International Journal of Advancements in Computer Networks and Its Security– IJCNS
Volume 4 : Issue 4 [ISSN 2250-3757]
Publication Date : 27 December,2014

Distributed Data Clustering in Peer-to-Peer Networks: A Technical Review

[Rasool Azimi, Hedieh Sajedi]

***Abstract*—Clustering as one of the main branches of data mining, has gained an important place in the different applied fields. On the other hand, Peer-to-Peer (P2P) networks with features such as simplicity, low cost communication, and high availability resources, have gained a worldwide popularity in the present days. In P2P network, high volumes of data are distributed between dispersed data sources. In such distributed systems, sending data to a central site is not possible, due to processing, storage and handling costs. This issue makes the need for specific methods for distributed data clustering in P2P networks. This paper makes a technical review of distributed data clustering approaches in P2P network, to understand the research trends in this area.**

***Keywords*—Peer to Peer Networks, Distributed Data Mining, Distributed Data Clustering**

I. Introduction

Identify clusters, is an important factor in the analysis of large datasets. Generally, for extracting data, eliminating duplicate data, and making usable these data, several techniques have been proposed as data mining methods. As a result, data mining has emerged as an important area of research. Distributed computing environments such as P2P networks, have separated and diffuse data sources. Due to the large volume of computing and communications and network bandwidth limitations, privacy reasons, or because of the huge amount of distributed data, it's essential that the processing of data be performed using a distributed approach, without aggregate data to a centralized location [1]. Furthermore, the proposed clustering algorithms for P2P networks are encountered with the challenges such as the dynamic of P2P networks, scalability and fault tolerance in these distributed environments. In this paper, we present the technical review of the most important distributed clustering algorithms in P2P networks. The rest of the paper is organized as follows, In Section II, we describe DDM approaches. In Section III, we discussed distributed clustering methods. In Section IV, we take a look at P2P networks. In Section V, we investigate the most important works of distributed clustering in P2P networks, due to the type of P2P networks. In section VI, we briefly compare the approaches reviewed in this paper, given - the challenges of distributed data clustering in P2P networks. Finally, In Section VII, we present conclusions and directions for future work.

Rasool Azimi
Islamic Azad University, Science and Research, Qazvin Branch
Iran

Hedieh Sajedi
University of Tehran
Iran

II. Distributed Data Mining

In a P2P network, the data have been placed on the peers in a distributed way and analysis and supervision of this distributed data sources, requires data mining technology, which is designed for distributed applications, and it's called Distributed Data Mining (DDM) [2]. As stated in TABLE I, overall, DDM operations can be performed in four approaches [6].

TABLE I.

No.	DDM Approach
1	Bringing the data to a central site, then apply centralized data mining on the collected data
2	Performing local mining at each site to produce a local model and transmit them into a central site in order to form a global model
3	Selecting a set of representative data objects and transmit them to a central site, to combine them and performing data mining on the global representative dataset
4	Interactive performing data mining operation by local sites, without the help of a central site

Types of DDM approaches

The fourth approach, unlike the previous three approaches, does not include a central site for facilitation of data mining operation. Thus, this approach is belongs to P2P networks. It's a fully distributed data mining operation, through the interactions between peers in order to produce a general model, which represents the mining result of the entire dataset, can be applied.

III. Distributed Data Clustering

Distributed data clustering, aims to extract potentially useful information from large datasets by grouping similar data, and separating dissimilar data according to some criteria of dissimilarity between data items. In a distributed environment, it needs to be done when the data cannot be concentrated on a single site, for example, for reasons of security concerns or due to bandwidth limitations or due to high volumes of distributed data [1].

A. *Data Clustering approaches*

As stated in TABLE II, generally, clustering algorithms can be classified into six categories [12]. In [5], provides an overview of clustering algorithms. Nevertheless, most works in

the area of distributed data clustering in P2P networks, apply the first two clustering methods, which are reviewed in the following.

TABLE II.

No.	Clustering Method	Summary of Approach
1	Density-based clustering	Based on connectivity and density functions
2	Partitioning-Based clustering	Construct random partitions and then iteratively refine them by some criterion
3	Hierarchical clustering	Create a hierarchical decomposition of the set of data (or objects) using some criterion
4	Grid-based clustering	Based on a multiple-level granularity structure
5	Model-based clustering	A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
6	Constraint-based clustering	Clustering by considering user-specified or application-specific constraints

Types of clustering approaches

1) *Density-based clustering method*

Density-based clustering method has been developed based on the concept of density. In this method, cluster growth, will continue as long as the density in the "neighborhood" exceeds some threshold; that is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise and discover clusters of arbitrary shape [12].

