# Comparative Study of Web Pages Classification Methods

**Karunendra Verma[1], Dr. Prateek Srivastava[2], Dr.Prasun Chakrabarti[3]**

[1, 2, 3] Dept of CSE

[1, 2, 3] Sir Padampat Singhania University, Udaipur,India

***Abstract***- *The web is a huge database of information and there is a need for web pages classification to facilitate the indexing, search and retrieval. Web page classification is significantly different from traditional full text classification because of the existence of some additional information provided by the HTML structure. It has been analyzed that there are various way of web page classification but classification take higher time to compute with lesser accuracy. So as to make web page classification method, there is need to design efficient algorithm in order to reduce time and increase web page classification result.*

*Data Mining is an automated or semi automated exploration and analysis of large volume of data in order to reveal meaningful patterns. The term web mining is the discovery and analysis of useful information from World Wide Web that helps web search engines to find high quality web pages and enhances web click stream analysis. One branch of web mining is web structure mining. The goal of which is to generate structural summary about the Web site and Web pages. Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. In this paper various web page classification strategies are comparing.*

***Keywords***- Web mining, Web content mining, Web structure mining, and Web usage mining

## I. INTRODUCTION

Now a day World Wide Web becomes very popular and interactive for transferring of information. The web is huge, diverse and active and thus increases the scalability, multimedia data and temporal matters. The growth of the web has outcome in a huge amount of information that is now freely offered for user access. The several kinds of data have to be handled and organized in a manner that they can be accessed by several users effectively and efficiently.

The web is a collection of interrelated files on one or more Web servers. Web mining is the application of data mining techniques to extract knowledge from Web data including Web documents, hyperlinks between documents, usage logs of web sites etc.

Web mining broadly divided into three categories: web structure mining, web content mining and web usage mining.

### 1.1 Web Structure Mining

Web structure mining [1] is the process of using graph theory to analyze the node and connection structure of a web site.

According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

The goal of the Web Structure Mining is to generate the structural summary about the Web site and Web page. It tries to discover the link structure of the hyper links at the inter - document level. Based on the topology of the hyperlinks, WSM classify web pages and generates related patterns, such as the similarity and the relationships between different Web sites. There are various work in this domain of Structure-Based Classification [8] of Web Documents .Classification of web page content is important to many tasks in fetching the web information and organizing the web directories and focused creeping. But the huge content of web which are present today bring a great challenge to web page classification as compared with the traditional text base classification, but the present interconnectivity the nature of hypertext also provides function that can assist the process.

Type of WSM

There are number of algorithms based on the link analysis. Out of them three algorithms Page Rank, weighted pager rank and HITS are discussed.

### a) Page Rank Algorithm

This algorithm [2] is used to determine the importance of website pages and it is developed by Brin and Page at Stanford University. It works by counting the number and quality of links to determine a rough estimate of how important the website is. It allocates a numerical weight to each element of a hyperlinked set of documents. The link from one page to another page is considered as a vote. Not only is the number of votes that a page receives important but the importance of pages that casts the vote is also important.

### b) HITS (Hyper -link Induced Topic Search) algorithm

HITS is a link analysis algorithm that rates web pages developed by Jon Kleinberg [2]. It is also knows as Hubs and Authorities. A good hub represented a page that pointed to many other pages and a good authority represented a page that was linked by many different hubs.

### c) Weighted Page Rank algorithm

It is develop by Wenpu Xing and Ali Ghorbani [2]. It is an addition of Page Rank algorithm. Page Rank and HITS algorithms treat all links equally when distributing rank scores. But this algorithm considers the importance of both in-links and out-links of the pages and distributes rank scores [9] based on the popularity of the pages.

