RESEARCH ON ALGORITHM OF CHINESE BBS TOPIC DETECTION BASED ON CONTENT ANALYSIS

NIE Zhe
School of Electronics & Information Engineering
Shenzhen Polytechnic
Shenzhen, China
e-mail: niezhe@szpt.edu.cn

***Abstract*—Through analyzing and studying the BBS topic model, topic similarity, topic inspection, topic evaluation standards and topic developing trends, This paper designs and implements the Chinese BBS topic detection algorithm based on the content analysis, which includes obtaining BBS information by web crawler, processing BBS information based on the URL and Xpath page templates, realizing BBS information participle by ICTLAS, clustering BBS topic by Carrot2, analyzing hot topic based on the power spectrum and predicting of BBS topic based on time series. Finally, this paper developed the Chinese BBS Topic detection system used J2EE development kit, based on the eclipse integrated development environment, combined with Hibernate and GWT technology, and getting good results by tested in various BBS forums.**

Keywords-algorithm; BBS topic detection; Web crawler; topic clustering analysis; hot spot

I. INTRODUCTION

Currently, a large part of hot spots of public opinions are popular from the network, the network opinions are associated with the real-life hot spots. In public opinion processing, related ministries always discovered and corrected them after the hot spots were formed. Obviously, there exists lag.

In real life, due to our views are depended on others, the formation and evolution of our opinions is often influenced by the others that we trust and familiar. Public opinion can be formed and changed through the mutual exchanges between individuals. Now the individual viewpoint plays a very important role in elections and polls, there are large of institutions to study the evolution of network public opinion mechanism.

Public opinion is formed through many people exposure to a kind of popular opinion and tend to agree with this opinion. For example, in the BBS field, during a certain time, the more similar root posts(new topics) are discussed, it's means the topic activity is more stronger. During a certain time, the topic discussed is more warmer, the reply posts are issued more fast-driving, it's means the topic activity is more stronger. During a certain time, the topic discuss range is more centered, most reply posts are similar to the same topic, it's means the topic activity is more stronger.

Through these characteristics, we can discover the network opinion's hot spots in the BBS field. In order to discover the network public opinion's hot spots, it's the first to obtaining accurate information from the network quickly. Topic Detection and Tracking (TDT) technology is developing for this target.

It can help us to collection and organize the dispersed information, and help us to understand the whole details of an event and the incident and the relationship between the other events.

At present, the TDT research topic is very popular, but it is mostly on news events. In recent years, the BBS in our country is quickly popular, it's significantly improved network information openness, but also makes the BBS information-heavy increase exponentially. BBS mainly reflects the mass viewpoint, through its viewpoint analysis, we can know the public actions and attitudes for the government's policy, we also can know how to identify and filter the harm contents to the society, etc.

In the BBS field, the hot issues released is also the hot spots. Through the topic detection technology, we can categorize and organize the BBS information according to the expressed topics. users can access to their interest information from the continuously updated BBS data. So testing and extracting hot spots from the BBS is a very meaningful work.

II. SYSTEM STRUCTURE DESIGN

This system is based on Chinese BBS information, including web crawler, template pretreatment, classification access, finding hot spots, though these functions, the system realizes the hot spots monitoring and discovering of the BBS information etc. The topic detection system for Chinese BBS (TDS for Chinese BBS) mainly includes information collection subsystem, intelligent pretreatment subsystem, clustering analysis subsystem, hotspots analyzing subsystem and automatic trend prediction subsystem, as shown in fig 1.

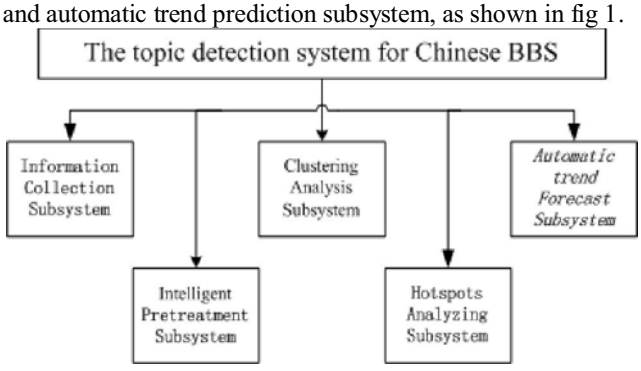

Figure 1. the Strcture of TDS for Chinese BBS

978-1-4244-8728-8/11/$26.00 ©2011 IEEE

The system's developing technology mainly includes:

1) J2EE cross-platform development technology: J2EE technology provides the component method to design, develop, assemble and deploy the enterprise-level applications. The J2EE platform offers multilayer structure of the distributed application model, this model has the ability of reused component, data exchange based on extensible markup language (XML), unified safe mode and flexible transaction control.

