ISSN- 2350-0530(O) ISSN- 2394-3629(P)
Impact Factor: 2.035 (I2OR)
DOI: 10.29121/granthaalayah.v4.i2.2016.2821

INTERNATIONAL JOURNAL OF RESEARCH – GRANTHAALAYAH
A knowledge Repository

Science

EMPLOYING AGENTS IN DESCRIPTIVE MINING

Dr. R. Sabitha [*1], Dr. S. Karthik [2]
[*1] Associate Professor, Department of IT, Info Institute of Engineering, Coimbatore, INDIA
[2] Professor & Dean, Department of CSE, SNS College of Technology, Coimbatore, INDIA

ABSTRACT

Agent technology and Data Mining have emerged as two of the prominent areas in information sciences. An effort has been activated towards the interaction and integration between agent technology and data mining which is referred to as "AGENT MINING". Data Mining is the process of extracting interesting information or patterns from large volumes of data. Agents comprise a powerful technology for the analysis, design and implementation of autonomous intelligent systems that can handle distributed problem-solving, cooperation, coordination, communication, and organization in a multiplayer environment. This agent uses information technology to find trends and patterns in an abundance of information from many different sources. The user can sort through this information in order to find whatever information they are seeking. Intelligent agents are today accepted as powerful tools for data mining in a distributed environment. The interaction and integration between agent and mining has potential to not only strengthen either side, but generate new techniques for developing more powerful intelligence and intelligent information processing systems. This paper discusses how agents are used in the various descriptive models of Data Mining. The various challenges and methodologies are analyzed and it clearly indicates the need for and the promising potential of agent mining for the mutual enhancement of both fields and for the creation of super-intelligent systems. Even though many researchers have been committed, more efforts are required to develop techniques and systems in practical perspectives.

Keywords:
Agent mining, Multi Agents, distributed data mining, KDD, Clustering, Association rule mining.

Cite This Article: Dr. R. Sabitha, and Dr. S. Karthik, "EMPLOYING AGENTS IN DESCRIPTIVE MINING" International Journal of Research – Granthaalayah, Vol. 4, No. 2 (2016): 111-120.

1. INTRODUCTION

1.1.OVERVIEW

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information

in their data warehouses. It is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns or models in data to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. [1]

Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments. [1]

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases.
- Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step.

Data mining is not specific to one type of media or data. Data mining is applicable to any kind of information repository like Flat files, Relational Databases, Data Warehouses, Transaction Databases, Multimedia Databases, Time-Series Databases, and World Wide Web. [2]

1.2.THE DATA MINING PROCESS

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
Impact Factor: 2.035 (I2OR)

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. [4]

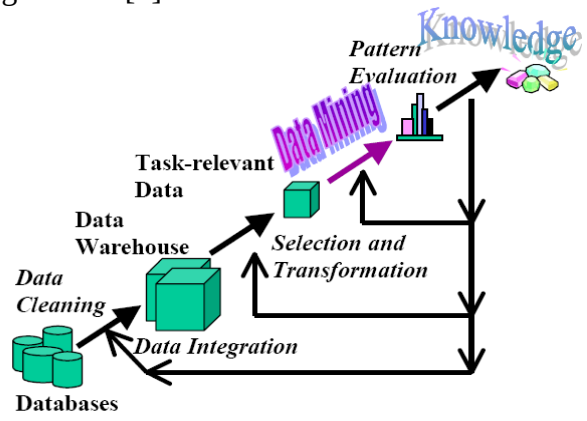

Figure 1.1: The process of data mining

Data pre-processing or data cleaning or data preparation is also a key part of data mining. Quality decisions and quality mining results come from quality data. Data are always dirty and are not ready for data mining in the real world. For example, data need to be integrated from different sources; data contain missing values. i.e. incomplete data; data are noisy, i.e. contain outliers or errors, and inconsistent values (i.e. contain discrepancies in codes or names); data are not at the right level of aggregation.

