Chapter 1
Re-Evaluation of On-Line Hot Topic Discovery Model

Hui-min Ye  
University of Vermont, USA

Sushil K. Sharma  
Ball State University, USA

Huinan Xu  
Ernst & Young, USA

ABSTRACT
As a major medium for information transmission, Internet plays an important role in diffusing and spreading news on web. Some governments attach great importance and pay lot of effort trying to detect, track the development of events and forecast emergency on internet. On the basis of the researches in the field of topic detection and tracking, we proposed a model for hot topic discovery that would pick out hot topics by automatically detecting, clustering and weighting topics on the websites within a time period. We also introduced a topic index approach in following the growth of topics, which is useful to analyze and forecast the development of topics on web.

INTRODUCTION
The web has become indispensible part of modern life and has unbelievable influence in our real society. It spreads and provides sources of hundreds of millions of news and information, in which some topics that are growing in interest over time have great impact on our real life, sometimes even affect the development of events in a way. In order to pick out these influencing news and topics on the web, we built an intelligent system that can automatically and effectively discover hot topics embedded on the web within a period. Earlier we took bulletin board system (BBS) into consideration in our model in terms of discovering hot topic on internet. It’s true that some messages or topics on BBS are concerned with the mainstream news or topics, however, we discover later on that most messages on BBS are too trivial and irrelevant to serious issues, the meaningful hot topics selected from BBS are almost amount to nothing compared with large number of data collection, which means it contributes little in picking out hot
topics. Therefore in our modified topic model we ignore the contribution of messages from BBS.

The algorithm we proposed for hot topics discovery is based on the principle of Term Frequency * Proportional Document Frequency (TF*PDF). Research on TF*PDF (Term Frequency * Proportional Document Frequency) algorithm has been described in (Khoo and Ishizuka, 2001a, 2001b). The algorithm has been adapted in our article in a way that assigns heavy weight to those topics that discussed in many documents from many sources concurrently. Based on the principle of stock index, we use topic index to manifest the developing process of hot topics.

The hot topic discovery model takes a collection of data as input and identifies topic areas that are growing in importance with collected information (Allan et al. 2000, Fiscus & Doddington, 2002). It includes two stages, the first stage is topic detection and clustering stage, the second stage is hot topic discovery and generation of topic index. In the following sections, we first briefly depict the first stage in which we adopt existing algorithm to realize this task of topic detection and clustering. The second stage explains the modified model we propose to detect hot topics and the topic index method to follow the growth of topics. Then we discuss and compare the results of experiments based on two models. The flow of information in the system is illustrated in Figure 1.

The objective of topic detection is to identify topically related stories without positive or negative training stories. It is basically a problem related to clustering, where the goal is to group stories discussing the same event (Wayne, 2000). Detection is similar to tracking with the exception that no any training stories are provided for a particular topic, and all the topics that are mentioned in stream have to be identified (Allan et al., 2003, Allan, et al, 1999, Yang et al. 1999, Yang et al, 2000). Topic detection can be accomplished using any clustering algorithm as long as it is on-the-fly and nonoverlapping clustering (Wayne, 2000, Martin et al., 1997). In the process of detection we make use of tf*idf weighting scheme and the cosine similarity metric. The idf component of the weighting is based on incremental statistics to emulate the online nature of the task.

**Document Representation**

A document is represented with a $tf\cdot idf$ feature weights where

$$tf = \frac{t}{t + 0.5 + 1.5 \frac{dl}{dl_{avg}}}$$  \hspace{1cm} (1)

$$idf = \log\left(\frac{N + 0.5}{df}\right)$$ \hspace{1cm} (2)

---

**Figure 1. System information flow topic detection stage**

![Figure 1. System information flow topic detection stage](image)