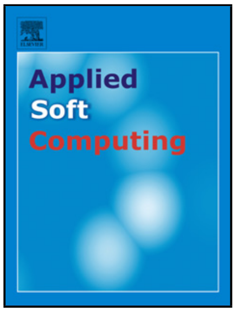

Title: A Distributed Data Clustering Algorithm in P2P Networks

Author: Rasool Azimia Hedieh Sajedib Mohadeseh Ghayekhlooa

PII: S1568-4946(16)30614-7
DOI: http://dx.doi.org/doi:10.1016/j.asoc.2016.11.045
Reference: ASOC 3935

To appear in: *Applied Soft Computing*

Received date: 27-1-2015
Revised date: 14-11-2016
Accepted date: 28-11-2016

A Distributed Data Clustering Algorithm in P2P Networks

Rasool Azimi[a], Hedieh Sajedi[b,*], Mohadeseh Ghayekhloo[a]

[a] *Young Researchers and Elite Club, Qazvin Branch, Islamic Azad University, Qazvin, Iran*
[b] *Department of Computer Science, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran*

[*] Corresponding author at: Department of Computer Science, College of Science, University of Tehran, Tehran, Iran. Tel.:+982161112915

E-mail addresses: hhsajedi@ut.ac.ir (Hedieh Sajedi), r.azimi@qiau.ac.ir (Rasool Azimi), m.ghayekhloo@qiau.ac.ir (Mohadeseh Ghayekhloo)

ABSTRACT

Clustering is one of the important data mining issues, especially for large and distributed data analysis. Distributed computing environments such as Peer-to-Peer (P2P) networks involve separated/scattered data sources, distributed among the peers. According to unpredictable growth and dynamic nature of P2P networks, data of peers are constantly changing. Due to the high volume of computing and communications and privacy concerns, processing of these types of data should be applied in a distributed way and without central management. Today, most applications of P2P systems focus on unstructured P2P systems. In unstructured P2P networks, spreading gossip is a simple and efficient method of communication, which can adapt to dynamic conditions in these networks. Recently, some algorithms with different pros and cons have been proposed for data clustering in P2P networks. In this paper, by combining a novel method for extracting the representative data, a gossip-based protocol and a new centralized clustering method, a Gossip Based Distributed Clustering algorithm for P2P networks called GBDC-P2P is proposed. The GBDC-P2P algorithm is suitable for data clustering in unstructured P2P networks and it adapts to the dynamic conditions of these networks. In the GBDC-P2P algorithm, peers perform data clustering operation with a distributed approach only through communications with their neighbours. The GBDC-P2P does not need to rely on a central server and it performs asynchronously. Evaluation results demonstrate the superior performance of the GBDC-P2P algorithm. Also, a comparative analysis with other well-established methods illustrates the efficiency of the proposed method.

Keywords: Distributed data mining, Data clustering, Gossiping, Overlay, Peer-to-Peer network.

1. Introduction and Related Works

Peer-to-Peer (P2P) computing or networking is a distributed application architecture that is used as a common method for the applications involving data exchange between distributed resources. In such applications, large volumes of data are distributed across several data sources. Due to privacy concerns, bandwidth limits and the large amount of data, it is difficult to collect peer's data on a central server and perform data mining over the whole data, centrally. In fact, processing of this distributed data should be done in a distributed way without concentrating the whole data in a central place [1], [8].

In P2P networks, we are facing the challenges such as dynamics of the network, complex structure of the network and distributed data. The data at each peer is constantly changing and some peers may not be present in the network all the time. The new peers can join the network while old peers leave at any time [19].

Generally, several techniques known as data mining methods have been proposed for extracting data, finding unsuspected relationships between data, and transforming it to an understandable structure for further use. Traditional data mining algorithms assume that all data are concentrated in a location. Nowadays, with the increasing development of communication systems, there is a high volume of data to be distributed and this volume of data is increasing gradually [3]. On P2P networks, the data have been placed on the peers in a distributed way. Analysis of this distributed data sources needs data mining technology, which is designed for distributed applications, called Distributed Data Mining (DDM) [2].

[*] Corresponding author at: Department of Computer Science, College of Science, University of Tehran, Tehran, Iran. Tel.:+982161112915

E-mail addresses: hhsajedi@ut.ac.ir (Hedieh Sajedi), r.azimi@qiau.ac.ir (Rasool Azimi), m.ghayekhloo@qiau.ac.ir (Mohadeseh Ghayekhloo)