Revista Eletrônica de Sistemas de Informação
ISSN 1677-3071

v. 10, n. 1

2011 - Edição temática sobre governo eletrônico

doi:10.5329/RESI.2011.1001

Sumário

(cc) BY
Este trabalho está licenciado sob uma Licença Creative Commons Attribution 3.0 .

ISSN: 1677-3071

Esta revista é (e sempre foi) eletrônica para ajudar a proteger o meio ambiente, mas, caso deseje imprimir esse artigo, saiba que ele foi editorado com uma fonte mais ecológica, a *Eco Sans*, que gasta menos tinta.

DETECÇÃO DE CARTÉIS EM LICITAÇÕES PÚBLICAS COM AGENTES DE MINERAÇÃO DE DADOS

DETECTION OF CARTEL FORMATION IN GOVERNMENT BIDDINGS USING DATA MINING AGENTS

(artigo submetido em setembro de 2010)

Carlos Vinícius Sarmento Silva
Controladoria-Geral da União (CGU)
SAS, Qd 01, Bl A, Edifício Darcy Ribeiro, CEP 70.070-905, Brasília/DF, Brasil
carlos.silva@cgu.gov.br

Célia Ghedini Ralha
Departamento de Ciência da Computação – Universidade de Brasília (UnB)
Caixa Postal 4466, CEP 70.904-970, Brasília/DF, Brasil
ghedini@cic.unb.br

ABSTRACT

The Office of the Comptroller General (CGU), as the central agency of Brazil's Federal Government Internal Control is responsible for the fiscalization and auditing to fight and prevent corruption. However, some activities such as government purchasing fraud detection are limited by the difficulty of finding effective solutions, considering the huge volume of data, with millions of finantial registers. In such a context, the proccess of knowledge discovery may take advantage of Data Mining techniques, including classification, clusterization and association rules; which associated to multiagent system enrich the processing power through the interation and distribuiton of data mining agents. Thus, this research work used data mining agents with association rules and clusterization techniques to identify cartels, acting in fraud detection. As a research result, more than one hundred association rules were discovered, of which ten have strong evidence of cartelization, proving the usefulness of the approach to support the work of government auditing.

Key-words: government auditing; data mining; multi-agent system, data mining agents.

RESUMO

A Controladoria-Geral da União (CGU), como órgão central do sistema de controle interno do Poder Executivo Federal do Brasil é responsável pela realização de atividades de auditoria e fiscalização, visando à prevenção e o combate à corrupção. No entanto, algumas atividades como a detecção de cartéis em licitações é limitada, pela dificuldade de encontrar soluções efetivas em grande volume de bases de dados com milhões de registros de transações financeiras. Nesta seara, algumas áreas de Ciência da Computação apresentam bons resultados no processo de descoberta de conhecimento com uso de técnicas de mineração de dados, tais como classificação, clusterização e regras de associação, as quais, associadas à área de Sistema Multiagente, ampliam o poder de processamento de forma distribuída e interativa com agentes de mineração de dados. Neste sentido, esta pesquisa utiliza agentes de mineração de dados com regras de associação e clusterização para a solução do problema de detecção de cartéis em licitações. Como resultado da pesquisa foram descobertas mais de cem regras de associação, das quais dez apresentam fortes indícios de cartelização, comprovando a utilidade da abordagem como suporte ao trabalho de auditoria governamental.

Palavras-chave: auditoria governamental, mineração de dados, sistema multiagente, agentes de mineração.

1 INTRODUÇÃO

Atualmente, a Administração Pública Federal (APF) no Brasil mantém a maioria de seus processos apoiados por sistemas computacionais. Um exemplo de grande importância no cenário nacional é o Sistema Integrado de Administração Financeira do Governo Federal (SIAFI), o qual registrou no ano de 2009 um bilhão de transações financeiras com 24 mil unidades gestoras. Outro exemplo vital do atual governo foi o Portal da Transparência, criado e mantido pela Controladoria-Geral da União (CGU), o qual mantém atualmente mais de um bilhão de registros que totalizam cerca de 7,6 trilhões de reais em gastos do Governo Federal, tais como transferências de recursos, gastos diretos e uso de cartões corporativos (CGU/PR, 2010 e 2011).