### 1.2 Web Usage Mining

Web Usage Mining [18] is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

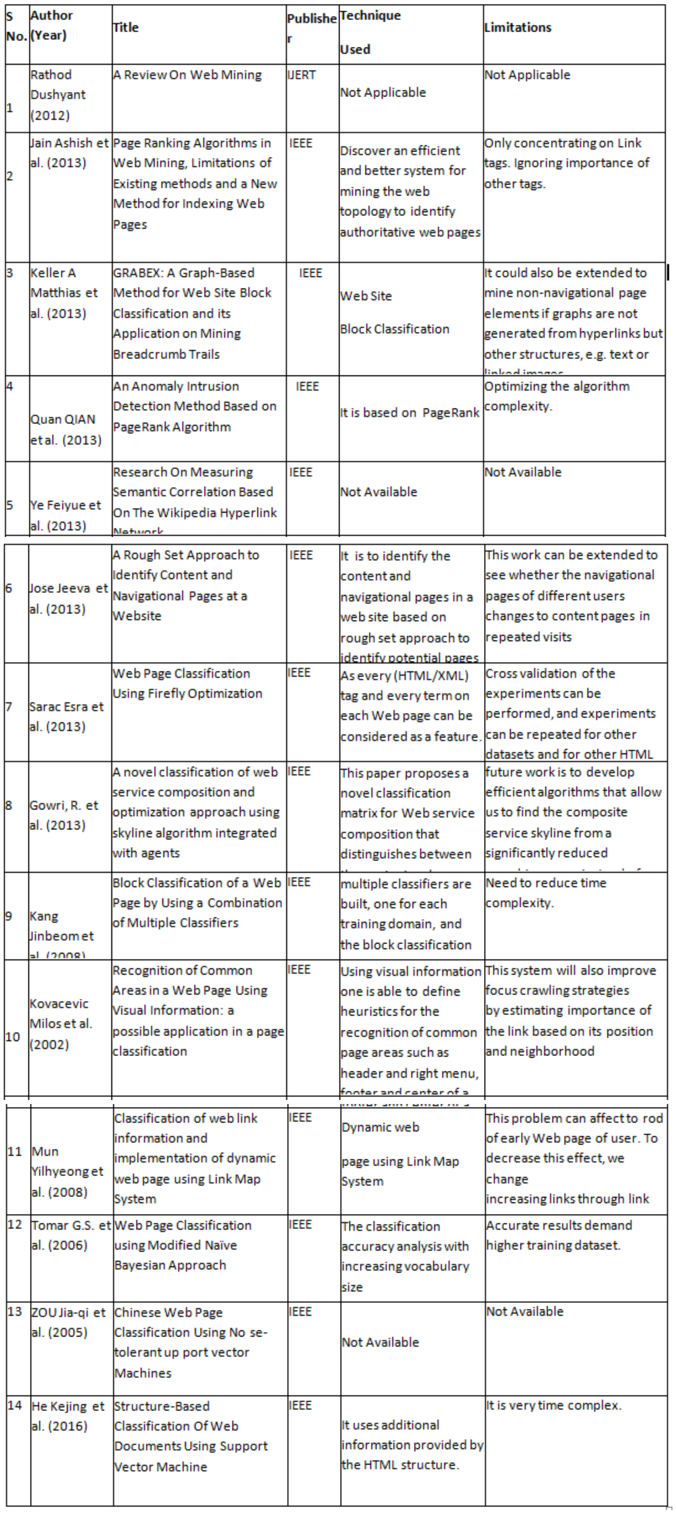
Web usage mining itself can be classified further depending on the kind of usage data considered:

1. Web Server Data
2. Application Server Data
3. Application Level Data

### 1.3 Web Content Mining

Web content mining [2] is the mining, extraction and integration of useful data, information and knowledge from Web page content.

