IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 3 (Nov. - Dec. 2013), PP 44-47
www.iosrjournals.org

A Survey of Agent Based Pre-Processing and Knowledge Retrieval

Sapna Pujara[1], Dr. Kanwal Garg[2], Mr. Bharat Chhabra[3]
[1](*Research Scholar, DCSA, Kurukshetra University, Kurukshetra, India)*
[2](*Assistant Professor, DCSA, Kurukshetra University, Kurukshetra, India)*
[3](*Assistant Professor, DCSA, Government College, Safidon , Jind, India*

Abstract: *Information retrieval is the major task in present scenario as quantum of data is increasing with a tremendous speed. So, to manage & mine knowledge for different users as per their interest, is the goal of every organization whether it is related to grid computing, business intelligence, distributed databases or any other. To achieve this goal of extracting quality information from large databases, software agents have proved to be a strong pillar. Over the decades, researchers have implemented the concept of multi agents to get the process of data mining done by focusing on its various steps. Among which data pre-processing is found to be the most sensitive and crucial step as the quality of knowledge to be retrieved is totally dependent on the quality of raw data. Many methods or tools are available to pre-process the data in an automated fashion using intelligent (self learning) mobile agents effectively in distributed as well as centralized databases but various quality factors are still to get attention to improve the retrieved knowledge quality. This article will provide a review of the integration of these two emerging fields of software agents and knowledge retrieval process with the focus on data pre-processing step.*

Keywords: *Data Mining, Multi Agents, Mobile Agents, Preprocessing, Software Agents.*

I. Introduction

Knowledge discovery in databases is a rapidly growing field, its growth can be figured out from the increasing interest of researchers & its practical, social as well as economical needs. The Knowledge Discovery in Databases process comprises of a few steps to get knowledge by processing raw data. The iterative process consists of the steps viz. Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and finally Knowledge representation to visually represented the results of data mining. It is common to combine some of these steps together. As data cleaning and data integration can be combined together and known as a pre-processing phase to obtain target data. Data selection and data transformation can also be combined where the selection is done on transformed data. To get this filtered and so called pre-processed data which is used for data mining is our area of focus. This step is significant in the sense that it may play a crucial role to resolve the data quality problems which arise during data collection. As data collection methods are loosely controlled, which results in incomplete, inconsistent, noisy, missing values, etc.(Dasu & Johnson 2003). If there is much irrelevant & redundant information is present in the initially collected data then knowledge retrieval during mining process will not be meaningful. Therefore the need for data cleaning which produces final training set increases significantly for accurate results.

The manual process of data pre-processing becomes tedious as size of data grows and its quality degrades, so the process of data pre-processing needs to be automated to the extent possible. It is also a legitimate option to get it done with the help of software agent. As studies have been conducted on the integration of these two fields of software agents & knowledge discovery.

II. Software Agents

The idea of an agent originated with john McCarthy in the mid-1950's, and the term was coined by Oliver G. Selfridge a few years later, when they were both at the Massachusetts institute of technology (Kay 1984).

A software agent may be assumed as a piece of code/program having autonomy, persistence, proactiveness, reactiveness and some others among its major characteristics as shown in Fig. 1. In the process of data mining, once the task of discovering some knowledge is delegated, it is followed by ascertaining the goal, fixing the strategy and behaving proactively to complete the necessary actions.

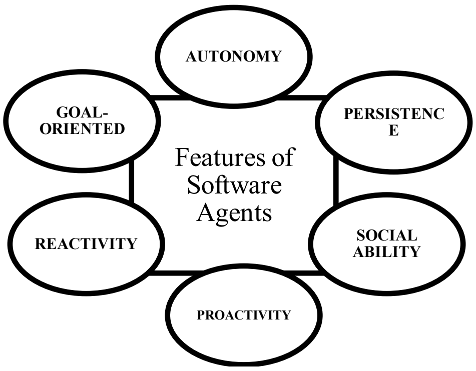

Figure 1 Features of Software Agents

Software agents are being touted & used for various diverse applications as electronic commerce, computer games, personalized information management, interface design and management of complex commercial and industrial processes.