The main part of data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. The choice of a particular combination of techniques to apply in a particular situation depends on both the nature of the data mining task to be accomplished and the nature of the available data. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of the structure of the data during which time knowledge is acquired. Once knowledge has been acquired this can be extended to larger sets of data working on the assumption that the larger data set has a structure similar to the sample data.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.

- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation for which the answer is not known. This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. Once the model is built it can then be used in similar situations for which the answer is not known. [3]

1.3.THE MAIN DATA MINING TECHNIQUES

The most commonly used techniques in data mining are:

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships. [5, 6]

1.4.THE MAIN DATA MINING TASKS

Knowledge discovery (learning from data) comes in two flavours: directed (supervised) and undirected (unsupervised) learning from data.

The six main activities of data mining are:
Classification (examining the feature of a newly presented object and assigning it to one of a predefined set of classes);

Estimation (given some input data, coming up with a value for some unknown continuous variable such as income, height, or credit-card balance);

Prediction (the same as classification and estimation except that the records are classified according to some predicted future behavior or estimated future value);

Affinity grouping or association rules (determine which things go together, also known as dependency modeling, e.g. in a shopping cart at the supermarket - market basket analysis);

Clustering (segmenting a population into a number of subgroups or clusters); and

Description and visualization (exploratory or visual data mining).
The first three tasks – classification, estimation and prediction – are all examples of directed knowledge discovery (supervised learning). In supervised learning the goal is to use the available data to build a model that describes one particular variable of interest, such as income or response, in terms of the rest of the available data ("class prediction").

The next three tasks – affinity grouping or association rules, clustering, and description and visualization – are examples of undirected knowledge discovery (unsupervised learning). In unsupervised learning no variable is singled out as the target; the goal is to establish some relationship among all the variables ("class discovery"). Unsupervised learning attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. [7]

2. AGENT TECHNOLOGY

Agent mining refers to the methodologies, technology, tools and systems that synthesize multi agent technology, data mining and knowledge discovery, machine learning and other relevant techniques such as statistics and semantic web for better addressing issues that cannot be tackled by any single technique with the same quality and performance.

The concept of agent mining is interchangeable with other concepts such as agents and data mining interaction and integration, agent mining interaction and integration, or data mining and multi agent integration. The interaction and integration between agent technology and data mining and machine learning come from the intrinsic challenges, needs and opportunities faced by the constituent technologies both respectively and mutually.

New challenges are appearing with the emergence of new computing mechanisms such as behaviour computing, cloud computing, and social computing. Agent mining brings about multi-fold advantages to multi agent systems, data mining, and machine learning, as well as new derived theories, tools and applications that are beyond any individual technology. For instance, in a multi agent system, agents can collaborate with one another to work towards a mutual goal per predefined policies. With data mining and machine learning, the analysis of a counterpart's historical behaviours and the detection of current behaviours can effectively optimize agent collaboration performance and enhance capabilities for tackling exceptions and conflicts. By the same token, cloud analytics may be greatly enhanced by engaging automated cooperation between agent-based computing components.

The adaptability of norms and protocols may be enhanced in a social system through learning results from distributed agents in historical and real time. These examples illustrate the potential of agent mining in handling challenges in individual communities, and bringing about new opportunities for creating new technologies, tools and systems that cannot be delivered with any

single technology. For example, the collaboration of multiple trading agents, in which each contributes an optimal trading strategy learned from the historical market data, can lead to a better trading performance. [9]

3. OVERVIEW OF THE PREDICTIVE MODELS

3.1.ASSOCIATION RULE MINING

The association mining task can be stated as follows:
Let I be a set of items and D be a database of transactions, where each transaction has a tid and contains a set of items. A set of items is also called an itemset. An itemset with k items is called a k-itemset. The support of an itemset X, denoted as σ(X) is the number of transactions in which it occurs as a subset. The input data (D) for most ARM algorithms comprises N columns describing a binary valued set of attributes A, and M transactions such that each transaction describes some subset of A. The k length subset of an itemset is called a k-subset. An itemset is maximal if it is not a subset of any other itemset. An itemset is frequent if its support is more than a user-specified minimum support (min_sup) value. The set of frequent k-itemsets is denoted as Fk. If the support of an item set is greater than a given support threshold, "min_sup", the itemset is said to be large or frequent.

