INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):183-200, April 2017.

Model of Network Topic Detection Based on Web Usage Behaviour Mode Analysis and Mining Technology

M. Chen

Mo Chen
1. Business College of Beijing Union University
A3, Yanjingdongli, Chaoyang district, Beijing, 100025, P.R. China.
mo.chen@buu.edu.cn
2. School of Information, Renmin University of China.
No.59, Zhongguancun Street, Haidian District, Beijing, 100872, P.R. China.
chenmoky@sohu.com

Abstract: This research has caught researchers' wide attention for detecting network topic exactly with the arrival of big data era characterized by semi-structured or unstructured text. This paper proposes a model of network topic detection based on web usage behaviour mode analysis and mining technology taking Web news as object of research. The author elaborates main function and method proposed in this model, which include the analysis module of Web news instance clicking mode, the analysis module of Web news instance retrieval mode, the analysis module of Web news instance seed and the analysis module of similar Web news instance supporting topics. Based on these functions and methods, the author elaborates main algorithm proposed in this model, which include the mining algorithm of Web news seed instances and the mining algorithm of similar Web news instances supporting topics. These functional algorithms have been applied in processing module of model, and focus on how to detect network topic efficiently from a large number of web usage behaviour towards to Web news instances, in order to explore a research method for network topic detection. The process of experimental analysis includes three steps, firstly, the author analyses the precision of topic detection under different method, secondly, the author completes the impact analysis of Web news topic detection quality from the number of Web news instances concerned and seed threshold, finally, the author completes the quality impact analysis of Web news instances mined supporting topic from the number of Web news instances concerned and probability threshold. The results of experimental analysis show the feasibility, validity and superiority of model design and play an important role in constructing topic-focused Web news corpus so as to provide a real-time data source for topic evolution tracking.

Keywords: web usage behaviour, network topic detection, clicking mode analysis, retrieval mode analysis.

1 Introduction

With the arrival of big data era, the field of information technology and Internet has developed a challenging stage so far. According to survey of TeckTarget that is a global leading professional IT network media [15] [2] [22], it has shown that the number of enterprises' data has broken through PB level with development of network, social media, business and other fields. Based on data existed and existing, people should think how to analyse complicated network data showing a tendency of explosive growth [4] [5], which have been concerned and are characterized by semi-structured or unstructured text, nevertheless, in whole process of cognizing network data, detecting topic exactly and effectively is the important and critical application direction.

In a mass of network data, the number of Web news released has reached EB level with events that continue to take place in social [6] [7], which shows the 4V features of big data, it is volume, variety, velocity and value [15] [9]. Based on these features above, Web news should reflect high currency and reliability, on the basis of which, the topic contained in Web news should be condensed quickly and its path of evolution should also be tracked nearly in real time. However, how to detect network topic efficiently from a large number of web usage behaviour towards to Web news instances, it has become an urgent problem solved to construct a topic-focused Web news corpus so as to provide real-time data source for topic evolution tracking.

This paper proposes a model of network topic detection mainly containing four processing modules based on web usage behaviour mode analysis and mining technology taking Web news as object of research. The author elaborates function, method and technology on every processing module of the model in detail, which have been used or completed, and focuses on how to detect network topic efficiently from massive web usage behaviour towards to Web news instances. This process of research does key contribution for exploring a method for network topic detection, this experimental analysis results show the feasibility, validity and superiority of model design and implement.

2 Related works

In recent several years, some scholars have conducted some research about network topic detection method using different theory and technology. For example, Yang et all. survey research on the method of topic link detection based on improved information bottleneck theory [10], in this paper, a method of representing text is proposed, which can divide text into several sections of sub-topic features based on the regular pattern of semantic distribution and improve information bottleneck theory, then, the text represented by the attributes is utilized to do topic link detection, the experimental results have shown that this method has a fast convergent rate, and can improve the performance of topic link detection system. Suhara, Yoshihiko and others survey research on the method of information detection based on sentence-level topic [11], in this paper, the text sentence-level diversity features based on the probabilistic topic model is proposed, an information content classifier is also constructed combining features proposed, the experimental results show that this method outperforms the conventional methods. Pang, JB and others survey research on the method of unsupervised web topic detection using a ranked clustering-like pattern across similarity cascades [12], in this paper, a method using a clustering-like pattern across similarity cascades is investigated from the perspective of similarity diffusion, a topic-restricted similarity diffusion process is also proposed to identify real topic from a large number of candidates efficiently, the experimental results demonstrate that this approach outperforms the state-of-the-art methods on several public data sets, those works are related to author's research direction of network topic detection and application.

In recent several years, some scholars have also conducted certain research about method and technology of web usage behaviour analysis and mining. For example, Dziczkowski, Grzegorz and others survey research on the opinion mining approach for web user identification and clients' behaviour analysis [13], in this paper, an approach based on statistical analysis of natural language is proposed, three different methods are used for classifying opinions from clients' data, two new methods are introduced based on linguistic knowledge, in order to assign a mark dependent upon the client's emotions and opinions described in comments, the effect of experiments demonstrates that the system developed can carry out an evaluation and rating of opinions. Karakostas, Bill and others survey research on the MapReduce architecture for web site user behaviour monitoring in real time [14], in this paper, a MapReduce style architecture is proposed, where the processing of event series from the web users is performed by a number of cascading mappers, reducers, local to the event origin, the experimental results show that this architecture is capable to carry out time series analysis in real time for very large web data sets based on the actual events instead of resorting to sampling or other extrapolation techniques. Zhang, YH and others survey research on the new replacement algorithm of web search based on user behaviour [15], this paper analyses the search ranking based on the user behaviour investigated mass distribution of information on the website, then proposes a replacement algorithm for web search, the simulation experimental results show that this approach under the search algorithm can reduce the execution time of retrieval effectively, and the optimal parameter selection for this blocking organization can be discussed continuously, those works are related to author's research direction of web usage behaviour application for analysis and mining.

Based on the analysis of the related research on network topic detection method and web usage behaviour application technology, experts and scholars have studied on two directions, but the research of constructing a network topic detection model based on web usage behaviour mode analysis and mining technology taking Web news as an object of analysis according to its attention and usage trait is missing. Therefore, this paper proposes a model of network topic detection based on web usage behaviour mode analysis and mining technology mainly, in order to explore how to detect network topic accurately.

3 Problem definition and notations

With rapid development of information and network technology, there are many types of network information, such as short text of micro blog, short, moderate or long text of Web news, long text of document and so on, while the biggest difference is structure of text content among them. This paper selects Web news as the object of research in view of ensuring high adaptability that the model of network topic detection based on web usage behaviour mode analysis and mining technology should have, in order to achieve the ideal effect of topic detection in the aspects of analysis precision and so on, this research provides scientific method for constructing and validating model of network topic detection.

