ISSN 2320-9984

Volume 1, Issue 3, October 2013 ISSN: 2320-9984 (Online)

International Journal of Modern Engineering & Management Research

Website: www.ijmemr.org

Improvement in accuracy and performance factor of Hierarchical Data Divisive Soft Clustering Algorithm

Samta Khobragade
Department of Computer Science Engineering
M.Tech.(CSE), SRIST Jabalpur(M.P.)
Email : samta.khobragade@gmail.com

Prof. Prateek Gupta
Department of Computer Science Engineering
M.Tech. (CSE), SRIST Jabalpur (M.P.)
Email : pguptacs@gmail.com

Abstract— *The process of grouping a set of physical or abstract object into classes of similar objects is called clustering. There are several techniques and algorithms are used for extracting the hidden patterns from the large data sets and finding the relationships between them.*

The main novelty of the Hierarchical Data Divisive Soft Clustering (H2D-SC) algorithm is that it is a quality driven algorithm, since it dynamically evaluates a multi-dimensional quality measure of the cluster to drive the generation of the soft hierarchy. Specifically, it generates a hierarchy in which each node is split into a variable number of sub-nodes.

Cluster at the same hierarchical level share a minimum quality value: cluster in lower levels of the hierarchy have a higher quality, this way more specific clusters (lower level clusters) have a higher quality than more general clusters (upper level clusters). Further, since the algorithm generates a soft partition, a document can belong to several sub-clusters with distinct membership degrees. The proposed algorithm is divisive, and it is based on a combination of a modified bisecting K-Means algorithm with a flat soft clustering algorithm used to partition each node.

Index Terms—*Web Data Mining, Clustering, XML, XPATH, XSL, ANT Clustering.*

1. INTRODUCTION

Clustering techniques have been widely applied in the information retrieval (IR) and information filtering (IF) context. In the IR and IF contexts it is desirable that a document is assigned to more than one cluster to a distinct degree. Soft clustering techniques generate overlapping clusters. Almost all clustering methods assume that each item must be assigned to exactly one cluster and are hence partitioned. However, in a variety of important applications, overlapping clustering is a technique where items are allowed to be members of two or more discovered clusters. The characteristic of soft clustering approaches with respect to the property of robustness i.e. the performance of a system should not be affected drastically due to outliers (bad observations).

Another issue when clustering textual documents is need to generate a hierarchy, i.e. a taxonomy-like structure of clusters, so as to provide a classification of documents organized into topics of distinct granularity. This allows analyzing the contents of a collection at distinct levels of specificity. Actual hierarchical clustering techniques are inadequate since they generate a dendrogram graph, in which at each node a cluster (or two clusters) is (are) split (merged) into two child cluster (one parent cluster). From the dendrograms several flat partitions can be obtained by specifying distinct thresholds on the minimum intra-cluster similarity.

Therefore, from dendrograms one can derive cluster representing topics of distinct granularity, but at the cost of further processing.

Several soft and hierarchical clustering algorithms have been applied in the IR context to address the above issues, they still present problems. Mainly they are not scalable in terms of documents and terms, they do not generate soft hierarchies, they need several parameters in input to tune the clustering process.

Soft and hierarchical clustering: Soft clustering techniques have been defined to generate overlapping clusters, i.e. clusters to which elements belong with distinct degrees: they can be classified into probabilistic and fuzzy approaches.

Probabilistic or model-based methods assume that clusters can be represented by a parametric distribution. The most common probabilistic method is mixture of Gaussians and the most important algorithm m used to deal with this problem is called Expectation-Maximization (EM).

While flat approaches create a flat partition of clusters, hierarchical clustering algorithm generates a dendrogram i.e. a hierarchy in which each cluster is composed of two sub-clusters. Hierarchical clustering techniques can be classified into two sub-categories, divisive and agglomerative, on the basis of the approach used to create the hierarchy of clusters (either top-down or bottom-up).

Divisive clustering algorithms use an iterated cluster bi-sectioning approach. It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold.

Agglomerative algorithms, on the contrary, build the hierarchy by initially assigning each documents to a distinct cluster, and by iteratively merging pairs of clusters to obtain a single all-inclusive cluster. Probabilistic agglomerative clustering algorithm work by merging clusters so that the final likelihood of membership of the object to the cluster is maximized.

