ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 5, May 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

An Efficient Approach for Improving the Performance of Information Retrieval in Web by Reranking the Search Engine Results

R. R. Keole[1]
Research Scholar
Dept. of Information Technology
H. V. P. M. College of Engg.& Tech
Amravati – India

Dr. P. P. Karde[2]
Supervisor & Head of Department
Dept. of Information Technology
Govt. Women's Residential Polytechnic
Yavatmal – India

Dr. V. M. Thakare[3]
Professor & Head of Department
P.G.Dept. of Computer science
Sant Gadge Baba Amravati University
Amravati – India

Abstract: ***The web is huge, diverse, dynamic, widely distributed global information service center. With the rapid growth of the web, users get easily lost in the rich hyperlink structure. User rely on the web for information, but the currently available search engines often gives a long list of results, much of which are not always relevant to the user's requirement. Providing relevant information to the users to cater to their needs is the primary goal. Therefore, finding the content of the web and retrieving the users' interests and needs from their behavior have become increasingly important. The search engine uses these ranking methods to sort the results to be displayed to the user. In that way user can find the most significant and useful result first. Information Retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Therefore in this paper we are proposing an Approach to combine web content, web structure & web usage mining for Enhancing Web Search Engine Results Delivery.For Web Content mining the textual content of web pages is captured through extraction of frequent words using a term-based weighted technique will be combined with hyperlinks using Weighted Page Rank algorithm of Web structure mining which takes into account the importance of both the in-links and the out-links of the pages & Web server log files to discover useful information of user. Finally, the Search result is optimized by re-ranking the result pages. This proposed system proves to be efficient as the pages desired by the user will be on the top priority in the result list and also optimize the query performance in terms of query results. The proposed work will focus on the problem of improving the performance of information retrieval in web search engine results.***

Keywords: web content mining, web structure mining, web usage mining, and web search engine.

I. INTRODUCTION

With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. A considerably large portion of information present on the World Wide Web (www) today is in the form of unstructured or semi-structured text data bases. The www instantaneously delivers huge number of these documents in response to a user query. However, due to lack of structure, the users are at a loss to manage the information contained in these documents efficiently. The www continues to grow at an amazing rate as an information gateway and as a medium for conducting business. Web mining is the extraction of

interesting and useful knowledge and implicit information from artifacts or activity related to the www [1, 2].The World Wide Web which contains several billions of information is still growing at a very faster rate as most of the people use the internet for retrieving interesting document. Searching is considered a very important aspect of the Internet [1, 2]. In the age of Google, Yahoo!, Bing and others where each one of them is trying to outdo the other in terms of performance for their search engines, it is apparent that search and its related technologies is an important research area. There are tens and hundreds of search engines available but some are popular like Google, Yahoo, Bing etc., because of their crawling and ranking methodologies. The search engines download, index and store hundreds of millions of web pages. They answer tens of millions of queries every day. So Web mining and ranking mechanism becomes very important for effective information retrieval [26].Keeping information organized is an important issue to make information retrieval easier. Although the information we need is sometimes available on the Web, this information is only useful if we have the ability to find it. Current search engines return lists of ranked url's with their title and their snippet, but still fail to find relevant contents and to present them in an organized way, therefore the user is required to go through the extensive list of the retrieved results to satisfy its needs [11]. Most search engines perform in a query-triggered way that is mainly on a basis of one keyword or several keywords entered. Sometimes the results returned by the search engine don't exactly match what a user really needs due to the fact of the existence of the homology.

Therefore, the main problem is regarding the retrieval of relevant web pages. One of the solutions to this problem is the combine approach of Web Usage Mining, Web Content Mining and Web Structure Mining for Enhancing Search-Result Delivery. Therefore the main aim of the proposed research work is to design an efficient approach to improve the performance of information retrieval in web search engine results that is able to reorder the web documents effectively [11].

II. Related Work

Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges. Oren Etzioni [11] was the person who coined the term Web Mining first time. Web mining is also a cross point of database, information retrieval and artificial intelligence.

Characteristics of web and various issues on web content mining are presented in [1].

R. Kosala & H. Blockeel presented a survey related to the research in the area of Web mining, they focused on the term Web mining and suggested three Web mining categories.

