Entropy Measurement and Algorithm for Semantic-Synaptic Web Mining

Hiteshwar Kumar Azad
Department of Computer Science and Engineering
National Institute Of Technology Patna
Patna, India
Email: azad07it17@gmail.com

Kumar Abhishek
Department of Computer Science and Engineering
National Institute Of Technology Patna
Patna, India
Email: kumar.abhishek@nitp.ac.in

***Abstract*—Semantic-Synaptic web mining [1] aims to integrate the best idea from the semantic web and synaptic web with web mining at low entropy. Semantic entropy can help researchers to decide not only how to work with words, but which words to work with. Many researchers are struggling to upgrade the result of web mining by capitalizing semantic structure of the web so that one can get the relevant and efficient information from the web, but efficient and relevant information extracting from the web still faces a big challenge. Semantic-Synaptic web mining present a novel mining technique which interlinks the web of data to different data sources available on the web which have low entropy, so that one can find out the most relevant and accurate information on the web. This paper proposes an algorithm for semantic-synaptic web mining and presents a method for measuring the entropy of web pages using information content.**

Keywords*—*Entropy, Semantic Web, Semantic-Synaptic web, Semantic-Synaptic web mining, Web Mining.

I. Introduction

The past few decades, the success of World Wide Web has established by the fast-exploiting research area semantic web, synaptic web and web mining. They integrate each other efficiently and each gives a contribution to fix a new challenge comes in the path of the great success of World Wide Web. The widespread acceptance of World Wide Web all over the world and today web has turned to be the biggest information sources available on this planet. The Majority of the web contains a huge amount of unstructured data, which can only appreciated by humans, but due to the huge amount of data they can handle effectively by machines only. The semantic web is an ontological based technique which makes the web content more semantic and machine-understandable and synaptic web makes the connection between web content, while web mining deals the remaining part by automatically discover and gather the valuable information hidden in these web content, and making it feasible for extracting the relevant and efficient information from the web. The success of World Wide Web can't avoid the contribution of *words* in the web content and *connection* between webs content because it is very important for us, which and how to work with word and what should be the connection between webs content. The aim of improving the effectiveness of web contents, *entropy* is introduced in information theory, which can help researchers to decide not only how to work with words, but which words to work with. Further, for improving the electronic connection between people, data sets and real world to the online world, *synaptic web* is introduced.

978-1-4799-4674.7/14/$31.00

The aim of this technical document is *what should be the future of web mining and how the integration of semantic web and synaptic web at low entropy could be beneficial for us?* For the purpose of improving the web mining, first semantic-synaptic web mining with its component are described. Then an observation will be performed to see how the combination of semantic web and synaptic web could be profitable at low entropy web contents. In the next section an algorithm for semantic-synaptic web mining is proposed and finally a method of measuring the entropy of web pages using information content is presented.

II. Related Work

Oren Etzioni came up with the term web mining [2] in 1996 and arise a question: Is information on the web amply structured to provide effective web mining? Papers [3,4, 5] performed research in the web mining and advised to distinguish the web mining into three areas depending on which kinds of data to be mined. Some researchers integrate the content mining with structure mining to improvement of mining technique as mentioned in [6, 7, 8], but majority of researchers doesn't agree with their classification [3]. The most recognize categories of web mining today are WCM, WSM and WUM. In depth analysis of methods and application, see [9, 10]. Tim Berners-Lee in 2001 came up with a novel ontological approach, which construct the web more semantic and machine interpret-able, known as semantic web [11].The aim of semantic web is not only recommends to access the information on the web by direct link or by search engines but also to support its usages easily. [12] Focus the semantic web and web mining can suitable for each other: semantic web enhance the structure and effectiveness of mining technique, while web mining support to construct the semantic web.[13] present an outlook of where the semantic web and web mining coincide and how a perfect integration could be beneficial. Many researchers are struggling to upgrade the result of web mining, linked data [19] is one of their sequential contributions. In 2006 Berners-Lee turns up with the broad guidelines on how to use the advanced web technologies to link the data on the web from different data sources so that linked data provides a most relevant data on the web [20, 21, 22, 23]. For entropy measurement in information theory, the term *entropy*

credit to the Shannon and known as Shannon entropy [14], where entropy has explained in the context of information content. Resnik [15] uses concept's frequency for measuring the entropy, while Melamed [16] uses the words for measuring the entropy.

