IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014
ISSN : 2348 - 6090
www.IJCAT.org

Clustering Methods of Web Pages: A Survey

[1] Aliya Edathadathil, [2] Syed Farook.K

[1] Mtech in the Department of CSE, MES Engineering College , Kuttippuram

[2] Asst .Prof in the Department of CSE, MES Engineering College, Kuttippuram

Abstract - It is a very difficult task to the web users with the existing search engines like Google, Yahoo, MSN, Ask, etc. That the information related to the entered query returns a long list of results or a long list of snippets. So that it is a difficult task to the user to go through each title, snippet and even sometimes link of the search results are irrelevant. Data clustering is the process of organizing similar data in to different groups and the members of those groups are similar in some way. In the case of webpage clustering we can organize the web pages in to different groups and it is very easy to the web users to search the web pages. Here performance of the different clustering techniques presented and compared.

Keywords - **Clustering of web pages, suffix tree clustering, graph partitioning method, K-means clustering, K-Medoid, Density based clustering.**

1. Introduction

Web Page clustering is the method of unsupervised learning. Clustering means partitioning the data set into different groups and the members of those groups are similar in some way. Internet is undoubtedly the fastest and easiest mode of access for unlimited resources of information. But the same reason is disabling the increasing efficiency of accessing information. It is aptly said that Internet is an unorganized, unstructured and decentralized place of accessing information. The web pages are increasing in billions since times.

However due to the rapidly growing and unstable characteristic of the web, such directories often point to outdated ,even not existing documents. Clustering of search results is a special technique in data mining using which the retrieved results are organized into meaningful groups enlightening the user work. It is not only posing challenges in the field of data mining but also in the areas of information Retrieval systems and in data warehousing

2. Literature Survey

Several techniques exist in the literature to address clustering of web pages. Researchers are always trying to improves the performance and efficiency of the clustered web pages.

2.1 Suffix Tree Clustering (STC) Technique

Suffix tree clustering technique is one of the main web page clustering technique[1]. Clustering would form groups of documents with similar features. By showing thematic groups the clustered results can increase readability and easy access. Suffix tree is one among the data structure techniques that facilitates easy search options to the user disabling the useless links.

In suffix tree clustering technique, first we have to create the suffixes of the corresponding string that we have entered in the search box of the search engine. For this purpose we consider a tree for storing these strings in which there is one node for every common prefix. The strings are stored in extra leaf nodes. We call this tree as a 'Trie'[1]. The suffix tree is a tree like data structure representing all suffixes of a string. consider an example string T= DATAMINING. The tree process the string symbol by symbol from left to right, and has always the suffix tree for the scanned part of the string ready.

We can make the suffix tree of the word DATAMINING as shown in the figure 1.To make these suffixes prefix-free we add a special character, say 'dollar' at the end of T. . So our word DATAMINING become DATAMINING$. so that the resultant leaf nodes formed are (n+1) in number where n is the length of the string.

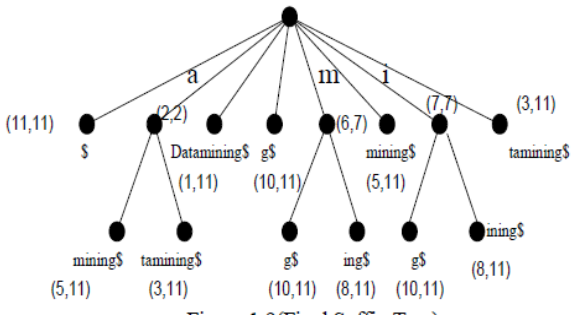

Fig 1. Suffix tree representation of the string 'datamining'

A. Suffix Tree Algorithm

There are mainly two phases for the algorithm[2]. In the first phase we have to scan the string and partition the suffixes. The second phase consist of the following steps.

1. START build suffix tree
2. Populate suffixes from current partition
3. Sort the suffixes on first symbol using temp
4. Then make leaf node to the tree
5. Push the unevaluated nodes to the stack
6. While stack is not empty
7. Pop the node
8. repeat until all the nodes are evaluated
9. end

After the creation of suffix tree of the particular string we have to apply the algorithm to cluster the web pages related to the entered string. The suffix tree clustering algorithm (STC) involves mainly five steps:

1). preparing the documents: Retrieving the document snippets from Google and parsing and stemming the results.

