Knowledge Extraction for Semantic Web using Web Mining with Ontology

Dipali Panchal
P.G Student,
Department of Information Technology, Mumbai University, PIIT, New Panvel, India

Sharvari S. Govilkar
Assistant Professor
Department of Computer Engineering, Mumbai University, PIIT, New Panvel, India

ABSTRACT

Today, web is growing rapidly, the users get easily lost in the web's rich hyper structure. The primary goal of the web site owner is to provide the relevant information to the users to fulfill their needs. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed. This paper presents two web mining techniques namely, web content mining and web usage mining in the process of extracting conceptual relationships, and applying web structure mining process which focus on fully-structured form. Web structure mining has used in developing ontologies and help to retrieving process.

Keywords

Web mining, web content mining, web usage mining, web structure mining, ontology.

1. INTRODUCTION

The World Wide Web is a rich source of information and continues to expand in size and complexity [1]. The mass amount of information becomes very hard for the users to find, extract, filter or evaluate the relevant information. This issue lifts up the attention to the obligation of some technique that can solve these challenges. Web mining can be easily used in this direction to carry out the problem with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), and Machine Learning etc. These techniques can be used to discuss and analyze the useful information from web. Dealing with these aspects, there are some challenges we should take it into account as follow.

The following challenges in Web Mining are:

1) Web is enormous.
2) Web pages are partially structured.
3) Web information stands to be miscellany in meaning.
4) Degree of quality of the in sequence extracted.
5) Winding up of knowledge from information extracted.

This paper is organized as follows- Web Mining is introduced in Section 2. We cite the past literature in section 3.The related works are discussed in section 4. Proposed Methodology in section 5. The areas of Web Mining i.e. Web Content Mining in section 6, Web Usage Mining in section 7, and Web Structure Mining are discussed in Section 8 and Ontology on web log in section 9, Evolution in section 10 and finally section 11 discussed conclusion.

2. WEB MINING

Web mining is the application of data mining techniques to discover the patterns from the web. The main objective of web mining is to develop more intelligent tools for potentially help the user in finding, extracting, filtering and evaluating valuable information and resources. Web mining techniques could be used to solve the above problems directly or indirectly. The absolute process of extracting knowledge from Web data is follows

a. Resource finding: the task of retrieving / discovery of locations of unfamiliar files on the network.

b. Information selection and pre-processing: Robotically selecting and pre- processing definite from information retrieved Web resources.

c. Generalization: automatically discovers general patterns at individual web sites as well as across multiple sites.

d. Analysis: Rationale and interpretation of the mined patterns.

3. LITERATURE SURVEY

This research is a clear evidence to show how the two fast developing research areas semantic web and web mining met each other closely. After all the study, we have applied various techniques to build the proposed system for very efficient results. Below given are the details of survey we did on the way to propose this system.

Jayatilaka A.D.S [1] proposed combines web content mining with web usage mining in the knowledge extraction process. Therefore, both the web user's and web author's perspectives are captured with respect to the web content, which ultimately leads to extraction of more realistic set of conceptual relationships. The evaluation results prove the effectiveness of the proposed methodology. This research is a clear evidence to show how the two fast developing research areas semantic web and web mining meet each other closely.

Miguel Gomes da Costa Júnior Zhiguo Gong and T.Nithya [2][3] proposed the knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and Web based data warehousing. He provided an introduction of Web mining as well as a review of the Web mining categories. This paper focus on one of these categories: the Web structure mining.

Gerd Stumme, Andreas Hotho [4] proposed Web usage needs to take into account not only the information stored in server logs, but also the meaning that is constituted by the sets and sequences of Web page accesses. In Web usage mining, the primary Web resource that is being mined is a record of the requests made by visitors to a Web site, most often collected in a Web server log. The content and structure of Web pages, and in particular those of one Web site, reflect the intentions of the authors and designers of the pages and the underlying information architecture. The actual behavior of the users of these resources may reveal additional structure.

