Recuperating Link Structure of Website using MST

Prateek Dwivedi
SRMCEM, Lucknow
UP, India

Pushkar Singh Negi
SRMCEM, Lucknow
UP, India

Raj Gaurang Tiwari
SRMCEM, Lucknow
UP, India

ABSTRACT

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in find the desired information resources, and to track and analyze their usage patterns. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge. Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web.

Our aim is to utilize the concept of using an efficient MST approach to assist easier web navigation with involvement of clustering concept. In this paper we analyze and thereby make use of implementing the same concept in it, which works on the fact of constructing a minimum spanning tree of a point set i.e. nodes and removes edges that satisfy a predefined criterion.

Keywords

Cluster Heads, Gateway Nodes, MST (Minimum Spanning Tree), CHMST(Cluster Head minimum spanning tree).

1. INTRODUCTION

Clustering is a division of data into groups of similar objects, called clusters. The objects in a cluster are similar between themselves and dissimilar compared to objects of other clusters. Clustering methods are broadly classified as hierarchical and partitioning clustering. Hierarchical clustering views each data point as a node and at each iterative step merges two neighboring nodes to form a new node. A tree is constructed in a bottom up fashion after $n - 1$ steps, where n is the number of data points. For merging two nodes different linkage methods can be used which decide the neighboring pair of nodes to merge. The problem with hierarchical clustering is that it decides the nodes to be merged locally on the basis of some form of linkages without taking a global objective into consideration and once nodes are merged they cannot be separated in later steps.

In contrast to the hierarchical clustering algorithm, partitional clustering find a single partition of the patterns instead of a clustering structure. It usually generates clusters by evaluating a criterion function which is defined locally or globally and attempts to recover the natural clusters present in the patterns. The advantage of partitional clustering methods is that they are especially appropriate in the analysis of large data sets. The problem with partitional algorithm is the setting of parameter for the number of desired output clusters. The graph-theoretic clustering is one of the partitional clustering techniques to partition the given dataset. The well-known graph-theoretic clustering algorithm is based on building of the minimal spanning tree (MST) of the data, and then deleting longer edges from the MST to generate clusters.

In cluster analysis, one of the most important issues is the measure used to evaluate the quality of the clustering results that are produced. This measure can then be used to compare the solutions from different algorithms. This paper is intended to study and compare different clustering based on Minimum Spanning Tree. The results are to be compared on different basis.

2. RELATED WORK

There is a large literature and comprehensive work plan being expended already on obtaining useful results for minimum spanning tree based approach to easier web navigation. The work of Zahn [1] presents a segmentation method based on the minimum spanning tree (MST) of the graph.

This method has been applied both to point clustering and to image segmentation. The edge weights in the graph are based on the differences between points. The segmentation criterion in Zahn's [1] method is to break MST edges with large weights. The work done in this field by Ramakrishnam Raju & Valli Kumari [8] provides useful and deep insights to this. They used validity index thus in estimating optimal number of clusters (k) for division of nodes into various clusters. According to their work validity index makes use of COMPACTNESS & ISOLATION as the basis. Compactness measures the internal cohesion among the data elements whereas isolation measures separation between the clusters. The compactness is measured by Intra-cluster distance and separation is measured by Inter-cluster distance. Intra-cluster distance is the average distance of all the points within a cluster from the cluster centre whereas Inter-cluster distance is the minimum of the pair wise distance between any two cluster centers.

VALIDITY INDEX= INTRA C.D/INTER C.D

As a goal of reaching to this predefined number of clusters (k) criteria the usage of Linkage algorithms is relevant. For n samples, agglomerative algorithms [9] begin with n clusters and each cluster contains a single sample or a point. Then two clusters will merge so that the similarity between them is the closest until the number of clusters becomes one or as given by the user.

1. Start with n clusters, and a single sample indicates one cluster.
2. Find the most similar clusters *Ci* and *Cj* then merge them into one cluster.
3. Repeat step 2 until the number of cluster becomes one or as specified by the user.

