World Applied Sciences Journal 29 (Computer Sciences, Engineering and Its Applications): 241-247, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.29.csea.2275

Dynamic Peer-To-Peer Distributed Document Clustering and Summarization

R. Udayakumar, V. Khanaa and K.P. Thooyamani

School of Computing Science, Bharath University - 73, India

Abstract: The main objective of this paper is to provide cluster summarization of huge text document. Mining process includes the sharing of large scale amount of data from various sources, which gets concluded at the mined data. In distributed data mining, adopting a flat node distribution model can affect scalability, modularity, flexibility which are being overcome by using dynamic peer to peer document cluster and cluster summarization. The Dynamic P2P document cluster and cluster summarization architecture is based on a multilayer overlay network of peer neighborhoods. Dynamically created peer-to-peer systems are proving statements about the evolution of the system while nodes are continuously joining and leaving the group, because the system operates for an infinite time. The rate at which nodes consume resources to maintain the system state are being used rather than the performance measure based on runtime. It's because he runtime performance is uninformative. The clustering algorithm being used is called CBC (Clustering By Committee),which produces higher quality clusters in document clustering tasks as compared to several well known clustering algorithms. Within a certain level of the hierarchy, peers cooperate within their respective neighborhoods to perform P2P clustering. For document clustering applications, the system summarizes the distributed document clusters using a distributed key-phrase extraction algorithm, thus providing interpretation of the clusters.

Key words: Distributed data mining · Distributed document clustering · Hierarchical peer-to-peer networks · Document summarization

INTRODUCTION

Generally, data mining (sometimes called knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data [1, 2]. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. 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 DDM can be identified.

First Approach: This is very commonly used and it brings the data to the centralized site and then centralized mining is applied. This consequently leads to huge bottleneck at the centralized site. Additionally there may be delay, (Figure 1) collision and high computation cost involved at the centralized site to perform the mining process.

Second Approach: This approach performs local mining on the respective sites and the local models are pooled at a central site. The central site now merges all the local models to the global model. (Figure 2) [3]. Ensemble methods use this technique. This approach may not scale well with many sites and suffers from many disadvantages of the first approach since again computation cost is consumed in merging all the local models to one global model, but it is a better solution than pooling the data.

Corresponding Author: R. Udayakumar, School of Computing Science, Bharath University - 73, India.

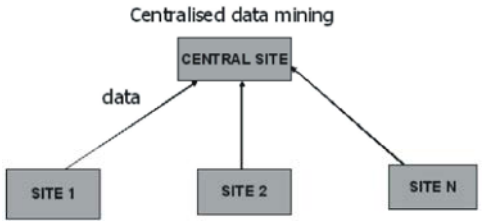

Fig. 1: First approach

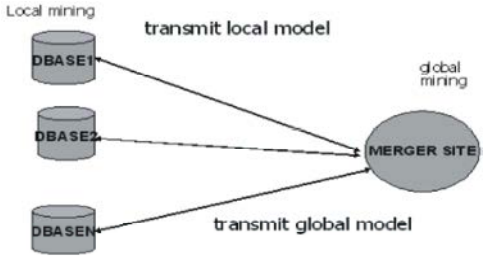

Fig. 2: Second approach

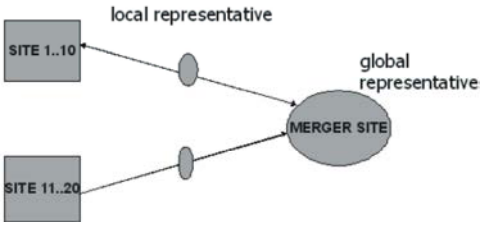

Fig. 3: Third approach

Third Approach: This approach first selects a small set of representative data sets from the local site and then transmits them to the global site (Figure 3) which combines them to global representative set and then performs mining process [5].

All previous three approaches involve a central site to facilitate the DDM(Distributed Data Management) process. But we must go for a different solution which may not involve centralized operation like a peer-to-peer (P2P) system.

Fourth Approach: This approach is called the hierarchically distributed P2P Clustering (HP2PC) [1] involving a hierarchy of P2P neighborhoods, in which the peers in each neighborhood are responsible for building a clustering solution, using P2P communication, based on the data they have access to. The HP2PC model is based on static hierarchical structure using peer network.

