# AGMI - AN AGENT-MINING TOOL AND ITS APPLICATION TO BRAZILIAN GOVERNMENT AUDITING

Carlos Vinícius Sarmento Silva
*Controladoria-Geral da União, SAS, Qd 01, Bl A, Edifício Darcy Ribeiro, Brasília, DF, CEP 70.070-905, Brazil*

Célia Ghedini Ralha
*Computer Science Department, University of Brasília, POBox 4466, Brasília, DF, CEP 70.904-970, Brazil*

Keywords: Agent-mining tool, Brazilian government auditing, AGMI, e-Government.

Abstract: This paper presents research combining two originally separated areas increasingly interrelated: distributed multi-agent systems and data mining. In our approach, we prove the interaction features in a bilateral and complementary way, since we have defined an integrated architecture and developed a prototype, which has been used in a government auditing study case. In Brazil, government auditing is performed by the Office of the Comptroller General (CGU), where several approaches are being used to prevent and fight corruption. However, some activities such as government purchasing fraud detection are limited by the difficulty in finding effective ways to implement. Considering data mining perspective, we have used different model functions, such as clusterization and link analysis with association rules. Our approach integrating multi-agent and data mining techniques resulted in expressive discovered knowledge, which would help detection of cartels acting in public bidding processes at CGU.

## 1 INTRODUCTION

The CGU is the agency of the Federal Government in charge of assisting the president of Brazil in matters which, within the executive branch, are related to defending public assets and enhancing management transparency through internal control activities, public audits, corrective and disciplinary measures, corruption prevention and combat, and coordinating ombudsman's activities (http://www.cgu.gov.br/english/).

Nowadays, a large volume of information has been produced and stored by the Brazilian government information systems. Considering only 2009, the Federal Accountability System (SIAFI) registered one billion of financial transactions. All this data are normally used to support the preparation and execution of government auditing. In this way, CGU has driven efforts to apply technologies in order to promote transparency and prevent corruption. However, the analysis of the available data to produce useful knowledge to auditing activities is a hard task. Data mining (DM) and knowledge-discovery in databases (KDD) are playing an important role, especially when integrated to other computational techniques to analyze and explore information.

In the past decade, intelligent agents/multi-agent systems (MAS) and DM/KDD have emerged as two increasingly interrelated research areas, opening space to the agent-mining interaction and integration (AMII) research field. This new field has driven efforts from both sides to find benefits and complementarities to both communities (Cao, 2009; Ralha, 2009).

In the context of CGU and AMII research field, this paper presents an agent-mining tool – AGMI. AGMI has been tested using the Brazilian Government Auditing data, in order to help to prevent corruption through KDD. The Brazilian Federal Bidding database used is entitled ComprasNet (http://www.comprasnet.gov.br/).

The rest of this paper is presented as follows: in Section 2, we discuss the problem at the Brazilian Government Auditing domain; in Section 3, we present AGMI architecture and prototype; in Section 4, we present the experimental results and the discovered knowledge; and finally, in Section 5, we conclude and suggest future work.

Vinícius Sarmento Silva C. and Ghedini Ralha C..
AGMI - AN AGENT-MINING TOOL AND ITS APPLICATION TO BRAZILIAN GOVERNMENT AUDITING.
DOI: 10.5220/0003333905350538
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 535-538
ISBN: 978-989-8425-51-5

## 2 THE PROBLEM

In general, the identification of cartels is a difficult task since it requires analysis of several public bidding processes, which usually exceeds the scope of only one government department. Cartels can operate in various government departments, cities and even states of the Federation. Furthermore, the analysis of data from databases using Structured Query Language (SQL) queries is also impractical because of the exponential solution space. Thus, the problem consists in creating an efficient way to identify groups of companies which might be suspected of practicing cartels in public bidding processes.

In (Silva and Ralha, 2010), a solution using Association Rules to help solving the problem of cartels detection in public bidding processes is proposed. The proposal is due to the fact that this technique is useful to find strong relationships among attributes. Thus, it is possible to apply this technique creating a dataset, so each attribute is a boolean value, which indicates the participation or not of the company in each bidding process of the database. The dataset for association rule technique must be constructed as the matrix $A$ consisting of $m$ rows and $n$ columns, where $m$ is the total of bidding processes from database and $n$ is the total of companies from database.

$$a_{i,j} = \begin{cases} true & \text{if } j \text{ has participated of } i \\ false & \text{if } j \text{ has not participated of } i \end{cases}$$

$$1 \leq i \leq m; 1 \leq j \leq n;$$
where $i$ is a bidding and $j$ is a company

Thus, we expect to obtain rules with transitive properties, like the following: $company_A = true,\ company_B = true \rightarrow company_C = true$.

However, tasks like dataset preparation, the execution of different DM algorithms and rule evaluation are adequate to be integrated to the MAS approach, considering either the distribution aspects of the automated tasks or the parallel execution of algorithms. Thus, we have to propose a way to improve the time execution of the tasks, since it took us weeks to perform manually the tasks of DM using Weka.