III. Semantic-Synaptic Web Mining

Semantic-Synaptic web mining [1] aims to combine the best idea from the semantic web and synaptic web with web mining at low entropy web pages. Actually semantic-synaptic web mining is entropy based mining technique, which interlinks the data on the web to different data sources available on the web, which have low entropy web contents. Lower the entropy of web contents, higher the semantic similarity between web contents, which provides us the most relevance and efficient data on the web. Details will be discussed in the following sections.

A. Semantic web

The semantic web was coined by Tim Berners-Lee [11], the founder of the World Wide Web to take the World Wide Web much further and develop a distributed system for knowledge representation and computing. The excellent achievement of the present World Wide Web presides over a new challenge where a bulky amount of data understandable by humans only because the machine support is restricted. Tim Berners-Lee recommends to improving the web content by machine interpret-able information which helps the client in his work. The aim of semantic web is not only recommends to accessing the information on the web by direct links or by search engines but also to support its usages easily. Today search engines are previously quite effective, but due to the huge collection of unstructured data on the web still respond unsatisfactory and inadequate list of data. Machine interpret-able information can leads the search engine to significant web pages which enhance the accuracy and recall. To achieve this goals the semantic web will be built up in multi-layered structure where each layer sequentially increases functionality, and lower layer support the upper one To achieve this goals the semantic web

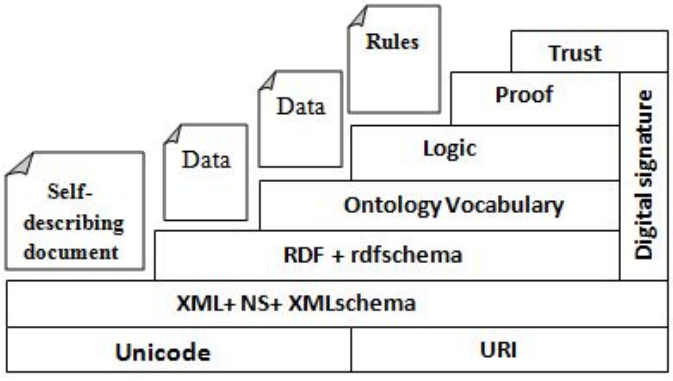


Fig. 1. The Semantic web layers structure.

will be built up in multi-layered structure where each layer sequentially increases functionality, and lower layer support the upper one. As shown in figure 1.

B. Synaptic Web

In the human brain, brain cells or neurons are connected to each other, which are at the root of brilliance. These chemically medicated connections are known as *synapses* [17]. The metaphor of the biological term *synapse* (connection between neurons) in the real world is used and it is observed that these connections have morphed in amazing track and the nature of connections are altering at an accelerating rate.

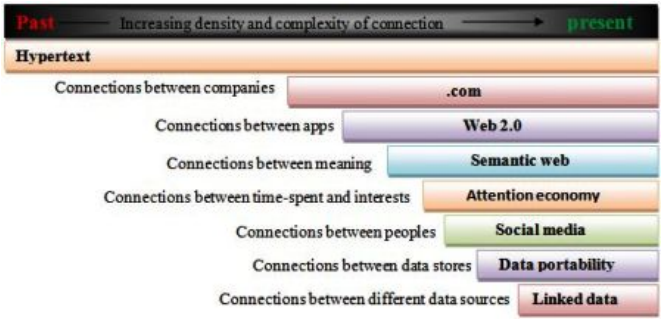


Fig. 2. Contribution of synaptic web in the present scenario.

As the web is a huge collection of data and the synaptic web established the connection between the data on the web.The *synaptic web* in the connections between objects (content/ information) is more important than objects themselves.

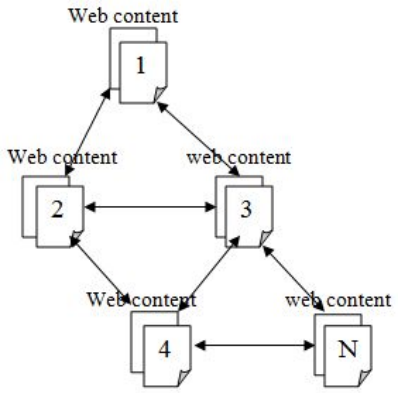


Fig. 3. Synaptic web.

The synaptic web [24] is the evolution of document delivery platform to a communication platform.In the synaptic web, filtering is more important than search. It is the next generation mechanism to navigate the document discovery issue. While search is contracting the incredible amount of data on the web to a limit set of web pages, filtering is about contracting the torrent of streams, networks and links that are matches from your criteria. It is about specifying and steadily refining your word view so that you can find the most relevant and accurate data on the web.

