Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LINK BASED CLUSTERING OF WEBPAGES IN INFORMATION RETRIEVAL

[1]SARASWATHY.R, [2]SEETHALAKSHMI.R
[1]M.Tech, School of Computing, SASTRA University, Thanjavur, India.
[2] Professor,School of Computing,SASTRA UNIVERSITY,Thanjavur,India.
E-mail: [1]srnrj7@gmail.com, [2]rseetha123@cse.sastra.edu

ABSTRACT

Recently, web search engine plays a vital role in information retrieval. This is the most common tool used for information retrieval on web.However current status of web search engine is far from satisfaction. The results obtained does not contain needed information.To enhance the performance of web search engine, in this paper,Link based search engine for information retrieval using hierarchal clustering of web pages is introduced. In traditional search engine the results obtained is based on ranking mechanism.In link analysis, the relevance of web pages is obtained based on the occurrence of each inlink and outlink of particular web page. Then the web pages are grouped by filtering out irrelevant pages to increase the relevance rate and to reduce the computational time. The experiments of results show that the link analysis along with cluster of web search results is promising.

Keywords: *Web Pages, Link Analysis, Hierarchal Cluster, Information Retrieval, Web Search Engine.*

1. INTRODUCTION

Nowadays, web search engines are used as effective tools for information retrieval. However users are not satisfied with the search engine results. Because the results are not accurate and does not contain the needed information.Thusit takes time to filter out the irrelevant pages. Therefore one has to improve the performance of web search engine.The performance of web search engine can be improved by using link analysis techniques with clustering approach. This link analysis of web pages helps to filter out irrelevant pages and make the search engine to increase its performance.

The proposed approach combines link analysis with hierarchical cluster to make the web search engine to fetch the relevant result based on user query.The link analysis is performed by considering both in link and outlink of the particular webpage. Then the page which has more number of outlink to the relevant pages is considered as the central page. By using this page all other relevant pages are clustered, based on the similarity measure. The grouping of similar page is carried out by using extended hierarchal cluster method.

Many traditional clustering techniques usethe term frequency measure as the parameter for clustering.The main advantage of the proposed link based clustering approach over traditional method is hyperlink between web pages,which provides useful information to group the relevant pages. And we can apply these proposed techniques to any languages whereas traditional techniques cannot beappliedto other languages rather than English. Thus the proposed techniques overcome some of the disadvantage of traditional method to meet the requirements of web search users.

When the web pages are clustered using hierarchical cluster algorithm itincreases the relevance rate of search results by filtering out irrelevant pages. After the clustering process when user submit query to the search engine they get most relevant pages based on user query without irrelevant pages. So, the proposed approach increases the relevancy rate and reduces the computational time.

2. RELATED WORK

Many works [1][2][3][15][18] were aiming to improve efficiency of the web search engine by

Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

assuming link analysis techniques or term frequency method.

Early search engine mainly compares the content similarity of the query and the index pages. A page owner can repeat some words and add many related words to boost the ranking of the page to make the page relevant to a large number of queries. The hyperlink based search algorithm such as page rank and HITS [1] were reported. Both these algorithms are related to social networks. They exploit the hyperlinks of the web to rank page according to their levels of 'prestige' and 'authority'. Page rank has emerged as dominant link analysis model. Page rank is global measure and query independent. The value of all the pages are computed and saved off-line rather than at the query time. And HITS is a search query independent. HITS first expand the list of relevant pages that are returned from search engine. Then produces two rankingsauthority ranking and hub ranking of the expanded set of pages. The following figure1 shows the authority and hub pages.

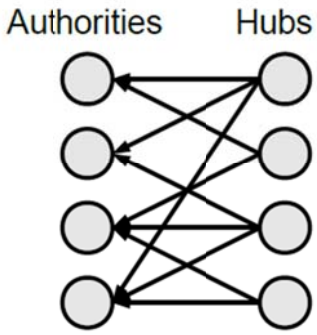

Figure1. *Potential hub pages and authority pages in search results*

The link analysis has been useful in many areas such as ranking, query results, crawling, computing of web pages reputation, finding related web pages, finding mirrored host and computing of web pages. And analyzing of hyperlink structure of web has been started and lead to significant improvement in web information retrieval. Hyperlink provides information of web information retrieval. This form of retrieving information is known as link analysis. Link analysis algorithm has been widely used in web information retrieval. However current link analysis algorithm will work only on flat link graph, ignoring the hierarchal structure of the web. Hierarchical structure produces better cluster, and detailed description and comparisons could be found in [5][10][13].

