Document Clustering in Distributed Environment

R.Brintha
Department of Computer Science
Pondicherry University
Karaikal

S.Bhuvaneswari
Head
Department of Computer Science
Pondicherry University
Karaikal

ABSTRACT
Document clustering has emerged as a widely used technique with the increase in large number of documents that is getting accumulated day by day in various fields like news groups, government organizations, Internet and digital libraries. Document clustering is the process of grouping similar documents into clusters . A good document clustering algorithm should have high intra-cluster similarity and less inter- cluster similarity. i.e the documents with the clusters should be more relevant compared to the documents of other clusters.

In this paper, the implementation of document clustering in distributed environment based on peer to peer network architecture is reviewed. The documents in local site are clustered using K-means algorithm. Hierarchical clustering is obtained when clusters in each peer combine to form the next level of cluster. This process repeats until a global cluster is formed and is made available in all the peers. These clustered documents find its application in search engines.

Keywords
Clustering, Distributed Knowledge Discovery, k-means algoriothm, Supernodes, Intercluster, Intracluster

1. INTRODUCTION
Data mining in distributed environments is known as DDM, and sometimes as Distributed Knowledge Discovery (DKD). The central assumption in DDM is that data are distributed over a number of sites and that it is desirable to derive, through data mining techniques, a global model that reflects the characteristics of the whole data set.

Huge data sets are being collected daily in different fields; e.g., retail chains, banking, biomedicine, astronomy, and so forth, but it is still extremely difficult to draw conclusions or make decisions based on the collective characteristics of such disparate data. Four main approaches for performing Document Clustering in distributed environment can be identified.

- Centralized clustering
- Distributed clustering
- Hierarchical clustering
- Peer-to-Peer clustering

2. LITERATURE SURVEY

2.1 Document Clustering Methods
The process of determining the finite set of category of the objects to which the similar dataset may suit is known as Clustering. The applicability of clustering is manifold, ranging from market segmentation and image processing through document categorization and Web mining [1].

Document Clustering is increasingly widespread. It is finding application in browsing, in improving the similarity of search tools and in automatically generating thesauri. In query analysis clustering has been used for transforming a free text query into a fuzzy Boolean constraint. [5].

There are a couple of crucial ideas that must be taken into account while considering clustering in distributed environment.
1. The first of which is the idea of minimizing data transmission
2. A second point to consider is data privacy.

Clustering techniques can be broadly divided into three main types: overlapping (so called nonexclusive), partitional, and hierarchical.

2.1.1 Hierarchical Clustering
Hierarchical techniques produce a nested sequence of partitions, with a single, all-inclusive cluster at the top and singleton clusters of individual points at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level (or splitting a cluster from the next higher level). The result of a hierarchical clustering algorithm can be graphically displayed as a tree, called a dendogram. This tree graphically displays the merging process and the intermediate clusters.

Consider the behavior of scientific book editor, who needs to organize multiple research papers into a single book volume, with a hierarchical table of contents. Typically, research papers, even when coming from the same area, are written (1) in multiple writing styles, (2) on different levels of detail, and (3) in reference to different aspects of an analyzed area. Instead of searching for identical words and counting their occurrences, like many well-known computer-based text clustering techniques do [2]–[4], the human brain usually remembers only a few crucial keywords representing senses, which provide the editor with a compressed representation of the documents. These senses are then used to fit a given research paper into a book organization scheme, reflected by the table of contents. In our work, we replace editor's knowledge with ontology and use it to discover common senses that can then be used to organize documents.

2.1.2 Classical Partitioning Method: k-Means
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 the cluster's centroid or center of gravity. [5].

Algorithm: k-means

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Inputs

- k: The number of clusters
- D: a data set containing n objects

Outputs

- A set of k clusters

Method

1. arbitrarily choose k objects from D as the initial cluster centers;
2. **repeat**
3. (re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. Update the cluster means, i.e, calculate the mean value of the objects for each cluster;
   until no change;

3. PROPOSED WORK

3.1 Hierarchical Peer-to-peer clustering:

The HP2PC model is based on static hierarchical structure that is designed up front, upon which the peer network is formed. The goal is to achieve a flexible DDM model that can be tailored to various scenarios. The proposed model is called the **Hierarchically Distributed P2P Clustering (HP2PC)**. HP2PC is a hierarchically distributed P2P architecture for scalable distributed clustering of horizontally partitioned data. A scalable distributed clustering system should involve hierarchical distribution. A hierarchical processing strategy allows for delegation of responsibility and modularity.

