SEMANTIC WEB MINING: OPTIMUM APPROACH

Mamta, Dr. Vijay Rana

Deptt of Computer Science , Arni University

1. INTRODUCTION

With the explosive growth of the web, the users need assistance of search engines and automatic personalization tools for discovering relevant content. In particular, a web user can find documents about a specific topic by submitting a query to a search engine and clicking on the results [26]. Web search is helpful when the users have a clear idea of what they want; however the existing centralized platform and usually a large server can't ensure the scalability, non-redundancy and interpretability of information provided to users [21]. To achieving this interoperability between dissimilar information systems is extremely tedious, complex and error-prone task. *Therefore the need for research activities in web management & enhancements by developing a standard, flexible but intelligent, adaptive and distributed framework for the support of heterogeneous infrastructure is apparent.* The current scenario demands the delegation of intelligence of web to a smaller but more intelligent community of components known as Web Mining and Semantic Web.

In the last few years, a lot of attention has been devoted to improving web search and recommender systems through data mining. It allows sifting through large quantities of data for useful information. Web mining refers to the process of inferring knowledge from web data [16]. It is a multidisciplinary effort that inherits concepts and techniques from fields such as information retrieval, statistics, machine learning, and others. 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 [10. 26]. 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.

Although the Web Mining has been the vision for the next generation of the web, where information is desired to be useful not only for the people but also for the computers but one major obstacle in implementing Web Mining is that machines don't have the kind of vocabulary that people have [13, 20]. Since people make use of language from the very early years of their lives so it easy for them to make connections between different words and concepts and to make inferences based on contexts. But this is not the case with computers. To make computers able to understand meaning of words and relationship between different words, they must have documents describing all the words and logic to make necessary connections. Web Mining achieves this goal through Semantic Web, where Semantic Web (SW) implies an Intelligent Web i.e. a meaningful web [4, 5, 6]. It aims to make computers understand the meaning of information on the web pages rather than merely presenting them to users [28]. The idea is to make World Wide Web (WWW) intelligent and machine readable by providing tools to find, exchange and interpret information to a limited extent by adding metadata.

Our research work is related to the field of Semantic Web Mining. In particular, we analyze data stored in web search engines' logs to discover usage patterns, and the aim is to enhance performance of search tools as well as to help users to find information on the web.

2. SEMANTIC WEB

The Semantic Web (SW) aims towards transformation of information oriented web into knowledge oriented web. The semantic web is a vision which extracts information from the web and crafts it possible to facilitate machines to identify intricate human queries. It brings the proposal of structuring information accessible across the web in a significant way enhancing search technique and thus resulting user satisfaction [1, 2].

The idea of SW was coined by Tim-Berner Lee in 'The Scientific Americans' [9] in 2001. The article described the evolution of a web that comprised largely of documents containing data and information for computers to manipulate. Since the time of inception and publication of the original SW vision by Lee et.al [9]. Berners-Lee proposed four versions of Semantic Web architecture. All of the reference architecture versions were presented by Berners-Lee in different presentations; they were never published in literature or included as part of World Wide Web Consortium (W3C) Recommendation. Table 1 given below provides brief description of the various layers of the semantic web architecture.