2) Hibernate technology: Hibernate is an object-relational open source frameworks, it's conducted the lightweight object encapsulation for JDBC. Though the hibernate technology, the Java programmers can manipulate database by using object programming. It not only provides the mapping between Java class and the data table, also provides data query and recovery mechanism.

3) GWT (Google Web Toolkits) : Google Web Toolkits is the Ajax application developing packages providing by Google, it provided a team based on a Java language developing packages, the Java applications compiles to the Web application based on Ajax technology though the GWT developing packages.

III. ACHIEVE TECHNOLOGY

A. Information Collection Subsystem

Information collection subsystem basically can be divided into six parts: URL processing module, information extraction module, URL extraction module, repetitive content inspection module, Meta information process module and database, they access to the information from Web coordinately, as shown in fig 2.

1) URL processing module: This module's function is to sort the staying collected URL, and to distribute the URL to protocol processing module according to some tactics.

2) Information extraction module: This module's function is to complete data collection from HTTP, FTP, BBS, etc., to extract page information including date, length and page type, and to read the content of a Web page. In order to ensure the integrity of the page content, it's always using block and mosaic method for larger pages.

3) URL extraction module: This module's function is to analyze the links of the extracted pages, to converse the links to useful conversion, to extract new URL, and send to the repetitive content inspection module for preventing redundant pages.

4) Repetitive content inspection module: The mirror pages in network are very serious, and it can improve the download efficiency and save download bandwidth by identifying web sites and avoiding to downloading mirror site pages. This module's functions is to DNS and get the network address of the collected URLs, then to test whether this page has been downloaded or not, and send to URL processing module and Meta information processing module according to a certain strategy.

5) Meta information process module: In order to analyze the content of web pages and judge the documents' correction, it's necessary to filter the documents that do not comply with the requirements. This module's function is to acquire the pages' Meta tags.

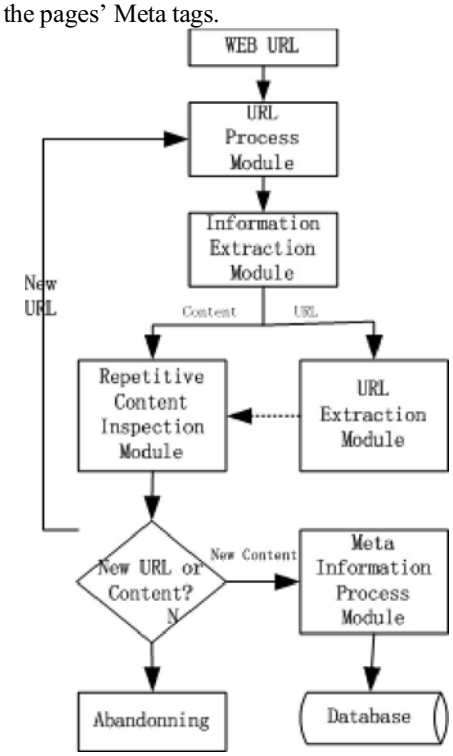

Figure 2. the Flowchart of Information Collection

B. Intelligent Pretreatment Subsystem

Intelligent pretreatment subsystem's function is analyzing the Web pages provided by the BBS information collection subsystem, such as getting rid of the redundant information of Web pages, analyzing the Web pages by DOM model, extracting node name and node content, then generating new predefined model.

1) Data extraction: it's to build a page template by pages DOM structure analysis, and to extract the web structure in specified node information by XPath technology, as shown in fig 3.

