International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.15, No. 5/6, November 2025

A COMPARATIVE ANALYSIS OF LINK-BASED ALGORITHMS FOR EFFECTIVE WEB STRUCTURE MINING

V Indumathi

School of Computer Studies-UG, RVS College of Arts and Science, Sulur, Coimbatore.

ABSTRACT

The World Wide Web has become one of the most valuable resources for data recovery and information releases since it has the largest collection of data and numerous pages or reports. Advances in web mining are the key to unlocking information on the Internet. There are three types of web mining: web content, web structure, and web usage. One of these classes, Web structure mining is the focus of this research. A significant role in the web mining process is played by web structure mining. This paper discusses the experimental results for Link Based Ranking Algorithms and clarifies Web Mining techniques and certain well-known tactics used in Web structure mining.

KEYWORDS

Web structure Mining, Page Rank Hits, SimRank, Web mining.

1. INTRODUCTION

Web is a system of overall level, continually changing and non-structured [1]. The Web is the biggest information source on the planet. Web mining expects to concentrate and mine valuable information from the Web. It is a multidisciplinary field including information mining, AI, regular language handling, measurements, databases, data recovery, sight and sound, and so on. The measure of data on the Web is gigantic, and effectively open. The information doesn't come just from the contents of the web pages yet additionally from the extraordinary component of Web, its hyperlink structure and the assorted variety of contents. Examination of these qualities regularly uncovers intriguing examples and new information which can be useful in expanding the effectiveness of the clients, so the methods which are useful in extricating information present on the web is an intriguing region of examination. These procedures help to extricate information from Web information, in which in any event one of structure or usage (Web log) information is utilized in the mining process. In this paper, we right off the bat give a review on by and large Web mining ideas and advances, at that point, center on Web structure mining in detail.

2. WEB STRUCTURE MINING

Inside the Internet, average of Web-pages may have the structure incorporates the hubs and some hyper-joins as edges that are interfacing the related pages. The proposed WSM approach is a system of finding the structure of data store over the Web. This kind of structure mining approach is focused over the hyperlink situated structure of a Web procedure. Different articles in this are connected with one another by any way. Just by actualizing the essential procedures additionally by assuming that a few occasions are not reliant so they may cause off base ends. This kind of

mining structure is an approach of applying the diagram hypothesis so as to watch the hub along with their structure of associations of the websites.

This structure situated mining can encourage the clients to recover the significant records by breaking down connection situated structure of Web content. What's more, a few issues for these sorts of mining so as to work with any structure of gave hyper-joins in the Web. In this examination of Link is likewise a space for scientists. Furthermore, the Web may comprise of not just of pages, yet additionally of hyperlinks guiding any page toward the some other kind of pages. It finds the particular structure of connections that is hyperlinks given at some degree of between reports and furthermore so as to produce the basic end in regards to any Web-webpage or any Web-page. This idea is utilized for recovering pages that are not just applicable but at the same time are of high caliber, or legitimate on the subject. However, by the rising consideration inside the Web mining idea, the examination of the structure have additionally been expanded additionally their endeavors can have created in the most recent sort of exploration area alluded as the Link-Mining [8], that is additionally positioned in the content inside the examination of connection, additionally the hyper-text is put and played out the web mining through some learning and the inductive kind of programming approach and through diagram mining forms. This can be additionally sorted into two distinct sorts based on the sort of structure data applied as in Fig. 1.

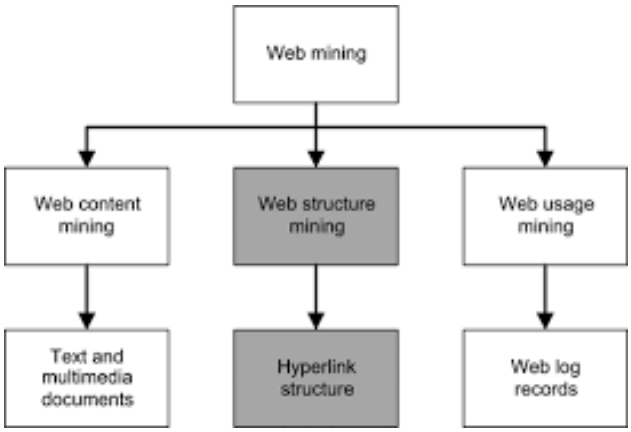

Figure 1: Web Structure Mining Techniques and Applications

3. TYPES OF WEB STRUCTURE MINING TECHNIQUES

a) Hyperlinks

This is the unit of structure which may joins an area inside the Web-page with different areas, that are has a place with the comparative Web-page or over the diverse kind of Web-pages. Any kind of hyperlink which may interface with the various segments of comparable page is eluded as Intra Document-Hyper-interface likewise the hyper-connect which may joins the two kinds of pages is eluded as an Inter-Document Hyperlink. There has been a critical collection of work on hyperlink investigation, of which [23] give an exceptional study.