Table 1: Description of layer in Semantic Web Architecture

Name of layer	Description
Unique Identification Mechanism	It ensures that the terms and concepts are tied to a unique definition and refer to an address on the web.
Syntax Description Language	This layer contains languages necessary to make information available on web pages, machine readable. XML (eXtensible Markup Language) is one of the most popular SDL, since it has been recommended as the language for the publication of web document in W3C-2006.
Meta-data Data Model	This layer is based on RDF which is framework for describing information about resources on the web, using XML tags. RDF/RDF Schema facilitates semantic interoperability. Recently W3C has recommended RDF to standardize the definition and use of metadata descriptions of Web-based resources
Ontology	This layer contains the ontology of the domain under consideration, where ontology is an explicit specification of conceptualization. Ontologies provide a common platform for understanding of topics and communication between people and application systems.
Rules	This layer contains rules for converting a document from one RDF schema into another one. This layer includes inference rules without having negation. Using inference rules, we can draw inferences on similarity of two properties, defined in two different schemas.
Logic Framework	This layer provides facility of writing logic into documents thus provides rules for deduction of one type of document into another type. This layer includes Predicate Logic and Quantifiers so as to facilitate deductions. Knowledge Interchange Format (KIF) is the language used to specify logic in this layer.
Proof	This layer involves proof checking for validation of identity. The proof will be a chain of assertions and reasoning rules with pointers to all the supporting material.
Trust	This layer requires that the reasoning system must include signature verification system. This will result into a system which can express and reason about relationships across the whole range of public key based security and trust systems.

3. WEB MINING

Data mining is the computational process of discovering interesting patterns in large datasets. It allows the non-trivial and automatic extraction of implicit and potentially useful information from very large amount of data. With the explosion of content and services available online, the web has recently become a fertile area for data mining. Web mining consists in the application of data-mining techniques to data, artifacts, and activities related to the web [3, 15]. It is a dynamic and wide research area, which draws methodologies from statistics, database, information retrieval, and some branches of artificial intelligence such as machine learning and natural language processing (NLP). Web mining is generally divided into three main subareas [27], corresponding to three different knowledge-discovery domains.

As shown in figure 1, SWM technique can be classified into three forms i.e. Web Content Mining (WCM), Web Structure Mining (WSM) [25] and Web Usage Mining (WUM).