According to N. Jennings & M. Wooldridge, 1996 software agents can be *simple* which execute on pre-specified rules & assumptions, *sophisticated* which execute some specified task on user's request and *predictive* which executes when it finds it appropriate without being explicitly asked. Some special features of software agents helps it to carve its niche among other major streams like ubiquitous computing, distributed processing, object-oriented systems and artificial intelligence.

III. Related Work

(Davies et al. 1995) has introduced Agent-K to work over distributed databases which offers a simple production rule mechanism for programming of agents & these rules respond to the KQML (Finin et al. 1993) messages & the current state of agent. Here, different approaches for knowledge integration has also been discussed. As admitted by the author himself that this approach of knowledge integration turns out to be quite heavy in terms of computation. Further work will require to focus on maximum usage of distributed resources and least network traffic.

(Seydim 1999) represented a study on agent model in the context of information retrieval, filtering, classification and learning along with communication framework for the interchange of information between multiple mobile agents(Harrison, Chess, Kershenbaum 1995) working over distributed systems for knowledge discovery.

(Vassiliadis. 2000) developed ARKTOS, an automated tool for data cleaning and transformation in data warehouse environments. Even though specialized tools for are already available designed by Ardent Software and Data Mirror corporation etc., this ARKTOS tool is well equipped with one graphical interface and two declarative languages closely related with XML and SQL. The tool contains primitive operations for Extraction-Transformation-Loading and more especially cleaning primitives like primary key violation, reference violation, domain mismatch and others.

(Knoblock & Craig 2004) Introduced the set of software agents for travel planning by retrieving information from web. These agents provide interactive interface as user is provided with all the choices & monitors all the aspects of a trip. Finally, this software perform mining over all information to help the user in their decision making.

(Zghal et al. 2005) has introduced the agent framework for data mining of spatial data by combining the different algorithms of data mining & features offered by the multi-agent systems. Authors also resented the architecture of Computer Aided Spatial Agent Mining Mart Environment (CASAMME) and a CASE tool (2003) based on multi agent system.

(Ong et al. 2005) has discussed the problems of wrong assumptions associated with data mining algorithm i.e. static view of data & a stable execution environment as contrast to dynamic view of data and execution environment. As a solution to this, author has introduced a new methodology of designing stream-based algorithms with mobile agents. The proposed multi agent system has been implemented also over stream based data in their Matrix project, where the used the concept of summary structures to develop several mobile agents and they work under the coordination of a special agent called plan and coordinate agent, whose responsibility is to form the execution plan based on current (dynamically altered) data set and dispatch the necessary instructions and information to several mobile agents sitting over distributed locations of data. The experimental results show the speed up attained which is roughly closer to be linear.

(Bach et al. 2005) proposed retrieval of public data spread over web with software agents for business intelligence. Software agent work for retrieval of data from the Data base of stolen cars in Croatia, the data thus collected is also analyzed and various reports are prepared stating risk involved in different classes and brands of cars to help for better decision making in insurance company.

(Nurmi. 2005) has presented an architecture for distributed data pre-processing. Although the approaches for the development of the context aware application mobile agents may be implemented at application level or middleware level, the architecture proposed by the author enlightens and take the advantage of both levels. The intelligent agents described under this architecture does the sensing of data and then to preprocess the collected data, pattern recognition techniques like handling missing values, outlier removal and normalization etc. are implemented. The architecture framework presented by the authors is initiated with recognition phase, followed by decision making phase. Authors emphasis on the first phase which actually includes several subtasks like feature extraction, feature selection, classification etc. Even though, at first glance the architecture seems to focus on just preprocessing phase whereas it is also claimed that it may actually be implemented over distributed ubiquitous environment with a little touch to interface.

(Tudor et al. 2009) has emphasized the use of software agents to figure out the relevant information so that academic organizations may focus their activities on improving management quality by using knowledge. Here data mining & software agents are combined to work on knowledge management in academic environment with the help of multiple agents at multiple levels(educational, research, administrative, secretarial, others).

