AN EFFECTIVE WEB PAGE RANKING BY APPLYING DIMENSION REDUCTION OVER WEB LOG DATA

[1]Monika Sahu, [2]Dr Megha Mishra, [3]Dr Vishnu Kumar Mishra
MTECH Student, Senior Assistant Professor, Associate Professor
Dept of Computer Science& Engineering
Shri Shankaracharya Technical Campus,(Shri Shankaracharya Group Of Institutions), Junwani Bhilai, Chhattisgarh, India

Abstract—Nowadays showcase is loaded with various inquiry apparatuses over web having perceptible decent variety as far as working and the end indexed lists. Given a query, seek devices ordinarily restore a substantial number of pertinent website pages. To be more successful, the returned pages must be positioned by their pertinence as for the client's question. Page Rank and Weighted Page Rank Algorithms give the proficient outcome however these calculations are question autonomous calculations as these depend on the web usage mining.Web Usage Mining techniques focuses the problem of fetching social patterns from one or more web usage logs.In this paper we will provide a comparisonamong different web usage mining techniques upon web log data emphasis on how web usage mining techniques helps to do effective web page ranking. Proposed algorithm applies dimension reduction over web log data which will reduce the time complexity of existing algorithm and increase the accuracy of web page ranking.

Keywords—Web log;

I. INTRODUCTION

The World Wide Web is the gathering of many interlinked hypertext reports which are gotten to by means of the Internet. Internet searcher is the one which encourages client to look through the archives in this expansive store according to the client's inquiry. These sought archives can be in hundreds which prompts the issue of choosing which report ought to seem in the first place, second et cetera. Keeping in mind the end goal to choose this, the idea of Web Page Ranking is utilized which fulfills the undertaking of giving a rank to each site page through a calculation and the site page with higher rank will seem first. Internet searcher has a procedure that goes from Crawling, Indexing, Searching and Ranking of data.

A crawler or spider is a program that seeks websites and scans the contents and other information of their pages in order to generate copy of the visited pages for a search engine index to provide fast searches. This process is called crawling for providing new or updated data that has been submitted by website owners. Indexer is a program that extracts the terms from web pages and creates an alphabetical order of terms. It also contains extra information such as URL of the page, frequency and position of the terms. It is also called as "Search Engine Database".

Search engine searches the web page in the index created by the indexer in response to the user query and aligns the web page links as per their page rank. The page link with higher page rank will be appeared earlier than the page link with lower page rank. Page Rank is a numeric value representing how important a page is on the web.There are various ranking algorithms which are based on web mining.

Web mining having different categories as follows:

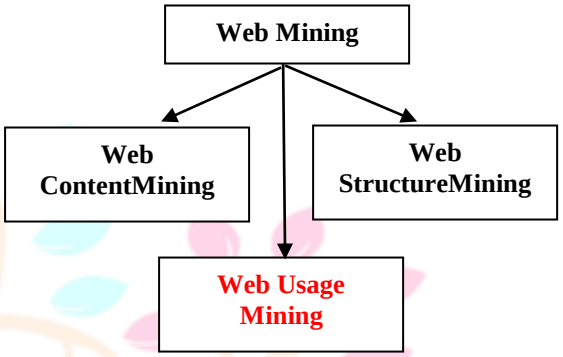

Fig.-1Categories of Web Mining

Web Usage Mining:It is the application of data mining techniques to determine interesting usage patterns from Web log data, in order to appreciate and better serve the needs of Web based applications. Usage data reveals the intention, identity and origin of Web users along with their browsing behavior at a Web site [V.Chitraa et. Al.]. Web Usage Mining consists of three phases as shown in Figure-2 which are named data processing, pattern discovery and pattern analysis.

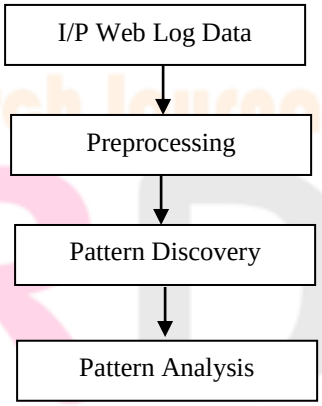

Fig.-2. Phases of Web Usage Mining

Data preprocessing has a fundamental role in Web Usage Mining applications. The preprocessing of web logs is usually complex and time demanding. It comprises four different tasks: (i) the data cleaning, (ii) the identification and the reconstruction of users' sessions, (iii) the retrieving of information about page content and structure, and (iv) The data formatting.