No presente cenário governamental o volume de dados produzidos e armazenados pelos diversos sistemas de computação tem aumentado expressivamente, sendo a informatização dos diversos setores do governo a causa primária do aumento na produção de dados digitais do país. Os dados provenientes dos sistemas computacionais federais, especialmente do SIAFI, são utilizados pelos órgãos de auditoria governamental para planejamento e execução de suas tarefas de fiscalização e auditagem de aplicação dos recursos públicos. Podemos encontrar estatísticas sobre o uso do SIAFI no site da Secretaria do Tesouro Nacional (STN, 2011).

No âmbito do Poder Executivo Federal, a CGU tem direcionado esforços no sentido de utilizar tecnologias em análises de dados para desenvolvimento de ações voltadas à promoção da transparência e à prevenção da corrupção. A maior dificuldade, porém, reside em correlacionar esses dados para geração de conhecimento útil para os auditores. Desta forma, as alternativas atualmente se restringem a consultas aos sistemas em casos pontuais ou preparação de amostras estatísticas que diminuem o universo para um conjunto reduzido de informações proporcional à capacidade operacional do Órgão.

Para lidar com grandes volumes de dados, a utilização de técnicas de mineração de dados (MD) tem se mostrado de grande valia na obtenção de informações e no processo de descoberta de conhecimento (WITTEN e FRANK, 2005). Estas técnicas pertencem a um ramo da Ciência da Computação conhecido como Descoberta de Conhecimento em Base de Dados ou *Knowledge Discovery in Database* (KDD), a qual, associada à subárea de Sistema Multiagente (SMA), tem se apresentado como abordagem útil no processamento distribuído de grandes bases de dados com uso de agentes de mineração de dados (CAO, 2009).

O uso de agentes de mineração de dados para o problema encontrado no trabalho de auditoria, especificamente em análises de processos de licitação, procurando identificar a ocorrência de rodízio de licitação, se mostra de fundamental importância. Ressaltamos que a prática de rodízio é ilegal e traz prejuízos significativos ao Erário, sendo que a detecção de grupos suspeitos de praticar rodízio de licitação é bastante difícil, não havendo formas determinísticas que auxiliem eficazmente nesta tarefa.

doi:10.5329/RESI.2011.1001008

Neste artigo, apresentamos a aplicação de agentes de mineração de dados para descoberta de regras que possibilitem a detecção de cartéis em licitações públicas. A descoberta automática de regras baseada nos padrões comportamentais dos envolvidos nas licitações podem viabilizar o direcionamento eficaz do trabalho dos auditores públicos.

O artigo está estruturado da seguinte forma: a Seção 2 apresenta as definições necessárias para melhor compreensão do problema; a Seção 3 expõe o problema de detecção de cartéis em licitações públicas; a Seção 4 propõe a solução com uso de agentes de mineração com técnicas de regras de associação e clusterização; a Seção 5 apresenta os experimentos realizados com mineração de dados; enquanto a Seção 6 utiliza agentes de mineração para realização das mesmas tarefas; a Seção 7 avalia os resultados e a descoberta de conhecimento; e finalmente, a Seção 8 apresenta as conclusões e fornece sugestões de trabalhos futuros.

2 CONCEITOS IMPORTANTES

O Conselho Administrativo de Defesa Econômica (CADE) é a Autarquia responsável por investigar e punir as empresas que se unem na prática do cartel. Essa prática configura tanto ilícito administrativo punível pelo CADE, nos termos da Lei nº 8.884/94, quanto crime, punível com pena de 2 a 5 anos de reclusão, nos termos da Lei nº 8.137/90 (JUSTIÇA, 2010).

Licitação pode ser entendida como um procedimento democrático para se contratar com o poder público, sendo que este procedimento tem como objetivo garantir a observância do princípio constitucional da isonomia, selecionando a proposta mais vantajosa para a APF (BRASIL, 1993).

Segundo Di Pietro (2009), licitação é o procedimento administrativo pelo qual um ente público abre a todos os interessados, que se sujeitem às condições fixadas no instrumento convocatório, a possibilidade de formularem propostas dentre as quais selecionará a mais conveniente para a celebração do contrato. Como licitação é o meio mais comum de realização de despesa pública, as atividades de auditoria governamental dão especial atenção à análise dos processos de licitação e contrato. Essa preocupação se dá pelo fato de que o envolvimento de recursos financeiros possibilita e até mesmo incita a criação de esquemas ilícitos.