The HP2PC model is based on static hierarchical structure that is designed up front, upon which the peer network is formed. Using the HP2PC model, we can partition the problem in a modular way, solve each part individually, and then successively combine solutions if it is desired to find a global solution.

The model lends itself to real-world structures, such as hierarchically distributed organizations or government agencies. In such scenario, different departments or branches can perform local clustering to draw conclusions from local data. Parent departments or organizations can combine results from those in lower levels to draw conclusions on a more holistic view of the data. HP2PC is a hierarchically distributed P2P architecture for scalable distributed clustering of horizontally partitioned data. The communication between nodes is accomplished by their supernodes. Supernodes are representative from each node. The notion of a node accompanied by a supernode can be applied recursively to construct a multilevel overlay hierarchy of peers; i.e., a group of supernodes can form a higher level neighborhood, which can communicate with each other on that particular level of hierarchy.

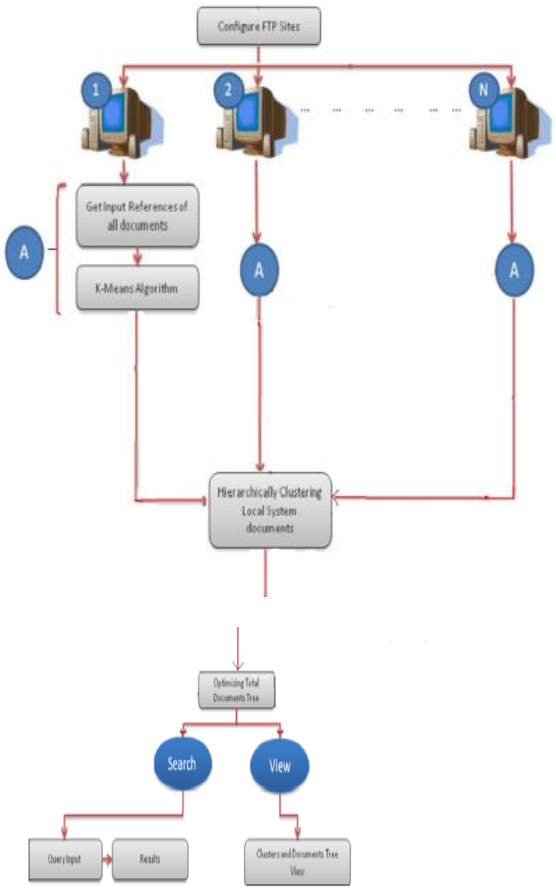

Fig 1: Framework

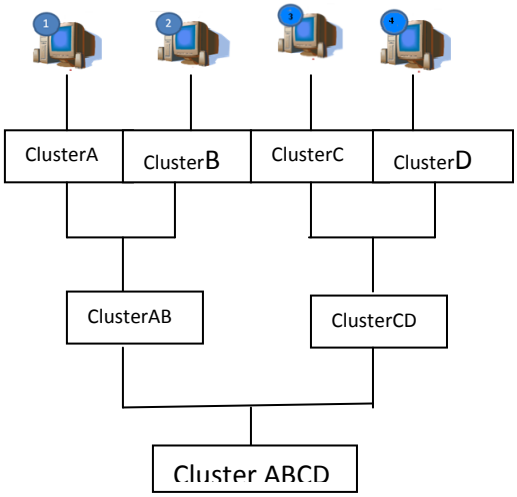

Fig 2: Logical Overview of Hierarchical Clustering

3.2 INPUT/ OUTPUT DESCRIPTION

3.3.1 Input Description
1. Input files of the form .doc or .txt

3.3.2 Output Description
1. Global Cluster containing all the files
2. Searching and viewing required files

3.3.3 Parameter Description
Node Characteristics
1. Node name
2. Directory of each node
3. Cluster id

Cluster Characteristics
1. Cluster id
2. No of clusters
3. Euclidean Distance
4. Document names
5. Document id
6. Document Path
7. Node name
8. Data Points
9. Frequent words