3.1 Usage feature analysis

Users can search and browse Web news from different dimensions, granularities and frequencies, which have been extracted and analyzed, these processes have been elaborated in previous journal article published by authors [16], [17], [18], [19]. In the process of searching and browsing Web news, the user usage behaviour can be recorded, which not only explains Web news features used by users, but also contains the concerning topics hidden in Web news instances. Therefore, based on the analysis of Web news usage features, it is conducive to discover knowledge hidden in massive Web news, detect topic that the users are concerned about, track a series of events occurring in topic, and comb out process of event evolution.

From the perspective of global usage, every Web news instance concerned by users can be seen as a node in the range of Web news websites with authority, and the node set of some relevant instances supporting social events can be considered as a topic, each topic can also trigger a series of events, therefore, when users directly concEren a series of topics reflected by the social event, not only browse multi Web news instances content that support the topic, but also browse a series of events triggered by the topic [21]. From the perspective of local usage, when users search for Web news, in addition to input keywords that are related to the social events reported by Web news, but also input semantic keywords that may appear in Web news title or content, core event, core event occurring time, core event occurring location, subjective or objective object of the core event, and relation event information triggered by the core event [21]. Therefore, in the process of topic detection towards to Web news, if the usage features of Web news can be considered, then it can be mined for social events reported by Web news, topics concerned by social events, Web news instances supporting topics, which will provide a Web news topic corpus with high quality for Web news topic evolution analysis.

3.2 Topic detection norm analysis

Based on the analysis towards to the usage features of Web news, if logical relationship need to be mined from behaviour data among the social events reported by Web News, the topics concerned in social events, Web news instances supporting topics, then the data and norms adopted should be specified in the process of topic detection [22], [23], which include user behaviour records, Web news clicking frequency based on S-U, Web news clicking frequency corresponding to URL, Web news clicking rate corresponding to URL and so on.

In the application process of Web news topic detection for social events, users can input the interesting keywords searched, when clicking submit request, Web news page will show multitudinous title, releasing time, releasing source and content link of Web news instances, and when users click on the contents link of the Web news instance, the application platform will record user names, the search information submitted, the behaviour clicking on Web news instances, the usage time and other data items, in which the search information submitted is expressed in English, but processed in this paper by the way of Chinese.

Based on the Web news usage behavior above, (s, u) of $S - U$ information can express retrieval keywords and Web news instances URL contained in behavior synchronously, $fq(s, u)$ can express Web news instances clicking frequency based on $S - U$, which explains the number of (s, u) appearance in a certain time period, $fq_i(u)$ can express Web news clicking frequency corresponding to URL, which explains the number of the Web news instance appearance in the i particle size of a certain time period, $fq(u)$ can express Web news instance clicking frequency in a certain time period as shown in formula 1, $rt_i(u)$ can express Web news clicking rate corresponding to URL as shown in formula 2.

$$fq(u) = fq_1(u) + \ldots + fq_i(u) + \ldots + fq_n(u) \tag{1}$$

$$rt_i(u) = \frac{fq_i(u)}{fq(u)} \tag{2}$$

Based on the Web news usage behavior above, it can be converted into a graph $G = (S, U, E)$, in which S can express the set of retrieval keywords submitted, U can express the set of Web news instances URL clicked, E can express the set of edges between S and U, the edge (s, u) can express behavior of clicking the Web news instance after submitting search request for users, whose weight value is $fq(s, u)$ corresponding to it.

$S(u)$ can express the set of S, which directly connect with u in G, $U(s)$ can express the set of U, which directly connect with s in G, $D(s)$ can express the degree of node in search request, which is the number of Web news instance nodes that is connected to retrieval requirement, D(u) can express the degree of the Web news instance node, which is the number of retrieval requirement nodes that is connected to the Web news instance as shown in figure 1, in which the S set is expressed in English, but processed in this paper by the way of Chinese.

3.3 Problem Notations

In this section, the author provides notations used in model and algorithms based on practical value and application direction of Web news topic detection. Let $NewsSet$ be a set of Web news instances, which is a data source using Web news for user and contains a large number of instances in Web news websites with authority. Let $UserBehavior$ be a set of records using Web news

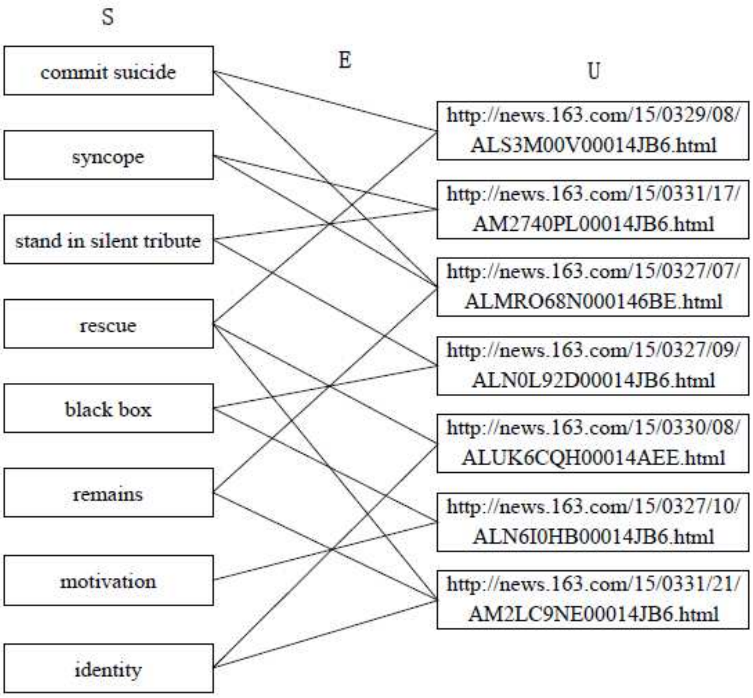

S
E
U
commit suicide
syncope
stand in silent tribute
rescue
black box
remains
motivation
identity
http://news.163.com/15/0329/08/ ALS3M00V00014JB6.html
http://news.163.com/15/0331/17/ AM2740PL00014JB6.html
http://news.163.com/15/0327/07/ ALMRO68N000146BE.html
http://news.163.com/15/0327/09/ ALN0L92D00014JB6.html
http://news.163.com/15/0330/08/ ALUK6CQH00014AEE.html
http://news.163.com/15/0327/10/ ALN6I0HB00014JB6.html
http://news.163.com/15/0331/21/ AM2LC9NE00014JB6.html

Figure 1: The behavior transition diagram based on Web news usage for users

behavior for users, which is also a data source of Web news topic detection, the specific notations are as follows.