These plays a vital role in reduction of navigation cost/time factors to suit and assist user specific needs which we will describe later how these can be improved upon effectively. The distances between each pair of clusters are computed to choose two clusters that have more opportunity to merge.

There are different methods to calculate the distances between the clusters *Ci* and *Cj*.

X1, X2, ... , Xk =Observations from cluster 1
Y1, Y2, ... , Yk = Observations from cluster 2
d(x, y) = Distance between a subject with observation vector *x* and a subject with observation vector *y*.

The methods for calculating distance between clusters are called linkage methods [5] and are as shown below:

Single Linkage: The distance between the two closest members of two clusters is,

$D_{12=}{}^{min}{}_{ij}d\ (X_i, Y_j)$

Complete Linkage: The distance between the two farthest members of two clusters is,

$D_{12=}{}^{max}{}_{ij}d\ (X_i, Y_j)$

Average Linkage: This method involves looking at the distances between all pairs and averages all of these distances.

$D_{12}=1/\ (kl)\ {}_{i=1}{}^{k}\Sigma\ {}_{j=1}{}^{l}\Sigma\ d\ (X_i, Y_j)$

In addition to it, the work of S. J. Peter & S. P. Victor in their paper entitled S. P. Victor, S. J. Peter, *A Novel Algorithm for Minimum Spanning Clustering Tree [10]* cited in related work with what is called as INCONSISTENCY MEASURE for cluster formation. There exist numerous ways to divide clusters successively, but there is not suitable choice for all cases. Zahn [1] proposes to construct **MST** of point set and delete inconsistent edges – the edges, whose weights are significantly larger than the average weight of the nearby edges in the tree. Zahn's inconsistent measure is defined as follows. Let *e* denote an edge in the **MST** of the point set, *v1* and *v2* be the end nodes of *e, w* be the weight of *e.* A *depth neighborhood* ***N*** of an end node *v* of an edge *e* defined as a set of all edges that belong to all the path of length *d* originating from *v*, excluding the path that include the edge *e.* Let *N1* and *N2* be the depth *d* neighborhood of the node *v1* and *v2.* Let ŴN1 be the average weight of edges in *N1* and *σN1* be its standard deviation. Similarly, let ŴN2 be the average weight of edges in *N2* and *σN2* is its standard deviation. The inconsistency measure requires one of the three conditions hold:

1. w > ŴN1 + c x σN1 or w > ŴN2 + c x σN2
2. w > max(ŴN1 + c x σN1 , ŴN2 + c x σN2)
3. w > f max (c x σN1 , c x σN2)
where *c* and *f* are preset constants.

All the edges of a tree that satisfy the inconsistency measure are considered inconsistent and are removed from the tree. This result in set of disjoint sub trees each represents a separate cluster. In partitional graph clustering [8] terminology, the objective is to eliminate in consistent edges leading to clustering where the disconnected sub graphs lead to resulted clusters.

Now this is being used by us in our formal approach of cluster formation.

The MST Algorithm: [8]

```
Input: S the point set.
Output: number of clusters and validity index
Let $e_1$ be an edge in the MST constructed from S
Let W be the weight of $e_1$
Let σ be the standard deviation of the edge weights in MST
Let $S_T$ be the set of disjoint sub trees of MST
Let $n_c$ be the number of clusters

1. Construct an MST from S
2. Compute the average weight of Ŵ of all the Edges from MST
3. Compute standard deviation σ of the edges MST
4. $S_T = ø$; $n_c = 1$; C = ø; validity ratio=Inf;
5. Repeat
6. For each $e_l$ Є MST
7. If (We > Ŵ + σ)
8. Remove el from MST
9. $S_T = S_T$ U { T' } // T'' is new disjoint
Sub tree (regions)
10. $n_c = n_c+1$
11. Compute the center Ci of Ti
12. End
13. Compute Validity Index
14. Until (all the edges whose length We > Ŵ + σ are removed)
15. While (Validity index decreasing)
16. Compute the center Ci of each Ti Є $S_T$
17. Merge the two clusters whose distance between
their centers is minimum.
18. $n_c = n_c-1$
19. Compute Validity Index
20. End
21. $n_c = n_c+1$
22. Return $n_c$ , Validity Index values.
```

The MST algorithm[8] does not require the user to specify the parameters to terminate the algorithm. The algorithm terminates when the condition We > Ŵ + σ, is not satisfied. The second phase of the algorithm i.e. merging continues as long as the validity ratio is decreasing. In this algorithm the user intervention is minimized. The final number of clusters and validity index of clustering are returned finally.

