International Journal of
Science and Engineering Investigations vol. 5, issue 57, October 2016
ISSN: 2251-8843

Web Mining: A Review

Ihab Zaqout[1], Ali Mahdi[2], Mohammed Alhabbash[3]
[1,2,3]Dept. of Information Technology, Faculty of Engineering & Information Technology, Al Azhar University Gaza, Palestine
([1]i.zaqout@alazhar.edu.ps, [2]loosho_ali@hotmail.com, [3]e_i_2010@hotmail.com)

***Abstract*-**The World Wide Web, or simply Web, represents one of the largest sources of information in the world. We can say, perhaps, that any subject we think has probably become exists on a page in the Web. Information on the Web comes on different shapes and types, such as document texts, images and video clips. However, extraction of useful information, without the help of some Web tools, is not a trivial process. Here comes the role of Web mining, which provides tools that help us to extract useful knowledge from Web data. In this paper, we will provide an overview of Web mining and discuss the well-known applied algorithms of the three types of Web mining named content, structure and usage. Some future directions for this area provided.

Keywords- *Data mining, Web mining, Web content mining, Web structural mining, Web usage mining.*

I. INTRODUCTION

In the computer world, the data is a very interesting area. It continues to grow and expand dramatically, and it is important for us to find useful information from these huge data. The overall process of analyzing data sets, to find understandable and useful information to the owners of the data, called data mining. In the past few years, most of the data owned by institutions in the structured data stores such as relational databases. This data can easily be accessed for the purposes of mining using a variety of data mining techniques [1]. However, the nature of the data has changed dramatically since the advent of the Internet, which has the features and characteristics set it apart from structured data. These characteristics can be summarized as follows [2, 3]:

- The huge volume of data on the Web and continues to grow exponentially.
- Web contains data from various types and formats. This includes structured data, such as the table, and semi-structured data such as documents and Extensible Markup Language (XML), and unstructured data such as text in Web pages, multimedia data such as images and movies.
- Heterogeneity of information on the Internet. Authors from around the world are participating in the construction of Web content. As a result, you may find the content of similar or identical pages.
- Web data has links with hyperlinks, which means that Web pages are linked together so that anyone can navigate through the pages within the same site or across different sites. These links can tell us how information is organized between the pages within the site, and the strength or weakness of the relationship between the pages across different locations.
- Noise of information on the Web. The reasons for this are two issues. First, the typical Web page usually contains a lot of information, such as the main page, links, ads, and more. Thus, the page has no specific structure. Second, there is no control on the quality of information, meaning that anyone downloading content on the Web can be, regardless of its quality or its quality.
- A large part of the content on the Internet is dynamic; this means that often the information is updated continuously. For example, whether information is updated continuously.
  Website contains e-commerce sites that enable people to perform many of the purchases, transfer money, and more. This type of sites needs to provide customers with services such as computerized recommendation system.
- Web is not just a data and information. Nowadays, a virtual community, where people, organizations, and even computerized systems can communicate and interact with each other is the Web.

All of these characteristics make the process of extracting data on the network more challenging, and at the same time give us opportunities to discover useful knowledge and value of the Web. Because of the wide range of data types, traditional data mining techniques became inadequate [2]. This led to the crystallization of a need to develop new techniques and algorithms aimed mining data on the Internet.

Section 2 describes the related work while section 3 discusses the methods used to implement the three types of web mining. Conclusion and future work is provided in section.

II. RELATED WORK

The aim of Web mining is how users access to the data that generated from browsing Web. With the increasing growth of the Web, process of searching for information has become a

difficult and complex. Here comes the role of Web mining, which provides tools that help us to extract useful information from Web data. There are many data mining techniques and algorithms used by the Web mining to extract useful information form usage log file from the Web.

In [4], they used Weighted Page Rank (WPR) algorithm in the retrieval of relevant information in accordance with the user query based on the user's browsing behavior. A new Web content mining (WCM) algorithm called Page Content Rank (PCR) which depends on the significance of words in a page proposed by [5], while a Hybrid Page Rank (HPR) algorithm to get relevant results from the Web search engines depending on the basis of content and link structure of the Web pages proposed by [6, 7].

