International Journal of Future Generation Communication and Networking
Vol. 9, No. 11 (2016), pp. 81-94
http://dx.doi.org/10.14257/ijfgcn.2016.9.11.08

A Novel Deep Web Data Mining Algorithm based on Multi-Agent Information System and Collaborative Correlation Rule

Hongpu Sun and Qianru Hu

Hebei University, Baoding, Hebei, 071002, China

Abstract

Along with the rapid advancement of Internet technology and machine learning science, the data mining techniques have been widely applied on the web page information pattern analysis issues. To enhance the traditional mining algorithms theoretically and numerically, we propose the novel deep web data mining algorithm based on multi-agent information system and collaborative correlation rule in this manuscript. Firstly, we review the latest web mining methodologies to serve as the comparison objects. Then, we introduce the revised agent based algorithm. MAS consists of more than one agent, MAS using parallel distributed processing technology and modular design thought and the complex system is divided into relatively independent agent subsystem. Later, we combine the AdaBoost method to propose the collaborative correlation rule. As the combination, we use the mentioned two techniques to form the optimized and enhanced deep web data mining algorithm with the implementation of programming languages. The experimental result proves the feasibility of our approach and compared with other contemporary state-of-the-art algorithms, our method outperforms and achieves better accuracy with low time-consuming.

Keywords: *Multi-Agent System, Data Mining, Deep Web, Collaborative Correlation Rule, Web Crawler, Information System, Feature Selection*

1. Introduction

Data mining, as the name suggests is from the large, incomplete, noisy, fuzzy data in the process of digging out useful information and knowledge. The information and knowledge is implicit, previously unknown and have potential value for decision-making. Along with the computer technology and information technology application is more and more widely, the enterprise every year to accumulate a large amount of data, using data mining technology in such a large amount of data we can find out the valuable knowledge, rules, or a high level of information provide the basis for the decision-making, so that the data warehouse to become a rich resources and reliable service for enterprise decision makers. According to the literature review, the data mining task could be divided into the following steps. (1) To understand and define problems. Data mining is not a general analysis of the process, and is not a simple data mining algorithm is applied to the database and then gets some. A question may have multiple solutions at the same time, but also consider efficiency and effectiveness. (2) Data collection and extraction. Collection and extraction of the related data from a database that are generally according to the algorithm and the user of proposed mission, determine the attribute domain of interest, by using the statistical methods such as sampling, the operation of a variety of data collection. (3) The purification of the data and data interpretation. As a result of the input data where there may be mistakes and inaccurate, and data mining is to discover internal relation of data, which requires data cleansing and understanding. (4) Data engine. Data mining is an inner link, looking for data search process, different solutions need to form different dynamic database search process, in the face of the large amount of data to data engine to control and optimize the data flow. (5) Planning

ISSN: 2233-7857 IJFGCN

algorithm. Dig algorithm will affect the quality of mining mode, but how to choose the optimal algorithm has yet to form a forming theory, it is more necessary to find effective algorithms. (6) The results of preliminary evaluation. Model is used to assess the mining method depends on how good the practical problems to be solved, given only the accuracy of the patterns is meaningless as the real test can only be in the actual application, some actual detection is not possible or very difficult, so relevant experts in the field of evaluation has become one of the an important result evaluation method. (7) To refine the data and problems. Data mining is an iterative process as from simple to complex process that requires repeated data elaboration until the satisfied result achieved [1-4].

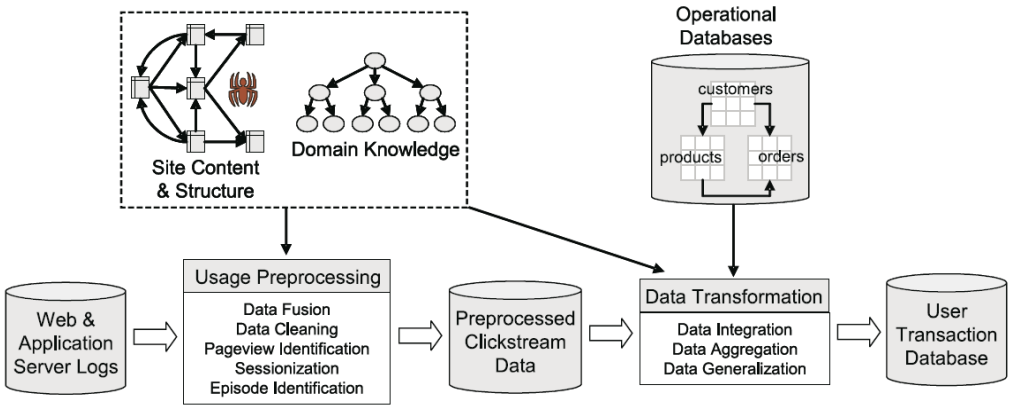

Figure 1. The Data Mining Steps and Procedures

The figure one shows the basic steps of the data mining technology mentioned. However, according to the literature review, there are still some drawbacks and challenges needed for better solution. (1) Because of the relationship between database and data warehouse has been widely used, for them to develop effective data mining system is very important. However, the other databases may contain complicated data object, hypertext and multimedia data, data space, time or transaction data. (2) Many large capacity of database, data widely distributed and some data mining algorithm computational complexity is factor in the development of the parallel and the distributed data mining algorithm. (3) Multiple abstraction layer interactive knowledge mining: because it is difficult to know exactly what can be found in the database, data mining process should be reciprocal. In the face of so much data, the existing statistical methods and so on all had a problem of the low accurate rate [5].