The other branch of soft clustering techniques is represented by algorithm based on fuzzy set theory. Fuzzy approaches are less demanding than probabilistic ones in terms of basic assumptions, and this is one of the reasons that motivated their application to textual document clustering. According to these techniques each object can belong to more than one cluster, with membership degree in [0, 1]. One of the most important fuzzy clustering algorithm is Fuzzy C-Means (FCMs).

FCM is a method of clustering which allows one piece of data to belong to two or more clusters and it is frequently used in pattern recognition.

2. EXISTING SYSTEM

Clustering has been widely applied in data analysis to group similar objects. Many algorithms are either similarity- based or Model-based. In general, the former requires no assumption on data densities but simply a similarity function, and usually partitions data exclusively into clusters. In contrast, model-based methods apply mixture model to fit data distributions and assign data to clusters probabilistically. This soft clustering is often desired, as it encodes uncertainties on data- to-cluster assignments. However, their density assumptions can sometimes be restrictive.

In contrast to flat clustering, hierarchical clustering makes intuitive senses by forming a tree of clusters. A main problem of flat clustering algorithms is that they explicitly require the number K of clusters to be generated as an input parameter. A common practice to estimate this parameter is to run the clustering algorithm several times with distinct values of K, and then selecting the K value that minimizes a validity measure. Several validity measures have been proposed for the FCMs algorithm.

Another common flat soft clustering algorithm is Expectation Maximization (EM) algorithm. It is a model-based clustering since it tries to recover the original model from the data. The model defines the clusters and the degrees of membership of documents to clusters. The EM algorithm has been applied in the context of information retrieval since it has been proved to be effective. Its advantage is that it is fast, scalable and easy to implement. But it also suffers from some disadvantages, such as the need of stating a priori the number of cluster to generate and the instability with respect to the starting seeds.

The K- means algorithm takes the input parameter, K, and partitions a set of n objects into K clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity.

The K-means algorithm proceeds as follows:

First, it randomly selects K of the objects, each of which initially represents a cluster mean or center. For each of the remaining object, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster, this process iterates until the criterion function converges.

Finally the bisecting K-means (BKMs) algorithm is a divisive hierarchical algorithm that generates a dendrogram of clusters. This algorithm starts with a single cluster containing all the documents in D, and works iteratively by dividing the clusters into two sub-clusters by using the basic K-Means. The bisecting step is repeated for a fixed number of times, and then the clusters with the highest overall similarity are selected. The iteration stops when clusters contain a single document or the desired number of clusters is reached. The main limitation of this algorithm is that it generates crisp partitions, and each cluster is divided into two sub-nodes.

The problem statement of this research work is to find out the method for information retrieval using Hierarchical Data Divisive Soft Clustering (H2D-SC) algorithm with reduced time complexity.

The Hierarchical-Data Divisive Soft Clustering Algorithm (H2D-SC):

This algorithm allows users to specify three optional input parameters: the desired minimum DMin and maximum DMax size of the clusters in terms of number of documents, and MAXL, the desired maximum number of hierarchy's levels. If they are not specified, the algorithm assumes that a cluster must contain at least two documents, while there is no upper limit. By specifying DMin and DMax the user declares that he/she wants clusters containing a number of documents greater that DMin and smaller than DMax. A user does not really want clusters neither containing thousands of documents, nor containing just a single or a couple of documents. The maximum number of hierarchy levels MAXL can also be specified depending on the utility of the result.

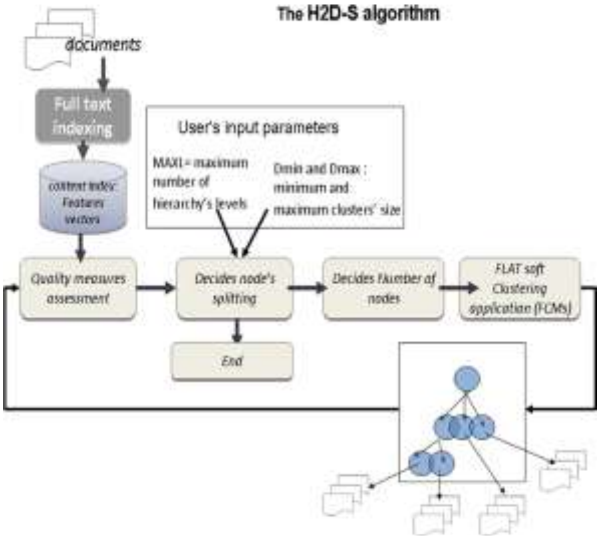
The H2D-S algorithm
documents
Full text indexing
content index: Features vectors
User's input parameters
MAXL = maximum number of hierarchy's levels
Dmin and Dmax : minimum and maximum clusters' size
Quality measures assessment
Decides node's splitting
Decides Number of nodes
FLAT soft Clustering application (FCMs)
End

Figure 1 : The H2D-S Algorithm

Description of the H2D-SC algorithm:

The initial cluster at the root of the hierarchy coincides with the whole collection i.e. C=D.