This paper [5] proposes a method for making the K-Means algorithm more effective and efficient; so as to got better clustering with reduced complexity for discovering content from web pages using web content mining. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups, usually multidimensional is classified into groups (clusters) such that members of one group are similar according to a predefined criterion. The proposed algorithm uses standard deviation that reduces the time to make the cluster in simple k-mean.

Tatyana IVANOVA Technical University of Sofia, College of Energy and Electronics, Botevgrad Bulgaria [6] proposed and discussed the architecture of the ontology learning module for extension of integrated development environment for learning objects, known also as Learning resource management and development system by integration of semantic technologies.

The Authors Sivakumar and Ravichandran K.S As given in [7] semantic A Review on Semantic-Based Web Mining and its Applications. Author survey the Semantic-based Web mining is a combination of two fast developing domains Semantic Web and Web mining. Our approach is supported by our integrated the current challenges of the World Wide Web (WWW). The idea is to improve the results of Web Mining by making use of the new semantic structure of the Web and to make use of Web Mining for creating the Semantic Web.

The authors Gerd Stumme, Andreas Hotho, Bettina Berendt have studied the combination of the two fast-developing research areas Semantic Web and Web Mining [8]. We observed how Semantic Web Mining can improve the results of Web Mining by exploiting the new semantic structures in the Web; and how the construction of the Semantic Web can make use of Web Mining techniques. The ultimate goal of Semantic Web Mining: “a better Web” for all of its users, a “better usable Web”. We have focused to enable search engines and other programs to better understand the content of Web pages and sites. This is reflected in the wealth of research efforts that model pages in terms of ontology of the content, the objects described in these pages.

The author Gautam R. Raithatha has analyzing different web mining techniques for extracting knowledge from web data for creating semantic web [9]; from this paper we have concluded that the unstructured data present on the web can be scanned to create ontologies to populate the knowledge base of the search engine. The information inserted in this knowledge base is in structured form that computers can understand. This information from the knowledge base can be used by the computers to better serve the web user’s query. Thus we can add in proposed system how semantics to the current web by extracting knowledge and creating ontologies to create the semantic web.

The authors Govind Murari Upadhyay, KanikaDhingra have proposed Data mining techniques and web content mining tools are used to extract useful information or knowledge from web page contents [10] .By these techniques we can make our search of contents over the web faster and exact. From this paper we studied various web content mining techniques and uses of Web Content Mining and applied in our proposed system.

The authors ShailyG.Langhnoja, Mehul P. Barot and, Darshak B. Mehta have presented Web Usage Mining is application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [12] . Analyzing data through web usage mining can help effective Web site management, creating adaptive Web sites, business and support services, personalization, and network traffic flow analysis and so on. Lots of research has been done in this field while this paper emphasizes on finding user pattern in accessing website using web log record. The aim of this paper is to find user access patterns based on help of user’s session and behavior. Web usage mining includes three phases namely pre-processing, pattern discovery and pattern analysis. In this paper we studied combined effort of clustering and association rule mining is applied for pattern discovery in web usage mining process in our system. This approach helps in finding effective usage patterns.

The authors Gerd Stumme, Andreas Hotho have proposed Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining [13]. Web Mining aims at discovering insights about the meaning of Web resources and their usage. Given the primarily syntactical nature of data Web mining operates on, the discovery of meaning is impossible based on these data only. Therefore, formalizations of the semantics of Web resources and navigation behavior are increasingly being used. This fits exactly with the aims of the Semantic Web: the Semantic Web enriches the WWW by machine process able information which supports the user in his tasks. In this paper, from this paper we observed the interplay of the Semantic Web with Web Mining, with a specific focus on usage mining.

The author Yan Wang have survey the researches in the area of Web mining with the focus on the Web Usage Mining [14]. Three recognized types of web data mining are introduced generally. Around the key topic of this paper - usage mining, we provide detailed description of the three phases of the process. An example of usage mining system is given to illustrate the overall usage mining process. Moreover, the research of major applications of usage mining personalization and navigation pattern discovery are discussed. Finally, we wrap up this paper with the most controversial topic -the user privacy.