In this research, we propose the novel deep web data mining algorithm based on the multi-agent information system and the collaborative correlation rule. The reminding of the paper follows the structure as listed. We firstly review the contemporary web data mining methods with the comparison of the general performance. Then, we introduce the primary features and the theoretical analysis of the multi-agent information system and collaborative correlation rule, respectively. Later, through combining the discussed modification, we propose the novel web data mining algorithm. To verify the feasibility of our method, we simulate the method with discussion and prospect in the final section.

2. The Review of the Web Data Mining Methods

Various forms of documentation and users to access information on the Web is constitutes the object of the Web data mining. Web includes three types of data: the Web page data, the structure of the Web and Web log files. The corresponding mining algorithms could also be separated into the categories. (1) Web text mining. The object of the Web content mining is made up of unstructured data semi-structured data and

structured data. Unstructured text data and constitute the main components of the Web. (2) Web multimedia mining. Media data mining is an important area of data mining, multimedia data correlation, or other not directly stored in multimedia database model. Generalized multimedia data mining is the mining of the knowledge description of image, video, including the text data mining. (3) Web structure mining. Web structure mining is the basic idea of the Web as a directed graph, the vertex is a Web page, the page is hyperlinked diagram between the edges as then using graph theory the topology of the Web is analyzed. The following figure 2 shows the existing Web data mining methods and its primary characteristics [6-11].

Level	User intent classification													
Level 01	Informational						Navigational		Transactional					
Level 02	Directed		Undirected	Find	List	Advice	Navigation to trans-actional	Navigation to inform-ational	Obtain	Download		Search engine results page		Interact
Level 03	Closed	Open						Online	Off-line	Free	Not free	Links	Other	
Prior studies							*Corresponding labels*							
Carmel et al. (1992)								Browsing (search-oriented, review, scan)						
Navarro-Prieto, Scaife, Rogers (1999)	Fact finding			Exploratory										
Choo and Turnbull (2000)	Formal search	Informal search	Monitoring					Undirected viewing						
Morrison et al. (2001)	Find		Explore monitoring	Collect										
Rozanski et al. (2001)	Single mission do it again quickies		Information please loitering	Just the facts		Quickies		Surfing						
Sellen et al. (2002)	Finding			Information gathering				Browsing				Transacting		
Broder (2002)			Informational					Navigational				Transactional		
Bodoff (2004)								Browsing (navigating, current awareness, undirected, scanning)						
Rose and Levinson (2004)	Informational directed closed	Inform-ational directed open	Inform-ational undirected	Inform-ational locate	Inform-ational list	Inform-ational advice		Navigational	Resource obtain		Resource download		Resource interact	
Teevan et al. (2004)			Orienteering					Teleporting						
Kellar et al. (2007)	Fact finding (looking for specific information)		Information gathering					Browsing				Transactions		
	Fact finding (monitoring)													

Figure 2. The Existing Web Data Mining Methods and Its Characteristics

3. The Multi-Agent Information System

3.1. The Basic Knowledge of Multi-Agent System

Task allocation problem first appeared in all kinds of production, planning, and flexible manufacturing system is the typical combinatorial optimization problem. Because traditional the task allocation problem of goal is to make full use of distributed multiprocessor system parallelism, so a problem is decomposed into more child the mission, through the concurrent execution of subtasks makes problems can in the shortest possible time to complete. In this kind of system of task allocation is mostly based on the following set: system all processor can handle all indifferently tasks, namely for the same task, all processor should return the same result, therefore, same task only assigned to a practitioner. Compared with traditional task allocation problem, the task allocation of multi-agent system has its particularity, mainly reflected in two aspects. (1) In MAS to allow the Agent to accept multiple tasks at the same time, forming the corresponding task list, so as to make full use of the MAS parallel solving ability. (2) Use different method to design agent on the same issue of possible inconsistencies, therefore, to improve the processing precision of the tasks, a task assigned to more than one agent in the MAS parallel solution [12].

MAS besides can integrate the data in the data source it can also operate indirect business operation system of data in data source. MAS consists of more than one agent, MAS using the parallel distributed processing technology and modular design thought, the complex system is divided into relatively independent agent subsystem, done through the basic cooperation and competition between agents for complex problem solving. MAS of agent of the members of the organization structure provide the interaction framework, to provide the overall view and related information for solving problems, and reasonable distribution of tasks. The members of the agent are autonomous as they may be using different design methods and programming language development. They perform their duties in MAS can communicate with each other at the same time to obtain information, work together to complete the relatively complex task. The figure 3 shows the classical scripting programming deployment of the MAS.

```
N-MAS_Prog      =  "Agents: "    (<agentName> <agentProg> [<nr>])+ ;
                   "Facts: "                 <bruteFacts>
                   "Effects: "               <effects>
                   "Counts-as rules: "       <counts-as>
                   "Sanction rules: "        <sanctions>;
<bruteFacts>    =  <b-literals>;
<effects>       =  ("{"<b-literals>"}" <actionName> "{"<b-literals>"}")+;
<counts-as>     =  ( <literals> "⇒" <i-literals> )+;
<sanctions>     =  ( <i-literals> "⇒" <b-literals> )+;
<agentName>     =  <ident> ;
<agentProg>     =  <ident> ;
<nr>            =  <int> ;
<actionName>    =  <ident> ;
<b-literals>    =  <b-literal> {"," <b-literal>} ;
<i-literals>    =  <i-literal> {"," <i-literal>} ;
<literals>      =  <literal> {"," <literal>} ;
<literal>       =  <b-literal> | <i-literal> ;
<b-literal>     =  <b-atom> | "not" <b-atom> ;
<i-literal>     =  <i-atom> | "not" <i-atom> ;
```