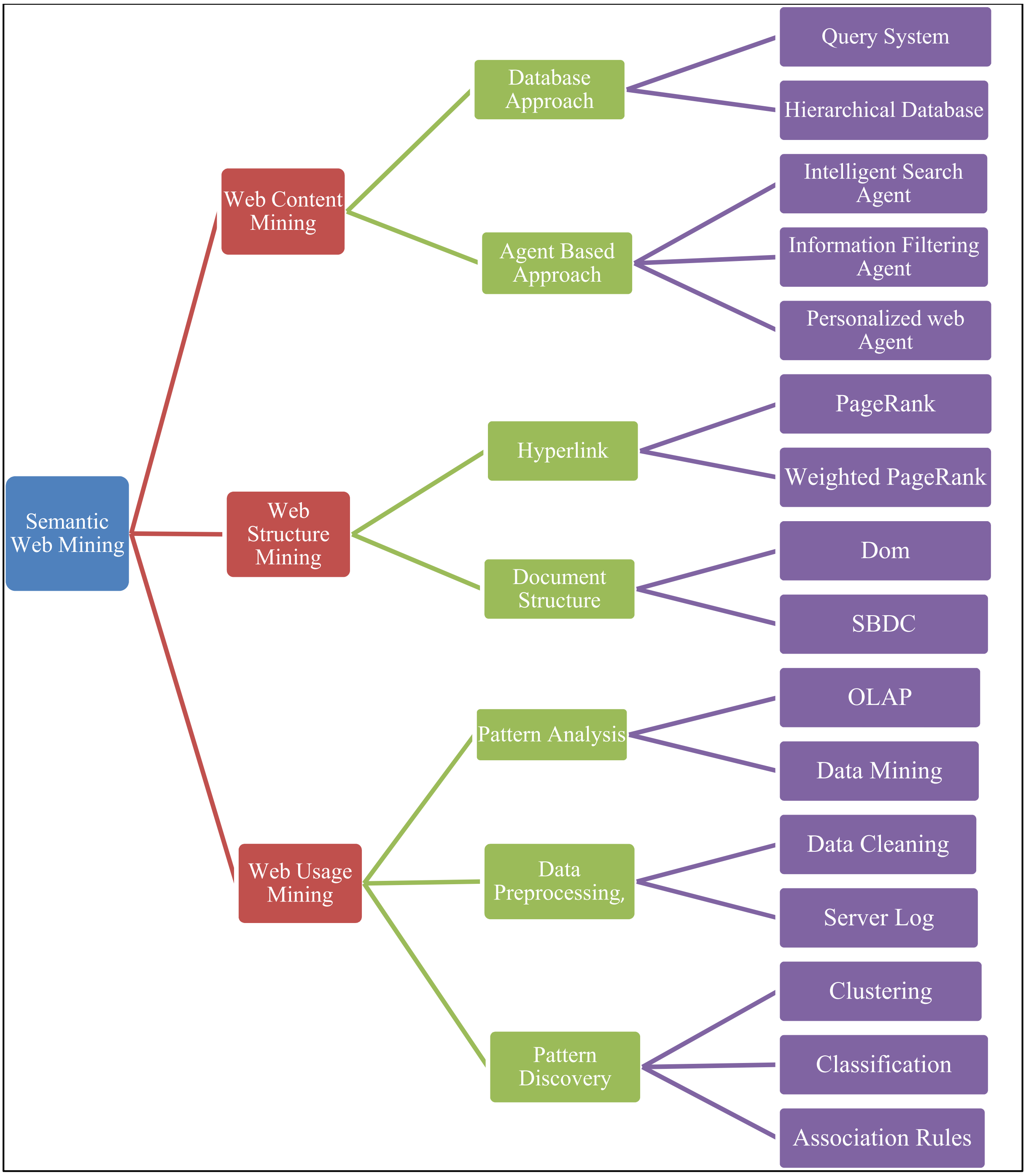

Figure 1: Classification of SWM

3.1 Web Content Mining (WCM)

WCM is a preeminent technique to find valuable contents and documents from the web. The web contents are describing in two forms: text and multimedia contents. Text content further consists of semi-structure content likes HTML data and unstructured text [25]. On other side multimedia content contains of picture, sound, tape and structured articles that provide semantic description. The current development of WCM technique have encouraged developers to make more intelligent approaches for knowledge accessing, such as Information Retrieval (IR) [7] and Database approach (DB) [14]. IR uses intelligent agent approach [17] to enhance the information searching and extracting the information from the users inferred or solicited profiles. DB uses database approaches to determine the data on the web and integrate them so that more difficult queries could be searched. The rapid growths in WCM techniques have allowed system to increase knowledge deliverance of information through combining of several approaches such as agent based approach and database approach [17].

3.2 Web Structure Mining

Web structure mining is the way of discovering valuable information from the interconnected hypertext document on the web. This technique is work with the topology of hyperlinks consisting of web pages as nodes and hyperlinks as edges. It is appropriate technique to calculate the relatedness of each web page [19]. Web structure mining executes on two phases: hyperlink and document structure.

3.3 Web Usage Mining

Web usage mining is an innovative approach to automatically identify the user interaction patterns from web services and measures user behavior, when the user works on the web. It helps to identify type of contents in which user are more interested. Today various business firms and e-commerce societies are follows these rules for evaluating life time value of client and gives better link according their browsing behaviors. Web usage mining retrieves desire knowledge from server log, proxy log, browser log and managed databases. Web server log contains the history of page log and proxy server executes between customer browser and web server. It works on three forms such as data pre-processing, pattern discovery and pattern evolution.

4. THE INESCAPABLE RELATIONSHIP BETWEEN SEMANTIC WEB AND WEB MINING

With the large amount of information available on the web, it is more complicated to extract desirable information from the existing knowledge system. At that instances semantic web and web mining techniques plays an important role to mining valuable information form web. Semantic web is an extension of current web and web mining technique is an extension of data mining technique. Semantics can improve the results of Web Mining by taking advantage of structures in the Web. Web Mining can improve the Semantic Web by finding new semantic structures to enrich the semantics. The central idea of this work is to proposed new innovative model that is called semantic web mining (SWM).