The pattern discovery phase relies on various statistical methods and data mining algorithms to detect interesting patterns. Most of the commercial applications of Web Usage Mining exploit consolidated statistical analysis techniques. In contrast, research in this area is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of web usage data. Most of this research effort focuses on three main paradigms: association rules, sequential patterns, and clustering.

Further in this paper we will in section II we will elaborate different literature and will give tabular comparison among literature, in section III we will discuss problem identified in earlier web page ranking scheme and some bottlenecks, in section IV we will elaborate our proposed methodology, further in section V will give dataset description and experimental evaluation, at last section we will conclude.

II. LITERATURE SURVEY

B.Naveena Devi et. al. [Elsevier 2012] said that the importance of web usage mining is unquestionable with the rising importance of the web not only as an information portal but also as a business edge. Web access logs contain abundant raw data that can be mined for web access patterns, which in turn can be applied to improve the overall surfing experience of users. By taking into consideration we have mainly focused on designing of web usage mining intelligent system for clustering of user behaviors using agglomerative clustering algorithm.Author introduces a web usage mining intelligent system to provide taxonomy on user information based on transactional data by applying data mining algorithm, and also offers a public service which enables direct access of website functionalities to the third party.

Mehmet Lutfi ARSLAN et. al. [ARXIV 2013] attempts to develop an online reputation index of Turkish universities through their online impact and effectiveness. Using 16 different web based parameters and employing normalization process of the results, we have ranked websites of Turkish universities in terms of their web presence. This index is first attempt to determine the tools of reputation of Turkish academic websites and would be a basis for further studies to examine the relation between reputation and the online effectiveness of the universities.

Jun Yu et. al. [IEEE 2014] propose a new multimodal hypergraph learning based sparse coding method for the click prediction of images. The obtained sparse codes can be used for image re-ranking by integrating them with a graph-based schema. We adopt a hypergraph to build a group of manifolds, which explore the complementary characteristics of different features through a group of weights. Unlike a graph that has an edge between two vertices, a set of vertices are connected by a hyperedge in a hypergraph. This helps preserve the local smoothness of the constructed sparse codes. Then, an alternating optimization procedure is performed and the weights of different modalities and sparse codes are simultaneously obtained using this optimization strategy. Finally, a voting strategy is used to predict the click from the corresponding sparse code.

Nidhi Grover et. al. [IJERT 2010] compare two popular web page ranking algorithms namely: HITS algorithm and PageRank algorithm. The paper highlights their variations, respective strengths, weaknesses and carefully analyzes both these algorithms using simulations developed for both. Author concluded that both page rank and HITS algorithm are different link analysis algorithms that employ different models to calculate web page rank. Page Rank is a more popular algorithm used as the basis for the very popular Google search engine. This popularity is due to the features like efficiency, feasibility, less query time cost, less susceptibility to localized links etc. which are absent in HITS algorithm. However though the HITS algorithm itself has not been very popular, different extensions of the same have been employed in a number of different web sites.

V.Chitraa et. al. [IJCSA 2014] proposed a novel approach based on Fuzzy C Means in fuzzy environments is proposed to cluster the web user transactions. This approach is groups the similar user navigation patterns. The algorithm enhances the FCM, and Penalized FCM clustering algorithm by adding Posterior Probability to find highest membership for a member to add in a cluster. Classification is carried out by SVM and RVM for classifying a new user to a particular group. The method is experimented and evaluated and found it is better method for clustering than the existing methods.

Neeraj Raheja et. al. [IJCSI 2014] proposes an approach for web usage mining based upon web log partition. It takes less time and provides popular results in accordance with the existing approach. Some more results may be obtained if the number of cluster formed are changed i.e. from 4 clusters formed in our approach can be changed to 6, 8 or more. However recall and precision may be affected by changing the number of clusters i.e. either may be improved or decayed.

S. No.	Author/Paper title/Year	Name of Algorithm/Tool/ Method(discussed /Implemented)	Description	Accuracy
1.	Jihyun Lee et. al./Effective ranking and search techniques for Web resources considering semantic relationships/Elsevier 2013	Weighting measure for the semantic relationship	Propose a novel ranking method which considers the number of meaningful semantic relationships between a resource and keywords as well as the coverage and discriminating power of keywords. In order to improve the efficiency of the search, author prune the unnecessary search space using the length and weight thresholds of the semantic relationship path.	90% and Fast Processing due to pruning algorithm applied.
2.	Zhifeng Bao et. al./Effective XML Keyword Search with Relevance Oriented Ranking/IEEE 2009	XML TF*IDF ranking strategy	Propose specific guidelines that a search engine should meet in both search intention identification and relevance oriented ranking for search results. Then based on these guidelines, author design novel formulae to identify the search for nodes and search via	F-measure 0.40

