About this proposal
This doctorate research proposal document describes the working title of the research proposal and general overview of the area. The research plan mentioned in this document may be modified based on the approval on this documented research proposal.
## CONTENTS

1. RESEARCH TITLE .......................................................... 1

2. ABSTRACT .................................................................... 1

3. LITERATURE SURVEY .................................................. 2

4. PROBLEM DEFINITION ................................................ 4

5. PROPOSED METHODOLOGY ........................................ 4

6. APPLICATIONS ........................................................... 5

7. REFERENCES .............................................................. 6
1. **Research Proposal Title**

Distributed Data Mining Using Mobile Agents

2. **Abstract**

Many data mining applications assume that data is stored in data warehouse. The data warehouse contains the enterprise wide data which is historic in nature and integrated from heterogeneous data sources. This traditional approach may not work efficiently, if data is present at geographically separated locations as in case of multinational chain of retail stores Wal-Mart, where huge transaction databases are available at many locations. Sometimes, transmitting large amounts of data to a data centre is expensive and even impractical. To overcome this problem, many distributed and parallel data mining solutions are developed. The data mining algorithm traditionally assumes that data is stored centrally; it resides in memory and is static in nature. However, this assumption is no longer valid with the development of there are many data sources have come online such as weather and climate data, stream data such as financial data of stock exchanges etc. The nature of this data is dynamic and the data changes with high speed. This data is stored at multiple locations and it becomes increasingly expensive to bring it at one central location.

Hence to overcome these problems, distributed data mining is thought to be more efficient approach. Recently due to higher interest in business intelligence, the research in distributed data mining has picked up pace. In case of distributed data mining we can use to approaches such as either data or computation is thought to be distributed. In varying situations, the distributed data mining faces different challenges of resource constraints such as limited computation power or limited bandwidth, privacy and communication overhead. Since distributed data mining covers application requiring specific algorithms, it is difficult to provide a unified framework of algorithms.

Multi agent system consists of mobile agents that deal with complex application involving distributed approach. The collective and individual behaviour of these mobile agents depend on observed data from these distributed data sources. The distributed data mining faces the problem of analyzing distributed data from distributed sources and gives solution for many different analysis and mining operations. Since multi agent systems are also distributed systems, combining DDM with MAS for data intensive applications is possible. The mobile agent can collaborate with other agents to share their learning in order to improve the overall efficiency of data mining.
3. Literature Review and related work

3.1 Distributed Association Rule Mining

There are many distributed association rule mining solution proposed which are using multi agent approach. In [7] a framework is proposed which consists of Central site and n number of distributed sites consisting of horizontally partitioned transaction data set. Central site acts as launching station for mobile agents. Mobile as well as stationary agents are stored at agent pool at this site. Central security agency at this site assigns legal certificate to every mobile agent which is verified at host for authenticity. There are five agents proposed in this architecture, of which three are mobile agents while two agents are stationary intelligent agents. Mobile agents are Local frequent itemsets generator Agent, Local knowledge generator agent, Total frequent itemsets collector agent, Global frequent itemsets generator agent, and Global knowledge generator agent. The framework proposed is being implemented and results are being validated.

3.2 Multi Agent Data mining framework for Clustering

In [6] the Multi Agent Data Mining (MADM) framework is described for multi agent based clustering. The framework describes two approaches, K-means and KNN. The MADM framework comprises of four kinds of agents. User agents are interfaces between end users and the MADM environment. The agents are responsible for obtaining the input from user, spawning cluster agents in order to perform the clustering task and presenting the derived clustering result. Data agents are the “owners” of data sources. There is one to one relationship between data agents and data sources. Data agents are used by clustering agents to access data. Clustering agents are “owners” of clusters. Group of clustering agent can be thought of as representing clustering algorithm. A number of clustering agents are spawned, as required, by a user agent in order to perform some clustering task. Clustering agent has two principal functions (i) initial “start” cluster configuration (ii) cluster refinement. Validation agents are a special type of agent that performs validation operations on clustering results. Each validation agent is owner of a technique for measuring the “goodness” of a given cluster configuration. The MADM framework to achieve multi-agent based clustering can be extended to include other clustering algorithm in future. The particular feature of framework is that it enables agents to negotiate so as to improve on an initial clustering.
3.3 Multi Agent System for Information Retrieval