Definition 1. Given a set of NewsSet, it can denote using $NewsSet = \{ns_1, \ldots, ns_i, \ldots, ns_k\}$, the range of i is between one and $k.ns_i.url$ stores address of the Web news instance, $ns_i.title$ stores title of the Web news instance, $ns_i.pubtime$ stores releasing time of the Web news instance, $ns_i.pubsource$ stores releasing source of the Web news instance, $ns_i.content$ stores content of the Web news instance, $ns_i.keyword$ stores keywords of the Web news instance, the extraction process of these information has been elaborated in previous articles published by authors [16], [17].

Definition 2. Given a set of UserBehavior, it can denote using $UserBehavior = \{ub_1, \ldots, ub_i, \ldots, ub_n\}$, the range of i is between one and $n.ub_i.username$ stores user names using Web news, $ub_i.searchword$ stores keyword retrieving Web news, $ub_i.url$ stores URL of the Web news instances clicked by users, $ub_i.systemtime$ stores system time using Web news for users.

Definition 3. Based on the definition and notations above, the problem that the Web news topic detection model needs to solve is to detect topic set contained in massive Web news instances from massive Web news usage behavior, and mine set of Web news instances that can support relevant topic, this result can denote using $TopicURL = \{tu_1, \ldots, tu_i, \ldots, tu_k\}$, the range of i is between one and $k.tu_i =< Topic, Topicurl >$, $tu_i.Topic$ can express topic description detected, $tu_i.Topicurl$ can express the set of Web news instances URL mined, which can support relevant topic detected.

4 The design of network topic detection model

In the era background of big data development, it has become an important research direction to detect network topic exactly in Web text mining field through the process of defining detection targets, extracting valuable network information, analysing web user usage behaviour, mining potential topics and applying topics detected and so on.

Based on this process, the model of network topic detection based on web usage behaviour mode analysis and mining technology taking Web news as object of research is divided into four modules, which include the analysis module of Web news instance clicking mode, the analysis module of Web news instance retrieval mode, the analysis module of Web news instance seed and the analysis module of similar Web news instance supporting topics as showed in figure 2.

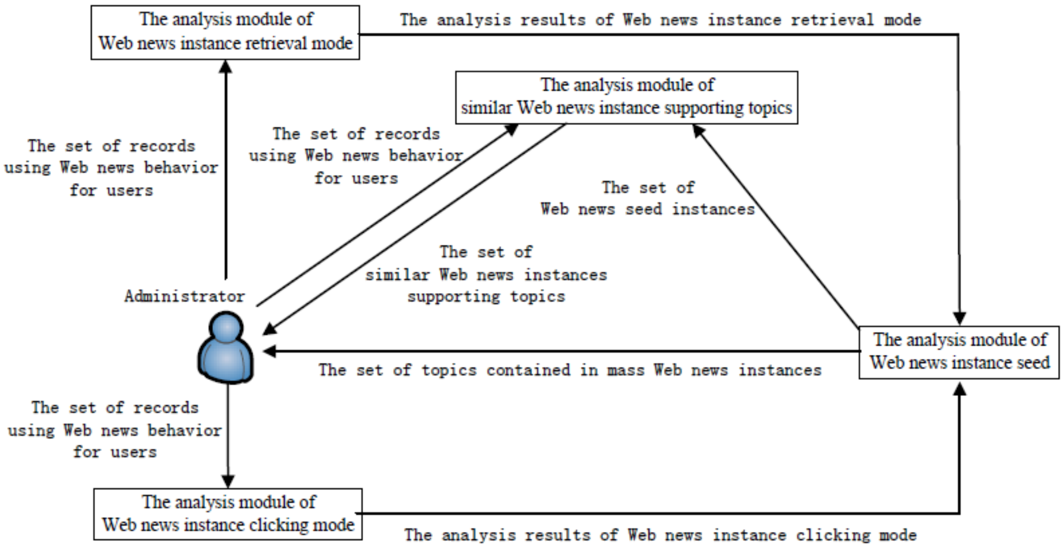

The analysis module of Web news instance retrieval mode
The analysis results of Web news instance retrieval mode
The analysis module of similar Web news instance supporting topics
The set of records using Web news behavior for users
The set of records using Web news behavior for users
The set of Web news seed instances
The set of similar Web news instances supporting topics
Administrator
The analysis module of Web news instance seed
The set of topics contained in mass Web news instances
The set of records using Web news behavior for users
The analysis module of Web news instance clicking mode
The analysis results of Web news instance clicking mode

Figure 2: The model of network topic detection

4.1 The analysis module of Web news instance clicking mode

The inputting content of this module is set of records using Web news behaviour for users, the outputting content of this module is analysis results of Web news instance clicking mode, the main function of this module is to analyse the outbreak and concern mode of Web news instances according to records using Web news behaviour for users, and infer clicking mode of Web news instance.

Based on analysis of Web news usage behaviour, in a certain period of time, the Web news instance mays show a clear outbreak mode regarded as a kind of sensor for social event topic aggregation in records of user's Web news usage behaviour, which indicates whether the Web news instance is able to describe relevant topic reflected in social events. Therefore, in this module, firstly, administrator should use the outbreak mode of the Web news instance to measure whether it is able to describe corresponding topic reflected in social events, whose value is expressed with BR(u) that is a conjecture to sharpness of clicking rate changes, and its range is between 0 and 1 as shown in formula 3, this formula cites characteristic of entropy.

$$BR(u) = 1 - (-(rt_1(u)log_n rt_1(u) + \ldots + rt_i(u)log_n rt_i(u) + \ldots + rt_n(u)log_n rt_n(u))) \qquad (3)$$

In the above formula, n represents the number of continuous granularity components clicked for Web news instances, the granularity can be set at hour or day for unexpected events in society, while the granularity can be set at week or month for normal events in society, if the fluctuations of clicking rate is not strong in the granularity setting for the Web news instance, then $BR(u)$ value is smaller, if the fluctuations of clicking rate is strong in the granularity setting for the Web news instance, then $BR(u)$ value is bigger.

Although using formula 3 can explain outbreak mode of the Web news instance, but its value cannot be absolutely certain that the key information of the Web news instance can describe topic in social events, because the $BR(u)$ value may be one, while the Web news instance is only once clicked by users. Therefore, in this module, secondly, administrator should use the concern mode of the Web news instance to measure whether it is able to describe corresponding topic reflected in social events again, whose value is expressed with CR(u), and its range is between 0 and 1 as shown in formula 4.

$$CR(u) = \frac{log(fq(u)) - Min_{u_i \in U}(log(fq(u_i)))}{Max_{u_i \in U}(log(fq(u_i))) - Min_{u_i \in U}(log(fq(u_i)))} \tag{4}$$

Because the process clicked has a characteristic of power law distribution for Web news instances, so in the above formula, the clicking frequency of the Web news instance is transformed to logarithm, if the Web news instance can be concerned by more users, then $CR(u)$ value is larger, whereas, $CR(u)$ value is smaller. Based on the above measurement of outbreak and concern mode for Web news instances, in this module, finally, administrator should use formula 5 to infer results of the Web news instance clicking mode.

$$ClickMode(u) = BR(u)CR(u) \tag{5}$$

4.2 The analysis module of Web news instance retrieval mode

The inputting content of this module is set of records using Web news behaviour for users, the outputting content of this module is analysis results of Web news instance retrieval mode, the main function of this module is to analyse the degree distribution and similar mode of Web news instances according to records using Web news behaviour for users, and infer the retrieval mode of Web news instance.