Dentre esses esquemas ilícitos está a formação de cartel, um acordo explícito ou implícito entre concorrentes que visa, principalmente, à fixação de preços ou quotas de produção, à divisão de clientes e de mercados de atuação. O objetivo dos participantes do cartel é, por meio da ação coordenada entre concorrentes, eliminar a concorrência, com o consequente aumento de preços e redução de bem-estar para o consumidor. O rodízio de licitações é uma prática típica da atuação em cartel e se dá quando um grupo de empresas se organiza criminosamente com a finalidade de dividir as licitações entre si, elevando o preço de contratação com a APF, trazendo, consequentemente, danos ao Erário.

A CGU, como Órgão Central de Controle Interno, mantém parceria com o CADE para que as investigações de prática de cartéis no âmbito da Administração Pública sejam mais eficientes. Dessa forma, sempre que a CGU encontra indícios de práticas de rodízio de licitações em suas auditorias, o processo pode ser encaminhado ao CADE para que este tome as providências cabíveis. Independente da decisão do CADE, a CGU também pode punir as empresas suspeitas através da Declaração de Inidoneidade, que as impede imediatamente de participar de licitações e contratar com a APF. Os processos licitatórios eivados deste vício poderão também ser anulados pela CGU, baseada no artigo 90 da Lei 8.666/93, que reconhece como crime o ato de "frustrar ou fraudar, mediante ajuste, combinação ou qualquer outro expediente, o caráter competitivo do procedimento licitatório, com o intuito de obter, para si ou para outrem, vantagem decorrente da adjudicação do objeto da licitação" (BRASIL, 1993).

2.1 MINERAÇÃO DE DADOS

É importante distinguir o que é uma tarefa e o que é uma técnica de mineração. A tarefa consiste na especificação do que queremos buscar nos dados, que tipo de regularidades ou categoria de padrões quer-se encontrar. Já a técnica de mineração consiste na especificação de métodos que nos garantam como descobrir os padrões que nos interessam. Dentre as principais técnicas utilizadas em mineração de dados, temos técnicas estatísticas, técnicas de aprendizado de máquina e técnicas baseadas em crescimento-poda validação. A seguir, descrevemos de forma sucinta as principais tarefas e técnicas de mineração relacionadas a este trabalho.

Segundo Tan *et al.* (2005), as tarefas de MD são geralmente divididas em duas categorias principais, a de tarefas preditivas e a de tarefas descritivas. A primeira objetiva predizer o valor de um atributo particular baseado nos valores de outros atributos. O atributo a ser predito é conhecido como *alvo* ou *variável dependente*, enquanto os atributos usados para fazer a predição são conhecidos como *explanatórios* ou *variáveis independentes*. Já a segunda categoria tem como objetivo derivar padrões como correlações, tendências, grupos, trajetórias e anomalias as quais sumarizam as relações subjacentes nos dados.

Na definição de Frawley *et al.* (1992), KDD é uma extração não trivial de informações implícitas, previamente desconhecidas e potencialmente úteis de uma base de dados. Dessa forma, a aplicação de KDD tem sido utilizada em diversas áreas tanto no campo da pesquisa, quanto no dos negócios e também nas esferas governamentais (FAYYAD *et al.*, 1999). No processo de KDD, a MD tem o papel de extrair padrões ou modelos dos dados através da aplicação de algoritmos específicos para esse fim.

Existem diversas técnicas de MD, tais como classificação, clusterização, regras de associação, regras de sequência, regressão, sumarização, entre outras, as quais podem ser úteis no processo de KDD. A classifica-

ção, por exemplo, é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados.

Passaremos a detalhar melhor as tarefas descritivas com uso de clusterização e tarefas preditivas com regras de associação. A priorização destas escolhas baseia-se no fato de englobarmos as duas principais categorias de tarefas de MD, conforme definido por Tan *et al.* (2005), as quais demonstraram ser úteis ao problema de correlacionamento de informações no âmbito de rodízio de licitações.

Segundo Jain e Dubes (1988), a clusterização é a tarefa descritiva onde se procura identificar um conjunto finito de categorias ou *clusters* para descrever uma informação, podendo ser mutuamente exclusivas, ou não. No processo de clusterização os objetos são agrupados de acordo com suas similaridades.

A técnica de regras de associação, por sua vez, consiste em descobrir relações fortes entre determinados atributos. Busca detectar padrões em forma de regras que associam valores de atributos em um determinado conjunto de dados. Essas regras são expressas em forma de conjunções do seguinte tipo:

$$a1=v1, a2=v2, ..., am=vm \rightarrow am+1=vm+1, am+2=vm+2, ..., an=vn$$

onde *a* é um atributo do conjunto de dados e *v* é o valor do atributo identificado na regra.

