International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228325

A Comprehensive Study on SomeWeb Mining Algorithms

Monimoy Ghosh, Asoke Nath*
Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, West Bengal, India

Article Info
Volume 8, Issue 2
Page Number : 319-326

Publication Issue :
March-April-2022

Article History
Accepted: 10 April 2022
Published: 25 April 2022

ABSTRACT

The tremendous growth of Webtechnologies in the past three decades has made it the largest publicly accessible data source in the world. With the ever-increasing volume of data on the Web, it is getting difficult and time-consuming to discover informative knowledge and patterns. Finding intelligent and user-requested data from unstructured and inconsistent material on the internet is a difficult undertaking. Web mining is the application of data mining techniques to discover patterns and structures and extract knowledge from the World Wide Web. It extracts structured and unstructured data from web pages, server logs, and link structures using automated methods. To categorize and rank search results, a variety of Web Mining techniques are commonly employed, including PageRank, Weighted PageRank, and HITS. The motive behind this paper is to present and analyze the currently important algorithms for ranking web pages such as PageRank, Weighted PageRank and HITS.

Keywords : Web mining, World Wide Web, Web search rank, Page rank and Weighted Page rank, HITS.

I. INTRODUCTION

The World Wide Web is a global repository of information and continues to expand in size and complexity. As the Web grows daily, obtaining that information becomes more and more tedious. The fundamental challenge is regulating semi-structured or unstructured Web information, which is difficult to regulate and enforce a framework or standards. A set of Web pages lacks a unifying structure and shows manifold authoring styles and content variation. This heterogeneous and dynamic nature of the Web increases the complexity of dealing with information from different perspectives of knowledge seekers such as business analysts and web service providers. Analysing and fetching relevant data from large databases calls for automated extraction tools, through which user-queried data can be fetched from billions of pages over the internet leading to the discovery of relevant information and interesting patterns.

Web mining is the application of data mining techniques to discover patterns and structures to extract knowledge from the World Wide Web. It uses automated methods to extract both structured and unstructured data from web pages, link structures, and server logs.

II. WEB MINING

Web mining is the process of using the Data Mining technique in order to automatically discover or extract the information from web documents.[1] It consists of the following sub-tasks:

1)**Resource finding:** This is the process of retrieving data from multimedia sources on the Web, such as news, stories, forums, blogs, and the text content of HTML pages acquired by eliminating the HTML tags, whether it is online or offline.

2) **Information selection and pre-processing**:This entails selecting and pre-processing specific information from retrieved online resources automatically. This procedure converts the data that was originally retrieved into information.

3) **Generalization:**It automatically detects general patterns on particular websites as well as across several sites. In generalisation, data mining and machine learning techniques are applied.

4) **Analysis:** It involves validating and interpreting the patterns that have been mined. It is very essential in pattern mining. In the information on the knowledge finding process on the web, a human plays a critical part.

III. WEB MINING CATEGORIES

According to analysis targets, there are three main sub-categories of web mining[2]:

- Web Content Mining (text, image, records, etc.)
- Web Structure Mining (hyperlinks, tags, etc.)
- Web Usage Mining (HTTP logs, app server logs, etc.)

A. Web Content Mining

Web content mining analyses web content such as text, multimedia data, and structured data (within web pages or linked across web pages) to understand the content of web pages, provide scalable and informative keyword-based page indexing, entity/concept resolution, web page relevance and ranking, web page content summaries, and other valuable information related to web search and analysis.