2) *Partition-based clustering method*

A Partition based clustering method, given a dataset of n data tuples, makes k partitions of data where each partition represents a cluster ($n>k$). That is, the clustering of data into k groups which together satisfy the following requirements: 1- Each group must have at least one object. 2- Each object must belong to exactly one group [12].

B. *Distributed Data Clustering Applications*

In general, cluster analysis plays an important role in almost every area of science and engineering, including bioinformatics, market research, privacy and security, image analysis, web search, health care and many others. Due to the rapid increase in the number of autonomous data sources, there is a growing need for effective approaches to distributed data clustering. Nowadays, distributed Data Clustering has wide-ranging applications such as in P2P file sharing systems, mobile ad-hoc networks, P2P sensor networks and so forth.

IV. P2P NETWORKS

In recent years, highly decentralized networks such as P2P have emerged as the most interesting, challenging and innovation rich areas in computer networking. The nature of these networks requires efficient, local, scalable and self-stabilizing algorithmic solutions [3].

A. *Types of P2P Networks*

P2P networks, based on how peers relate to each other, and how their content is shared, are classified into Unstructured and Structured networks [9]. In Unstructured P2P networks, each peer in the logging time, randomly chooses a set of other peers as its neighbors, and connected to them. When leaving the network, it's enough that the peer hang up this connections. In Structured P2P networks, there is a regular topological structure. Usually the regular topology of a distributed hash table (DHT) has been used to build and manage the network.

B. *P2P Data Mining*

P2P Data Mining (P2PDM) is a special type of DDM where there is no notion of centralization to the mining process and all peers are regarded as peers. P2PDM scenarios typically exist when no single entity owns or has jurisdiction over the entire dataset. In those cases, the data are naturally distributed over a large number of peers, and the goal is to find a clustering solution that takes into account the entire dataset [8].

V. DISTRIBUTED DATA CLUSTERING ALGORITHMS IN P2P NETWORKS

In this section, we review the most important works in the area of distributed data clustering in P2P networks. In the first section, reviews the approaches provided in the area of distributed data clustering in Structured P2P networks, and then in the next section we review the approaches provided in the area of distributed data clustering in Unstructured P2P networks.

A. *Distributed Data Clustering Algorithms in Structured P2P Networks*

In this section, we review related works on distributed data clustering in Structured P2P networks: In [13], the PENS algorithm is introduced to cluster data stored in P2P networks. In order to clustering the whole data entities stored in P2P networks, each peer performs clustering operation, on the data entities that are mapped in its region on the CAN overlay by using the DBSCAN algorithm [7]. This algorithm uses the Hierarchical Cluster Assembly (HCA) to aggregate the clusters in smaller regions, to form clusters in larger regions, and then these clusters aggregate in the same way to form clusters in larger regions. The main shortcoming of this method is that, since it's based on the CAN overlay, mapping data may cause to create communication overhead in large datasets. A brief review of the PENS algorithm is presented in TABLE III.

TABLE III.

Objective	Data clustering in P2P Networks with a hierarchical cluster assembly
Strengths	- Deal with P2P system and investigate density-based clustering;
Weaknesses	- Lack of scalability in large datasets with a high degree of distribution; - Need for global synchronization;

Brief review of the PENS algorithm [13]

In [1], a method is introduced for clustering multidimensional data in Structured P2P networks, based on Density. In short, we call this algorithm MDP2P-Scan. First, each peer calculates the local density and then the maximum of the peer region is determined. This maximum includes the mean local density and its location in the form of the geometric center of the region. After that, in each peer, the maximums of all of the regions and their locations are used to associate each local data to a maximum. This association is determined on the basis the ratio between the value of each maximum and its distance from the points in the whole region is maximized and, in this way, the clusters take shape by mapping each object on a maximum. In this paper, a method is introduced for clustering data in multidimensional P2P networks that has a routing mechanism, can partition the data space, and can search, without the need for a re-management of the network and without changing or endangering the fundamental services of P2P systems. Since this algorithm is based on the overlaying of the CAN, MURK-CAN, and MURK-SF mapping data may cause to create communication overhead in large datasets. A brief review of this work is shown in TABLE IV.

