SEMANTIC WEB MINING and IMPROVING PERFORMANCE

Zahid Javed, Aqsa Hameed, Amna Umer* and Werdah Abbas

Department of Computer Science, University of Agriculture Faisalabad Pakistan *Corresponding aauthor's email: <u>amna.umer1119@gmail.com</u>

Data mining is a procedure of understanding knowledge from such huge data. Our main problem is to mine information from web because there is huge amount of data on web which is unstructured way. Semantic webs offer us a smart service in which we can synchronize and arrange the data over the web in discipline manner. Due to enormous research on web mining, many techniques and applications have been introduced. But our main challenge is still the same, that is, how to gather accurate and necessary data on the user demand. Our three main goals and challenges are accurate data, necessary data and according to the user demand. In this research paper, we are discussing concepts of web data mining we included web data mining techniques, challenges algorithms used in web data mining and we have given the concepts of semantic web data mining. At the end we are moving towards our main goal that is to improving performance of web mining and take out useful data in well refined form. Keywords: Semantic Web; Web Mining; Intelligent Agent; Web Mining Techniques; Web Mining Challenges;

Web Mining Algorithms; Content Extraction.

INTRODUCTION

From last few years with the increase of usage of internet many organization and business start relying on the internet. Many resources and information is available on internet so that web has generated the huge amount of data. Usage of internet is at its peak. According to the paper [15] an analysis shows that how many activities occur in one second on the internet. This is shown in Table 1.

Online Occurrence	Number
Global Internet transfer	25 TB
New websites created	9
SPAM emails sent	1800000
Instagram photos uploaded	5000
Skype calls made	1500
Tweets tweeted	4000
Dropbox files uploaded	10000
Google searches made	45000
YouTube videos viewed	92000
Facebook likes made	55000

TABLE I. ONLINE OCCURRENCES HAPPENING EACH SECOND

Data available on web but all the data is in rough form and it is useless until it is not organized in good manner. User required very less data from this huge amount of data which is only related to its required data, remaining data is useless for him and user has no concern with it. Data mining permits an enterprise to habit the bulk volumes of data that it is assembled and organized to help business and to support decision making. Data mining has three major modules clustering Association rules or classification and sequence Analysis. Web data mining is a presentation of data mining. Web mining is a knowledge retrieval infrastructure which exploits the data mining technique to extract information automatically from web. Web mining has introduced many new research fields like Natural Language processing (NLP), Machine Language (ML), Information Retrieval (IR) and statistics. Due to massive, different,

formless nature and self-motivated data properties, web research has introduced many challenges like scalability, multimedia and temporal issues.

This paper is divided into five major sections. Section I is about web mining technique. Section II summarizes the web mining challenges. Section III includes web mining algorithms. Section IV proposes architecture for improving performance in web mining.

RELATED WORK

Oren Etzioni was the scientists who introduced the web mining for the very first time in his paper [1]. He classified the web mining into 3 categories in his paper. Due this new technique, new other fields were opened for the researchers to do search. After that authors of [2][3][4][5] had also described these 3 categories by using different names. After that researcher of [6][7][8] have combine the content and structure mining to improving web mining. In 2001, Bemer-Lee [9] had introduced a novel approach in which web had become more semantic and more understandable machine which was named as "semantic web". After that semantic web was started emerging with new generation web. In 2006, Bemers-Lee turned up with the principle of linked data in paper [10]. It gave the guideline, that how can we use the standard web technology to link data from multiple sources and which data publisher have been initiated to recognize the web of data.

WEB DATA MINING TECHNIQUES

A. Web content mining

Web content mining is the procedure of mining valuable facts from the constituents of web documents. Web is dynamic and contents keep on changing continuously. Facts that are conveyed to the user are introduced in it. Text, images, structured record like lists, tables audio and video can be included. Web pages can be organized in a tree structure page. In which so many HTML and XML tags could be included. Recently researchers have introduced new approaches on web content mining which are being used to mine the valuable data. The approaches are Information Retrieval (IR) and Database Approach (DB). 'IR' utilizer is the intelligent agent approach to improve the information searching and filtering.