A new search result ranking algorithm proposed by [8] based on Web pages and tags clustering, and use several evaluating methods to assess and contrast with Google. In [9], they presented a new technique to extract data from Web pages through called top-k Web pages that describes top-k instances of any particular topic. A new approach based on hybrid clustering methods for the Web usage mining, called (WUM) process contains three phases: pre-processing, data mining and result analysis [10].

Web mining has also been seen as an ideal tool to detect and prevent fraud and threats. Thus, research issues include techniques that can be used to develop ways to detect many types of fraud and threats are discussed in [11, 12].

III. WEB MINING

Web Mining aims to find and extract useful information from the Web data, which include hyperlinks structures, Web pages content, records of using Web [2]. Web mining process divided into the following phases [13]:

- Resource finding: to retrieve and collect Web documents.
- Information selection and pre-processing: to determine the specific data and transform it into an appropriate form of treatment.
- Generalization: to discover patterns and identify them.
- Analysis: To check the validity of the information those has been extracted and work to represent it in an appropriate manner.

As shown in fig.1, Web Mining can be divided into three types [14-21]:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

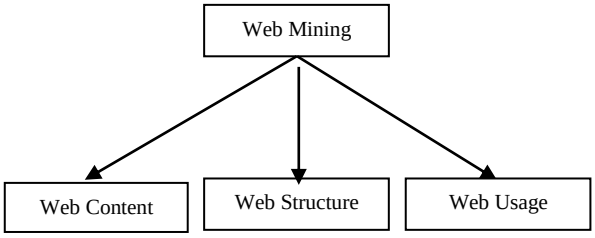

Figure 1. Types of Web mining

A. Web Content Mining

Web content mining uses the contents of Web pages to extract useful information. Classification and assembling of Web pages according to their subjects are examples of tasks that fall within the Web content mining [4]. These functions are similar to those used by [2] in the extraction of traditional data. However, there are jobs on the Internet that are not considered traditional extraction tasks. Examples include extracting customer feedback and posts from social networks, customer reviews, product specifications and extract in addition to emotions analysis and many more.

1) Term Frequency – Inverse Document Frequency

Term frequency–inverse document frequency (TF-IDF) is often the coefficient used in information retrieval and text mining. This parameter is a statistical measure used to assess the importance of the word in the presence of a particular document in the texts, but is offset by the frequency of the word in the corpus. Importance increases relatively by increasing the number of times a word or term appear in the document. The search engines use different forms of TF-IDF as a central tool in the assessment and arrangement of documents according to relevance given a user query.

Example 1: Here are two simple text documents:

1) Ali likes to watch movies. Osama likes movies too.

2) Ali also likes to watch football games.

Based on these two text documents, a list is constructed as follows:

[“Ali", "likes", "to", "watch", "movies", "also", "football", "games", “Osama", "too"]

Term count in each document d_1 and d_2 is computed as follows:

(1) [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]

(2) [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

TF(t, d) = the number of times that term t occurs in document d.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) * \text{IDF}(t, D) \quad (1)$$

TF("Ali", 1) = 1

IDF("Ali", D) = $\log\frac{2}{2} = 0$

TF-IDF ("Ali", 1, D)= 1*0 = 0

TF("too", 1) = 1

IDF("too", D) = $\log\frac{2}{1}$ = 0.301
TF-IDF("too", 1, D) = 1*0.301 = 0.301
TF-IDF("too", 2, D) = 0, because Term count "too" in d_2 = 0.

1) Genetic Algorithms

It is the method of optimization and research. This method can be classified as one of the evolutionary algorithms that rely on the mimicking the natural work of the Darwinian perspective. It uses technical search to find exact or approximate solutions to achieve optimization. It classifies global search heuristics. It is also a certain class of evolutionary algorithms known as the evolutionary computation, which uses bio-inspired operators such as inheritance, mutation, selection and crossover.