P2P Networks: P2P networks can be unstructured or structured. Unstructured networks are formed arbitrarily by establishing and dropping links over time and they usually suffer from flooding of traffic to resolve certain requests. Structured networks, on the other hand, make an assumption about the network topology and implement a certain protocol that exploits such topology.

This paper has been developed to introduce a dynamic structure extension to this model. Using the Dynamic peer to peer distributed document cluster and cluster summarization(DP2PC)model, partition the problem in a modular way, solve each part individually, then successively combine solutions if it is desired to find a global solution dynamically. This will avoid some problems in the DDM such as high communication cost usually associated with a structured, fully connected network and uncertainty in the network topology usually introduced by unstructured P2P networks.

The paper is organized as follows. Section II presents a brief review of the related literature and offers a historical perspective. Section III describes Dynamic Document Clustering and Summarization Architecture and Section IV describes about the implementation of this paper. Finally, We conclude the paper in Section V with a short summary and a few remarks on future work.

Related Work

Basic Definitions: Data:Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll and accounting
- Nonoperational data, such as industry sales, forecast data and macro economic data
- Meta data - data about the data itself, such as logical database design or data dictionary definition

Document Clustering: Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. Document clustering is the act of collecting similar documents into bins. It is a more specific technique for unsupervised document organization and fast information retrieval [4, 6].

Example: Web search engine returns thousands of pages in response to query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to group the retrieval documents into a list of meaningful categories [7, 8].

A hierarchical clustering method named RACHET (Recursive Agglomeration of Clustering Hierarchies by Encircling Tactic) for analyzing multi-dimensional distributed data. A typical clustering algorithm requires bringing all the data in a centralized warehouse [9, 10]. This results in $O(nd)$ transmission cost, where n is the number of data points and d is the number of dimensions. For large datasets, this is prohibitively expensive. RACHET applies a hierarchical clustering algorithm locally at each site. For each cluster in the hierarchy it maintains a set of descriptive statistics, which form a condensed summary of the data points in the cluster. The local dendrograms along with the descriptive statistics are transmitted to a merging site, which agglomerates them in order to construct the final global dendrogram. Experimental results show that RACHET achieves good quality of clustering compared to a centralized hierarchical clustering algorithm, with minimal communication cost. In contrast, RACHET runs with at most $O(n)$ time, space and communication costs to build a global hierarchy of comparable clustering quality by merging locally generated clustering hierarchies. RACHET employs the encircling tactic in which the merges at each stage are chosen so as to minimize the volume of a covering hyper sphere. For each cluster centroid, RACHET maintains descriptive statistics of constant complexity to enable these choices. RACHET's framework is applicable to a wide class of centroid-based hierarchical clustering algorithms, such as centroid, medoid and Ward [2].

Multi-Agent Systems (MAS) offers an architecture for distributed problem solving. Distributed Data Mining (DDM) algorithms focus on one class of such distributed problem solving tasks-analysis and modeling of distributed data. This paper offers a perspective on DDM algorithms in the context of multiagent systems. It discusses broadly the connection between DDM and MAS. It provides a high-level survey of DDM, then focuses on distributed clustering algorithms and some potential applications in multi-agent-based problem solving scenarios. It reviews algorithms for distributed clustering, including privacy preserving ones. It describes challenges for clustering in sensor network environments, potential shortcomings of the current algorithms and future work accordingly.It also discusses confidentiality (privacy preservation) and presents a new algorithm for privacy-preserving density-based clustering [3].

The problem of combining multiple partitioning of a set of objects into a single consolidated clustering without accessing the features or algorithms that determined these partitioning. It first identifies several application scenarios for the resultant 'knowledge reuse' framework. The cluster ensemble problem is then formalized as a combinatorial optimization problem in terms of shared mutual information. In addition to a direct maximization approach, it proposes three effective and efficient techniques for obtaining high-quality combiners (consensus functions). The first combiner induces a similarity measure from the partitionings and then reclusters the objects. The second combiner is based on hypergraph partitioning. The third one collapses groups of clusters into meta-clusters which then compete for each object to determine the combined clustering. Due to the low computational costs of our techniques, it is quite feasible to use a supra-consensus function that evaluates all three approaches against the objective function and picks the best solution for a given situation. It evaluates the effectiveness of cluster ensembles in three qualitatively different application scenarios: (i) where the original clusters were formed based on non-identical sets of features, (ii) where the original clustering algorithms worked on non-identical sets of objects and (iii) where a common data-set is used and the main purpose of combining multiple clustering's is to improve the quality and robustness of the solution. Promising results are obtained in all three situations for synthetic as well as real data-sets [4].