b) Document-Structure

Moreover, the content introduced in the Web-pages may likewise mastermind in the tree sort of structured example, this is subordinate over the few kinds of HTML-labels or the XML-labels in any page of web. Mining endeavors here have concentrated on naturally removing the archive

object model structures out of records [5, 7]. The WSM is likewise a classification sort of the web-mining process for the information, which is a sort of hardware that is applied so as to perceive the relationship in the middle of the Web-pages that are associated through the data or might be with the direct association. What's more, this kind of information structure is presented through the providing the web engineering model by means of the procedures of database for the Web-pages. So this kind of association may empower any web crawler so as to pull the information which is connected with the query for search which is straightforwardly associated with the connecting page of any Web-website on which the content is set. This kind of work might be performed by the utilization of the creepy crawlies filtering process over Web-webpage, additionally recovering content from the home-pages, after which is joins the point by point by means of the connections of reference all together to acquire some specific pages front that are having a few wanted details.[7]

4. MACHINE LEARNING METHODS

Future research can use WSM to analyze non-formal organizations and find customer groups with typical information sources and applications, like given in Fig.2:

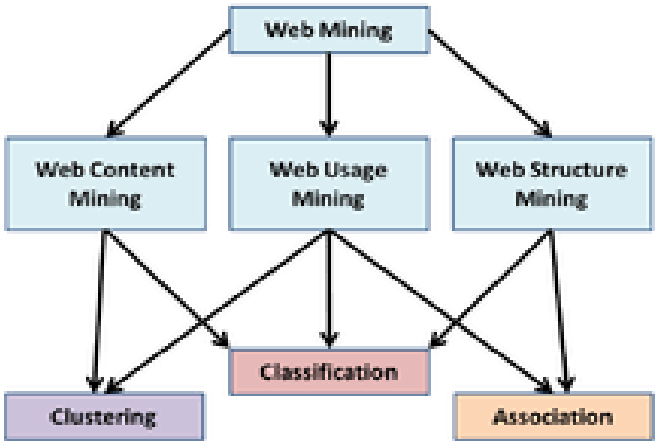

Figure 2: Data Mining techniques applied in Web Mining

A. Clustering

A small number of items on the WWW are Web pages that link to other pages via links, and these items primarily lack an integrated structure [2]. The clustering technique's primary goal is to group similar pages together according to the type of structure information used, such as hyperlinks, document structure, and link analysis. Thus, through keyword association and extracted contents, clustering enables linked Web pages to determine the relationship of other related pages and enables users to access the best information [4].

B. Classification

There are two types of classification in Web data: Link Based Classification and Content Based Classification. Classification is an automated Data Mining technique that aims to assign class properties from the set of a preset set of classes.

- **Link-Based Classification:** The most recent advancement in Web mining is link-based classification, also known as hyperlink classification. Its primary goal is to forecast the

category of a webpage based on the characteristics of the link (e.g., anchor text, other HTML tags, and links between Web pages). A hyperlink is a structural element that links one place on a Web page to another, either inside the same page or to other pages that are part of one or more Web sites. There are two categories of link (hyperlink) structures: intra-document hyperlinks and inter-document hyperlinks as in Fig. 3. While intra-document hyperlinks link various sections of the same Web page together, inter-document hyperlinks link distinct Web sites together [5].

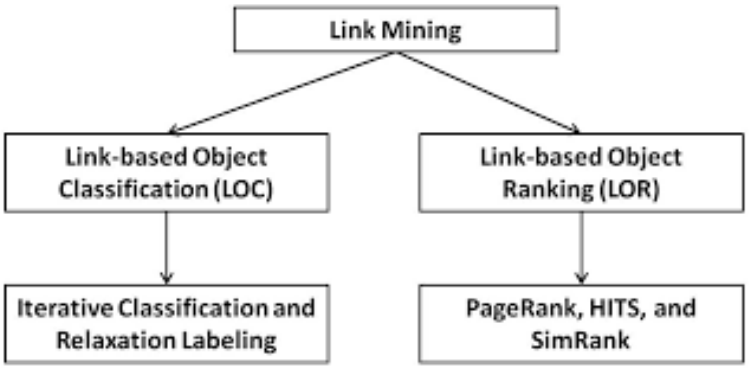

Figure 3: Approaches in Link based Web Structure Mining

- **Content Based Classification:** The goal of content-based classification is to organize Web pages according to the contents of links (hyperlinks) (the anchor text of the link). Because the class of Web sites may change due to the links that are dispersed throughout the Web page, each Web page is assigned to a class based on the words that appear in their links. This methodology is necessary for assigning the names in an iterative manner [3].