- 1) *Agent Base Approach*: This technique is work as a software entity which is working for others behalf.
- Database Approach: This approach plays a central role to manage semi structured data in more organized form. It uses data mining tools and query language to summarizing the data. The tools that DB approach uses are hierarchical Database and Query systems.
- B. Web Structure Mining

Web mining is an instrument that is used to categorize the association concerning web pages related by straight link association or information. This joining permits the search engines to abstract data related to search query by associating directly web pages from web sites. Web structure mining operates on inter documents level by using hyperlink as a getaway to others interlinked pages. It helps us to improve data quality related to search query. This is done mainly using pages ranks and related hyperlinks. Web structure mining has been divided into two main categories:

1) Hyperlink: Hyperlink is a structure element which is used to connect a web leaf in diverse locality in identical web pages or with diverse web pages. Hyperlink, which is inside the document is called Intra –hyperlink and is in web itself is called inter-hyperlink. It uses so many links analysis models like pages ranks, weighted pages rank, and Hypertext Induced Topic Search Algorithm.

 Document structure: It organizes the contents in a tree structure which use so many HTML and XML tags. The main purpose of this is to automatically retrieve the Document Object Model (DOM).

C. Web Usage Mining

Web usage mining is the method of mining beneficial data from server logs. It is a process of discovery that what is user seeing for on the internet. Some user like textual contents and some refer images and multimedia data. The best usage pattern is used to discover in this technique so that we can understand the need of web based application in better way. Web usage mining; extract the data from different servers. It included logo server, proxy server, browser log and organizational database. Web usage mining is additional categorized on the basis of data usage type. It includes Application Level Data. Web server data and Application Server Data.

CHALLENGES

On the matter of searching data, web users always are like drowning in an 'ocean' of information. Much information faces the problem of overloading. Web mining introduces many challenges and problems in the field of search. In this section we will discuss them one by one.

A. Finding relevant information

To find out the web specific information is a problem. When user search on the net then either they directly search the web document or uses the search engine, when user write any query on search engine so they enter into many keywords related to that information and search engine returns several ranked pages which are related to the query. It will be occur with two problems. First one is low precision which is occur due to return of a lot of irrelevant pages by search engine and second one is low recall which is occur due to deficiency of compatibilities of keys. So finding relevant information is still a challenge in research.

B. Finding needed information

As discussed before when user search on a search engine it get many irrelevant web pages as a result. The data that user needed is difficult to analyze. For example a user want to search python programming language then user search for single word query python the result may be about the kind of snake not about the language. This result misleads the user this is also a problem to web mining.

C. Unstructured Information

There is a lot of information on web that is mostly in an unstructured format. This is the main unresolvable issues on web mining system and cause of this issues it has weak operational techniques and tools which are designed to convert structured information into useful knowledge when we use them on unstructured information, it becomes in affective. The word "unstructured" has been used in different ways when we talk about relational database system then it means that data is not saved in row and table format. In information system, this mechanizes shows the information which has no pre-defined structure and it can't be utilized in any computer system directly. Unstructured data exist in most organizations like blogs, emails etc. this issue must be addressed so that we can share unstructured information and get useful information through this.

ALGORITHM USED IN WEB MINING

Web mining quickly gathers and assimilates information from several web sites. Collecting information from many separate sources causes many issues. To deal with these issues, computing power is required not the brain power. One solution is use to AI algorithm. Here we discussed some algorithms used to extract information from web sites.

A. Association

In this algorithm, we imply association relationship between objects that occur together to another in a database. If there is set of transactions and every transaction is a set of strings called items then an relationship rule is am appearance from X to Y, where X and Y are the set of stuffs. By imposing the association rule, we mean that a transaction of database X tends to contains y.

B. Classified Algorithm

In this algorithm we classify data into different classes or groups and develop a explanation or ideal for each class in a database allowing to the features that are present in a set of class labeled working out data. There are many classified methods available and these are statistical method, database oriented method, neural networks, decision tree method, rough set etc.

C. Sequential Analysis

In this algorithm we have a look on sequential designs called data arrangement. Every data sequence is an ordered list of transaction where every item is a set of items. Typically time is associated with every transaction. Problem is that we have to find out the sequential pattern to the users with specific minimum support where the support of that sequential pattern is the percentage of that data which contain the pattern.

