Analysis of Web Mining Technology and Their Impact on Semantic Web

Vijay Rana
Research Scholar, Punjab Technical University, Jalandhar, Punjab, India.
vijay.rana93@gmail.com

Dr Gurdev Singh
Professor, Gurukul Vidyapeeth Institute of Engineering & Technology, Patiala, Punjab, India.
singh_gndu@yahoo.com

ABSTRACT

Web mining gives an innovative direction for scientific research and pushing web technology to toward making the meaning information and exploits some data mining techniques to automatically extract valuable information from the World Wide Web. It makes an environment where the information available on the web can be semantically interpreted. Web mining assembles more feature to built web personalize interaction and customizing a web site according to the requirements of users, obtaining advantage of the knowledge attained from the study of the user's browsing behavior. However, the existing heterogeneous environment and usually a large server can't guarantee the reliability and making unstructured information. Therefore necessitate for research activities in web mining to extending a standard but intelligent, adaptive and distributed framework for the support of heterogeneous environment is apparent and semantic web is most promising techniques at that instances. In this work we have tried to through the light on concept of web mining techniques.

Keywords: Web Content Mining, Web Structure Mining, Web Usage Mining, Semantic Web Mining.

I. INTRODUCTION

The web is an interactive medium to retrieve and manipulate information on the internet. With the rapid growth of information sources existing on the web, it has become necessary to exploit some automated techniques to find the valuable information on web and summarized their usage patterns [8]. These essential aspects confer the rising demand of intelligent system, which able to extract desired information and Web mining is most suitable tool in these circumstances. Web mining is a knowledge retrieval infrastructure, which exploits data mining technique to automatically extract desired information from the World Wide Web. The term web mining has been used in three different ways, web content mining, web structure mining and web usage mining. Web content mining is a vital tool to discover useful information of web contents [5]. Web structure mining is the process of extracting knowledge from the interrelated hypertext document on the web and web usage mining is process of extracting user browsing behavior and access patterns.

The web mining is a highly research topic of several web communities such as Information Retrieval, Machine Learning, NLP and AI [15]. It is combination of two active research areas: Data Mining and World Wide Web, which make the new technologies and infrastructure components obligatory to build a web that handles its users wisely [8]. However, the existing centralized infrastructure and usually a large server can't guarantee the reliability and non-redundancy of information provided to users. Therefore require for research activities in web mining to extending a standard, flexible but intelligent, adaptive and distributed framework for the support of heterogeneous environment is apparent. Thus the obligation for predicting user needs in order to enhance the usability and user retention of a Web site can be addressed by semantic web [20]. Semantic web is an innovative approach with web mining system to improve the search and extraction pattern of web. It gets the vision of structuring information accessible across the web in a meaningful way improving search mechanism and thus resulting user satisfaction. It makes an environment where the information available on the web can be semantically interpreted.

This paper has been broadly divided into four sections**.** Section 2 gives a brief overview of web mining categories. Third section highlights the challenges in web mining system. Fourth section defines the working of semantic web mining system and finally section 4 concludes by presenting open research challenges. The upcoming section provides an insight into the present scenario web mining categories.

II. WEB MINING CATEGORIES

In this section we have presented classification of web mining such as web content mining, web structure mining and web usage mining.

2.1 Web Content Mining

Web content mining is a superlative tool to discover of useful information from the web contents and documents. These web contents contain two types of entities: text and multimedia contents. Text content consists of semi-structure data such as HTML documents and unstructured text, i.e Knowledge Discovery in Text (KDT) [15]. On other side multimedia content consists of image, audio, video and structured records, which formulate semantic annotations.

The recent development of web content mining technique have encouraged researchers to develop more intelligent tools for information extraction, such as Information Retrieval (IR) and Database approach (DB). IR utilizes intelligent agent (AI) approach to improve the information searching and filtering the information from the users inferred or solicited profiles. DB employs to summarize the data on the Web and combine them so that more complicated queries could be performed. The exponential development in web content mining tool have allowed system to enhanced knowledge deliverance of information through integration of various technologies such as agent based approach and database approach.