The authors L.K. Joshila Grace1, V.Maheswari2, Dhinaharan Nagamalai3, gives a detailed look about the web log file, its contents, its types, its location etc [15]. Added to this information it also gives a detailed description of how the file is being processed in the case of web usage mining process. The various mechanisms that perform each step in mining the log file is being discussed along with their disadvantages. The additional parameters that can be considered for Log file entries and the idea in creating the extended log file is also discussed briefly. The extended work is to combine the concept of learning the user’s area of interest.

The author Ahmed Sultan Al-Hegami has proposed a web usage mining approach for generating the ontology [16]. This approach is consisting of three stages beginning from using an existing or semi-automatically built ontology intended to enhance information retrieval. Using Web Usage Mining methods like classification and sequential pattern mining techniques to analysis the log files to extract pattern. From this paper concept we have generated Web log ontology and using the extracted knowledge.

4. RELATED WORK

The World Wide Web has grown in the past few years from a small research community to the biggest and most popular way of communication and information dissemination. Every day, the WWW grows by roughly a million electronic pages, adding to the hundreds of millions already on-line. WWW serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content and software.

The continuous growth in the size and the use of the WWW imposes new methods for processing these huge amounts of data. Moreover, the content is published in various diverse formats. Due to this fact, users are feeling sometimes disoriented, lost in that information overload that continues to expand. Web mining is a very broad research area emerging to solve the issues that arise due to the WWW phenomenon. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. Attributes include HTML tags, word appearances and anchor texts. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval.

Due to the unstructured and semi-structured nature of Web pages, it is a challenging task in categorizing and extracting content from the Web. In this ontology [7] plays a major role. Ontology is being represented as a set of concepts and their inter-relationships relevant to some knowledge domain. The knowledge provided by ontology is extremely useful in defining the structure and scope for mining Web content. Ontology is defined as an explicit specification of a set of objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them.

5. METHODOLOGY

Eliminating the human intervention could contribute significantly for the rapid growth of semantic web. However, these methodologies should not compensate the accuracy of the extracted semantics for full automation. Proposed methodology contains four main stages. Following techniques has used to extract the information from past behavior of users.

