© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 07 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

DWPR ALGORITHM FOR PAGE RANKING USING WEB LINK ANALYSIS

[1]Anshu Baliyan, Research Scholar, Department of CSA, CDLU Sirsa
[2]Avininder singh, Assistant professor, Department of CSA, CDLU Sirsa

Abstract: This research paper studies the importance of data mining in the field world wide web. The main objective of the thesis is to design the algorithm for efficiently rank the web page based multiple parameters that are necessarily required to rank the web page. Data or information of the web refers to that what type of data should be included; data should have different types so that it can attract the user like audio, video, images, animations and graphics. Usage information includes the data present on server logs and web logs. This information is very beneficial for web developers. Web developers take advantage of such information for making the web more interesting and more useful to user by analyzing the usage information. Depending upon the categorization of WWW data, web mining is also has three categories i.e. web usage mining, web content mining and web structure mining or web link analysis is used to deal with the complex web information.

Keywords: WWW, DWPR, Web link analysis, Data mining, web structure mining

[1] INTRODUCTION

Mining is an amazingly valuable term in numerous fields. In laymen terms, mining can be defined as extraction of valuable liquids, minerals, gases or other geological minerals which exists as veins, liquids, seams, and ore bodies in the earth. Ores that are discovered by the mining contains clay, limestone, coal, metals, oil, gemstone, potash, rock salt and gravel. Performing the mining in the field of data is termed as Data Mining. Data miningis slightly different from information retrieval in the way of extracting the data. Data mining is to extract the useful data from the managed set of data while information retrieval is the retrieval of the relevant information resources from the large set of information resource.

Mining the data is very important to solve many hazardous problem related to the performance of the large databases maintained by large size organizations like colleges, companies, banks, hospitals, universities, railways etc. These organization's databases have large amount of data that is also complex to deal without data mining. This huge collection of data is very complicated to handle in the absence of data mining. It may be referred to extract useful patterns and trends form the large databases for use in analysis and decision making. This is where data mining is useful.

DATA MINING

Data Mining (Also called as Knowledge Discovery) is process of extracting useful pattern or information or knowledge from the large collection of data. This huge data is stored in Data Warehouse, it is a term given to a large database. Data warehouse may contain billions, millions or trillions of data which is collected in several years (Han &Kambert 2001).

High level information, interesting knowledge, or some relevant pattern and required data can be extracted by knowledge discovery. This discovered information can be used in decision making, query processing, process control, information management and in many other processes.

Data mining is a field in which researchers of many other fields like knowledge acquisition, database system, machine learning, artificial intelligence etc show interest. Other application like World Wide Web, online services use many techniques of data mining. Data mining is basically way to take advantage of historically collected huge data. This is useful to analyze historically data in a meaningful way. Mining of data can be done by firstly digging the data from the data warehouse then analyzing the different set of data extracted in digging process and then finally obtaining the meaning of the data. Huge collection of historical data present in every organization whether it is an educational institutions, companies, banks, railways, hospitals, world wide web etc (Chen & Han 1996).

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 07 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

Before understanding the details of data mining it is necessary to understand the difference between data, information and knowledge.

DATA, INFORMATION AND KNOWLEDGE

Data is any unprocessed facts, text, numerical or any such collection of words that can be further processed by computer. This data is growing day by day explosively and presented in many different formats and different databases in organization. Some forms of data are operational data (Accounting, cost, sales, payroll etc.), non-operational data (Forecast data, macro economic data etc.) and Meta data (data about data).

[2] LITERATURE REVIEW

In recent years the growth of the World Wide Web exceeded all expectations. Today there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task. So, the World Wide Web is a fertile area for data mining research. (Pandia 2011)

Link-based analysis of the Web provides the basis for many important applications—like Web search, Web-based data mining, and Web page categorization—that bring order to the massive amount of distributed Web content.(Caverlee 2009)

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

[3] TOOLS AND TECHNOLOGY

From the literature survey it is clear that working with the link structure of web is a very complex task. The main challenges are to handle the huge, dynamic, heterogeneous data as well as the links of the WWW. Appling data mining techniques to deal with complexity of the web is very fertile area for researchers. Web mining is the application of data mining which automatically determinesand extracts information from documents and services present on WWW. There exist many algorithms for web structure mining. Detailed study of web structure mining or web link analysis algorithms reveals that each algorithm is based on various parameters. Each algorithm has a different parameter which makes that algorithm different from other.

JAVA ENVIRONMENT:

JAVA is a very simple, secure, robust, portable, dynamic, and multithreaded language. JAVA is class based and object oriented and developed to have as minimum implementation dependencies as possible. JAVA facilitates application developers to develop “write once, run anywhere” (WORA) code. It means that application that works on one platform can work on another without recompilation. These are some main and important features of JAVA which makes it a very useful language for many areas.

[4]PROPOSED WORK

To provide an efficient algorithm for web structure mining it is necessary to work upon all the possible parameters. Once the algorithm is designed, it should be analyzed over some specific domain of websites. In this research we propose an algorithm, and then analyze its behavior over the education domain websites.