## II. REVIEW OF LITERATUR



| S No. | Author (Year) | Title | Publisher | Technique Used | Limitations |
|---|---|---|---|---|---|
| 1 | Rathod Dushyant (2012) | A Review On Web Mining | IJERT | Not Applicable | Not Applicable |
| 2 | Jain Ashish et al. (2013) | Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages | IEEE | Discover an efficient and better system for mining the web topology to identify authoritative web pages | Only concentrating on Link tags. Ignoring importance of other tags. |
| 3 | Keller A Matthias et al. (2013) | GRABEX: A Graph-Based Method for Web Site Block Classification and its Application on Mining Breadcrumb Trails | IEEE | Web Site<br>Block Classification | It could also be extended to mine non-navigational page elements if graphs are not generated from hyperlinks but other structures, e.g. text or linked images |
| 4 | Quan QIAN et al. (2013) | An Anomaly Intrusion Detection Method Based on PageRank Algorithm | IEEE | It is based on PageRank | Optimizing the algorithm complexity. |
| 5 | Ye Feiyue et al. (2013) | Research On Measuring Semantic Correlation Based On The Wikipedia Hyperlink Network | IEEE | Not Available | Not Available |
| 6 | Jose Jeeva et al. (2013) | A Rough Set Approach to Identify Content and Navigational Pages at a Website | IEEE | It is to identify the content and navigational pages in a web site based on rough set approach to identify potential pages | This work can be extended to see whether the navigational pages of different users changes to content pages in repeated visits |
| 7 | Sarac Esra et al. (2013) | Web Page Classification Using Firefly Optimization | IEEE | As every (HTML/XML) tag and every term on each Web page can be considered as a feature. | Cross validation of the experiments can be performed, and experiments can be repeated for other datasets and for other HTML |
| 8 | Gowri, R. et al. (2013) | A novel classification of web service composition and optimization approach using skyline algorithm integrated with agents | IEEE | This paper proposes a novel classification matrix for Web service composition that distinguishes between | future work is to develop efficient algorithms that allow us to find the composite service skyline from a significantly reduced |
| 9 | Kang Jinbeom et al. (2008) | Block Classification of a Web Page by Using a Combination of Multiple Classifiers | IEEE | multiple classifiers are built, one for each training domain, and the block classification | Need to reduce time complexity. |
| 10 | Kovacevic Milos et al. (2002) | Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification | IEEE | Using visual information one is able to define heuristics for the recognition of common page areas such as header and right menu, footer and center of a | This system will also improve focus crawling strategies by estimating importance of the link based on its position and neighborhood |
| 11 | Mun Yilhyeong et al. (2008) | Classification of web link information and implementation of dynamic web page using Link Map System | IEEE | Dynamic web<br>page using Link Map System | This problem can affect to rod of early Web page of user. To decrease this effect, we change increasing links through link |
| 12 | Tomar G.S. et al. (2006) | Web Page Classification using Modified Naïve Bayesian Approach | IEEE | The classification accuracy analysis with increasing vocabulary size | Accurate results demand higher training dataset. |
| 13 | ZOU Jia-qi et al. (2005) | Chinese Web Page Classification Using No se-tolerant up port vector Machines | IEEE | Not Available | Not Available |
| 14 | He Kejing et al. (2016) | Structure-Based Classification Of Web Documents Using Support Vector Machine | IEEE | It uses additional information provided by the HTML structure. | It is very time complex. |

## III. RESEARCH GAP

We found from the papers that the classification is done on the dataset of web structure, is optimized by Structure Based web document analysis. There are various ways to perform the web structure based classification [8] even beside of this describe technique. Undoubtly, web structure based classification give better in association of the feature selection results because it finds various features of the dataset's records. But while using this technique with simple web structure based classification, there are two concerns or says areas where there is scope of improvement lies. These concerns are as follows:

Web structure based Classifications itself takes higher time to compute.

Result of simple web structure based classification is not that optimized.

And these concerns are because of the accuracy of the classification method itself. In this study it is proposed to improvise these factors to get better efficiency.

## IV. CONCLUSION

A large number of web classification algorithms have been proposed till date. This paper has summarized and compared some of these algorithms. Although each algorithm as its merits and demerits, we tried to find the best algorithm using various evaluation measures. In this paper we focus the research area of Web mining, focusing on the category of Web structure mining. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

## V. ACKNOWLEDGMENTS

I would like to thank to all the people those who have help me to give the knowledge about these research papers and I thankful to my guides Dr. Prateek Srivastava & Dr. Prasun Chakrabarti , Professor, Department of Computer Science & Engg. ,Sir Padampat Singhania University with whose guidance I would have completed my research paper and make it to published, finally I like to thank to all the website and IEEE paper which I have gone through and have refer to create my review paper successful.

