

## An Advanced Partitioning Approach of Web Page Clustering utilizing Content & Link Structure

Ruma Dutta<sup>1,3</sup>, Indranil Ghosh<sup>1</sup>, Anirban Kundu<sup>1,3</sup>, Debajyoti Mukhopadhyay<sup>2,3</sup>

<sup>1</sup>Netaji Subhash Engineering College, Garia, Kolkata 700152, India

rumadutta2006@gmail.com, ig\_cse@yahoo.co.in, anik76in@gmail.com

<sup>2</sup>Calcutta Business School, Diamond Harbour Road, Bishnupur 743503, India

debajyoti.mukhopadhyay@gmail.com

<sup>3</sup>WIDiCoReL, Green Tower, Block C, Flat 9/1, Golf Green, Kolkata 700095, India

doi: 10.4156/jcit.vol4.issue3.9

### Abstract

*Clustering of non-homogenous documents has become an increasing challenge and opportunity with the huge proliferation of World Wide Web. It has become difficult to retrieve the desired information without proper clustering of Web-page with the increase in information on the WWW. Several new ideas have been proposed in recent years. Among them partitioning approach is still widely used clustering approach for its simplicity. This paper proposes a partitioning approach to cluster the Web-page based on information provided by the hyperlink structure of Web-pages and also by the content of the Web-pages. The proposed approach of Web-page clustering exhibits better result than K-medoid partitioning clustering approach as the centroids are chosen by HITS Algorithm. The partitioning approach like K-mediod, K-means require number of clusters apriori. It has been observed that the performance of these approaches depend on the initial selection centroids of the clusters. These two problems have been solved by the approach proposed in this paper. Experimental result supports our approach as better concept.*

### Keywords

*Clustering, Hits Algorithm, K-mediod*

### 1. Introduction

Web has recently become a powerful platform with popularizing and spreading of Internet applications. Indexing or searching millions of non-homogenous documents and retrieving the desired information has become an increasing challenge and opportunity with the rapid growth of WWW. The structure of the Web is composed of a huge pool of documents and links between them. Currently, Web documents present human readable contents. Many Information are

increasing in the Web in every day because the millions of new Web sites and Web-pages are continuously added to the Web. It is impossible to make proper meaningful indexing and retrieve desired information within minimum time span without proper clustering.

Web-page clustering is a technology which puts related Web pages into groups and is useful for categorizing, organizing, and refining the search results. The goal of this paper is to develop a technique which will guide user to retrieve desired information with proper clustering of Web-page in WWW. This paper emphasizes to present an efficient algorithm for clustering large set of Web- pages. Partitioning clustering approaches are facing two disadvantages. One is in all these algorithms number of clusters has to be mentioned in advance and another is the performance of the clustering algorithm depends on initial selection of the centroids. These two problems have been solved in this paper by considering hyperlink structure of the Web-pages. Experimental result shows that our approach is better in terms of clustering performance of the Web-pages.

Section 2 of this paper discusses the related work of Web-page clustering. Section 3 presents the proposed approach. Section 4 shows the experimental work and section 5 concludes the paper.

### 2. Related work

Clustering is a classical problem in data mining research. There are lots of algorithms developed that can be categorized into partitional algorithm (K-means etc.), hierarchical algorithm (HAC)[6], density-based algorithm (DBSCAN), grid-based algorithm and graph based algorithm etc [1]-[6], [7], [9]. There is a great deal of work done previously in clustering, including K-means [5], CLARANS [5] etc. In the Information Retrieval (IR) community, the Scatter/Gather algorithm [8] is aimed at re-organizing document

search results by examining document contents. It is similar to K-means in that it requires pre-set cluster number, which is a requirement that we do not assume in our paper.