TABLE IV.

Objective	Distributed Data Clustering in Multi-Dimensional P2P Network
Strengths	- No need for a specific reorganization of the network; - Providing an accuracy guarantee; - No change in the routing mechanism, the data space partition among peers and the search capabilities;
Weaknesses	- Lack of scalability in large datasets with a high degree of distribution; - Need for global synchronization;

Brief review of the MDP2P-Scan algorithm [1]

In [8], a modular, flexible, and scalable hierarchical distributed approach, called HP2PC, is suggested for P2P networks, which provided a novel architecture and clustering method for these networks. The distributed clustering strategy in a neighborhood is like parallel K-Means algorithm provided in [11] and then the resulted centroids by main peers will transfer to the higher level. A main peer, cooperate as a higher-level neighborhood with its peers, to produce centroid collection for its neighborhood. This process continues hierarchically until the production of a centroid collection in the root of hierarchy. The algorithm will be finished when there's no more change in the entities assignment. The disadvantage of this method is that clustering quality decreases noticeably for each aggregation level, because of the random grouping of peers at each level. Therefore, quality decreases significantly for large networks. A brief review of HP2PC algorithm is shown in TABLE V.

TABLE V.

Objective	A hierarchical distributed approach for P2P Network
Strengths	- No global synchronization required; - Scalability; - Solving Message Flooding Problem in P2P Network;
Weaknesses	- Not considering the dynamics of P2P network; - Significantly Reducing clustering quality in large networks;

Brief review of the HP2PC algorithm [8]

B. Distributed Data Clustering Algorithms in Unstructured P2P Networks

In this section, we review related works on distributed data clustering in Unstructured P2P networks: In [16], a K-Means algorithm for distributed data clustering on large-scale dynamic networks, called P2P K-Means, is represented. The algorithm begins with a peer that randomly produces primary centroids and sends them with the termination threshold to all its immediate neighbors and initiates the first iteration. When a peer receives the primary centroids and the termination threshold, it sends them to other its neighbors, and starts the first iteration. Afterward, in each iteration, each peer receives from its neighboring peers, centroids, and the number of clusters of each iteration. Then, using this information and iterating them with local data, this peer produces centroids for the next iteration. If the distance between new centroids and centroids of the previous iteration is more than the termination threshold, the next iteration will begin. Otherwise, the peer will enter the termination phase. This algorithm is communication efficient and robust to network or data changes. In this approach, the achieved accuracy is relatively high, but in contrast, the communication cost is also high. A brief review of P2P K-Means algorithm is shown in Table VI.

TABLE VI.

Objective	Distributed data clustering on large-scale dynamic P2P Network
Strengths	- No global synchronization required; - Being communication-efficient; - Being robust to network or data changes;
Weaknesses	- High cost of communication;

Brief review of the P2P K-Means algorithm [16]

In [10], a partition-based distributed clustering algorithm called LSP2P K-Means has been proposed. In summary, the LSP2P K-Means algorithm is the frequent iterations of improved K-Means algorithm in every peer N_1. Each peer N_i

receives, in each iteration, centroids and the number of clusters related to iteration *l* from neighboring peers. This peer, then produces centroids for the next iteration using this information and its integration with local data. If the distance between centroids of the new clusters and centroids of the previous iteration is greater than the threshold, the next iteration is started. Otherwise, that peer will reach the termination state. Based on the results of the experiment, this algorithm manifested excellent scalability, yet the performance of the algorithm was not affected by network size. In discussing the communication, it's shown the communication complexity in general increases linearly with the network size. The main problem with this algorithm is that, it's not possible to guarantee the analytical accuracy of the algorithm. A brief review of this algorithm is presented in TABLE VII.

TABLE VII.

Objective	A distributed K-means clustering algorithms based on local synchronization of P2P Network
Strengths	- No global synchronization required; - Considering the dynamics of P2P network; - Able to deal with topology changes and loss of data;
Weaknesses	- Clustering accuracy not be guaranteed; - Not robust to outliers;

Brief review of the LSP2P K-Means algorithm [10]

In [15], a fully decentralized density-based clustering algorithm, called GoSCAN, is represented which is capable of clustering dynamic and distributed datasets without requiring central control or message flooding. This method is based on DBSCAN Algorithm [7] and actually compatible with the dynamic nature of Unstructured P2P networks. GoSCAN algorithm can be broken into two major tasks: 1- identifying core points, and 2- forming the actual clusters. These two operations are executed in parallel employing gossip-based communication. Design of gossip-based communication methods, allowed the parallel execution of the two tasks, which gradually increased the algorithm accuracy. GoSCAN allowed each peer to find an individual trade-off between quality of discovered clusters and transmission costs. Experimental results have shown that GoSCAN allows effective clustering with efficient transmission costs. The weakness of this method is that, the networks with fewer peers appear to have higher accuracy compared to networks with larger number of peers. A brief review of the GoSCAN algorithm is presented in TABLE VIII.