(Moemeng et al. 2010) has introduced an agent-based distributed data mining platform named i-Analyst consisting of software packages & development kit for the improved performance of data mining algorithms by maintaining security & privacy of the data. The example results itself reveals the significance of the agents in enhancing the execution performance.

(Singh et al. 2011) discussed & compared five agent development toolkits: JADE, VOYAGER, ZEUS, AGLET & ANCHOR developed by different groups. The comparison has been drawn on the basis of standards followed, security mechanism, agent mobility and migration scheme etc. Authors deduced that Jade (open source) agent development toolkit is most balanced toolkit. Voyager is a commercial tool, Anchor provides good security, Aglet lacks security and scalability but these three doesn’t comply with FIPA standards. On the other hand, Zeus supports FIPA standards but doesn’t provide agent mobility.

(Jayabrabu et al. 2012) proposed the automated process of data mining for better visualization with the integration of multi-agent system to detect those new and hidden patterns which otherwise are not available to less domain user (novice and specialist users). The automated clustering of relevant data set by these agents leads to good input cluster to mine on, which ultimately returns the better correlated output which are visualized by link charts instead of traditional data mining visualization methods like graphs, pie charts or histograms etc.

IV. Present Scenario

The success story of software agents is itself advocated by the above mentioned developments in the field of data mining. Going through present applications of data mining & software agents as an integrated technology, it is revealed that these fields are proving successful with their efficient outcomes or results. Software agents are designed to perform data mining in various fields like travelling agents, academic management, business intelligence, data-stream mining, distributed data mining & many more areas. Software agents are also capable for collecting valuable data from the Web.

These strong features of software agents are exploited by researchers to achieve the specified goal autonomously & in a predictive manner which is the demand of present customers who want to get knowledge as much as possible for decision making. This increasing demand for better decision support is solved by availability of knowledge discovery products, in the form of research prototypes developed at various universities as well as software products from commercial vendors like WEKA, ORANGE and etc.

For this various authors presented various platforms for the designing of the different types of agents. Many toolkits (JADE, AGLET, ZEUS & ANCHOR) are available which helps in development of software agents.

V. Future Scope

The present situation strongly suggests that the use of intelligent mobile agents is inevitable for the achievement of rationale outcomes in the field of data mining especially distributed one. The work carried on so far by various researchers show the successful implementation of these software agents in unique domains like academic, management and ubiquitous computing. Although the availability of different tools in market focus on Extraction -Transformation-Loading including data cleaning primitives, each one having its own limitations viz. limited transformation primitives, non-availability of impact analyzer, non-availability of optimizer to prefer one efficiency criteria over the other. These open issues motivate the researchers to keep on putting hard efforts for the balanced discovery of intelligent mobile agents based automated tools containing rich variety of primitives for preprocessing and simple point and click type front end. Our future efforts will focus on summarizing the specific problem areas during preprocessing using intelligent mobile agents. Our following objective aims to find the role of mobile agents for elimination of data cleaning problems existing with data collected from same or different sources. Where both classes have their own inherent problems at schema level or instance level. Better quality of data terminates in a good knowledge outcome since the knowledge discovery is performed on the best quality of data. Other issues among multi-agent systems comes up during agent selection and knowledge integration where some agents have to wait for others to finish their assignment. Other objective focuses the assessment of the available automated tools of data pre-processing where all these will be studied deeply and a comparative analysis will be performed on all those methods to figure out the best one.

VI. Conclusion

Data pre-processing is a very important step in the data mining which has a huge impact on the results but often neglected. Analyzing data that has been carefully screened for problems of data pre processing will produce more accurate & reliable results. The output of data pre-processing is the final training set on which user performs knowledge retrieval process to get desired results. These results should provide user some useful and trustworthy knowledge which will support them in decision making. Therefore data should be pre-processed in automatic manner every time before performing knowledge discovery.

The above review presented is far away from full or comprehensive, although it describes the basic role of intelligent mobile agents in the process of data mining especially preprocessing. There are numerous other application areas of multi agent systems with learning (intelligent) ability like MALE (Sian. 1991), ANIMALS (Davies. 1993) etc., which we could not describe here for the sake of length of the paper. The entire process of preprocessing with agents should be with parallel consideration of scalability and ultimately these intelligent agents should result in resolving the inconsistencies in static or dynamic data and also reduce the size of data. The intelligent mobile agent possessing these characteristics would definitely lead to a drastic drop in the manual preprocessing activities and commercially viable and successful tool.