Clustering is a challenging topic in Web data management too as depicted in [1]. Web data clustering is the process of grouping Web data into “clusters” such that similar objects are in the same class and dissimilar objects are in different classes. Its goal is to organize data circulated over the Web into groups / collections in order to facilitate data availability and accessing, and at the same time meet user preferences. Suffix-Tree is another closely related clustering method. Its input is also portions of the document contents and researchers in this area did not consider both the content and hyperlink structure of Web-pages. Therefore, the main benefits include: increasing Web information accessibility, understanding users’ navigation behavior, improving information retrieval and content delivery on the Web.

Hyperlink structure is a special feature in WWW. This feature can be utilized for finding the associative relation between Web-pages and used to obtain high quality search result like PageRank and HITS algorithm. While the term-based algorithm only considers the content of the Web-pages, the content-link approach considers both the content and link information of the Web-pages [10]. T.H. Haveliwala et. al. proposed a technique LSH (Local-Sensitive-Hash) for clustering the entire Web which mainly emphasized on scalability of clustering. Snippet-based clustering algorithm is another algorithm which is used for the similarity measures between Web-pages for clustering and is used sometimes to remove duplicates. Applying the technique of association rule to term vector is another approach for clustering. This approach can automatically produce the Web-pages without actually measuring the similarity. All these approaches use similarity of the contents of the Web-pages but differ in clustering method.

Many works have been concentrated on link analysis. HITS algorithm is one of them which is explained later in this paper. This paper proposes one partitioning algorithm where centroids are based on HITS algorithm, thus utilizes the information gathered from link analysis and similarity measures is used to determine the cluster of Web-pages.

## 2.1 Existing Partitioning approaches

Partitioning approaches of clustering obtain single level partition of objects. These approaches usually are based on greedy heuristics that are used iteratively to obtain optimum solution. Given  $n$ -objects, these

approaches make  $k \leq n$  clusters of data and the relocation iterative methods are used. It is assumed that one object falls into only one cluster and each cluster contains atleast one object.

The two common partitioning cluster approaches are K-means and K-mediod approach. K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. K-means method picks  $k$  seeds as centroids of the  $k$  clusters. The seeds are chosen randomly. The distance (mainly Euclidean distance) are computed of each object from each of the centroids. Each object to the cluster is assigned to the cluster it is nearest to. Then the centroids of the clusters is calculated by computing the means of the attribute values of the objects in each cluster. This process is going on until the cluster membership is unchanged. K-means is not suitable for categorical attributes.

In K-mediods algorithm, also known as PAM algorithm, initially a random set of  $k$  items is chosen to be the set of mediods. Then, at each step, all items from the input dataset that are currently mediods are examined one by one to check if they should be mediods. The algorithm checks whether any mediod objects can be replaced by any non-mediods objects. In doing so, the algorithm judges whether new mediods exhibits better quality of clusters.

These two methods are used for numerical attributes. Web is a set of documents which are converted to vectors to apply these algorithms. But the main problem with these approaches are determining  $K$  and the initial set of means or mediods which is addressed in this paper.

## 3. Proposed Approach

This paper proposes an approach to cluster the Web-page based on information provided by the hyperlink structure of Web-page as well as by the content of the documents of the Web- page. The process has two parts

- I. Determination of centroids
- II. Determination of clusters of other Web-pages

### 3.1 Determination of centroids

According to well known Kleinberg's HITS (Hyperlinked Induced Topic Search) algorithm, two types of Web-pages are used in WWW: Authorities and Hub Web-pages. The Web-pages present in the authority Web sites are known as authority Web-pages. Authorities are Web-pages that are recognized as providing significant, trustworthy, and useful

information on a topic. The Web-pages present in the hub Web sites are know as hub Web-pages. Hubs are index Web-pages that provide lots of useful links to relevant content Web-pages. The authority score of a Web-page is proportional to the sum of hub scores of Web-pages linking to it, and conversely, its hub score is proportional to the authorities scores of the Web-pages to which it links. In matrix notation, this translates to the following pair of equations:

$$\vec{a} = E^T \vec{h} \quad (i)$$

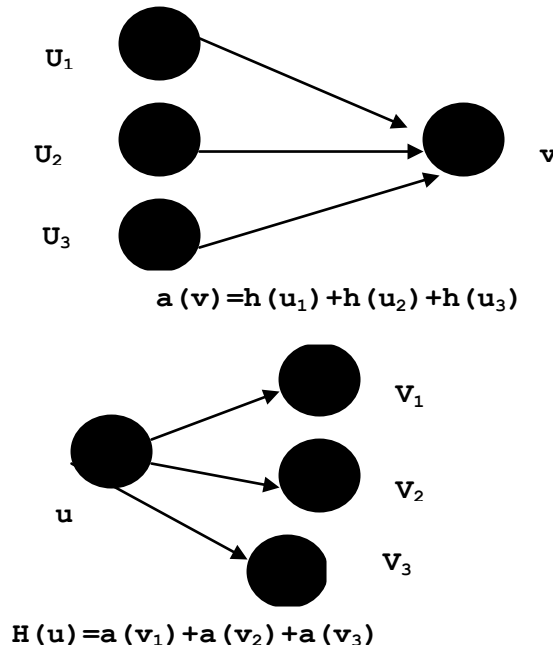
$$\vec{h} = E \vec{a} \quad (ii)$$

Where,  $\vec{a}$  is the authority score of a Web-page

$\vec{h}$  is the hub score of a Web-page

E is the edge set of the Web

Pictorially, authority and hub scores can be determined by Figure 1



**Figure 1:** Relationship of Hub and Authority Web-pages

Here we consider that the information provided by the hyperlink structure of “authorities” and “hub” Web-pages and also by the content of the documents of the “authorities” Web-page and the Web-page for which cluster have to be decided. The Web-pages which have the higher authority score more than that of the average of the graph are considered as centroids. Like there are three Web-pages in the graph p1, p2, p3 and their authority scores have been calculated by a(p1), a(p2), a(p3). Then their average authority score is a(av) = (a(p1) + a(p2) + a(p3)) / 3. If a(p1) > a(av) and a(p2) > a(av) and a(p3) < a(av) then p1 and p2 are considered as centroids. So, by the property of hyperlink structure of the Web, we have been able to determine the centroid and like other partitional

algorithms we need not give the number of clusters in advance.

Example 1: Let us suppose there is a collection of 10 Web-pages with following in-links and out-links given in Table 1.1 [11]

**Table 1.1:** In-links and out-links of the Web-pages

Web-page	In-Links	Out-links
A	B	G, H, I, J
B	C, D, E, F	A, G, H, I, J
C	None	B, G, H, I, J
D	None	B, G, H, J
E	None	B, H, I, J
F	J	B, G, I, J

G	A, B, C, D, F, H, J	H, I
H	A, B, C, D, E, G, I, J	G, I, J
I	A, B, C, E, G, H, J	H
J	A, B, C, D, E, F, H	F, G, H, I

Now, Hubs and authorities are computed iteratively. The result for that, with normalization is shown in Table 1.2. Here,  $x_0$  is the authority weight at start,  $y_0$  is the hub weight initially. Similarly  $x_1$  is the

authority weight at iteration 1 and  $y_1$  is the hub weight at iteration 1.  $x_2$  and  $y_2$  are authority and hub weights at iteration 2.  $X_0, X_1, X_2$  are normalized authority weights and  $Y_0, Y_1, Y_2$  are normalized hub weights. To obtain normalized values, original authority weights are divided by the value of the last row (SQRT) of respective columns. The SQRT is the square root of the sum of squares. Like for  $x_0$ , the sum of squares are 10. SQRT is the square root of 10, that is 3.16. After iteration 2, the sum of the authorities is 2.23. So the threshold value, according to our approach is .223. So, the centroids are B,G, H, I, J.