TABLE VIII.

Objective	A fully decentralized density-based clustering algorithm for P2P networks
Strengths	- No need to Global synchronization; - Compliance with dynamic datasets; - Able to clustering dynamic and distributed datasets without requiring central control or message flooding;
Weaknesses	- Only applicable for small-scale P2P networks; - The impossibility of fault-tolerant;

Brief review of the GoSCAN algorithm [15]

In [14], a fully distributed K-Means algorithm (Epidemic K-Means) has been proposed to solve distributed randomly K-Means issue. It's the first data mining approach, which is suitable for P2P overlay networks. At any time, each peer may be one of two states, *ACTIVE* and *CONVERGED*. When active, the corresponding node performs two phases at each K-Means iteration. In the first phase, the calculation is performed on the dataset, such as Sequential K-Means. For any given point x, the nearest cluster's centroid is calculated and local cluster partitions are determined. In the second phase, the overall datasets are calculated by using the gossip-based aggregation protocols. Iteration of K-Means algorithm is continued until the size the total error is less than the threshold value previously set**.** Below this threshold, the peer state has changed to *CONVERGED* state. The statistical guarantee of gossip-based protocol ensures that results within a bounded approximation error at each peer of the system are consistent with static and faultless networks. However, the performance of this approach decreases when losing messages and failure peers in the asynchronous networks. A brief review of the Epidemic K-Means algorithm is presented in TABLE IX.

TABLE IX.

Objective	A decentralized formulation of the distributed K-Means clustering algorithm for P2P networks
Strengths	- No global synchronization required; - Accurate, scalability and fault tolerant under unreliable network conditions;
Weaknesses	- Not considering the dynamics of P2P network; - Reduce the efficiency of the algorithm when losing messages and failure peers in the asynchronous networks;

Brief review of the Epidemic K-Means algorithm [14]

VI. DISCUSSION

Given the challenges ahead in the area of distributed data clustering in P2P networks, Challenges such as dynamics of P2P networks, scalability, synchronization, guarantee the accuracy of clustering and fault tolerant, as stated in TABLE X, we compare the algorithms reviewed in the previous section with each other.

A. *Dynamics of P2P Networks*

The dynamics of P2P networks is of paramount importance, as the peers may join or leave the network at any time. Among the algorithms were reviewed, two algorithms, P2P K-Means [16] and LSP2P K-Means [10] are considered the dynamics of these networks.

B. *Scalability*

The proposed algorithms for clustering in P2P networks should be scalable in terms of handling varying data size or varying number of peers. Among the algorithms were reviewed, P2P K-Means [16], LSP2P K-Means [10], HP2PC [8] GoSCAN [15] and Epidemic K-Means [14] are capable of providing scalability.

C. *Synchronization*

In P2P networks any attempt to synchronize between the entire networks is likely to fail due to connection latency, or limited bandwidth in case of sensor networks. Thus, any algorithm developed for P2P network should not take the route of global synchronization [4]. However, some of the algorithms we reviewed in Section 5, such as PENS [13] and MDP2P-Scan [1] need for global synchronization.

D. *Guarantee the Accuracy of Clustering*

Algorithms such MDP2P-Scan [1], HP2PC [8], Epidemic K-Means [14], GoSCAN [15] guarantee the accuracy of clustering. Under moderately or highly skewed cluster distributions, these approaches are expected to fail in approximating the ideal solution and in guaranteeing consistency over the network peers.

E. *Fault Tolerant*

Among the methods that were reviewed, only Epidemic K-Means algorithm [14] is capable of fault tolerance. In fact, this algorithm is accurate and fault tolerant under unreliable network conditions and is suitable for asynchronous networks of very large and extreme scale.

TABLE X.