PROPOSED PROCESS TO ACHIEVE THE OBJECTIVES:--

A step by step process to achieve the first objective is:

- Analysis of all the existing algorithms.
- Clearly identify the areas which can be enhanced to give better result.
- Identify the parameter on which the new algorithm can be depends on.
- Propose an algorithm that includes the identified parameters.

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 07 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

- Compare the proposed algorithm with the result of existing algorithms.

A process to analyze the result of proposed algorithm on education domain websites:

- Select 5 websites of education domain.
- Detailed study of the structure of the websites.
- Simplify the structure of selected websites up to the maximum possible limit.
- Apply the proposed algorithm on simplified structure of the websites to analyze the behavior of the algorithms.

[5]RESULT AND DISCUSION

The existing and the proposed algorithms are developed in JAVA environment and results are compared. The algorithms have been implemented on websites in the education domain. Each website has been converted into graphical representation (nodes and links) and the algorithms are implemented on them.

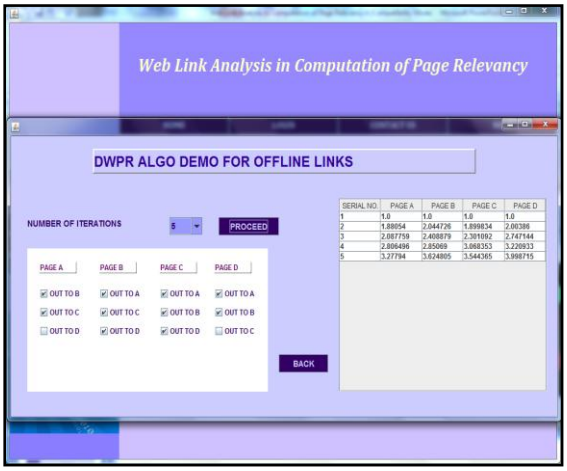

Fig 1. Snapshot of results of the DWPR Algorithm

We have taken five websites of AKGEC, Galgotia University, BITS Plani, NIT and MNNIT. These all websites are analyzed and simplified up to the maximum level it could be and then all algorithms are applied on these websites. All these websites are from the education domain. However the structure for the websites is diverse and this has facilitated better verification.

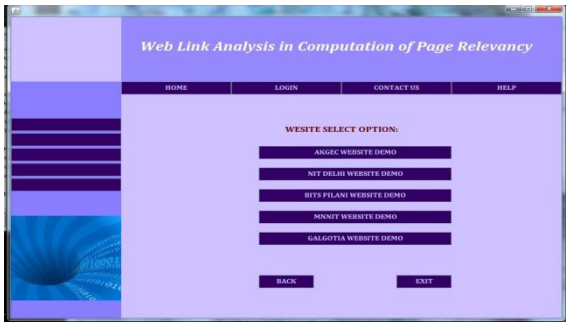

Fig 2. Snapshot of interface a page containing all 5 structures of websites

Fig 3: Snapshot of simplified structure AKGEC website

Fig 4: Snapshot of results of Page Rank Algorithm applied on AKGEC website

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 07 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