Segundo Han e Kamber (2006), a técnica de regras de associação diferencia-se da técnica de classificação na capacidade que aquela tem de prever padrões com qualquer atributo e não só com a classe selecionada. Diferentes regras de associação expressam diferentes regularidades subjacentes no conjunto de dados, cada uma predizendo coisas diferentes. A qualidade das regras obtidas por essa técnica é medida pelo suporte e confiança. O suporte da regra é a probabilidade da regra se repetir no conjunto de dados. A confiança da regra é o percentual de instâncias preditas corretamente pela regra. Destaca-se como um dos algoritmos mais populares para aplicação dessa técnica o *Apriori*, apresentado em Agrawal e Srikant (1994).

Segue a definição formal de regras de associação, segundo Han e Kamber (2006):

Seja $I = \{I_1, I_2, ..., I_M\}$ um conjunto de itens e *D* os dados da base contendo transações formadas por itens do conjunto *I*. Sejam também *A* e *B* conjuntos de itens. Uma regra de associação é uma implicação da forma $A \rightarrow B$ onde $A \subset I$, $B \subset I$, e $A \cap B = \varphi$. A regra $A \rightarrow B$ se aplica no conjunto de transações *D* com suporte *s*, onde *s* é o percentual de transações em *D* que contém $A \cup B$, isto é, a probabilidade $P(A \cup B)$. A regra $A \rightarrow B$ tem confiança *c* no conjunto de transações *D*, onde *c* é o percentual de transa-

ções em D contendo A que também contém B, isto é, a probabilidade condicional $P(B|A)$.

2.2 SISTEMA MULTIAGENTE

Segundo Wooldridge (2009), um agente é uma entidade computacional capaz de perceber o ambiente onde está e agir autonomamente sobre ele com o objetivo de alcançar metas específicas. Podemos citar como características de um agente inteligente: autonomia, reação, interação e iniciativa.

Wooldridge e Jennings (1995) citam como capacidades especiais de agentes inteligentes a reação, proatividade e capacidade social, conforme segue:

- reação - agentes inteligentes são capazes de perceber seus ambientes, e responder às mudanças que ocorrem neles em tempo hábil no intuito de satisfazer os seus objetivos projetados.
- proatividade - agentes inteligentes são hábeis para tomar decisões no intuito de satisfazer os seus objetivos projetados.
- capacidade social - agentes inteligentes são capazes de interagir com outros agentes (e possivelmente humanos) no intuito de satisfazer os seus objetivos projetados.

De acordo com suas características e capacidades específicas, existem diversos tipos de agentes definidos na literatura. Wooldridge (2009) fala de agentes puramente reativos, agentes com estados, agentes orientados a objetivo e agentes orientados por utilidades.

Um SMA é um sistema em que muitos agentes interagem e agem em um ambiente por meio de comunicação e coordenação. Os agentes podem decidir cooperar por benefícios mútuos ou podem competir para atender seus próprios interesses (RUSSEL e NORVIG, 2010). A vantagem de utilizar um SMA abrange aspectos tais como a distribuição de recursos e controle, descentralização dos dados, comunicação assíncrona, entre outras.

2.3 AGENTES DE MINERAÇÃO

As áreas de agentes inteligentes e MD se desenvolveram separadamente. Enquanto o estudo dos agentes inteligentes visava o comportamento autônomo e independente dos agentes, a MD, de forma mais abrangente, lidava com a descoberta de conhecimento em grandes bases de dados. A princípio, as duas áreas seguem com metas e objetivos distintos, porém, há vários aspectos de ambas as áreas que coincidem, tais como interação usuário-sistema, papéis humanos, modelagem dinâmica, fatores de domínio, fatores organizacionais e sociais, entre outros. Por causa disso, ambas as áreas potencializam o avanço da Inteligência Artificial, o processamento de informações, serviços e sistemas inteligentes, possibilitando claramente a interação e integração entre agentes e MD, e esta integração é conhecida como Integração e Interação

entre Agentes de Mineração (*Agent Mining Interaction and Integration - AMII*) (CAO *et al.*, 2009).