2). Suffix tree construction: Inserting the strings associated with each document on the suffix tree.

3). merging the clusters: Combining the similar nodes of suffix tree.

4). labeling the clustering: Generally a label for each cluster.

5). scoring clusters: Ranking the clusters.

Advantages of suffix tree clustering includes: (i) it uses less time for clustering and (ii)it provide overlapping, i.e, a document can appear in more than one cluster. This is in contrast to most other clustering algorithms.

2.2 Graph Partitioning Method

Graph partitioning method is another kind of webpage clustering technique. A partitioning method is one of the earliest clustering methods to be used in web usage mining [3]. They used an incremental algorithm that produces high quality clusters. Each user session is represented by an n dimensional feature vector, where n is the number of web pages in the session. The value of each feature is a weight, measuring the degree of interest of the user in the particular web page.

The core element of this system is a new clustering method, called cluster mining, which is implemented in the Page Gather algorithm. Page Gather receives user sessions as input. Using these data, the algorithm creates a graph, as signing pages to nodes. An edge is added between two nodes if the corresponding pages co-occur in more than a certain number of sessions. Clusters are defined either in terms of cliques, or connected components.

Generally, several pretreatment tasks need to be done before performing web mining algorithms on the Web server logs. For this work, these include data cleaning, user differentiation and session identification. These preprocessing tasks are the same for any web usage mining problem.

A . Navigation Pattern Modeling

To model navigational patterns we use the proposed algorithm[4]. In our contribution the degree of connectivity in each pair of pages depends on two main factors: the time position of two pages in a session and the occurrence of two pages in a session. We proposed an algorithm for modeling the pages accesses information as an undirected graph M= (V, E). The set V of vertices contains the identifiers of the different pages hosted on the Web server. We propose a weight measure for approximating the connectivity degree of each two web pages in sessions. First, introduce two concepts related to this measure, Time Connectivity" and "frequency".

B. Web Page Clustering Algorithm Based on Graph Partitioning Method

The algorithms mainly consist of four steps.

1. Using DFS starting from a vertex and searching for connected component reachable from this vertex.
2. Once such component has been found the algorithm checks if there are any nodes not considered in the visit.
3. If so DFS is again applied starting from one of the nodes not visited
4. repeat the procedure until all the connected nodes are visited

2.3 K-Means Clustering Method

K-Means clustering is one of the simplest unsupervised learning methods among all partitioning based clustering methods[5]. In K means clustering method a centroid is defined to represent each partition. Centroid is a point in space that represents the cluster representative. This method classifies a given set of n data objects in to K clusters, where K is the number of desired clusters and it is required in advance. A centroid is defined for each cluster and assign each data to the cluster which is closest to the centroid. The distance between centroid and the data are calculated by using any distance measurement method. Commonly used distance measurement method is Euclidean distance measurement. After the first iteration the centroids are

changed and repeat the process until centroids do not change.

The algorithm of K-Means clustering consist of following steps:

Input
- K, the number of clusters to be partitioned and n, set of n data.

Output
- Set of K clusters.

Steps
- Arbitrarily choose k centroids as the initial cluster centers.
- Repeat,
  - Assign each item to the cluster which has the closest centroid.
  - Update the centroid i.e., calculate the mean value of the objects for each cluster.
- Until no change

The main advantages are:
- Relatively scalable and efficient in processing large data Sets.
- Easy to understand and implement.

The main disadvantages are:
- Applicable only when the mean of a cluster is defined.
- Need to specify k, the total number of clusters in advance.
- Unable to handle noisy data and outliers.
- Result and total run time depends upon initial partition.
- Not that much efficient when the data are uncertain.

2.4 K-Medoid Clustering

K Medoid clustering is another type of partitioning based clustering method[6]. This method is same as that of K Means clustering method, the one difference is that here we are taking medoids instead of taking the centroids[1]. Medoid is most centrally located data in the dataset used to represent the cluster representatives. All other procedures are same as that of K means clustering.

