# Similarity Matrix Approach in Web Clustering

**Dr.R.Surendiran**
*Assistant Professor*
*Department of Computer Science*
*Mass College of Arts & Science,Kumbakonam*

***Abstract:***

*In the modern world internet plays a main role in everyone life style which makes life easier and time cosuming process compare with eariler days. Internet comes in every activities like purcahse, booking , investment,entertainment,etc but how far the reliablity of search results in the search engines which provides for the user queries is quite questionable thing. In this research work propose a novel approach in web clustering enchancement thorugh the similarity metrix concept in adddition to the existing concept. The results of the research work proves that proposed work efficent then the existing methods.*

**Keywords:***Similarity metrix, Adajacency matrix, HITS, Clustering*

## I. INTRODUCTION:

In Trendy world web technology plays a huge role and useful role but internet growed and growing such as large data storage which leads the information overload meanwhile number of users to the internet growing day by day in rampant manner. The availability of large information set,Nature of web,Nature of search engine are complex to fetch the relative information of user query.Metasearch engines , Search engines and Web Directories have beendeveloped over a period of time in order to satisfy the users requirement quickly and easily.

In General user searching for information submits a query contain by a fewkeywords to a search engine such as Google. The search engine performs approximate and similairty matching betweenthe query terms and the keywords of each web page and presentsthe results to the user. These long lists results of URLs, which are very hardto search. Furthermore, users without domain expertise are not familiar with theappropriate terminology thus not submitting the right query terms, leading to the retrieval of more irrelevant pages.

This scenario require the need for the development of new methods to help userseffectively view, trace and search the available web documents, with theultimate goal of finding those best matching their requirement. One of the techniquesthat can play an important role towards the achievement of this objective isdocument clustering. The increasing importance of document clustering and thevariety of its applications has led to the development of a wide range ofalgorithms with different quality.

Existing document clustering systems retrieve the big list of documents based on the ranking and hitting which makes the user forced to gothrough the unwanted information and also loosing their valuable time. probably all the search engine follows this concept which leads user in the confusion mode.

This research work examine the document clustering is enough to group the related data's. Numerous numbers of algorithms are proposed to work with offline but only few algorithms are working in realtime search engine in effective manner. Clustering should be minimal set because there are millions of queries were exuected in the search engine. The accuracy level of the search engines designed with ranking which makes user to find their informations are too difficult. Search engines design should be consider the following charactersitics. Such as

- Relevancy.
- Snippet-tolerance.
- Speed.
- Incrementality.
- Browsable Summaries.
- Overlap.

In our research work we have proposed a novel clustering method to improve the occuracy level.

## II. PROPOSED METHOD :

The World Wide Web has a varaitey of documents and complex structure which contains both textual web documentsand the hyperlinks that connect them. Combination of web documents and hyperlinks can be drawn as directed graph in that web document denoted as vertices and hyperlinks denoted as edges. Algorithms were developed to utilize the directed grpah to extract the information of hyperlinks and documents from the graph. For example HITS algorithm purely on hyperlink information to retrievethe most relevant information: authority and hub documents for a user query.

Moreover, if the hypertext source contains of several topics, authorityand hub documents may cover the most popular topics and leave out theless popular ones. To overcome this issue we will partion the hypertext collection into sub groups and then present the search as list of topics. We need to cluster the web document in two ways one is textual based another one is hyperlink based. In this research we utilized the similarity based clustering method and normalized cut.

While clustering in addition to the textual contents, we have enchanced the clustering using the concept of hyperlink between documents and co-citation patterns between the documents for new similairty metrics for measuring the topical homogeneity of web documents.hyperlink played a major role in the similarity metric and the textual content used to strength the hyperlink.

The similarity metric will help to construct the weighted graph which will goes to the clustering method based on normalized cut.