Essentially, it[1][4] is the scanning and mining of text on a Web page in order to determine the content's relevance to the search query. After clustering web pages through structure mining, this scanning is done, and the results are based on the level of relevance to the recommended query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query. Text mining is aimed at extracting specific information from customer search results in search engines. This enables the entire Web to be scanned in order to retrieve cluster content, as well as the scanning of specific Web pages inside those clusters. Despite the fact that search engines have the potential to deliver thousands of links to Web pages related to the search topic, this type of web mining allows for the elimination of irrelevant information. When used in conjunction with a content database devoted to a specific topic, web text mining is extremely successful. Online universities, for example, use a library system to retrieve publications relevant to their general topics of study. This unique content database allows for the extraction of just information relevant to certain subjects, resulting in the most targeted search engine results. The results are of greater quality since only the most relevant information is provided. This boost in productivity is attributable to the usage of text and visual content mining. The major purposes of this type of data mining are to collect, categorise, arrange, and deliver to the user requesting the information the best possible information available on the WWW. This system is required for scanning the numerous HTML documents, graphics, and text on Web pages. The generated data is fed into search engines in order of

relevance, resulting in more useful search results for each query.

B. Web Structure Mining

Web structure mining is the process of analysing nodes and connection structures on the Web using graph and network mining theory and algorithms. It extracts patterns from hyperlinks, which are structural components of a web page that connect it to another place. It can also extract information from a page's document structure (e.g., analyse the treelike structure of page structures to describe HTML or XML tag usage). Both types of web structure mining aid in the comprehension of web material and may also aid in the transformation of web information into more structured data sets.

In other words, it[1][4] is a tool for determining the relationship between Web pages that are linked via information or a direct link. This structure data is discoverable through the provision of web structure schema using database mechanisms for Web pages. This link enables a search engine to send data about a search query directly to the connecting Web page from the Web site where the content is hosted on. This is accomplished by using spiders to search Web sites, retrieve the home page, and then link the information using reference links to bring up the precise page containing the necessary information. Because of the large volume of information on the Internet, structure mining helps to solve two major issues. The first of these issues is search results that are irrelevant. Since search engines frequently only allow for low precision parameters, the relevance of search results can be misrepresented. The second of these issues is the inability to index the massive volume of information available on the Web. With content mining, this results in a poor level of recall. This reduction is due in part to Web structure mining's function of determining the model underlying the Web hyperlink structure. Structure mining's main goal is to discover previously unknown linkages between Web pages. Web structure data mining allows corporations to link information from its own website in order to facilitate navigation and cluster content into site maps. This enables users to find the information they need using keyword association and content mining. The hierarchy of hyperlinks is also determined in order to route related content within the sites to the relationship of competitor links and connections through search engines and third-party co-links. This allows for the clustering of related Web sites in order to determine their relationship. Structure mining allows for the determination of similar structures of Web sites on the WWW by grouping and identifying underlying structures. This information can be used to project online content similarity. The capacity to preserve or develop a site's information to allow web spiders to access it in a higher ratio is then provided by the recognised similarities. The larger the number of Web crawlers, the more beneficial to the site because of related content to searches.

3) Web Usage Mining

Web usage mining is the technique of obtaining relevant information (such as user clickstreams) from server logs. It discovers patterns relating to general or specific groups of users; recognises users' search habits, trends, and associations; and anticipates what users are looking for on the Internet. It aids in the improvement of search efficiency and effectiveness while also promoting items and services, presenting information to various groups of people at the appropriate moment. Web search companies do web usage mining on a regular basis to improve their service quality.

Web usage mining [5] collects information about how people access Web pages. This usage data contains the paths to the Web pages that have been visited. This data is frequently collected automatically by the Web server and stored in access logs. Other important information provided by CGI scripts includes referral logs, user subscription information, and survey logs. This area is critical to firms' total use of data mining for internet/intranet-based applications and information

access. Companies can use usage mining to get useful information about the future of their business function potential. Some of this data can be gathered from a combination of lifetime user value, product cross-marketing techniques, and the success of promotional campaigns. The obtained usage data enables organisations to deliver more effective results for their company, resulting in increased revenue. Usage data can also be valuable for honing marketing skills that will help the company outsell its competitors and promote its services or products more effectively. Usage mining is beneficial not only to firms that use online marketing, but also to e-businesses that rely exclusively on traffic generated by search engines. The usage of this type of web mining aids in the gathering of essential data.