The algorithm of K-Medoid clustering consist of following steps:

Input
- K,the number of clusters to be partitioned and n,set of n data items.

Output
- Set of K clusters.

Steps
- Arbitrarily choose K objects as the initial cluster medoids.
- Repeat,
  - Assign each item to the cluster which has the closest medoid.
  - Update the medoid.
- Until no change.

The main advantage is:
- More robust than k-means in the presence of noise and outliers.

The main disadvantages are:
- Relatively more costly.
- Need to specify k, the total number of clusters in advance.

2.5 Density Based Clustering Method (DB Clustering)

DB clustering is another data clustering technique. In the partition based clustering, the total n data are to K number of clusters and we have to define the total number clusters in advance. But in the case of DB clustering we do not need to define the total number of clusters in advance [7]. DB clustering is work as follows. To find a cluster, DB clustering starts with an arbitrary point p and retrieves all points density reachable from p wrt NEps(Neighbourhood points with maximum radius) and MinPts(Minimum number of points in an NEps-neighbourhood). If p is a core point, this procedure yields a cluster wrt. NEps and MinPts . If p is a border point, no points are density-reachable from p and DB clustering visits the next point of the database. Continue this procedure until all the data are processed [8].

The algorithm of DB clustering consist of following steps:

Input
- n, Set of n data items.

Output
- Set of k Clusters.

Steps
- Arbitrary select a point p
- Retrieve all points that are density-reachable from p.
- If p is a core point, then a cluster is formed.
- If p is a border point, then no points are density-reachable from p and DB Clustering visits the next point of the database.
- Continue the process until all of the points have been processed.

The main advantage is:
- Do not need to specify number of clusters in advance.

The main disadvantage is:
- Need to specify NEps and MinPts, which can be difficult

in practice.

3. Performance Analysis

Here various clustering algorithms of web pages are discussed .In the partitioning based clustering method like, K-means, K-medoid algorithm need to specify total number of clusters in advance. But in the case of Suffix tree clustering method, Graph partitioning clustering method and Density based clustering method the clusters are not need to specify in advance. In the case of partitioning based the initial partition affect the result and total run time. The performance comparison between the above methods are presented in the table 1.

TABLE I
PERFORMANCE ANALYSIS OF DIFFERENT METHODS

Different settings	STC	Graph Partitioning	K-Means	K-Medoid	DB clustering
Advanced specification of no: of clusters	No	No	Yes	Yes	No
Initial partition affect result and run time	No	No	Yes	Yes	No
Overlapping	Yes	No	No	No	No
Time	Less	More	Less	Less	Less

4. Conclusion

It can be concluded that partitioning based clustering methods are suitable for spherical shaped clusters in medium sized data sets. The choice of clustering algorithm depends on both the type of data available and on the particular purpose and chosen application. The K-means, and K medoid clustering, we need to specify the total number of clusters in advance. In the case of suffix tree clustering ,graph partitioning clustering and DB clustering we do not need to specify the total number of clusters in advance. In the case of DB clustering it is difficult to calculate NEps and MinPts. The researchers are trying to improve the clustering method to cluster the web pages effectively and accurately.

References

[1] D. F. Manne Suneetha and S. M. Z. Pervez, "Clustering of web search results using suffix tree algorithm."" in ICETECT 2011,IEEE 2011

[2] P. B. Guihong Cao, Dawei Song and E. Simon, "Suffix tree clustering on post retrieval documents." springer, 2008.

[3] N. M. Mehrdad Jalali and A. Mamath, "A new clustering approach based on graph partition for navigation pattern mining." IEEE, 2008.

[4] R.Barglia and M. Silvestri, "An online recommender system for large web sites," IEEE Computer Society. 2004 International Conference on (WI'04), 2004.

[5] Dr. T. Velmurugan, "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points" IJCTA, 2012.

[6] T. Feder and D.H. Greene, "Optimal Algorithms for Approximate Clustering" Proc. Ann. ACM Symp. Theory of Computing (STOC), 1988.

[7] Morteza Haghir Chehreghani, "Improving density-based methods for hierarchical clustering of web pages" Elsevier, 2009.

[8] Martin Ester, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise ",ACM, 2010.