### *a) Result set preparation :*

In our WWW era accurate information reterival is very difficult process. The main question is how to efectively organize such a large number of retrieved web documents for a user query according to their relevance ? . for example web document ranking algorithm HITS are unable to slove this issue. HITS will retrive most popular topic and many less popular documents were eliminated wich may be the requirement of user. To over come this problem before ranking which is necessary togroup the retrieved web documents into distinct topic areas, then return the rankeddocuments for each group according to their relevance.

If twoof web documents have very small text matches, it is less chance that they comes undersame group or topic, however they are connected by a hyperlink. Therefore, to improve the quality of the graph representation, the text information can be incorporated into the linkgraph as a factor of edge weight. The co-citation is another important factor betweentwo web documents based on how many other web documents create hyperlinks toboth of them.

### *b) Similarity metrics :*

Very next step we have to classify the data objects in to the subgroups which is based on similarity concept between the objects. It leads to the objects within the group are near identical.Edge weights are represented by similarity between nodes in the graph representation.

We introduce the concept of link structure, textual information to calculate the similarity metrics which yields the weight matrix **W.** Link structure play a vital role and textual information is included as enhance the links.

Link graph *G = ( V, E)* , Adjacency matrix will be defined as

$$Adjmat = \begin{cases} 1 \; if (i,j) \in E \; or \; (j,i) \in E, \\ 0 \; otherwise \end{cases}$$

Directionality of the hyperlinks isignored in the adjacency matrix of the link graph. Linkstructure provides us with rich information on the content.

Existing methodology to incorporating the textual informationis to evaluate the similarity between a user query and the anchor text present in the web document but in our approach

- Utilizes the entire text of a web document,not just the anchor text.
- Measures the textual similarity Hij between two webdocuments i, j.
- Uses Hijas the strength of the hyperlink between web documents i, j.

ThereforeHij properly calculate the importance of an individual hyperlink.

We represent each web document as a vector in the vector space model of Information Retrieval then compute the similarity between them. The high similarity, the more likely the two documents fall with the same topic.

For each element of the vector we use the standard tf.idf weighting: *tf(i; j) ∗ idf(i)*. *tf(i; j)* is the *Term Frequency* of word i in document j, representing the numberof occurrences of word i in document *j*. *idf* is the Inverse Document Frequencycorresponding to word i, defined as

$$Idf(i) = \log\left(\frac{number\ of\ total\ documents}{number\ of\ documents\ containing\ word\ i}\right)$$

Since the term vector lengths of the documents vary, we use cosine normalizationin computing similarity. That is, if x and y are vectors of two documents d1 and d2,then the similarity between d1 and d2 is:

$$s(d1, d2) = s(d2, d1) = \frac{\sum_i x_i \; y_i}{\| x \|_2 \| y \|_2}$$

Where $\| x \|_2 = \sqrt{\sum_i x_i^2}$

## III. EXPERIMENT:

We have teseted our algorithm for the keyword *sivaji* which gives three distinguish meaning. First one Chhatrapati Shivaji Maharaj, was an Indian warrior king and a member of the Bhonsle Maratha clan, second thing The famousthe great indian actor sivajiganesan, last one the famous indian film sivaji starrer by rajinikanth.

Now we are ready to apply our clustering method on the data source. Each and every cluster has a huge number of web documents, In our experiment we have choosen only the most important webdocuments among a large set. The most important documents has been selected ineach cluster are determined based on the HITS algorithm.

First we have to derive the link graphs with the help of text based search engine with query terms. Search engine returns a set of URLs which has top ranking for the relevant

query.In our research we have restircted upto 40 URLs because of complexity issue of implmentation for base set. So that we can maintain overall data set within the limit. After that we can expand the base set with the relevanted document. Link graph were drawn as direct graph which is easy to convert in other format of matrix.

Now complete list of URL set available, Next step we have to run the a web crawler to filter the text information of these web documents. In this research experiment has limited the length of document words 300. Remaining words are discarded if the document contain more than 300 words. In the above process stop words were eliminated using stemming algorithm.

## IV. CLUSTERING :

Now we can apply our algorithm into our data set. Then we should apply hits algorithm in the cluster which has obtained. Cluster has no special meaning which is just the sequence of query terms.