Figure 3. The Classical Scripting Deployment of Multi-Agent Information System

3.2. The Enhanced Multi-Agent Information System

In the field of the MAS consultation, the thought method and technology in the field of the other disciplines are vital to the development of the field. Although these fields are not able to directly solve the problem of the AI, but what they consider a wide range of the research questions, provided by the technical method for the design of MAS often played an important role. Game theory is the study of confrontation or mathematical theories and methods of the competitive nature of the phenomenon, which has become an important mathematical tool dealing with MAS negotiation. Game theory provides a rich model for DAI problems, and to evolve to a lot of new research methods [13-14].

Reinforcement learning from the animal learning parameter perturbation adaptive control theory and its basic principle is: if instance of agent behavior strategies lead to environment is reward, so this behavior strategy after agent trend will be strengthened. To enhance the basic performance, the problem can be abstracted as the formula one.

$$V(s) = \max Q(s,a) \tag{1}$$

Where the $Q(s,a)$ represents the prediction value, in the MAS negotiation mechanism, the alliance is a set of equal, collaborative, shared some set of task agent and it divided according to certain strategy to complete the task. In order to make the system achieve global optimal, task allocation process need to exhaust all the possible alliance as system to list all possible alliance first, and then to tentative assignment of each task, finally get the global optimal solution. The formula 2 describes the mentioned procedure.

$$Q(s,a) = (1-\delta)Q(s,a) + \delta(r + \gamma V(s')) \tag{2}$$

From the point of view of machine learning will strengthen learning technology is applied to the MAS methods fall into the two categories: one is the MAS system as a computable learning agent, the other is the reinforcement learning mechanism of each agent has its own, through interaction with other agent suitable to speed up the learning process. Most of the MAS reinforcement learning research belongs to the latter. To change the condition and for the global optimal strategy, we optimize the $V(s)$ as the follows [15].

$$V(s) = \max_{a \in A} \min_{o \in O} Q(s, a, o) \tag{3}$$

Further, when the introduction of communication mechanism in the process of negotiation, the parties can cooperation by all parties can accept the solution, but the premise need each agent promises or establish mutual trust relationship between the individuals. In the use of traditional game theory dealing with MAS negotiation, usually in a static environment is discussed, the model is applied to the dynamic environment of DAI will be larger. As a result, needs further discussion on the basis of the static model of the consultation process dynamic circumstances which are listed as the following aspects.

- An alliance allows for the multiple tasks at the same time, the former is called cross alliance, which is called the multitasking. Complex alliance means cross unions and the multitasking are allowed to exist.
- For the task sequence can, in turn, generate the global optimal coalition, but still only consider an agent can only join a union and an alliance can only take on a task as is bound to cause agent ability and the waste of resources.
- Many particle swarm collaborative optimization of complex union serial generation algorithm, in an attempt to generate the complex alliance through the complete tasks at the same time to maximize total revenue of the system.
- To face the value consultation in the field of the model, by defining a state about the environment may be the value of a function to determine the goals of the agent, the goal of this agent implicit function values achieve maximum state of environment.
- All the agents are not on the logical operations at the same time, but there is order and to get other agent has completed strategy choice.

In particle swarm, only send messages to other particles global extreme value, this is a one-way flow of information, the whole process of searching for updates is to follow the current process of the optimal solution rapidly to the point of basic convergence, the convergence of particle swarm could easily lead to rapid convergence to the local optimal domain and cause premature convergence. The formula 4 defines the optimized form.

$$x_i = \langle \chi_1, ..., \chi_n \rangle \to C_i \tag{4}$$

Take information positive feedback method to improve the global search ability of particle swarm, the main idea is: subgroup searches independently, each subgroup according to their own search to date to the most advantage of the fixed group of speed and position of particle and compare the most advantages, find out the global optimal particle, the "law of the jungle, the survival of the fittest" exit strategy, the global optimal particle to replace each subgroup of the worst particles, but each subgroup own global extrema. With each iteration of the global optimal particle as guidance information affect the evolution of each subgroup, but does not alter behavior of each subgroup blindly direction. The introduction of population information positive feedback and is, in fact, by the independent groups coding mode of restructuring, increase the number of the corresponding coding mode, so as to avoid the core occurrence of the premature convergence. The figure four demonstrates the modification.

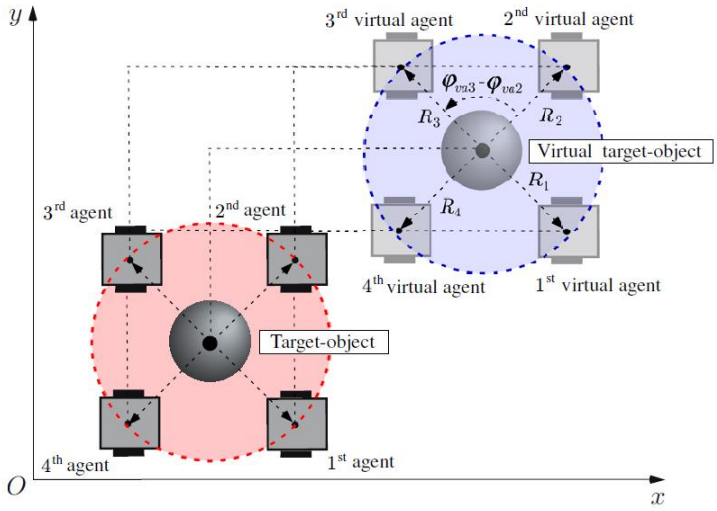

Figure 4. The Agent Topology Distribution for the Overall Performance Enhancement

4. The Optimized Collaborative Correlation Rule

4.1. The Cooperative Association Analysis

Correlation method based on machine learning is a kind of the real-time event correlation method. The basic idea is: using machine learning method for security event data correlation analysis, and generate the corresponding sequence. Collaborative forensic computing refers to from target system found in all available resources, and related information, interpretation, analysis with the evidence to determine the causation.