1. Web Content Mining.
2. Web Usage Mining.
3. Web Structure Mining.
4. Ontology.

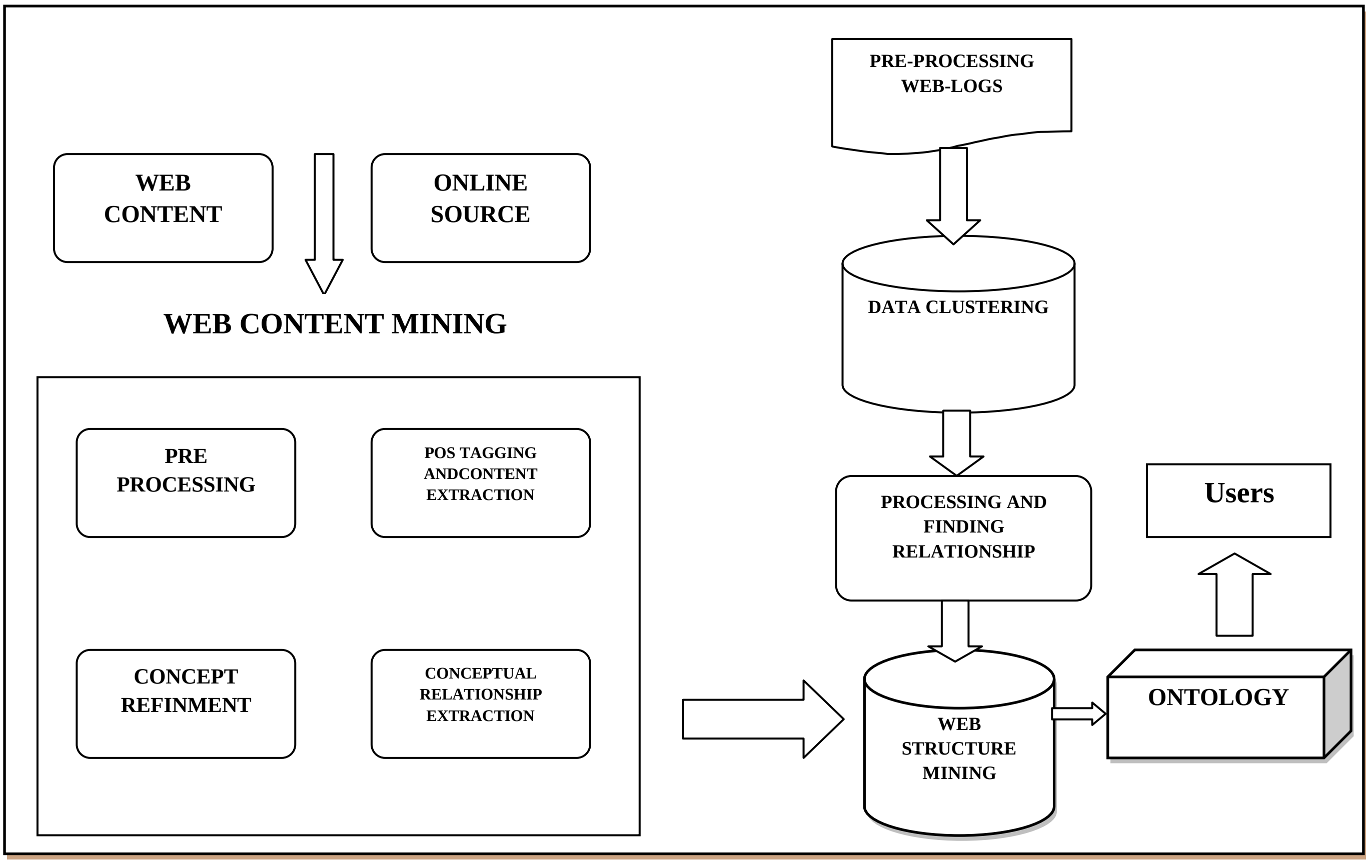

Fig 1: Proposed Architecture

6. WEB CONTENT MINING

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. Web content mining phase discovered important concept only and rest of filtered out throw following phases. This phase accepts the input documents and generates the preprocessed sentences. This phase apply stemming procedure, remove stopping word and apply POS tags of each sentences from input documents for each word of sentences and generate list of common concepts which is most frequently occurred for identify concept relationships and extraction.

