2014 Fourth International Conference on Communication Systems and Network Technologies

Semantic-Synaptic Web Mining: A Novel Model for Improving the Web Mining

Hiteshwar Kumar Azad
Dept. of CSE, NIT Patna
Patna, India
azad07it17@gmail.com

Kumar Abhishek
Dept. of CSE, NIT Patna
Patna, India
kumar.abhishek@nitp.ac.in

Abstract— **Web mining is the application of data mining technique to automatically discover and gathered information from web documents and services which can be in structured, unstructured or semi-structured form. It is used to understand user behaviour, evaluate the effectiveness of a particular web and find out the relevant and efficient results from the web. Accuracy and relevance of information extracting from the web is the most significant issue of concern for the realization of web mining. The idea is to improve the accuracy and relevance of information extracting from the web. This paper proposes a novel model for improving the web mining, we hypothesize that web mining is Semantic-Synaptic web mining. Semantic-Synaptic web Mining interlinks the web of data to different data sources at low entropy (Information Theory). This paper combines the best ideas from the semantic web and synaptic web at low entropy and constructs the architecture of Semantic-Synaptic web mining.**

Keywords-Web mining; SemanticWeb Mining; Synaptic web; Synaptic web Mining ;

I. INTRODUCTION

The past few decades, have seen an explosive growth of information available on the web. Today web has turned to be the largest information sources available in this planet. Absolutely web is a huge, explosive, distinct, influential and mostly unstructured data storehouse, which provide incredible amount of information. The tremendous growth of data sources available on the web and information overloading makes information retrieval a tedious and difficult for users to access relevant information efficiently. The user is only concerned about a small but relevant information of the web and not interested in the rest of the information contained in web, because the expected search results will be immersed by the traditional search engines which are based on the keywords; other, since the majority of web contains huge collection of unstructured data, which lead to the traditional data mining results will be unsatisfactory. Several web mining techniques have been proposed in order to resolve these problems and enhance the efficiency of information rehabilitation and relevance of information extracting from the web. The techniques aim to develop a new methodology to effectively extract and mine useful information from these web pages and allows user to easily locate intended knowledge or information from huge data on the web. The accuracy and relevant information extracting from the web still faces a big challenges. The web mining technique has been an active and popular research field. It is a converging area from several research association, such as information retrieval, Database technologies, Artificial intelligence, Machine learning and also psychology and statistics as well.

In this paper, first traditional web mining is discussed and then semantic web mining, a more valuable web mining which have the better improved and efficient result. Lastly a discussion is made for a proposed novel model which portrays the present, past and future of web mining and gives a most effective idea for accuracy and relevance of information extracting from the web, called semantic-synaptic web mining. Finally construction and explanation of the proposed semantic-synaptic web mining architecture is made.

II. RELATED WORKS

Oren Etzioni first came up with the term of web mining in his paper [1]. Oren Etzioni pointed out a question: whether effective web mining is feasible in practice? He also suggested dividing the web mining into three processes. This paper opened up a new progressive research field. R.Kosala etal [2] performed research in the area of web mining and suggested dividing the web mining to three categories depending on kind of data to be mined. The authors of [3] claims the web involves three types of data: data on the web (content), web log data (usage) and web structure data. The author of [4] categories the data type as content data, structure data, usage data and profile data. While the author of [5] classified the web mining into web usage mining, web text mining and user modeling mining. Some researchers combine the content and structure mining to improve the web mining technique's strengths as mentioned in [6][7][8]. F. Sebastini [6] and S. Chakarbarti[7] talks about web content mining in details, and J. Furnkranz[8] has done surveys work in web structure mining. Even though majority of researchers don't agree with their classification. However today the most recognized categories of web mining are web content mining, web structured mining, web usage mining. In 2001 Berners-Lee came up with a novel ontological approach which makes the web more semantic and more understandable to machine is known as Semantic web [9]. Semantic web is emerging as the next generation web, with a semantically rich language such as ontology which provides the most relative data and well defined web structure. B. Berendt and colleagues [10] discuss the semantic web and web mining can fit together: web mining results help to build

978-1-4799-3070-8/14 $31.00 © 2014 IEEE
DOI 10.1109/CSNT.2014.96

the semantic web and the semantic web improve the infrastructure and effectiveness of web mining. In 2006 Berners-Lee turn up with the principles of Linked data [11] and provide ample guidelines on how to uses the standard web technologies to links between data on the web from different sources and which data publishers have initiated to recognized the web of data. This guidelines widened by technical documents [12][13]. In practice linked data are unable to provide the most relevant and efficient data over the web because when user want to find out a particular data on the web, it results in many linked data at one instant of time, which would be much noisy and irrelevance from our choice. There has been work done so far on the Linked Data [14][15][16][17] but none of these have talked about noisy and irrelevancy of linked data.