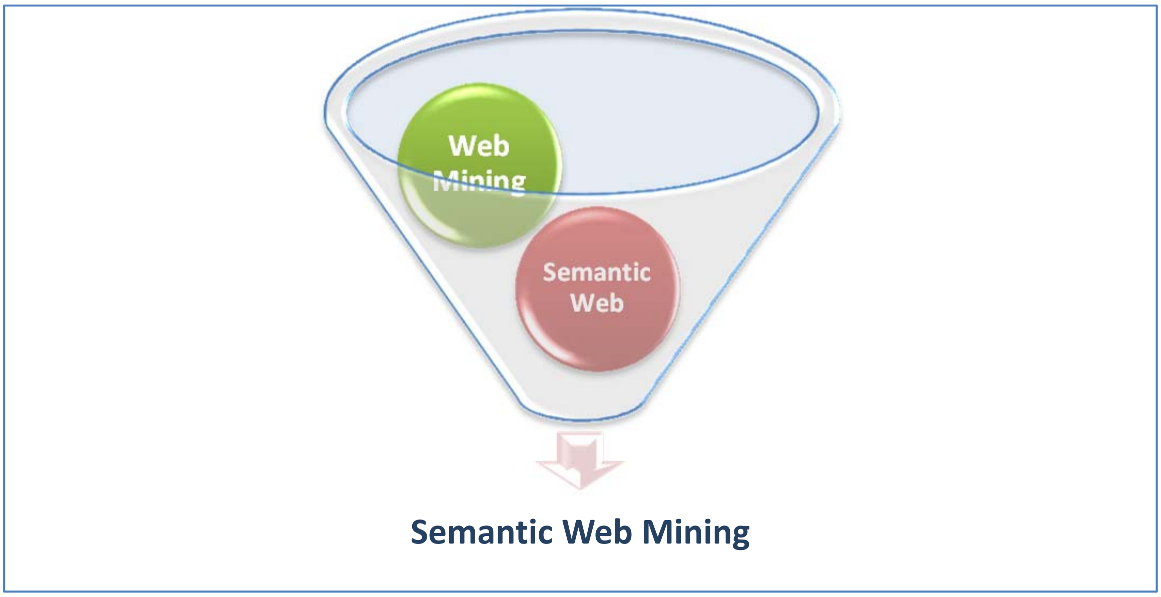

Figure 2: High Level View of semantic web mining

The objective of SWM is to get a higher understanding of user behavior at the time of Web surfing in order to better support for the users on the Web. The Semantic Web mining offers to add structure to the Web, while Web Mining can learn implicit structures. This is an interesting way for Semantic Web Mining to create itself as the dependence between the Semantic Web and Web Mining increases. The resulting research benefits many areas of industry such as “e-activities”, health care, privacy and security, and knowledge management and information retrieval. Semantics can improve the results of Web Mining by taking advantage of structures in the Web. Web Mining can improve the Semantic Web by finding new semantic structures to enrich the semantics. The application of each area to the other creates a feedback loop, where the goal of Semantic Web Mining is realized. This in turn will create a more usable Web and may help in transforming the Web into the Semantic Web

5. LITERATURE SURVEY

Extensive research has been done in the area of semantic web, ontology and agent technology. This section highlights the work of eminent researchers and explores the challenges, which still need to be addressed.

Tim Berners-Lee et.al [9] in their visionary article laid down the foundation of Semantic Web. They gave a new direction for the information oriented WWW to be knowledge oriented in future. Their work enlightened the powerful role of agents and ontologies in semantic web.

Rana and Juneja proposed an ant based system for semantic web [26], where ants are hypothetical sophisticated agents that carry information on behalf of its users. This work provides a brief introduction of semantic web and ant agent. The inspiration of ant agent has been obtained from natural insects because ants have logical abilities to retrieve relevant results on distributed environment like web. The main objective of this work was implementation of semantic web by deploying ant behavior as a software program to find meaningful information on web.