The data mining task is to generate all association rules in the database, which have a support value greater than min_sup. These association rules are frequent.

This task can be broken into two steps:

1. Find all frequent itemsets. Given m items, there can be potentially 2m frequent itemsets. Efficient methods are needed to traverse this exponential itemset search space to enumerate all the frequent itemsets.
2. Generate confident rules. This step is relatively straightforward to generate association rules of the form X - > Y, for all frequent itemsets X and Y. [4]

Mining Association Rules using Agents:

If the top level management needs to mine novel information in the process of decision making, there are two options. The first one is to transfer data to a single database and mine it on that database. The second option is to mine them independently and still generate information for the combination of the data in multiple databases. The architecture of a data mining system using intelligent agents is presented in the following figure 3.1.The development of distributed rule mining is a challenging and critical task since it requires knowledge of all the data stored at different locations and the ability to combine partial results from individual databases into a single result [10]. The individual databases have to be analyzed to generate rules to make local decisions. It would be easier for the organization to make decisions based on the rules generated by the individual branches, rather than using the raw data. If the raw data from each of the individual databases were sent to a single database to generate the rules, certain useful rules, which would aid in making decisions about local branches, would be lost. If the raw data from all the databases were transferred to a single database then each of the individual branches would

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
Impact Factor: 2.035 (I2OR)

not be generating the rules with respect to its data. In such a case the organization may miss out certain rules that were prominent in certain branches and were not found in the other branches similar to the above example. Generating such rules would aid in making decisions about specific branches.

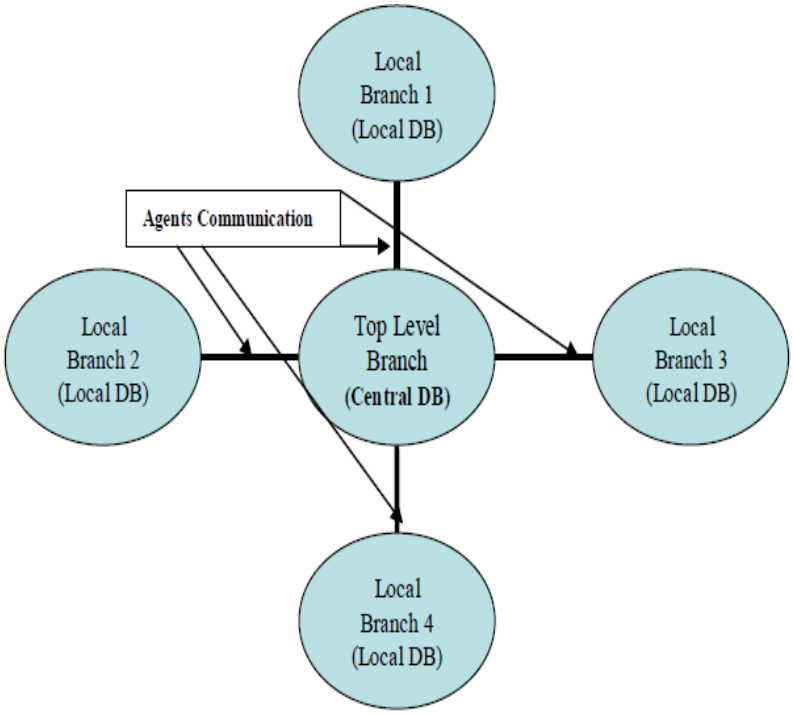

Figure 3.1: Mining rules in a distributed environment using intelligent agents.

The patterns in multi-databases are divided into the following classes:

- Local patterns: Local branches need to consider the original raw data in their datasets so they can identify local patterns for local decisions.
- High-vote patterns: These are the patterns that are supported by most of the branches and are used for making global decisions.
- Exceptional patterns: These patterns are strongly supported by only a few branches and are used to create policies for specific branches. [10].