Faced with record structure and the storage methods of the different heterogeneous log data sources, the first thing to do is data cleaning. Main work includes: in the process of data cleaning to establish a unified time format and complete time synchronization, dealing with invalid values and the missing value, extract the needed fields, finally in the unified database, the formation of the original data source.

After completion of the data cleaning data network security can on demand the analysis but the structure of the logging, caused huge obstacle to multi-source collaborative analysis, data fusion is imperative. Information visualization and the visual analysis technology is a new research field of multidisciplinary integration, it will be the massive high-dimensional data in graphics way express, and provide the effective means of interaction and improve people's cognitive ability, to quickly find data in the implicit rules, patterns and trends. Visualization technology through a graphical mapping and interactive methods, help analysts perceive the information in the network security data more efficiently, more quickly identify abnormal events and the general attack characteristics [16].

Based on a variety of heterogeneous data in network security feature level fusion, realized from the big data to small data conversion and also established the basis of analyzing multi-source data together. Network security incidents and the characteristics of network security time-series data two unified format level data will be used as input data source of visual analysis tools of network security, so that we can better play to the role of the visual analysis technology to help users efficiently grasp the overall network security status.

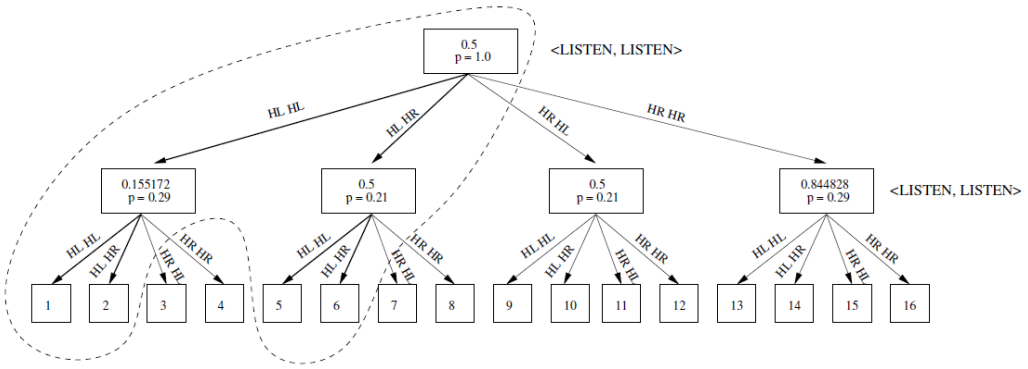

Figure 5. The Cooperative Association for the Data Sets

4.2. The AdaBoost Assisted Collaborative Correlation Rule

AdaBoost algorithm according to different training sets the same weak classifier, then put these on the different training set classifier together, constitutes a stronger the final classifier. Theory has been proved that as long as each weak classifier classification abilities is better than random guesses, when its number tends to infinity, the strong classifier error rate will tend to zero. The algorithm itself by changing data distribution, every time it according to the training set of classification of each sample is correct, the overall classification accuracy as well as last time to determine the weights of each sample. For multiple target classification, common methods are not level and level two ways. The hierarchy method in composed of the target expression feature space, with linear or nonlinear function split target set.

AdaBoost algorithm adjusted Boosting algorithm, ability to weak classifier error obtained by learning adaptability adjustment: use AdaBoost classifier can eliminate some unnecessary training data characteristics, and focus on the key training data, and the greater the difference between each weak classifier as the integrated strong classifier better generalization ability. The formula 5~6 show the principles of the technique [17-19].

$$K\left(x_i, x_j\right) = \sum_{i=1}^{R} \alpha_i \chi_i \left(x_i, x_j\right) \tag{5}$$

$$\chi_i \left(x_i, x_j\right) = \exp\left(-\frac{1}{2\varepsilon_l^2} \left\| x_{i,j} - x_{j,l} \right\|^2 \right) \tag{6}$$

Of single classifier AdaBoost algorithm, the consolidation process is particularly important is the weight of each member classifier has to combine. It is represented as the component classifier candidates to probability that the training set. Assumptions, a sample points have been classified accurately, is under constructing a training set, it reduce selected probability. On the contrary, if a sample point is not correct classification, the weight of it can improve accordingly. The calculation step could be reflected from the formula 7.

$$f\left(x\right) = \sum_{t=1}^{T} \alpha_t h_t \left(x\right) \tag{7}$$

Based on similarity and algorithms in the field of the application of the correlation of the consideration here only the traditional algorithm and the template matching calculate methods comparing with evolutionary algorithm. Traditional algorithm has high recognition rate and recognition rate, but the vast amounts of weak feature selection process to make training takes seriously, the algorithm converges slowly. In the evolutionary algorithm made up of positive and negative samples to generate positive and negative weighting template to each level of the weak classifier, and according to the weak classification error rate adjust sample weight distribution to construct new weak classifier, until get a satisfactory strong classifier.