			nodes of a query, and present a novel XML TF*IDF ranking strategy to rank the individual matches of all possible search intentions.	
3.	B.Naveena Devi et. al./Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce/ Elsevier 2012	Agglomerative clustering algorithm	Author introduces a web usage mining intelligent system to provide taxonomy on user information based on transactional data by applying data mining algorithm, and also offers a public service which enables direct access of website functionalities to the third party.	-
4.	Jun Yu et. al. [IEEE 2014] /Click Prediction for Web Image Reranking Using Multimodal Sparse Coding/IEEE 2014	Multimodal Sparse Coding	Experimental results on real-world data sets have demonstrated that the proposed method is effective in determining click prediction. Additional experimental results on image re-ranking suggest that this method can improve the results returned by commercial search engines.	Effective in determining click prediction.
5.	Dawei Yin et. al./Ranking Relevance in Yahoo Search/KDD '16, August 13-17, 2016, San Francisco, CA, USA	Base ranking function	Introduce three key techniques for base relevance – ranking functions, semantic matching features and query rewriting. Author also describe solutions for recency sensitive relevance and location sensitive relevance. This work builds upon 20 years of existing efforts on Yahoo search, summarizes the most recent advances and provides a series of practical relevance solutions.	-
6.	Petar Ristoski et. al./Mining the Web of Linked Data with RapidMiner/Preprint submitted to Journal of Web Semantics May 11, 2015	RapidMiner	Author introduce the RapidMiner Linked Open Data extension. The extension hooks into the powerful data mining and analysis platform RapidMiner, and offers operators for accessing Linked Open Data in RapidMiner, allowing for using it in sophisticated data analysis workflows without the need for expert knowledge in SPARQL or RDF.	-

III. PROBLEM IDENTIFICATION

After going to various literature we have come across some bottle necks of existing web page ranking algorithms.

Search engine searches the web page in the index created by the indexer in response to the user query and aligns the web page links as per their page rank.The page link with higher page rank will be appeared earlier than the page link with lower page rank. Page Rank is a numeric value representing how important a page is on the web.

Existing web page ranking techniques based on inbound and outbound of web pages, which does not provided accurate web page rank, further our ranking algorithms such as weighed web page ranking, which evaluates weight of outbound link known as weighted page ranking:

$$WL(V,u) = I(u) / \sum_{peR(v)} I(p)$$

Where I(u) and I(p) represent the number of incoming links of page u and page p respectively. R(v) represents the reference page list of page v.

We have identified some bottle necks of existing algorithms which are as follows:

- PageRank works by calculating the number and quality of web links to a page to govern a rough approximation of how significant the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. Existing algorithms are not capable of bypass search engine optimization (SEO) tools, SEO tools bluff the crawler programs and increases the page rank of websites.
- Web access dataset has gigantic number of features which are irrelevant or redundant.
- There is difficulty of inadequate increase in dimension of web access data set.
- Existing algorithm does not consider user behavior for evaluating web page ranking.
- Huge dataset size decreases the computation performance.
- Less appropriate web log data reduces the performance of page ranking algorithm.
- Lesser value of precision of existing algorithms.

Precision and recall are the most collective procedures for estimating an classification systems. Precision is the percentage of returned documents that are targets, whereas recall is the percentage of target documents reverted.

Precision = Good messages kept / All messages kept ……. (1)

Recall =Good messages kept / All good messages …… (2)

F-measure = 1 / (average (1/precision, 1/recall)) …… (3)

IV. PROPOSED METHDOLOGY

To overcome the problems of existing algorithms we have proposed an algorithm which will reduce the time complexity, increases the value of precision and by applying preprocessing over input dataset elimination of missing data and redundant data removal has been done.

