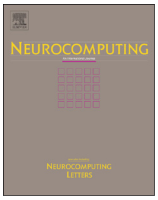
Neurocomputing 167 (2015) 596–603

Contents lists available at ScienceDirect

Neurocomputing

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



Research and application of public opinion retrieval based on user behavior modeling

Baocheng Huang, Guang Yu*

School of Management, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 28 December 2014
Received in revised form
25 March 2015
Accepted 12 April 2015
Communicated by Xudong Zhao
Available online 30 April 2015

Keywords:
Public opinion analysis
Public opinion data organization
Information retrieval

ABSTRACT

This paper designs a system of network public opinion analysis based on the specific way of "matching Topic with Opinion" which can organize public opinion data, avoid the redundant data and retain the original information structure of the opinion. And this article proposes a user model founded on the user access behavior to scientifically classify and represent the relevant theory of the users' retrieval behavior, then to analyze and manage the retrieval results which can to provide the accurate and relevant search results of public opinion information for users.

1. Introduction

Many companies have been developed the application system of network public opinion monitoring with the rapid growth of Web data and the special requirements of public opinion analysis. We learned from the study of human behavior science that the human behavior can be expressed as a multi-function by combined human internal requirements with external environment actually, and the human retrieval behavior is part of the purpose behavior category. The man is conscious and unconscious to identify a target when they retrieve information (blind searching almost does not in human behavior). And the humans get satisfaction after reaching the retrieval target [1]. The behavior of human is to treat action as a criterion, correspondingly, a series of actions making one's behavior. Therefore, if we can grasp the user retrievals behavior, we can speculate the user retrievals target and can predict the next retrieval step, which help to greatly improve the accuracy of search engine and the efficiency of retrieval and to simplify the retrieve steps for users [2].

Development trend in user behavior model has began to show up. There are the following characteristics. That available search behavior analysis methods could only analyze one single type of search behavior, which leads to the problems that the user interest could not be effectively get, a joint analysis method of multiple search behaviors was proposed. By combining the analysis of page dwell time, mouse click times etc. Other than one single type of user behavior were gained for user interest analysis, and joint analysis of high dimension data composed of multiple user behaviors was realized with timeliness ensured for online behavior analysis. The analysis method of text from the network for the user's behavior is more effective [11,12,19,23].

2. Architecture

This chapter designs the architecture of SEB prototype system (in Fig. 1), according to the characteristics of the network public opinion data and the specific demand of search engine for the network public opinion analysis. The SEB prototype system includes the web crawler, webpage analysis, word segmentation, query expansion and sorting and others module. The main sources of SEB information data are from the network public opinion data, likely, blog, comments, weibo and other news to comply with the requirements of network public opinion monitoring [3]. The Web crawler module is responsible for grasping the Internet public opinion information, and saving it to the original webpage in the database for further processing. The webpage analysis module is to identify and remove the noise content from WebPages and to extract the text information and title, keywords, anchor text and the chain of inside and outside information. The analysis work of blog and Weibo are more than common which need including authors, publication time, keep abreast of the corresponding responses and so on. Depending on the URL, crawl time, grab the host and crawl threads, each webpage can generate itself webpage Docid as a unique identification. Then we can accord the digital signature Contentid to judge the repeatability degree of WebPages [4,9,10,14,18].

* Corresponding author.
E-mail address: yug@hit.edu.cn (G. Yu).

http://dx.doi.org/10.1016/j.neucom.2015.04.029
0925-2312/© 2015 Elsevier B.V. All rights reserved.