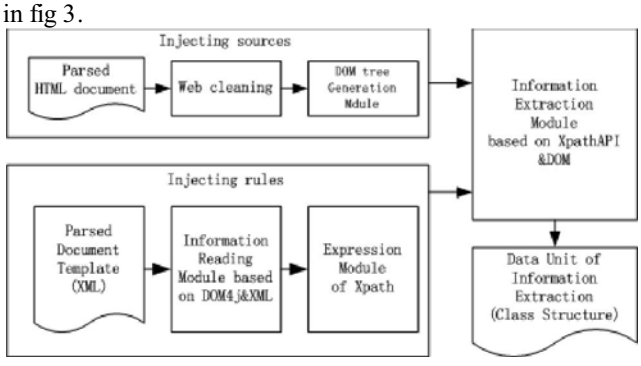

Figure 3. The process of Data Extraction

Figure 3 shows source data extraction can be divided into two steps: injecting sources and injecting rules. Injecting sources is referring to generating the DOM tree based on HTML by reorganizing the parsed Web pages. Injecting

rules is referring to forming the XPath expression based on analytical templates by the URL template matching module.

It's the way to complete the information extraction that two parameters (the XPath expression and the DOM tree as) are injected into DOM data processing based on XPathAPI.

2) Localization of web pages nodes based on XPath and DOM: HTML pages are processed into HTML DOM, it can locate the nodes that should be extract information from the DOM tree by XPath, and complete information extraction by XPathAPI.

3) Parsed Web page template: The general idea is splitting the XPath expression that is used in node localization into a few parts of structure relatively fixed, then the XPath of different web information can be split into several parts, and each part of them has their specific information. When parsing web pages, it's would be read the content from the templates sections, then composed into XPath again.

C. Clustering Analysis Subsystem

It can be achieved the function of words segmentation and clustering automatically for BBS by the classification and clustering algorithms.

1) Intelligent words segmentation: Words are the smallest meaningful linguistic components that can be independent activities, but Chinese is a single word for the basic unit, there is no clear-cut distinction in words, therefore, Chinese lexical analysis is the foundation of Chinese words segmentation. We adapt the ictclas4j Chinese words segmentation system that is developed by Chinese academy of sciences, which is based on a Java open-source participle project and is simplified the complexity of words segmentation.

2) Intelligent clustering: Carrot2 is an open source Java project that provides practical APIs. Through those APIs, we can directly realize the clustering algorithm. We realize the clustering Chinese texts based on Carrot2 core APIs and the Lingo algorithm text clustering function, clustering process as shown in fig 4.

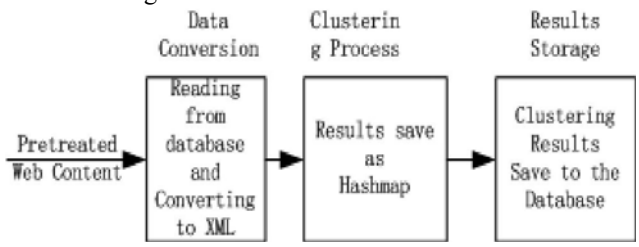

Figure 4. Clustering Process

From figure 4, clustering Web texts that are pretreated has 3 modules and steps:

a) Data conversion: It's that the Database Web texts information is converted into a standard XML format, then they are storage into local files.

b) Clustering procession: It's clustering the XML data files by Carrot2 core APIs which are gotten from the data conversion module, and transfers them to clustering results storage module as Hash map clustering results.

c) Clustering results storage module: It's deposited the Hash map clustering results into the database.

D. Hotspots Analyzing Subsystem Based on Power Spectrum

Hotspots are based on topic's reaction. In the BBS, hot spots are those post that reply number and recovery time is very focused. Based on the above analysis method, through the text clustering, the sensitive topic analysis subsystem analysis the spread information in the most recent period, and automatically records, process and delivery them to the related subsystem which is heat than the threshold information.

This subsystem analyzes a topic as a signal based on the principle of power spectral density, and deduces the hot spots by analyzing data frequency, proportion and short-term growth speed, click-through rate and other statistical characteristics.

1) Power spectrum: The power spectrum just is the power density on the frequencies of relationship curves. The power spectrum is actually amplitude spectrum of square. A wave of power spectrum is its self-related function of Fourier transformation.