Based on analysis of behaviour diagram G using Web news for users, it can be found that the degree of Web news instances shows a characteristic of power law distribution, so in this module, firstly, administrator should use the degree distribution mode of Web news instance to measure whether it is able to describe corresponding topic reflected in social events and express it using $DR(u)$ as shown in formula 6.

$$DR(u) = \frac{log(d(u)) - Min_{u_i \in U}(log(d(u_i)))}{Max_{u_i \in U}(log(d(u_i))) - Min_{u_i \in U}(log(d(u_i)))} \tag{6}$$

Although using formula 6 can explain degree distribution mode of the Web news instance, but its value cannot be absolutely certain that the key information of the Web news instance can describe topic in social event, because the reason of generating Web news instance degree is that users search it using keywords. Therefore, in this module, secondly, administrator should use the similar mode of Web news instance to measure whether it is able to describe corresponding topic reflected in social events again, whose value is expressed with $SS(u)$, in order to solve problem of existing sparse records in user clicking behaviour as shown in formula 7.

$$SS(u) = \frac{2}{n(n+1)} Sum(\frac{Sum(s_{ik}(u)s_{jk}(u))_{k \in dataitem}}{\sqrt{Sum((s_{ik}(u))^2)_{k \in dataitem}}\sqrt{Sum((s_{jk}(u))^2)_{k \in dataitem}}})^{n}_{i<=j} \tag{7}$$

Based on the above measurement of degree distribution and similar mode for Web news instances, in this module, finally, administrator should use formula 8 to infer results of the Web news instance retrieval mode.

$$SearchMode(u) = DR(u)SS(u) \tag{8}$$

4.3 The analysis module of Web news instance seed

The inputting content of this module is analysis results of Web news instance clicking and retrieval mode, the outputting content of this module is sets of topic contained in massive Web news instances and Web news seed instances, the main function of this module is to infer set of Web news seed instances according to analysis results of Web news instance clicking and retrieval mode, and describe corresponding topic referring to Web news key information researched in previous job.

In this process, firstly, administrator should use formula 9 to mine set of Web news seed instances, its weight value is more than or equal to seed threshold, secondly, based on releasing time of Web news, the Web news seed instances should be sorted in set, finally, the corresponding topic should be described using key information of the Web news seed instance, in following experiment, the optimal value of seed threshold will be analysed.

$$SeedURL(u) = ClickMode(u)SearchMode(u) \tag{9}$$

4.4 The analysis module of similar Web news instance supporting topics

The inputting content of this module is set of Web news seed instances and records using Web news behaviour for users, the outputting content of this module is set of similar Web news instances supporting topics, the main function of this module continues to mine set of similar Web news instances with topic described according to sets of Web news seed instances, the corresponding topics described, the records using Web news behaviour for users that have been analysed above.

In this process, administrator should take Web news seed instances as core, and use the probability of first transfer from the Web news instance to itself as possibility that whether it is able to support topic described. If the Web news seed instance is set su, then the variable t_u indicates that whether the Web news instance is able to support topic described by su, the variable t_s indicates that whether the retrieval keyword is able to support topic described by su. If the key information of the Web news instance can support topic described by su, then $t_u = 1$, otherwise $t_u = 0$, if the retrieval keyword can support topic described by su, then $t_s = 1$, otherwise $t_s = 0$. For each Web news seed instance mined, in initial state, t_{su} is set one, then $P(t_{su} = 1) = 1$, and the probability is set zero for any other Web news instances supporting topic described by su, in this way, $P(t_s = 1)$ can be used to calculate the probability arriving su to itself, which is directly linked to the search keyword with it.

$$P(t_s = 1) = Sum(\psi_{su}P(t_u = 1))_{u:(s,u) \in E} \tag{10}$$

$$\psi_{su} = \frac{fq(s,u)}{Sum(fq(s,u_i))_{(s,u_i) \in E}} \tag{11}$$

In formula 11, ψ_{su} expresses transfer probability form the search keyword to the Web news instance, based on this probability, the value of $P(t_u = 1)$ can be calculated by using the following formula for all other Web news instances that are directly connected to it.

$$P(t_u = 1) = Sum(\psi_{us} P(t_s = 1))_{s:(s,u)\in E} \tag{12}$$

$$\psi_{us} = \frac{fq(s,u)}{Sum(fq(s_i,u))_{(s_i,u)\in E}} \tag{13}$$

In formula 13, ψ_{us} expresses transfer probability form the Web news instance to the search keyword, when $P(t_u = 1)$ is greater than or equal to probability threshold, then the Web news instance can be found to support topic described by su, in order to mine set of similar Web news instance that can support topics, in the following experiment, the optimal value of probability threshold will be analysed.

5 The design of network topic detection algorithm

Based on model design of network topic detection above, in this section, the author designs the mining algorithms of Web news seed instances and similar Web news instances supporting topics, in order to make sure that the topic detection has a high accuracy for Web news, in following experiments, the precision of algorithms will be analysed, the optimal value of parameters will be determined.

5.1 The mining algorithm of Web news seed instances

The key information of Web news have been expressed using Web news information extraction and analysis method researched in previous job [16], [17], [18], but what topics the users concern are still unknown in the face of massive Web news released based on social events. Therefore, this algorithm mainly uses set of records using Web news behaviour for users and results of Web news information extraction and analysis, and through analysing Web news instance clicking and retrieval mode to mine set of topic contained in massive Web news instances.

5.2 The precision analysis of topic detection under different method

This experiment compares precision of Web news topic detection under Web news instance clicking mode analysis method, Web news instance retrieval mode analysis method and the method proposed in this paper. As shown in figure 3, the precision represents quality of Web news topic detection using three methods, the red column expresses precision change situation of Web news topic detection using Web news instance clicking mode analysis method that is called DClickMode method in this chart, from its trend, it can be known that the quality of Web news topic detection is not high only through a single analysis for Web news instance clicking mode with increasing number of Web news instances concerned, although the precision has a certain improvement, but the maximum can only float on the 62.8% the blue column expresses precision change situation of Web news topic detection using Web news instance retrieval mode analysis method that is called DSearchMode method in this chart, from its trend, it can be known that the quality of Web news topic detection is not also high comparing with DClickMode method only through a single analysis for Web news instance retrieval mode with increasing number of Web news instances concerned, although the precision has also a certain improvement, but the maximum can also only float on the 63%, the green column expresses precision change situation of Web news topic detection using method proposed by this paper in this chart, from its trend, it can be known that the quality of Web news topic detection has been significantly improved, because of integrating two analysis methods of Web news instance clicking and retrieval mode, while the quantity of Web news instances concerned is less, although the difference of precision is not big comparing with other two methods, the distance of precision is constantly widening among other two methods with increasing number of Web news instances concerned, the maximum can float on the 75.2%, this experiment shows that the quality of Web news topic detection using method proposed in this paper is higher than DClickMode and DSearchMode method.