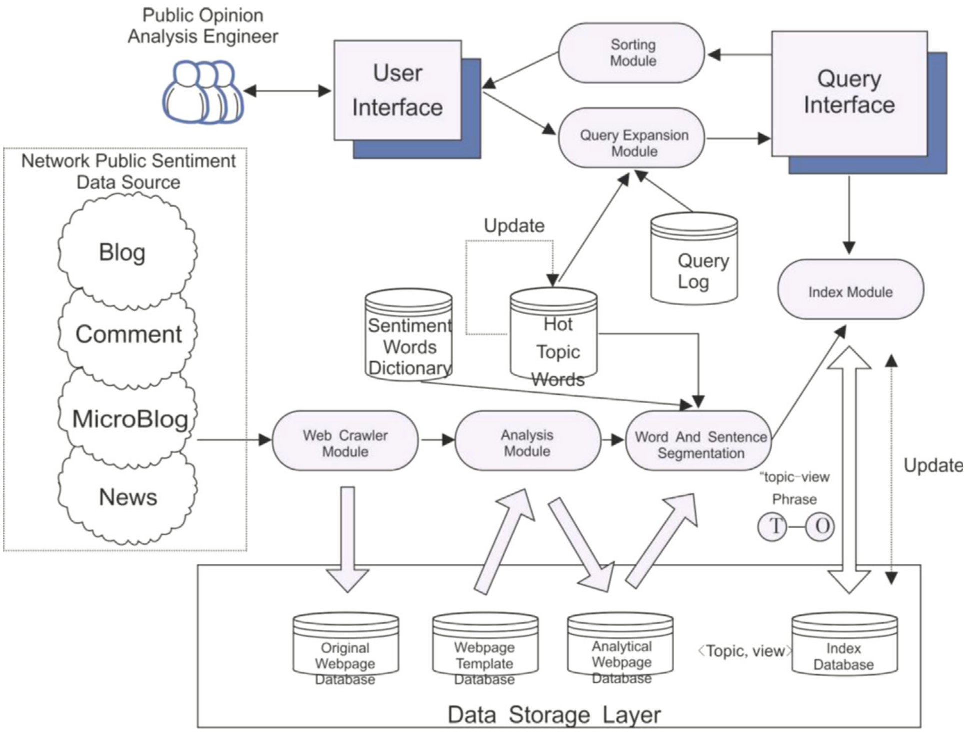


Fig. 1. The architecture of SEB prototype system.

Different from traditional search engine, the view on Internet public opinion data has structural characteristic, SEB need decompose the content of text into sentence firstly. After this, the word segmentation module is considering the word emotional dictionary and hot topics thesaurus to retain the word pairs by matching "topic—Opinion" and to remove the irrelevant words. In order to improve the effective retrieval, the SEB in the index module for the pairs words from "topic—Opinion" to establish an inverted index and store it in the data storage layer [5]. Meanwhile, in response to the real-time requirements of public opinion monitoring, the hot topics thesaurus regularly add the new words and hot words into thesaurus, and the index data in the SEB system also to be renewed periodically.

The end-users of Internet public opinion analysis submit queries to the SEB system. According to the query words, the query expansion module uses the hot topic thesaurus and the related words in query log for expanding the original query terms to obtain more relevant information in search results. Query interface receive the user queries and help to query the other relevant results in the index file. Considering the relevance of topic and the consistency of opinion of the query results, the Sorting module is showing the network public opinion analysis results to users [6].

2.1. The web crawler module

Depending on the generated contents from the blog, comments and microblogging and other network users, SEB takes four crawl types to ensure the integrity and real-time of public opinion data.

(1) The Initial Crawl. Taking the breadth-first strategy to grasp the URL in Seed URL file layer by layer. Then storing these WebPages into the original webpage database and these analysis results into analysis webpage database.
(2) The White list Crawl. Taking the breadth-first strategy to grasp the URL in White list file layer by layer. When grasping data in the White list Crawl, the system only collect itself domain Web pages and exclude others information. Such as, there is one white list file URL: http: //weibo.com, then only crawl the Web pages in Sina Weibo site layer by layer and ignore the other site URL which links from Sina Weibo site.
(3) The new discovery URL crawl. There are the URLs in White list crawling process, which not in the White list scope, and not in the original webpage database. The procedure of new discovery URL crawl is the same with the initial crawl.
(4) The Update crawl. There are the URLs in White list crawling process, which not in the White list scope, but in the original webpage database. The system will crawl this URL and judge the repeatability of the new webpage contents and the original webpage contents. If they are different, then it indicates that the webpage has been updated and need to overwrite the original webpage contents. The system can record the frequency of update, which equal to (Last crawl time—initial crawl time)/crawl times. And the system will automatically scan the update cycle of the URL database, and grasp the updated URL. When crawling just need to grasp the WebPages which is the URL points, don't need to layer by layer.

It is important to grasp the latest hot topics for network public opinion monitoring. To ensure the down load data timely and completely, the system update its crawl methods. The SEB system designs the update crawl methods shown in Fig. 2.

2.2. The webpage analysis module

As for the original webpage, SEB take different webpage analysis methods according to different types of network data. (Shown in Fig. 3).

In Fig. 3, the webpage analysis steps as followings:

(1) Taking preliminary analysis and preprocessing for the WebPages of the original webpage database and using a new format to storage. The storage formats include the web URL,