## 3 AGMI ARCHITECTURE AND PROTOTYPE

DM has many characteristics, which make interaction and integration to MAS possible by combining KDD through the different DM techniques. In this section, we present our architecture of cooperative DM, which uses two frameworks: Java Agent Development Framework (JADE), a Java framework to support the development of MAS (Bellifemine et al., 2007); and Data Mining for Agents (DMA), a Weka-based framework, developed to enable integration of DM in a multi-agent environment.

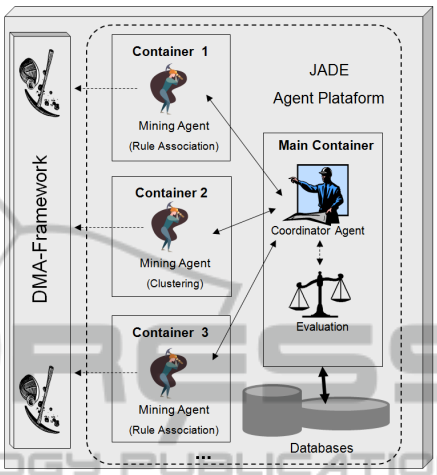

Figure 1: AGMI Architecture Schema.

Figure 1 presents the structural schema of AGMI. The agents run on JADE platform, and may be distributed among different physical hosts. The model uses three different types of agents: Mining Agent, Coordinator Agent and Evaluator Agent.

Mining Agent – responsible for running specific DM algorithms. Thus, we have multiple agents performing the same or different services of DM, whereby the DM services are classified by DM techniques (e.g. Association Rule Service, Clustering Service).

Coordinator Agent – responsible for overall coordination of the DM activities executed by the mining agents. It prepares datasets for the agents and coordinates the interactions allowing the cooperation among different agents. It also has the ability to hire a mining agent to perform a service required. When the service required is provided by more than one agent, the coordinator chooses the agent that has the best profile to perform the task, based on memory and processing capacities. Agent services can be found in JADEs Yellow Pages – a service provided by JADE to publish the services of registered agents (Directory Facilitator).

Evaluator Agent – responsible for evaluating the found knowledge before being presented to the user. In general, DM algorithms work with evaluating criteria (e.g. support and confidence in association rules). However, it is possible to adopt other criteria to discovered knowledge. Thus, we can adapt this agent

to perform a specific function or just define the minimal values of covering and accuracy of rules resulting from mining algorithms.

# 4 EXPERIMENTAL RESULTS

In order to test AGMI, we used real data from ComprasNet – the Brazilian Federal Bidding online system. The used data is relative to all bidding processes of a specific type of service contracted by Federal Executive agencies, between years 2005 and 2008, including all states of Brazil (26 states + the Federal District). The database includes 26,615 records, 2,701 bidding processes and 3,051 companies. Each record in the dataset represents one bid of a company in a specific bidding process.

## 4.1 First Experiment

We started our experiments with AGMI using three agents: the coordinator, the evaluator and the rule association agent. The Rule Association Agent used the *Apriori* algorithm, available in Weka framework. The algorithm has been adapted to DMA in order to be used in a multi-agent environment. In previous tests, before AGMI prototype was implemented (Silva and Ralha, 2010), we found that a rule with high lower bound of support might just imply the presence of big companies in bidding processes (frequent itemsets). Thus, setting a high lower bound of support in this algorithm can suppress the appearance of several good rules, with real features of cartels. On the other hand, high lower bound of confidence ensures the selection of good rules.

With the help of experts, in our first experiment with AGMI, we set the lower bound of support to get rules with 9 occurrences on the database, and the lower bound of confidence was set in 90%. We have also defined Equation 1 for evaluating the rules obtained through the DM process. This function was implemented in the Evaluator Agent for measuring the quality of the rules, and means the probability of one company of the suspicious group winning a bidding process. The higher the Rule Quality (*RQ*) value, the more suspicious the group is of cartel practicing.

$$RQ = 100.\frac{V(C)}{Sup. \times Inst.} \quad (1)$$

In Equation 1, $Sup.$ is the rule support value; $Inst.$ is the total number of instances; $C$ is the company set from rule; $B$ is the bidding processes where the company group $C$ has participated; and $V(C)$ is the number of victories in $B$ of any company in $C$.

We ran our tests using two computers: Host A (Intel Core 2, 2.40 GHz, 2.00 GB RAM) and Host B (Intel Pentium Dual, 1.86 GHz, 2.00 GB RAM). The coordinator and evaluator agent were set in Host A, and the association rule agent was set in Host B. In this experiment, we found 128 rules, and the execution time was 29 minutes. The top 100 rules scored an average of 16.56 in *RQ* evaluation. The average of support, on the other hand, was 30.98.

The top 10 rules scored an average of 33.00 in *RQ* evaluation, and the average of support was 26.90. The best rule, according to Equation 1 had 46 points of *RQ* and 26 of support. As noticed, the top ten rules have a smaller average support than the top 100 rules. Validating the results with auditing experts, we conclude that higher limits of support do not guarantee better rules.