In [6] Z. Lu & H. Zha States those different users may have different search goals when they submit a query to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. Here a novel approach to infer user search goals by analyzing search engine query logs was proposed.

Ida Mele [7] proposed that Data, stored in server logs, represents a valuable source of information. The research focuses on two important issues: improving search-engine performance through static caching of search results, and helping users to find interesting web pages by recommending news articles and blog posts.

A novel approach using weighted technique [9] is introduced to mine the web contents catering to the user needs. Experimental results prove that the performance of the proposed approach in terms of precision, recall and F-measure is high when compared to other search engine results. Algorithm used is Relevancy and Weight based approach.

A new method is presented in [14] to identify navigation related Web usability problems based on comparing actual and anticipated usage patterns. The actual usage patterns can be extracted from Web server logs and then applying a usage mining algorithm to discover patterns among actual usage paths.

W. Xing and A.Ghorbani [17] proposed a Weighted Page Rank (WPR) algorithm which is an extension of the Page Rank algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page

evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links.

In [18] S. Nimgaonkar and S. Duppala presented detailed study about web Mining and web content mining also present a comprehensive survey of some of the techniques of web content mining used in real time for the extraction of structured and semi-structured data.

T. Joachims [23] proposed an approach to automatically optimizing the retrieval quality of search engine using click-through data stored in query logs and the log of links the users clicked on in presented ranking.

R.Bhushan and R.Nath [24] presented a web recommendation approach based on learning from web logs and recommends user a list of pages which are relevant to him by comparing with user's historic pattern. Search result list is optimized by re-ranking the result pages.

A framework presented in [38] uses hyperlinks for topic detection by means of clustering techniques, considering the corpus as a directed graph and document clusters as topics.

A method proposed in [39] models both document contents and link structure in a unified manner, though they employed different representations for each one. Textual contents were represented by means of TF-IDF, the link structure was employed by using a directed graph where nodes are documents and arcs are probabilistic links between them.

An Approach to hybridize web content and web structure mining to improve the performance of web mining was proposed by S. Byreddy and R. Aluvalu in [40].

[41] Proposed an approach to integrate Web content mining into Web usage mining. The textual content of web pages is captured through extraction of frequent word sequences, which are combined with Web server log files to discover useful information and association rules about users' behaviors.

III. PROPOSED METHODOLOGY

Most of the research work is focused only on web usage mining, web content mining or web Structure mining for Enhancing Search-Result Delivery, for improving the performance of Information Retrieval in web search engine results. Combine approach of Web Usage Mining, Web Content Mining and Web Structure Mining for Enhancing Search-Result Delivery is not considered. In this paper emphasis will be given on information retrieval based on the combine approach of Web content– free text, Web structure–hyperlinks, and Web usage-web log data [14], to increase relevancy in retrieval of information from web .For Web Content mining a term-based weighted technique will be used to mine the web contents. In web structure mining, Weighted Page Rank algorithm (WPR) takes into account for the importance of both the in-links and the out-links of the pages which distributes rank scores based on the popularity of the pages. Web Logs will record the user activities on search results to infer user search goals by analyzing search engine query logs. This technique provides a framework to discover different user search goals for a query by proposed feedback sessions constructed from user click-through logs. The Search result is optimized by re-ranking the search result pages. This proposed system proves to be efficient as the pages desired by the user will be on the top priority in the search result list.