III. TRADITIONAL WEB MINING AND SEMANTIC WEB MINING

A. Traditional Web Mining

The term Web mining was coined by O. Etzioni [1] in 1996 to designate the use of data mining techniques to automatically discover web documents and services, deduce information from web resources and bring to light a general pattern on the web. Over the years, web mining research has been extended to cover the use of data mining and similar techniques to discover resources, patterns and knowledge from the web documents and services.

Web mining is not simply the application of information retrieval and text mining techniques to the web pages, it is also pertains non-textual data such as web server logs and other transaction based data. Chen and Chau [18] stated that web mining depends heavily on data mining and text mining techniques, but the techniques employed for web mining are not totally based on data mining or text mining. Techniques, such as web link structure analysis is unique to web mining. Web mining can be considered as a subfield of data mining but not as subfield of text mining because some web data are not textual.

TABLE I CLASSIFICATION AND APPLICATION OF RETRIEVAL & MINING TECHNIQUE

Purpose→ Information Sources↓	**Retrieving known data or documents effectively & efficiently**	**Finding new patterns or knowledge previously unknown**
Any data	Data Retrieval	Data Mining
Textual data	Information Retrieval	Text Mining
Web related data	Web Retrieval	Web Mining

Web mining can be viewed as consisting of following subtasks [1] :

(a) Information Retrieval (Resource Discovery): the task of automatically retrieve all relevant information from web.

(b) Information Extraction and Pre-processing: automatically extracting and pre-processing specific information from the retrieved web resources without human interaction.

(c) Generalization: automatically determine general patterns at individual web sites and across multiple sites.

(d) Analysis: validation and analyzing the mined pattern.

Today the most recognized categories of web mining are:

i. Web content mining
ii. Web structure mining
iii. Web usage mining

Web content mining basically deals with extracting useful information and knowledge from contents/ data/ documents/ services. The web document mainly contains various types of data, such as text, image, audio, video, metadata and hyperlinks.

Web structure mining is used to generate structure summary about the web site and web page. It tried to discover the link structure of the hyperlinks at the inter document structure unlike web content mining, which pertains to intra document structure. Based on the analysis of the hyperlinks, web structure mining will categorize the web pages and create the information, such as the connection and similarity between web sites.

Web usage mining deals with behaviour, when users interact with web. It tries to discover the meaningful information from the insignificant data derived from the interactions of the users while surfing on the web. Web usage mining is sometimes referred as log mining because it pertains mining the web server logs. Web usage mining is categories in three distinctive phases: Pre-processing, Pattern discovery and Pattern analysis.

B. Semantic Web Mining

Semantic web mining is a sequence of improving the quality of web mining, in which we combine the two fast developing research areas: Semantic web and Web mining. [10] discuss the semantic web and web mining can fit together. Web mining enables the semantic web vision and the semantic web technique improves web mining's effectiveness. Semantic web is designated to extend the current web by providing well defined meaning of information to grant better co-operation and machine understandable for computer to process.

The semantic web [9] was first outline by T. Berners-Lee to makes the web more semantic and more understandable to machine. They defined the semantic web as "Semantic web is a web of data that can be processed directly and indirectly by machine and described the relationship between things and the property of things". According to Berners-Lee, Semantic web is a multi-layered architecture, in which each layer gradually increases functionality and lower layer support the upper one, as shown in Table II.

TAB LE II SEMANTIC WEB ARCHITECTURE

Layers	Name	Description
Layer 7	Trust	Establishment of trust connections between users.
Layer 6	Proof	Basis of logic, it authenticates statements in order to connection.
Layer 5	Logic	Provides the axioms and inference rules and basis for the intelligence service.
Layer 4	Ontology vocabulary	A richer language for describe the various types of resources and the connections between resources.
Layer 3	RDF+ rdfschema	Used to describe the resources on the web and types.
Layer 2	XML+NS+XMLschema	Used to represent the information content and structure.
Layer 1	URI and Unicode	URI provide the standards for identifying and locating resources. Unicode processing resource to encoding.

Semantic web describes the purpose of semantic knowledge, which can be applied to different web mining.