In Density Based Distributed Clustering(DBDC) each site carries out the DBSCAN algorithm, a compact representation of each local clustering is transmitted to a central site, a global clustering representation is produced from local representations and finally this global representation is sent back to each site. A clustering is represented by first choosing a sample of data points from each cluster. The points are chosen such that: (i) each point has enough neighbors in its neighborhood (determined by fixed thresholds) and (ii) no two points lie in the same neighborhood. ThenK-means clustering is applied to all points in the cluster, using each of the sample points as an initial centroid. The final centroids along with the distance to the furthest point in their K-means cluster form the representation (a collection point, radius pairs).

The DBSCAN algorithm is applied at the central site on the union of the local representative points to form the global clustering. This algorithm requires an _ parameter defining a neighborhood. The authors set this parameter to the maximum of all the representation radii [5].

A number of challenges (often conflicting) arise when developing DDM methods:

- Communication model and complexity,
- Quality of global model and
- Privacy of local data.

It is desirable to develop methods that have low communication complexity, especially in mobile applications such as sensor networks, where communication consumes battery power. Quality of the global model derived from the data should be either equal or comparable to a model derived using a centralized method. Finally, in some situations when local data are sensitive and not easily shared, it is desirable to achieve a certain level of privacy of local data while deriving the global model

Communication between nodes in distributed clustering algorithms can be categorized into three classes (in increasing order of communication cost):

- Communicating models
- Communicating representatives and
- Communicating actual data.

The first case involves calculating local models that are then sent to peers or a central site. Models often are comprised of cluster centroids, e.g., P2P K-means, cluster dendograms, e.g.,RACHET [2]. In the second case, nodes select a number of representative samples of the local data to be sent to a central site for global model generation, such as the case in the DBDC algorithm [5]. The last model of communication is for nodes to exchange actual data objects; i.e., data objects can change sites to facilitate construction of clusters that exist in certain sites only, such as the case in the collaborative clustering scheme [11, 12].

The Dynamic Peer-To-Peer Document Cluster and Cluster Summarization Architecture: DP2PC is a dynamic document clustering architecture for scalable distributed clustering of horizontally partitioned data. Peers in a neighborhood can communicate directly but not with peers in other neighborhoods. Each neighborhood has a super node. Communication between neighborhoods is achieved through their respective super nodes. This model reduces flooding problems usually encountered in large P2P networks.

From the point of view of DP2PC architecture (Figure 4), we are adopting a peer-to-peer framework. Each node in the network has access to part of the whole document collection. The network is a connected graph, in which every node is connected to every other node.

Various peers connected to the network send their documents to the necessary peer. Now after downloading the documents from the neighboring peers, the central peer starts the clustering and summarization process. Central peer has to summarize the document and send it to the corresponding peers. First process is that the key terms are to be extracted. This is achieved by using the k-means algorithm. Next, we have the clustering. Here, we collect all the key terms and then we group them such that we can give a correct description about the summary.

Now the clustered document will be grouped. Next we will match the key terms with the document and finally take out the sentences that were related with the key terms. Now we shall take the sentences and put them according to the clustered group. After the document is summarized we shall save a copy for ourselves and then it is shared to the corresponding peers [13, 14].

Detailed Description: This paper is divided into the following four categories:

- Group creation in peer-to-peer networks
- Communication using Distributed Data Mining
- Document flocking
- Document Summarization using distributed key phrase extraction

Group Creation in Peer-To-Peer Networks: In this module, we present a menu option in the front end for creating a group. Under the Group menu we present three options namely Create, Join and Remove. Creation of a group is mainly done for the efficient communication of peers and also enables the file sharing between the two peers. The peer that creates the group is considered to be the head of the group.