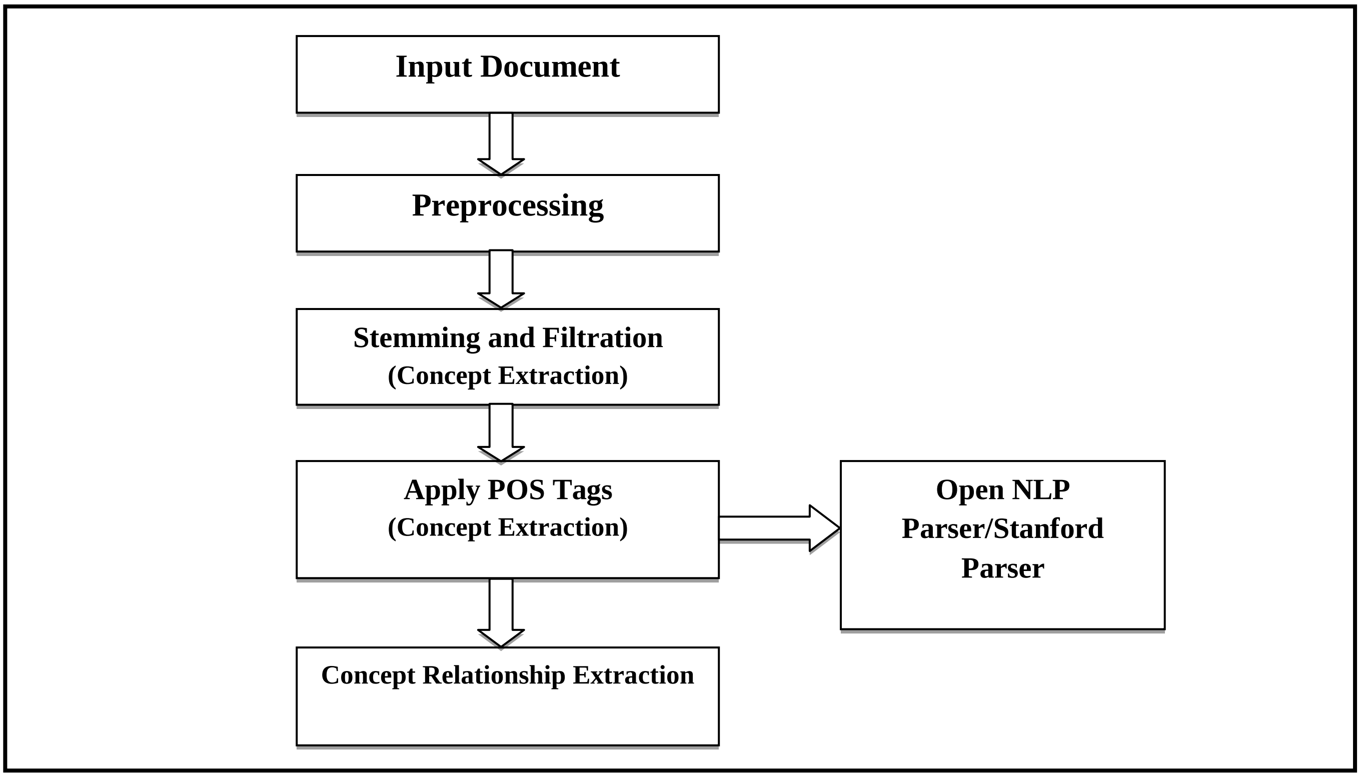

Fig 2: Procedure of web content mining

7. WEB USAGE MINING

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications.

It tries to make sense of the data generated by the web surfer’s sessions/behaviors. While the web content and structure mining utilize the primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web.

The web usage data includes the data from web server logs, proxy server logs, browser logs, and user profiles. (The usage data can also be split into 3 different kinds on the basis of the source of its collection: on the server side (there is an aggregate picture of the usage of a service by all users), the client side (while on the client side there is complete picture of usage of all services by a particular client), and the proxy side (with the proxy side being somewhere in the middle).Registration data, user sessions, cookies, user queries, mouse clicks, and any other data as the results of interactions [9].

Web usage mining analyzes results of user interactions with a web server, including web logs, click streams. Web usage mining also known as web log mining.

Web usage mining process can be regarded as a three-phase process consisting [10]:

1. Preprocessing/ data preparation - web log data are preprocessed in order to clean the data – removes log entries that are not needed for the mining process, data integration, identify users, sessions, and so on
2. Pattern discovery - statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns.
3. Pattern analysis phase - discovered patterns are analyzed, and filter out the uninteresting rules/patterns.

After discovering patterns from usage data, a further analysis has to be conducted. The discovered rules and patterns can then be used for improving the system performance / for making modifications to the web site.

The purpose of web usage mining is to apply statistical and data mining techniques to the preprocessed web log data, in order to discover useful patterns and help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service.