IV. PROPOSED MODEL

The enormous amount of unstructured information available on the web, results in inaccuracy and irrelevant information while user extract desired information from the web. For resolution of this issue, this paper proposed a novel model, which is based on mainly three things: First **Semantic web**, a technique to manage content and process with creation and use of semantic metadata. Second **Synaptic web**, synapse is a biological term, it is the connection between different neurons in the brain, same as in the synaptic web like the human brain the synaptic connections between objects (Content/ Information) are more important than the object themselves makes the smarter web.

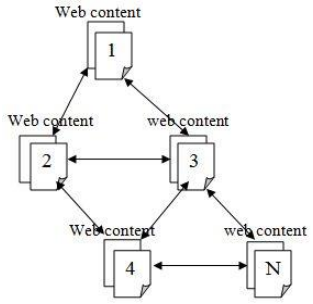

Fig.1 Synaptic Web

Third and last **Entropy**, in information theory the term generally refers to the Shannon entropy [19], is a measurement of uncertainty and inconsistency in random variable, which evaluate the information content in a message. Mostly, the information content [20] is the uncertainty of each event. For measuring the semantic similarity in taxonomy of web, information content provides quite reasonable results. Higher the information content of the web taxonomy, higher the uncertainty and higher the uncertainty, higher the irrelevance, so if we want to relevance data from the web we prefer low information content in web means low entropy. Lower the entropy, higher the semantic similarity of the web content from the different data sources on the web.

The main focus is at low entropy in the combination of semantic and synaptic web but the interest of this work is to know the conclusion of different combination of the web at low and high entropy in content and connection layer as shown in figure 2.

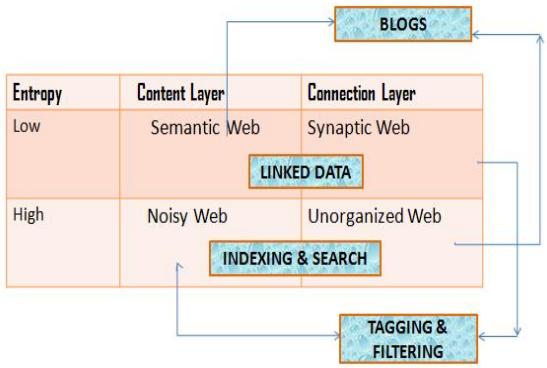

Fig. 2 Different combinations of web at low and high entropy

As shown in figure 2, at low entropy in the content layer, content must be in structured and metadata form, which realize the semantic web, otherwise content will remain unstructured and irrelevant, which realize the noisy web. In the connection layer at low entropy each content element is connected to the other like neurons in the human brain, which describe the significant relations between the content element as same as synaptic web, otherwise connections between the content are in ad-hoc state that provide us an unorganized web.

The analysis of web at high and low entropy becomes more impressive when we combine their domains.

(a) Semantic-Synaptic Web: The most organized and ideal form of web in which content and connections grant the information more relevance and efficient, which is machine understandable and user friendly as like link data.

(b) Noisy-Unorganized Web: The most unorganized web, which contents are unstructured and irrelevance. It is a key based indexing and search which grant the noisy and irrelevance information.

(c) Semantic-Unorganized Web: Content are in structured form but its connection through the web is not well. For the example, most blogs post have the authentic semantic structure but their connections by the hyperlinks have not the authentic relation.

(d) Noisy-Synaptic Web: Content are in unstructured and irrelevance form but it connecting the relevant information with a significant relations as same as tagging, filtering and recommendation of data.

After the study of all combinations of domain the question is obvious: Which one of the domain grants user the

most relevance and efficient result from the web? Based on the above domain's description it can be easily illustrated that the best choice is semantic-synaptic domain.

A. Semantic-Synptic Web Mining

In this proposed web mining the best ideas from semantic web and synaptic web at low entropy are combined. Semantic web mainly concentrate on web content, it makes the content more relevance and efficient and synaptic web concentrate on connectivity of web content, it interlink the web content to different content sources. All works have been done at low entropy so that information content of web content must be low. The low information content leads to web content that have higher semantic similarity from the different data sources on the web and user get the most relevant and accurate data. After the combination of above ideas from semantic web and synaptic web at low entropy, we get a novel mining technique called Semantic-Synaptic web mining, which provides us a most relevant and accurate data on the web.