C. Association

Since World Wide Web Worms now index the anchor text parts (language that appears in hyperlinks) of the original Web pages, information contained in hyperlinks is crucial for retrieving results in search engines. Authorities' pages and hubs pages are the two categories that could be obtained in response to a client's request [3]. Using an algorithm known as HITS (Hyperlink Induced Topic Search), which addresses the problems of Web search engines, each Web page assigns two scores: an authority score and a hubs score.

While pages that link to a large number of authority Web pages are referred to as center pages and their useful resources on the Web, authority pages are those that contain important information on the subject of the inquiry. Therefore, the scores identified which page is a wonderful authority page if it contains links to many great authorities, and a reasonable center point page if it contains links to many great authorities. The pages are positioned according to Web search engine score esteems [3].

5. RESULT AND DISCUSSION

WSM's primary goal is to uncover previously hidden connections between the pages that belong to one or more websites. Link topology mining and link URL mining are the two distinct techniques used in WSM; they both make use of various unprocessed data and techniques. The Web is viewed as a chart in topology mining, with the hubs being web sites and the edges being connections between them. To create a more precise link patterns model, URL mining is

combined with link topologies for the source and target pages. A nonparametric indicator of the degree and direction of relationship between two web results on an ordinal scale is the Spearman correlation coefficient. The sign rs, or rho, is used to represent it.

Table 1: Comparison of Link Based Object ranking using Mean Spearman Correlation

No of Query	Mean Spearman Correlation		
	Page Rank	Hits	SimRank
10	0.13	0.15	0.19
15	0.19	0.21	0.28
20	0.24	0.28	0.36
25	0.31	0.37	0.48
30	0.43	0.49	0.56

Table 1 show the comparison of link based ranking algorithm based on the metric Mean Spearman Correlation across the different no of query. As the number of queries increases from 10 to 30, the Mean Spearman Correlation improves for all three algorithms, indicating that more query data leads to more stable and reliable ranking outcomes. SimRank consistently outperforms PageRank and HITS in all query sizes, indicating it produces rankings that more closely match the reference ranking. The experimental results demonstrate that SimRank is the most effective link-based ranking algorithm among the three in producing rankings aligned with the reference. As the number of queries grows, all methods improve, but SimRank exhibits superior scalability and correlation strength, making it a strong candidate for applications requiring high-accuracy web structure mining and ranking.