Table 1. Comparative study

Author	Techniques/Focus	Advantages	Limitations
Jianhan Zhu[3]	Assist users in navigating web sites efficiently by learning link structures and user behaviors.	Mining conceptual link hierarchies and predictions using Markov Model for web site link structure.	Its application in Clustering aspect and access view.
Miguel Gomes et. Al[4]	It surveys the area of web mining especially web structure mining.	Providing useful starting point for further research.	Necessity of introduction of a web site link structure concept.
Oleksandr Grygorash et. Al[6]	Detection of clusters with irregular boundaries based on k-partition of set of points.	Study and application of MST concept in web.	Application of MST in various clusters being formed.
Ramakrishnam Raju et. Al[8]	Study of data clustering concept in minimum spanning tree.	Focus on different types of clustering classification's.	Irregularity of clustering boundaries in data mining .
Soumen Chakrabarti[14]	Exploiting the hyperlink structure of the www for information discovery and categorization.	Concept of HITS	Learning of link patterns.
William B. March et. Al[18]	Analysis of EMST in clustering & algorithms.	Usage of adaptive algorithm for optimal runtime bound.	Optimality problem still in question.

3. OUR APPROACH TO EASIER WEB NAVIGATION

Based on efficient and effective works cited above we can divide the work proposed into three major steps as follows:

- Construction of MST.
- Formation of clusters.
- Establishment of Cluster Heads for link setup and communication.

We could use the known algorithms like single source shortest path algorithm, all pair shortest path algorithm etc for construction of minimum spanning tree for given set of nodes (N) and set of edges (E) with predefined or variable weights (W) assigned to them.

Now, for efficient cluster formation we take the help of measure of inconsistency[8] here. The clustering basis also may depend upon similarity or relevance of web pages to satisfy user's navigational needs of web pages or on basis of weight of edges, so that the heavy weight edges belonging to a specific cluster cannot belong to other clusters. To address this problem Wenpu Xing and Ali Ghorbani in their paper entitled weighted page rank algorithm [17] cited Page Rank (Popularity based results for in links and out links) and HITS algorithm (Authoritative pages and Hubs concept).

The novelty of the proposed idea lies in Cluster head and cluster gateway mechanism incorporated in our algorithm. It is similar to that of CBRP Protocol being used in mobile Computing environment. Here the entire workload of finding the appropriate destination by the source node is distributed among the cluster heads which already maintain the routing tables at their location and automatically provide link establishment as & when required.

Also the cost of bidirectional link traversal in web semantics is reduced by introduction of binary or logical variable ch_value. After particular point of time it's the responsibility of the cluster head invoking the request for destination to the other related cluster heads in vicinity of it for determination of ch_value, each of whose value for each of the cluster heads is stored as an entry to the routing table.

EFFICIENT CLUSTERING BASED NAVIGATION TECHNIQUE INVOLVING MST

As per precondition of subject and plan it is moderately favorable to go through with analytical research originally. On reaching of gratifying work it can be preceded with investigational research as to set off, persuade and come out with the predicted consequences related to the recital of recommendation systems.

3.1 Graph-Based Approach

We take a graph-based approach to segmentation[12]. Let G = (V; E) be an undirected graph with vertices Vi€ V , the set of elements to be segmented, and edges (Vi; Vj) € E corresponding to pairs of neighboring vertices. Each edge (Vi; Vj) €E has a corresponding weight w((Vi; Vj)), which is a non-negative measure of the dissimilarity between neighboring elements Vi and Vj . In the case of image segmentation, the elements in V are pixels and the weight of an edge is some measure of the dissimilarity between the two pixels connected by that edge (e.g., the difference in intensity, color, motion, location or some other local attribute). However, the formulation here is independent of these definitions. In the graph-based approach, a segmentation S is a partition of V into components such that each component (or region) corresponds to a connected component in a graph G = (V; E). In other words, any segmentation is induced by a subset of the edges in E. There are different ways to measure the quality of segmentation but in general we want the elements in a component to be similar, and elements in different components to be dissimilar. This means that edges between two vertices in the same component should have relatively low weights and edges between vertices in different components should have higher weights.