Segundo Cao *et al.* (2009), a interação entre MD e agentes, onde o processo de descoberta de conhecimento é auxiliado e enriquecido por agentes de softwares inteligentes, é conhecida *como Multi-agent-Driven Data Mining*. Esta integração pode auxiliar em vários aspectos o processo de descoberta de conhecimento, tais como problemas de administração dos dados, interação e supervisão humana no processo de MD, seleção de dados, enriquecimento do conhecimento através da combinação de técnicas distintas de MD, entre outros.

3 PROBLEMA

O problema se resume em identificar de forma eficaz e eficiente grupos de empresas suspeitos de praticar cartéis em licitações públicas. A identificação de cartéis em licitações analisando bases de dados é uma tarefa bastante difícil. Isso se dá porque a análise das combinações das empresas numa base de dados é de complexidade exponencial. Assim, a utilização de linguagens de consultas tais como SQL (*Structured Query Language*) é impraticável na solução desse problema. Desta forma, as atividades de auditoria se limitam somente à confirmação de suspeitas normalmente levantadas após denúncias.

Além disso, a atuação de um cartel normalmente extrapola o escopo de apenas um órgão da APF. Cartéis podem atuar em vários órgãos, cidades, e até mesmo estados da Federação, necessitando de cruzamento de dados para análise computacional, o que exclui soluções determinísticas eficientes e eficazes para solução do problema.

4 PROPOSTA DE SOLUÇÃO

A proposta de solução para o problema definido se dá através da utilização da técnica de regras de associação. Isso se deve ao fato de essa técnica ser útil para encontrar relações fortes entre atributos. O problema de detecção de grupos de empresas suspeitos de praticar rodízio em licitações pode então ser adaptado de forma que cada processo licitatório se torne um registro em uma base de dados, tendo como atributos as empresas participantes daquela licitação.

A estratégia usada para procurar associação entre empresas é organizar os *datasets* de forma que cada fornecedor da base de dados – empresa participante de licitação - seja um atributo booleano e cada instância seja um processo de licitação. Assim, para cada licitação, o atributo relativo a um determinado fornecedor é preenchido com o valor 'sim', caso aquele fornecedor tenha participado do certame, ou 'não', caso contrário.

A preparação dos *datasets* para regras de associação se resume então em construir a matriz *A* formada por *m* linhas e *n+1* colunas tal que:

m = número total de licitações da base de dados

n = número total de fornecedores da base de dados

$$a_{i,j} = \begin{cases} \text{sim, se fornecedor } j \text{ participou da licitação } i \\ \text{não, se fornecedor } j \text{ não participou da licitação } i \end{cases}$$

$a_{i,n+1}$= vencedor(i)

Tal que $1 \leq i \leq m$; $1 \leq j \leq n$;

vencedor(i) = CNPJ da empresa vencedora da licitação *i*

Dessa forma, espera-se obter regras do tipo:

fornecedorA = sim, fornecedorB = sim ➔ *vencedor = fornecedorC*

O preenchimento da coluna de vencedores pode ser também eliminado produzindo regras do tipo:

fornecedorA = sim, fornecedorB = sim ➔ *fornecedorC = sim*

5 EXPERIMENTOS COM MINERAÇÃO DE DADOS

Os experimentos com MD envolvendo a matriz descrita no item anterior foram recentemente apresentados em Silva e Ralha (2010). A base de dados utilizada para realização dos experimentos foi extraída do sistema *ComprasNet*, por meio do qual são realizados os pregões eletrônicos do Governo Federal (MPOG, 2010). Esses dados são relativos a todas as licitações para contratação de um determinado tipo de serviço na modalidade Pregão para órgãos do Poder Executivo Federal entre os anos de 2005 e 2008, em todos os estados da Federação. Cada registro da base de dados representa a participação de uma empresa em uma licitação. Informações sobre a base de dados utilizada nos experimentos são apresentadas na Tabela 1.

A ferramenta utilizada para execução dos algoritmos de MD foi o *Weka*, em sua versão 3.6.1. Este software foi desenvolvido na Universidade de Waikato na Nova Zelândia e foi escolhido pelos vários algoritmos de MD que disponibiliza (WAIKATO, 2010).

Tabela 1. Base de dados utilizada nos experimentos preliminares.

Informações	Total
Registros	26615
Licitações	2701
Empresas	3051
Empresas que já ganharam pelo menos 1 licitação	1162
Empresas que já ganharam pelo menos 5 licitações	121

Fonte: dados da pesquisa.