2.1.1 Agent-Based Approach:

Agent based approach an agent is an autonomous software entity, which has the ability to perform his task on the behalf of others. They may able to accomplish their job and exhibit key attributes like mobility, learning ability, user preferences and search patterns. It also utilizes to identify the topic representation by a web document and locate web pages across different servers that are similar. Agent based content mining system performs their task the following three types.

- ***Intelligent Search Agent*** *[15]:* there are numerous intelligent systems have been developed for retrieving knowledge oriented information and interpret with the discovered information such agents are Harvest and ShopBot. Harvest [1] system is utilized to attain specified domain information about of scrupulous types of document. ShopBot [9] system employ for retrieving product information from various vendor websites and using it to get general information of the product domain.
- ***Information Filtering Agent****:* some intelligent system work to retrieving, filtering and categorize information automatically such as HyPersuit [3]. Hypersuit system used for making cluster hierarchies of hypertext documents.
- ***Personalized web Agent***: is a process of customizing a web site according to the requirements of specific users, obtaining advantage of the knowledge attained from the study of the user's browsing behavior in association with other information gathered in the Web context, structure and user profile data, such as WebWatcher and Firefly [9,11].

2.1.2 Database Approach

Database approach plays a central role to managing the semi-structured data into more organized form, so that better information management and querying on the web become possible. It also uses query languages and data mining tools to summarize it. These tools are:

- ***Hierarchical Database*** *[15]:* multilevel database approach utilize in web mining to retrieve relevant information from hypertext document. This approach works on two level of data hierarchies, lowest level and higher level. The lowest level of database consists of semi-structured information of the hypertext document and higher level retrieved metadata information from lower level and organized them into object oriented database format.
- ***Query System****:* with the abundance of information on web, it is more difficult to extract desired information form web. It has become necessary to employ some valuable techniques so that vital information can be achieved. Web mining system utilizes standard database query languages such as SQl, W3QLand WEBLOG to retrieve information from dynamic and semi structured information sources.

2.2 Web Structure Mining

Web structure mining is the process of discovering knowledge from the interrelated hypertext document on the web. This process is based on the topology of hyperlinks, which consists of web pages as nodes and hyperlinks as edges. It is valuable technique to calculate the quality rank or similarity of each web page. Web structure mining performs on two levels: hyperlink and document structure.

2.2.1 Hyperlink

Hyperlink is a structure element, which connects a web page to different locations either within the same web page or different web pages [5]. The hyperlink is within the web document (Intra-document hyperlink) and hyperlink is within the web itself (Inter-document hyperlink). Hyperlink approach used hyperlink analysis tool to improve the search quality mechanism [13]. It also utilizes several link analysis models to examine the link quality such as PageRank, Weighted PageRank and HITS (Hypertext Induced Topic search) algorithms.

- ***PageRank***: is a metric for ranking hyperlink documents based on their quality and calculates the numerical weight of each hyperlink documents on the basis of citation analysis. Functionality of PageRank algorithm allied with link structure of web pages. If web page holds essential links towards it then the links of this page towards the other page are also to be viewed as important pages. In additional back-link utilized to measuring ranking score of PageRank algorithm. It gives a more advanced way to calculate the importance or similarity of a web page than simply counting the number of pages that are linking to it, which is called back-links. If a back-link appears from a relevance page, then that back-link is given a higher weighting than those back-links appears from non essential pages.
  The Eiq -1 is depicting the PageRank mathematical term [14].

$$PR(n) = A \sum_{X \epsilon F(n)} \frac{PR(x)}{K_X} \quad (1)$$

Where n represents the web page, F(n) is the set of pages that point to n, PR(n) and PR(x) are rank attains of page n and x. Kx specify the number of outgoing links of page x. A is an aspect used for normalization. In Eiq-1 show the direct link of www and maximum user do not follow that, so that modified version of this is depicted in Eiq-2 [14].

$$PR(n) = (1 - B) + B \sum_{X \epsilon F(n)} \frac{PR(x)}{K_X} \quad (2)$$

In Eiq-2 B is a depending factor that is frequently set to 0.85. B can be deliberation of as the viewpoint of users' following the direct links and (1 – B) as the page rank distribution from non-directly linked pages.