3.3.4 Database Description
(i) Table: FTPSites
1. FTPSiteName
2. Directory

(ii) Table: Localdocs
1. Document id
2. Document Name
3. Document Path
4. FTPSite
5. Cluster
6. Point

3.3 Proposed Algorithm-

Distributed Document Clustering Using Hierarchical Architecture

```
For each system
    Configure FTP
   Invoke FTP client object
    For all input files
       If input file is word document or text file
       Docs ← input file
       End if
     End for
   For each input file
      Performer Porter stemmer Algorithm
      Find the frequent word
    End for
  Perform K means clustering
End for
  For all nodes available
     Perform hierarchical clustering
  End for
```

3.4 HIERARCHICAL CLUSTERING

The Hierarchical clustering is done with the help of local tree obtained in each node after the local clustering (K- Means). After the local clustering is done in each system, a local tree is obtained, with the local clusters formed as its immediate children and the documents in each cluster as the children of the cluster. Now, during hierarchical clustering this local tree acts as the representative and they communicate with each other and form the next hierarchy of clusters with the documents in the systems that are communicating to form a global cluster.

3.5 SEARCH AND VIEW

The documents which reside in various systems can be searched by giving the keywords or the whole file name. Once we give the keyword, it searches the hierarchical (global) tree formed, which contains the documents as its children. If any match is found with the documents, those documents along with the path are sent to the requested system. If no match is found, then no documents will be sent.

4. IMPLEMENTATION

The k-means and hierarchical clustering algorithm have been implemented using Java Netbeans software. The databases used to store the nodes or systems connected in LAN and their corresponding files and clusters are created using Mysql. The components used in the clusterization process are the following.

- **Stemmer** - Stems the words in the document by removing the stop words and prefixes and suffixes of the word

- **Frequent word finder** - Finds the most frequently occurring word in the document

- **Word counter** - Counts the total number of words and calculate the word frequency

4.1 Performance Analysis

The performance of the clustering algorithm is measured based on the clustering speed. Experiments were carried out to compare the performances of centralized and hierarchical clustering algorithms by varying the number and domain of the documents. The variation of clustering speed with the change in number of documents is studied for these algorithms. By comparing the clustering speed for various no of documents, it is observed that the Hierarchical clustering algorithm has relatively good performance and the graph is shown below.

