Expert Systems with Applications 39 (2012) 11642–11656

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A multi-agent data mining system for cartel detection in Brazilian government procurement

Célia Ghedini Ralha [a,*], Carlos Vinícius Sarmento Silva [a,b]

[a] Computer Science Department, University of Brasília, P.O. Box 4466, 70.904-970 Brasília, Brazil
[b] Controladoria-Geral da União, SAS, Qd 01, Bl A, Ed. Darcy Ribeiro, 70.070-905 Brasília, Brazil

ARTICLE INFO

Keywords:
Multi-agent data mining system
Cartel detection
Brazilian government procurement
AGMI
Multi-agent
Distributed data mining
Database knowledge discovery

ABSTRACT

The main focus of this research project is the problem of extracting useful information from the Brazilian federal procurement process databases used by government auditors in the process of corruption detection and prevention to identify cartel formation among applicants. Extracting useful information to enhance cartel detection is a complex problem from many perspectives due to the large volume of data used to correlate information and the dynamic and diversified strategies companies use to hide their fraudulent operations. To attack the problem of data volume, we have used two data mining model functions, clustering and association rules, and a multi-agent approach to address the dynamic strategies of companies that are involved in cartel formation. To integrate both solutions, we have developed AGMI, an agent-mining tool that was validated using real data from the Brazilian Office of the Comptroller General, an institution of government auditing, where several measures are currently used to prevent and fight corruption. Our approach resulted in explicit knowledge discovery because AGMI presented many association rules that provided a 90% correct identification of cartel formation, according to expert assessment. According to auditing specialists, the extracted knowledge could help in the detection, prevention and monitoring of cartels that act in public procurement processes.

1. Introduction

As the massive volume of data stored in distributed infrastructures (e.g., data warehouses) continues to grow, useful analyses of data in connection with government audits have become increasingly difficult. In addition, traditional centralised data mining (DM) methods no longer meet auditors' needs because of security, privacy, data confidentiality and infrastructure limitations (such as network bandwidth). Addressing this problem means utilising different computer science fields, such as distributed data mining (DDM), knowledge discovery in databases (KDD), multi-agent systems (MAS) and other distributed computing technologies (e.g., grid, cloud, etc.). To address this issue, the main goal of this research project is to find a technological solution to the problem of generating and extracting useful knowledge from the vast amounts of information used by auditors in the process of corruption prevention. Thus, this research applies two increasingly interrelated research areas, MAS and DDM/KDD. The motivation for using MAS is that collective intelligence has to be developed through the analysis of DDM because the underlying task is not completely decomposable and/or the computing resources are limited. MAS is particularly useful when applied to DDM/KDD in the context of sharing resources and expertise. Furthermore, because KDD is concerned with extracting hidden knowledge from data, including data that is widely distributed in many different forms and in multiple databases, the integration of MAS offers an even greater advantage.

The integration of agent technology (including individual agents, MAS and societies) into DDM technologies expanded into the agent-mining interaction and integration (AMII) research field, where complementary benefits are sought from both communities (Cao, 2009; Ralha, 2009). In the literature, we find many different approaches to this integration that can be bidirectional: (i) the agent-driven DM approach, where DM with a number of discrete and dependent tasks can use agents to regulate, control and organise potentially distributed activities in the knowledge discovery process, and (ii) the DM-driven agent approach, where the discovered knowledge items can constitute building blocks for agent intelligence.

This paper presents AGMI, an agent-mining tool that focuses the AMII field in two directions, uniting an agent-driven DM approach with a DM-driven agent approach when the discovered knowledge develops agent's intelligence and ability to mine databases through validated rules. As proof of concept, AGMI is utilised in a test case in an audit of the domain of the Brazilian government procurement, where the goal of the audit is to detect cartel formation in government procurement. We believe that the conceptual framework of

* Corresponding author. Tel.: +50 61 3107 0468.
E-mail addresses: ghedini@cic.unb.br (C. Ghedini Ralha), carlos.silva@cgu.gov.br (C.V. Sarmento Silva).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.04.037

AGMI is also useful in other domains, as we discuss further in Section 6. The rest of the paper is divided as follows. Section 2 discusses research motivation, while Section 3 briefly presents the MAS and DM areas of research. Section 4 presents an overview of the related work. Section 5 presents the AGMI tool, including its conceptual framework, implementation process and knowledge evaluation aspects. Section 6 outlines the use of AGMI in the Brazilian government procurement test case. Section 7 discusses important aspects of the research, and Section 8 concludes the paper and identifies directions for future work.