## 4.2 Second Experiment

In the First Experiment, with only one mining agent in the system (Rule Association), we couldn't set the lower bound support to less than 24. This happened due to the lack of memory resources available in our machines. As lower the bound of support is set, as more memory is consumed by the rule association algorithm. Furthermore, it's quite possible to have rules with support less than 24 with real characteristics of a cartel. Thus, we needed to introduce a strategy to divide our space to seek the groups more accurately. For this, clustering is an appropriate DM technique to do such activity.

In addition to the Rule Association Agent we have used Clustering Agent, which implemented the Expectation-Maximization (EM) algorithm to discover clusters considering the companies and the Brazilian states. According to (Han and Kamber, 2005), in EM algorithm, each object is assigned to each cluster according to a weight representing its probability of membership. For the experiments with the Clustering Agent we have included it in Host A.

With the addition of the Clustering Agent, we found the regions of public biddings, and the companies that participated of those biddings in each region. Seven clusters of states were found, and, for each cluster region, the rule association agent ran his technique searching associated companies. The found clusters were: 1 - {AM, PA, AC, RO, AP, RR, MT, MS, RJ, ES and MG}; 2 - {RS, SC, PR}; 3 - {BA, SE}; 4 - {PE, PB, RN, CE, PI, MA} 5 - {TO, DF, GO}; 6 - {SP}; 7 - {AL}.

We found in this experiment 6,150 rules. In the top 100 rules, we found an average of 69.71 in RQ, and 9.78 of support. In the top 10 rules, the RQ

average was 89.70, and the average of support was 9.40. The time spent in the execution was 75 minutes. Based on this result, we can notice that the addition of the Clustering Agent improved the results dramatically. Considering the top 10 rules of both experiments, the values for quality of rules has grown up more than 150%.

If we compare the average of the Top 100 rules, the Second Experiment presents an average more than 4 times greater than the First Experiment's average. With the addition of the Clustering Agent in our tests, the search space was divided, and so it was possible to reduce the lower bound of support generating more and better rules. Thus, we improved a lot the quality of rules produced by the system. This proves the potential to integrate different DM techniques to SMA.

We also expected that all clusters were made up by states with common borders. In general, the companies do not act in states far from each other. However, the results have shown one cluster with large proportions and with some very disconnected states (Cluster 1). For more information about the geographic position of the Brazilian states, visit http://www.ibge.gov.br/estadosat/.

### 4.3 Discovered Knowledge

The experiments present lots of rules, specially the second one (Section 4.2), due to the value of lower bound of support. And several rules presented by the Evaluator Agent after the execution showed lots of company groups with evidence of cartel acting. The Clustering Agent result also revealed the trends of companies' participation in public bidding in Brazil. Following, we present some of the best rules as example of discovered knowledge:

- In 9 different public biddings in the same state, and in just one government agency, one rule has pointed the participation of two specific companies. In spite of the fact that both companies have participated in all biddings, just one of the companies has won all of the biddings. When we analyzed the history of the loser company, we found that it had only participated in those 9 exactly bidding processes, evidencing a possible simulation of competition to hide the cartel characteristics. Probably the company was created just to simulate competition in the public biddings, where competition is mandatory.
- A group with cartel characteristics, made up by 4 companies, has acted in Region 2 in 10 different public biddings, and has won 9 of these. Two companies of that group have appeared in other 7 different groups, in the same region, discovered by other rule association.

## 5 CONCLUSIONS AND FUTURE WORK

Apart from being in charge of inspecting and detecting frauds in the use of federal public funds, the CGU is also responsible for developing mechanisms to prevent corruption. Thus, in this article we presented the Agent-Mining Tool – AGMI, a MAS based tool to support distributed DM activities which was used at the Brazilian Government Audition domain.

In addition to enable the combination of distinct DM techniques, in order to improve the KDD process, AGMI also includes the DMA (Weka-based framework) made up by a set of DM algorithms adapted to multi-agent environment.

We tested AGMI as a proposed solution to the problem of detecting cartels in public biddings. Two experiments were conducted and the results proved that AGMI is a potential solution for the problem. Several association rules indicating evidences of cartel acting in public biddings were found. Besides, the combination of DM techniques enabled an increase of 150% in the quality of the top ten association rules.

For future work, we will study the inclusion of other DM techniques and the reduction of the overall processing time, through different data preparation methods and its automation. We are also studying other mechanisms and heuristics to use, in order to improve our agents interaction protocol, and to give more autonomy to coordinator and evaluator agents during the KDD process.

## REFERENCES

Bellifemine, F. L., Caire, G., and Greenwood, D. (2007). *Developing Multi-Agent Systems with JADE*. Wiley.

Cao, L. (2009). Introduction to agent mining interaction and integration. In Cao, L., editor, *Data Mining and Multi-agent Integration*. Springer US.

Han, J. and Kamber, M. (2005). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ralha, C. G. (2009). Towards the integration of multiagent applications and data mining. In Cao, L., editor, *Data Mining and Multi-agent Integration*. Springer US.

Silva, C. V. S. and Ralha, C. G. (2010). Utilização de técnicas de mineração de dados como auxílio na detecção de cartéis em licitações. In *XXX Congresso da Sociedade Brasileira de Computação (SBC/WCGE)*, Porto Alegre, RS, Brazil.