3. PROPOSED SYSTEM

In proposed link analysis of web page extended hierarchical cluster are used for efficient search results in web search engine. The cluster used here is more efficient to cluster the data points of web page which is arbitrarily distributed. Initially the similarity measure webpage is obtained based on inlink and outlink of the webpage. Then these pages are clustered based on similarity measure by using hierarchical cluster.

3.1Link analysis on web page

Link analysis is based on analyzing the inlink and outlink of each web page obtained by search results of user query. Initially all inlink and outlink of each web page is extracted by using Meta search engine based on user query.

3.2 Duplicate web page

The web contains many duplicate or mirrored web pages which misleads clustering process.The duplicates can be removed by considering common links shared by web pages. If two web pages are said to be duplicate if they have at least eight out links and most of them are common. The page which has majority of common links is said to be duplicate and its corresponding inlinks and out links will be removed. Here the similarity of web page is calculated bytraditional cosine measures by analyzing common linksbetweentwo compared web pages. Then the web pages are clustered by using the extended hierarchical cluster. The link analysis of web pages is shown in figure 1.

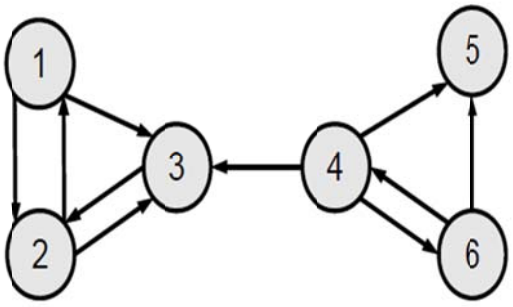

Figure 2. *In link and out link on web pages*

3.3 Hierarchal cluster method

Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Eventhough there are many clustering algorithms developed for web page, thecluster method used in this context is the extension of hierarchical cluster.InK-means clustering the cluster is based on the center point. It clusters N data points into one level.The disadvantage is that the quality and the structure of final clusterare based on initial centroid. Hierarchical cluster is contrast to k-means which creates nested sequence of partition.In the below figure 3, shows the hierarchical structure of web pages. The extension of hierarchical cluster method does not consider all the WebPages as the standard cluster. Here the high quality pages whose sum of inlink and outlink are at least two is taken for clustering. This helps to filter irrelevant pages from others.

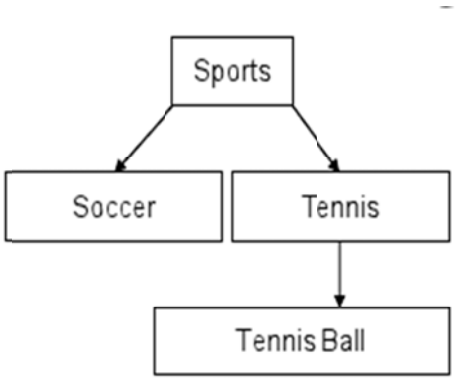

***Figure 3**. Hierarichical structure of web pages*

ALGORITHM

Input D ::={$d_1, d_2, d_3, \ldots d_n$}

1. Calculate similarity matrix SIM [i,j]
2. Repeat
3. Merge the most similar two clusters, k and L to form new cluster KL
4. Compute similarities between KL and each of the remaining cluster and update SIM [i,j]
5. Until there is a single (or specified number) cluster
6. Output: dendogram of clusters

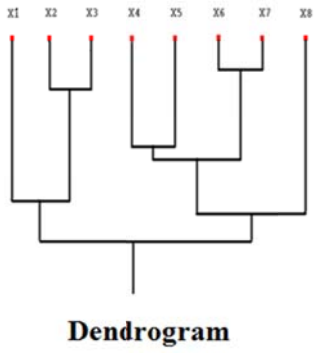

***Figure 4.** Dendrogram of cluster*

In above figure 4, shows the dendogram of web page cluster. Thus the cluster must include the pages by considering both similarity and common links shared by web pages. If the web page does not meet the requirements then it is taken as new cluster. Like this, the process is iteratively executed until it converges and form base cluster. Finally base cluster can be combined by using merging threshold. By analyzing the majority shared web pages, merging threshold is calculated.Thus the proposed technique helps to produce the web pages that are most relevant by eliminating irrelevant web pages according to user query. The web search engine has improved its performance and the computational time is reduced.