In [2] a multi-agent framework for retrieving information in real time from Micro blogs during unexpected event is proposed. MASIR is composed of six kinds of agents. The Stream Retrieval Agent (SRA) remains connected to twitter during an event and search in real time. The SRA is responsible for extracting identity of the user sharing information on the topic related to event. It sends the identity of such users to historic listener agent manager. The Historic Listener Agent Manager (HLAM) receives the list of users sent by SRA and assigns each user a Historic Listener Agent. The HLA extracts historic information related to user and saves the information in historic and social information base. Historic Listener Agent (HLA) also extracts social information such as list of followers and followees to the information base. Prominent User Detector (PUD) detects most prominent user by calculating and updating the prominence score of each user. Stream Listener Agent Generator (SLAG) manages the tracking process of most prominent users during event. The Stream Listener Agent (SLA) keeps listening to users profile and detects new updates in real time.

3.4 Multi-agent System for Intelligent retrieval and processing of Information

In [1] a solution for determining the polarization of text document based on classifying feature vectors is proposed. A multi-agent sentiment analysis algorithm based on classifying feature vectors consist of typical stages associated with document classification process such as preprocessing, feature extraction, and classification of extracted vectors. Large number of analyzed documents and their various sources makes it a good case for applying multi-agent approach to this problem.

A Multi-agent sentiment analysis algorithm based on lexicon of words is also proposed in this work. Lexicon contains list of words with appropriate polarization value. Every word in lexicon is labeled as positive negative or neutral more complex lexicon assigns numerical values to every word. Fantastic could be assigned as 1.0 or 9.0 while boring could be equal to -1.0 to -9.0. The proposed solution consists of manager agent, preprocessing agent, classifier agent. The proposed multi agent information retrieval approach can be applied to distributed environment with huge text repositories.

3.5 Java Agent DEvelopment Framework

JADE (Java Agent DEvelopment Framework) is a software framework to develop agent-based applications in compliance with the FIPA specifications for interoperable intelligent multi-agent systems. The goal is to simplify the development while ensuring standard compliance through a comprehensive
set of system services and agents. Besides the agent abstraction, JADE provides a simple yet powerful task execution and composition model, peer to peer agent communication based on the asynchronous message passing paradigm, a yellow pages service supporting publish subscribe discovery mechanism and many other advanced features that facilitates the development of a distributed system. The JADE architecture consists of notion of agent, container, platform, Main container, AMS and DF.

4. Problem Definition

Due to advances in communication and computing technology there are multiple sources available at many compute nodes. The distributed data mining approach provides more scalable and practical approach for mining as it involves large number of data sites. Focusing on distributed data mining, this research aims to apply mobile agent paradigm to distributed data mining to analyze and retrieve underlying hidden patterns.

5. Methodology

The traditional data mining system assumes that the data is located at central source. Normally this source of data is an enterprise data warehouse. In the agent based distributed data mining, there is an agent locally present on each site. The agent carries out the data mining task locally. The locally learned model is integrated to generate the global model.

Fig: System Overview
Figure above shows the generic outline of the multi agent distributed data mining system. The proposed system consists of user agent which is given a specific data mining task by the end user. The user agent is responsible for spawning the specific task agent. The task agents persists till the task given to them is completed. They in turn spawn the data mining agent. The data mining agents are supposed to have varying data mining algorithm. The data mining agents uses data agents which act as conduit to and fro from data source. The validation agents are special type of agents that perform the validation operation on data mining results.

6. Applications

Examples of applications where agent based distributed data mining can be applied are Wireless area network, internet, remote sensing data and astronomical data.

6.1 Adhoc wireless sensor network

In wireless sensor adhoc network there is constraint of bandwidth. Hence it is not possible to transfer voluminous data across the network to a central node. So it needs distributed execution of data mining primitives at different nodes. Potential application for Wireless adhoc network is intrusion detection system.

6.2 Remote sensing data mining

The data collected from different satellites is stored across geographically separated places. This data is voluminous in nature. It is expensive and impractical to store all this data at central location. The agent based distributed data mining on this data base may be more efficient and scalable for mining these databases.
7. References


