International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-01, APR 2021

A Review Paper on Web Mining: Web Structure Mining

Riddhi Doshi, Assistant Professor Department of Computer Engineering, St. Vincent Pallotti College of Engineering and Technology Nagpur India, riddhidoshi1996@gmail.com

Vivek Kute, Associate Professor Department of Computer Engineering, St. Vincent Pallotti College of Engineering and Technology Nagpur India, vivekkute@rediffmail.com

Abstract Incorporation of roughly more than 70 million webpages everyday led Web moving up tremendously. The wide adoption of Internet has fundamentally altered the ways in which we communicate, conduct businesses and make purchase. This expansion of the web has been resulted in a massive amount of freely accessible data. The large amount of web data encapsulates various beneficial data patterns and relationship among them, which needs to be to mined and identified. Thus, web mining, an extension of data mining plays important role. It integrated various technologies in research fields including statistics, informatics, knowledge discovery and many more. Based type of data, web mining gets classified into three categories namely web content mining, web usage mining and web structure mining. Web structure mining analyses the structure of the web and discovers the relationships among websites and webpages.

In this paper we have scrutinized web structure mining meticulously along with different web structure mining algorithms.

Keywords — Web Mining, Web structure Mining, Data Mining, PageRank, HITS, Weighted Page Rank, Link Analysis

I. INTRODUCTION

Web mining makes use of data mining strategies to significantly identify and mine facts from web documents along with utilities. Exponential increase in web libraries, web portals and number of archived documents makes manual bifurcation of each document challenging as well as exquisite. Data mining can administrate this issue well [1]. Web mining methodologies manages pulling facts and fetching data optimally. Here the data mining strategies are utilized to enhance web utilities [2]

Web mining has been the fundamental consolidation of data mining and World Wide Web. It cites to the whole mechanism of determining probably appropriate as well as previously unknown data or facts from web information [2]

Web mining is divided into following four sub-tasks:

1. Resource Finding: the mechanism of fetching information i.e. each of two online or offline from intermedia reference on Internet [1] [2]
2. Information Selection and Pre – processing: the mechanism of transforming distinct initial data fetched in previous task [1] [2]
3. Generalization: the mechanism of automatically determining primary patterns with respective web sites along with across numerous sites [1] [2]
4. Analysis: the mechanism of verifying or classifying the determined patterns in previous task [1] [2]

In the first section we have discussed about web mining and its methodologies. Second section is brief about Web Structure Mining followed by different web structure mining algorithms in next section. Finally, we summarize the paper.

II. WEB MINING METHODOLOGIES

Web mining methodologies can be classified into three different categories:

- Web Content Mining: also referred as Text Mining, the process of examining and mining of text, image as well as graph of web page to identify the importance of that content to particular search request. It's more effective when used with in relation to a content database dealing with specific topic [1]. Analogous to both data mining as well as text mining the web content mining makes use of data mining methodologies as most of the data available on Internet is in text format [3] [4] [5]
- Web Usage Mining: Web usage mining refers to the procedure used to discover the browsing style by surveying the user's navigational nature. It considers secondary data on Internet. So this type of web mining implicates automatic generation of significant access patterns along different web servers [3] It's acquisition of web page access information for web sites. Usage mining generates the route that dominates to most accessed web pages. It makes use of information collected in

 DOI : 10.35291/2454-9150.2021.0160

access logs through web servers [1]. Web usage information can be captured through proxy servers, client applications and web servers efficiently [3] [4] [5]

- Web Structure Mining: Web Structure mining targets discovery of skeletal summary of web sites along with web pages. Link information forms to be the core centre of structure mining [2] [4]

III. WEB STRUCTURE MINING

Web structure mining discovers the correlation between web pages associated through information or immediate link relation. It's a mechanism to develop a prototype of hyperlink structure of particular webpage. It produces web structures overview with reference to particular website along with webpage. It also attempts to create the link structure of all hyperlinks at inter document level. Web structure mining relates to web content mining since a web page which is composed of hyperlinks and makes use of real as well as primary data [3]. This web structure information is predictable from the arrangement of web structure blueprint over database approach for webpage. It also lets web browser to fetch information relevant to search request precisely to the connected webpage from website where particular data resides. Web crawlers examine the website, fetching significant home page then associating the information over multiple reference links to retrieve the particular webpage comprising appropriate information. The main idea is to uproot previously unknown correlation between webpages [1]. Hence, the primary target of web structure mining is to root out the underlying knowledge.