IV. Web Mining Algorithms

The Web mining technique provides additional information through hyperlinks whereby different documents are connected [6]. We can view the web as a directed labelled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as a web graph. For example, the graph in Figure 1 is a directed graph with 3 vertices and 6 edges.

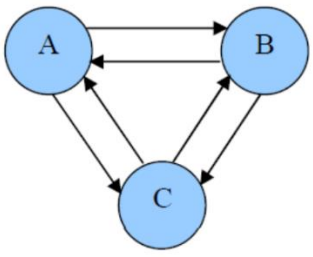

Figure 1: An example of a directed graph G with 3 vertices and 6 edges.[11]

There is a number of algorithms based on link analysis. In the present study the authors will discuss three important algorithms Page Rank [6], Weighted Page Rank[7] and HITS (Hyperlink Induced Topic Search) [8] which are discussed below :

A. PageRank

During their Ph.D. at Stanford University, Brin and Page devised the PageRank algorithm based on citation analysis [6]. The popular search engine, Google, uses the PageRank algorithm. In Web search, they used citation analysis to treat incoming links as citations to the Web pages. However, just applying citation analysis techniques to a large number of Web documents does not yield effective results. As a result, PageRank offers a more sophisticated method of determining the value or relevance of a Web page than just counting the number of pages connecting to it (also known as "back links"). If a back link comes from an "important" page, it is given a higher weighting than back links from non-important pages. In a nutshell, a link from one page to another can be thought of as a vote. However, not just the share of votes a page receives is essential, but so is the "importance" or "relevance" of the ones who cast those votes.

Assume any arbitrary page A has pages T1,T2,T3… to Tn pointing to it (incoming link). PageRank can be calculated by the following equation:

$$PR(A) = (1-d) + d\left(\frac{PR(\mathrm{T}_1)}{C(\mathrm{T}_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \quad (1)$$

where,

the parameter ‘d’ is a damping factor, usually set to 0.85 (to stop the other pages from having too much influence, this total vote is “damped down” by multiplying it by 0.85).

$C(T_i)$ is defined as the number of links going out of page T_i.

The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person

will continue is d, and it follows that he will skip a page with a probability of 1-d.

The PageRanks form a probability distribution over the Web pages, so the sum of all Web pages' PageRank will be one. PageRank can be calculated using a simple iterative algorithm and corresponds to the principal eigenvector of the normalized link matrix of the Web.

Consider the following webgraph G1.

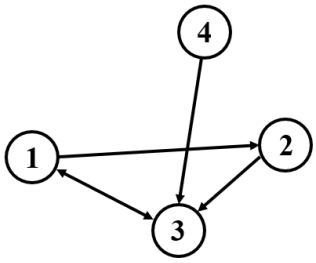

Figure 2:A web graph of 4webpages for PageRank illustration

If we apply the PageRank algorithm on the above set of 4 webpages, we get the following result after 40 iterations: [1.490110.783301.57660 0.15000], where the numbers indicate the PageRanks of the pages 1 to 4 respectively.

B. Weighted PageRank

The Weighted PageRank (WPR) algorithm, introduced by Wenpu Xing and Ali Ghorbani [7], is an extension of the PageRank algorithm. Instead of distributing the rank value of a page evenly across its outbound linked pages, this algorithm assigns a higher rank value to the more significant pages. Each outgoing link is assigned a value proportional to its relevance. The importance of the incoming and outgoing links is assigned in terms of weight values and is represented as $W^{in}(m, n)$ and $W^{out}(m, n)$ accordingly. $W^{in}(m, n)$ is the weight of link(m, n) calculated based on the number of incoming links of page n and the numbers of incoming links of all reference pages of page m, as shown in equation (2).

$$W_{m,n}^{in} = I_n / \sum_{p \epsilon R(m)} I_p \quad (2)$$

$$W_{m,n}^{out} = O_n / \sum_{p \epsilon R(m)} O_p \quad (3)$$

where I_n and I_p are the numbers of incoming links of page n and page p respectively. R(m) denotes the reference page list of page m. $W^{out}(m, n)$, as shown in (3), 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, where O_n and O_p are the number of outgoing links of page n and p respectively. The formula, as proposed by Wenpu et al, for the WPR is, as shown in (4). which is a modification of the PageRank formula.