Integration of the multiple classifiers is learning to solve the problem of the same training, when to deal with the new data, the conclusion of each classifier to synthetically in some way. This method can overcome the classifier of the training set the fitting problem, improve the generalization ability, and as good as possible to deal with new data. Center method based on the following assumptions: each sample and it is the center of the vector similarity than with the rest of the class at the center of the vector similarity, that is to say, think center method, the center of the class vector can be used as a representative of the vector and the prototype and the similarity threshold value of the each class are the same. Due to all kinds of data distribution in shape, density and other differences, which often do not agree with the actual data distribution, the basic hypothesis of many data distribution does not comply with the method of center to center deviation method of classification. In the following figure 6, the comparison of the AdaBoost and other techniques is demonstrated [20].

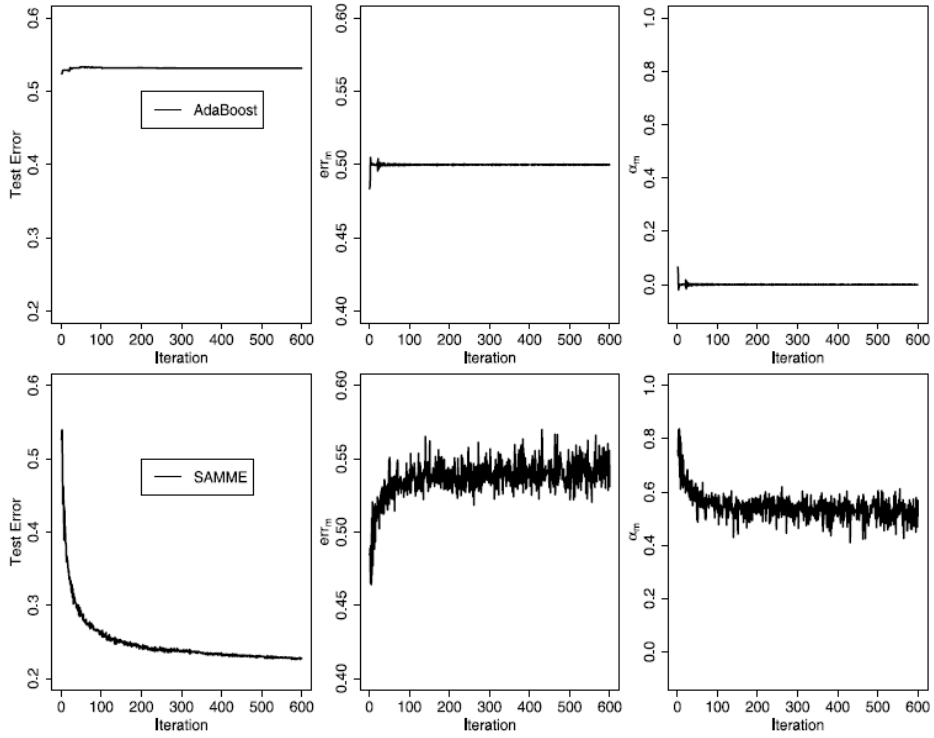

Figure6. The Robustness and Stability of the AdaBoost Demonstration

5. The Novel Web Data Mining Algorithm

Web data mining is the unknown, hidden dug up from a Web page, the decision-making process of the potential knowledge or rules. It is to put the data mining technology for the resources on the Internet which is a kind of practical new technology. To achieve the better and higher feasibility of the mining algorithm, we should consider the following conditions.

- Web content mining. The element object with both text and more than text data, such as graphics, images, multimedia data. Both of the structured data from the database, and useful HTML or XML tags semi-structured data and no free text structure. Web content mining can help users to search information can filter the useless information according to the users' search terms.
- Web use mining. Records from the server Web use mining user access logs, or from the user's browsing information extracting interesting knowledge model,

through the analysis of these data can help us understand the user behavior patterns, hidden in the data to make a predictive analysis.

- Web structure mining. From the organizational structure of the Web and basic Web structure mining process of link relation is interesting knowledge. Mining structure and Web structure of the page, which can be used to guide the page classification and clustering to find authoritative page, thus improve the retrieval performance.

Deep of Web page classification is not only to examine the content of the page and their factors, such as the structure of the semantic and shall be provided according to the user's cognitive ability and the personality characteristics at a deeper level classification basis and classification results, users in the shortest time accurately for the interest and attention of the information on the Web. As shown below, we demonstrate the systematic structure [21].