In this paper, comparative studies of different controllers are studied and performance is evaluated according to time domain functions. It is observed that all controllers able to maintain the set point at the desired value but ZN-PID ,Fuzzy based controllers has slight overshoot, Model Reference Adaptive controller has no overshoot and settles quickly. So it conclude that Model Reference Adaptive Controller is the best controller then other controllers

## REFERENCES

[1] Rathod Dushyant.2012.A Review On Web Mining. International Journal of Engineering Research and Technology (IJERT).

[2] Jain Ashish, Sharma Rajeev , Dixit Gireesh and Tomar Varsha.2013. Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages. International Conference on Communication Systems and Network Technologies, (pp. 640-645), IEEE .

[3] Keller A Matthias and Hartenstein Hannes,GRABEX: A Graph-Based Method for Web Site Block Classification and its Application on Mining Breadcrumb Trails.WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT) , (pp. 290-297), IEEE .

[4] Quan QIAN ,Jianyu LI , Jing CAI , Rui ZHANG and Mingjun XIN .2013. An Anomaly Intrusion Detection Method Based on PageRank Algorithm. International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, (pp. 2226-2230), IEEE .

[5] Ye Feiyue, Zhang Feng ,Luo Xiangfeng andXu Lingyu.2013. Research On Measuring Semantic Correlation Based On The Wikipedia Hyperlink Network. (pp. 309-314),IEEE .

[6] Jose Jeeva and SojanLal P.2013. A Rough Set Approach to Identify Content and Navigational Pages at a Website.,IEEE.

[7] Sarac Esra, Ozel Selma Ayse .2013.Web Page Classification Using Firefly Optimization. Innovations inIntelligent Systems and Applications (INISTA), IEEE International Symposium.

[8] He Kejing and Cli henyang .2016.Structure-Based Classification Of Web DocumentsUsing Support Vector Machine. (pp. 215-219 ), Proceedings of CCIS2016, IEEE .

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

[10] Aliakbary Sadegh, Hassan Abolhassani,Rahmani Hossein and Nobakh Behroo.2009.Web Page Classification Using Social Tags. International Conference on Computational Science and Engineering ,(pp. 588-593), IEEE.

[11] Gowri, R. And Lavanya, R.2013.A novel classification of web service composition and optimization approach using skyline algorithm integrated with agents. Computational Intelligence and Computing Research (ICCIC),(pp. 26-28), IEEE .

[12] Kang Jinbeom and Choi Joongmin.2008.Block Classification of a Web Page by Using a Combination of Multiple Classifiers. Networked Computing and

Advanced Information Management Volume 2,(pp. 290-295) ,IEEE .

[13] Kovacevic Milos , Diligenti Michelangelo , Gori Marco and Milutinovic Veljko.2002.Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification .(pp. 250-257 ), IEEE .

[14] Mun Yilhyeong,Lee Minkyung and Cho Dongsub.2008.Classification of web link information and implementation of dynamic web page using Link Map System . Granular Computing, (pp. 26-28) IEEE .

[15] Tomar G.S. , Verma Shekhar and Jha Ashis.2006.Web Page Classification using Modified Naïve Bayesian Approach .IEEE TENCON 2006 Hong Kong, (pp. 14-17).

[16] ZOU Jia-qi , CHEN Guo-long and GUO Wen-zhong.2005.Chinese Web Page Classification Using No se-tolerant up port vector Machines.Natural Language Processing and Knowledge Engineering, (pp.785-790 ),IEEE NLP-KE.

[17] Levering Ryan , Cutler Michal , and Yu Lei .2008.Using Visual Features for Fine-Grained Genre Classification of Web Pages . Proceedings of the 41st Annual Hawaii International Conference on System Sciences,(pp. 1-10 ),IEEE