- ***Weighted PageRank***: algorithm is an extension of PageRank algorithm, which assigns the larger rank values to the most important pages rather than partitioning it evenly. Every out-link page attain their proportional value from in-link and out-link popular pages. The in-link page represented as $W^{in}(a,b)$ or out-link represented as $W^{out}(a,b)$ and $W^{n}(a,b)$ is a weight of link (a,b), a is number of in-links [16].

$$W^{in}_{(a,b)} = \frac{I_b}{\sum_{k \in R(a)} I_k} \quad (3)$$

$$W^{out}_{(a,b)} = \frac{O_b}{\sum_{k \in R(a)} O_k} \quad (4)$$

I_b, O_b and I_k, O_k are the numbers of incoming link of page b and k. P (a) is a references page of page a. In equation 5 show the best algorithm of PageRank:

$$WPR(b) = (1 - n) + n \sum_{a \in K(b)} WPR(a) W^{in}_{(a,b)} W^{out}_{(a,b)} \quad (5)$$

2.2.2 Document Structure

Document structure organizes the web contents in form of tree-structure format and used several HTML and XML tags inside the web pages [5]. Its main objective to automatically retrieving documents object model (DOM) from the entire documents. Document structure approach basically works in two forms:

- Structure Based Document Clustering [7]: used to extracting exact document structure information after clustering process.
- Discovery of browsing Pattern: technique used to analysis of hyperlink web pages, which accessed by different sessions and detecting the user interest.

2.3 Web Usage Mining

Web usage mining is an imperative technique to automatically discovery the user interaction patterns from web servers and predicts user behavior, when the user works with the web. It helps to determine nature of contents in which user are more interested. Most business organization and e-commerce website follows these rules for managing life time value of customer and provides best link according their browsing behaviors.

Web usage mining extracts information from server log, proxy log, browser log and organizational databases. Web server log maintains the history of page request and proxy server does work among client browser and web server. It consists of three phases such as preprocessing, pattern detection and pattern evaluation.

2.3.1 Data Preprocessing

Data preprocessing is performed on raw data to transform it into the data abstraction, which necessary for pattern discovery. It done in three forms: information cleaning, user & session determination and path examination.

a. ***Data Cleaning***: process is used to eliminate inconsistent or unnecessary items from web server logs.
b. ***User and Session Identification***: is used to identify the user and session log.
   - ***User Identification***: is concern to verify who access the website and which pages are accessed. This process complete in many ways such as IP address, Cookies and Direct Authentication. In [11] authors have elaborated the following step for user identification:
   - Dissimilar IP values in the IP address field characterize different users [17].
   - If the IP address is similar and different browser operation performed that indicates different users [4].
   - Similar user may stay web more than once at different point of time, so a time heuristic is used to divide those intervals into different user session.
   - ***Session Identification***: is process of identify the set of pages visited by the user within time of particular visit to a website [19]. It can be single or multiple sessions within that visit. Therefore user has been recognize, then done portioning of the session,

that is called sessionization (session-reconstruction) [19]. Sessionization can be done in three times, where two methods depend on time and one depends on navigation.

- Page Visiting Time: analysis number of page visited by particular user at a specific time.
- Page Stay Time: is used to summarize differences among two time stamps.
- Navigation Pages: process works on web pages connectivity.

c. ***Path Completion***: due to local caching, proxy servers, agent cache and "Post" techniques, some important web usage record does not recorded in log file. Their effect are making problem in study of browsing behavior of users. This problem can be resolved by inferring cache hits.

2.3.2 Pattern Discovery

Once user transaction has been verified, a variety of data mining techniques are applied to discover the user browsing patterns. These techniques are statistical analysis, association rules, clustering and classification.