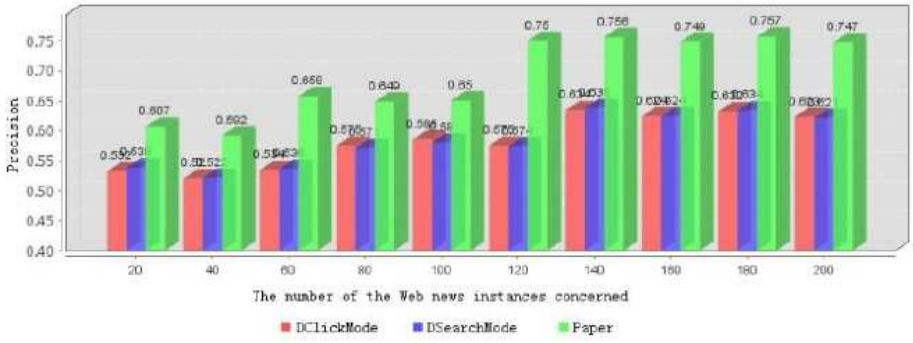


Figure 3: The quality of Web news topic detection under different methods

5.3 The impact analysis of Web news topic detection quality from the number of Web news instances concerned and seed threshold

Under the increasing number of Web news instances concerned, this experiment analyses precision change situation of Web news topic detection through adjusting seed threshold. As shown in figure 4, the precision represents quality of Web news topic detection with increasing number of Web news instances concerned in Y axis through adjusting seed threshold in X axis, when the threshold value is certain, from this graph, it can be known that the quality of Web news topic detection can increase slowly until a relatively stable trend with increasing number of Web news instances concerned, the main reason is that when the number of Web news instances concerned is less, the data and its relationship among them are relatively simple in the process of analysing clicking and retrieval mode, when the number of Web news instances concerned is increasing, the link relationship will exist among data in the process of analysing clicking and retrieval mode, which is conducive to topic detection, so that the precision of Web news topic detection is gradually increasing until more stable, when the number of Web news instances concerned is certain, from this graph, it can be known that the quality of Web news topic detection shows a trend of increasing firstly and then decreasing with increasing of threshold value, the main reason is that when the threshold value is less, a part of inaccurate or approximate accuracy Web news topic is likely to be detected as accurate Web news topic, when threshold value increases to a certain stable interval, only a small number of approximate accurate Web news topic are likely to be detected as accurate Web news topic, when threshold value is increased to a certain value, a part of accurate Web news topic may not be detected, this experiment shows that when the number of Web news instances concerned is 160, and seed threshold is 0.75, the quality of Web news topic detection can reach the highest value, which is close to 78.5%.

5.4 The mining algorithm of similar Web news instances supporting topics

Although the topics concerned by users have been detected using the mining algorithm of Web news seed instances, but what Web news instances supporting these topics the users concern

Algorithm 1 MiningSeedTopic

```
Input: UserBehavior, NewsSet, Threshold, InitialTime, T;
Output: TopicURL
MiningSeedTopic(UserBehavior ub, NewsSet ns, SystemData s);
BEGIN
UserRecord u[]=ExtractRecord(ub);
double br,cr,cm,dr,ss,sm;
GroupUserRecord<u> gur[];
TopicURL tu=new TopicURL();
gur=GroupByURL(u);
for i do=0 to gur.size()-1
    br=CalculateBR(gur[i]);
    cr=CalculateCR(gur[i],u);
    cm=br*cr;
    dr=CalculateDR(gur[i],u);
    ss=CalculateSS(gur[i]);
    sm=dr*ss;
    if c thenm*sm>=s.getThreshold()
        tu.add(ns.getFT(gur[i].url),gur[i].url);
    end if
    if g thenetCurrentTime()-s.getInitialTime()>=s.getT()
        ReSort(tu);
        ReAdjust(s.getThreshold());
    end if
end for
END
```

are still unknown in addition to Web news seed instances. Therefore, this algorithm mainly uses sets of records using Web news behaviour for users and Web news seed instances analysed to mine set of similar Web news instance which can support topics.

6 The experimental analysis and results

In this section, the author carries out experimental analysis and shows experimental results in order to validate feasibility, validity and superiority of model proposed in this paper, in this process, the author adopts experimental environment towards to event of German A320 plane crash shown as follows. The processor is dual core, the memory is 32G, the language of computer programming design is Java, its version is Java SE Development Kit 8, the platform of experimental design and implementation is MyEclipse 2015, the platform of experimental data storage and management is Microsoft SQL Server 2016.

6.1 The quality impact analysis of Web news instances mined supporting topic from the number of Web news instances concerned and probability threshold

Under the increasing number of Web news instances concerned, this experiment analyses precision change situation of Web news instances mined supporting topic through adjusting probability threshold. As shown in figure 5, the precision represents quality of Web news instances

Algorithm 2 MiningSimilarTopicURL

```
Input: TopicURL, UserBehavior, Threshold, InitialTime, T.
Output: TopicURL
MiningSimilarURL(TopicURL tu, UserBehavior ub, SystemData s)
BEGIN
double pts,ptu;
for i do=0 to tu.size()-1
    SearchWord sw1,sw2;
    WebNewsURL wnu1,wnu2;
    sw1=IsExist(tu[i].get("Topicurl"),ub);
    while s dow1!=NULL
        for j do=0 to sw1.size()-1
            pts=Calculate(sw1.get(j).position,tu[i].get("Topicurl"),ub);
            if p thents>=s.getThreshold()
                sw2.add(sw1.get(j));
            end if
            ub.set(sw1.get(j).position,pts);
        end for
        for j do=0 to sw2.size()-1
            wnu1=IsExist(sw2.get(j).position,ub);
            if ( thenwnu1=IsEqual(wnu1,wnu2))!=NULL
                for k do=0 to wnu1.size()-1
                    ptu=Calculate(wnu1.get(k).position,sw2.get(j).position,ub);
                    if p thentu>=s.getThreshold()
                        wnu2.add(wnu1.get(k));
                    end if
                    ub.set(wnu1.get(k).position,ptu);
                end for
            end if
        end for
        sw1=IsExist(wnu2,sw2,ub);
    end while
    tu[i].set(wnu2);
end for
if g thenetCurrentTime()-s.getInitialTime()>=s.getT()
    ReSort(tu);
    ReAdjust(s.getThreshold());
end if
END
```