5.1 APLICAÇÃO DA TÉCNICA DE REGRAS DE ASSOCIAÇÃO

Foram preparados dois *datasets* no intuito de aplicar as técnicas de regras de associação para detecção de grupos suspeitos de fazer rodízio de licitações. Foi escolhido o algoritmo *Apriori*, por ser um algoritmo seminal da técnica de regras de associação, além de ser um dos melhores algoritmos para esta técnica de MD, conforme Wu *et al.* (2007). O algoritmo evita a explosão combinatória de regras baseando-se no princípio de que se um item não é frequente, nenhum de seus superconjuntos será tampouco frequente.

O *dataset* inicial foi construído contemplando todas as licitações da base e todos os fornecedores. Já o segundo *dataset* contemplou apenas os fornecedores que já tinham participado de pelo menos 2 licitações (Tabela 2). O filtro aplicado no segundo *dataset* foi útil no sentido de focar nas empresas que tinham mais probabilidade de estar envolvidas em algum cartel de licitação.

Regras que indicam a não participação de fornecedores não trazem, a princípio, nenhum resultado de interesse para o problema de rodízio de licitações (e. g.: fornecedor_A=não, fornecedor_B=sim ➔ forncedor_C=não), desta forma, os *datasets* para execução do algoritmo de regras de associação foram preparados para que os resultados trouxessem apenas regras que contemplassem a participação de fornecedores em processos de licitações. A Tabela 2 mostra o resultado da execução do algoritmo *Apriori* nos *datasets* preparados.

Tabela 2. Resultados da execução do *Apriori* para os dois *datasets.*

	Instâncias	Atributos	Suporte mínimo	Confiança mínima	Nº de regras
Dataset 1	2701	3051	1%	70%	294
Dataset 2	2370	1086	1%	80%	145

Fonte: dados da pesquisa.

Valores altos na configuração do suporte mínimo para execução do algoritmo não garantem boas regras para identificação de cartéis. O suporte alto de uma regra pode representar apenas a detecção de grandes fornecedores que participam com freqüência de processos de licitação. No entanto, a configuração de um suporte mínimo alto para execução do algoritmo pode suprimir a aparição de diversas regras boas, com reais características de cartéis. Por isso foram definidos valores baixos de suporte na execução do algoritmo de regras de associação como mostra a Tabela 2. Essa diminuição no valor do suporte mínimo fez com que o número de regras aumentasse, necessitando de uma função de avaliação adicional para filtrar as melhores regras, que será apresentada na Seção 7.

5.2 APLICAÇÃO DA TÉCNICA DE CLUSTERIZAÇÃO

A aplicação de regras de associação em dados de todo o país deixou o espaço de soluções bastante esparso. Foi verificado que muitas vezes os

fornecedores não se restringem necessariamente às regiões macroeconômicas. Isso pode ser exemplificado pela situação dos estados de Mato Grosso do Sul, Goiás e Tocantins. Embora Mato Grosso do Sul e Goiás pertençam à mesma região, é mais provável que os fornecedores do estado de Goiás atendam o estado de Tocantins, por causa da proximidade geográfica, apesar de Goiás e Mato Grosso do Sul estarem na região Centro-Oeste e Tocantins na região Norte. Para resolver esse problema, foi aplicada a técnica de clusterização para mapear os grupos comuns de atuação dos fornecedores. O algoritmo utilizado para a descoberta não supervisionada de *clusters* foi o EM (*Expectation-Maximization*).

EM é uma extensão do paradigma *k-means*, que associa um objeto ao cluster que lhe é mais similar, baseado na média encontrada. O algoritmo pode associar objetos a mais de um *cluster* definindo a probabilidade daquele objeto pertencer aos *clusters* associados. Essa característica é interessante na análise de regionalização de mercados de licitações. Pode ser que um estado tenha 57% de chance de pertencer a uma região de licitação e 43% de chance de pertencer à outra. Ao se tratar de cartéis, a consideração das duas regiões na análise é importante. Por esse motivo, o algoritmo EM pareceu mais propício na descoberta de regiões geográficas formadas pelos fornecedores de licitações (HAN e KAMBER, 2006).

O algoritmo EM foi executado com as 26.615 instâncias da base, considerando os fornecedores que participaram das licitações nos estados onde elas ocorreram. O algoritmo apresentou como resultado 10 *clusters*, conforme apresentado na Figura 1. Note que a maioria dos *clusters* encontrados têm como característica a proximidade geográfica. A Tabela 3 mostra a distribuição das instâncias nos *clusters* encontrados.