References

[1] Bach, M P, Vlahovic, N, & Knezevic, B 2005, September, “Public data retrieval with software agents for business intelligence”, *in proceedings of the 5th wseas int. Conf. On Applied Informatics*, pp. 15-17.

[2] Bellifemine, F, Poggi, A, & Rimassa, G 2001, “Developing multi-agent systems with JADE”, *Intelligent Agents VII Agent Theories Architectures and Languages*, pp. 42-47.

[3] Dasu, T, & Johnson, T 2003, “Exploratory data mining and data cleaning”, *Wiley-Interscience*, vol. 442.

[4] Davies, W 1993 “ANIMALS: A distributed, heterogeneous multi-agent learning system”, *MSc Thesis, University of Aberdeen*.

[5] Davies, W H, & Edwards, P 1995, “Agent-based knowledge discovery”, *In Working Notes Of The Aaai Spring Symposium On Information Gathering From Heterogeneous, Distributed Environments. Stanford University, Stanford, Ca Winton*.

[6] Harrison, C G, Chess, D M, Kershenbaum, A 1995 “Mobile Agents: Are they a good idea?”, *IBM Research Report, T.J.Watson Research Center, NY.*

[7] Jayabrabu, R, Saravanan, V, & Vivekanandan, K 2012, “Software agents paradigm in automated data mining for better visualization using intelligent agents”, *Journal Of Theoretical And Applied Information Technology*, 39(2).

[8] Jennings, N, & Wooldridge, M 1996, “Software agents”, *IEE Review*, 42(1), 17-20.

[9] Knoblock, C A 2004, “Building Software Agents For Planning, Monitoring, And Optimizing Travel”, *University Of Southern California Marina Del Rey Information Sciences Inst.*

[10] Moemeng, C, Zhu, X, Cao, L, & Jiahang, C 2010, “I-Analyst: An agent-based distributed data mining platform”, *In Data Mining Workshops (ICDMW), 2010 IEEE International Conference, IEEE*, pp. 1404-1406.

[11] Nurmi, Petteri, Przybilski, Michael, Linden, Greger & Floreen, Patrik 2005, " An Architecture For Distributed Agent-Based Data Preprocessing".

[12] Ong, K L, Zhang, Z, Ng, W K, & Lim, E P 2005, “Agents and stream data mining: a new perspective”, *Intelligent Systems, IEEE,* 20(3), pp. 60-67.

[13] Seydim, A Y 1999, “Intelligent agents: A data mining perspective”, *Southern Methodist University, Dallas*.

[14] Sian, S 1991, "Extending learning to multiple agents: issues and a model for multi-agent machine learning (ma-ml)" , *in proceedings of the European Working Session On Learning - ewsl91, Y. Kodratroff (ed.), Springer- Verlag*, 458-472.

[15] Singh, A, Juneja, D, & Sharma, A K 2011, “Agent development toolkits”.

[16] Tudor, I, & Ionita, L 2009, “Intelligent agents as data mining techniques used in academic environment”, *In The 4th International Conference On Virtual Learning* ,vol. 156, pp. 380-384.

[17] Vassiliadis, Panos, Vagena, Zografoula, Skiadopoulos, Spiros, Karayannids, Nikos, Sellis, Timos 2000 "ARKTOS: A tool for data cleaning and transformation in data warehouse environments", *Bulletin Of The IEEE Computer Society Technical Committee On Data Engineering*.

[18] Witten, I H, Frank, E, Trigg, L E, Hall, M A, Holmes, G, & Cunningham, S J 1999, “ Weka: Practical machine learning tools and techniques with Java implementations.”

[19] Zghal, H B, Faiz, S, & Ghezala, H B 2005, “A framework for data mining based multi-agent: An application to spatial data”, *World Academy of Science, Engineering and Technology*.