The applications are:

- Extract statistical information and discover interesting user patterns.
- Cluster the user into groups according to their navigational behavior.
- Discover potential correlations between web pages and user groups
- Identification of potential customers for ecommerce
- Enhance the quality and delivery of Internet information services to the end user.
- Improve web server system performance and site design.
- Facilitate personalization

8. WEB STRUCTURE MINING

As the web is growing rapidly, the users get easily lost in the web's rich hyper structure. The primary goal of the web site owner is to provide the relevant information to the users to fulfill their needs. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed. Web Structure Mining plays an important role in this approach [2][3].

The goal of Web structure mining is to generate structured summary about the website and web page. It tries to discover the link structure of hyper links at inter document level. As it is very common that the web documents contain links and they use both the real or primary data on the web so it can be concluded that Web structure mining has a relation with Web Content Mining [3].

In proposed system, structure mining applied to the web usage data which generated through access log. Therefore web author could easily analyze and filtered the contents.

9. ONTOLOGY

Ontology is used for knowledge sharing and reuse. It improves information organization, management and understanding. Ontology has a significant role in the areas dealing with vast amounts of distributed and heterogeneous computer based information, such as World Wide Web, Intranet information systems, and electronic Items [6].

The ontology learning process can be converting in to a fully automated process which will totally eliminate the human involvement in ontology learning. Ontology learning process is a sequence of steps, including terminology extraction, identification of grammatical characteristics of extracted terms and it meaning, determining it type as ontology elements, and placing them in the right place of the learned ontology [11].

OWL ontology languages allow users to write explicit, formal conceptualization of domain models. They full fill the main requirements of ontology languages such as

- A well-defined syntax
- A formal semantics
- Convenience of expression
- Efficient reasoning support

The subclass relationship between OWL and RDF property

OWL builds on RDF and RDF schema and uses RDF‘s XML based syntax. OWL can be defined with XML based syntax, Abstract syntax which makes use of RDF and Graphical Syntax.

Proposed architecture applied to generate ontology based on web log which described in web ontology language (OWL).Web authors will be helped by analyzing user's browser's history of web sites which are frequently visited by people and the items which are popular in market and the same is easily identified by this application. Web structure mining has used in developing ontologies and help to retrieving process.

10. EVALUATION

We have used two straight forward measures which derived from information retrieval namely Memory Usage and Parameter that is Time of execution in millisecond.

In first evolution graph that is a measure of execution time and memory usage of content extraction from web usage mining process throw access log where quantity of results returned is measured.

In second evolution graph that is measure of ontology creation time based on web log where quality of the results obtained is measured.

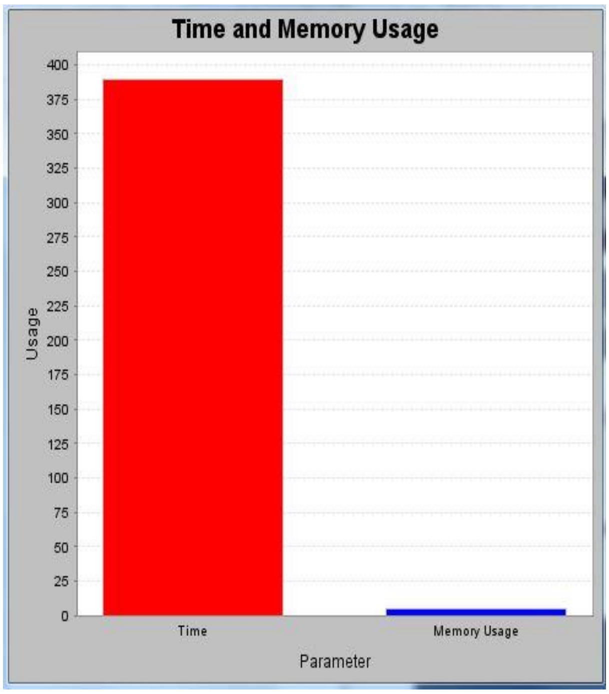

Fig 3: Execution Time and Memory Usage of web usage process based on Access log(Time In millisecond and Memory usage in mb)