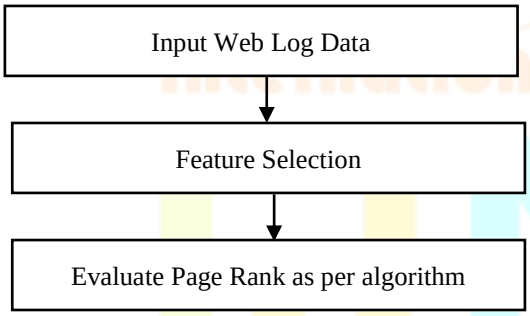

Fig.-3Earlier System

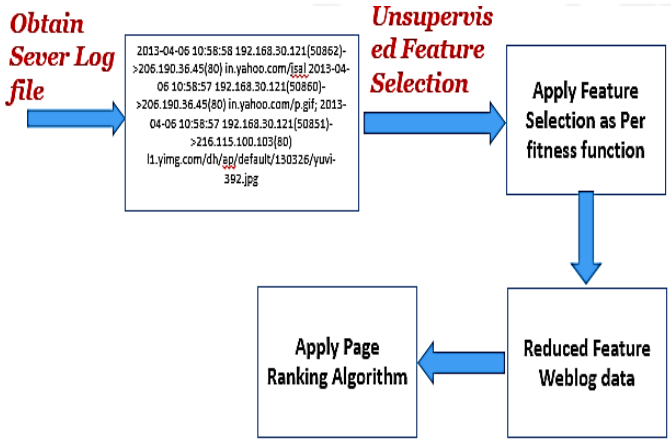

Fig.-4 Proposed System Flow Chart

Proposed Algorithm

Step-1. Input web log dataset.

Step-2. Apply preprocessing algorithm.

// as we have to concentrate on some of the parameters of web log henceforth we need to suppress rest fields.

Step-3. Apply dimension reduction over web log data and selecting feature as per fitness function.

Step-4. Calculating web page ranking of each URL of web log by multiplying dapping factor.

P(r) = Damping Factor* Frequency of Page* Time duration

Step-5. Calculating the value of precision

Damping (or Dampening) factor

To understand damping factor let's take the same simple set-up that we used when we looked at the initial values: Page1 links to all pages - and all pages link back.

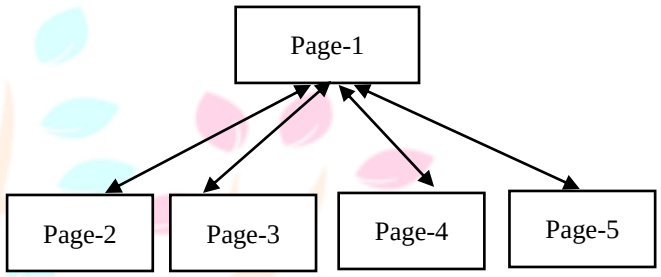

Fig.-5 Link of Web Pages

Some observation on damping factor are:

A low damping factor (= much damping) will make calculations easier. Since the flow of PageRank is dampened the iterations will quickly converge.

- A low damping factor (= much damping) means that the relative PageRank will be determined by PageRank received from external pages - rather than the internal link structure.
- A high damping factor (= little damping) will result in the site's total PageRank growing higher. Since there is little damping, PageRank received from external pages will be passed around in the system. It will not grow forever though - the maximum limit is Inbound PageRank * d/(1-d).

The value recommended in the original paper is 85%. Google also uses 0.85 as damping factor.

V. RESULT AND DISCUSSION

For implementation of existing algorithm we have used JDK 1.8 64 bit, and Mysql for storing reduced dimension dataset. We have used apache web usage dataset.

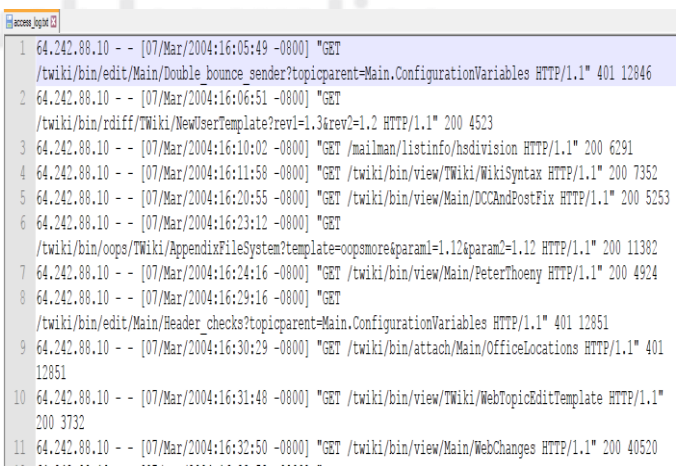

Fig.-6 Apache Web Access dataset

Fig.-7 Main UI of Proposed System

Fig.6 is the snippet of the of our proposed system in which there is option to browse the web access data set file, then process then process that dataset through earlier approach and proposed approach, at finally calculation of value of precision of both systems.