A. Methodology Used In Web Content Mining

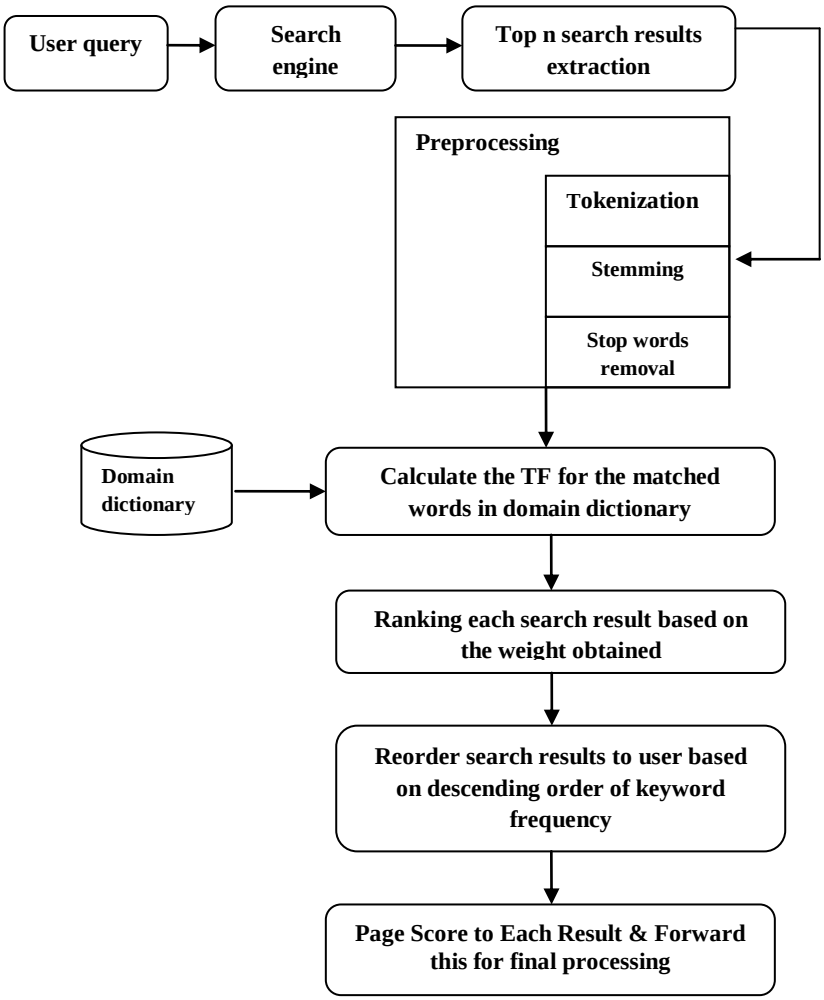
The first module is where the user gives the input query to obtain the search results. Based on that query top n search results are retrieved or extracted from the search engine. Most of the documents retrieved from the search engine may or may not be relevant to the user query.

The second module is pre-processing. The various steps involved in pre-processing are stemming, stop word removal and tokenization. Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form –

generally a written word form. Stop words are common words that carry less important meaning than keywords. Tokenization is the process of breaking a stream of text up into words, phrases, symbols.

The third module is the term frequency calculation. Statistical parameters such as a term frequency (TF) are calculated. For this every result is individually analyzed based on keywords and content. The calculations depend on the user query. After pre-process, Dictionary is built for user query with related words. The words present in the document are compared with the words present in the domain dictionary. So the words that are matched with the dictionary are taken for the term frequency calculation based on weighted technique. The importance increases proportionally to the number of times a word appears in the document.

The fourth module is the relevance calculation using the statistical methods; the normalized value of each result is sorted in descending order to get the most relevant content for the user query. Re-ordered results are sent back to the user so that the top most pages are more relevant for the user query. Finally results are forwarded for further final processing.



Algorithm for Mining Web Content

Algorithm: Relevancy and Term Based Weighted Approach

Input: User Query & Top N Search Engine Results

Output: Reordered Search Engine Results

Step 1: Extract search Engine results SR_i for the user query

Where $1 < i < N$

Step 2: Preprocess user query and Extract root words RW_j

Where $1 < j < N$

Step 3: Construct Dictionary D for the user query RW_j named as

UK_i (User Query Keywords)

Step 4: Preprocess and Extract the Keywords from Search Engine Results SR_i named as KSR_i (Keywords present in search Engine results SR_i),

Step 5: Compute Keyword Strength of KSR_i i e. Calculate total number of Keywords in SR_i

$T(KSR_i) = \sum KSR_i$

Step 6: Compare each Keyword KSR_i against Dictionary D. ie UK_i (User Query Keywords)

$[UK_{i} == KSR_i]$

If $UKi_{==} KSR_i$ match is found then award strength to KSR_i denoted by Term frequency TF_i

If $(UK_{i} == KSR_i)$

$TF_i = TF + 1;$

Else

$TF_i = 0$ (ie. award 0 strength to KSR_i)