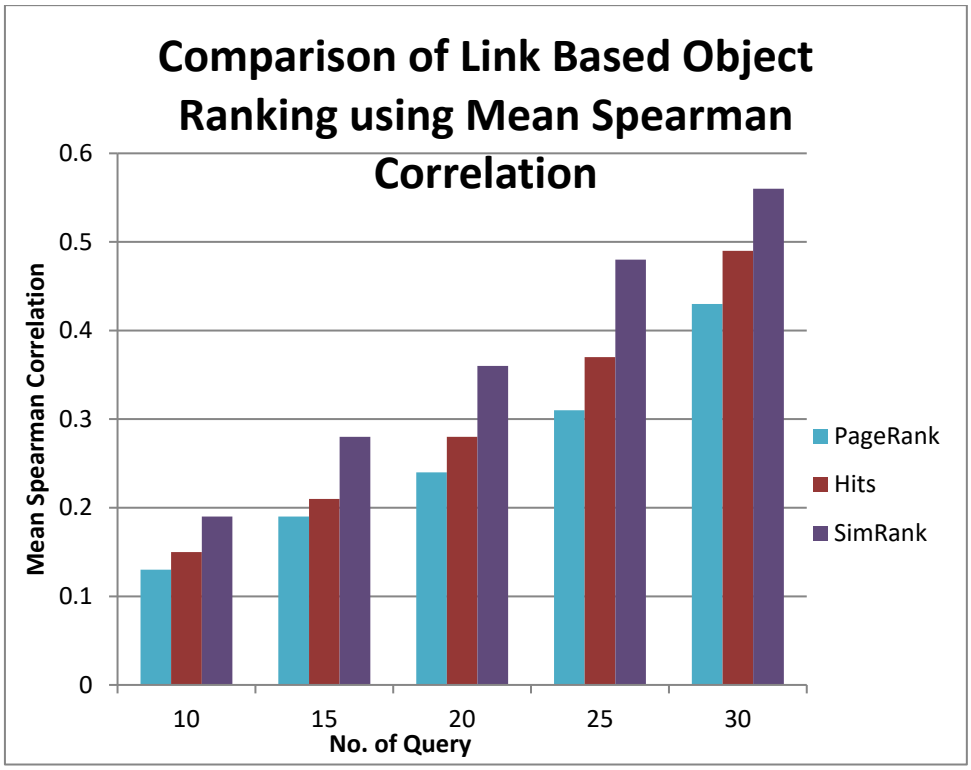

Figure 4: Comparison of Link Based Object Ranking using Mean Spearman Correlation

Fig. 4 shows the pictorial representation of comparing the 3 link based ranking algorithm based on metric Mean Spearman Correlation. SimRank is the most effective link-based object ranking algorithm among the three, particularly when more query data is available. HITS perform moderately well, better than PageRank but not as well as SimRank. PageRank, while foundational, trails behind in correlation strength, especially at higher query counts. It confirms that query volume significantly impacts ranking quality, and algorithms like SimRank that consider structural similarity are better suited for more complex web structure mining tasks.

6. CONCLUSION

In this paper it is closed with quickly portraying the fundamentals of PC innovation with its commitment inside different areas of mining of information over web alongside its different sorts and furthermore features barely any encouraging locales of the future examination. In this paper given the concise depiction about the web-structure kind of mining process (WSM) alongside its different working and furthermore clarifies types. This paper have portrayed about the different strategies of the web-mining idea. This sort of mining approaches has been demonstrated helpful in the business world. The report also discusses the methods used to extract information from different types of online data and how this information may be used for mining. Finally Link Based objects ranking methods were implemented and compared using the performance metrics Mean Spearman Correlation. Among the Ranking algorithms SimRank shows improved results.

REFERENCE

[1] Sharma, N., Boghey, R., & Prasad, R. (2024). A Review on Data Mining Issues, Solution & Techniques. International Journal For Multidisciplinary Research, 6(4). https://doi.org/10.36948/ijfmr.2024.v06i04.26654

[2] Pradeep, A. (2023). Web Mining: Opportunities, Challenges, and Future Directions. 1–6. https://doi.org/10.1109/conit59222.2023.10205913

[3] Pradeep, A. (2023). Web Mining: Opportunities, Challenges, and Future Directions. 1–6. https://doi.org/10.1109/conit59222.2023.10205913

[4] Sial, A. H. (2024). Web Content Mining: A Review on Concepts, Techniques, and Tools. https://doi.org/10.20944/preprints202407.2339.v2

[5] Choudhary, L., & Swami, S. (2023). Exploring the Landscape of Web Data Mining: An In-depth Research Analysis. Current Journal of Applied Science and Technology. https://doi.org/10.9734/cjast/2023/v42i244179

[6] Chopra, N. K., Mohan, C. R., Chaudhary, S., Kasar, M. M., Suryawanshi, T., & Dubey, S. K. (2025). Data Mining Techniques for Web Usage Mining. 259–270. https://doi.org/10.1002/9781394272464.ch18

[7] Prameswari, A. (2022). Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey. Mobile Information Systems, 2022, 1–19. https://doi.org/10.1155/2022/7519573

[8] Basit, M. Q., & Albarry, S. (2024). Web mining algorithms for social media analytic. International Journal of Computing, Programming and Database Management. https://doi.org/10.33545/27076636.2024.v5.i1a.94

[9] Knowledge Discovery in Web Usage Patterns Using Pageviews and Data Mining Association Rule (pp. 233–247). (2022). Smart innovation, systems and technologies. https://doi.org/10.1007/978-981-19-2541-2_19

[10] Marchi, V., Apicerni, V., & Marasco, A. (2021). Assessing Online Sustainability Communication of Italian Cultural Destinations – A Web Content Mining Approach (pp. 58–69). Springer, Cham. https://doi.org/10.1007/978-3-030-65785-7_5

[11] Li, S., Liu, F., Zhang, Y., Zhu, B., Zhu, H., & Yu, Z. (2022). Text Mining of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review. Mathematics, 10(19), 3554. https://doi.org/10.3390/math10193554