3.2.CLUSTERING

Data analysis procedures can be dichotomized as either exploratory or confirmatory, based on the availability of appropriate models for the data source, but a key element in both types of procedures is the grouping, or classification of measurements based on either

i. goodness-of-fit to a postulated model, or
ii. Natural groupings (clustering) revealed through analysis.

Cluster analysis is the organization of a collection of patterns into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. An example of clustering is depicted in Figure 3.2.

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
Impact Factor: 2.035 (I2OR)

The input patterns are shown in Figure 1(a), and the desired clusters are shown in Figure 1(b). Here, points belonging to the same cluster are given the same label. [1]

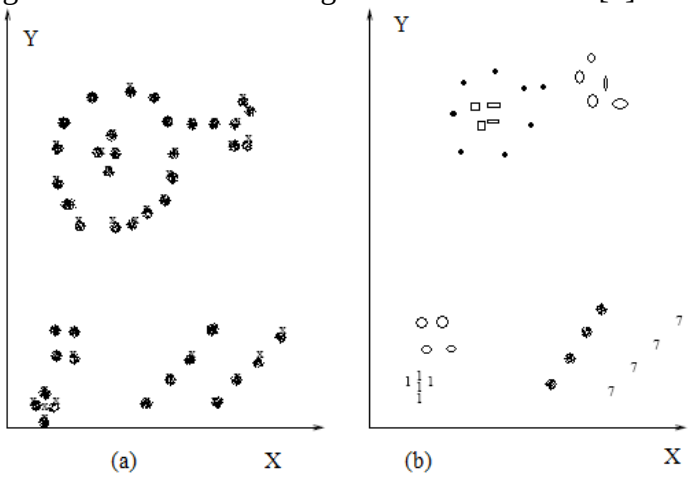

Figure 3.2: Example for Clustering

There are different approaches to cluster the data. At the top level, there is a distinction between hierarchical and partitional approaches (hierarchical methods produce a nested series of partitions, while partitional methods produce only one). There are various approaches to cluster the data namely, Agglomerative, Divisive, Monothetic, Polythetic, Hard, Fuzzy, Deterministic, Stochastic, Incremental and Non-incremental. [1,8]

Clustering use agents to facilitate data privacy and support distributed data clustering.
The proposed MADM framework, as noted above, comprises four categories of agent:

1. User agents.
2. Data agents.
3. Clustering agents.
4. Validation agents.

User agents are the interface between end users and the MADM environment. The agents are responsible for obtaining the input from the user, spawning clustering agents in order to perform the clustering task and presenting the derived clustering result. To the above list of specific MADM agents we can also add a number of housekeeping agents that are utilised within the MADM framework. Data agents are the "owners" of data sources. There is a one-to-one relation-ship between data agents and data sources.

Data agents can be thought of as the conduit whereby clustering agents can access data. Clustering agents are the "owners" of clusters. Groups of clustering agents can be thought of as representing a clustering algorithm. With respect to this paper the K-means and K-Nearest Neighbour (KNN) clustering algorithms have been adopted; however, our collections of clustering agents could have been configured to perform some alternative form of clustering (for example hierarchical clustering).

A number of clustering agents will be spawned, as required, by a user agent in order to perform some clustering task. Thus, each clustering agent represents a cluster and is responsible for

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
Impact Factor: 2.035 (I2OR)

selecting a record from a data set and determining whether that record would belong to its cluster or not. The number of clustering agents therefore depends on the number of clusters (K). In the case of the K-means algorithm the number of clusters is predefined; thus, by extension, the number of clustering agents that will be spawned will also be predefined. In the case of the KNN approach only one initial clustering agent will be spawned; then, as the KNN algorithm progresses further clustering agents may be created.

Clustering agents collectively have two principal functions: (i) initial generation of a "start" cluster configuration, and (ii) cluster refinement. Validation agents are a special type of agent that performs validation operations on clustering results. Each validation agent is the "owner" of a technique or measuring the "goodness" of a given cluster configuration. In the current system validation agents consider either cluster cohesion or cluster separation or both. A possible configuration for the proposed MADM framework, incorporating the above, is presented in Figure 3.3.