Step 7: For $TF_i \neq 0$

Calculate Percentage % of Term Frequency TF occurred in SR_i

$PTF_i = TF_i \times 100 / T(KSR_i)$

Step 8:Compute Total Relevancy for each SR_i by multiplying each PTF_i (ie multiplying each Percentage of Term Frequency for ex. If input query is java programming language and if percentage of occurrence is java =10%, programming = 20% & language = 0% then Final Percentage $FPTF_i$ of Term Frequency is 0)

If $FPTF_i > 0$

SR_i = Relevant

Else

SR_i = IR-Relevant

Step 9: Sort Relevant SR_i List L1 in Descending order of PTF_i

Step 10: Sort IR-Relevant SR_i List L2 in Descending order of PTF_i

Step 11: Concatenate L1 & L2 for Final output of Reordered search results

B. Methodology Used In Web Structure Mining

In web structure mining, Weighted Page Rank algorithm (WPR) takes into account for the importance of both the in-links and the out-links of the pages which distributes rank scores based on the popularity of the pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as $W^{in}_{(m, n)}$ and $W^{out}_{(m, n)}$ respectively.

Algorithm: Weighted Page Rank

Input: User Query q, Set of pages p, Damping factor d =0.85, Inlinks & Outlinks of pages p.

Output: Rank score of WPRR.

1. Initialize all pages with initial rank is 1

2. for (i ϵ p, 1 ≤ i ≤ p) //Set of pages p

{

3. Calculate $I_{p \& Op}$ // Where I_p denote the number of incoming links & O_p the number of outgoing links

$W^{in}_{(m, n)} = I_n / \Sigma I_p$ // $W^{in}_{(m, n)}$ denotes to the incoming links

$P \epsilon Re_{(m)}$

$W^{out}_{(m, n)} = O_n / \Sigma O_p$ // $W^{out}_{(m, n)}$ denotes to the outgoing links

$P \epsilon Re_{(m)}$

4. $WPR(n) = (1-d) + d \Sigma WPR(m) W^{in}_{(m, n)} W^{out}_{(m, n)}$

5. Allocate Page Rank to each Page

WPR algorithms provide pages in the sorting order according to their ranks to users for a given query. So the order of relevant pages and their numbering are very important for users in the resultant list.

C. Methodology Used In Web Usage Mining

Web Logs are important information repositories, which record user activities on the search results due to these following reasons. One can keep track of previously accessed pages of a user. These pages can be used to identify the typical behavior of the user. Frequent access behavior for the users can be used to identify needed links to improve the overall performance of future accesses. Common access behaviors of the users can be used to improve the actual design of web pages. Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

Our approach is to organize search results by aspects learned from search engine logs. Given an input query, the general procedure of our approach is:

1. When user clicks the url out of the search results list, the information about user query and the clicked url are stored in the server log. Get this related information from search engine logs. All the information forms a working set.

2. Learn aspects from the information in the working set. These aspects correspond to users' interests given the input query. Each aspect is labeled with a representative query.

3. Categorize and organize the search results of the input query according to the aspects learned above.

4. When next time user enters the query, the results are retrieved from search engines and compared with the data saved in the server log and rank the search results accordingly so that users can reach effortlessly what they are looking for.

IV. Reconstruction of Search Results and Expected Outcome

For Web Content Mining the process is implemented using the Algorithm Relevancy and Term Based Weighted Approach described in the proposed methodology. Here input is User Query & Top n Search Engine Results and output is Reordered Search Engine Results with their relevancy score.

For Web Structure Mining, WPR value of each search engine result is calculated using the WPR Algorithm. The web documents are sorted in the descending order of their Weighted Page Rank value calculated using WPR Algorithm. Therefore, the highest WPR value calculated is at the top and least at the bottom.

For Web Usage Mining, when a user enters the search query and gets search engine results it first check for the server log entries. If the log contains some entries for previously accessed data then the page score is given on the basis of data present in the server log. The log contains the URL of the selected result and number of times he made the click on that particular URL. The log is updated automatically whenever the user selects the web page.