**Table 1.2:** Computing hubs and authorities with normalization

Page	x0	X0	y0	Y0	x1	X1	y1	Y1	x2	X2	y2	Y2
A	1	0.32	1	0.32	0.32	0.06	1.26	0.33	0.42	0.08	1.91	0.37
B	1	0.32	1	0.32	1.26	0.26	1.58	0.42	1.42	0.28	1.99	0.39
C	1	0.32	1	0.32	0	0	1.58	0.42	0	0	2.19	0.42
D	1	0.32	1	0.32	0	0	1.26	0.33	0	0	1.68	0.33
E	1	0.32	1	0.32	0	0	1.26	0.33	0	0	1.72	0.33
F	1	0.32	1	0.32	0.32	0.06	1.26	0.33	0.33	0.06	1.72	0.33
G	1	0.32	1	0.32	2.21	0.45	0.63	0.17	2.42	0.47	0.97	0.19
H	1	0.32	1	0.32	2.53	0.51	0.95	0.25	2.42	0.47	1.44	0.28
I	1	0.32	1	0.32	2.53	0.51	0.32	0.08	2.58	0.5	0.47	0.09
J	1	0.32	1	0.32	2.21	0.45	1.26	0.33	2.42	0.47	1.51	0.29
SQRT	3.16	1	3.16	1	4.94	1	3.79	1	5.15	5.15	1	5.16

### 3.2 Determination of Clusters

In this phase, the cluster to which Web-pages other than the centroids belong to, are determined. Here like all other documents, Web-pages are considered as vectors. Then the cosine similarity (described in Section 3.2.2) value between the Web-pages and each centroid is calculated. The Web-page belongs to that cluster whose centroid has maximum cosine similarity value with the Web-page. Suppose  $W \in \{WP\}$  where  $\{WP\}$  is the set of Web-pages and there are n no. of clusters. Centroids are  $p_i$  of cluster  $C_i$ , where  $i=\{1,2,3\}$ . The distance between  $W$  and  $p_i \forall i, D_i$  are calculated.  $W \in C_i$  for which  $D_i$  is maximum. That means the distance between the

Web-page  $W$  and centroids of all clusters are calculated and the centroid which is most similar to the Web-pages gives the cluster of the Web-page  $W$ .

#### 3.2.1 Representation of documents as Vectors

Every document has some words (terms). Among them some are stop words like “and”, “a”, “the”, “an”. These stop words are excluded from the list of terms.

Every term has frequency of occurrence. This idea is exploited to represent the documents as vectors.

Example 2: Let us suppose, there are two documents  $d_1$  and  $d_2$ . The document  $d_1$  contains “Cat, dog and cow are domestic animals. The girl has a cat and a dog.”. Other document  $d_2$  contains “Cow, dog and goat are domestic animals. The girl has a dog”.  $D_1$  contains the terms cat, dog, cow, domestic, animal, girl with frequencies 2, 2, 1, 1, 1, 1 respectively.  $D_2$  contains terms dog, cow, goat, domestic, animal, girl with frequencies 2, 1, 1, 1, 1, 1. First step of the process is to sort all the terms found in two documents. So the terms after sorting are animal, cat, cow, dog, domestic, girl, goat. The vector for  $d_1$  is (1, 2, 1, 2, 1, 1, 0) and for  $d_2$  (1, 0, 1, 2, 1, 1, 1).

#### 3.2.2 Cosine Similarity

Cosine Similarity is the most popular distance measure between two document vectors. The formula of cosine similarity between two document vectors is given in (iii)

$$\text{Cos}(d1, d2) = \frac{\text{dot}(d1, d2)}{\|d1\| \|d2\|} \quad (\text{iii})$$

where  $\text{dot}(d1, d2) = d1[0]*d2[0] + d1[1]*d2[1] + \dots$   
 $\|d1\| = \sqrt{d1[0]^2 + d1[1]^2 + \dots}$

