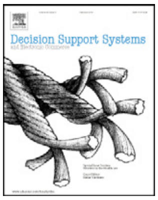
Decision Support Systems 88 (2016) 85–97

Contents lists available at ScienceDirect

Decision Support Systems

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



A novel trend surveillance system using the information from web search engines

Ze-Han Fang, Chien Chin Chen *

Department of Information Management, National Taiwan University, Taiwan

ARTICLE INFO

Article history:
Received 2 November 2015
Received in revised form 15 April 2016
Accepted 2 June 2016
Available online 11 June 2016

Keywords:
Trend surveillance
Learning to rank
Data mining
Feature selection

ABSTRACT

Web search engines are becoming a major platform for the general public to access information. It has been suggested that because the search patterns of search engine users are correlated with emerging events, the query log of search engines has the potential for trend surveillance, such as monitoring outbreaks of epidemics. Many trend surveillance studies have investigated the use of query logs and have strived to identify query terms suitable for trend surveillance. Most of these works select representative query terms by consulting domain experts or by preparing a large text corpus for feature selection. The process of these approaches, however, is too costly to make the trend surveillance methods adaptable to different topics. In this paper, we propose an adaptive trend surveillance method. We developed a simple and effective feature selection algorithm, called *TF-LTR*, which leverages the document returned by search engines and the frequency of the terms in the returned documents to select representative query terms of trending topics. Specifically, we investigated pair-wise learning to rank models in order to measure a term's discriminative power in making a document rank higher in the returned document list. The discriminative power is combined with the term frequency which denotes the on-topic degree of a term to measure a term's representativeness against a trending topic. Representative terms and their query frequencies are applied to a state-of-the-art data mining model to enhance the effectiveness of trend surveillance. The experimental results based on trending topics of different domains show that our trend surveillance method performs well and the ranking information of search engines are helpful for trend surveillance. In light of this, the proposed method can provide effective support for government officials and authorities in order to help them to respond to fast-changing events and topics, and to make appropriate decisions.

1. Introduction

A trending topic is a long-running event which is highly associated with people's life and activities, and has an index that depicts the topic's status (development). Because trending topics are usually associated with the concerns of individuals and authorities, trend surveillance systems that periodically predict the status of a trending topic are thus important for countries and organizations to help decision makers make appropriate decisions in response to fast-changing national and international situations. For instance, a health surveillance system that systematically collects health-related data from different areas of a country enables a government to monitor the health status of the public [1–4]. Disease outbreaks can be detected that help the government determine in a timely manner where, when, and how to allocate health resources in order to achieve the best epidemic control performance. Financial surveillance systems can assist organizations in understanding domestic and global financial trends which enable the establishment of appropriate business policies [5–8]. There is a great deal of evidence showing that trend surveillance is indispensable and that decision making processes can easily be misguided and problematic without such surveillance systems [9–12].

To construct a reliable surveillance system, representative indicators should be identified. Financial surveillance systems normally make use of business indices such as industrial production, stock price index, and manufacturing sales to measure the economic status of a country [7,13–15]; and health surveillance systems are always based on the infection number of a certain disease reported by medical institutions [16,17]. While the indicators are effective, in practice, their collection normally involves long data processes that delay the announcement of a trend's status [5–7]. The announced trend status thus lags, thereby possibly increasing the uncertainty of decision making. For instance, in the United States, officials generally take more than one month to compile economic indices, which seriously delay the announced economic status [6,7,18,19]. To remedy the problem, how to choose reliable and timely indicators is a practical and important research target.

Recently, due to the rapid development of the Internet, many studies (e.g., [1–5,20–22]) have utilized web search engines for trend surveillance. This is because when important events happen, people generally search the web first to acquire the desired information [18,22,23], which becomes user behavior that is logged on search engines (i.e., in

* Corresponding author.
E-mail addresses: d03725003@ntu.edu.tw (Z.-H. Fang), patonchen@ntu.edu.tw (C.C. Chen).

http://dx.doi.org/10.1016/j.dss.2016.06.001
0167-9236/© 2016 Elsevier B.V. All rights reserved.

the query logs), and which in turn corresponds well with the development of trending topics [18,20,21]. In this light, the query logs can be potentially used for efficient trend surveillance. In the past, query logs were the private asset of search engine companies and could not be accessed by the general public. However, they are now accessible for the retrieval of the latest search information through a number of online web services. For example, Google trends,[1] which was launched in 2006, provides information on how often a particular search term is queried relative to the total search volume in a particular time period across various regions of the world. Since the web service provides the up-to-date search behavior of users, it attracts many researchers to develop trend surveillance systems using search engine query logs. For instance, Eysenbac [1] observed a high correlation between the usage of epidemic-related terms queried on web search engines and the intensity of epidemics, and found that search engine query logs are effective healthcare surveillance indicators; Chen and Tsai [5] validated that the frequency count of business-related queries are highly correlated with the status of a business cycle, and leveraged query terms to develop a business cycle surveillance system; Li et al. [21] constructed an ontology framework to choose unemployment-related queries, and applied the queries to a support vector regression model to predict future unemployment rates. Basically, the success of the surveillance systems depends on the quality of the selected query terms: the surveillance systems cannot predict a trend status correctly if the query terms are off-topic. Some systems (e.g., [4, 21]) thus consult domain experts to compile query terms relevant to trending topics. However, the manual compilation takes time. While many systems (e.g., [2,3,5]) employ techniques of feature selection to identify query terms automatically, the feature selection techniques require a large document corpus. The query log of each term in the corpus needs to be downloaded and examined in order to acquire terms representative of trending topics; for this reason, the computational cost is high. Since the manual and automatic query term identifications are costly, the existing systems generally are specific to a single topic.