All the above ranking Score by Using Content Mining, Using Structure Mining and Using Usage Mining are considered for the final Score and Ranking Using Average Page Score of Content, Structure & Usage Mining will be considered as the final ranking.

V. CONCLUSION

The proposed system improves the performance of information retrieval in web search engine results. For Web Content mining a term-based weighted technique is used to mine the web contents with the help of the algorithm relevancy and term based weighted approach, where input is user query & top n search engine results and the output is reordered search engine results. In web structure mining, Weighted Page Rank algorithm (WPR) takes into account for the importance of both the in-links and the out-links of the pages which distributes rank scores based on the popularity of the pages and provides web pages in the sorting order according to their ranks to users for a given query. So the order of relevant pages and their numbering are very important for users in the resultant list. For Web Usage Mining web logs are considered. The log contains the url of the selected result and number of times the user made the click on that particular url. Using the click data the page score is decided. All the above ranking Score by Using Content Mining, Using Structure Mining and Using Usage Mining are considered for the final Score. Finally the Search result is optimized by re-ranking the result pages as per the final page score.

ACKNOWLEDGEMENT

Ranjit Keole is thankful to Dr.P.P.Karde, Head of Department, Dept. of Information Technology, Govt. Women's Residential Polytechnic, Yavatmal, for his constant support and helping out with the preparation of this paper. He is also thankful to the Dr. V.M.Thakare, Professor & Head of Department, P.G.Department of Computer science, Sant Gadge Baba Amravati University, Amravati for being a constant source of inspiration.

References

1. Bing Liu, Kevin Chen- Chuan Chang, and Editorial: Special issue on Web Content Mining, SIGKDD Explorations, Volume 6, and Issue 2.
2. G.Poonkuzhali, K.Thiagarajan, K.Sarukesi and G.V. Uma, SignedApproach for Mining Web content Outliers, Proceedings of World Academy of Science , Engineering and Technology, Vol.56,2009,PP 820-824.
3. Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey,ACM SIGKDD, July 2000.
4. R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.
5. H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.
6. Z. Lu, H. Zha, X. Yang, W. Lin, Z. Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions," Proc. IEEE Transactions on Knowledge and Data Engineering, pp. 502-513, 2013.
7. I. Mele, " Web Usage Mining for Enhancing Search –Result Delivery and Helping Users to Find Interesting Web Content," ACM SIGIR Conf. Research and Development in Information Retrival (SIGIR '13), pp. 765-769, 2013.
8. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl., 1(2):12, 2000.
9. P. Sudhakar, G. Poonkuzhali, R. Kishor Kumar, "Content Based Ranking for Search Engines," Proc. International Multi Conference of Engineers and Computer Scientists (IMECS 12), 2012.