- *Initialization*: First, many individual solutions are randomly generated as primary forms of chromosomes. Chromosomes size depends on the nature of the problem, but usually there are several hundred or thousands of possible solutions. Traditionally, chromosomes are randomly generated to cover the full range of possible solutions on search spaces. Occasionally, the solutions may be "seeded" in areas where optimal solutions are likely to be found.
- *Crossing Over*: During each consequent generations, a collection of the current chromosomes is selected to produce a new generation. Selection relies on the optimization function. There is another way by choosing a random group of chromosomes, but this process may take a very long time
- *Mutation*: It is the process of generating a second generation of chromosomes that have been picked up during the selection process and then repeat crossing over and mutation until we get the children.

Example 2: Let X is session and Y chromosome as shown in Table 1.

- *Initial population/ Reproduction*

TABLE I. GENETIC ALGORITHM EXAMPLE

Chromosome	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Y_1	1	0	1	0	1	1	0	1	1	1
Y_2	1	1	1	1	1	0	1	0	1	1
Y_3	0	1	0	1	0	1	1	1	1	1
Y_4	0	0	1	0	1	0	0	0	0	0
Y_5	1	0	0	0	0	0	0	0	0	1
Y_6	1	1	0	0	1	0	0	1	1	0
Y_7	0	0	0	1	0	0	0	0	0	0

The percentage of the Fitness function support of each individual is calculated as follows:

$$F(S)\% = \frac{total\ number\ of\ true\ bit\ in\ chromosome}{total\ length\ of\ chromosome} = \frac{TDAc}{TLc}\ X\ 100 \quad (2)$$

Where,

TDAc = Total dominants alleles in chromosome and TLc = Total length of chromosome

Chromosome-Y_1 = (7/10)*100 =70%
Chromosome-Y_2 = (8/10)*100 =80%
Chromosome-Y_3 = (7/10)*100 =70%
Chromosome-Y_4 = (2/10)*100 =20%
Chromosome-Y_5 = (2/10)*100 =20%
Chromosome-Y_6 = (5/10)*100 =50%
Chromosome-Y_7 = (1/10)*100 =10%

We assume minimum support fitness = 20%. So Chromosome-Y_7 do not satisfy the minimum support, so this not selected in initial population.

- *Crossing Over*

Step 1:
Chromosome-Y_1: 1 0 1 0 1 1 0 1 1 1
Chromosome-Y_2: 1 1 1 1 1 0 1 0 1 1
Offspring (child) {Y1, Y2} → Y: 1 0 1 0 1 0 0 0 1 1
Fitness value of offspring = (5/10)*100 = 50% satisfies minimum threshold
Result: Two set frequent content = {Y_1, Y_2}

Step 2: Suppose we are performing crossover between {Y1, Y2} → Y chromosome and Chromosome-Y_3.
Offspring (child) {Y_1, Y_2} → Y: 1 0 1 0 1 0 0 0 1 1
Chromosome-Y_3: 0 1 0 1 0 1 1 1 1 1
Offspring (child) {Y_1, Y_2, Y_3}→Y:0 0 0 0 0 0 0 0 1 1
Fitness value of Offspring (child) {Y_1, Y_2, Y_3} → Y = (2/10)*100 = 20%
Offspring {Y_1, Y_2, Y_3} → Y is survivable.
Result: Three frequent content set = {Y_1, Y_2, Y_3}, as per property it's all two item subsets are frequent = {{Y_1, Y_2}, {Y_1, Y_3}, {Y_2, Y_3}}.

Step 3: Suppose we are performing crossover between {Y_1, Y_2, Y_3} → Y chromosome and Chrosome-Y_4.
Offspring (child) {Y_1, Y_2, Y_3}→Y:0 0 0 0 0 0 0 0 1 1
Chromosome-Y_4: 0 0 1 0 1 0 0 0 0 0
Offspring (child) {Y_1, Y_2, Y_3, Y_4} → Y: 0 0 0 0 0 0 0 0 0 0
Fitness of Offspring (child) {Y1, Y2, Y3, Y4} = 0%
Thus offspring did not survive and 4-itemset {Y_1, Y_2, Y_3, Y_4} is not frequent
Result: It means chromosome-Y_4 may be may not be associated with some subset ancestor of partner individuals.