In search engine domain, user wants to incur most accurate and publicly renowned outcomes, instead some inappropriate webpage, so it is very eminent to enlist the most accurate webpage at the top of search result. Structure mining is developed on this arbitration approach to obtain a lot of information and assist users with path and suggestion of most significantly accurate web page [6] [7] [8]

A. Types of Web Structure Mining

Web Structure Mining is divided into three categories depending upon the structured data used:

1. Web Graph Mining
2. Web Information Extraction
3. Deep Web Mining

All the above mentioned categories have different ways of handling main data, processing methods and application areas. [6]

1. Web Graph Mining: Web is considered as directional graph where webpages are viewed as nodes and hyperlinks among them are the edges. Internet contributes secondary details about how distinct webpages are related to each other by hyperlink. There has been a vital body to function on reviewing the features of webgraph and deriving useful data from it. Graph compiles valuable perception into web algorithm for scanning, community exposure, web cleaning and subcultural development that represent its expansion [2]

2. Web Information Extraction (Web IE): Information Extraction is the function of automatically fetching structured and unstructured system readable document. Web information mining targets upon automatically fetching structures with distinct authenticity as well as granularity across the webpage. The structure of web content is encapsulated in one page which is called as Interpage Structure. Content inside a webpage have inbuilt structure formed on multiple HTML and XML tags within webpage. [2]

3. Deep Web Mining: Alongside web pages that are easily approachable by consequent hyperlink, web also inholds huge number of non - accessible content. This sheltered segment of web is called as Deep Web or Hidden Web. It consists of huge amount of online web databases. Deep web incorporates tremendously huge amount of high - quality structured information [2]

IV. WEB STRUCTURE MINING ALGORITHMS

A. Partitioning Algorithm

This algorithm is based on clustering approach. In this algorithm initially a webpage is illustrated as vector within whole hyperlinked range. This range forms a huge dimensional vector as large number of website exists. To scale down this size and minimize entanglement evaluate only first part of URL amid initial clustering loop. Then re – cluster every cluster with the help of whole URL. Here webpages are grouped as M clusters (N - the number of processing units). The algorithm is carried out on following steps:

1. Depict every webpage by its hyperlink
2. Take account of all base URL from all webpages (as A).
3. Portray each page by A-Dimensional vector.
4. Implement K-means clustering for M groups.
5. Evaluate extent of every cluster.
6. If Standard deviations of extent of clusters are not below 10% then return to step 4 and execute with distinct initial value.
7. Now take account of all the URLs in single cluster (say UB)
8. Demonstrate every page by B-Dimensional vector.
9. Implement K-means clustering for M groups.
10. Compute length of all clusters
11. If Standard deviation of size of cluster is not below 10% then return to step 9 and execute with distinct initial value.

 DOI : 10.35291/2454-9150.2021.0160

12. Compute the distance across each cluster and clusters with minimum distance are located within same machine.

This algorithm consists of approximately 80% of all the referenced pages that have been referenced by pages within particular cluster. The processing time of the algorithm inflates linearly as processing units are increased. [2]

B. PageRank Algorithm

This algorithm was developed by Brain and Page at Standford University. It advances the concept of citation analysis. It gives efficient method that can quantify the significance of webpage by simply estimating number of pages that are linked. These links are known as Backlinks. If a Backlink is from authoritative page then this link is given higher preference. This link from one page to another is measured as vote. Page inheriting as well as forming the votes both are vital [3]. Algorithm begins by translation of every URL from database to an integer value. Each hyperlink is collected in a database with the help of integer value generated before to determine particular webpage. Before initiating iteration hyperlink structure is organized with respect to parent – id and withdrawing suspended links. The primary assignment must be elected to accelerate concurrence. For an instant of time weights are conserved in memory and the prior weights are obtained on disk in sequential manner. After concurrence, suspended links are set back and rankings are re-valuated [1]. Page and Brain proposed a formula to calculate the page rank of page W as stated below:

PageRank (W) = (1-df) + d (PageRank (W1)/O (W1) + ---------- + PageRank (Wn / O(Wn)) [3]

Here, PageRank(Wi) is PageRank for page Wi that refers to page W, O(Wi) is no. of outlinks on Page Wi and df is damping factor (Used when stopping too many influential pages) [3]

The steps for implementing this algorithm include:

1. Set the rank of each page within the structure by 1/m where m i s the total number of pages to be ranked. Then these pages are represented by an array W [j] = 1/m

2. Select a damping factor ranging in 0 to 1 i.e 0<df<1

3. For each node j, repeat the following: Let PageRank ( ) be an array of m elements which represents PageRank of each individual web page. PageRank [j] = 1-df

For (all pages k such that k links to j) do PageRank[j] = PR[j] + df*W[k]/Kn where Kn = number of outlinks of j

4. Update values of A as W[j] = PageRank[j]

Repeat from step 3 until the rank score converges [9].