Example 3: Let us consider the documents d1 and d2 of example 2. The vector of d1 is (1, 2, 1, 2, 1, 1, 0) and the vector for d2 is (1, 0, 1, 2, 1, 1, 1). The dot(d1, d2) is  $1*1 + 2*0 + 1*1 + 2*2 + 1*1 + 1*1 + 1*1 = 9$   
 $\|d1\| = \sqrt{1^2 + 0^2 + 1^2 + 2^2 + 2^2 + 1^2 + 1^2 + 0^2} = 3.464$   
 $\|d2\| = \sqrt{1^2 + 0^2 + 1^2 + 2^2 + 1^2 + 1^2 + 1^2} = 3$   
 Cosine similarity =  $9 / (3.46 * 3) = .866$

### 3.2.3 Algorithms

Following algorithms are used to find the centroids and clusters of the Web-pages. Algorithm 1 determines the clusters of all the Web-pages which are initially considered for clustering. These are base clusters. Algorithm 2 is used to find the cluster of any new Web-page added to the Web site.

#### Algorithm 1: Finding Base Cluster

Input: Database of Web Pages (WP= {wp1, wp2 ... wpn})

Output: Set of Clusters (C= {c1, c2... ck})

Step 1: While (WP are present)

Step 2: Calculate Authority weight (aw) for WP using HITS algorithm.

Step 3: Loop End

Step 4: Calculate threshold (th) value from

$$\text{Step 2 using } \frac{\sum_{i=1}^n aw_i}{n} \text{ where } i=1,2,\dots,n$$

Step 5: Check the value of authority weight of the Web-pages to see if it is greater than threshold value : if yes, then consider the WP according to aw value as Centroids, If no , then consider the WP according to aw value as Other Web Page(OWP).

Step 6: Calculate Cosine Similarity value between Centroids and OWP.

Step 7: The Cosine Similarity value of the Centroids which will be most similar to the OWP , under consideration will be considered as the cluster of the OWP.

Step 8: Stop.

#### Algorithm 2: Finding Cluster

Input: Base Cluster, Database of New Web Pages (NWP= {nwp1, nwp2 ... nwpm})

Output: Set of Clusters (C= {c1, c2... ck})

Step 1: Add New Web Pages (NWP) to previous Web pages which present in database.

Step 2: Follow Steps 1 to 8 of Algorithm1.

Step 3: Stop.

## 4. Experimental Result

The experiments have been carried out for different Web sites. One of the sample Web site is <http://www.ihrmcal.org> url. In experiments, the centroids and the clusters have been found out using the proposed approach. Then the centroids and the number of clusters have been changed and cost has been found out like K-mediods. The results of our experiments related to these Web-pages are given below.

**Table 1:** Cosine values of the Web-pages

Centroid Web-page	Other Web-page	Cosine values
2	1	0.029412
2	5	0.001090
2	6	0.000817
2	9	0.000516
3	1	0.000000
3	5	0.000437
3	6	0.000203
Centroid Web-page	Other Web-page	Cosine values
3	9	0.000205
4	1	0.000000
4	5	0.000367
4	6	0.000196
4	9	0.000148
7	1	0.000000
7	5	0.000277
7	6	0.000187
7	9	0.000142
8	1	0.000000
8	5	0.000229
8	6	0.013889
8	9	0.001998

Based on Cosine values found out from Table 1, Clustering of Web-pages are given in Table 2.

**Table 2:** Clustering of Web-pages

Cluster ID	Centroid	Other Web-pages
1	2	1,5
2	3	
3	4	
4	7	
5	8	6,9

**Total Cost:** 0.046389

The experiments have been carried out by changing the centroids and number of clusters. The cost have been calculated for all the cases. The result have been tabulated in Table 3. It has been concluded from the result shown in Table 3, the cost of proposed approach is maximum and hence our approach is better than other partitioning approaches in the area of Web-page clustering.