Consider the following webgraph G2.

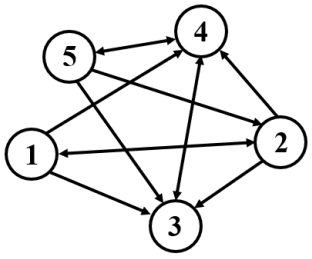

Figure 3:A web graph of 5 webpages for Weighted PR illustration

If we apply the Weighted PageRank algorithm on the above set of 5 webpages, we get the following result after 5 iterations: [0.756300.757081.021610.941290.77353], where the numbers indicate the Weighted PageRanks of the pages 1 to 5 respectively.

C. Hyperlink Induced Topic Search (HITS)

The HITS algorithm ranks web pages by analysing their inbound and outbound links. Kleinberg [8] distinguishes between two types of Web pages: **Hubs** and **Authorities.** Authorities are pages that contain

essential information. Hub pages serve as resource lists, directing users to authoritative sources. As a result, a good hub page on a subject links to many authoritative pages on that topic, and a good authority page links to many good hub pages on the same topic. Hubs and Authorities are shown in Figure 2.

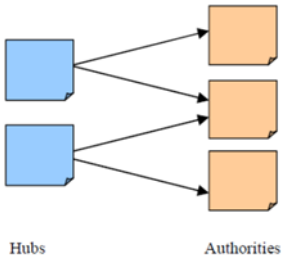

Figure 4: Hubs and Authorities [9]

According to Kleinberg, a page can serve as both a hub and an authority. This circular link leads to the development of the HITS iterative algorithm. The HITS algorithm treats the WWW as a directed graph G(V,E), where V represents pages and E represents links.[10] The HITS algorithm determines a web page's rating by assessing its textual contents in relation to a particular query. After the collecting of web pages, the HITS algorithm focuses solely on the structure of the web, ignoring the textual content. The HITS algorithm consists of two basic steps. The sampling step comes first, followed by the iterative step. In the Sampling step, a set of relevant pages for the submitted query is selected, i.e. a subgraph S of G with a high authority page count is retrieved. This approach begins with a root set R, a subset of S, with the understanding that S is small, rich in relevant pages regarding the query, and comprises the majority of the good authorities. The second stage, the Iterative step, uses equations (5) and (6) to discover hubs and authorities based on the outcome of the sampling step.

$$H_p = \sum_{q \epsilon I(p)} A_q \qquad (5)$$

$$A_p = \sum_{q \epsilon B(p)} H_q \qquad (6)$$

where H_p is the hub weight, A_p is the Authority weight, I(p) and B(p) denotes the set of reference and referrer pages of page p. The page's authority weight is proportional to the sum of the hub weights of pages that it links to it [8]. Similarly, a page's hub weight is proportional to the sum of the authority weights of pages that it links to.

Consider the web-graph G3.

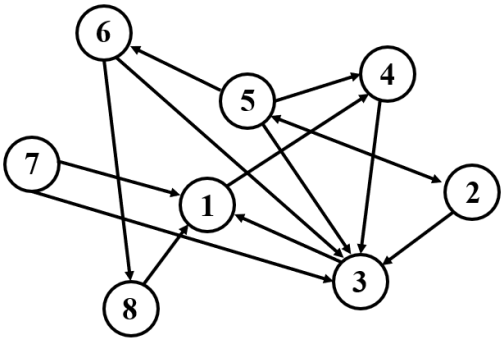

Figure 5:A web graph of 8 webpages for HITS illustration

If we apply the HITS algorithm on the above set of 8 webpages, we get the following result after 3 iterations:

Hub Scores [5 9 4 13 22 1 11 4]

Authority Score [13 15 27 11 5 9 0 3]

where the numbers indicate the Hub Scores and Authority Scores of the pages 1 to 8 respectively.