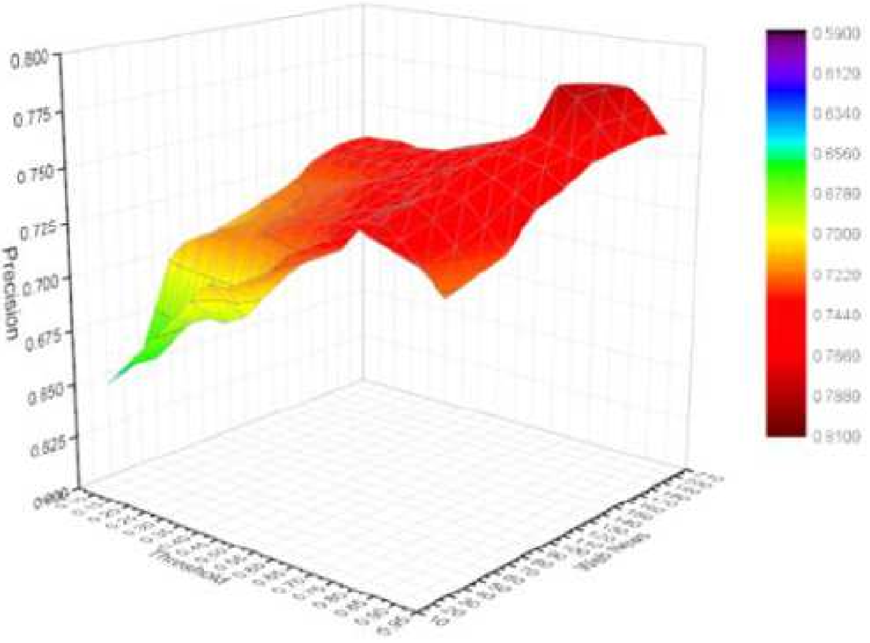


Figure 4: The changing trend of precision with number of Web news instances concerned and seed threshold

mined supporting topic with increasing number of Web news instances concerned in Y axis through adjusting probability threshold in X axis, when the threshold value is certain, from this graph, it can be known that the quality of Web news instances mined supporting topic can increase slowly until a relatively stable trend with increasing number of Web news instances concerned, the main reason is that when the number of Web news instances concerned is less, the data and its relationship among them are relatively simple in the process of analysing similar Web news instances supporting topic, when the number of Web news instances concerned is increasing, the link relationship will exist among data in the process of analysing similar Web news instances supporting topic, which is conducive to instance mine, so that the precision of Web news instances mined supporting topic is gradually increasing until more stable, when the number of Web news instances concerned is certain, from this graph, it can be known that the quality of Web news instances mined supporting topic shows a trend of increasing firstly and then decreasing with increasing of threshold value, the main reason is that when the threshold value is less, a part of inaccurate or approximate accuracy Web news instances are likely to be mined, when threshold value increases to a certain stable interval, only a small number of approximate accurate Web news instances are likely to be mined, when threshold value is increased to a certain value, a part of accurate Web news instances may not be mined, this experiment shows that when the number of Web news instances concerned is 140, and probability threshold is 0.7, the quality of Web news instances supporting topic can reach the highest value, which is close to 75.7%.

6.2 The process analysis of detecting Web news topic

The author illustrates effectiveness of Web news topic detection method implemented in this paper, As shown in figure 6, in this experimental webpage, firstly, users can choose the social event occurred that is German A320 plane crash, secondly, users can choose releasing time of Web news reporting the social event chose, thirdly, users can choose place, object or core event related

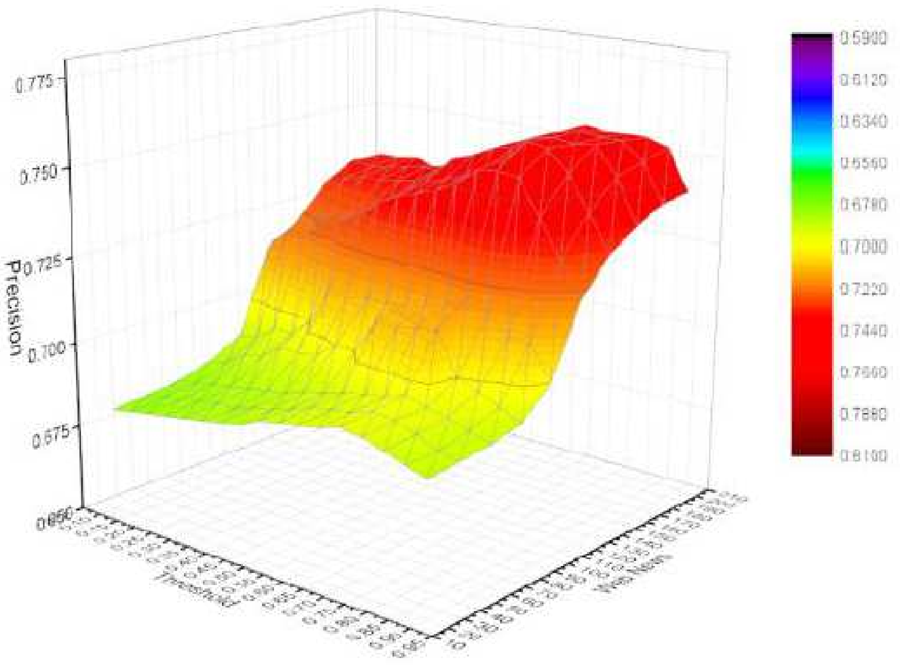
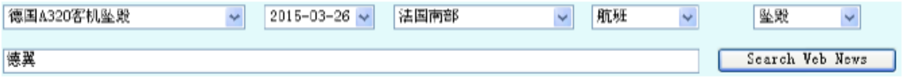


Figure 5: The changing trend of precision with number of Web news instances concerned and probability threshold

to the social event chose, whose selection range is imported from result of Web news extraction and analysis, fourthly, users can input keywords in textfield component again, finally, when users click Web news search button, the experimental platform will display webpage as shown in figure 7. In this webpage, the users can browse retrieval results of Web news instances according to retrieval condition chose or inputted that include title, releasing time, leading content and keyword of the Web news instances. When users are concerned about the Web news instance, and then click its title link, the experimental platform will display corresponding webpage that is linked source URL, at the same time, it will also record the current user name, the search keywords submitted, URL of clicking the Web news instance, usage time and other information in order to detect Web news topic.

As shown in figure 8, in this experimental webpage, firstly, users can choose the social event occurred that is German A320 plane crash, secondly, users can choose the number of topic detected that need to be displayed, finally, when users click submit button, the experimental platform will display top N topics description that have been detected, these topics are sorted in accordance with releasing time of corresponding Web news seed instance, in addition, it will also show set of Web news instances supporting every topic and category information of each topic, Web news instances are also sorted in accordance with its releasing time in set, when users are concerned about the Web news instance supporting topic, then click on its title link, the experimental platform will show corresponding Web news instance browsing webpage.