- ***Statistical Analysis***: is an appropriate technique to retrieve information about visitors to a particular websites. It performs statistical analysis operation such as frequency, mean, medium, page view, viewing time and length on the basis of navigational path.
- ***Association Rules*****:** is discovery of those pages, which are most frequently referenced in a single server sessions.
- ***Clustering****:* is a method of group simultaneously a set of items having equivalent characteristics. It consists of two following types of clustering methods.
  - Usage Clustering [6]: used to determine those groups of users, which browsing pattern are similar. It helps to make a better market segmentation in e-commerce applications.
  - Page Clustering: process is determines groups of pages, which have similar contents.
- **Classification**: is an accomplish method to grouping the data into several predefined classes [17]. This is one of best technique for optimal future website, where developers make a particular content, which are belonging to a particular user class and categories. Classification can be done using decision tree classifier, naïve Bayesian classifier and K-nearest neighbor classifier.

2.3.3 Pattern Analysis

Pattern analysis approach comes to play an important role in web mining process for extracting user desired information from pattern discovery tool. The exact meaning of pattern analysis process is knowledge discovery phase. This approach used various knowledge extraction systems such as OLAP, Knowledge Query Mechanism, Visualization technique and smaller but more intelligent community of components know as intelligent.

- ***OLAP*** (Online Analytical Processing): is used for decision support system and extract knowledge from web server.
- ***Visualization***: is a vital approach for defining graphical tools and assigning color of different values.
- ***Knowledge Query Mechanism***: is a process of hold the feature of query languages such as SQL.
- ***Intelligent Agent***: performs much faster and handle more transaction in given time [20].

Upcoming section highlights the challenges that arise in web mining system.

III. CHALLENGES

The upper section emphasizes the detail that the web mining system has comes a long way and it is broadly utilized in dissimilar areas of research. A summarized view at the preceding section reveals the unfolded challenge which is listed below:

3.1 Unstructured Information

To managing the unstructured information on the web is one of the key unsolved issues among web mining system. The main cause for this unanswered problem is their weak operational techniques, means those systems and related tools, which have established to successfully converting structured information into knowledge intelligence that systems are ineffective when we use to execute the same on unstructured information.

The word unstructured information elaborates in different ways such as in the context of relational database systems, it refers to data have not stored in rows and columns. In information system, it refers to mechanize information that does not have any pre-defined structure and cannot directly utilize by computer system. The most of data in business organization exits as unstructured format, usually illustrating in e-mail, blog, discussion forum, wikis, official and business management. The data in the type of audio and tape files is presented in enormous amount across the world. There is no clear pattern existing while we browse these files. This issue must be addressed so that unstructured information can share and collaboratively exploit distributed, heterogeneous knowledge in a scalable way.

3.2 Semantic Interpretability

Web mining is a knowledge retrieval infrastructure, which exploits data mining technique to automatically extract desired information in heterogeneous environment such as web. Currently human are able to understand unstructured query on web on the basis of their past experience and other side software agents are used by web mining system to searching and retrieving information on web.

However certain challenges are inhibit in the implementation of software agents in web mining infrastructure and one of the most noticeable challenges being the issue of semantic interpretability. Semantic interpretability rises question, how to enable agents to acquire consistent knowledge from other agents, those agents are doing work in dissimilar domains and formulating use of different ontologies. Here problem arises when different ontologies used by different domain agents, e.g some agents are working in medical domain and other agents are working in financial domain, how they will share common knowledge among them. Several kinds of semantic interpretability need to be handled such different language representation, lack of communication protocol and ambiguity. This issue needs to be taken care of. Next section defines the working of web mining system.

IV. SEMANTIC WEB MINING (SWM)

The term semantic web implies an intelligent web, which process the information not only for human but also for computers, where machines can interpret and exchange information on the web and raising the probability of relevant information retrieved. The initial goal of semantic web formulates service information more supportive and enhances search mechanisms thus resulting in user satisfaction. However this vision is more complicated due to machine don't have the kind of vocabulary that people have.