C. Entropy

In information theory the term *entropy* credit to the Shannon and known as Shannon entropy [14], is a measure of ambiguity and inconsistency in a random variable, which

determine the predicted value of the information contained in a message. Entropy is the average uncertainty in a random variable, which is similar to its information content. For measuring the semantic similarity in taxonomy of web, information content grants quite reasonable results [18]. However lower the information content in message, lower the uncertainty and lower the uncertainty, higher the relevancy in the result. The choice should be at low information content in web content, which is equivalent to the low entropy. Lower the entropy, higher the semantic similarity between the web content, which are available at different data sources on the web.

Entropy can help researcher to decide not only how to work with words, but which words to work with. Details about measuring the entropy of web pages will be discussed in next section.

D. Semantic-Synaptic Web

The combination of two particular domain semantic web and synaptic web at low entropy,an influential platform as semantic-synaptic web is obtained.

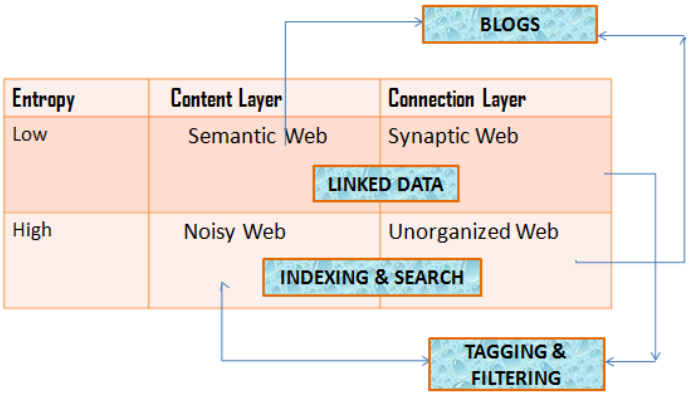


Fig. 4. Different combination of web at low and high entropy.

Semantic-synaptic web is a most organized and ideal form of the web in which all contents is connected to the other relevant content. Semantic-synaptic web provides well defined meaning of information which is connected to the other information source on the web to grant better co-operation and machine understandable for computer to process. Azad and Abhishek [1] present the details about the different combination of the web at low and high entropy in content and connection layer as shown in figure 4.

The combination of the semantic-synaptic web to the web mining technique at low entropy of web contents, a most effective web mining technique known as semantic-synaptic web mining is obtained. Semantic-synaptic web mining is entropy based mining technique in which web pages are distributed at a hierarchical range of entropy, where page having lowest entropy is set as root and the pages have second lowest range of entropy set as the level first and are connected to the root page. The pages having next lowest range of entropy are connected to the first level of web pages and so on, as shown in figure 5. The root web page provides the most relevant and accurate data on the web and pages having at the first level provides the next most relevant and accurate data on the web and so on.

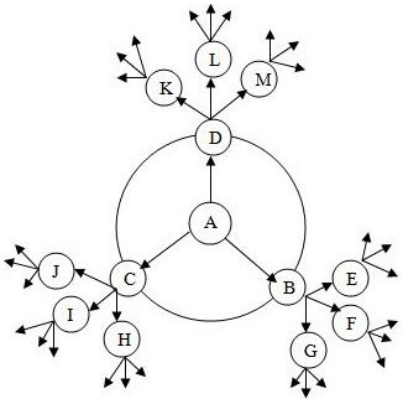


Fig. 5. Semantic-synaptic web mining structure.

IV. Algorithm for Semantic-Synaptic Web Mining

The system first randomly cluster the relevant web pages from web server (first phase) and measuring the entropy of all web pages (second stage). The page having the lowest entropy E_{t_0} set the web pages as root position, while page having less than the highest entropy E_{t_n} it go to the next step. Where if page_Entropy comes in a particular range of threshold entropy (E_{t_i}), connect the web pages as adjacent level i.

TABLE I. Algorithm: Semantic-Synaptic Web Mining Using Page Entropy

First Phase
begin
Clustering relevant web pages or document and services
end

Second Phase
begin
Calculate the web pages entropy
$page\_Entropy = E_{t_0}$,set the web page as root position
while $(page\_Entropy < E_{t_n})$
for i=1 to n
if$(E_{t_{i-1}} < page_Entropy <= E_{t_i})$then
Connect the web pages as adjacent level i
end for
end

V. Measuring the Entropy of Web Pages Using Information Content

For measuring the entropy of web pages, inconsistency of words in the web pages can be calculated through the principle of information theory. In information theory, Shannon [14] introduce an idea for determine the expected value of information contained in a message. Information content is inconsistency of every occurrence, which is well known as entropy.