2. Motivation

Currently, enormous volumes of data are being produced and stored in computer systems around the world. The Brazilian Government Information Systems (BGIS) are a good example of this type of computerisation of an enormous volume of data. For example, in 2009, one BGIS, the Integrated Financial Management System of the Federal Government (SIAFI – Sistema Integrado de Administração Financeira do Governo Federal), registered and stored one billion financial transactions.[1] In terms of growth, in 1998, the average number of transactions (data searches and insertions) executed monthly at SIAFI was 39 million. This number reached 63.8 million by 2002 and 87 million in 2009, resulting in a 223% increase over the last 10 years.

All these financial data are used to support the preparation and execution of government auditing. The Office of the Comptroller General (CGU - Controladoria Geral da União)[2] is the agency of the Federal Government of Brazil that assists the president in defending public assets and enhancing management transparency through internal control activities, public audits, corrective and disciplinary measures, corruption prevention and combat and coordinating ombudsman activities. As the internal audit unit and the anti-corruption agency of the Brazilian Federal Government, CGU implements the important action of corruption prevention by applying different technologies to promote transparency and prevent corruption.

Our research has focused on the formation of cartels in the public procurement processes in Brazil. Identifying cartels is difficult because it requires the analysis of several public bidding processes, which usually exceeds the scope of a single government agency. Cartels can operate in various government departments, cities and even states of the Federation, which demands the sophisticated analysis of massive datasets.

In addition, the analysis of data from databases typically proceeds with languages using queries such as those found in the Structured Query Language (SQL),[3] which renders analysis impractical because of the exponential solution space. Therefore, auditing activities involved in detecting cartels are generally limited to confirming suspicions (normally after denunciation), given the difficulties inherent in the process of cartel detection.

Consider the goal of analysing all the possible groups of companies for cartel detection. We have to prepare all the possible combination of companies with at least two companies, as shown in Eq. (1), where n is the total number of companies in the database.

$$\sum_{i=2}^{n} C_i^n = 2^n - n - 1 \tag{1}$$

Thus, the brute-force algorithm runs in $O(2^n)$, and there is no deterministic way to identify cartels effectively because the solution space is exponentially related to the number of companies that participate in the bidding processes being analysed. Thus, the problem consists of creating an efficient process to identify groups of companies that might be suspected of practicing cartels in public procurement processes.

Because the volume of data to be treated is very large and the solution space is exponential, DM techniques are adequate to address the problem of analysing and understanding the massive datasets. However, the DM process demands a substantial amount of work to prepare the data, especially considering the use of different DM algorithms. Furthermore, DM algorithms alone cannot address the problem of workload distribution, considering the runtime execution: for example, through the use of parallel execution, which is important when handling massive datasets. Therefore, together with DM techniques, the MAS approach is important because it speeds up the execution time in a distributed and parallel manner. It is also essential to allow the use of rational agents to prepare data for DM with different algorithms and with the autonomy to analyse and improve the algorithms' results.

Table 1
Database used in the experiment.

Information	Number
Records	26,615
Procurement processes	2701
Companies	3051
Companies with at least 1 victory	1162
Companies with at least 5 victories	121

2.1. Initial problem analysis

When we first considered the problem of cartel detection in public procurement processes, we proposed a solution using association rule mining (ARM) because this technique is useful for finding strong relationships between attributes. The database used in this study was the Brazilian Federal Procurement system ComprasNet.[4] ComprasNet is a large database with information on all the procurement processes of the various types of services that are contracted by the Federal Executive agencies in all twenty-seven states of the Federation (Federation Unit – FU) of Brazil. Table 1 presents information about the dataset used in our experiments. This dataset records all the procurement processes to contract a specific type of service between the years 2005 and 2008. Each record in the dataset represents one bid from one company in the procurement process.

Most procurement processes consist of the following basic attributes: (i) participant companies or suppliers and their bids; (ii) negotiated object or service; (iii) government agency; (iv) city/state; and (v) winner. Table 2 describes the ComprasNet data related to the last procurement of each participant (DF – Federal District and MG – state of Minas Gerais).

To apply the ARM technique to associate only companies/suppliers, we have created a new dataset from the Table 2 data, which is presented in Table 3. Note that each column A, B, C, D, E, F, G represents supplier participation in each procurement process through a Boolean attribute, which registers a company's participation or nonparticipation in the procurement process.

The dataset for the ARM technique must be constructed as the matrix A, consisting of m rows and n columns:

$$m = (\text{total number of procurement processes from the database})$$
$$n = (\text{total number of companies from the database})$$

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

[1] The SIAFI official site – http://www.tesouro.fazenda.gov.br/siafi/index_mapa.asp.
[2] The CGU official site -http://www.cgu.gov.br/english/default.asp.
[3] MySQL official site – http://www.mysql.com/.
[4] The ComprasNet official site - http://www.comprasnet.gov.br/.