The current web structure is a huge repository of unstructured information that usually understood by human being only. In these circumstances semantic web plays an important role by providing machine interpretable semantic that makes the data machine understandable and Web mining addresses the semantic interpretability issue by using (automatically) semi-structure technique to extract hidden knowledge from vast amount of web data [16]. The combination of these two massive approaches make new web infrastructure that is called Semantic Web Mining (SWM). The figure 1 is characterizing the working of SWM. SWM works on the three bases, first phase facilitates the semantic interpretability among unstructured data such as -mail, blog, discussion forum, and wikis and utilizes intelligent approaches to eliminate the inconsistency [10]. In second phase is used to retrieving knowledge from the interrelated hypertext documents. It is a process of study design of conceptualization and ranking hyperlink documents based on their quality. Third phase utilizes to extract desire information from knowledge discovery phase and gave agreeable results. The comprehensive description of SWM has given below:

4.1 Content Extraction

Content extraction is the process of extracting semantic information from the unstructured data such as email, images, audio, video and presentation. Content data corresponds to the collection of facts a web page was proposed to communicate to the users and utilizes intelligent agent (AI) and Information Retrieval (IR) approaches to improve the information finding and filtering the information from the users inferred or solicited profiles. IR employs to summarize the data on the Web and to combine them so that more complicated queries other than the keywords based search could be performed. After this process knowledge discovery phase are implemented for extracting valuable information.

4.2 Knowledge Discovery

Knowledge discovery is the process of discovering knowledge from the interrelated hypertext document on the web. This process is based on the ontology matching and PageRank algorithm.

- ***Ontology Matching***: finds correspondence between semantically related entities of ontologies and determine the set of synonym concepts which are parallel in meaning but have different names or structures [21]. These correspondences can be used for various tasks, such as ontology integration, question answering, data conversion, etc. It is a competent technique to study design of conceptualization, where conceptualization is an apparition of different words, which reveal the objects and their relationship with other entities. When that finds the semantic correspondence among two ontologies, it computes their weight by PageRank algorithm.
- ***PageRank***: is a metric for ranking hyperlink documents based on their quality and calculates the numerical weight of each hyperlink documents on the basis of citation analysis. It helps to analysis of hyperlink web pages, which accessed by different sessions and detecting the user interest. After this process knowledge discovery phase measure those web pages, which are most relevant to the user query.

4.3 Information Retrieval System

Final phase of SWM system is obtaining relevant information, which are achieving through OLAP, Agent system, query languages and data mining.

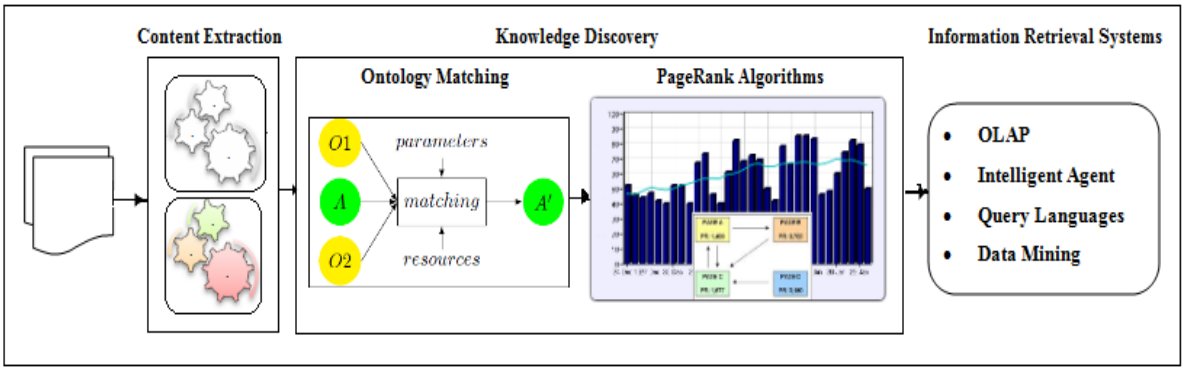

Figure 1: Semantic Web Mining Infrastructure

V. CONCLUSION

In this work we have tried to throw light on concepts of web mining system with brief survey of its mining systems and their problem domain. Web mining is vital future of World Wide Web, where it provides lot of feature for making knowledge oriented web pages according to user browsing behavior. The objective of this work is exploits the web mining technique in wisdom web that is our future work.