3.2 Efficient clustering based navigation technique involving MST steps

With the ongoing advent of efficient clustering based navigation techniques involving MST based approach; we have formulated the following steps based on analytical observations to suit the needs of improved user navigation when it is pertaining to World Wide Web. This may be summarized by following steps:

With the ongoing advent of efficient clustering based navigation techniques involving MST based approach; we have formulated the following steps based on analytical observations to suit the needs of improved user navigation when it is pertaining to World Wide Web. This may be summarized by following steps:

3.3 Our CHMST Algorithm:

Let e be set of edges in the MST constructed from S.
Let W be the weight of edge e.
Let ∂ be the standard deviation of the edge weights in MST

► Input: Point set S (web pages), edge set, edge weights (links).
► Construct a MST from the given point set S.
► Formation of Clusters:

* Let initial number of clusters is one (include all point set).

Compute average weight $\hat{W}$ and Standard deviation ∂ of given edges.

* For each edge weight W, if W> $\hat{W} + \partial$ then remove inconsistent edges.

* And this way clusters are formed incrementing number of cluster by one each time until k clusters are formed.

► Let CHi be a cluster head for each cluster i.

► The CH_i of the current cluster containing node requesting for destination will search for destination page in own cluster, if successful; performs link setup , communication and finally link termination.

►If unsuccessful, the CHi of the current cluster will transmit the *dest_id* of required web page to the cluster head of the remaining clusters.

►Each cluster head will check for the requested id recursively and accordingly will set *ch_value* i.e. false or true.

► The invoking cluster head CH_i will check for this *ch_value* after particular time frame and update the same in each of their routing table's.

►if *ch_value*=false then CH_i will skip that cluster else will search in that particular cluster for requested page where the returning *ch_value*=true.

► Process the data transmission between sender and receiver page and then finally terminate the link.

4. RELEVANT EXAMPLE

To demonstrate this by an example, after constructing the MST of the given nodes and performing clustering, there may be suppose three clusters with three cluster heads (CH_i) and two gateway nodes (GN_i) and other connected nodes. Let say a source node S in cluster 1 want to communicate to destination node D in cluster 3. Then instead of searching the desired destination and unnecessarily tracing different routes, which consumes time, to other irrelevant nodes; it would just broadcast the destination id of the destination to each of the cluster heads. Since the cluster heads keep all the information of the associated nodes in the cluster, it will send a true message as a reply to the source if it contains the destination node within its cluster boundary else will return the false message. So, cluster head 2 will return false message whereas cluster head 3 will return true message to cluster head 1.Now the search stress on cluster head 1 gets reduced and the sender on receiving the true message by that cluster head 3 would establish communication link. And instead of irrelevant communication, it would just access the path to the destination D determined by the replying cluster head.