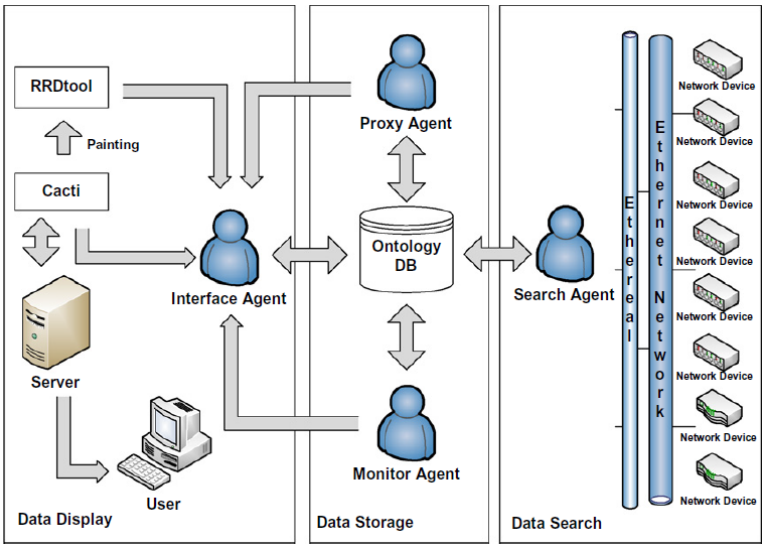

Figure 7. The Web Data Pattern Mining System Architecture

Users browse the web, with some kind of information needs, and then follow the hyperlink to guide, reading a web page. This information needs to interest in some concept. If we can from the user browsing path dig up the implicit information in the analysis of demand, we can than dig up the user's interest in the end can be modified to fit the page with the user's site structure. Markov prediction method is a method to predict the occurrence probability of the event and it is based on the Markov model, according to the events of the current status to predict the future every moment or period of a prediction method as formula 8.

$$P\left(q_t = s_j \mid q_{t-1} = s_i, q_{t-2} = s_k, \cdots\right) \tag{8}$$

Hidden Markov model is a dual random process, with state and the observation sequences. Concrete cannot be observed, the state of the sequence of state transition probability, that is, the state of the model conversion process is hidden, observable events of stochastic process is a random function of the hidden state transition process. User access path through the above analysis can use the HMM model. The concept of the web page contains as observation status symbol set, concept model of the observed sequence into the output sequence, the user access path via the web set as the implicit state set. Calculate the output probability concept output sequence, the greater the output probability, show that the concept of the more likely is the needs of users in the access path concept, contains the concept of large amount of information of web pages that will be on the site layout the position of the obvious [22].

```
<owl:Class rdf:ID="ContextExample"
 <owl:UnionOf rdf:parsetype="Collection">
  <owl:Class rdf:resource="#IdentityContextGreenPatient"/>
  <owl:Class rdf:resource="#LocationContextGreen"/>
  <owl:Class rdf:resource="#TimeContext1418"/>
 </owl:UnionOf>
</owl:Class>
<owl:Class rdf:ID="MusicServiceAction">
 <rdfs:subClassOf rdf:resource="& action;AccessAction"/>
 <rdfs:subClassOf>
   <owl:Restriction>
     <owl:onProperty rdf:resource="&action;accessedEntity"/>
     <owl:allValuesFrom rdf:resource="#MusicService"/>
   </owl:Restriction>
 </rdfs:subClassOf>
</owl:Class>
<policy:Authorization rdf:ID="AuthC271">
 <policy:activatedBy rdf:resource="#ContextExample"/>
 <policy:controls rdf:resource="#MusicServiceAction"/>
 ...
</policy:Authorization>
```

Figure 8. The Core Scripting Code Snippet for User Data Mining

As demonstrated in the figure 8, we show the core scripting code snippet for the user data mining. Although the users browsing process in the Web space is a browse purpose, cultural background, hobbies, and other factors influence the complex process, there are a lot of the differences, observation of a large number of users browsing process can be found, however, some users browsing process showed the same or similar characteristics, such as they browse the Web pages are basically the same, the order of the various Web pages are similar, this phenomenon caused the study of classification of the Web users. Users through the user classification, the same category with the same model to describe it, and a greater difference between the different categories of users browsing process, describe their characteristics in different model is more reasonable. In the formula 9, we define the feasibility measurement tool for the final evaluation.

$$W_{p_i(r_j)} = \left(f_{ji} \,/\, f_{\max} \right) / \sum_{k=1}^{m} \left(\frac{f_{jk}}{f_{\max}} \right)^2 \tag{9}$$

Backward inference is a means of the knowledge reasoning in artificial intelligence, in the form of association rules for reasoning, describe the association rules are showing the basic relationship between the URL support and primary confidence of quantitative standard, the quantitative standard is a kind of probability measure that is certain.

6. Experiment and Verification

In this section, we conduct experimental analysis on the proposed system. Our experiment can be generally separated into the following steps. (1) Add the word segmentation dictionary function. (2) In the crawlers and the indexing applications to join the relevant classification algorithm, can be in the classification of the web page collection and processing phase. Use of spiders crawling web pages, then the index of the web page, and store the location of the web pages, in order to provide a text-based search and return to the search results. (3) According to different query requirements, find corresponding index file, read database corresponding page sorting value, descending sort by value output. (4) For the query keywords, and find out the keywords of all web pages, and then descending order according to the order value. About the quality of the order of evaluation is mainly based on two things: search results page and the query item

correlation; High correlation web site in the search results list, position on the sorting quality better more. (5) Compare the experiment result with the others.