D. Clustering Algorithm

Clustering Algorithm is grounded on Probability and enlargement based. In probability step, cluster membership of every case is calculated and then in maximization step, using these cluster membership parameter of model are estimated. The author of paper [13] performed a case study analysis on clustering algorithm. They conducted many experiments based on different parameter. They had also concluded and discussed the results. These experiments show that clustering algorithm is very efficient and scalable. Here we refer some research of paper [13]. The case study analyzes the users which include following list of measurements.

- 1) Most requested pages
- 2) Least requested pages
- 3) Top exit pages
- 4) Most accessed directories
- 5) Most downloaded files
- 6) New versus returning visitors
- 7) Summary of activity for exam period
- 8) Summary of activity by time increment
- 9) Number of views per each page
- 10)Page not found

SEMANTIC WEB MINING

The term semantic web implies an intelligent web which is not only for human being but also processes the information for the computer. It also increase the probability of relevant information retrieval and machines are also capable of interpret and exchange information on web. But this is complex due to involvement of machine because machine has not that vision which human beings have. Semantic web is used to make data understandable by the machine. In the field of web mining, researchers have solved the issues of semantic interpretability using semi structure by technique to extract useful data from huge amount of rough data. This combination has introduced one new technique called semantic web mining. Figure 1 is summarizing working SWN. In first phase, data is being extracted by the unstructured data. So that in consistency of data could be removed and utilize the intelligent approach. In second phase, we are retrieving the knowledge from the linked web pages. In third phase, we are getting the desire information from the result of second phase to give proper results. Infrastructure of semantic web mining is explained in Fig. 1.

A. Content Extraction

Content extraction is the process of extracting semantic information from the unstructured data such as email, audio, video, images, blogs, and presentation. Web content mining is used in this phase we use intelligent agent (AI) and information retrieval approaches to get the content data from web pages as discussed before. IR approach improves and filters the information.

B. Knowledge discovery

It is the process of discovering knowledge from the linked web pages on the web. This process uses ontology matching and page ranked algorithm.

- Ontology Matching: It find out direct relationship between entities of ontology that is related semantically and also the set of synonyms concepts which are parallel in meaning but have different names or structure. This relationship is used in many things, like data conversion, ontology, integration, question answering etc. this is a competent technique, which is used to study the design of conceptualization is vision of different words which reveal the objects and their relationship with other entities. When we find out the relationship between 2 ontologies then we can compute the weight by page rank algorithm.
- Page Rank: Page rank is a mechanism for ranking the web pages based on their usage, quality and content access and finds the numerical weight of each page on the basis of citation analysis.

After this process knowledge discovery phase measure those web pages which are most relevant to the user query.

C. Information Retrieval System:

Finally in last stage SWM system is acquiring related information that is attained by using OLAP, Agent system query language and data mining technique.





PROPOSED MODEL

There is a lot of unstructured data available on web and it is very difficult to analyze the data under the data under a common structure and to view them. We have considered both web mining and semantic web mining which are being simplified by a semantic agent. This model support the well-structured semantic network and unstructured real world network situation. Proposed model is represented in figure and it has following steps:

Step1: In this step, Query is directed to the query processor through query interface. Query process is the sub component of data server that processes the request of the user.

Step 2: Query processor call the both traditional query and intelligent engine through interface engine parallel with user parameters. Interface stop controller allows the user to stop the mining process if required. Query engine is a service that takes the search request and gives back the result to caller after evaluating and executing it. It works between the client and underlying data source as an intermediary layer. It interprets the client search request and informs that how to access the data source. Query engine sends the initial results to the interface engine and then the result are transferred to the RDF database.

Step 3: to perform agent based searching an initial ontology should be build and through this initial ontology, there is need to gather many concepts related to the web object to construct this initial ontology. In many cases, it is necessary to use gathered data for specialized clustering algorithm. Ontology model is used to merge the knowledge of experts with environment. Ontology level is saved in ontology library for future use.

Step 4: When agent receives the query parameter from query processor through interface engine then it checks the RDF DB if it has the user desired results then agent will send these result to the user through interface engine. On the other hand, agent will find out the relationship between the user query and other web entities from the ontology library and will make an ontology base with relational entities, if the desired result not find in RDF DB.

Step 5: Ontology base has saved those possible nodes which are related to the user request and collected from the agent acquired from the knowledge of ontology base. Resource acquisition module collects task related information form web. The main problem during the data acquisition is that much irrelevant information is acquired. The total performance of this model is depends upon the performance of this model of data acquisition.

