INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, e-ISSN 1841-9844, 14(3), 311-328, June 2019.

Research on Key Technology of Web Hierarchical Topic Detection and Evolution Based on Behaviour Tracking Analysis

M. Chen

Mo Chen*
Business College of Beijing Union University
A3, Yanjingdongli, Chaoyang District, Beijing, 100025, P.R. China
*Corresponding author: mo.chen@buu.edu.cn

Abstract: In the development background of today's big data era, the research direction of Web hierarchical topic detection and evolution characterized by the semi-structured or unstructured data has caught wide attention for academicians. This paper proposes an idea of Web hierarchical topic detection and evolution based on behaviour tracking analysis taking the network big data as the research object, and expounds main implementation methods, which include the instance analysis of the usage mode, the instance analysis of the seed, the set analysis of similar instance supporting the topics, the set analysis of similar instance supporting the events, the evolution analysis of the event, and expounds the algorithm of Web hierarchical topic detection and evolution based on behaviour tracking analysis. The process of experimental analysis is organized as follows, first of all, the experiment analyses the quality of topic detection, the accuracy rate with the number of instance concerned and the seed threshold variation trend, the accuracy rate with the number of instance concerned and the probability threshold variation trend, secondly, the experiment analyses the quality of topic evolution, the accuracy rate with the variation trend of parameter adjustment, the accuracy rate with the number of instance concerned and the similar threshold variation trend, finally, the experiment analyses the time consuming to solve main research problem under different method, the qualitative result of topic detection and evolution under different data set. The results of experimental analysis show the idea is feasible, verifiable and superior, which plays a major role in reconfiguring Web hierarchical topic corpus and providing an intelligent big data warehouse for the network information evolution application.

Keywords: Web hierarchical topic, topic detection, event evolution, behaviour tracking analysis.

1 Introduction

In the development background of Web text mining technology and big data era, so far, the field of intelligent technology has also developed into a more challenging stage [7, 13, 28], the network has become one of services, which can transmit most popular information for users. According to deep survey, the number of network data has gone through EB level in different domain [14, 18, 19, 27]. Academicians should cogitate how to analyse intricate big data, nevertheless, it is an important and key application direction for researching Web hierarchical topic detection and evolution based on behaviour tracking analysis.

In the network big data, the number of Internet news is showing explosive growth as a kind of flow resource with on-going events, which has shown 5V features of volume, variety, value, velocity and veracity [3, 23, 26]. Based on above characteristics, the Internet news should reflect high currency and reliability [5], on this basis, the topic of Internet news should be quickly detected, and its evolution path should be tracked in real time. However, how to research Web hierarchical topic detection and evolution based on behaviour tracking analysis, it has become an urgent problem to build a Web hierarchical topic corpus and provide a real-time big data source for the network information evolution application.

Through researching the literature related to topic detection and evolution technology, this paper proposes an idea of analysing the process for Web hierarchical topic detection and evolution, and expounds main implementation methods and algorithms following with interest Web hierarchical topic detection and evolution from behaviour tracking analysis. This process does important contribution for researching a method of analysing the detection and evolution for Web hierarchical topic, the results of experimental analysis show that the implement of this idea is feasible, verifiable and superior.

2 Related works

In recent years, some scholars have done certain research about the technology of Web topic detection and evolution. A statistical model is proposed [2], in this model, it can combine context with related topics by jointly modelling the topic word with the hash tag and the time stamp, in order to detect and track interpretable topics over time along with their distribution of the hash tag, in this technical context, the experiment demonstrates that this model effectively reveals the process of topic detection and evolution by using the real dataset, this model is different from the traditional topic mining model, it shows serious improvement due to this fact that the distribution of the metadata containing in user content generated can be analysed, so the whole research result does main contribution in the area of topic detection and evolution in the context of the statistical analysis. A topic detection and evolution method is proposed by analysing the semantic word shift, the topic trend, and the evolving dynamic using the data set [4], in this method, it can merge and split local topics in different time periods, in order to track the process of knowledge transfer among topics, in this technical context, the experimental results show that the process of topic detection and evolution usually follows pattern from adjusting status to mature status, and sometimes with readjusting status, this method is different from the statistical analysis, it shows serious improvement due to this fact that the word migration via topic channels has been defined, and three migration types of non-migration, dual-migration, and multi-migration are better to understand topic detection and evolution, so the whole research result does main contribution in the area of topic detection and evolution in the application direction of information retrieval. A topic detection and evolution method is proposed [30], which is called the citation-content-latent Dirichlet allocation method, in this method, it can account for the document citation relation and the content of document itself via a probabilistic generative model, this model can deal with the citation and text information, and its parameters are estimated by a collapsed Gibbs sampling algorithm, in addition, a topic detection and evolution algorithm is designed, which can run in two steps of the topic segmentation and the topic dependency relation calculation, this model and algorithm have been tested by using the online dataset, in this technical context, the experimental results demonstrate that the implementation of the model and algorithm can more effectively detect important topics and reflect topic evolution process comparing with the topic tracking in the knowledge transfer context, so the whole research result does main contribution in the area of topic detection and evolution for designing the model and algorithm. A topic detection and evolution framework is proposed based on the probabilistic topic model [31], in this framework, firstly, the notations, the terminology, and the basic topic mining model is introduced, secondly, three technologies of the topic detection and evolution are applied, which are the discrete time topic detection and evolution, the continuous time topic detection and evolution, and the online topic detection and evolution, thirdly, the application of this framework is discussed, in this technical context, the comparative experiments are completed for different technologies of the topic mining, this framework shows serious improvement than single probabilistic model and does main contribution in the area of the topic detection and evolution performance evaluation. A topic detection and evolution method is proposed based on the analysis of the content similarity or dissimilarity using the textual material [12], in this method, the graph-theoretical technology is applied, in order to deal with the network relationship among the content-similar topics, in this technical context, the explanatory experiment more effectively illustrates usefulness of the approach using the online news articles in different situations, so the whole research result does main contribution in the area of topic detection and evolution in the context of the network analysis.

Based on above analysis of the literature about the technology of Web topic detection and evolution, through comparing the difference among technologies and analysing the main contributions in the research area, if want to research the process of Web topic detection and evolution, in addition to using Web mining technology based on the structure and content, but also using Web mining technology based on the behaviour tracking. Therefore, in recent years, some scholars have also done certain research about the technology of Web behaviour tracking analysis.

