A Practical System for Harvesting and Monitoring Hot Topics on the Web

Xiaojun Wan, Jianwu Yang
Institute of Computer Science and Technology & Key Laboratory of Computational Linguistics,
MOE, Peking University, Beijing 100871, China
{wanxiaojun, yangjianwu}@icst.pku.edu.cn

ABSTRACT

This poster briefly describes a practical system named *FounderWISE* for harvesting and monitoring hot topics on the Web. *FounderWISE* consists of five components: Web crawler, text classifier, topic detector, topic summarizer and topic analyzer. In this poster we present two key components of topic detector and topic analyzer. The system has been successfully deployed in a few Chinese major government departments.

Categories and Subject Descriptors:

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *miscellaneous*

General Terms: Algorithms, Design, Performance

Keywords: Topic detection, information diffusion, topic evolution, *FounderWISE*.

With the rapid increase of text documents on the Web on even a single day, it is becoming harder for users to obtain useful information by reading each single document. Specific users want to know and monitor hot (or popular) topics on the Web by simply browsing, rather than by reading every document in the large web collection. Actually, every day's hot topics are usually reported in various web sites, maybe differently in particular documents. Aggregating the documents in a large number of different web sites and identifying the hot topics will alleviate users' burden to reading particular documents. Furthermore, it will greatly help users understand the topics and obtain useful information by discovering topic-related knowledge.

Topic detection and tracking (TDT) [1] is one of such techniques to detect events or topics from a news text stream. However, almost all previous works on TDT focus on topic detection on a small benchmark TDT corpus. Because the corpus is very small as compared with the Web collection, and the large-scale Web collection will bring a few new challenges for existing TDT techniques, the techniques cannot be applied directly to detect topics on the Web. In the industrial field, *BaiduNews* (news.baidu.com) and *GoogleNews* (news.google.cn) are considered as two popular commercial Chinese news aggregation services. The techniques underlying the two systems are unknown due to commercial privacy. Whereas, the two systems simply group news articles into news topics and the features of the systems are limited.

In this poster, we present a practical system named *FounderWISE* to incorporate novel features for harvesting and monitoring hot topics on the Web. *FounderWISE* can provide not only a polished list of hot topics but also the topic-related knowledge to users, thus greatly facilitate users to monitor and understand Web topics. The system consists of five components: a Web crawler to download web pages and extract document metadata, a text classifier to categorize web pages into fine-grained classes, a topic detector to identify and rank topics within each class, a topic summarizer to produce updated summaries for topics, and a topic analyzer to discover and present topic-related knowledge. Figure 1 gives the system architecture, and this poster focuses on the topic detector and topic analyzer.

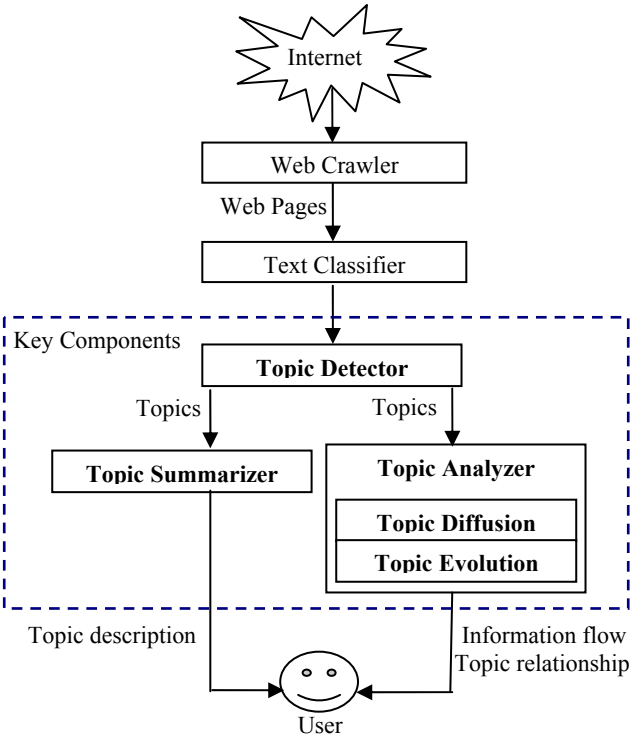

Figure 1: System architecture of *FounderWISE*

Topic Detector: The single-pass clustering algorithm (INCR) is the most widely used algorithm for online topic detection. It sequentially processes the input documents, one at a time, and grows clusters incrementally. However, the algorithm has a few problems when dealing with the huge amount of web data and providing practical services to users. Several improvements over the algorithm have been proposed in *FouderWISE* to better deal with practical data and meet users' requirements. First, a measure is proposed to evaluate the popularity of a topic and the topics can be ranked according to their popularity score. The measure is based on such features as document number, document timestamp and document visual weight. A number of unpopular topics are filtered out and only the hot topics are reserved, which can improve the detection efficiency as well. In addition, a novel step is added in the detection algorithm to merge the topics belonging to the same topic and reassign a small number of documents into

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

appropriate topics. This step can improve the detection results. Finally, a relevance feedback mechanism is designed to allow users to filter out the uninteresting topics and emphasize on the interesting topics.