Step 6: Resource acquisition module detects and collect the resource nodes of closet characteristics and store them into the RDF DB.

Step 7: Semantic web mining is used to mine the data from RDF DB for better output and then these output are sent back to the agents.

Step 8: For more related information, output sematic web mining filtered through many processes.

Step 9: In this final step, agent sends the save relational results to the interface engine from RDF DB. Then result ranking engine will ranks the output and then sends the results to the user. This process is so efficient when user do not have parameters to find the desired output from web.

Full working of this model is explained by a diagram in Fig. 2.



Fig 2.Proposed Web Mining Model

CONCLUSION

Due to unstructured nature if web there is a lot of data that is not related to the user interest. When user search a keyword in search engine then it returns so many websites because the keywords define in between them and second problem is the ambiguous keywords. Web mining is a technique through which we can filter data and then it can be presented to user. This paper devotes the concept of web mining and semantic web mining and concludes that by using web mining model under semantic agent. We can improve performance of data mining over web and outputs that are provided to the user by filtering process.

REFERENCES

- Oren Etzioni, *The World Wide Web: Quagmire or* gold mine, Communications of the ACM, Vol.39 (11), Pp.65-68 (1996).
- R. Kosala and H.Blockeel, Web Mining Research: A survey.SIGKDD:SIGKDD Explorations: newsletter of the special interest Group (SIG) on knowledge discovery and data mining, ACM, vol.2, Pp. 1-15 (2000).
- S. K. Madria, S. S. Bhow mick, E. P. Lim etal,

Research Issues in Web Data Mining. In proceeding conference, Dawak, 99, Pp.303-3012, (1999).

- R. Cooley, *The web usage mining: Discovery and Application of Interesting patterns from web data*, Ph.D. thesis, Dept. of computer Science, university of Minnesota, May 2000.
- M. Spiliopoulou, *Data mining for the web*. In proceeding of Principles of data mining and knowledge Discovery, Third European Conference, PKDD 99, Pp.588-589.
- F. Sebastini, *Machine Learning in Automated Text Categorization*. Tech. report B4-31, Istituto di Elaborazione dell'Informatione, Consiglio Nazionale delle Ricerche, Pisa, (1999).
- S. Chakarabarti, "Data Mining for Hypertext: A Tutorial Survey", ACM SIGKDD Explorations, Vol. 1, no. 2, pp. 1-11, 2000.
- J. Fumkranz, "Web Structure Mining: Exploiting the graph Structure Of the World Wide Web", Osterreichische Gesellschaft fur Artificial Intelligence (OGAI), vol. 21, no.2, Pp. 17-26 (2002).
- T. Berners-Lee, J. Hendler and O. Lassila, "TheSemantic Web", Scientific American. Vol. 284(5), Pp. 34-43(2001).
- T. Berners-Lee. "Linked Data—Design Issues", 2006;

of the 4rth International Conference on

http://www.w3.org/DesignIssues/LinkedData.html Hiteshwar Kumar Azad, Kumar Abhishek. Semantic-Synaptic Web Mining: A Novel Model for Improving the Web Mining. In proceeding

Communication Systems and Network

Technologies 2014 IEEE.

- Vijay Rana, Dr Gurdev Singh. Analysis of Web Mining Technology and Their Impact on Semantic Web. International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH14) 28 & 29 November 2014.
- Samia Jones, Omprakash K. Gupta. *Web Data Mining: A Case Study*. Communications of the IIMA 2006 Volume 6 Issue 4.
- Y.Raju, Dr. D. Suresh Babu. A Novel Approaches in Web Mining Techniques in Case of Web Personalization. International Journal of Research in Computer Applications and Robotics Vol.3 Issue.2, Pg.: 6-12 February 2015.
- Abzetdin Adamov. Data Mining and Analysis in Depth. Case Study of Qafqaz University HTTP Server Log Analysis. Applied Research Center for Data Analytics and Web Insights (CeDAWI) Qafqaz University, Baku, Azerbaijan.
- Sumaiya Kabir, Shamim Ripon, Mamunur Rahman, Tanjim Rahman. *Knowledge-Based Data Mining Using Semantic Web*. 2013 International Conference on Applied Computing, Computer Science, and Computer Engineering.