A collaborative tagging model is proposed [24], in this model, the technology of the usage behavior tracking analysis is applied to the information extraction direction for web user query in the ontology environment due to different structure data, which is based on the idea of the block acquiring page segmentation, in order to retrieve the tag-based information, in this technical context, the comparative experiments are completed regarding the average precision rate, the time cost, and the storage space rate with existing information retrieval model, it shows that the application of this model can assure research more effectiveness, so the whole research result does main contribution in the area of the behavior tracking analysis. A tracking method of the usage behavior is proposed [17], in this method, a relational graph is established by mining the temporal and causal information among aggregated HTTP request, and an algorithm is designed and implemented for primary request identification, which is a critical task of web usage mining, in order to demonstrate higher value and effectiveness, in this technical context, the experimental result shows that it is a more useful method of analysing a large-scale dataset for the real-world Web access log, so the whole research result does main contribution in the direction of mining value for information available. A web usage mining method is proposed [1], this method can be applied to solve the problem of developing more accurate and efficient recommendation systems, the traditional data protection mechanism focuses on the access control and the secure transmission, which provide only security against malicious the third parties, but not the service provider, so this method can mine efficiently and intelligently data, in order to guide and track the users' usage behavior, and does main contribution in the application direction of the modern E-business. A web usage mining method is proposed [11], in this method, the state-of-the-art session identification technology is used in terms of limitation, feature, and methodology, which provide a structured overview of the research development, in this technical context, the comparative experiments critically review existing session identification technology, the whole research result does main contribution in highlighting the limitation and related challenge, identifying the area where further improvement is required, in order to complement the performance of existing technology. A web usage mining method is proposed [21], in this method, an algorithm is also designed for the data cleaning and filtering using the web log dataset, this algorithm mainly complete the process of preprocessing and clustering the usage behavior, which consists of the use cases for the data cleaning and filtering, the user and session identification, in this technical context, the experiments are carried out to obtain the aggregate clustering results, through this process, two datasets of web usage log are collected and processed, so the whole research result does main contribution in the area of web usage behavior analysis.

Based on above analysis of the literature about the technology of Web behavior tracking analysis, through comparing the difference among technologies and analysing the main contributions in the research area, the scholars have studied two research directions including the technology of Web topic detection and evolution, but do not fully take into account the hierarchical series induced by the usage behavior, if do not fully consider this point, then ignore the process to track the usage behavior from the angle of topic detection and evolution. So this paper mainly takes web usage behaviour record of news big data as the research source, utilizes the method of analysing web usage behaviour tracking to perfect news big data corpus for topic detection and evolution research, and proposes an idea of analysing the process for Web hierarchical topic detection and evolution to solve difficult problems existing in current research status.

3 Problem definition and notation

Under the background of Web big data development, users can retrieve and browse Web news from different dimension, granularity and frequency, which has been analysed and evaluated [15,20,22]. In this process, the time sequence trajectory of users' behavior can be recorded. These data not only record the characteristic of Web news that users use, but also contain the topics reflected in Web news, and the events that are generated based on these topics. Therefore, the knowledge hidden in Web news big data can be mined, the topics that users are concerned about can be detected, a series of events under the topics can be tracked, and the evolution process of events can be combed out based on the process of analysing Web news usage characteristic.

Based on the analysis of Web news structure, contents and semantic feature, every Web news that users concern can be regarded as an instance node in every authoritative Web news network from the perspective of global usage. The social events supported by set of some related nodes can be considered as a topic, and a series of events will be created under each topic. In this way, when users are concerned about a series of topics reflected in a social event, they can not only browse many Web news instances supporting topics, but also browse a series of events that are generated by this topic. When users are concerned about an event, it can also browse more than one Web news instance content that supports this event. From the perspective of local usage, when users retrieve Web news instances, besides inputting the keywords related to social events reported by Web news, they can also input the keywords with five tuple semantic description. Therefore, during the analysis process of Web news topic detection and evolution, the social events can be mined utilizing structure, contents and semantic feature, a series of events caused by the topics can be mined focusing on mining Web news instances supporting the events, the logical hierarchical relationship can be mined between mining objects, which will construct a multi-level structural corpus, it can represent visual topics and events of Web news with a high degree of quality.

Based on the analysis of Web news utility feature, users can retrieve Web news in the search process with a high degree of currency, which is called the time accuracy reporting Web news core events. Users can retrieve Web news with a high degree of truthfulness, which is called the releasing source reliability of Web news. Users can retrieve Web news of high currency with a high degree of truthfulness, therefore, during the analysis process of topic detection and evolution, it should weigh the factors of Web news currency and authenticity, and provide Web news utility instances supporting the topics and events from the perspective of users' utility characteristic for Web news.

Based on the behaviour tracking analysis of web usage characteristic for user, $S-U$ can link the search keywords and the URL instances synchronously, in the $S-U$ relation, S represents the set of the keywords, U represents the set of the URL instances. As shown in the formula 1, $fq(s,u)$ can indicate the clicking frequency of instances, $fq_i(u)$ can indicate the clicking frequency of homologous URL, $fq(u)$ can indicate the clicking frequency of instances in a certain period. As shown in the formula 2, $rt_i(u)$ can indicate the clicking rate of homologous URL.

$$fq(u) = \sum_{i=1}^{n} fq_i(u) \tag{1}$$

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

$NewsSet$ can be defined using the $\{ns_1, ..., ns_{i-1}, ns_i, ns_{i+1}, ..., ns_k\}$, the range of array is from one to k. $ns_i.url$ defines the url instance address, $ns_i.title$ defines the instance title, $ns_i.pubtime$ defines the instance releasing time, $ns_i.pubsource$ defines the instance releasing source, $ns_i.content$ defines the instance contents, $ns_i.keyword$ defines the instance keywords. $UserBehavior$ can be defined using the $\{ub_1, ..., ub_{i-1}, ub_i, ub_{i+1}, ..., ub_n\}$, the range of array is from one to n. $ub_i.username$ defines the user name, $ub_i.searchword$ defines the keywords, $ub_i.url$ defines the URL clicked, $ub_i.systemtime$ defines the system time.

Based on above definition and notation, the issue should be solved to mine topics contained in the instances, mine instances set supporting related topics. From research angle of mining Web news instance set supporting topics concerned by users, the issue should be solved to mine events that are happening under these topics, analyse the instance set supporting these events, so as to reflect the evolution process of Web news topics continuously. This result can be denoted using the $TopicURL$, it can be represented using the $\{tu_1, ..., tu_{i-1}, tu_i, tu_{i+1}, ..., tu_m\}$, the range of array is from one to m, tu_i can be represented using the $< Topic, Topicurl, Event, Eventurl >$, $tu_i.Topic$ can express the topic description detected, $tu_i.Topicurl$ can indicate the seed instances URL set supporting related topics, $tu_i.Event$ can express the description of events under topics mined, $tu_i.Eventurl$ can indicate Web news instance of events supporting topics detected.

4 The analysis of Web hierarchical topic detection and evolution

In view of the problem definition and notation, this paper proposes an idea of analysing the process for Web hierarchical topic detection and evolution shown in figure 1. This framework is used for completing the analytical process for Web hierarchical topic detection and evolution, which include the instance analysis of the usage mode, the instance analysis of the seed, the set analysis of similar instance supporting the topics, the set analysis of similar instance supporting the events, and the evolution analysis of the event. The algorithms are designed for completing the functions and methods of this framework, which include the algorithm of analysing the process for Web hierarchical topic detection, the algorithm of analysing the process for Web hierarchical topic evolution.

4.1 The algorithm of analysing the process for Web hierarchical topic detection

The algorithm of analysing the process for Web hierarchical topic detection is implemented by designing two methods of the usage mode analysis and the topic series construction, the inputting content of this algorithm is the result of semantic five tuple description analysis and utility evaluation for Web news instances, and the user usage behaviour record set, the outputting content of this algorithm is the similar and sequential set of Web news instances that can support corresponding topics.