At the first step, the algorithm decides whether the cluster has to split by computing formula. To avoid useless computation we exploit the fact that the constraints imposed on the cohesion, the specificity, the fuzziness and the entropy are aggregated by OR.

At the root level the constraint on the minimum mass cardinality of a cluster is always satisfied since it is unlikely that a user specifies as minimum size of a cluster the dimension of the collection (generally Dmin << |D|). But this condition is not sufficient for deciding the splitting which can still be prevented by checking all the considered quality measures.

To compute the cohesion at the toot level the average document of the collection is assumed as the centroid vector of the initial cluster. Then the splitting of the root level is guaranteed by the fact that Dmin << |D| and the constraint on the cohesion is satisfied due to the

Initial setting with maximum value 1.

In the next levels, also the other quality dimensions may play a role in the splitting decision. Thus, we evaluate the other constraints by taking into account both their utility and their computational costs, by stopping the computations when the first encountered quality measure has degraded.

The number of sub-cluster KCL+1 of CL to be generated by the divisive approach is determined as follow:

KCL+1 = KminC+ round [(1- cohesion (CL)) (KmaxC – KminC)]

Where Kmin and Kmax are defined based on the input parameters Dmin and Dmax as follows:

[KminC= mass_cardinality (CL)] / Dmax and [KmaxC = mass_cardinality(CL)] / Dmin

The less cohesive is the cluster, the more it makes sense splitting it in a greater number of sub- clusters so as to try to generate higher quality clusters. Further, we want to keep the value KCL

Below the maximum desired limit KmaxC and above the minimum Kminc. Once KCL is set, the convergence of the algorithm is achieved by only re-clustering the documents belonging to the mass of C by applying a soft clustering algorithm.

Complexity analysis of the H2D-SC algorithm:

The complexity of the algorithm depends on several parameters:

| D | = number of documents

N = dimension of the features space (number of index terms)

DMin = minimum size of a cluster (optional parameter, DMindefault =2)

DMax = maximum size of a cluster (optional parameter, DMaxdefault = | D |)

MAXL = the maximum number of levels of the hierarchy (optional parameter, MAXLdefault = 2)

ITER = maximum number of iterations of the flat algorithm (optional parameter, ITER default = 5)

It also depends on the complexity of the flat algorithm used to partition each node. For the flat algorithms, FCMs, EM or KMeans, given Kc, determines on the basis of DMax and DMin as specified in the above formula.

The time complexity is:

O (N* | D | * KC * ITER)

The memory complexity is:

O (| D | * N + | D | * KC)

and

The disk complexity is:

O (N* | D | * ITER)

(Where | D | * N is the size of the data and | D | * KC is the size of the partition matrix).

Since at each iteration of the flat algorithm it can be necessary to read from the disk the dataset.

3. PROPOSED WORK

The Hierarchical-Data Divisive Soft Clustering Algorithm (H2D-SC): This algorithm allows users to specify three optional input parameters: the desired minimum DMin and maximum DMax size of the clusters in terms of number of documents, and MAXL, the desired maximum number of hierarchy's levels. If they are not specified, the algorithm assumes that a cluster must contain at least two documents, while there is no upper limit. By specifying DMin and DMax the user declares that he/she wants clusters containing a number of documents greater that DMin and smaller than DMax. A user does not really want clusters neither containing thousands of documents, nor containing just a single or a couple of documents. The maximum number of hierarchy levels MAXL can also be specified depending on the utility of the result. Based on DMin and MAXL and on the assessment of cluster's quality the soft hierarchy is generated by applying a divisive approach consisting of the following steps:

Step 1: Initially all the characteristics to be considered for cluster division is to be decided for document and documents are compressed in a grouped variable. This is consisting of two parts

- Characteristics
- Compressed Contents

Step 2: At the beginning all documents are viewed as full members of unique huge cluster C, the whole collection (i.e. all documents belong to the initial cluster C=D with membership value equal to 1).