Step 4: Suppose we select two individuals {Y_1, Y_2} and chromosome-Y_4
Offspring (child) {Y_1, Y_2} → Y: 1 0 1 0 1 0 0 0 1 1
Chromosome-Y_4: 0 0 1 0 1 0 0 0 0 0
Offspring (child) {Y_1, Y_2, Y_4} → Y= 0 0 1 0 1 0 0 0 0 0
Fitness of Offspring (child) {Y_1, Y_2, Y_4} = (2/10)*100 = 20%

Result: $\{Y_1, Y_2, Y_4\}$ is three set frequent content and $\{Y_1, Y_4\}$, $\{Y_2, Y_4\}$, $\{Y_1, Y_2\}$ as two set frequent content sets

Step 5: Suppose we select Individuals $\{Y_1, Y_2\}$ and chromosome-Y_5
Offspring (child) $\{Y_1, Y_2\}$ → Y: 1 0 1 0 1 0 0 0 1 1
Chromosome-Y_5: 1 0 0 0 0 0 0 0 0 1
Offspring (child) $\{Y_1, Y_2, Y_5\}$→Y:1 0 0 0 0 0 0 0 0 1
Fitness of (child) $\{Y_1, Y_2, Y_5\}$ → Y = (2/10)*100 = 20%
Result: Offspring is survivable and $\{Y_1, Y_2, Y_5\}$ is a frequent three content set and it's all subsets $\{Y_1, Y_5\}$, $\{Y_2, Y_5\}$, $\{Y_1, Y_2\}$ are frequent.

Step 6: Suppose we select two individuals $\{Y_1, Y_2\}$ and chromosome-Y_6
Offspring (child) $\{Y_1, Y_2\}$ → Y: 1 0 1 0 1 0 0 0 1 1
Chromosome-Y_6: 1 1 0 0 1 0 0 1 1 0
Offspring (child) $\{Y_1, Y_2, Y_6\}$→Y:1 0 0 0 1 0 0 0 1 0
Fitness of offspring (child) $\{Y_1, Y_2, Y_6\}$ → Y = (3/10)*100 = 30%
Result: Offspring is survivable and $\{Y_1, Y_2, Y_6\}$ is a three set frequent content and it's all subsets $\{Y_1, Y_6\}$, $\{Y_2, Y_6\}$, $\{Y_1, Y_2\}$ are frequent

Step 7: Suppose we select two chromosome-Y_3 and Chromosome-Y_6
Chromosome-Y_3 = 0 1 0 1 0 1 1 1 1 1
Chromosome-Y_6 = 1 1 0 0 1 0 0 1 1 0
Offspring $\{Y_3, Y_6\}$ → Y: 0 1 0 0 0 0 0 1 1 0
Offspring (child) $\{Y_3, Y_6\}$ → Y = (3/10)*100 = 30%
Result: Offspring is survivable and $\{Y_3, Y_6\}$ is a two set frequent contents.

The selection process of population is based on the fitness value and child parent sibling relationship.

B. Web Structural Mining

Web structural mining uses the hyperlink structure of the Web as a source of information in the process of mining [15]. Hyperlinks represent one of the special features of the network, as well as Web-based. All Web pages are linked to each other through links so that the user can navigate from page to page through them. Web structure mining is designed to extract abstract useful information from hyperlinks structure on the network for many purposes [2]. Some of the techniques used in the Web structure mining are inspired from the analysis of social networks in which we can find certain types of pages, such as hubs, authorities and communities based on the incoming and outgoing links.

1) PageRank

A Webpage's ranking is determined by analyzing the ranking of all the other Webpages that link to the Webpage in question [14]. It is calculated as,

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{L(T1)} + \cdots + \frac{PR(Tn)}{L(Tn)}\right) \quad (3)$$

Where,

PR is a page rank score, *L*() is the number of out links, and *d* is the damping factor. The damping factor is used to stop other pages having too much influence. The total vote is "damped down" by multiplying it to 0.85.

Example 3: Assume there are three interrelated Webpages as depicted in fig.2.