According to web usage behaviour record, the explosive and attention mode of Web news instances are analysed, in order to infer the click mode of Web news instances, the degree distribution and similarity mode of Web news instances are also analysed, in order to infer the retrieval mode of Web news instances. In accordance with these results of analysing web usage

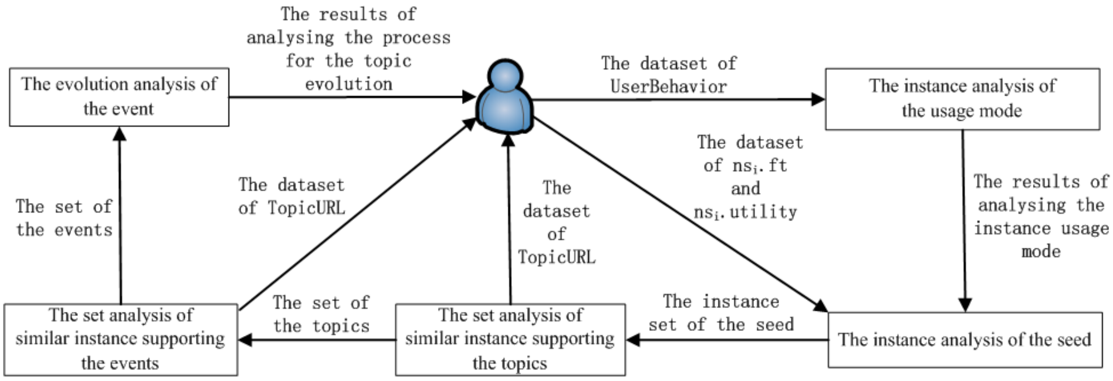
The results of analysing the process for the topic evolution
The evolution analysis of the event
The dataset of UserBehavior
The instance analysis of the usage mode
The dataset of ns_i.ft and ns_i.utility
The results of analysing the instance usage mode
The set of the events
The dataset of TopicURL
The dataset of TopicURL
The set analysis of similar instance supporting the events
The set of the topics
The set analysis of similar instance supporting the topics
The instance set of the seed
The instance analysis of the seed

Figure 1: The framework of analysing the process for Web hierarchical topic detection and evolution

mode, the seed set of Web news instances can be mined, the similar set of Web news seed instances can also be mined. Referring to the utility feature that have been analysed [6], the semantic five tuple can describe the topics under the time series.

For the process of analysing the explosion mode, the execution process can be viewed a sensor for the social event. The process of analysing the mode quotes the entropy characteristic, and the analytical result is a speculation about the sharpness of the click rate change. For the process of analysing the attention mode, the execution process makes up for the problem existing in the burst mode, that is, when the research instances are followed by web users, the measurement standard will be an absolute phenomenon. Based on the analysis of the usage mode, the click mode of Web news instances can be inferred shown in the formula 3.

$$\begin{aligned} ClickMode(u) = (1 - (-\sum_{i=1}^{n} rt_i(u) \times log_n rt_i(u))) \times \\ \times \frac{log(fq(u)) - Min_{u_i \in U}(log(fq(u_i)))}{Max_{u_i \in U}(log(fq(u_i))) - Min_{u_i \in U}(log(fq(u_i)))} \end{aligned} \tag{3}$$

In the formula 3, n indicates the number of the granular unit that the instances are followed. For the sudden event, it can be set in day. For the normal occurrence event, it can be set in week or month. If the fluctuation is not large for the click rate of Web news instances, then the attention mode is smaller. If Web news instances have obvious fluctuation, then the attention mode will be large. According to the click process of Web news instances, it has the power law distribution characteristic. Therefore, the click frequency of Web news instances has carried on logarithm transformation.

For the process of analysing the degree distribution mode, the degree of Web news instances presents the power law distribution, so its logarithmic transformation can be executed. For the process of analysing the similarity mode, it not only makes up for the problem existing in the degree distribution mode, which ignores the degree origin of Web news instances through retrieval keyword, but also solves the problem of the sparse record in the click behaviour of web user. Based on the analysis of the usage mode, the retrieval mode of Web news instances can be speculated shown in the formula 4.

$$
\begin{aligned}
SearchMode(u) = \frac{2}{n(n+1)} \times \frac{log(d(u)) - Min_{u_i \in U}(log(d(u_i)))}{Max_{u_i \in U}(log(d(u_i))) - Min_{u_i \in U}(log(d(u_i)))} \times \\
\times \sum_{i \le j}^{n} \frac{\sum_{k \in dataitem}^{\infty}(s_{ik}(u)s_{jk}(u))}{\sqrt{\sum_{k \in dataitem}^{\infty}(s_{ik}(u))^2}\sqrt{\sum_{k \in dataitem}^{\infty}(s_{jk}(u))^2}}
\end{aligned}
\tag{4}
$$

Based on the formula 3 and 4, the set of Web news instances can be mined by using the formula 5, $SeedURL(u)$ should be larger than or equal to the seed threshold. As shown in the equation 6, $< t_i, p_i, o_i, ce_i, re_i > (ns_i.uc)$ represents the semantic five tuple description for the seed instance. For the utility evaluation result, the sort sequence of the utility can be executed. In the subsequent experiments, the optimal value of the seed threshold will be analysed.

$$SeedURL(u) = ClickMode(u) \times SearchMode(u) \tag{5}$$

$$
\begin{aligned}
TimeSeries(ns) = (< t_1, p_1, o_1, ce_1, re_1 > (ns_1.uc), ..., \\
< t_i, p_i, o_i, ce_i, re_i > (ns_i.uc), ..., < t_n, p_n, o_n, ce_n, re_n > (ns_n.uc))
\end{aligned}
\tag{6}
$$

Algorithm 1 Method 1 Analysing the Usage Mode

```
Input: UserBehavior, Threshold;
Output: TopicURL;
LET $UserRecord \leftarrow UserBehavior$;
LET $GroupUserRecord \leftarrow GroupByURL(u), TopicURL \leftarrow \phi$;
$For\ each\ gur[i](0 \le i \le gur.size() - 1)\ Do$
  $SeedURL \leftarrow Calculate\ clickmode\ and\ searchmode$;
  $If\ SeedURL \ge Threshold\ Then$
    $tu.add(SeedURL)$;
  $End\ If$
$End\ For$
```

For the construction of the sequence topic, the execution process uses the probability for its first transfer, in order to judge whether it is similar to the seed instance supporting topics by using Web news instances and taking Web news seed instance as the research center. If su indicates the seed instance, then the variable t_u expresses that whether the instance can support the topic of su, the variable t_s expresses that whether the search keyword can support the topic of su. If the seed instance is able to support the topic of su, then $t_u = 1$, conversely, $t_u = 0$, if the search keyword is able to support the topic of su, then $t_s = 1$, conversely, $t_s = 0$. In initial status, t_{su} is one, $P(t_{su} = 1)$ is one, the probability is zero. As shown in the formula 7 and 8, $P(t_s = 1)$ is able to calculate su probability, $P(t_u = 1)$ is also able to be recalculated, when $P(t_u = 1)$ is larger than or equal to the probability threshold, the Web news instances are able to be found, so as to mine the similar instance set. In the subsequent experiments, the optimal value of the probability threshold will be analysed.