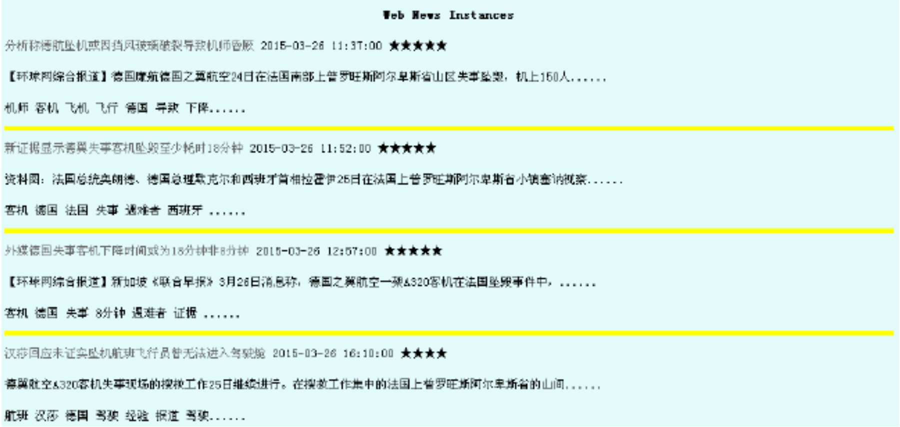
As shown in figure 9, in this experimental webpage, firstly, users can choose the social event occurred that is German A320 plane crash, secondly, users can choose the number of topic detected that need to be displayed, finally, when users click submit button, the experimental platform will show top N topics description detected by the way of time axis, these topics are sorted in accordance with releasing time of corresponding Web news seed instance, in addition, it will also show set of Web news instances supporting every topic, Web news instance are also sorted in accordance with its releasing time, when users are concerned about the Web news instance supporting topic, then click on its title link, the experimental platform will show corresponding Web news instance browsing webpage.



德国A320客机坠毁 | 2015-03-26 | 法国南部 | 航班 | 坠毁

德翼

Figure 6: The searching webpage of Web news



Web News Instances

分析称德航坠机或因挡风玻璃破裂导致机师昏厥 2015-03-26 11:37:00 ★★★★★

【环球网综合报道】德国廉航德国之翼航空24日在法国南部上普罗旺斯阿尔卑斯省山区失事坠毁，机上150人......

机师 客机 飞机 飞行 德国 导致 下降......

新证据显示德翼失事客机坠毁至少耗时18分钟 2015-03-26 11:52:00 ★★★★★

资料图：法国总统奥朗德、德国总理默克尔和西班牙首相拉霍伊25日在法国上普罗旺斯阿尔卑斯省小镇塞讷视察......

客机 德国 法国 失事 遇难者 西班牙 ......

外媒德国失事客机下降时间或为18分钟非8分钟 2015-03-26 12:57:00 ★★★★★

【环球网综合报道】新加坡《联合早报》3月26日消息称，德国之翼航空一架A320客机在法国坠毁事件中，......

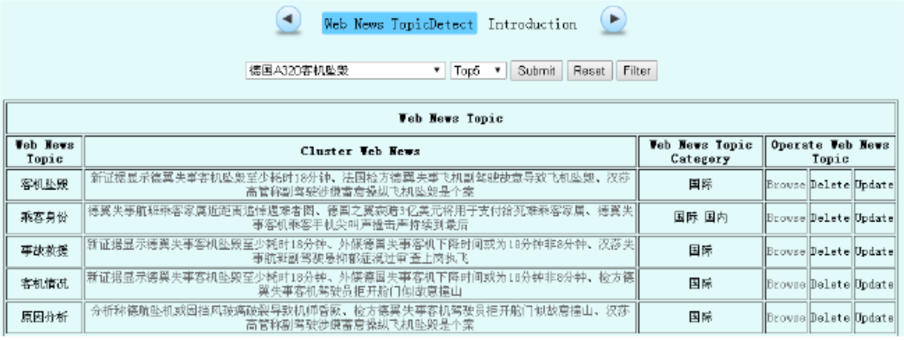
客机 德国 失事 8分钟 遇难者 证据 ......

汉莎回应未证实坠机航班飞行员曾无法进入驾驶舱 2015-03-26 16:10:00 ★★★★

德翼航空A320客机失事现场的搜救工作25日继续进行。在搜救工作集中的法国上普罗旺斯阿尔卑斯省的山间......

航班 汉莎 德国 驾驶 经验 报道 驾驶......

Figure 7: The clicking webpage of the Web news instance



德国A320客机坠毁 | Top5

Web News Topic

Web News Topic	Cluster Web News	Web News Topic Category	Operate Web News Topic		
客机坠毁	新证据显示德翼失事客机坠毁至少耗时18分钟、法国检方德翼失事飞机副驾驶故意导致飞机坠毁、汉莎高管称副驾驶涉嫌蓄意操纵飞机坠毁是个案	国际	Browse	Delete	Update
乘客身份	德翼失事航班乘客家属近距离追悼遇难者图、德国之翼获赔3亿美元将用于支付给死难乘客家属、德翼失事客机乘客手机尖叫声撞击声持续到最后	国际 国内	Browse	Delete	Update
事故救援	新证据显示德翼失事客机坠毁至少耗时18分钟、外媒德国失事客机下降时间或为18分钟非8分钟、汉莎失事航班副驾驶患抑郁症通过审查上岗执飞	国际	Browse	Delete	Update
客机情况	新证据显示德翼失事客机坠毁至少耗时18分钟、外媒德国失事客机下降时间或为18分钟非8分钟、检方德翼失事客机驾驶员拒开舱门似故意撞山	国际	Browse	Delete	Update
原因分析	分析称德航坠机或因挡风玻璃破裂导致机师昏厥、检方德翼失事客机驾驶员拒开舱门似故意撞山、汉莎高管称副驾驶涉嫌蓄意操纵飞机坠毁是个案	国际	Browse	Delete	Update

Figure 8: The webpage of Web news topic detection

Conclusion

This paper completes a research on model of network topic detection based on web usage behaviour mode analysis and mining technology, which takes Web news as research object, takes web usage behaviour application technology as research core and executes process of defining detection targets, extracting valuable network information, analysing web user usage behaviour, mining potential topics and applying topics detected from point of innovation. This result is important and valuable for researchers in the same or related field. In the process of model research, design and implement, this paper proposes the mining algorithm of Web news seed instances and similar Web news instances supporting topics in order to eliminate shortcomings existing in previous traditional method.