Topic Analyzer: The topic analyzer adds two useful features to the system by learning information diffusion process within a single topic and computing evolution relationships between multiple topics. The results are visually presented to users.

By our analysis, documents are diffused or transmitted between web sites frequently. Some documents are directly copied or forwarded from one web site to another web site without any changes, and other documents are forwarded between web sites after minor revisions, e.g., addition or deletion of some texts, or rewriting of some sentences. Thereafter, many diffusion relationships exist between the documents within a topic, and the diffusion process for the topic is represented by all the diffusion relationships. We consider the diffusion identification as a binary classification problem between pairwise documents, and we use the SVMLight tool with meta-data features, cueword features and similarity features. The final F-score based on a small set of labeled document pairs reaches 80%. The details are presented in [2]. Figure 2 visually presents the diffusion process for an example news topic.

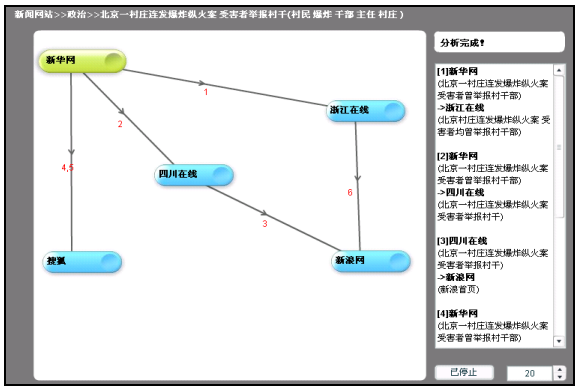

Figure 2: An example of diffusion process

FounderWISE stores hot topics for every single day in database and users can browse the historical topic list in a specific previous day. The hot topics usually evolve with time and the most hot topic may exist for more than one week or even one month. Our analysis shows that there exist two kinds of relationships between topics in different days: one is that two topics are reflecting the same event, i.e., the two topics are the same topic in essence; the other is that two topics are topically relevant but not the same, e.g., "Chinese president visits US" and "US president visits China". The "same" relationships can be identified in the topic detection algorithm because the same topics should have the same topic identifier in the topic detection algorithm. The "relevant" relationships can be identified by comparing the similarity between two topics in two adjacent days. If two topics have different identifiers and their cosine similarity value is larger than a preset threshold and, the topics are relevant with each other. The precision of the detected relationships can reach 95%, and the recall of the detected relationships can reach 86%.

Figure 3 gives an example of evolution graph for news topics detected between 2007-10-12 and 2007-10-18 for the politics category. Twenty hot topics are presented for each day and each topic is represented by a short horizontal bar. The two kinds of relationships between topics in adjacent days are denoted by two types of connection lines respectively: the horizontal connection line between topics denotes the "same" relationship between the topics, and the diagonal connection line between topics denotes the "relevant" relationship between the topics. A long horizontal line linking more than two topics means that the topic is very hot and has existed for a long time.

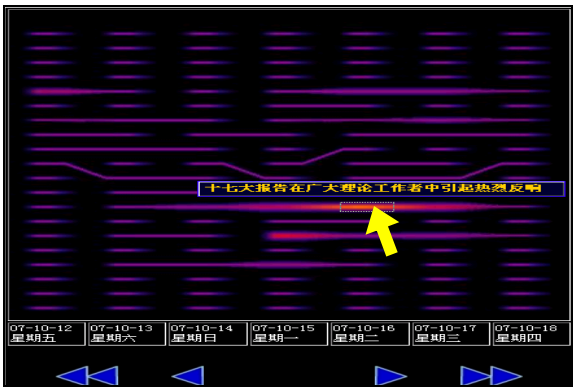

Figure 3: Sample topic evolution graph

Based on the above features, *FounderWISE* can effectively and efficiently identify hot topics and discover useful topic-related knowledge from more than 100,000 documents on a single PC server in everyday. *FounderWISE* has been successfully deployed in a few Chinese major government departments and a user study is performed and the results show several useful advantages of *FounderWISE* over *BaiduNews* and *GoogleNews*. Figure 4 gives the real user interface of *FounderWISE*.

Figure 4: User interface of *FounderWISE*

ACKNOWLEDGMENTS

We thank Bin Lu, Xiaojiang Huang, Dong Wang, Tao Feng and Rufeng Liang for system development. This work was supported by Beijing Nova Program (2008B03), NCET (NCET-08-0006), National High-tech R&D Program (2008AA01Z421) and National Development and Reform Commission High-tech Program of China (2008-2441).

REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron and Y. Yang. Topic detection and tracking pilot study: final report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] X. Wan and J. Yang. Learning information diffusion process on the web. In Proceedings of WWW 2007.