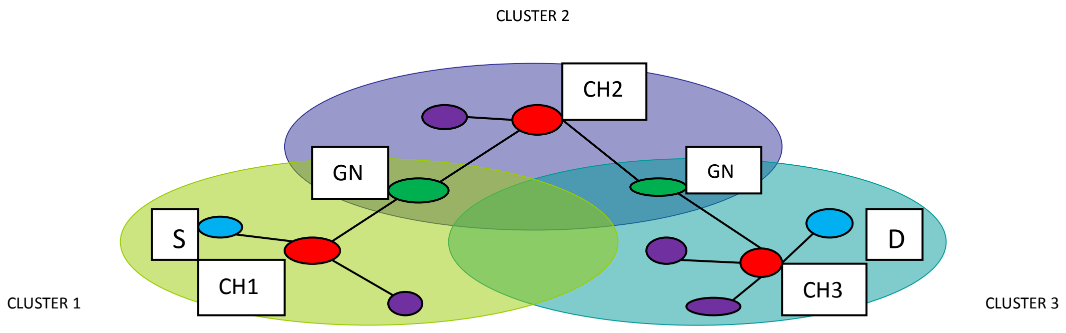

Fig 1: Web Page Arrangement

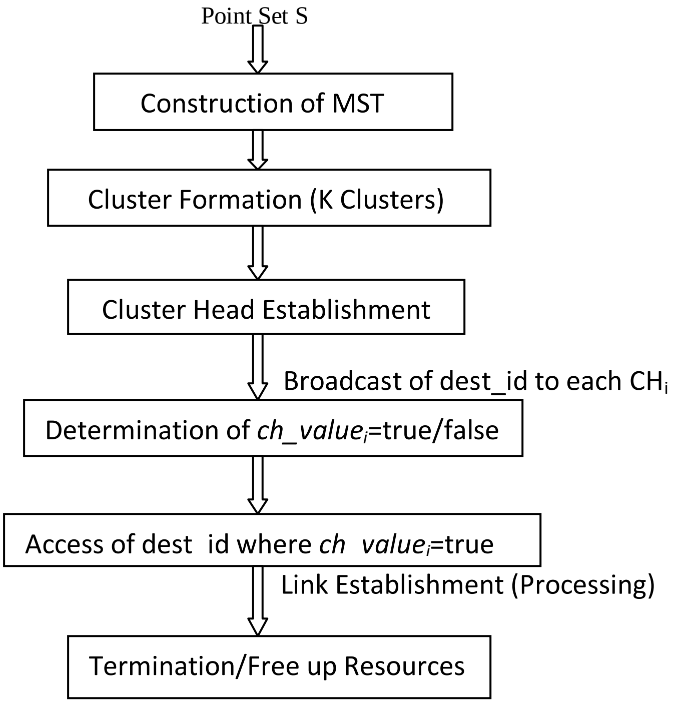
Point Set S
Construction of MST
Cluster Formation (K Clusters)
Cluster Head Establishment
Broadcast of dest_id to each CH$_i$
Determination of *ch_value$_i$*=true/false
Access of dest id where *ch value$_i$*=true
Link Establishment (Processing)
Termination/Free up Resources

Fig 2: Process Flow

5. APPLICATIONS

The said method could be used for:

a) Ordering documents matching a user query (ranking)
b) Deciding what pages to add to a collection
c) Page categorization
d) Finding related pages
e) Finding duplicated web sites
f) To find out similarity between them
g) Minimize the access time for a particular related web page in a set of data warehouse.
h) Enhance computational efficiency thereby lowering cost.

6. RELEVANCE OF WORK

1) ***Lower navigation cost*:** As compared to conventional algorithms implemented for such motive we have efficient technique due to the fact that instead of checking of resultant *ch_value=true/false* by the invoking cluster head repeatedly ,resulting from link failure or cluster head failure leading to unnecessary travel cost, to each of the other cluster heads, this responsibility entirely lies with the invoking cluster head to check this value after receipt of an acknowledgement signal in form of ch_value by other cluster heads and set up the link status to only that cluster head giving *ch_value=true.* So unnecessary cost is terminated by removal of connection to those cluster heads which do not contain the required destination node within their cluster boundary.

2) ***Application to mobile and fixed networks*:** This suggested algorithm may provide useful results ranging from World Wide Web semantics when related to fixed networks i.e. a wireless LAN or mobile computing environment by Cluster based routing Protocol (CBRP) in Wireless Application Environment.

3) ***Usage of proactive and active algorithms**:* As and when need arises it may be more suitable to environment to have an adaptive strategy employed. It means we can opt for proactive (Table Driven) routing algorithms or reactive (On Demand) algorithms to suit to user needs. This is left to subject and requirements of others researchers in the same area.

4) ***Triple step easier web navigation implementation:*** As already discussed we provide this approach by steps as formation of minimum spanning tree, Establishing clusters and cluster head based approach each of which greatly reduces the complexities and the area of search for the destination in same or varying LAN environments.