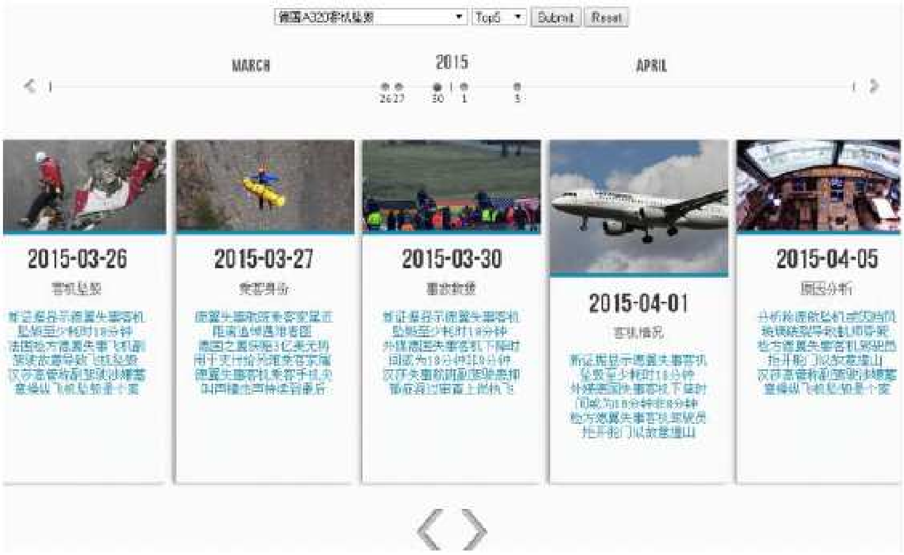


Figure 9: The browsing webpage of Web news topics

The experimental analysis and results of model do key contributions for feasibility, validity and superiority of network topic detection request, improve efficiency of understanding network information for users, enhance availability of websites, build scientifically and improve service functions of websites, and improve business operational efficiency and clicking rate of websites. In a word, the process of research, design and implement model of network topic detection has certain practical application value, which establishes real and exact foundation of corpus for continuative research and application on Web text mining direction.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant Nos.71572015, the National Natural Science Foundation of China under Grant Nos.71271209, the Project of Philosophy and Social Sciences Planning in Beijing under Grant Nos.13JGC090.

Bibliography

[1] Zhang Ji, Li Hongzhou, Gao Qigang, Wang Hai, Luo Yonglong, Detecting anomalies from big network traffic data using an adaptive detection approach, *Information Sciences*, 6(3): 96-97.

[2] Pandey Suraj, Nepal Surya, Cloud Computing and Scientific Applications-Big Data, Scalable Analytics, and Beyond, *Future Generation Computer Systems*, 29(7): 1774-1775.

[3] Zhu Zhiguo, A novel method for discovering frequent changing patterns from historical web access data, *ICIC Express Letters*, 8(9): 2443-2445.

[4] Nasomyont, Tamrerk, A study on the relationship between search engine optimization factors and rank on google search result page, *Advanced Materials Research*, 3(4): 1462-1464.

[5] Guo Yi, Chen Hao, Microblog user ranking based on PageRank and Hadoop, *WIT Transactions on Information and Communication Technologies*, 49(1): 1083-1085.

[6] Zhang Hongli, Huang Shouming, Web Information Extraction Method Based on MapReduce, *Journal of Anhui Science and Technology University*, 27(2): 72-74.

[7] Li Wen, Zheng Bangxi, Deng Wu, Research on Web Information Extraction Model Based on XML and DOM Technologies, *Journal of Dalian Jiaotong University*, 34(3): 96-98.

[8] Zhang Yaming, Tang Chaosheng, Information propagation model based on the dynamics of complex networks in mircoblogging, *Journal of Computational Information Systems*, 10(1): 443-445.

[9] Wu Jiagao, Zhou Fankun, Zhang Xueying, Research of the Extraction Method of Event Properties Based on the Combining of HMM and Syntactic Analysis, *Journal of Nanjing Normal University(Natural Science Edition)*, 37(1): 30-32.

[10] Yang Yuzhen, Liu Peiyu, Fei Shaodong, Zhang Chenggong, A topic link detection method based on improved information bottleneck theory, *Zidonghua Xuebao/Acta Automatica Sinica*, 40(3): 471-479.

[11] Suhara, Yoshihiko, Toda, Hiroyuki, Nishioka, Shuichi, Susaki, Seiji, Automatically generated spam detection based on sentence-level topic information, *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, 1157-1160.

[12] Pang Junbiao, Jia Fei, Zhang Chunjie, Zhang Chenggong, Unsupervised Web Topic Detection Using A Ranked Clustering-Like Pattern Across Similarity Cascades, *IEEE TRANSACTIONS ON MULTIMEDIA*, 17(6): 843-853.

[13] Dziczkowski, Grzegorz, Wegrzyn-Wolska, Katarzyna, Bougueroua, Lamine, An opinion mining approach for web user identification and clients' behaviour analysis, *IEEE Computer Society*, 79-84.

[14] Karakostas, Bill, Theodoulidis, Babis, A MapReduce architecture for web site user behaviour monitoring in real time, *DATA 2013 - Proceedings of the 2nd International Conference on Data Technologies and Applications*, 45-52.

[15] Zhang Yongheng, Feng Zhang, Fei You, A New Replacement Algorithm of Web Search Engine Cache based on User Behavior, *Applied Mathematics & Information Sciences*, 8(6): 3049-3054.

[16] Chen Mo, Yang Xiaoping, Research on Model of Network Information Extraction Based on Improved Topic-Focused Web Crawler Key Technology, *Tehnicki vjesnik/Technical Gazette*, 23(4): 49-54.

[17] Chen Xuegang, Research and realization of E-commerce monitor system based on focused web crawler, *Information Technology Journal*, 12(17): 4033-4035.

[18] Balla, Andoena, Real-time web crawler detection, *2011 18th International Conference on Telecommunications*, 428-430.

[19] Ahmadi-Abkenari, F, A clickstream-based web page significance ranking metric for web crawlers, *2011 5th Malaysian Conference in Software Engineering*, 223-225.

[20] Chen Mo, Yang Xiaoping, Liu Ting, A research on user behavior sequence analysis based on social networking service use-case model, *International Journal of u- and e- Service, Science and Technology*, 7(2): 1-4.

[21] Chen Mo, Yang Xiaoping, Sun Meng, Zhao Yun, Research on model of network information currency evaluation based on web semantic extraction method, *International Journal of Future Generation Communication and Networking*, 7(2): 103-105.

[22] Zhu Tao, Lin Yumin, Cheng Ji, Wang Xiaoling, Efficient diverse rank of hot-topics-discussion on social network, *Lecture Notes in Computer Science*, 8485(1): 522-524.

[23] Lu Ran, Xue Suzhi, Ren Yuanyuan, Zhu Zhenfang, A modified approach of hot topics found on micro-blog, *Lecture Notes in Electrical Engineering*, 269(1): 603-605.