In the following figure 9, we show the core code for mining the web information. The code could accurately catch the features and the corresponding index. In the figure 10~11, we show the accuracy and the distribution experiment on the proposed agent system. The result reflects that the agent performance is enhanced and the distribution is reasonable. In the figure 12, we show the comparison result, as our method performs better compared with others.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LIMES SYSTEM "limes.dtd">
<LIMES>
  <PREFIX> <NAMESPACE>http://www.w3.org/1999/02/22-rdf-syntax-ns#</NAMESPACE> <LABEL>rdf</LABEL>      </PREFIX>
  <PREFIX> <NAMESPACE>http://www.w3.org/2002/07/owl#</NAMESPACE>               <LABEL>owl</LABEL>      </PREFIX>
  <PREFIX> <NAMESPACE>http://data.linkedct.org/resource/linkedct/</NAMESPACE> <LABEL>linkedct</LABEL> </PREFIX>
  <PREFIX> <NAMESPACE>http://bio2rdf.org/ns/mesh#</NAMESPACE>                  <LABEL>meshr</LABEL>    </PREFIX>
  <SOURCE>     <ID> linkedct </ID>
               <ENDPOINT> http://data.linkedct.org/sparql </ENDPOINT>
               <VAR> ?x </VAR>
               <PAGESIZE> 5000 </PAGESIZE>
               <RESTRICTION> ?x rdf:type linkedct:condition </RESTRICTION>
               <PROPERTY> linkedct:condition_name </PROPERTY>                  </SOURCE>
  <TARGET>     <ID> mesh </ID>
               <ENDPOINT> http://mesh.bio2rdf.org/sparql </ENDPOINT>
               <VAR> ?y </VAR>
               <PAGESIZE> 5000 </PAGESIZE>
               <RESTRICTION> ?y rdf:type meshr:Concept </RESTRICTION>
               <PROPERTY> dc:title </PROPERTY>                                 </TARGET>
  <METRIC>     levenshtein(x.linkedct:condition_name, y.dc:title)              </METRIC>
  <EXEMPLARS>  70 </EXEMPLARS>
  <ACCEPTANCE> <THRESHOLD> 0.9 </THRESHOLD> <RELATION> owl:sameAs </RELATION>
               <FILE>diseases\_accepted.nt</FILE>                             </ACCEPTANCE>
  <REVIEW>     <THRESHOLD> 0.8 </THRESHOLD> <RELATION> owl:sameAs </RELATION>
               <FILE>diseases\_review.nt</FILE>                               </REVIEW>
</LIMES>
```

Figure 9. The Core Code for Mining the Web Information

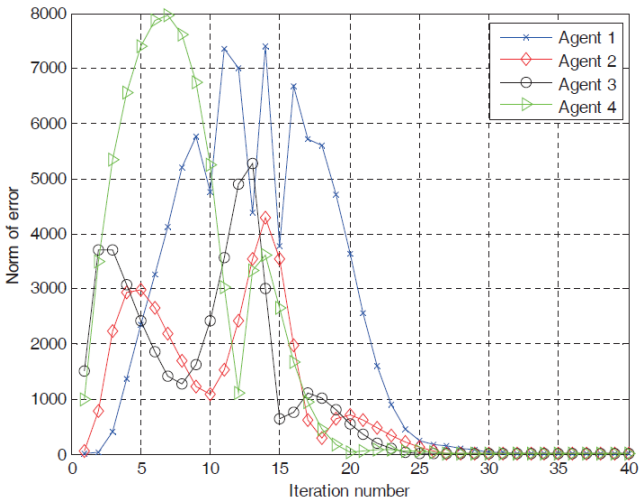

Figure 10. The Accuracy Experiment on the Proposed Agent System

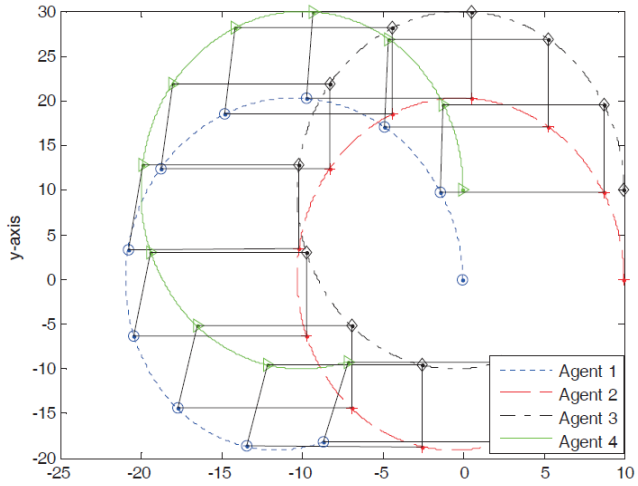

Figure 11. The Distribution Experiment on the Proposed Agent System

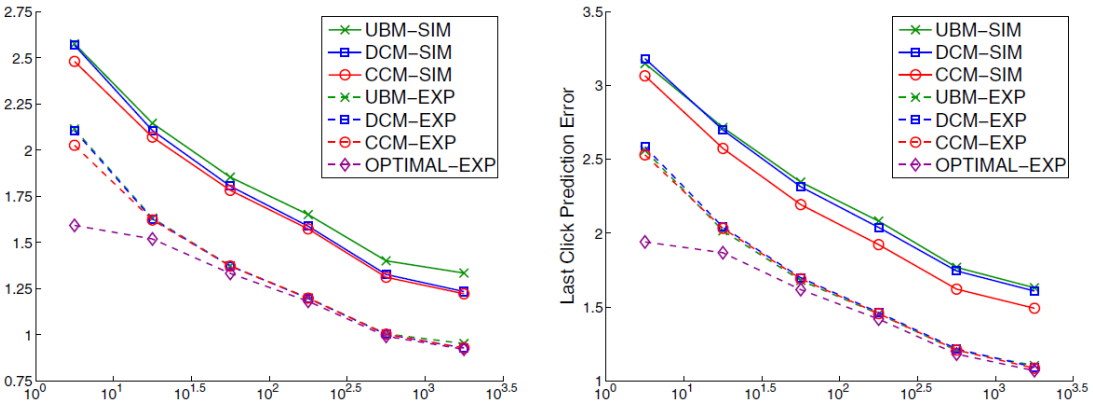

Figure 12. The Comparison Experiment on the Different Algorithms

7. Conclusion

Data mining is a kind of implicit extracted from a large database or the data warehouse predictive information technology, it can dig out the data to ask potential model, find out the most valuable information and knowledge, to guide business conduct or assist the study of science which model is the result of the mining algorithm are used to get the and is a simple description of a probability distribution knowledge or information is based on the model for processing and be easy to understand. In this paper, to enhance the accuracy of the traditional web data mining approaches, we propose the novel deep web data mining algorithm based on multi-agent information system and collaborative correlation rule. Deep page classification is not only to examine the content of the core page and their factors, such as the structure of the semantic and shall be provided according to the user's cognitive ability and the personality characteristics. With the introduction of the multi-agent information system and collaborative correlation rule, the mining algorithms will be able to deal with the complex tasks. Through the experimental simulation, we could conclude that the proposed algorithm holds better of the robustness and feasibility. The mining accuracy is enhanced while the time consuming is optimized. In the future, we will modify the mining scenario to test the efficiency of the method in different conditions and data sets.

References

[1] Y. F. Wang, "Mining medical data: a case study of endometriosis", Journal of medical systems, vol. 37, no. 2, **(2013)**, pp. 1-7.