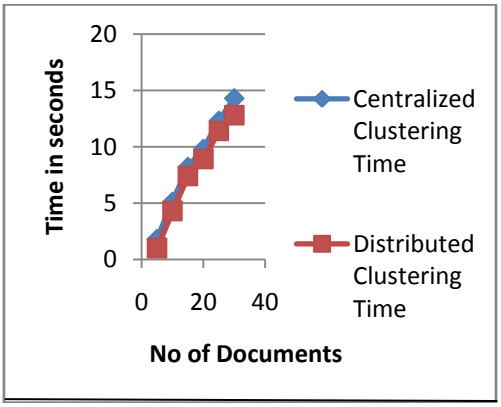

Fig 3: Performance of Centralized versus Distributed Clustering Algorithm

4.2 SNAPSHOTS

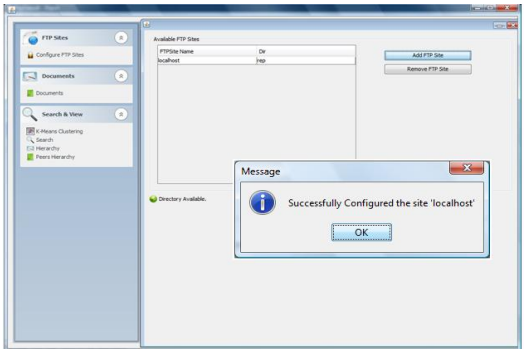

Fig 4: FTP Sites Configuration

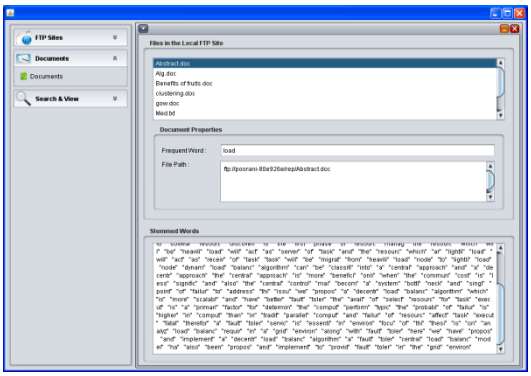

Fig 5: Retrieval of frequent word

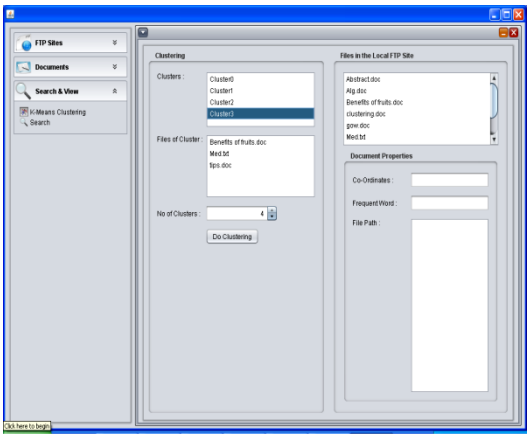

Fig 6: K-means clustering

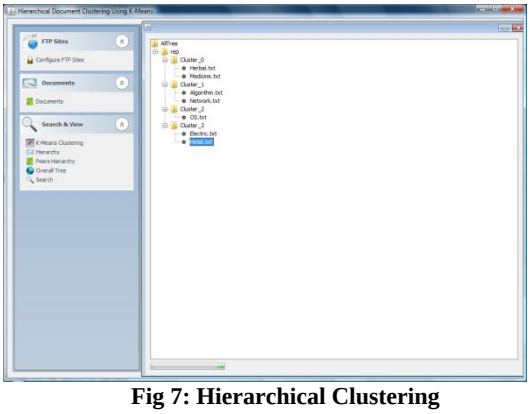

Fig 7: Hierarchical Clustering

5. CONCLUSION

A novel architecture and algorithm for distributed clustering, the HP2PC model, which allows building hierarchical networks for clustering data has been implemented. This is implemented and achieves comparable quality to its centralized counterpart while providing significant speedup and that it is possible to make it equivalent to traditional distributed clustering model. The importance of this contribution stems from its flexibility to accommodate regular types of P2P networks as well as modularized networks through neighborhood and hierarchy formation.

This model can be extended as dynamic, allowing nodes to join and leave the network, which requires maintaining a balanced network in terms of partitioning and height.

6. REFERENCES

[1] Yi Peng, Gang Kou, Yong Shi, Zhengxin chen , " A Hybrid Strategy for Clustering Data Mining Documents," IEEE international conference on data mining-workshops,2006

[2] Khaled M. Hammouda and Mohamed S.kamel, "Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization," IEEE transactions on knowledge and data engineering, vol. 21 , no.5, May 2009

[3] N.F. Samatova, G. Ostrouchov, A. Geist, and A.V. MelechkoRACHET: "An Efficient Cover-Based Merging of Clustering Hierarchies from Distributed Datasets," Distributed and Parallel Databases, vol. 11, no. 2, pp. 157-180, 2002.

[4] M.F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, no. 3, pp. 130-137, July 1980.

[5] Jiawei Han and Micheline Kamber, "Data Mining Concepts and techniques", Second Edition

[6] Hinrich Schiitze, Craig Silverstein "Projections For Efficient Document Clustering", Xerox Palo Alto Research Centre

[7] Douglass R.Cutting, David R. Karger, Jan O.Pedersen, John W.Tukey "Scatter/ Gather: A Cluster based Approach to Browsing Large Document Collections".

[8] Michael Steinbach, George Karypis and Vipin Kumar, "A comparison of Document Clustering Techniques", University of Minnesota.

[9] Debzani Deb, M.Muztaba Faud and Rafal A.Angryk,"Distributed Hierarchical Document Clustering", Motana State University, Bozeman,MT 59717, USA.