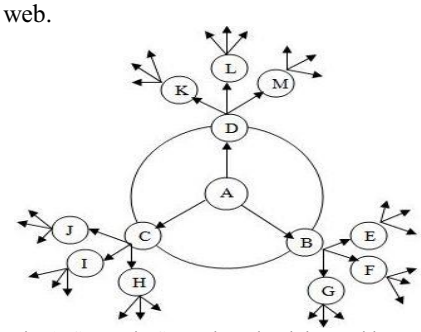

Fig. 3: Semantic-Synaptic web mining architecture

As shown in figure 3 all nodes shows the linked data on the web and outward arrows from the particular node indicates, that particular node is connected to nodes which have lowest entropy (particular range). As shown in figure node A is connected to nodes B, C and D, means B, C and D have lowest entropy among all nodes, so that it provides most relevant and accurate data to A. Now for node D, nodes K, L, and M are at lowest entropy with respect to node D, and the process continues in a similar fashion up above in the figure, similarly the structure continuing at nodes B and C. In simple words the nodes are to be distributed at a hierarchical range of entropy which results in relevant and accurate data from this proposed web mining comparison to the previous traditional web mining technique.

V. CONCLUSION

This paper first introduces the traditional web mining and proposed a novel model for improving the web mining, called Semantic-Synaptic web mining and construct the architecture of semantic-synaptic web mining. In this work two fast potent research areas semantic web and synaptic web at low entropy are combined and applied for mining the web and these results in an effective web mining technique which provides most relevant and accuracy of data on the web. Finally an architecture of Semantic-Synaptic web mining is proposed, in which web of data is distributed at a hierarchical range of entropy, so that most relevant and accurate data on the web is retrieved. The architecture depicts that the proposed technique is highly effective comparison to the previous traditional web mining.

REFERENCES

[1] Oren Etzioni. *"The world wide web: Quagmire or gold mine"*. Communications of the ACM, Vol.39(11), Pp.65-68 (1996)

[2] 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)

[3] 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).

[4] R. Cooley. "*The web usage mining: Discovery and Application of Interesting patterns from web data*", Phd thesis,Dept.of computer science, university of Minnesota, May 2000.

[5] 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.

[6] F. Sebastini, "*Machine Learning in Automated Text Categorization. Tech."* report B4-31, Istituto di Elaborazione dell'Informatione, Consiglio Nazionale delle Ricerche, pisa, (1999).

[7] S. Chakarabarti, "*Data Mining for Hypertext: A Tutorial Survey*", ACM SIGKDD Explorations, Vol. 1, no. 2, pp. 1-11, 2000.

[8] 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).

[9] T. Berners-Lee, J. Hendler and O. Lassila, "*The Semantic Web*", Scientific American. Vol. 284(5), Pp. 34-43(2001).

[10] B. Berendt, A. Hotho and G. Stumme," *Towards Semantic Web Mining* ", Proceeding of the First International Semantic Web Conference: The Semantic Web (ISWC 2002), Sardinia, Italy, vol. 2342, Pp. 264-278, (2002).

[11] T. Berners-Lee. "Linked Data—Design Issues", 2006; http://www.w3.org/DesignIssues/LinkedData.html .

[12] C. Bizer, R. Cyganiak and T. Heath, "How to publish Linked Data on the web", 2007; http://www4.wiwiss.fuberlin.de/bizer/pub/LinkedDataTutorial.

[13] L. Sauermann, R. Cyganiak, D. Ayers, M. Cool. " URIs for the semantic web", 2007; http://www.w3.org/TR/cooluris/ .

[14] C. Bizer, T.Heath and T.Berners-Lee, " Linked Data—The story so Far", Int'l. Semantic web & Information systems, to appear, 2009.

[15] C. Bizer, T.Heath, K. Idehen and T.Berners-Lee. Linked Data on the web (LDOW2008): Workshop Summary. In proceedings of the 17th International World Wide Web Conference, Beijing, China, 2008.

[16] C. Bizer, T.Heath, K. Idehen and T.Berners-Lee. Linked Data on the web (LDOW2009): Workshop Summary. In proceedings of the 18th International World Wide Web Conference, Madrid, Spain, 2009.

[17] C. Bizer, T.Heath, T.Berners-Lee and M. Hausenblas. Linked Data on the web (LDOW2011): Workshop Summary. In proceedings of the 20th International World Wide Web Conference, Hyderabad, India, 2011.

[18] Chen, H. and Chau, M., "*Web Mining: Machine Learning for Web Applications"*, Chapter 6, Annual Review of Information Science and Technology (ARIST), v38, p289-329 (2004).

[19] C. E. Shannon, "*A mathematical theory of communications*", Bell Systems Technical Journal, vol. 27, Pp. 379-423, (1948).

[20] Philip Resnik, "Semantic similarity in a Taxonomy: An Information-Based measure and its Application to Problems of Ambiguity in Natural Language", JAIR, 11, Pp. 95-130,(1999)