$$P(t_s = 1) = \sum_{u:(s,u) \in E}^{\infty} \frac{fq(s,u)}{\sum_{(s,u_i) \in E}^{\infty} fq(s,u_i)} \times P(t_u = 1) \tag{7}$$

$$P(t_u = 1) = \sum_{s:(s,u)\in E}^{\infty} \frac{fq(s,u)}{\sum_{(s_i,u)\in E}^{\infty} fq(s_i,u)} \times P(t_s = 1) \tag{8}$$

Algorithm 2 Method 2 Constructing the Topic Series

```
TopicURL, UserBehavior, Threshold;
TopicURL;
$For\ each\ tu[i](0 \le i \le tu.size() - 1)\ Do$
  $swset1 \leftarrow ExistSet(tu[i].getSet("Topicurl"), ub);$
  $While\ swset1\ is\ not\ null\ Do$
    $While\ each\ swset1\ is\ exist\ Do$
      $p(t_s) \leftarrow CalculateResult(swset1.getElement(j).position, tu[i].getSet("Topicurl"), ub);$
      $If\ p(t_s) \ge Threshold\ Then$
        $swset2.addSet(swset1.getElement(j));$
      $End\ If$
      $ub \leftarrow (swset1.getElement(j).position, p(t_s));$
    $End\ While$
    $While\ each\ swset2\ is\ exist\ Do$
      $wnuset1 \leftarrow ExistSet(swset2.getElement(j).position, ub);$
      $If(wnuset1 \leftarrow EqualSet(wnuset1, wnuset2))\ is\ not\ null\ Then$
        $While\ each\ wnuset1\ is\ exist\ Do$
          $p(t_u) \leftarrow CalculateResult(wnuset1.getElement(k).position,$
                                   $swset2.getElement(j).position, ub);$
          $If\ p(t_u) \ge Threshold\ Then$
            $wnuset2.addSet(wnu1.getElement(k));$
          $End\ If$
          $ub \leftarrow (wnuset1.getElement(k).position, p(t_u));$
        $End\ While$
      $End\ If$
    $End\ While$
    $swset1 \leftarrow ExistSet(wnuset2, swset2, ub);$
  $End\ While$
  $tu[i] \leftarrow wnuset2;$
  $Describe\ Topic\ tu[i];$
$End\ For$
```

4.2 The algorithm of analysing the process for Web hierarchical topic evolution

The algorithm of analysing the process for Web hierarchical topic evolution is implemented by designing two methods of the event series construction and the event evolution analysis, the inputting content of this algorithm is the result of analysing semantic five tuple for Web news instances, the user usage behavior record set and Web news topic set, the outputting content of this algorithm is the topic set that has been excavated under the events and the evolution result of analysing the events belonging to the topics.

According to the set of Web news instances that can support the topics mined, the user usage behavior record is used to analyse the time sequence of Web news instances that are followed, the similarity degree of the core events reported among Web news instances is calculated by using the result of analysing semantic five tuple description, and the evolution state of the events supported by the similar Web news instances and the topics belonging to the events can also be analysed.

The calculation result of the time series similarity can show the topics mined among Web news instances, the user can pay much attention to Web news instances that describe the same events, which have occurred within a certain time period, and the time sequence concerned is similar. On a certain granularity, the vector of the time sequence can be expressed as shown in the formula 9 representing the attention rate of Web news instances. In this formula, $rt_i(u_j)$ represents the attention rate of the instance u_j in i component for a granularity, n indicates the number of the granular units that the instances are continuously followed. For the sudden events, the granularity can be set in days, for the normality events, the granularity can be set in weeks or months.

$$\begin{aligned} TimeSeries(tu_i) = (\{rt_{11}(tu_{11}), ..., rt_{1j}(tu_{1j}), ..., rt_{1m}(tu_{1m})\}, ..., \{rt_{i1}(tu_{i1}),\\ ..., rt_{ij}(tu_{ij}), ..., rt_{im}(tu_{im})\}, ..., \{rt_{n1}(tu_{n1}), ..., rt_{nj}(tu_{nj}), ..., rt_{nm}(tu_{nm})\}) \end{aligned} \tag{9}$$

According to the semantic five tuple description of Web news instances supporting the topics, the core events can be extracted, the similarity degree can be calculated among the core events reported by Web news instance. As shown in the formula 10, $FS(fs_i, fs_j)$ can be calculated with a threshold, which is greater than or equal to the similarity threshold. The semantic five tuple is used to describe the core events reported by the aggregated Web news instance set, the Web news instances that represent the node of each event are arranged in ascending order according to the occurrence time of the core events. If the core events occur in the same time, then Web news instances are arranged in descending order according to its utility characteristic. The Web news instances that are clustered together can be arranged in ascending order according to the occurrence time of the core events, if the core events occur in the same time, then Web news instances are arranged in descending order according to its utility characteristic. $< t_i, p_i, o_i, ce_i, re_i >$ $(< T_i, E_i > .uc)$ expresses the seed topic and the event description, $< t_{ij}, p_{ij}, o_{ij}, ce_{ij}, re_{ij} >$ $(< T_i.uc >, < E_{ij}.uc >)$ expresses the seed topic of the evolution event description. In the subsequent experiments, the optimal range of analysing the parameters and the similarity threshold value will be analysed.

$$\begin{aligned} FS(fs_i, fs_j) = \alpha \times \frac{\sum_{m=1}^{n}(ts_{im}, ts_{jm})}{\sqrt{\sum_{m=1}^{n}(ts_{im})^2}\sqrt{\sum_{m=1}^{n}(ts_{jm})^2}} + \\ +\beta \times \frac{\sum_{m=1}^{n}(ce_{im}, ce_{jm})}{\sqrt{\sum_{m=1}^{n}(ce_{im})^2}\sqrt{\sum_{m=1}^{n}(ce_{jm})^2}} \end{aligned} \tag{10}$$

$$\begin{aligned} TopicURL(T, E) = (\{< t_{11}, T_1, E_{11} >, ..., < t_{1j}, T_1, E_{1j} >, ..., < t_{1m}, T_1, E_{1m} >\}, ..., \\ \{< t_{i1}, T_i, E_{i1} >, ..., < t_{ij}, T_i, E_{ij} >, ..., < t_{im}, T_i, E_{im} >\}, ..., \{< t_{n1}, T_n, E_{n1} >, ..., \\ < t_{nj}, T_n, E_{nj} >, ..., < t_{nm}, T_n, E_{nm} >\}) \end{aligned} \tag{11}$$

5 The experimental analysis and result

In the process of completing the experiments based on designing the algorithms, the experimental environment of the software and hardware is used as follows. Java language is used for the

Algorithm 3 Method 3 Constructing the Event Series

```
Input: TopicURL, UserBehavior, NewsSet, Threshold, Parameters;
Output: TopicURL;
$For\ each\ t[i](0 \leq i \leq t.size() - 1)\ Do$
    $Generate\ timeseriesvector$;
    $Fs \leftarrow Calculate\ similarity\ of\ timeseries\ and\ coreevent$;
    $If\ Fs \geq Threshold\ Then$
        $Adjust\ event\ under\ topic\ t[i]$;
    $End\ If$
    $Describe\ event\ under\ topic\ t[i]$;
$End\ For$
```

programming design to implement the algorithms, MyEclipse platform of the software research and development is used for the framework implementation, SQL Server of the database management system is used for web big data storage and process. The processor is Intel 2.40GHz, the memory is 32GB [**?**, 8, 9, 16, 29]. The experiments mainly use the standard data set for the social event of German A320 airliner crash, the data source is from massive Web news analysis corpus for the real data, the experimental analysis and result can verify feasibility and effectiveness of the research idea.