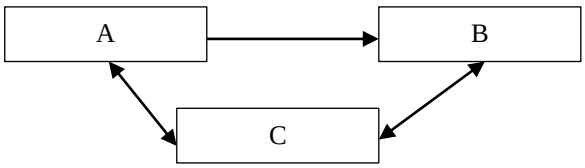

Figure 2. Web pages and their relationships

Let us assume the initial PageRank as 1.0 and do the calculation yields to the results Table 2. The damping factor *d* is set to 0.85:

PR(A) = (1-*d*) + *d**(*PR*(C)/L(C)) = 0.15 + 0.85*(1/2) = 0.575
PR(B) = (1-*d*) + *d**((*PR*(A)/L(A)) + *d**(*PR*(C)/L(C))) = 0.15 + 0.85*((0.575/2)+0.5) = 0.819
PR(C) = (1-*d*) + *d**((*PR*(A)/L(A)) + *d**(*PR*(B)/L(B))) = 0.15 + 0.85*((0.575/2)+(0.819/1)) = 1.091

PR(A) = (1-*d*) + *d**(*PR*(C)/L(C)) = 0.614
PR(B) = (1-*d*) + *d**((*PR*(A)/L(A))+*d**(*PR*(C)/L(C))) = 0.874
PR(C) = (1-*d*) + *d**((*PR*(A)/L(A))+*d**(*PR*(B)/L(B))) = 0.15+ 0.85*((0.614/2)+(0.874/1)) = 1.153

TABLE II. ITERATIVE CALCULATION FOR PAGERANK

Iterative	PR(A)	PR(B)	PR(C)
0	1.0	1.0	1.0
1	0.575	0.819	1.091
2	0.614	0.874	1.153

2) Weighted PageRank algorithm

Xing and Ghorbani [22] proposed a Weighted Page Rank (WPR) algorithm. This algorithm assigns rank values to pages according to their importance rather than dividing it evenly. The importance is assigned in terms of weight values to incoming and outgoing links.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_p} \quad (4)$$

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p} \quad (5)$$

Where,

I_n and I_p are the number of incoming links of page *n* and page *p*, respectively

R(*m*) denotes the reference page list of page *m*

$$WPR(n) = (1-d) + d \sum_{m \in B(n)} WPR(m) \,.\, W^{in}_{(m,n)} | W^{out}_{(m,n)} \quad (6)$$

Where,

W_{out} is the weight of link (m, n) calculated based on the number of outgoing links of page n and the number of outgoing links of all reference pages of m

o_n and o_p are the number of outgoing links of page n and p, respectively referring to the same hyperlink structure as shown in fig.2.

$WPR(A) = (1\text{-}d) + d*(WPR(C).W^{in}_{(C,A)}.W^{out}_{(C,A)})$
$WPR(B) = (1\text{-}d) +$
$d*(WPR(A).W^{in}_{(A,B)}.W^{out}_{(A,B)} + WPR(C).W^{in}_{(C,B)}.W^{out}_{(C,B)})$
$WPR(C) = (1\text{-}d) + d* (WPR(A).W^{in}_{(A,C)}.W^{out}_{(A,C)} +$
$WPR(C).W^{in}_{(B,C)}.W^{out}_{(B,C)})$
$W^{in}_{(C,A)} = I_A/(I_A + I_B)$ = 1/ (1+2) = 1/3
$W^{out}_{(C,A)} = O_A/(O_A + O_B)$ = 2/ (2+1) = 2/3

$WPR(A)$ = (1 - 0.85) + 0.85*(1*(1/3)*(2/3)) = 0.69
$WPR(B)$ = (1 - 0.85) + 0.85*(0.69*(1/2)*(1/3)) = 0.44
$WPR(C)$ = (1 - 0.85) + 0.85*(0.69*(1/2)*(2/3)) = 0.47

3) HITS algorithm

Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg [23]. Kleinberg identifies two different forms of Web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resource lists guiding users to authorities. Thus, many good hub pages on the same subject point a good hub page for a subject points to many authoritative pages on that content and a good authority page. The HITS algorithm treats WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. Fig.3 shows the hubs and authorities in Web.

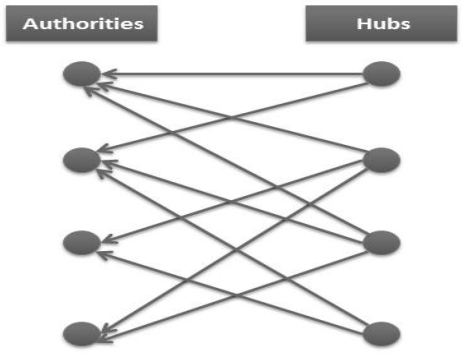

Figure 3. HITS algorithm [23].