10. X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.
11. O. Zamir and O. Etzioni, " Web Document Clustering: A Feasibility demonstration," ACM (SIGIR, 99) , pp. 46-54.,1998.
12. Guandong Xu," Web Mining Techniques for Recommendation and Personalization", Ph.D. dissertation, Victoria University, Australia, March 2008.
13. Ramakrishna, M.T. Gowdar, L.K. Havanur, M.S. Swamy (2010), "Web Mining: Key Accomplishments, Applications and Future Directions", International Conference on Data Storage and Data Engineering (DSDE), pp.187 – 191,2010.
14. Ruili Geng, Member, IEEE, and Jeff Tian, Member, IEEE," Improving Web Navigation Usability by Comparing Actual and Anticipated Usage ",IEEE Transactions On Human-Machine Systems, Vol. 45, No. 1, February 2015.
15. WangBin and LiuZhijing , "Web Mining Research" , in Proceeding of the 5th International Conference on Computational Intelligence and Multimedia Applications(ICCIMA'03) 2003.
16. Brin, S. and L. Page, The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst., 1998. **30**(1-7): p. 107-117.
17. Wenpu Xing and Ali Ghorbani," Weighted PageRank Algorithm", IEEE, 2004.
18. Satyajeet Nimgaonkar and Suryaprakash Duppala," A Survey on Web Content Mining and extraction of Structured and Semi structured data", ACM, 2012.
19. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl., 1(2):12{23, 2000.
20. X. Wang and C.-X Zhai, Learn from Web Search Logs to Organize Search Results,Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.
21. Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma," Learning to Cluster Web Search Results",ACM, 2004.
22. Hao Chen and Susan Dumais," Bringing Order to the Web: Automatically Categorizing Search Results", ACM, 2012.
23. T.Joachims"Optimizing Search Engines Using Clickthrough Data,"Proc.Eighth ACM SIGKDD Int. Conf. Knowledge Discovery &Data Mining (SIGKDD '02), pp. 133-142, 2002.
24. Ravi Bhushan and Rajender Nath, "Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques", Published by the IEEE Computer Society, IEEE 2012.
25. C. D. Manning, P. Raghavan, and H. Schtze "Introduction to Information Re-trieval",Cambridge University Press, 2011.
26. Duhan, N., A.K. Sharma, and K.K. Bhatia. Page Ranking Algorithms: A Survey. In Advance Computing Conference, 2009. IACC 2009. IEEE International. 2009.
27. Hsinchun Chen and Michael Chau, "Web Mining: Machine learning for Web Applications", Annual Review of Information Science and Technology 2003.
28. Salton, G., Wong, A., Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11):613-620, 1975.
29. Nicholas O. Andrews and Edward A. Fox, "Recent Development in Document Clustering Techniques", Dept of Computer Science, Virgina Tech 2007.
30. Xu, J. and Li, H. " AdaRank: A Boosting Algorithm for Information Retrieval, Proceedings of the 30th Annual International ACM SIGIR Conference,Amster-dam, Netherlands, 2011.
31. Ron Giles, How Search Engines Work, Available:http://www.website-consultant.co.nz/Website/Top+10+Search+Engine+Ranking+Factors/How+Search+Engines+work.html.
32. H. Cunningham, N. Fuhr, and B. Stein "Challenges in Document Mining, Dagstuhl Seminar Proceedings, vol.1, no.4, pp.65-99, Germany, 2011.
33. Thomas Mandl Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance, Proceedings of the 22nd ICML, Finland, 2007.
34. V. Fresno and A. Ribeiro. An Analytical Approach to Concept Extraction in HTML Environments. Journal of Intelligent Information Systems - JIIS. Kluwer Academic Publishers, 215-235, 2004.
35. Johnson, F., Gupta, S.K., Web Content Minings Techniques: A Survey, International Journal of Computer Application. Volume 47 – No.11, p44, June (2012).
36. Laxmi Choudhary and B. Shankar Burdak,"Role of Ranking Algorithms for Information Retrieval", International Journal of Artificial Intelligence and Applications (IJAIA), Vol.3, No.4, July 2012.
37. T.Munibalaji, C.Balamurugan, ―Analysis of Link Algorithms for Web Mining, International Journal of Engineering and Innovative Technology (IJEIT), ISSN: 2277-3754, Volume 1, Issue 2, February 2012, pp-81-86.
38. Garza Villarreal, S. E., Martínez Elizalde, L., and Canseco Viveros, A. Clustering hyperlinks for topic extraction: An exploratory analysis. In Proceedings of the 2009 Eighth Mexican International Conference on Artificial Intelligence, MICAI '09, pages 128–133, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3933-1, 2009.
39. Fersini, E., Messina, E., and Archetti, F. A probabilistic relational approach for web document clustering. Information Processing & Management, 46(2):117 – 130, 2010. ISSN 0306-4573., 2010.
40. S. Byreddy, RajaniKanth Aluvalu," Hybridization of Web Content and Structure Mining (HWCSM) Technique by means of Content Based Ranking Algorithm" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. VIII (Mar-Apr. 2014), PP 134-137.

41. S. Taherizadeh and N. Moghadam," Integrating Web Content Mining into Web Usage Mining for Finding Patterns and Predicting Users' Behaviors" International Journal Of Information Science And Management Year 2012.

AUTHOR(S) PROFILE

Prof. R. R. Keole, has received the M.E degree in Computer science & Engineering from Sant Gadge Baba Amravati University, Amravati, Maharashtra, in the year 2011. He is now pursuing his Ph.D from the same university under the guidance of Dr. P. P. Karde, Head of Department, Dept. of Information Technology, Govt. Women's Residential Polytechnic, Yavatmal.