CONCLUSION

The World Wide Web is a global information medium that people may read and write to using computers linked to the Internet. Web mining is the process of finding and analysing meaningful information on the World Wide Web. This document describes the three most used web mining algorithms: PageRank, Weighted PageRank, and HITS. Significant patterns about user behaviour on the web can be derived using web mining algorithms, improving the relationship between the website and its users. Web mining is a large field with a lot of work to be done. This study

may serve as an introductory insight into the field of web mining.

V. REFERENCES

[1]. Kosala, R. and H. Blockeel, "Web mining research: a survey". SIGKDD Explor. Newsl., 2000. 2(1): p. 1-5. DOI:10.1145/360402.360406

[2]. Han, J., Kamber, M., Pei, J. (2012). "Data Mining Concepts and Techniques". Elsevier/Morgan Kaufmann. 3rd edition. Netherlands.

[3]. Just, Jiri. "A Short Survey of Web Data Mining." (2013).

[4]. Osmar R. Za¨ane,"From resource discovery to knowledge discovery on the internet", Technical Report TR 1998-13, Simon Fraser University, 1998.

[5]. R.W. Cooley, "Web usage mining: Discovery and application of Interesting patterns from Web data", PhD thesis, dept of computer science, university of Minnesota, May 2000.DOI:10.1145/846183.846188

[6]. S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998. DOI: 10.1016/S0169-7552(98)00110-X

[7]. Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR "04), IEEE, 2004.

[8]. Kleinberg, J.M., "Authoritative sources in a hyperlinked environment". J. ACM, 1999. 46(5): p. 604-632. DOI: 10.1145/324133.324140

[9]. Mohamed-K HUSSEIN et al., "An Effective Web Mining Algorithm using Link Analysis", International Journal of Computer Science and Information Technologies, Vol. 1 (3), 2010, 190-197. DOI: 10.1.1.259.5389

[10]. C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "Link analysis: Hubs and authorities on the World". Technical report: 47847, 2001. DOI:10.1137/S0036144501389218

[11]. http://ianrogers.uk/google-page-rank

[12]. https://www.geeksforgeeks.org/weighted-pagerank-algorithm/ Date accessed: 15/05/2022

[13]. https://www.geeksforgeeks.org/hyperlink-induced-topic-search-hits-algorithm-using-networxx-module-python

[14]. https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af

AUTHOR'S PROFILE

Dr. Asoke Nath is working as Associate Professor in the Department of Computer Science, St. Xavier's College (Autonomous), Kolkata. He is engaged in research work in the field of Cryptography and Network Security, Steganography, Green Computing, Big data analytics, Li-Fi Technology, Mathematical modelling of Social Area Networks, MOOCs etc. He has published **259** research articles in different Journals and conference proceedings.

Mr. Monimoy Ghosh is a postgraduate student of the Dept. of Computer Science, St. Xavier's College, Kolkata. He is enthusiastic about Cyber Security, Web Development, Machine Learning, DBMS, Automata Theory, Compiler Design and Operating Systems. Being a computer science enthusiast,he is also a member of various online learning communities like Codechef, Coursera, edX, etc.

Cite this article as :

Monimoy Ghosh, Asoke Nath, "A Comprehensive Study on SomeWeb Mining Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 2, pp. 319-326, March-April 2022. Available at doi : https://doi.org/10.32628/CSEIT228325
Journal URL : https://ijsrcseit.com/CSEIT228325