Earlier System Output Values

S. No.	File Name	Time complexity	Precision
1.	log.txt	500ms	5%
2.	access_log.txt	5275ms	97%
3.	access_log_1.txt	5768ms	97%

Proposed System Output Values

S. No.	File Name	Time complexity	Precision
1.	log.txt	300ms	15%
2.	access_log.txt	5075ms	100%
3.	access_log_1.txt	5118ms	100%

Note: Files in a table are in increasing order of file size

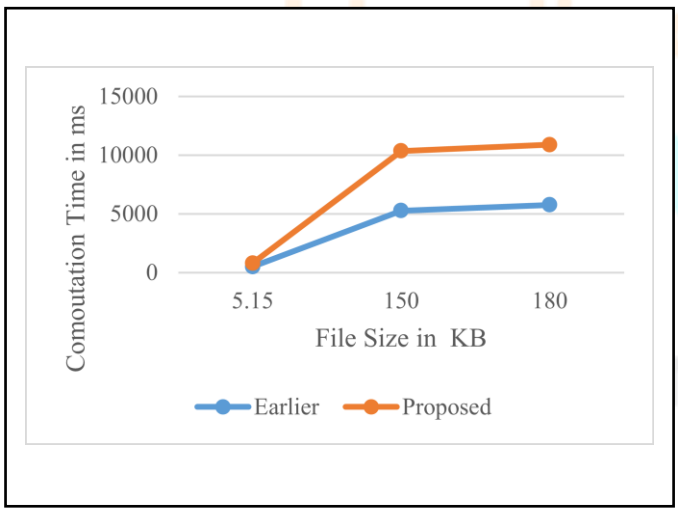

Fig-8 Performance comparison as per time complexity

VI. CONCLUSION

Web Usage Mining process is Pattern analysis. After discovering patterns from usage data, a further analysis has to be conducted. Using pattern discovery, the pattern set is found and then pattern analysis is performed to select the interesting patterns and to filter out uninteresting patterns.

In future we can integrate the existing page rank with our proposed algorithm. Further we can apply our algorithm to web image raking.

VII. ACKNOWLEDGMENT

Expression of giving thanks are just a part of those feeling which are too large for words, but shall remain as memories of wonderful people with whom I have got the pleasure of working during the completion of this work. I am grateful to "**Shri Shankaracharya Technical Campus Bhilai**" which helped me to complete my work by giving encouraging environment. I would like to express my deep and sincere gratitude to my supervisor, **"Dr. Megha mishra**"and **"Dr.Vishnu Kumar Mishra"**. the wide knowledge and his logical way of thinking have been of great value for me

REFERENCES

[1] Jihyun Lee et. al./Effective ranking and search techniques for Web resources considering semantic relationships/Elsevier 2013.

[2] Zhifeng Bao et. al./Effective XML Keyword Search with Relevance Oriented Ranking/IEEE 2009.

[3] B.Naveena Devi et. al./Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce/Elsevier 2012.

[4] Jun Yu et. al. [IEEE 2014] /Click Prediction for Web Image Reranking Using Multimodal Sparse Coding/IEEE 2014.

[5] Dawei Yin et. al./Ranking Relevance in Yahoo Search/KDD '16, August 13-17, 2016, San Francisco, CA, USA.

[6] Neeraj Raheja et. al./Efficient Web Data Extraction Using Clustering Approach In Web Usage Mining/IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 1, No 2, January 2014.

[7] Petar Ristoski et. al./Mining the Web of Linked Data with RapidMiner/Preprint submitted to Journal of Web Semantics May 11, 2015.

[8] Alani, H., Brewster, C., & Shadbolt, N. (2006). Ranking ontologies with AKTiveRank. In Proceedings of the 5th International Semantic Web conference. Lecture notes in computer science (Vol. 4273, pp. 1–15). Springer.

[9] Anyanwu, K., Maduko, A., & Sheth, A. (2005). SemRank: Ranking complex relationship search results on the Semantic Web. In Proceedings of the 14th international conference on World Wide Web (pp. 117–127). ACM (May).

[10] Anyanwu, K., & Sheth, A. (2003). q-Queries: Enabling ouerying for semantic associations on the Semantic Web. In Proceedings of the 12th international conference on World Wide Web (pp. 690–699). ACM (May).

[11] Castells, P., Fernandez, M., & Vallet, D. (2007). An adpatation of the vector-space model for ontology-based information retrieval. IEEE Transactions on Knowledge and Data Engineering, 19(2), 261–272 (February).

[12] Ceravolo, P., & Damiani, E. (2007). Bottom-up extraction and trust-based refinement of ontology metadata. IEEE Transactions on Knowledge and Data Engineering, 19(2), 149–163 (February).