Step 3: The decision whether to split C into more specific sub clusters are taken by assessing the quality of the cluster. Specifically, multiple endogenous and exogenous clusters' characteristics (properties) concur to the multidimensional cluster quality assessment decided in Step 1 are used.

Step 4: In the case the cluster C has been evaluated worth splitting, the number KC of sub-clusters to generate is determined depending on some cluster's properties.

Step 5: The algorithm applies a (soft) clustering algorithm (in the experiments the FCM algorithm was applied) having as input the number KC of clusters so as to generate the soft clusters of the next hierarchical level.

Step 5: The process iterates until no more cluster are evaluated worth splitting.

IMP:

- Characteristic collection and compression for each document is to be done for once and can be stored in secondary storage for future usage.
- Since data is compressed therefore the overall processing time taken will be less.

4. CONCLUSION

In this paper, we propose a new perspective of Web Mining through XML. Based on this, a novel WWW-oriented web recommendation system is proposed and shall be implemented. With the explosive growth of the World Wide Web, the amount of information available on-line is increasing rapidly. This certainly provides users with more options, but also makes it difficult to find the "right" or "interesting" information today. Web mining discovers user preference from the available data automatically and makes recommendations based on the extracted knowledge. More recently, a combination of web content, web structure and web usage mining has been studied and shows superior results in web recommendations.

The various research papers have been studied and decided the field. After reading and through guidance decided the final topic for implementation as dissertation work.

The studies of various algorithms of data mining has been done and since ANT Clustering algorithm has been least applied therefore it has been taken as field of application for this work.

REFERENCES :

[1] Arun K Pujari: Data Mining Techniques, Universities Press (India) Private Limited 2001.

[2] Jhon A. Hartigan:Clustering Techniques, Willey Publications 2005.

[3] Jaiwai Han and Micheline Kamber:. Data Mining concepts and Techniques, Elsevier 2006.

[4] Gloria Bordongna and Gabriella Pasi, National Research Council- IDPA, Dalmine (BG), Italy and University degli studi di Milono Bicocca, Italy. “**Hierarchical Data** Divisive Soft Clustering algorithm”.

[5] Liang Feng, Ming- Hui Qiu, Yu- Xuan Wang, Qiao- Liang Xiang, Yin- Fei Yang, Kai Liu, ECHO Laboratory, School of Communication and Information Engineering, Nanjing University of Posts and Telecommunication, Nanjing, Jiangsu, China and School of Computing, National University Of Singapore, Singapur. “**A fast divisive clustering** algorithm”.

[6] Arindam Banerjee Chase Krumpelman Joydeep Ghosh, Dept. of Electrical and Computer Engineering University of Texas, Austin, USA and Sugato Basu Raymond J. Mooney, Dept Of Computer Sciences University of Texas at Austin, USA.” **Model based** Overlapping Clustering”.

[7] Yanfei Zhao, “Study on Web Data Mining Based on XML”, 2012 International Conference on Computer Science and Information Processing (CSIP), 978-1-4673-1411-4 © 2012 IEEE.

[8] Xingyuan LI, Ningbo, China, Yanyan Wu, PING CHENG, “Research of Business Intelligence based on Web Accessing Data Mining”, The 7th International Conference on Computer Science & Education (ICCSE 2012) July 14-17, 2012. Melbourne, Australia, 978-1-4673-0242-5 © 2012 IEEE.

[9] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining”, Published by the IEEE Computer Society, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012, 1041-4347 © 2012 IEEE.

[10] Weigang Zuo,. Qingyi Hua, Weigang Zuo “The application of Web data mining in the electronic commerce”, 2012 Fifth International Conference on Intelligent Computation Technology and Automation, 978-0-7695-4637-7 © 2012 IEEE, DOI 10.1109/ ICICTA.2012.90.

[11] Jianli Duan, Shuxia Liu, “Research on web log mining analysis”, 2012 International Symposium on Instrumentation and Measurements, Sensor Network and Automation (IMSNA), 978-1-4673-2467-0/12 © 2012 IEEE.

[12] Jinyue Yang.Lin Yang, Customers’s intelligence:Kernel of CRM[J] Modernization of Management, 2002-07.