Other peers under the group can create another group also. This process is called hierarchically distributed peer to peer group creation. Automatically, when a person creates a group, he comes under the group. Other peers can join the group by clicking Group->join. Now the peer will be able to select which group he wants to join and join the necessary group.

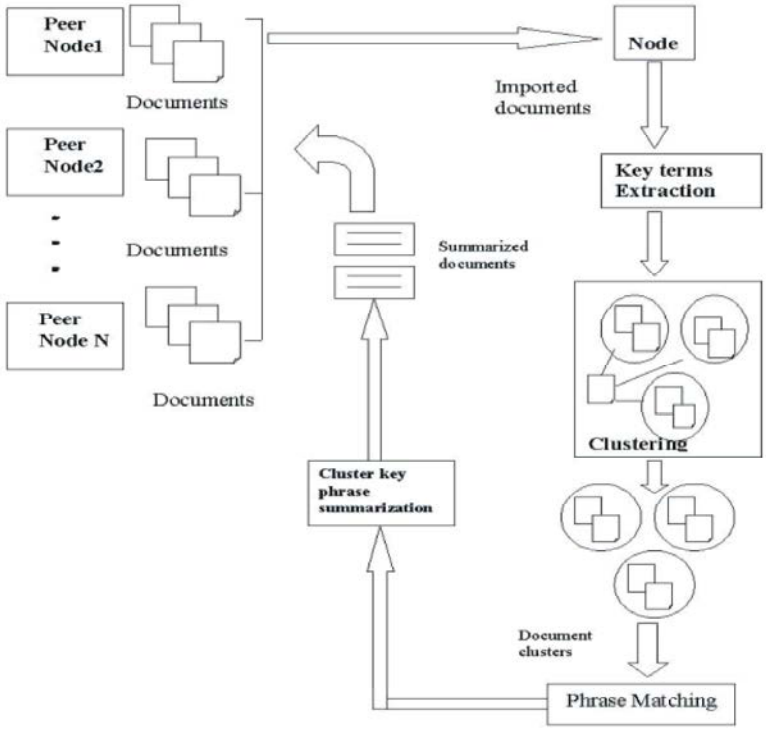

Fig. 4: DP2PC Architecture

When any peer joins the group, then the peer comes under the creator. When any particular peer wants to come out of the group he can do so by clicking group->remove. Now the peer is moved out of the group. Further all the peers connected to the group are informed about the removal of the peer [15].

Communication Using Distributed Data Mining: When the group creation is over, it means that the communication is possible for the peers to communicate with one another. When a message is sent through the Whiteboard chat client window, then the message will be able to reach all the peers connected to the group. Thus we enable a hierarchically distributed peer-to-peer communication.

This module is used to create a communication among peers. This enables the peers to share information. It helps to inform other peers that the peer has shared a particular document. Information is passed to a peer through this module.

Document Flocking: Document flocking refers to document clustering. This process can be done by applying the distributed K-means clustering algorithm. Clustering of document also serves as a searching tool for many documents that are needed. When this module is designed it presents a new means for searching and clustering various distributed documents. Clustering helps to extract main key words from the given document. We split out stop words from the document. Then we identify the key words from the text file. Since we are using only text files for document clustering & summarization, we need to remember that retrieval of information from them are a bit tough. Then using ontology, we can extract the necessary information

Document Summarization Using Distributed Key Phrase Extraction: Document summarization is the process of making a brief explanation regarding the document. This is done using the distributed key phrase extraction algorithm. Summarization of the document helps us to understand the document very easily such that we have a good idea about the contents and the key terms present in the document. This is done by collecting the keywords of any document by clustering process. Then after the keywords are being collected, the phrase with which the keywords are associated is extracted form the document. This process comes under the key-phrase extraction algorithm. Automatic keyphrase extraction from document clusters provides a very compact summary of the

contents of the clusters, which often helps in locating information easily. We have introduced an algorithm for topic discovery using keyphrase extraction from multi-document sets or clusters based on frequent and significant shared phrases between documents. Each candidate keyphrase l is assigned the following features:

Document Frequency: The number of documents in which the phrase appeared, normalized by the total number of documents.

Average Weight: The average weight of the phrase over all documents.