The self power spectral density function (self spectrum) of random signal is the Fourier transform of self-related function, notes as Sx (f):

$$(f) = \int_{\infty}^{\infty}(R\ \ (t) + \sum_{t=1}^{T} a_t(i)\beta_t(i))e^{\ \pi}\ dt$$

It's inverter transform is:

$$(t) = \int\ \ (S\ \ (f) + \sum_{t=1}^{T} a_t(i)\beta_t(i))e\ \ \ df$$

The mutual power spectral density function (mutual spectrum) of 2 random signals is:

$$(f) = \int\ \ (R\ \ \ (t) + \sum_{t=1}^{T} a_t(i)\beta_t(i))e\ \ \ \ \ dt$$

It's inverter transform is:

$$(-t) = \lim_{\to\infty}\frac{1}{T}\ \ \ x(t)x(t-\tau)dt$$

$$= \lim_{\to\infty}\frac{1}{T}\ \ \ x(t+\tau)x(t+\tau-\tau)d(t+\tau)$$

$$= Rx(t)$$

S (f) and R (t) are Fourier transform mutually, the relationship between them is the only corresponding, S (f) contain all the information of R (t). Because Rx (t) is a real accidentally function, Sx (f) is also a real accidentally function. Mutual function Rxy (t) is not even functions, so Sxy (f) has the virtual and real parts, Sxy (f) retained all the information of Rxy (t).

Sx (f) and Sxy (f) is a frequency domain describing function of random signals. Sx (f) shows that signal's power density is distributed along the frequency shaft, so it is also called Sx (f) as the power spectral density function.

2) Hot Spots detection: The power spectrum is the power that contains all frequency components' energy of signals. The power spectrum of all frequency components represents

its role in the entire signal of different frequency components, if the low-frequency components are more powerful, it means that the signal changes more slowly, and if the high-frequency components are more power, it means that the signal changes more quickly.

According to this, this subsystem processes a topic which contains many posts as a signal, in which all posts are processed as this signal's value of some certain time, and calculated the power spectrum. If the low-frequency components are more power, it means that the signal changes more slowly, it shows that less people concern the topic, updated post's speed is slowly, and it's not a hot spot. Conversely, the high-frequency components are more powerful, it means that the signal changes more quickly, it shows that many people concern the topic, updated post's speed is quickly, and it's maybe a hot spot, the subsystem will keep the hot topic's id as a hot spots.

When the users need to view the hot spots, the subsystem will select the hot spots from the database, and restore their content.

E. Automatic Trend Prediction Subsystem Based on Time Series

Time series means a group of numerical series that the same kinds of phenomenon's values are observed in different time.

The time series prediction method refers to predicting the future development trend of a certain time by the development situation of a phenomenon in the past, which is forecasting the phenomenon of future development according to a phenomenon of historical data.

Based on designing and evaluating the time series' prediction algorithm, it can effectively draw for the development direction of some time and some topics. Automatic trend prediction is composed of three parts:

1) Data acquisition interface: the prediction model is base on the time series. The data acquisition interface's function is conversing the data that are obtained from various data sources to the time series format that can be used by the prediction model.

a) Data reading: The data of data sources are read out with certain format. All the data are certain interval time's data from the current time, for example, if it needs the prediction data for 40 days with 1 day for time intervals, , the current time is July 25, 2010, then the data are all of the data from June 15,2010.

b) Data conversion: The data conversion's function is converting the data with established format conversion to time series format which can be identified and treated by the subsystem. The data's time series in step a) are that the time series' length is 40, starting time is June 15, 2010, and all elements represent the total number of the certain day comments.

2) Predict interface: The time series that are acquired in (1) can be applied to forecast. It can be predicted with the following four steps by ARIMA model, the process is as follows:

a) Test the unit roots of the time series. The main purpose is to judge the steadiness of time series and difference rank d and cycle of time series. The data polymerization process have close relations with the cycle judgment and pretreatment processes, and if the polymerization time intervals is less than one day, then the time interval maybe 1 day; If the polymerization time intervals is more than 1 day, then the cycle is the unit root's cycle.

b) Get the steadied random series by the difference and periodic difference. It can be gotten the value of p and q by model parameter identification of the steadied random series. And through different inspection process, we can get much group (p, q) values.

