similarity match between the news and the topics. The technique is implemented in a framework of dependency structure language model (DSLM). The experiments show remarkable improvements to the existing approaches.*

**Keywords:** Topic tracking, Temporal information extraction, Key Web contexts, Dependency structure language model

1. **Introduction.** Topic detection and tracking (TDT) is a relatively new issue in the information retrieval (IR) field. The TDT2000 project embraces five key technical challenges, namely topic segmentation, topic tracking, topic detection, first story detection, and link detection. Tracking and link detection are considered to be the primary tasks, representing core technology that is broadly applicable to many different TDT applications [16].

The topic tracking can be understood as an information filtering task in which the system is given one or more sample documents and is expected to spot all further documents discussing the topic of the samples [3]. Language modeling techniques have been found well suited to topic tracking task [2,5,11,15]. Lee et al. [5] proposed that the dependency structure language model (DSLM) should be used to overcome the limitation of unigram and bigram models in TDT. Its structure is described in terms of mathematical models, which need to be joined with some simple semantic information as an extension to it.

One contribution of this paper is to add temporal information to DSLM. Most TDT systems have been built based on the observation that a particular event appears in news stories within a certain time period and that a similar news story generated after a while is likely to refer to a new event. Temporal information is an important feature of a topic, and work has been done to utilize temporal information to enhance the effectiveness of TDT systems [4,6]. Li et al. [6] have proposed a strategy with time granularity reasoning for utilizing temporal information in topic tracking. They compared the method with