The sum of all page ranks across the webpages is one as it forms a probability distribution. Page rank of a particular webpage relies on the number of webpages directing to other webpage [3]. The algorithm is extraneous to the theme, if full uses PageRank value as domain assessment standards [6]. This algorithm makes use of offline computing which leads to least response time [6]. It assigns weights to link evenly which may lead to assigning similar authority to highly related webpage and lowly realted webpage. This avoiding the certain link situation develops irrationality. Dead ends and circular references will shorten the rank of the first page. If one particular cluster of page has no address to a different cluster of pages, then it leads to Spider Trap [9]. [10]

C. Weighted PageRank Algorithm

Weighted PageRank algorithm is the advanced version PageRank algorithm. This algorithm allots rank value to the most authoritative page rather than distributing the whole rank value of a page equally across all linked pages. [3] All the incoming and outgoing links are given priority with respect to allotted weight value. This evaluation is done on the account of number of links to a page and number of incoming links to all the reference pages. Let weight of link (x,y) be $W^{in}(x,y)$ evaluated depending on the no. of incoming links of page y along with the no. of incoming links if all orientations pages of pages x. [9]

$W^{in}(x,y) = I_y / \Sigma_{m \in A(x)} I_m$ Where Im, Iy are no. of inlinks of page m and y resp.

$W^{out}(x,y)$ is the weight of link (x,y) evaluated depending upon the no. of outlinks of page y and number of outlinks of all reference pages of x.

$W^{out}(x,y) = O_y / \sum_{m \in B(x)} O_m$ Where O_m, O_y are number of outlinks of page m and y resp.

Following are the steps to implement Weighted PageRank Algorithm:

1. Initialize the rank score of each web page in the structure by 1/x where x is total number if pages to be ranked. Then pages are represented as array of x elements. B[i] = 1/x

2. Select damping factor ranging in 0 to 1 0<d<1

3. For a given page i having calculate the values for $W^{in}(i,j)$ and $W^{out}(i,j)$ using above formula.

4. Repeat for each node: Let WPR be an array of x element which represent Weighted PageRank for each web page. WPR[j] = 1-d

For all pages i such that i link to j do WPR[j] = WPR[j]+d*B[i]* $W^{in}(i,j)$*$W^{out}(i,j)$

5. Update the values of B as B[i] = WPR[i]

Repeat from step 4 until the rank score converges. [5]

This algorithm is more efficient because it executes in crawl time instead of request time. The ranking value is attributed based on the significance of every page associated with it. The gaps in this algorithm are that as a result it reverts less unrelated page to a particular search request because here ranking depends on the structure of web instead of content. There are pages that have tendency to stay popular all along which does not guarantee the particular information at t the user's request. [9]

 DOI : 10.35291/2454-9150.2021.0160

D. Hilltop Algorithm

Hilltop Algorithm follows primary principle of PageRank algorithm. It regulates the sorted weight of search query outcome with the help of total count and quality of links similar to PageRank algorithm. The only difference is that this algorithm use the input provided by user to induce the authority of a particular page. This algorithm abetted Google with thorough technical amendments in search ranking. The algorithm begins by creating an apt page index. Then takes search query from user and discovers the most relevant expert page and score. After this target page is sorted and search results are acknowledged. Basically the algorithm works on two main process namely expert page search and target page sorting [6].

This algorithm is used with PageRank algorithm by Google for better goggle search results. The only advantage of this algorithm is it identifies the best quality webpages with scientific approach. Low operational efficiency and low scalability are the shortcoming affecting the result of the algorithm [6].

E. HITS Algorithm

HIST stands for Hyperlinked – Induced Topic Search. This algorithm is also known as hubs and authorities. It is a type of Link Analysis algorithm that rates Web pages. In this algorithm, Kliengerg gives two forms of webpages called as Hubs and Authorities [3] [11]. Hubs are the pages that interpret list of resources. Authorities are the pages which have vital content. HIST algorithm considers World Wide Web as a directed graph where pages are vertices and links corresponds to edge. Weight of Hub (Hw) and Authority (Aw) is calculated using following equation:

$Hp = \Sigma Aa$ where a belongs to I(p)

$Ap = \Sigma Ha$ where a belongs to B(p)

Authority weight of a page is directly proportional to thesum of hub weights of the pages that link to it. Hubs of a page is proportional to the sum of authority weight of pages that link to it [6].