5) ***Ease of implementation subject to future research work:*** As already visible this theoretical approach involves graph theory, data structures and computer networks in light of suitable simulating environment. So it depends upon scalability and implementation ease of proposed algorithm by the those trying to do so, so as to how they can make it more convenient and changing according to their needs. Hence, it may holds promising results if more work is done in this field.

7. CONCLUSION

The clustering aims at recognizing and digs out significant groups in underlying data. Thus based on a clustering criterion the data are grouped so that data points in a cluster are more similar to each other than points in different clusters. Since clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed. In this paper we discussed and compared different categories in which algorithms can be classified (partitional, hierarchical). We concluded the discussion on clustering algorithms by a comparative presentation using the validity ratio that MST clustering algorithm performs better than the classical K-means partitional algorithm. The experimental results also showed that the hierarchical trees produced by partitional algorithms are better than those produced by agglomerative hierarchical clustering algorithms. This paper demonstrated that MST based clustering algorithm outperforms other clustering algorithms, including Hierarchical and k-means on most of the benchmark data sets.

The purpose of the paper is to have a brief survey on clustering techniques and their applications in current web scenario. Our main motive was to analyze and relate the given clustering algorithm with nodes as web pages and links as edges and exploit concept of MST in it to render different clusters with least weight edges to support more suitable web navigation.

8. ACKNOWLEDGMENTS

This research work was supported by Mrs. Vinodini Katiyar, The Head of Department, Computer Science & Engineering Deptt., SRMCEM and under the project guidance of Mr. R. G. Tiwari. The author's would also like to express deep sense of gratitude, especially to the below mentioned authors of related research papers and to each and everyone who have directly or indirectly helped to achieve this goal.

9. REFERENCES

[1] C. Zahn, 2011, "Graph theoretical methods for detecting and describing gestalt clusters" .

[2] D. Avis,1986, "Diameter Partitioning", Discrete and Computational Geometry.

[3] Jianhan Zhu, 2003,Mining Website Link structures for Adaptive website navigation and Search.

[4] Miguel Gomes da Costa Junior, Zhiguo Gong,2005, Web Structure Mining: an Introduction.

[5] Margaret H.Dunham,2003, "Data Mining Introductory and Advance Topics", Low price Edition – Pearson Education, Delhi.

[6] Oleksandr Grygorash, Yan Zhou, Zach Jorgensen, Minimum Spanning tree Based clustering Algorithms

[7] Pedro F. Felzenszwalb, Daniel P. Huttenlocher, Efficient Graph-Based Image Segmentation.

[8] Ramakrishnam Raju, Valli Kumari, 2011,Comparison of parameter free clustering algorithm with hierarchical agglomerative clustering Algorithms.

[9] Ricardo Baeza –Yates ET. Al.2005, WIM: an Information Mining Model for the Web.

[10] S. P. Victor, S. J. Peter,2005-10, A Novel Algorithm for Minimum Spanning Clustering Tree.

[11] S. P. Victor, S. J. Peter, 2010,A Novel Algorithm for Informative Meta Similarity Clusters Using Minimum Spanning Tree.

[12] S. P. Victor, S. J. Peter, 2010,A Novel Minimum Spanning Tree Based Clustering Algorithm for Image Mining.

[13] S. Chidambaranathan, S. J. Peter, 2011,Detection of Outlier-Communities Using Minimum Spanning Tree.

[14] Soumen Chakrabarti, Byron E. Dom ET. Al.,1999, Mining the link structure of the World Wide Web.

[15] T. Asano, B. Bhattacharya, M. Keil, F. Yao,1988, Clustering algorithms based on minimum and maximum spanning trees.

[16] T. Karthikeyan, S. J. Peter, 2011,Outlier Removal Clustering through Minimum Spanning Tree.

[17] Wenpu Xing, Ali Ghorbani, Weighted PageRank Algorithm.

[18] William B. March, Parikshit Ram, Alexander G. Gray, Fast Euclidean Minimum spanning tree:Algorithm,Analysis and Applications.