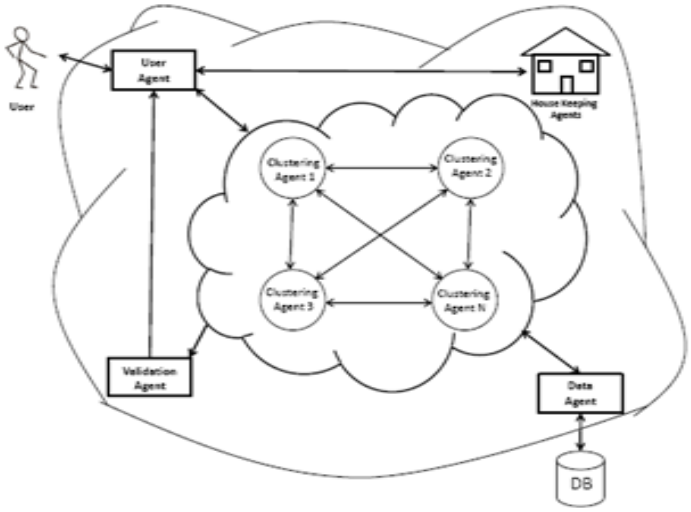

Figure 3.3: MADM framework

The Figure includes a User Agent, a collection of Clustering Agents, a Data Agent, a Validation Agent and some house- keeping agents. The directed arcs indicate communication between agents. Note that communication can be bidirectional or unidirectional and that the Data Agent directly communicates with each Clustering Agent. Intra-communication between Clustering Agents takes follows a protocol that permits negotiation about cluster exchange to take place. [10]

4. CONCLUSIONS & RECOMMENDATIONS

This paper discusses how agents are used in the various descriptive models of Data Mining. The various challenges and methodologies are analyzed and it clearly indicates the need for and the promising potential of agent mining for the mutual enhancement of both fields and for the creation of super-intelligent systems. Even though many researchers have been committed, more efforts are required to develop techniques and systems in practical perspectives.

5. *REFERENCES*

[1] *Pavel Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software, Inc.*
[2] *MURTAGH, F, Multidimensional Clustering Algorithms, Physica-Verlag, Vienna.1985.*
[3] *JAIN, A., DUBES, R, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.*
[4] *ZHANG, T., RAMAKRISHNAN, R. , LIVNY, M, BIRCH: an efficient data clustering method for very large databases, In Proceedings of the ACM SIGMOD Conference, Montreal, Canada, 1996, 103-114.*
[5] *Rakesh Agrawal., Johannes Gehrke., Dimitrios Gunopulos., Prabhakar Raghavan, Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, In Proceedings of the International Conference on Management of Data, (SIGMOD), volume 27(2) of SIGMOD Record, S`eattle,WA, ACM Press, USA, 1–4 June 1998 , 94–105..*
[6] *Sudipto Guha., Rastogi, R., Shim, K, CURE: A clustering algorithm for large databases, Technical report, Bell Laboratories, Murray Hill, 1997.*
[7] *WANG, W., YANG, J., MUNTZ, R, STING: a statistical information grid approach to spatial data mining". In Proceedings of the 23rd Conference on VLDB, Athens, Greece, 1997,186-195.*
[8] *Mihael Ankerst., Markus M., Breunig., Hans-Peter Kriegel., J ¨org Sander, OPTICS: Ordering Points To Identify the Clustering Structure", In Proceedings of the International Conference on Management of Data, (SIGMOD), volume 28(2) of SIGMOD Record, ACM Press, Philadelphia, PA, USA, 1–3 June 1996, 49–60.*
[9] *Cristian Aflori., Florin Leo, Efficient Distributed Data Mining using Intelligent Agents, Proceedings of the 8th International Symposium on Automatic Control and Computer Science, Iaşi, 2004.*
[10] *Santhana Chaimontree., Katie Atkinson., Frans Coenen, A Multi of Agent Based Approach to Clustering: Harnessing the Power, Conference Proceeding: 01/2011; In proceeding of: Agents and Data Mining Interaction - 7th International Workshop on Agents and Data Mining Interation, ADMI 2011.*