References	Algorithm Name	DDM Approach No.	Clustering Approach No.	Dynamics of P2P Networks		Scalability		Synchronization		Accuracy of Clustering		Fault Tolerant	
				Yes	*No*	*Yes*	*No*	*Global*	*Local*	*Yes*	*No*	*Yes*	*No*
[13]	PENS	4	1		☒		☒	☒			☒		☒
[1]	MDP2P-Scan	4	1		☒		☒	☒		☒			☒
[16]	P2P K-Means	4	2	☒		☒			☒		☒		☒
[10]	LSP2P K-Means	4	2	☒		☒			☒		☒		☒
[8]	HP2PC	4	2		☒	☒			☒	☒			☒
[15]	GoSCAN	4	1	☒		☒			☒	☒			☒
[14]	Epidemic K-Means	4	2		☒	☒			☒	☒		☒	

Comparison of distributed data clustering algorithms

VII. CONCLUSIONS

Due to the decentralized and peer-relying nature of P2P networks and the lack of central authority in P2P networks, all the works mentioned in this paper, use the fourth approach of DDM for interactive performing data mining operation, without the need for a central server. In terms of several main aspects, these works can be compared with each other. The first aspect is related to the type of clustering algorithms, which can be divided into two types based on density and partitioning. The second aspect is related to the type of P2P network. Some works have been employed for Structured P2P networks and some others for Unstructured P2P networks. Other high-priority issues include the scalability properties of the algorithms, considering the dynamic nature of P2P networks and challenge in requiring global or local synchronization between peers and finally, to guarantee the clustering accuracy. With regard to the issues raised, provide a fully distributed and scalable approach to data clustering in P2P networks is considered for future approach.

References

[1] S. Lodi, G. Moro, and C. Sartori, "Distributed Data Clustering in Multi-Dimensional Peer-to-Peer Networks," Australian Computer Society, Vol. 32, no 3, pp. 171- 178, 2010.

[2] B.H. Park, and H. Kargupta, "Distributed Data Mining: Algorithms, Systems, and Applications," Data Mining Handbook, 2002.

[3] M. Onus, "Overlay Network Construction in Highly Decentralized Networks," Ph.D. dissertation, Arizona State. Univ, United State, 2009.

[4] S. Datta, "Probabilistic Approximate Algorithms for Distributed Data Mining in Peer-to-Peer Networks," Ph.D. dissertation, Maryland. Univ, United State, 2008.

[5] A.k. Jain, M.N. Murty, P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, Vol. 31(3), pp. 265-323, 1999.

[6] K.M. Hammouda, and M.S. Kamel, "Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization," IEEE Transactions on Knowledge and Data Engineering, Vol. 21, no. 5, pp. 681-698, 2009.

[7] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A Densitybased Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In Proceedings of Knowledge Discovery in Database (KDD), pp. 226–231, 1996.

[8] K.M. Hammouda, and M.S. Kamel, "Models of distributed data clustering in peer-to-peer environments" Knowl Inf Syst, vol. 38, iss. 2, pp. 303–329, 2014.

[9] J. Buford, H. Yu, and E. K. Lua. "P2P Networking and Applications," Morgan Kaufmann, 2008.

[10] S. Datta, C. Giannella, and H. Kargupta, "Approximate distributed k-means clustering over a peer-to-peer network", IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 10, pp. 1372–1388, 2009.

[11] I.S. Dhillon, and D.S. Modha, "A Data-Clustering Algorithm on Distributed Memory Multiprocessors," Large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems, SIGKDD, pp. 245-260, 2000.

[12] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques," 2nd ed, Morgan Kaufmann Publishers, 2006.

[13] M. Li, G. Lee, W.C. Lee, and A. Sivasubramaniam, "PENS: An algorithm for Density-Based Clustering in Peer-to-Peer Systems," Proceedings of the 1st international conference on Scalable information systems, pp. 39, 2006.

[14] G. Di Fatta, F. Blasa, S. Cafiero, and G. Fortino, "Fault tolerant decentralised K-Means clustering for asynchronous large-scale networks," Journal of Parallel and Distributed Computing, vol. 73, no. 3. pp. 317-329, 2013.

[15] H. Mashayekhi, J. Habibi, S. Voulgaris, and M. van Steen, "GoSCAN: Decentralized scalable data clustering." Computing, Vol 95(9), pp. 759–784, 2013.

[16] S. Datta, C. Giannella, H. Kargupta, "K–Means Clustering over a Large, Dynamic Network," Proc. SIAM Int'l Conf. Data Mining, pp. 153-164, 2006.