$H_P = \sum_{q \in I(p)} A_q$ and $A_P = \sum_{q \in B(p)} H_q$ (7)

Where,

H_P = The hub weight

A_P = The authority weight

$I(p)$ and $B(p)$ denotes the set of reference and referrer pages of page p

C. Web Usage Mining

Web usage mining refers to practical usage patterns of discovery data on the Web. The raw data used in this process represents through user records, which recorded the interactions between the user and the Website logs. This includes data such as user clicks, the date and time of access, IP addresses, etc. Usage logs are usually found on servers as well as the server access logs and records of Web application [20]. Like the data extraction process in [18], Web usage mining often divides into three stages as shown in fig.4 namely: pre-treatment, discover patterns, and analyze patterns as discussed in [2].

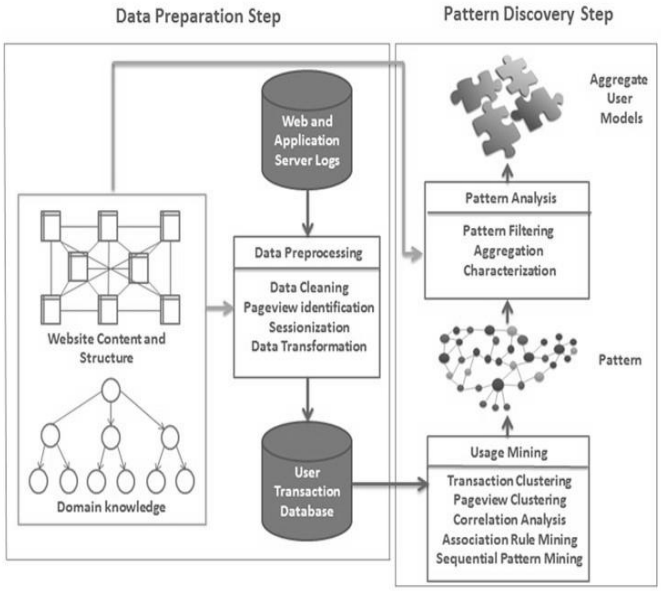

Figure 4. Three stages of Web usage mining [2]

In the first stage, pre-treatment, the data use transfer to the abstractions, which represents the interactions for a user within a Website. Other types of data can be involved at this stage, including the real data in the pages of the site, and the data that describes the structure of Web pages, and data representing the demographic information about users.

In the discovery of patterns stage, a wide range of styles and techniques is applied from various fields such as statistics, data mining, databases, and machine learning to detect hidden patterns that carries the behavior of users.

In the last stage, pattern analysis, patterns and statistics discovered in the previous stage are processed to filter rules or patterns that unnecessary. This phase can produce models that can be used as input for other applications such as visualization tools and Web analytics and recommendation systems. Often a

mechanism to inquire about the knowledge is required at this stage, such as Structured Query Language (SQL).

IV. CONCLUSION AND FUTURE WORK

Data Source as Web is very worthy for mining and extraction of knowledge. On the one hand, the richness and diversity of information on the Web make it a valuable source of information from which we can extract a lot of useful knowledge. On the other hand, it makes the process of mining more difficult and more complex than the traditional data mining process, especially in the absence of pre-defined structure. Many research issues have been studied extensively in the field of mining Web. However, as the network continues to grow in all dimensions, including the structure, content, and usage information, there are still many areas, which we can invest in researches.

Further researches can be done to expand the right Web standards and do calculation for these standards, so that they can study the different behaviors on the Internet. Using data users click, we can do more researches on the development of ways to extract decision-making process itself, not just the final results of this process; this can help to improve the process and see different parts of any process impacts the Web Standards.

REFERENCES

[1] D. Hand, H. Mannila, and R. Smyth, Principles of Data Mining, MIT Press, Cambridge, 2001.

[2] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer, 2011.

[3] M.G. Junior and Z. Gong, "Web Structure Mining: An Introduction", Proceedings of the IEEE International Conference on Information Acquisition, June 27 - July 3, 2005, Hong Kong and Macau, China, pp. 590 – 595.