In information theory, Entropy is known as the predicted average value of information content I_{x_i} associated with a random variable X. Entropy of a discrete random variable X is a function of probability mass function $p(x) = Pr[X =$

$x_i], x_i \in X$ where $i \in [1, n]$ and is defined as

$$E(X) = \sum_{i=1}^{N} p(x_i)I(x_i) = \sum_{i=1}^{N} p(x_i)\frac{1}{log(p(x_i))} = -\sum_{i=1}^{N} p(x_i)log \quad (1)$$

In the above equation (1) -ve sign shows the entropies are always +ve because probabilities are always between [0,1].

A. Method for Measuring the Entropy

For measuring the entropy, we are introducing the information content by Resnik [15] as a basic approach.

Step 1:

Clustering all taxonomy of the pages and counted the frequency of each word in taxonomy, which contained the concept c. formally,

$$Frequency(c) = \sum_{n \in words(c)} count(n) \quad (2)$$

Where words(c) is the set of words contained by concept c.

Step 2:

Measuring the Probability of words contains the concept ci simply as the relative frequency

$$p(c_i) = \frac{frequency(c_i)}{N} \quad (3)$$

Where N is the total number of words recognized in taxonomy.

Step 3:

After measuring the probability of words contain the concept c_i, we can easily calculate the information content by formula

$$I(c_i) = -log(p(c_i)) \quad (4)$$

According to the equation (4), information content decreases with the node's probability increases. In a hierarchical order of nodes, if there is a novel concept in a particular node then its information content will be 0 because it contains all child nodes then its probability will be 1.

Step 4:

According to the above equation (3) and (4), the probability $p(c_i)$of particular words containing the concept c_i and information content $I(c_i)$ of that word. Now applying the equation (1), measuring the individual entropy of each word and simply added them. Now we have an entropy of the web page E(T).

$$E(T) = \sum_{i=1}^{N} f(c_i)E(c_i) \quad (5)$$

Where $f(c_i)$ is frequency of words containing the concept c_i and $E(c_i)$ is the entropy of that word.

For example,in the following figure 6, the root node A contains all child nodes B, C, D, and E, so its information content is 0. As we see B has two child nodes D and E and probability of D is less than B hence the information content of B must be lower than D. however when we analyze the information content of B and C at the identical layer, B has two child nodes D, E while C has null node, hence the information content of B is lower than C.

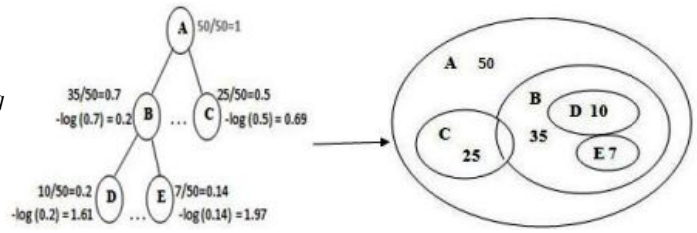
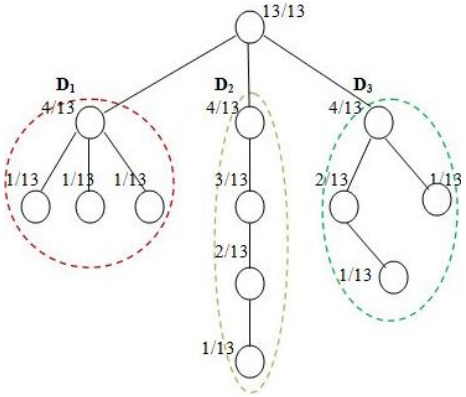


Fig. 6. Hierarchical structural of information content.

According to the information theory, information content of the nodes increases with increase the number of nodes in the hierarchical architecture. Basically entropy is the average information content of the taxonomy. Similarly analysis of the structural characteristics of web taxonomy by measuring entropy is carried out. Assume the taxonomy in figure 7 contains the domain taxonomy (D_1, D_2, D_3) and every individual domain consists of the same number of nodes but different hierarchical structure. .



$E(D_1) : 1.3768, E(D_2) : 1.711$ and $E(D_3) : 1.5075$

Fig. 7. Entropy of different hierarchical structure of taxonomy.