5.1 The qualitative analysis of the topic detection

As shown in the figure 2, the accuracy rate represents the quality of the topic detection by using three web usage behaviour mining processes. Firstly, the red column represents the accuracy rate of analysing the instance clicking mode, this quality is not high, although it has improvement, but the maximum is able to only arrive on about 0.64. Secondly, the blue column represents the accuracy rate of analysing the instance searching mode, this quality is not also high, although it has also improvement in several monitoring points, but the maximum is able to also only arrive on 0.63. Thirdly, the green column represents the accuracy rate of the algorithm designed in this paper, this quality is significantly improved, and the maximum is able to arrive on about 76

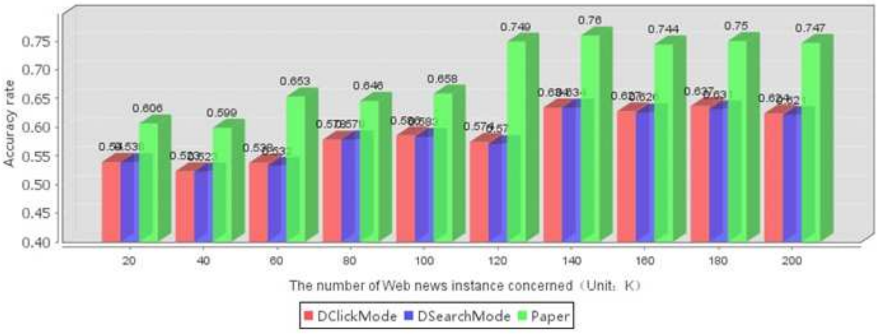

Figure 2: The qualitative analysis of the topic detection

5.2 The analysis of the accuracy rate with the number of the instances concerned and the seed threshold variation trend

As shown in the figure 3, the accuracy rate of the topic detection is indicated through the X and Y axis adjustment. The accuracy rate represents the quality of the topic detection, if the seed threshold is defined, then it is able to increase in a stable trend, because the number of the instances followed is less, the relationship among big data is simpler. If the number of the instances is increasing, then the relationship of the link exists, so the accuracy rate of the topic detection is increasing in a stable trend. If the number of the instances is defined, firstly, then the quality of the topic detection expresses an increasing trend, secondly, then it will decrease with the increasing threshold, because the threshold is less, the topics of inaccurate accuracy are able to be found. If the threshold can arrive at a stable range, then the topics of approximate accurate are able to be found. If the threshold can arrive at a value, then the accurate topics cannot be found. This experiment expresses when the number of the instances is one hundred and sixty, and the seed threshold is zero point seven five, the quality of the topic detection can get the highest about 0.78.

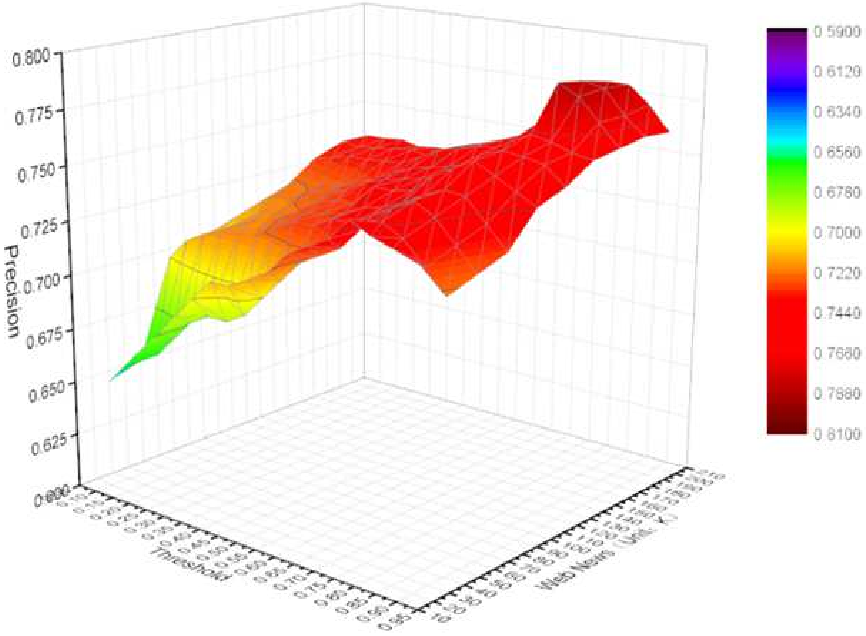

Figure 3: The analysis of the accuracy rate with the number of the instances concerned and the seed threshold variation trend

5.3 The analysis of the accuracy rate with the number of the instances concerned and the probability threshold variation trend

As shown in the figure 4, the accuracy rate represents the quality of the instances mined that support the topics by adjusting the probability threshold of the X axis and the number of the instances of the Y axis. The accuracy rate indicates the quality of the instances mined, if the threshold is defined, then the quality of the instances mined is able to increase in a stable trend, because the number of the instances followed is less, the relationship among big data is simpler. If the number of the instances concerned is increasing, then the relationship of the link exists, so the accuracy rate of the instances mined is increasing in a stable trend. If the number of the instances is defined, firstly, then the quality of the instances mined expresses an increasing trend, secondly, then it will decrease with the increasing threshold, because the threshold is less, the instances of inaccurate accuracy can be mined. If the threshold can arrive at a stable range, then the instances of approximate accurate can be mined. If the threshold can arrive at a value, then the accurate instances cannot be mined. This experiment expresses when the number of the instances followed is one hundred and forty, and the probability threshold is zero point seven, the quality of the instances mined can get the highest about 0.76.