Average Phrase Frequency: The average number of times this phrase has appeared in one document, normalized by the length of the document in words:

Average Phrase Depth: The location of the first occurrence of the phrase in the document.

Those features will be used to rank the candidate phrases.

CONCLUSION AND FUTURE WORK

This paper proposed a novel architecture and algorithm for Dynamic document cluster and cluster summarization model, which allows building hierarchical networks for clustering data. It demonstrates the flexibility of the model, showing that it achieves comparable quality to its centralized counterpart while providing significant speedup and that it is possible to make it equivalent to traditional distributed clustering models by manipulating the neighborhood size and height parameters.

The model shows good scalability with respect to network size and hierarchy height, degrading the distributed clustering quality significantly. 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. It also allows privacy within neighborhood boundaries (no data shared between neighborhoods). In addition, it provides interpretation capability for document clustering through document cluster summarization using distributed key phrase extraction. This project has been further planned to extend merging and splitting of complete hierarchies.

REFERENCES

1. Khaled, M. Hammouda and Mohamed S. Kamel, 2009. Fellow, IEEE "Hierarchically distributed peer-to-peer Document Clusterin and Cluster Summarization, 21(5): 681-698.
2. Samatova, N.F., G. Ostrouchov, A. Geist and A.V. Melechko, 2002. "RACHET: An Efficient Cover-Based Merging of Clustering Hierarchies from Distributed Datasets," Distributed and Parallel Databases, 11(2): 157-180.
3. Da Silva, J., C. Giannella, R. Bhargava, H. Kargupta and M. Klusch, 2005. "Distributed Data Mining and Agents," Eng. Applications of Artificial Intelligence, 18(7): 791-807.
4. Strehl, A. and J. Ghosh, 2002. "Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, 3: 583-617.
5. Pfeifle, M., H.P. Kriegel and E. Januzaj, 2004. "DBDC: Density Based Distributed Clustering," Proc. Ninth Int'l Conf. Extending Database Technology (EDBT '04), pp: 88-105.
6. Udayakumar, R., V. Khanna, T. Saravanan and G. Saritha, 2013. Cross Layer Optimization For Wireless Network (Wimax), Middle-East Journal of Scientific Res., ISSN:1990-9233, 16(12): 1786-1789.
7. Saravanan, T. and R. Udayakumar, 2013. Optimization of Machining Hybrid Metal matrix Composites using desirability analysis, Middle-East Journal of Scientific Research, ISSN:1990-9233, 15(12): 1691-1697.
8. Datta, S., C. Giannella and H. Kargupta, 2006. "K Means clustering over a Large, Dynamic Network," Proc. Sixth SIAM Int'l Conf. Data Mining (SDM '06), pp: 153-164.
9. Datta, S., K. Bhaduri, C. Giannella, R. Wolff and H. Kargupta,"Distributed Data Mining in Peer-to-Peer Networks," IEEE Internet Computing, 10(4): 18-26.
10. Hammouda, K. and M. Kamel, 2005. " Corephrase: Keyphrase Extraction for Document Clustering," Proc. IAPR Int'l Conf. Machine Learning and Data Mining in Pattern Recognition (MLDM '05), P. Perner and A. Imiya, eds., pp: 265-274.
11. http://www.springerlink.com.
12. Udayakumar, R., V. Khanaa and K.P. Kaliyamurthie, 2013. Optical Ring Architecture Performance Evaluation using ordinary receiver, Indian Journal of Science and Technology, ISSN: 0974-6846, 6(6): 4742-4747.

13. Udayakumar, R., V. Khanna, T. Saravanan and G. Saritha, 2013. Retinal Image Analysis Using Curvelet Transform and Multistructure Elements Morphology by Reconstruction, Middle-East Journal of Scientific Research, ISSN:1990-9233, 16(12): 1798-1800.
14. Saravanan, T., V. Srinivasan and R. Udayakumar, 2013. Images segmentation via Gradient watershed hierarchies and Fast region merging, Middle-East Journal of Scientific Research, ISSN:1990-9233, 15(12): 1680-1683.
15. Thooyamani, K.P., V. Khanaa and R. Udayakumar, 2013. Application of Soft Computing Techniques in weather forecasting: Ann Approach, Middle-East Journal of Scientific Research, ISSN:1990-9233, 15(12): 1845-1850.