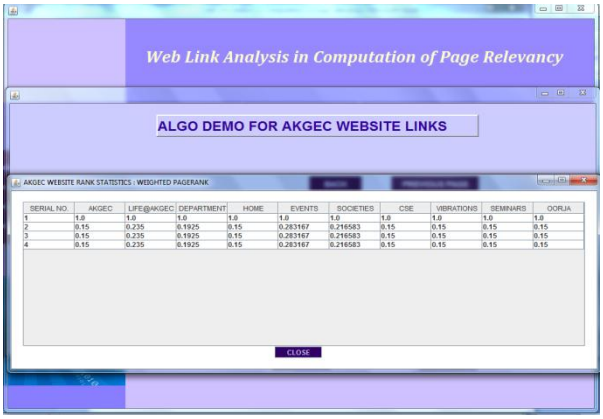

Fig 5: Snapshot of results of Weighted Page Rank applied on AKGEC website

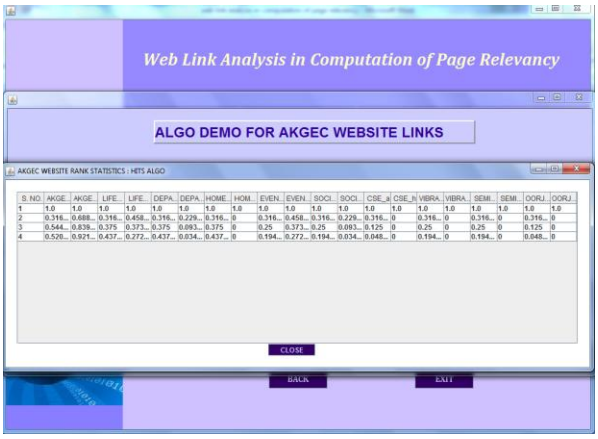

Fig 6: Snapshot of results of HITS Algorithm applied on AKGEC website

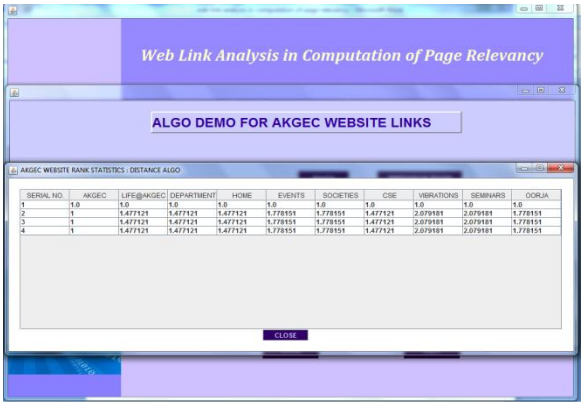

Fig 7: Snapshot of results of Distance Rank Algorithm applied on AKGEC website

Fig 8: Snapshot of results of DWPR Algorithm applied on AKGEC website

Comparison analysis for AKGEC Tree structure-

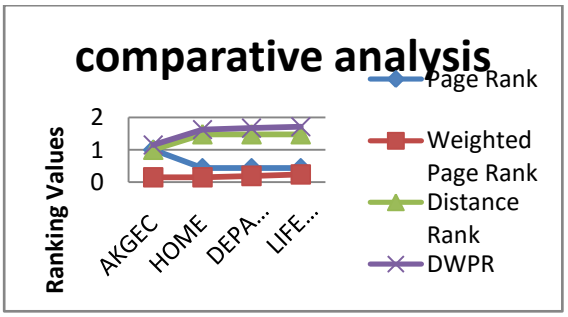

Fig 9 Comparative analysis for AKGEC Structure

[6] FUTURE SCOPE AND CONCLUSION

The study of web mining and algorithms used in web mining reveals that each algorithm has its own limitation. In this research we propose a new algorithm by considering maximum numbers of parameters. Analysis of the calculated results shows that the proposed algorithm gives better result among all algorithms.

New algorithm considers incoming links, outgoing links, distance, and content of the web page. These parameters are the basic parameters of a web page. The proposed algorithm achieves better page rank in same complexity.

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 07 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

The proposed algorithm is firstly analyzed on a graph with 4 nodes and with dynamic links. But when we talk about the WWW, the structures present on WWW are not always a graph structure. The structures of a websites are tree like structure. We have analyzed the proposed algorithm on some websites of education domain. We have taken five websites and then applied the proposed algorithm on tree like structures of all the websites. After this analysis we can conclude that the proposed formula works efficiently on tree structures also.

REFERENCE

[1] Abiteboul, S., Buneman, P., and Suciu, D., Data on the Web: From Relations to Semistructured Data and XML, Morgan Kaufmann, San Francisco, 2000.

[2] Bhamidipati, N.L et al., "Comparing Scores Intended for Ranking", In IEEE Transactions on Knowledge and Data Engineering, 2009.

[3] Bidoki, A.M.Z, Yazdani, N., "DistanceRank: An Iintelligent Ranking Algorithm for Web Pages", Information Processing and Management, ELSEVIER, 2007

[4] Borkar, V.R., Deshmukh, K., and Sarawagi, S., "Automatic Segmentation of Text into Structured Records," *Proc. ACM-SIGMOD Int'l Conf. Management ofData* (SIGMOD 2001), ACM Press, New York, 2001, pp. 175-186.

[5] Brin, S., Page, L., "The Anatomy of a Large-scale Hypertextual Web Search Engine", Proceedings of the Seventh International World Wide Web Conference, 1998.

[6] Chakrabarti, S. et al., "Mining the Web's Link Structure," *Computer*, Aug. 1999, pp. 60-67.

[7] Chakrabarti, S., Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data, Morgan Kaufmann, San Francisco, 2002.

[8] Chaudhary, R., Bhusry, M., 2013, "Analysis of Web Link Algorithms for Web Page Ranking", published in proceedings IEEE International conference ICCCCM, Allahabad, India.

[9] Chaudhary, R., Bhusry, M., 2014, "A New Contrive to Evaluate Web Page Ranking", IEEE International Conference on Electronics and Communication System, ICECS-2014, KCE, Coimbatore, India.

[10] Chaudhuri, S. and Dayal, U., "An Overview of Data Warehousing and OLAP Technology," *SIGMOD Record*, vol. 26, no. 1, 1997, pp. 65-74.

[11] Chen, M. S., Han, J. and Yu, P.S., 'Data Mining: An Overview from a Database Perspective', IEEE transaction on knowledge and data engineering, Vol 8, No. 6, December 1996.

[12] Clifton, C. and Marks, D., 'Security and Privacy Implications of Data Mining', Proc 1996 SIGMOD '96 Workshop Research Issues Data Mining and Engineering and Knowledge Discovery (DMKD '96), pp. 15-20, Montreal, Canada, June 1996.

[13] Duhan, N., Sharma, A.K. and Bhatia, K.K., "Page Ranking Algorithms: A Survey", in proceedings of the IEEE International Advanced Computing Conference (IACC), 2009.

[14] Fujimura, K., Inoue, T. and Sugisaki, M., "The EigenRumor Algorithm for Ranking Blogs", In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem, 2005.

[15] Han, J. and Kambert, M., 'Data Mining: Concepts and Techniques', Morgan Kaufmann, San Francisco,2001.