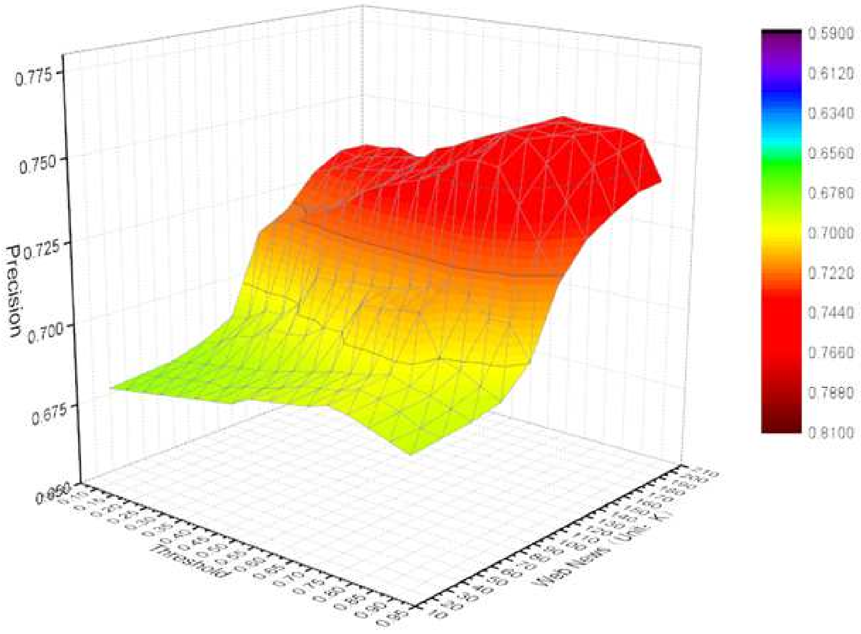

Figure 4: The analysis of the accuracy rate with the number of the instances concerned and the probability threshold variation trend

5.4 The qualitative analysis of the topic evolution

As shown in the figure 5, the accuracy rate represents the quality of the topic evolution by using three web usage behaviour mining processes. Firstly, the red solid line represents the accuracy rate of analysing the time series similarity, which shows that the accuracy rate is not high with the increase in the number of Web news instances concerned, although it has risen, but the highest can only arrive on about 0.64. Secondly, the blue solid line represents the accuracy rate of analysing the core event similarity, which shows that the accuracy rate is not also high with the increase in the number of Web news instances concerned comparing with analysing the time series similarity, and the accuracy rate has also a little decreasing slightly trend, the highest can only arrive on about 0.64. Thirdly, the green solid line represents the accuracy rate of the algorithm designed in this paper, which shows that the accuracy rate has greatly improved because of integrating the time series similarity based on the semantic evaluation and the similarity analysis for the core events. Although the accuracy rate is similar comparing with other two methods under the circumstance of less Web news instances concerned, but the accuracy rate has gradually widening the gap comparing with other two methods with the increase in the number of Web news instances concerned, the highest can arrive on about 0.75. So this experiment expresses that the quality of analysing the topic evolution is higher than other two methods by using the algorithm designed in this paper.

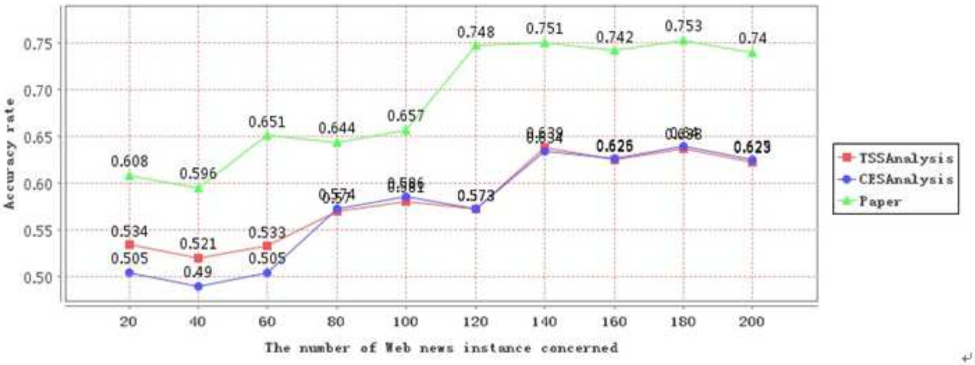

Figure 5: The qualitative analysis of the topic evolution

5.5 The accuracy rate of analysing the topic evolution with the variation trend for the parameter adjustment

As shown in the figure 6, the accuracy rate indicates the quality of analysing the topic evolution for Web news topics according to the parameter adjustment of the sequence event construction process. The red dashed line represents the accuracy rate, in which the Alpha parameter values are different aiming at the formula 10. From its trend, when the Alpha value is adjusted from 0.6 to 0.65, and the Beta value is adjusted from 0.35 to 0.4, the quality of analysing the topic evolution is more high and stable, and the accuracy rate is close to 0.70. In general, the parameter adjustment can make the quality of analysing the topic evolution more stable to the maximum for Web news topics, which accords with the experimental effect expected, and can determine the optimal range of the parameter.

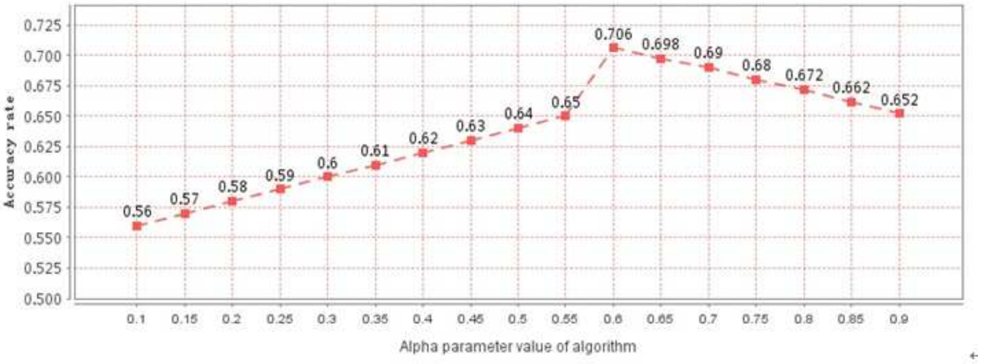

Figure 6: The accuracy rate of analysing the topic evolution with the variation trend for the parameter adjustment

5.6 The accuracy rate analysis with the number of the instances concerned and the similar threshold variation trend

As shown in the figure 7, the accuracy rate represents the quality of analysing the topic evolution by adjusting the similarity threshold of the X axis and the number of the instances of the Y axis. If the threshold is defined, then the quality of analysing the topic evolution is able to increase in a stable trend, because the number of the instances followed is less, the relationship among big data is simpler in the analytical process of the time series and core event similarity.

If the number of the instances followed is gradually increasing, then the relationship among big data adds also the semantic feature for analysing the process of the topic evolution, so its accuracy rate can increase. If the number of the instances is defined, firstly, then the quality of analysing the topic evolution can increase, secondly, then it will decrease with the increasing threshold, because the threshold is less, the inaccurate or approximate accurate analysis of the topic evolution mays be completed. If the threshold can increase to a stable range, then the approximate accurate result of analysing the topic evolution can be excavated. If the threshold can increase to a value, then the accurate result of analysing the topic evolution cannot be excavated. This experiment expresses when the number of the instances followed is one hundred and eighty, and the similarity threshold is zero point seven, the quality of analysing the topic evolution can get the highest about 0.76.