This algorithm is implemented in two phases:

1. Sampling Phase: In this phase basically webpage is viewed as a graph. This phase begins by developing subgraph in which hubs and authorities are identified. This main goal is create a subgraph which is rich in relevant authoritative pages. In index based searching, query is used to collect root set of page. As numerous pages are relevant to particular search topic, then at least some of them may be authorities. Extend root set into base set by considering all the pages that root set pages link to, unto designated cut-off size. Base set contains approx. 1000 to 5000 pages with corresponding links and it is the final result of Sampling Phase.

2. Weight Propagation Phase: In this phase, best hubs and authorities are fetched from base set. Assign non – negative authority weight and hub weight as aw and hw resp. Normalization is done so that total sum remains bounded. Now, initialize all a and h value to a uniform constant.

If a page is linked to many good hubs then authority weight is increased. Therefore value of aw for page w is updated to be the sum of ha across all pages a that links to w:

$aw = \Sigma ha$

Similarly, If a page is linked to many good authorities, hub weight is increased and updated.

$Hw = \Sigma aa$

The relationship between authority and hub leads this algorithm to dig into more authoritative pages. It gives good precision ratio for search query. Algorithm needs to perform some real-time calculations online which affect response speed. As there are many pages whose links are irrelevant to search request topic, but point to each other. HITS algorithm is likely to produce such page a high ranking, resulting in topic drift phenomena [6]. [11] [10]

V. FINDINGS

The whole web structure mining algorithm tries to gauge the importance of the website. The website can be artificially inflated to make this page important. The page ranking algorithm uses a random navigation model to calculate the score for a particular webpage. Sometimes references reduce the rank of the first page in the page ranking algorithm. While the weighted page ranking algorithm sometimes returns the least related pages. The hit algorithm rates a website based on hubs and permissions. However, it is difficult to identify whether a particular page should be considered a hub or an authority, and it sometimes returns irrelevant links. induces that HITS and the PageRank algorithm require a higher response time in online computing and real-time calculations, while they are opposite to the weighted PageRank algorithm. The response time increases linearly with the growth of the website. From the study conducted in this article, we found that in a given case, all algorithms return irrelevant links and the search result.

VI. SUMMARY AN D FUTURE SCOPE

With the appearance and improvement of internet mining, the technology is used not most effective withinside the search engine filed, however additionally in all factors of e-commerce and social media life. It extracts and find out expertise from big data. Web mining technology has emerged as the foundation for multitude of new internet technologies and created incalculable value. Web structure mining most effectively examines the relationships between web documents using the facts conveyed by the links in each and every document. This paper explains in about web mining and web structure mining in brief. It also looked over the different web structure mining algorithms.

Research can be done and algorithms can be modified in order to overcome the drawbacks mentioned under the finding section.

 DOI : 10.35291/2454-9150.2021.0160

REFERENCES

[1] 2. B. R. P. 1SK. Ahmad Mohiddin, "Web Mining: Methodologies, Algorithms and Applications," *International Journal of Computer Science And Technology*, vol. 3, no. 4, 2012.

[2] M. K. S. S Sundeep Kumar, "Web Pattern Analysis Using Web Structure Mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 5, 2017.

[3] T.Nithya, "Link Analysis Algorithm for Web Structure," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 8, 2013.

[4] G. S. V. K. Kavita Sharma, "Web Mining : Today and Tomorrow," *IEEE*.

[5] M. B. M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," *Indian Journal of Computer Science and Engineering*, vol. 1, no. 4.

[6] Y.-l. T. a. F. Qin, "Research on Web Association Rules Mining Structure," in *World Congress on Intelligent Control and Automation*, China, 2010.

[7] R. S. S. Qingyu Zhang, "Web Mining : A Survey of current research, techinques and software," *International journal of Information Technology and Decision Making*, vol. 7, no. 4, 2008.

[8] H. B. Raymond Kosala, "Web Mining Research: A Survey," *ACM SIGKDD*, vol. 2, no. 1.

[9] A. C. S. N. Ronak Jain, "Web Structure Mining using Link Analysis," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 6, 2015.

[10] K. P. Punit Patel 1, "A Review of PageRank and HITS Algorithms," *International Journal of Advance Research in Engineering, Science & Technology*, vol. 2, no. 1, 2015.

[11] K. L. Rinkal Sardhara, "Web Structure Mining : A Novel Approach to Reduce Mutual Reinforcement," in *3rd International Conference and Workshops on recent advances and Innovations in Engineering*, 2018.

 DOI : 10.35291/2454-9150.2021.0160