There are a total of 1123 URLs in this data set. Applying our algorithm, we get the following relevant clusters. The clusteres which hasmall size are not acounted.

**Maratiya sivaji: (Cluster 1)**

https://en.wikipedia.org/wiki/Shivaji
www.itstamil.com/chatrapati-shivaji.html
https://chhatrapatishivaji.wordpress.com/2008/10/29/shivaji-bhosle/
www.iloveindia.com › History › Medieval History of India
www.historydiscussion.net › Shivaji › Indian History › Kings › Life › Life of Shivaji

**Sivaji Ganesan : (Cluster 2)**

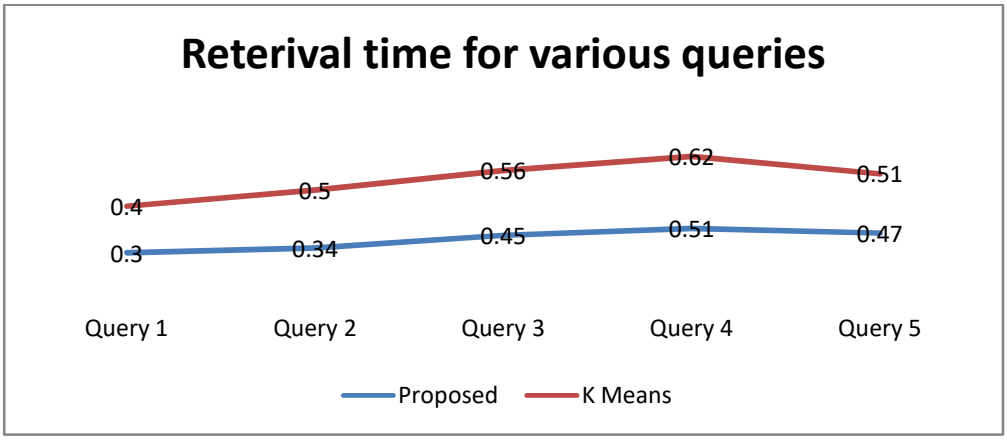
https://en.wikipedia.org/wiki/Sivaji_Ganesan
https://www.saddahaq.com/sivaji-ganesan-the-rule-book-of-acting-in-indian-cinema
https://www.britannica.com/biography/Sivaji-Ganesan
www.infoqueenbee.com/2013/10/biography-of-sivaji-ganesan-tamil.html
www.thehindu.com/features/metroplus/five...sivaji-ganesan/article7706802.ece

**Filim sivaji(Cluster 3)**

https://en.wikipedia.org/wiki/Sivaji_(film)
www.imdb.com/title/tt0479751
www.filmibeat.com/tamil/movies/sivaji-the-boss.html
www.rediff.com › Movies › Reviews

We can compare our results with existing well known method k-means algorithm . experiment showed that the similarity matrix based technique performs better than the K-means based algorithm. The following table of information can obtained from our experiment for the above query terms. Search engine we have used hotbot,webcrawler codeds that we write in Perl. Stemming algorithm used for stemming purpose.

| | Reterival time | Cluster accuracy | Relavancy of search terms | Valid Skipped link ratio | Small cluster omition ratio | Tightly connected % |
|---|---|---|---|---|---|---|
| K -Means | 0.5sec | 97.13% | 96.48% | 3.1% | 2.59% | 95.39% |
| Proposed method | 0.3sec | 99.2% | 98.87% | 1.62% | 1.37% | 99.45% |




Based on the performance table which clearly shows that proposed method better than k means algorithm in all the aspects. Diagrams shows that proposed mehtod increases the speed of reterival time for various search terms compare with exsiting one.

## V. CONCLUSION:

Modern world increases the internet usage and many people using search engines for various purposes for day today activities . Getting right information at right time leads to tremndous changes in business or life style. Our proposed method will be usefull to the society in the regular activities and also which help to the young resarchers for their academics to take it as base work.