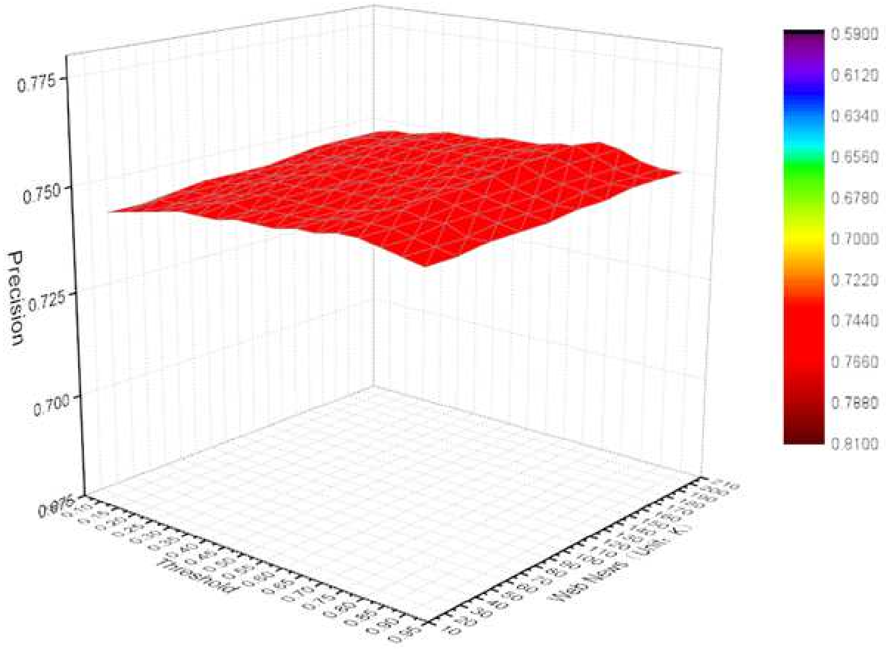

Figure 7: The accuracy rate analysis with the number of the instances concerned and the similar threshold variation trend

5.7 The time consuming analysis for solving main research problem under different methods

As shown in the figure 8, in view of German A320 airliner crash social event, the X axis represents massive Web news time released through the authoritative Web news network platform, the red solid line represents the time consuming for analysing Web hierarchical topic evolution under the method based on the content and description, the blue solid line represents the time consuming for analysing Web hierarchical topic evolution under the method based on the behaviour tracking. According to the change trend of two solid lines, the number of Web news instances has increased sharply in several time intervals with the progress of the event development, therefore, the time consuming has also increased sharply, in other time intervals, the time consuming has relatively stable trend. While the time consuming is lower for the blue solid line, because the analytical process of Web hierarchical topic detection and evolution uses the behaviour tracking method based on the semantic five tuple description and the utility evaluation. This experiment expresses that the non-deterministic problem can be solved efficiently using the method, which is proposed by this paper.

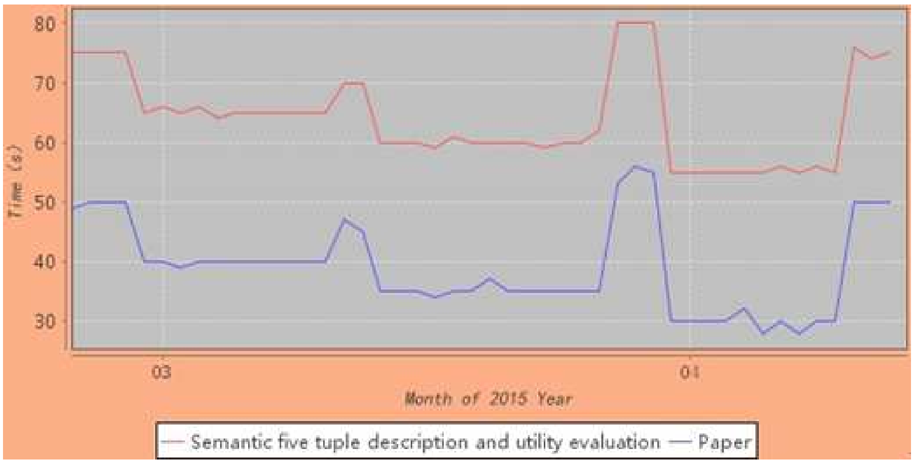

Figure 8: The time consuming analysis for solving main research problem under different methods

5.8 The qualitative analysis of Web hierarchical topic detection and evolution result under different data sets

As shown in the figure 9, the qualitative analysis of Web hierarchical topic detection and evolution result uses also other four standard data sets in addition to the data set of German A320 airliner crash social event, which include Shanghai Bund trample, Taiwan revival airliner falling river, Nepal 8.1 earthquake and Orient Star cruise overturn social event. According to the change trend of five columns, there is little difference in the qualitative analysis of Web hierarchical topic detection and evolution result at the start, development and end stage of five social events. This experiment shows that the analytical process of Web hierarchical topic detection and evolution is stable under different social events. Moreover, the qualitative analysis of Web hierarchical topic detection and evolution result has only little effect in different stages of the social events with the increase of the number of Web news instances.

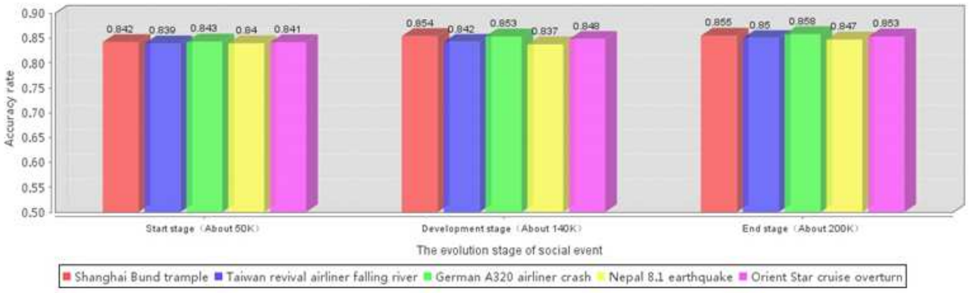

Figure 9: The qualitative analysis of Web hierarchical topic detection and evolution result under different data sets

6 Conclusion

This paper completes the research on an idea of analysing the process for Web hierarchical topic detection and evolution in view of the behaviour tracking technology taking the network big data of Web news as the processing object, this result of designing and implement is more valuable for the scholars in the research field. In the research process, this paper proposes the analytical algorithm of Web hierarchical topic detection and evolution, in which this paper has also proposed the methods of analysing the usage mode, the topic series construction, the event series construction and the evolution analysis for the events, so as to solve the problem existing in current research status. The results of experimental analysis show that the idea is feasible, verifiable and superior, which plays a major role in reconfiguring Web hierarchical topic corpus, improves understanding efficiency of the network big data, enhances the website availability, constructs and improves the website service function, improves the efficiency of the business operational and website clicking rate, provides an intelligent big data warehouse for the network information evolution application.

Funding

This paper is supported by National Natural Science Foundation of China under Grant Nos.71572015, Support Project of High-Level Teachers in Beijing Municipal Universities in the Period of 13th Five-Year Plan under Grant Nos.CIT&TCD201704072, Premium Funding Project for Academic Human Resources Development in Beijing Union University under Grant Nos.BPHR2018AS01.

Bibliography

[1] Ahila, S.S.; Shunmuganathan, K.L. (2016). Role of Agent Technology in Web Usage Mining: Homomorphic Encryption Based Recommendation for E–commerce Applications, *Wireless Personal Communications*, 87(2), 499-512, 2016.