4. EXPRIMENTAL ANALYSIS

Thus proposed system is verified by giving various user query and the results obtained are most relevant web pages. Initially query is given to meta search engine, it retrieves results obtained from other search engine. Then the duplicate web page is obtained by using cosine measure.Finally the web pages are clustered by using hierarchical clustering result is shown in table1, which improves relevancy rate and reduces computational time. This project is implemented in Microsoft Visual Studio and the backend is Microsoft SQL Server.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

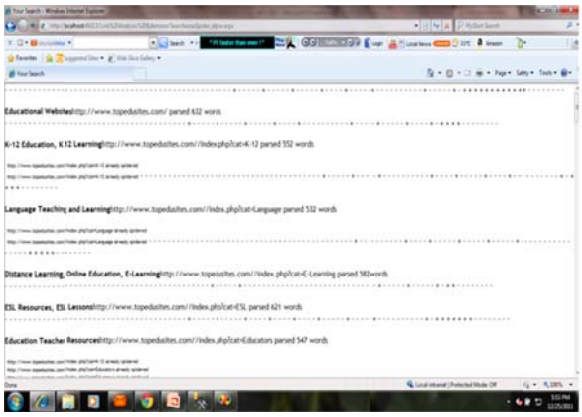

Figure5*. Links obtained based on user query*

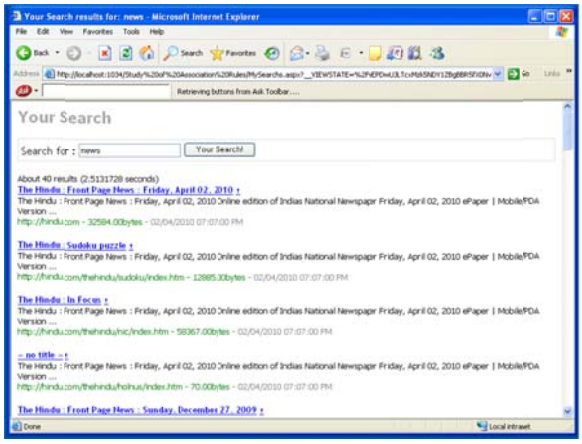

Figure 6. *Links without duplication*

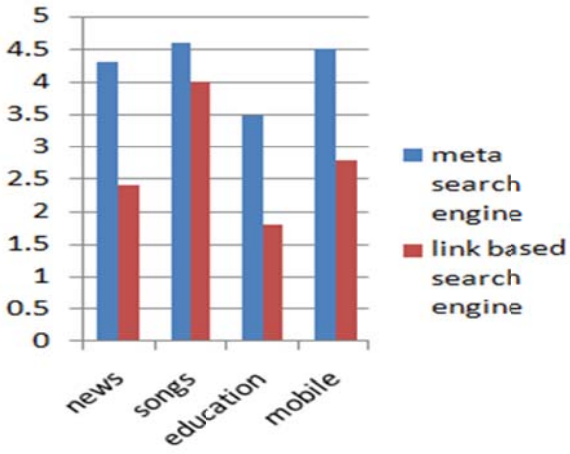

Figure 7.*Performance measure*

5. CONCLUSION

In this paper, an approach which introduces the link based clustering of web page techniques is dealt with. This improves the efficiency of the web search engine.Traditional method can be applied only to the Here the duplicate web pages are removed by analyzing the links between the web pages. Finally the web pages are clustered by using hierarchical cluster and means which creates nested sequence of partition. The proposed techniques efficiently improve the performance of web search engine by avoiding unwanted web pages and reduce the computational time.