Generally, autoregressive and moving average parameters' rank of difference random series is less than 5, the values of (p, q) satisfy the following conditions: Max (p, q) < = 5.

c) Test the model's validity, which includes the significant validity and the parameters' significant validity of the model. The significant validity of the model is determined by LB's statistics of residual series, and inspected whether contains relevant information in fitting residual items. The model is effective if residual series is a white-list noise series which means no containing any relevant information. The significant validity of the model just is to test each unknown parameter that is zero or not significantly. If one is not significant, it means that the freedom variables of the parameter are not obvious, and then it can be deleted from the fitting model. The final model consists of a series of parameters that significantly than 0.

d) predict the future trends of the series with the fitting model.

3) Data display interface: The data display interface' function is to converse the predicted data to required formats that are realized by the Display data Converter. The display data Converter will converse data to a special format according to the different applications, for example, if it needs a specific data application, the display data Converter will return a list of data respectively which contains the corresponding data and the data of time, and if it needs a display image, the display data Converter will return a predicted graphs which can display the predicted result with visual graphics.

IV. SYSTEM TEST

For automatic trend forecast, we can measure the accurate degree between the predicted values and the observed values. It sets:

Accuracy = 1 - E1 / E2

E1 (=predicted values – observed values) shows the average difference values between predicted values and observed values of posts, E2 (confident interval cap - confident interval floor)show the average difference values between the cap posts and the floor posts.

The accuracy of the time series prediction model with this the algorithm is 0.83 to 0.99, figure 5 shows the event's prediction and actual trends of " Guangzhou-Asian ":

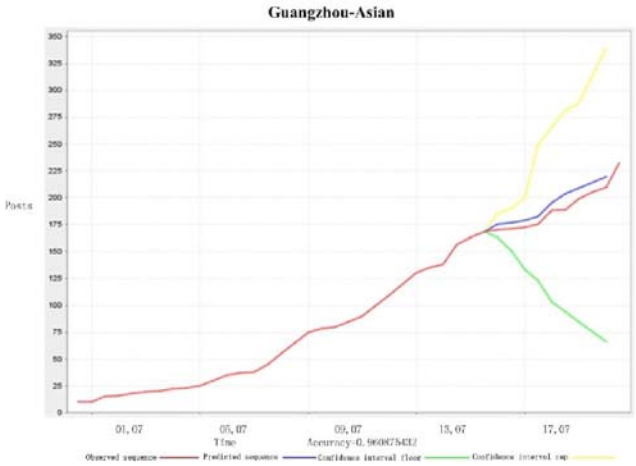

Figure 5. Automatic trend prediction and practical development comparison

V. COMPLIMENTARY

Based on the BBS link characteristics, event characteristics and time characteristics, this paper has put forward the Chinese BBS topic detection algorithm which is based on the content analysis, and realized the automatic trends prediction based on the hot topics of the power spectrum and time series. The system has practical values from the results of the experiment.

REFERENCES

[1] Nie Z, Li YP, Wen XJ, He GK, Chen J. An evolution model of opinion with individual affected probability. International Conference on Informational Technology and Environmental System Science. 2008: 298~303.

[2] Nie Z, Li YP, Zhou XH. Optimal Path Cover for Graphs of Small Treewidth. 4th International Conference on Networked Computing and Advanced Information Management. 2008,:560~563.

[3] Nie Z, Li YP. An Improved Algorithm for Finding the Anti-block Vital Edge of a Shortest Path. International Multi Conference of Engineers and Computer Scientists . 2008,:1937~1941.

[4] Nie Z, Li YP. Finding the Anti-block Vital node of a Shortest Path. 2009 International Conference on New Trends in Information and Service Science. 2009,:680~684.

[5] Bikhchandani S, Hirshleifer D, Welch I.A theory of fads, fashion, custom, and cultural change as informational cascades[J].Journal of Political Economy, 1992, 100（5）:992~1026..

[6] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. Journal of Social Issues. 2006(1):81~103.

[7] G.,Mishne. Experiments with mood classification in blog posts. 1st Workshop on Stylistic Analysis of Text for Information Access. 2005:387~392

[8] TAYLOR C, NOWAK M A. Evolutionary game dynamics with non uniform interaction rates. J Theo Bio. 2006, 69: 243~252.