REFERENCES

[1] C. M. Brown, B. B . Danzig, D. Hardy, U. Manber, and M. F. Schwartz, (1994), The harvest information discovery and access system. In Proc. 2nd International World Wide Web Conference, pp 182-188.

[2] COOLEY, R., TAN, P-N., AND SRIVASTAVA, J. (1999), WebSIFT: The web site information filter system. In Proceedings of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Boston, August).

[3] D. K. Giord, (1996) Hypursuit: a hierarchical network search engine that exploits content-link hpertexxt clustering. In The Seventh ACM Conference on Hypertext, pp 180-193.

[4] Dong D, (2009), "Exploration of Web Usage Mining and its Applications", IEEE, pp 1-5.

[5] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, (2005), "Web Mining - Concepts, Applications & Research Directions", Studies in Fuzziness and Soft Computing Volume 180- Springer, pp 275-307.

[6] Jaideep Srivastava, Robert Cooley, (2000), "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD, pp 1-12.

[7] Ke Wang, Liu, "Discovery of Typical Structures of Documents: A Road Map Approach", SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp146-154.

[8] Kosala R and Blockeel H, (2000), "Web Mining Research: A Survey", ACM SIGKDD, pp 1-15.

[9] M. Balabanovic, Yoav Shoham, and Y. Yun, (1997), "An adaptive agent for automated web browsing. Journal of Visual Communication and Image Representation- ACM", pp 127-158.

[10] Manoj Manuja & Deepak Garg, (2011) Semantic Web Mining of Un-Structured Data: Challenges And Opportunities, International Journal of Engineering (IJE), Volume (5) : Issue (3), pp 269-276.

[11] Neelam Tyagi, Simple Sharm, (2012), "Comparative study of various Page Ranking Algorithms in Web Structure Mining (WSM)", IJITEE, pp 14-19.

[12] P. Ravi Kumar and Ashutosh Kumar Singh, (2010), "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval", American Journal of Applied Sciences, pp 840-845.

[13] Prasanna Desikan, Jaideep Srivastava, Vipin Kumar, and Pang-Ning Tan, (2002), "Hyperlink Analysis: Techniques and Applications", pp 1-42.

[14] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell, (1995), Webwatcher: A learning apprentice for the world wide web. In Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments.

[15] R. Cooley, B. Mobasher, and J. Srivastava, (1997), "Web Mining: Information and Pattern Discovery on the World Wide Web", Ninth IEEE International Conference – IEEE, pp 1-10.

[16] Rana V, Singh G, (2014), "An Analysis of Semantic Heterogeneity Issues and their Countermeasures Prevailing in Semantic Web", International Conference on Reliability, Optimization and Information Technology - ICROIT 2014, IEEE Xplore, pp 80-85.

[17] Singh A, Mishra R, (2012), "EXPLORING WEB USAGE MINING WITHa SCOPE OF AGENT TECHNOLOGY", IJEST, pp 42834289.

[18] U. Shardanand and P. Maes, (1995), Social information filtering: Algorithms for automating

"word of mouth". In Proc. of Conference on Human Factors in Computing Systems (CHI-95), pp 210-217.
[19] V.Chitraa, (2011), "A Novel Technique for Session Identification in Web Usage Mining Preprocessing", IJCA, pp 23-27.
[20] Vijay Rana and Singh G, (2013) "Evaluation of an Intelligent Approach for Semantic Web", IJCT, pp 478-482.
[21] Vijay Rana, Singh G, (2014), "Analysis of Ontology Matching System and their Countermeasures", International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT -2014, IEEE, pp 86-92.
[22] Wenpu Xing and Ali Ghorbani, (2004), "Weighted PageRank Algorithm", IEEE, pp 1-10.