Tabela 3. Distribuição das instâncias por *cluster.*

Cluster	Instâncias	Percentual
1	8473	32%
2	3552	13%
3	2641	19%
4	2077	8%
5	1197	4%
6	1335	5%
7	2687	10%
8	461	2%
9	1457	5%
10	2732	10%

Fonte: MPOG (2010).

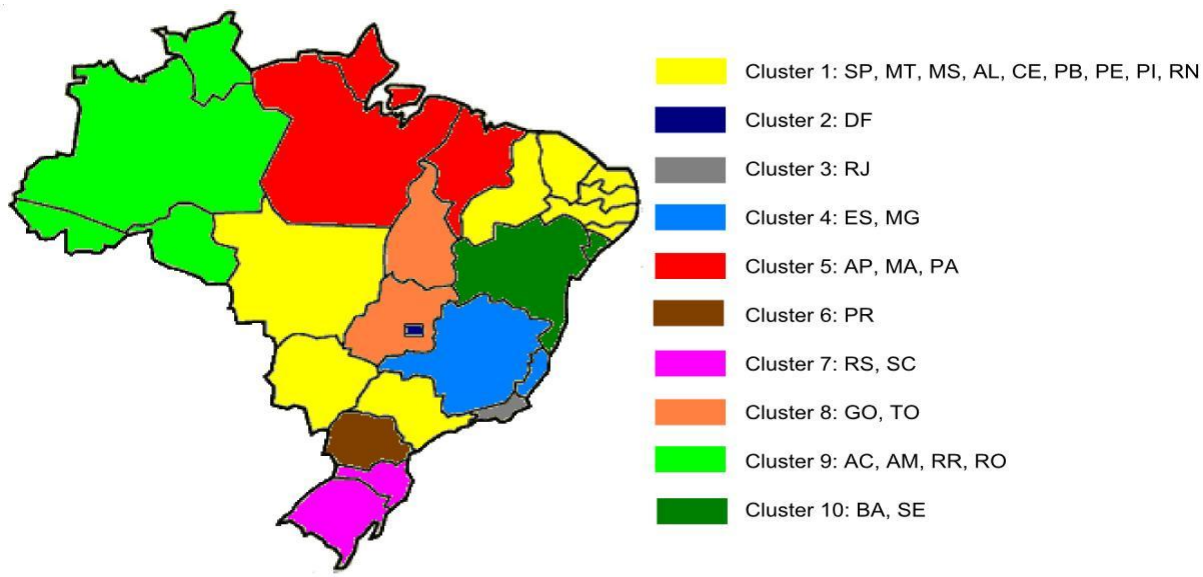

Figura 1. *Clusters* das Unidades da Federação descobertos no algoritmo EM.

Fonte: dados da pesquisa.

5.3 INTEGRAÇÃO DAS TÉCNICAS DE MINERAÇÃO DE DADOS

Foi executada novamente a técnica de regras de associação em cada região obtida através da clusterização, com o objetivo de detectar grupos suspeitos de praticar cartéis em regiões específicas de atuação. Os resultados deste experimento podem ser vistos na Tabela 4.

Tabela 4. Execução do *Apriori* para *datasets* de *clusters.*

Cluster	Instâncias	Atributos	Suporte mínimo	Confiança mínima	Regras
1	787	614	2%	80%	851
2	211	164	4%	80%	1406
3	261	166	3%	80%	100
4	194	257	5%	80%	86
5	134	168	6%	80%	115
6	98	152	9%	80%	2848
7	270	196	4%	80%	1679
8	94	118	1%	80%	3
9	211	204	4%	80%	22
10	134	259	10%	80%	5869

Fonte: MPOG (2010).

Note-se que foi mantido o grau de confiança mínima em 80%, sendo que o valor do suporte mínimo foi adaptado em cada *cluster* para que as regras obtidas na execução do algoritmo sejam válidas para pelo menos nove instâncias do *dataset*, isto é, para que o grupo de fornecedores apontado por uma regra de associação tenha atuado em pelo menos nove licitações. Desta forma, por exemplo, na região apontada pelo Cluster 1, formada pelos estados de SP, MT, MS, AL, CE, PB, PE, PI, e RN foi definido

um suporte mínimo de 2% do tamanho do *dataset* (787 instâncias), encontradas 851 regras; enquanto no Cluster 2, formado somente pelo estado do DF, o suporte mínimo foi de 4% do tamanho do *dataset* (211 instâncias), encontrando 1.406 regras, sendo que em ambos os casos houve pelo menos nove ocorrências da regra de associação no *dataset.*