[2] R. S. Santos, "A data mining system for providing analytical information on brain tumors to public health decision makers", Computer methods and programs in biomedicine, vol. 109, no. 3, **(2013)**, 269-282.
[3] N. Lathia and L. Capra, "Mining mobility data to minimise travellers' spending on public transport", Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, **(2011)**.
[4] D. P. McCloskey, "From market baskets to mole rats: using data mining techniques to analyze RFID data describing laboratory animal behavior", Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, **(2011)**.
[5] L. Bulysheva and A. Bulyshev, "Segmentation modeling algorithm: a novel algorithm in data mining", Information Technology and Management, vol. 13, no. 4, **(2012)**, pp. 263-271.
[6] O. Bryson and K. Muata, "A context-aware data mining process model based framework for supporting evaluation of data mining results", Expert Systems with Applications, vol. 39, no. 1, **(2012)**, pp. 1156-1164.
[7] M. Kaytoue, "Pattern Structures and Concept Lattices for Data Mining and Knowledge Processing", Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, **(2015)**, pp. 227-231.
[8] R. S. J. D. Baker, "Ensuring Reliability of Educational Data Mining Detectors for Diverse Populations of Learners", Presentation at CREA: Center for Culturally Responsive Evaluation and Assessment: Inaugural Conference, **(2013)**.
[9] T. Guo and J. V. Milanovic, "Probabilistic framework for assessing the accuracy of data mining tool for online prediction of transient stability", Power Systems, IEEE Transactions, vol. 29, no. 1, **(2014)**, pp. 377-385.
[10] M. Frank, "Mining permission request patterns from android and facebook applications", Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, **(2012)**.
[11] S. F. Shazmeen, M. M. A. Baig and M. R. Pawar, "Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis", Journal of Computer Engineering, vol. 10, no. 6, **(2013)**, pp. 01-06.
[12] Y. F. Wang, D. H. Wang and T. Y. Chai, "Active control of friction self-excited vibration using neuro-fuzzy and data mining techniques", Expert Systems with Applications, vol. 40, no. 4, **(2013)**, pp. 975-983.
[13] A. H. Wahbeh, "A comparison study between data mining tools over some classification methods", (IJACSA) International Journal of Advanced Computer Science and Applications, **(2011)**, pp. 18-26.
[14] M. Dhakar and A. Tiwari, "A novel Data mining based hybrid intrusion detection framework", Journal of Information and Computing Science, vol. 9, no. 1, **(2014)**, pp. 037-048.
[15] R. Gupta and M. P. Modise, "South African stock return predictability in the context data mining: The role of financial variables and international stock returns", Economic Modelling, vol. 29, no. 3, **(2012)**, pp. 908-916.
[16] K. Suto, "A Failure-Tolerant and Spectrum-Efficient Wireless Data Center Network Design for Improving Performance of Big Data Mining", Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st. IEEE, **(2015)**.
[17] B. Kositanurit, K. M. O. Bryson and O. Ngwenyama, "Re-examining information systems user performance: Using data mining to identify properties of IS that lead to highest levels of user performance", Expert Systems with Applications, vol. 38, no. 6, **(2011)**, pp. 7041-7050.
[18] S. Fong, "Quantitative analysis of trust factors on social network using data mining approach", Future Generation Communication Technology (FGCT), 2012 International Conference on IEEE, **(2012)**.
[19] J. W. G. Busse and Y. Yao, "Probabilistic rule induction with the LERS data mining system", International Journal of Intelligent Systems, vol. 26, no. 6, **(2011)**, pp. 518-539.
[20] M. Debeljak, A. Poljanec and B. Ženko, "Modelling forest growing stock from inventory data: A data mining approach", Ecological Indicators, vol. 41, **(2014)**, pp. 30-39.
[21] M. T. Khan, S. Qamar and L. F. Massin, "A prototype of cancer/heart disease prediction model using data mining", International Journal of Applied Engineering Research, vol. 7, no. 11, **(2012)**, pp. 1-6.
[22] Y. Yang, "Data flow modeling, data mining and QSAR in high-throughput discovery of functional nanomaterials", Computers & Chemical Engineering, vol. 35, no. 4, **(2011)**, pp. 671-678.