[2] Alam, M.H.; Ryu, W.J.; Lee, S. (2017). Hashtag-Based Topic Evolution in Social Media, *World Wide Web-Internet and Web Information Systems*, 20(6), 1527-1549, 2017.

[3] Aujla, G.S.; Kumar, N.; Zomaya, A.Y. (2018). Optimal Decision Making for Big Data Processing at Edge-Cloud Environment: An SDN Perspective, *IEEE Transactions on Industrial Informatics*, 14(2), 778–782, 2018.

[4] Chen, B.T.; Tsutsui, S.; Ding, Y.; Ma, F.C. (2017). Understanding the Topic Evolution in a Scientific Domain: an Exploratory Study for the Field of Information Retrieval, *Journal of Informetrics*, 11(4), 1175-1189, 2017.

[5] Chen, M.; Yang, X.P. (2016). Research on Model of Network Information Extraction Based on Improved Topic-Focused Web Crawler Key Technology, *Tehnicki vjesnik/Technical Gazette*, 23(4), 49–54, 2016.

[6] Chen, M.; Yang, X.P.; Sun, M.; Zhao, Y. (2014). Research on Model of Network Information Currency Evaluation Based on Web Semantic Extraction Method, *International Journal of Future Generation Communication and Networking*, 7(2), 103-116, 2014.

[7] Chen, Y.; Zhang, H.; Liu, R.; Ye, Z.W.; Lin, J.Y. (2019). Experimental Explorations on Short Text Topic Mining Between LDA and NMF Based Schemes, *Knowledge-Based Systems*, 163, 1–3, 2019.

[8] Dai, Y.; Wu, W.; Zhou, H.B.; Zhang, J.; Ma, F.Y. (2018). Numerical Simulation and Optimization of Oil Jet Lubrication for Rotorcraft Meshing Gears, *International Journal of Simulation Modelling*, 17(2), 318-326, 2018.

[9] Dai, Y.; Zhu, X.; Zhou, H.; Mao, Z.; Wu, W. (2018). Trajectory Tracking Control for Seafloor Tracked Vehicle by Adaptive Neural-Fuzzy Inference System Algorithm, *International Journal of Computers Communications & Control*, 13(4), 465-476, 2018.

[10] Du, J.; Sun, Y.; Ren, H. (2018). The Relationship of Delivery Frequency with the Cost and Resource Operational Efficiency: A Case Study of Jingdong Logistics, *Mathematics and Computer Science*, 3(6), 129-140, 2018.

[11] Fatima, B.; Ramzan, H.; Asghar, S. (2016). Session Identification Techniques Used in Web Usage Mining a Systematic Mapping of Scholarly Literature, *Online Information Review*, 40(7), 1033-1053, 2016.

[12] Gaul, W.G.; Vincent, D. (2017). Evaluation of the Evolution of Relationships between Topics over Time, *Advances in Data Analysis and Classification*, 11(1), 159-178, 2017.

[13] Jimenez-Marquez, J.L.; Gonzalez-Carrasco, I.; Lopez-Cuadrado, J.L.; Ruiz-Mezcua, B. (2019). Towards a Big Data Framework for Analysing Social Media Content, *International Journal of Information Management*, 44, 1–3, 2019.

[14] Kaseb, M.R.; Khafagy, M.H.; Ali, I.A.; Saad, E.M. (2019). An Improved Technique for Increasing Availability in Big Data Replication, *Future Generation Computer Systems-The International Journal of Escience*, 91, 493–497, 2019.

[15] Kausel, E.E. (2018). Big Data at Work: The Data Science Revolution and Organizational Psychology, *Personnel Psychology*, 71(1), 135-136, 2018.

[16] Kho, N.D. (2018). The State of Big Data, *Econtent*, 41(1), 11-12, 2018.

[17] Liu, J.; Fang, C.; Ansari, N. (2016). Request Dependency Graph: a Model for Web Usage Mining in Large-Scale Web of Things, *IEEE Internet of Things Journal*, 3(4), 598-608, 2016.

[18] Makkie, M.; Huang, H.; Zhao, Y.; Vasilakos, A.V.; Liu, T.M. (2019). Fast and Scalable Distributed Deep Convolutional Autoencoder for fMRI Big Data Analytics, *Neurocomputing*, 325, 20–22, 2019.

[19] Osman, A.M.S. (2019). A Novel Big Data Analytics Framework for Smart Cities, *Future Generation Computer Systems-The International Journal of Escience*, 91, 620–623, 2019.

[20] O'Halloran, K.L.; Tan, S.; Duc-Son, P. (2018). A Digital Mixed Methods Research Design: Integrating Multimodal Analysis with Data Mining and Information Visualization for Big Data Analytics, *Journal of Mixed Methods Research*, 12(1), 11-15, 2018.

[21] Pandian, P.S.; Srinivasan, S. (2016). A Unified Model for Preprocessing and Clustering Technique for Web Usage Mining, *Journal of Multiple-Valued Logic and Soft Computing*, 26(3), 205-220, 2016.

[22] Sagi, T.; Gal, A. (2018). Non-Binary Evaluation Measures for Big Data Integration, *VLDB Journal*, 27(1), 105-110, 2018.

[23] Tran, Q.T.; Nguyen, S.D.; Seo, T.I. (2019). Algorithm for Estimating Online Bearing Fault Upon the Ability to Extract Meaningful Information From Big Data of Intelligent Structures, *IEEE Transactions on Industrial Electronics*, 66(5), 3804–3806, 2019.

[24] Uma, R.; Muneeswaran, K. (2017). OMIR: Ontology-Based Multimedia Information Retrieval System for Web Usage Mining, *Cybernetics and Systems*, 48(4), 393-414, 2017.

[25] Wu, P.J.; Lin, K.C. (2018); Unstructured Big Data Analytics for Retrieving E-Commerce Logistics Knowledge, *Telematics and Informatics*, 35(1), 237-241, 2018.

[26] Yao, L.; Ge, Z.Q. (2019). Scalable Semisupervised GMM for Big Data Quality Prediction in Multimode Processes, *IEEE Transactions on Industrial Electronics*, 66(5), 3681–3684, 2019.

[27] Zhang, D. (2017). High-Speed Train Control System Big Data Analysis Based on Fuzzy RDF Model and Uncertain Reasoning, *International Journal of Computers Communications & Control*, 12(4), 11-15, 2017.

[28] Zhang, D.; Sui, J.; Gong, Y. (2017). Large Scale Software Test Data Generation Based on Collective Constraint and Weighted Combination Method, *Tehnicki Vjesnik*, 24(4), 1041-1050, 2017.

[29] Zhang, D.; Jin, D.; Gong, Y. (2015). Research of Alarm Correlations Based on Static Defect Detection, *Tehnicki vjesnik*, 22(2), 311-318, 2015.

[30] Zhou, H.K.; Yu, H.M.; Hu, R. (2017). Topic Discovery and Evolution in Scientific Literature Based on Content and Citations, *Frontiers of Information Technology & Electronic Engineering*, 18(10), 1511-1524, 2017.

[31] Zhou, H.K.; Yu, H.M.; Hu, R. (2017). Topic Evolution Based on the Probabilistic Topic Model: a Review, *Frontiers of Computer Science*, 11(5), 786-802, 2017.