In this paper, we propose a novel trend surveillance system using the information of search engines. We develop an efficient feature selection method, called *TF-LTR* (Term Frequency-Learning to Rank), which is adaptable to different trending topics. Instead of preparing a large document corpus, the feature selection method requires a small document corpus composed of a few top-ranking documents returned by search engines; the method leverages the ranking order of the documents and the frequency of the terms in the documents to select representative query terms relevant to a trending topic. Techniques of pair-wise learning to rank are employed to measure a term's discriminative power in making a document that is ranked higher in the ranking list. The discriminative power is combined with the term frequency which denotes the on-topic degree of a term to measure a term's representativeness against the trending topic. Representative terms are selected as the indicators of the trending topic and their query frequencies are incorporated into a state-of-the-art data mining method to train a surveillance model which monitors the development of the trending topic. Evaluations based on trending topics of different domains demonstrate that our surveillance system is able to accurately predict the status of various trending topics, and the selected query terms and their query frequencies reveal interesting human behavior patterns for different trending topics. Our experiment results show that *TF-LTR* feature selection method is robust and it outperforms other popular feature selection methods. We also demonstrate that representative query terms can be extracted efficiently and effectively from a small corpus by making use the ranking order of documents.

The remainder of this paper is organized as follows. Section 2 provides a review of related works. In Section 3, we present the proposed surveillance system, and then in Section 4 we evaluate the system's performance. Section 5 summarizes our conclusions.

2. Literature review

We begin this section with a review of the trend surveillance systems using search engine information. We also review a number of popular feature selection methods because the core of the proposed framework is feature selection that selects representative query terms for trend surveillance. These methods will serve as the baselines for the performance evaluation.

2.1. Trend surveillance systems using search engine information

Search engine information has been widely adopted to supervise trend status in different domains. Regarding epidemic surveillance, Eysenbach [24] first examined search engine information to inspect the outbreak of epidemics. He observed that epidemic-related searches are generally consistent with the development of epidemics and he presumed that the search frequency of epidemic-related query terms would be an effective indicator of epidemic surveillance. He subsequently used the correlation between the epidemic-related searches on Google and the intensity of epidemics, and demonstrated that the epidemic-related searches can accurately predict the outbreak of epidemics. Ginsberg et al. [2] also utilized Google's search information for epidemic surveillance. The authors scanned the search database of Google to identify query terms that could model the Centers for Disease Control (CDC) influenza-like illness (ILI) visit percentage in the United States. Forty-five query terms out of 50 million candidate searches were selected as indicators to develop a linear regression model which periodically predicts the inflection number of influenza. Fang et al. [4] modeled epidemic surveillance as a data classification problem and compared the surveillance performance of difference machine learning models using query logs of search engines. The authors evaluated various generative and discriminative classification models, and validated that generative models, such as the Naive Bayes model, normally classify the status of dengue accurately.

In addition to epidemic surveillance, many studies also employ query logs to monitor economics-related statuses. For instance, Askitas and Zimmermann [25] utilized the search information of Google Insights to predict the unemployment rate in Germany. The authors manually selected four sets of search queries that were relevant to the topic of unemployment. Their query frequencies were then considered as time series data to construct an error correction model. The experiment results showed that their model could predict the German unemployment rate with a high degree of accuracy. Vosen and Schmidt [22] observed that people who search for consumer goods are likely to purchase the goods, and thereby utilized the query logs of search engines to predict American Consumption Confidence. Their prediction model achieved a significant improvement over the traditional models which are usually based on economic variables, such as the Consumer Confidence Index (CCI) and the Michigan Consumer Sentiment Index (MCSI). Choi and Varian [18] applied query logs to the prediction of retail sales, vehicle sales, real estate sales, and travel package sales. They demonstrated that the predictions based on query logs are more accurate than the predictions based on the methods without using query logs. Chen and Tsai [5] investigated the query frequency of search engines to survey the status of business cycles. Rather than consulting domain experts, the authors employed a correlation coefficient to automatically retrieve query terms whose query frequencies were highly correlated with the status of the business cycle. The selected query terms and the corresponding query frequencies were incorporated into a Naïve Bayes model to predict the status of the business cycle. Data discretization techniques have also been implemented to reduce the sparseness of query frequencies. Li et al. [21] developed an ontology-based web mining framework to predict the unemployment rate. The authors consulted domain experts to construct a labor economics ontology from which query terms regarding labor economics concepts were extracted by means of the feature selection techniques.

[1] https://www.google.com/trends/.