**Table 3:** The cost for different cases after changing centroids and number of clusters

Case ID	Old centroids	New Centroids	Old Cost	New Cost
1	2,3,4,7,8	1,3,4,7,8	.046389	.045736
2	2,3,4,7,8	2,3,4,6,8	.046389	.035138
3	2,3,4,7,8	2,3,4,7,9	.046389	.034498
4	2,3,4,7,8	3,4,5,7,9	.046389	.008444
5	2,3,4,7,8	2,3,4,7,8,9	.046389	.044391
Case Id	Old centroids	New centroids	Old Cost	New Cost
6	2,3,4,7,8	1,2,3,4,7,8	.046389	.016977
7	2,3,4,7,8	1,2,3,4,5	.046389	.004788
8	2,3,4,7,8	3,4,5,6,7,8	.046389	.006446
9	2,3,4,7,8	1,3,5,7,9	.046389	.034291
10	2,3,4,7,8	2,3,4,5,9	.046389	.036046

The efficiency of the clusters has been measured using C-index which is given by

$$C\text{-Index} = \frac{S - S_{\min}}{S_{\max} - S_{\min}}$$

where, S is the sum of all the distances between items within a cluster (for N items there will be N\*N/2 distances,

$S_{\min}$  is the sum of N\*N/2 of the smallest distances between all items,

$S_{\max}$  is the sum of N\*N/2 of the largest distances between all items.

If the C-Index value is less, the cluster is better.

For the presented approach, C Index Value is 0.67852441

**Table 4:** C-Index value for all the 10 cases given in Table 3

Case ID	C-Index value
1	0.721567831
2	0.692562152
3	0.823546213
4	0.756231591
5	0.689753211
6	0.953128466
7	0.789236945
8	0.856291562
9	0.987289652
10	0.723856982

From the results presented in Table 4, we conclude that C-Index value for the presented approach is lowest as compared to the cases presented in Table 3.

Lower the C-Index value, better is the cluster. So, our approach of clustering is the best compared to the 10 cases of table 3.

## 5. Conclusion

The paper presented a method for Web-page Clustering. In doing so, it has exploited the hyperlink structure properties and also the contents of Web-pages. From experimental result, it has been seen that the centroids and the number of clusters, found from HITS algorithm shows a good promise. Also, it has been found that the presented approach exhibits lower C-index value. So this paper demonstrated that without sacrificing the quality of clusters, the disadvantages of other partitioning cluster algorithms can be removed.

## 6. References

- [1] A.Vakali, J. Pokorn'y, and T. Dalamagas, 2001, "An Overview of Web Data Clustering", EDBT Workshops 2004, pp. 597-606.
- [2] M. Sinka and D. Corne, "A large benchmark dataset for web document clustering", Management and Applications 87, 2002.
- [3] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering", AAAI-2000: Workshop of Artificial Intelligence for Web Search, 2000, Austin, pp.58-64.
- [4] A. Bouguettaya, "On-Line Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol.8, No.2, 1996, pp.333-339.
- [5] D.W. Albrecht, I. Zukerman, and A. E. Nicholson, "Pre-sending documents on the WWW: A comparative study",

IJCAI99 – Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence.

- [6] E.M. Voorhees., “Implementing agglomerative hierarchical clustering algorithms for use in document retrieval”, Information Processing & Management, pp:465-476, 1986.
- [7] E Rasmussen., “Clustering algorithms”, Information Retrieval, pp.419-442, Prentice Hall, Eaglewood Cliffs, N.J., 1992.
- [8] M. A. Hearst and J.O. Pedersen, “Reexamining the Cluster Hypothesis : Scatter/Gather on Retrieval Results”, In proceedings of the 19th Annual International ACM SIGIR Conference, Zurich, June 1996.
- [9] R.Ali, U.Ghani, A.Saeed, “Data clustering and its applications”, [http://members.tripod.com/asim\\_saeed/paper.htm](http://members.tripod.com/asim_saeed/paper.htm).
- [10] Y. Wang, M. Kitsuregawa, “On combining link and contents information for web page clustering”, 13th International Conference on Database and Expert Systems Applications DEXA2002, Aix-en-Provence, France (September 2002), pp. 902–913.
- [11] G.K. Gupta, “Introduction to Data Mining with Case Studies”, Prentice Hall of India, 2006.