As in fig.7 after the measurement of entropy of each domain D_1 has the lowest entropy because D_1 is in appropriate order and its each node has the identical information content excluding the root node. D_2 has the highest entropy among all domains because it consists of the child nodes with different information content, so D_2 is in most uncertainty among all three domain taxonomies.

VI. Conclusion

This paper makes a study of an entropy based mining technique known as semantic-synaptic web mining, where web pages are distributed in a hierarchical range of entropy. A discussion was done on the metaphor of the biological term *synapse*(connection between neurons) in the real world and observed how the nature of connections are changing at an accelerating rate and what is the contribution of the synaptic web in the present scenario. An algorithm for semantic-synaptic web mining and present a method for measuring the entropy of web pages using information content is proposed. The example provided in the last section which compares the

entropy for different hierarchical structure of taxonomy. In this work, the importance of low entropy over the web mining at the combination of semantic web and synaptic web is outlined.

It is expected, in the future, entropy will play a big role in making the web smarter and the contribution of the synaptic web in the real world will increase.

REFERENCES

[1] H. K. Azad, Kumar Abhishek, "Semantic-Synaptic web mining: A novel model for improving the web mining," in *IEEE International Conference on Communication Systems and Network Technologies (CSNT-2014)*, pp.454-457, 2014.

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

[3] 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.

[4] O.R. Zaane, "From resource discovery to knowledge discovery on the internet.," in *Technical Report TR 1998-13, Simon Fraser University*, 1998.

[5] J. Srivastava, R. Cooley, M. Deshpande, P. N. Tan, "Web usage mining: Discovery and application of usage patterns from Web data," in *SIGKDD Explorations*,1 (2),1223, 2000.

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

[7] S. Chakarabarti, "Data Mining for Hypertext: A Tutorial Survey,"in *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," in *Osterreichische Gesellschaft fur Artificial Intelligence (OGAI)*, vol. 21, no.2, pp. 17-26, 2002.

[9] Han, Kamber, "Data Mining. Concepts and Techniques," in *Morgan Kaufmann, San Francisco, LA*, 2001.

[10] D. Hand, H. Mannila, P. Smyth, "Principles of Data Mining, MIT Press,"in *MIT Press*, Cambridge, MA, 2001.

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

[12] B. Berendt, A. Hotho, G. Stumme, "Towards semantic web mining," in *I.Horrocks, J.A. Hendler (Eds.), The Semantic WebISWC 2002. First International Semantic Web Conference, Proceedings, volume 2342 of LNCS* ,Springer, pp. 264278, 2002.

[13] G. Stummea, A. Hotho, B. Berendt, "Semantic Web Mining State of the art and future directions," in *Journal of Web Semantic. Web Semantics: Science, Services and Agents on the World Wide Web 4*, pp. 124143. 2006.

[14] C. E. Shannon, "A mathematical theory of communication,"in *Bell System Technical Journal*, Vol. 27, pp. 379423, 623656, July, October, 1948.

[15] P.Resnik , "Using Information Content to Evaluate Semantic Similarity in a Taxonomy,"in *14th International Joint Conference on Artificial Intelligence*, [So 1.]: Springer, pp.448-453, 1995.

[16] I.D. Melamed, "Measuring Semantic Entropy," in *Proceedings of the SIGLEX Workshop on. Tagging Text with Lexical Semantics*,Washington, DC, 1997.

[17] $http://en.wikipedia.org/wiki/Synapse.$

[18] P. Resnik , "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language," in *journal of Artificial Intelligence Research*,Volume 11, pages 95-130.

[19] T. Berners-Lee, "Linked Data-Design Issues," in *http://www.w3.org/DesignIssues/LinkedData.html*, 2006.

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

[21] L. Sauermann, R. Cyganiak, D. Ayers, M. Cool, "URIs for the semantic web," in *Phttp://www.w3.org/TR/cooluris/* , 2007.

[22] C. Bizer, T.Heath and T.Berners-Lee, "Linked Data-The story so Far," *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):122, 2009.

[23] C. Bizer, T.Heath, K. Idehen and T. Berners-Lee, "Linked Data on the web," *Workshop Summary. In proceedings of the International World Wide Web Conference.*,LDOW 2008,2009,2010,2011,2012,2013.

[24] Khris Loux, Eric Blantz, Chris Saad, "The Synaptic Web," in *http://synapticweb.pbworks.com/w/page/8983891/FrontPage*, 1998.