Rana and Singh [29] have proposed a semantic web mining interface, which is competent to handle heterogeneity issue and provide meaningful information in non-redundant way. Their work focused on finding the most ambiguous words and finding the relatedness measure with other important keywords in the query. It also considered removing redundancy among keywords being matched, but performs little effort to enhance the social web and e-commerce activities.

Srivastava et.al in [16] described web mining technique that exploits the data mining approach to automatically find and retrieve desire information from the web and gave agreeable outcome to users. Further, authors divided a web mining technique into three basic forms content, structure and usage mining to extract meaningful information. Content mining is a superlative technique for retrieving meaningful information from the content of web resources. Structure mining is the process of retrieving knowledge from the interrelated hypertext on the web and usage mining an imperative technique of automatically searching and analysis the user interaction patterns with web servers. The main idea of their work is to provide summarized review of web mining system with their application areas.

Gracia et al. [18] proposed web based semantic relatedness technique that numerically computes the degree of semantic relatedness between different ontology terms. The authors utilize Normalized Google Distance (NGD) [11] measure to compute the relatedness degree of co-occurrence of words on web pages. The central objective of their work is to enhance desirable properties for maximum coverage and universality on web search results. However, this approach still requires developing tools for dynamic configuration of new results into it.

Singh proposed an Agent Based Semantic Web Mining (SWMS) [28] system that provides knowledge based response to the user. The work considered web mining technique with agent-based control for extracting valuable information from web contents. The web mining is performed at three levels: web structure mining, web content mining and web usage mining. Web Structure Mining is the approach used to evaluate the links among different web pages and web sites. Web Content Mining is used to retrieve knowledge from the web contents and it is also used to summarize, classify and cluster the web contents and web usage mining digging the usage of web contents from the logs maintained on web servers.

A knowledge based system [20] that could handle the semantic heterogeneity by using semantic, name and statistical techniques was proposed by Maree and his team. The key idea behind this work was to find semantic correspondence between the entities of inconsistent ontologies. Their work also enlightened the powerful role of semantics in ontology matching. This system lacks analyzes in terms of its efficiency& usability in various domains of practical interest.

A critical look at the above literature highlights the fact that there are numerous debatable issues which still need to be researched upon. The upcoming section lists few such issues.

REFERENCES

[1] Adam Pease, Ian Niles, John Li, (2002), The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications, AAAI Technical Report WS-02-11, pp 11-15.

[2] Aghaei S, Nematbakhsh M and Farsani H, (2012), Evolution of the World Wide Web: From Web 1.0 To Web 4.0, Vol 6, IJWesT, pp 1-10.

[3] Agirre, Edmonds. P, (2007). Word Sense Disambiguation: Algorithms and Applications, Springer Publishing Company, Incorporated 2007, New York, pp 1-364.

[4] Agostino Poggi and Michele Tomaiuolo, (2013), A DHT-based Multi-Agent System for Semantic Information Sharing, Studies in Computational Intelligence Volume 439- Springer, pp 197-213.

[5] Alain Leger, Lyndon J.B. Nixon, Pavel Shvaiko, Jean Charlet, (2005), Semantic Web applications: Fields and Business cases. The Industry challenges the research, IFIP-The International Federation for Information Processing Volume 188- Springer, pp 27-46.

[6] Andrea Moro and Roberto Navigli, (2015), SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking, SemEval 2015, Association for Computational Linguistics, pp 288-297.

[7] Andreas Hotho, Robert Jaschke, Christoph Schmitz and Gerd Stumme, (2006), Information Retrieval in Folksonomies: Search and Ranking, Proceeding ESWC'06, 3rd European conference on the Semantic Web: research and application, pp 411-426.

[8] Armando Vieira Redzebra Analytics, (2016),Predicting online user behaviour using deep learning algorithms, arXiv:1511.06247v3, pp 1-21.