Hierarchical clustering provides better result only for splitting small amount of data. However one fundamental limitation of this method is, dendogram gets increased and it leads to information losses when large amount of data needs to be clustered. Future work will be the implementation of efficient clustering algorithm thereby improving the efficiency of search engine for information retrieval.

ACKNOWLEDGMENT

The authors thank the authorities of their working organization and management for their support and the encouragement to pursue research in the chosen field of study.

REFERENCES:

[1] Kleinberg 98*Authoritative sources in a hyperlinked environment*. In proceedings of the 9thACM-SIAM Symposium on Discrete Algorithms (SODA), January 1998.

[2] Ravi Kumar *et. al*. 99*Trawling the Web for emerging cyber-communities*. In Proceedingsof 8th WWW conference, 1999, Toronto, Canada.

[3] Brin and Page 98*The anatomy of a large-scale hypertextual web search engine.* InProceedings of WWW7, Brisbane, Australia, April 1998.

[4] Oren Zamir and Oren Etzioni *99Grouper: A Dynamic Clustering Interface to Web SearchResults*. In Proceedings of 8th WWW Conference, Toronto Canada.

[5] Richard C. Dubes and Anil K.Jain**,** *Algorithms for Clustering Data***,** Prentice Hall, 1988

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[6] Oren Zamir and Oren Etzioni 97*Fast and Intuitive clustering of Web documents*,KDD'97, pp287-290

[7] Oren Zamir and Oren Etzioni 98*Web document clustering: A feasibility demonstration*. InProceedings of SIGIR' 98 Melbourne, Australia.

[8] Zhihua Jiang *et. al.Retriever: Improving Web Search Engine Results Using Clustering*http://citeseer.nj.nec.com/275012.html.

[9] Ron Weiss *et. al.* 96*Hypursuit: A Hierarchical Network Search Engine that ExploitsContent-Link Hypertext Clustering*. ACM Conference on Hypertext, Washington USA,1996

[10] Michael Steinbach *et. al.A Comparison of Document Clustering techniques*. KDD'2000.Technical report of University of Minnesota**.**

[11] James Pitkow and Peter Pirolli 97*Life, Death and lawfulness on the Electronic Frontier*.In proceedings of ACM SIGCHI Conference on Human Factors in computing, 1997

[12] Cutting, D.R. *et. al.92Scatter/gather: A Cluster-based approach to browsing largedocument collections*. In Proceedings of the 15th ACM SIGIR, pp 318-329; 1992

[13] A.V. Leouski and W.B. Croft. 96*An evaluation of techniques for clustering searchresults.* Technical Report IR-76 Department of Computer Science, University ofMassachusetts, Amherst, 1996

[14] Broder *et. al.* 97*Syntactic clustering of the Web*. In proceedings of the Sixth InternationalWorld Wide Web Conference, April 1997, pages 391-404.

[15] Bharat and Henzinger 98 *Improved algorithms for topic distillation in hyperlinkedenvironments.* In Proceedings of the 21st SIGIR conference, Melbourne, Australia, 1998.

[16] Chakrabarti *et. al.* 98*Automatic Resource Compilation by Analyzing Hyperlink Structureand Associated Text.* Proceedings of the 7th Worldwide Web conference, 1998.

[17] Florescu, Levy and Mendelzon 98*Database Techniques for the World-Wide Web: ASurvey*. SIGMOD Record 27(3): 59-74 (1998).

[18] Gibson, Kleinberg and Raghavan 98*Inferring Web communities from link topology*.Proc. 9th ACM Conference on Hypertext and Hypermedia, 1998.

[19] Agrawal and Srikant 94*Fast Algorithms for mining Association rules*, In Proceedings ofVLDB, Sept 1994, Santiago, Chile.

[20] M.M. Kessler**,** *Bibliographic coupling between scientific papers*, AmericanDocumentation, 14(1963), pp 10-25

[21] H. Small,*Co-citation in the scientific literature: A new measure of the relationshipbetween two documents*, J. American Soc. Info. Sci., 24(1973), pp 265-269

[22] yitong wang and masaru kitsuregawalink based clustering of web search results,institute of industrial science,the university of Tokyo(2000)

[23] M.Sathya,J.Jayanthi,N.Baskarlink based k-means clustering algorithm for information retrieval (2011)