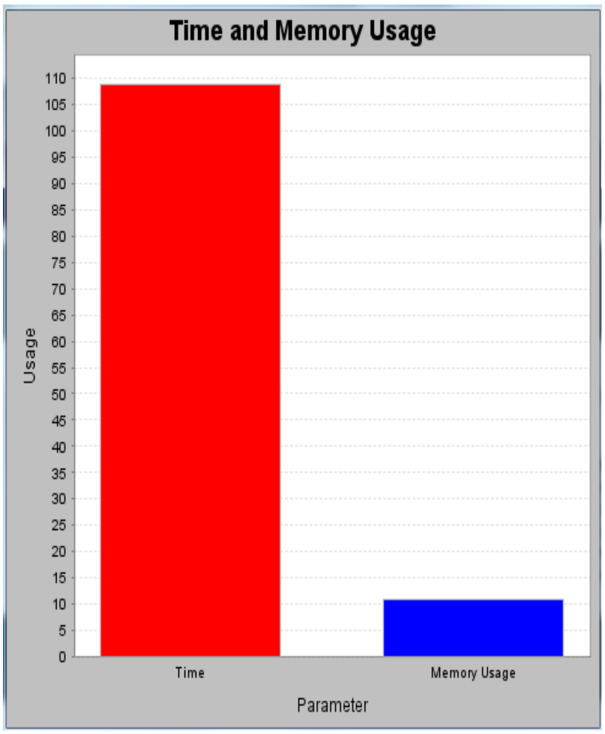

Fig 4 : Execution Time and Memory Usage of Ontology based on Web log. (Time in millisecond and Memory usage in mb)

11. CONCLUSION

This paper contains the introduction of web mining and its related techniques such as web content mining, web structure mining and web usage mining. The goal of search engines is to provide relevant information to the users to cater to their needs. Therefore, finding the content of the web, retrieving the user's interests and needs have become increasingly important.

Web structure mining has served as tool to improve retrieval performance, classification accuracy and for developing ontologies with profitable results which converts in to a fully automated process to totally eliminate the human involvements and ontology generated based on web log. This mechanism will helped to web authors by analyzing user's browser's history of web sites which are frequently visited by users and identified the products which are very popular in market.

12. REFERENCES

[1] Jayatilaka A.D.S and Wimalarathne G.D.S.P. 2011.Knowledge Extraction for Semantic Web Using Web Mining.

[2] Miguel Gomes da Costa and Júnior Zhiguo Gong. 2005. Web Structure Mining: An Introduction.

[3] T.Nithya. August 2013. Link Analysis Algorithm for Web Structure Mining.

[4] Gerd Stumme, Andreas Hotho. 2005. Web Usage Mining for and on the Semantic Web.

[5] Amita Verma, Ashwani kumar. January 2014. Performance Enhancement of K-Means Clustering Algorithms for High Dimensional Data sets.

[6] Tatyana. 2010. Ontology Learning and Management Capabilities. IVANOVA Technical University of Sofia, College of Energy and Electronics, Botevgrad Bulgaria.

[7] K. Sridevi1 and Dr. R. Umarani2. July – August 2012. A Survey of Semantic based Solutions to Web Mining. Department of Computer Science, Nehru Memorial College, Puthanampatti Trichy District, Tamilnadu, India,

[8] Ahmed Sultan Al-Hegam, Mohammed Salem Kaity . March 2014. An Ontology Framework based on Web Usage Mining. Sana'a University.

[9] RACHIT ADHVARYU. A Review Paper on Web Usage Mining and Pattern Discovery. Student M.E CSE, B. H. Gardi Vidyapith, Rajkot, Gujarat, India.

[10] V.Chitraa and Dr. Antony Selvdoss Davamani. 2010. A Survey on Preprocessing Methods for Web Usage Data. CMS College of Science and Commerce Coimbatore, Tamilnadu, India.

[11] Renáta Iváncsy, István Vajk. 2006. Frequent Pattern Mining in Web Log Data. Department of Automation and Applied Informatics.