[4] A. Sowmiya A and A. Gayathri, "Enhancement in Weighted Page Rank Algorithm for Ranking Web Pages", International Journal of Computer Technology & Applications, vol. 5, no. 1, pp. 140 – 143, 2014.

[5] J. Pokorny and J. Smizansky, "Page Content Rank: An Approach to the Web Content Mining", Proceedings of the IADIS International Conference on Applied Computing, vol. 2, Algarve, Portugal, February 22-25, 2005.

[6] M. Kaur and C. Singh, "A Hybri Page Rank Algorithm: An Efficient Approach", International Journal of Computer Applications (0975 – 8887) vol. 100, no. 16, pp. 58 – 63, 2014.

[7] M. Kaur and C. Singh, "A Hybrid Page Rank Algorithm using Content and Link Based Algorithms", Global Journal of Advanced Engineering Technologies, vol. 3, no. 2, pp. 160 – 163, 2014.

[8] C. Zhao, Z. Zhang, H. Li, and X. Xie, "A Search Result Ranking Algorithm Based on Web Pages and Tags Clustering", IEEE International Conference on Computer Science and Autoation Engineering (CSAE), 10-12 June 2011, pp.609-614.

[9] A. Joy and R. Remya, "Techniques for Web Mining of Various Forms of Existence of Data on Web: A Review", International Journal of Advance Research in Computer Science and Management Studies, vol. 3, no. 1, January 2015, pp. 279 - 281.

[10] N. Pushpalatha, "Hybrid Clustering Methods for Web Usage Mining", International Journal of Advance Research in Computer Science and Management Studies, vol. 3, no. 9, September 2015, pp. 228 - 232.

[11] J. Servastava, P. Dasikan, V. Kumar, Web Mining - Concepts, Applications, and Research Directions, Foundations and Advances in Data Mining, vol. 180 of the series Studies in Fuzziness and Soft Computing, pp 275-307, 2005.

[12] K. Sharma, G. Shrivastava and V. Kumar, "Web Mining: Today and Tomorrow", 3rd International Conference on Electronics Computer Technology (ICECT), vol. 1, pp. 399-403, 2011.

[13] R. Kosala and H. Blockeel, "Web Mining Research: A Survey", ACM SIGKDD, vol. 2, no. 1, pp. 1 – 15, 2000.

[14] R. Jain and G. N. Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications, vol. 13, no. 5, pp. 22 – 25, 2011.

[15] P. R. Kumar and A. K. Singh, "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval", American Journal of Applied Sciences, vol. 7, no. 6, pp. 840 - 845, 2010.

[16] G. Kaur and S. Aggarwal, "A Survey- Link Algorithm for Web Mining", International Journal of Computer Science & Communication Networks, vol. 3, no. 2, pp. 105-110, 2013.

[17] R. Malarvizhi and K. Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study", International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 2940 – 2945, 2013.

[18] G. Singh and P. Dixit, "A New Algorithm for Web Log Mining", International Journal of Computer Applications, vol. 90, no. 17, pp. 20 – 24, 2014.

[19] K. Tandele and B. Pansare, "Web Usage Mining with Improved Frequent Pattern Tree Algorithms", International Journal of Computer Science and Information Technology Research, vol. 3, no. 2, pp. 952-958, 2015.

[20] J. Srivastava, R. Cooley, M. Deshpande, and T. P- Ning, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, pp. 12-23, 2000.

[21] P. Mehtaa, B. Parekh, K. Modi, and P. Solanki, "Web Personalization Using Web Mining: Concept and Research Issue", International Journal of Information and Education Technology, vol. 2, no. 5, pp. 510-512, 2012.

[22] W. Xing and A. Ghorbani, "Weighted PageRank algorithm", CNSR '04 Proceedings of the 2nd Annual Conference on Communication Networks and Services Research, IEEE Computer Society Washington, DC, USA, pp. 305 – 314, 19 - 21 May, 2004

[23] J. Kleinberg, "Hubs, Authorities, and Communities". ACM Computing Surveys (CSUR), vol. 31, no. 4, 1999.