[9] Berners T, Hendler J and Lassila O, (2001), “The Semantic Web”, Scientific American: Feature Article, Vol.284, No5. pp. 1-4.

[10] Bohanec, M, (2001), What is decision support?, Proc.Information Society IS-2001: Data Mining and Decision Support in Action, (eds. Škrjanc, M., Mladenić, D.), Ljubljana, pp 86-89.

[11] Cilibrasi, R.L., Vitanyi, P.M, (2007), The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3), pp 370–383.

[12] Gen Teck Wei, Shirly Kho, Wahidah Husain, (2015), A Study of Customer Behavious Through Web Mining, Journal of Information Science and Computing Technologies, Vol.2, No.1, pp 103-107.

[13] Gerd Stumme, Andreas Hotho and Bettina Berend, (2006), Semantic Web Mining State of the art and future directions, Journal of Web Semantics – Elsevier, pp 124-143.

[14] Hai Leong Chieu and Hwee Tou Ng, (2002), A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text, AAAI, 2002, pp 786-791.

[15] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, (2005), Web Mining - Concepts, Applications & Research Directions, Foundations and Advances in Data Mining Volume 180 of the series Studies in Fuzziness and Soft Computing, pp 275-307.

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

[17] Jin Zhao, Min-Yen Kan and Yin Leng Theng, (2008), Math Information Retrieval: User Requirements and Prototype Implementation, ACM, 2008, pp 16-20.

[18] Jorge Gracia and Eduardo Mena, (2008), Web-Based Measure of Semantic Relatedness, WISE 2008, LNCS, Vol.5175, pp. 136–150.

[19] 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.

[20] Mohammed Maree and Mohammed Belkhatir, (2015), Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies, Knowledge-Based Systems, Vol 73, No. 3, pp199-211.

[21] 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.

[22] P.Saravana kumar/ R.Iswarya, (2014), Predictive Analysis of Users Behaviour in Web Browsing and Pattern Discovery Networks, International Journal of Latest Trends in Engineering and Technology (IJLTET), pp 239-245.

[23] Paul Warren, (2006), Knowledge Management and the Semantic Web: From Scenario to Technology, IEEE Computer Society, pp 53-59.

[24] Pedersen T, Lexical Semantic Ambiguity Resolution with Bigram-Based Decision Trees, Lecture Notes in Computer Science Volume 2004, 2001, pp 157-168.

[25] R. Cooley, B. Mobasher, and J. Srivastava, (1997), Web Mining: Information and Pattern Discovery on the World Wide Web, ICRI-1997, LNCS-1836, pp 1-10.

[26] Rana V, (2012), Blueprint of an Ant-Based Control of Semantic Web, Vol. 2, No. 4, IJoAT, pp. 603-612.

[27] Shrutilipi Bhattacharjee and Soumya K. Ghosh, (2015), Measurement of Semantic Similarity: A Concept Hierarchy Based Approach, International Conference on Advanced Computing, Networking and Informatics, pp 407-416.

[28] Singh A, Basim Alhadidi, (2013), Knowledge Oriented Personalized Search Engine: A Step towards Wisdom Web, International Journal of Computer Applications, Vol.76, No.8, pp. 1-9.

[29] Vijay Rana, Singh G, (2014) “Analysis of Web Mining Technology and Their Impact on Semantic Web”, CIPECH -2014, DOI: 10.1109/CIPECH.2014.7019035, IEEE Xplore, pp 5-11.

[30] Y Gupta, A Saini, AK Saxena , (2015), A new fuzzy logic based ranking function for efficient information retrieval system, Expert Systems with Applications, Elsevier, pp 79-86.

[31] Yan, T.W., Jacobsen, M., Garcia-Molina, H., and U. Dayal (1996), ―From User Access Patterns to Dynamic Hypertext Linking‖, Proc. 5th International World Wide Web Conference.