## VI. REFERENCES:

[1]Introduction to Data Mining, P.N. Tan, M. Steinbach, V. Kumar, Addison Wesley

[2] R.Surendiran, Rajan.K.P, SathishKumar.M: Study on the Customer targeting using AssociationRule Mining, International Journal on Computer Science and Engineering, Vol. 02, No. 07, 2010,ISSN: 0975-3397, Page 2483 - 2485

[3] R.Surendiran, K.Alagarsamy: A Novel Tree Based Security Approach for Smart Phones, International Journal of Computer Trends and Technology, Volume 3 Issue 6 – 2012, ISSN: 2231 - 2803, Page 787 - 792

[4] Porter, M.F., 1980. An algorithm for su.x stripping. Program 14, 130–137.

[5]An efficient k-means clustering algorithm: Analysis and implementation, T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, IEEE Trans. PatternAnalysis and Machine Intelligence, 24 (2002), 881-892

[6] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web communities from link topology. In Proc. 9th ACM Conference on Hypertext and Hypermedia (HyperText 98), pages 225–234, Pittsburgh PA, June 1998.

[7] K. Bharat and M. Henzinger.Improved algorithms for topic distillation in a hyperlinked environment. In Research and Development in Information Retrieval, pages 104–111, 1998.

[8] Dr.R.Surendiran: Development of Multi Criteria Recommender System, SSRG International Journal of Economics and Management Studies (SSRG-IJEMS) – volume4 issue1 January 2017, ISSN: 2393 - 9125, Page 28 - 33

[9] D. Butler. Souped-up search engines. Nature, 405:112–115, May, 2000.

[10] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text.In Proc. 7th International World Wide Web Conference, Brisbane, Australia, 1998.

[11] R.Surendiran, K.Alagarsamy: A Crtitical Approach for Intruder Detection in Mobile Devices, SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – Volume1 Issue4 - June 2014, ISSN: 2348 – 8387, Page 6 - 14

[12] W. A. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical recipes in C: the art of scientific computing(2nd ed.). Cambridge University Press, New York

[13] S. Chakrabarti, B. Dom, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. IEEE Computer, 32(8):60–67, 1999.

[14] R.Surendiran, K.Alagarsamy: Privacy Conserved Access Control Enforcement in MCC Network with
Multilayer Encryption, International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue5 - May 2013, ISSN: 2231-5381, Page 2217 - 2224

[15] Small, H., 1973. Co-citation in the scientific literature: A new measure of the relationship between twodocuments. J. Amer. Soc. Inform. Sci. 24, 265–269.

[16] Dr.R.Surendiran: Secure Software Framework for Process Improvement, SSRG International Journal of
Computer Science and Engineerin g (SSRG-IJCSE) – volume 3 Issue 12 – December 2016, ISSN: 2348 – 8387,Page 19 - 25

[17] J. Dean and M. Henzinger. Finding related pages in the World Wide Web. In Proc. 8th International World Wide Web Conference, Toronto, Canada, 1999.

[18]http://www.cs.cmu.edu/~cga/ai-course/kmeans.pdf 4.

[19] J. Kleinberg. Authoritative sources in a hyperlinked environment, 1997. Research Report RJ 10076 (91892), IBM.

[20]Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S., 1999. The web as a graph:measurements, models, and methods. Proceedings of the Fifth Annual International Computing andcombinatorics Conference, 1999. pp. 26–28.

[21] J. Kleinberg. Hubs, Authorities, and Communities.ACM Computing Surveys, 31(4es, Article No.5), 1999.

[22]http://www.cse.msstate.edu/~url/teaching/CSE6633Fall08/lec16%20k-means.pdf

[23] R.Surendiran, K.Alagarsamy: An Extensive Survey on Mobile Security and Issues, International Journal of Computer & Organization Trends – Volume2 Issue1 -2012, ISSN: 2249 - 2593, Page 39 - 46

[24] R.Surendiran, K.Alagarsamy: Privacy Conserved Access Control Enforcement in MCC Network with Multilayer Encryption, International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue5 - May 2013, ISSN: 2231-5381, Page 2217 - 2224