6 UTILIZAÇÃO DE AGENTES DE MINERAÇÃO

Os experimentos de MD realizados até o momento foram demasiadamente trabalhosos, em especial na fase de pré-processamento e tratamento dos dados. Podemos citar como fatos importantes para este trabalho: (i) a solução do problema que necessitou baixos limites de suporte para a aplicação da técnica de regras de associação e (ii) os testes retornarem muitas regras, dificultando a fase de interpretação e avaliação dos resultados. Essa intensa carga de trabalho demandou dedicação exclusiva de um analista de mineração durante uma semana de trabalho, incluindo a fase de pré-processamento, processamento com uso da ferramenta Weka e análise dos dados, com a finalidade de garantir resultados com qualidade durante os experimentos.

Foi então adotada a abordagem de SMA com a definição e implementação de quatro agentes responsáveis pela realização das tarefas de preparação dos dados, coordenação da interação entre os agentes, execução de técnicas de mineração e avaliação dos resultados, a saber:

- Agente coordenador - responsável pelas tarefas de pré-processamento e preparação dos *datasets* de entrada para os algoritmos de MD. Gerencia também as interações entre os agentes e solicita serviços dos agentes responsáveis pela MD.
- Agentes mineradores - executam algoritmo de MD, sendo utilizados dois agentes, um para execução do algoritmo *Apriori* e outro para EM.
- Agente avaliador - avalia as regras obtidas armazenando-as ou não na base de conhecimento dependendo da sua qualidade. Mantém a base de conhecimento ordenada para apresentação dos resultados ao usuário.

Para a implementação dos agentes foi utilizado o *framework* JADE (*Java Agents Development Framework)*. JADE é um *framework* implementado em Java para auxílio no desenvolvimento de aplicações utilizando agentes. Além da portabilidade resguardada pela linguagem Java, o *framework* permite a implementação de SMA de forma distribuída, o que facilita no processamento de grandes quantidades de informações (BELLIFEMINE *et al.*, 2007). A distribuição dos agentes nos *hosts* é mostrada na Tabela 5.

Tabela 5. Distribuição dos agentes nos *hosts.*

Máquina A	Intel Core 2, 2.40 GHz, 2.00 GB RAM	
Máquina B	Intel Pentium Dual, 1.86 GHz, 1.99 GB RAM	
	Máquina A	**Máquina B**
Agente coordenador	X	
Agente clusterização	X	
Agente regras de associação		X
Agente avaliador		X

Fonte: dados da pesquisa.

Para execução dos algoritmos de MD, utilizou-se a biblioteca do Weka 3.6.1, com algumas alterações para facilitar a integração com os agentes. A execução do teste com uso de agentes de mineração teve duração de 74 minutos e 57 segundos. A grande redução no tempo comparado com a semana de trabalho do analista de mineração em dedicação exclusiva, refere-se principalmente à automatização da preparação dos dados e avaliação das regras. No entanto, a distribuição das tarefas, possibilitando o processamento paralelo dos algoritmos de mineração auxiliou na redução do tempo. A distribuição dos agentes em *hosts* diferentes, fez com que tarefas de clusterização e regras de associação, que antes eram executadas sequencialmente, pudessem ser executadas em paralelo, aproveitando os recursos disponíveis e poupando tempo. Os resultados comprovam a utilidade da junção de duas abordagens importantes como MD e SMA, com uso de agentes de MD para problemas de descoberta de conhecimento em grandes bases de dados.

7 AVALIAÇÃO DOS RESULTADOS

Foi definido um método de avaliação das regras obtidas através do processo de MD com auxílio de especialistas auditores da CGU. A Equação 1 apresenta a avaliação utilizada.

$$M(F) = 100 * V(F) / (Sup. * Inst.) \quad (1)$$

Onde:

- Sup. = valor do suporte da regra;
- Inst. = número de instâncias do *dataset;*
- F = conjunto de fornecedores que figuram na regra;
- V(F) = número de vezes que algum fornecedor do conjunto F venceu as licitações em que todo o grupo participou conjuntamente;

Normalmente, taxas altas de suporte e confiança determinam o grau de qualidade de uma regra de associação. No entanto, foi encontrada uma regra de associação que mostrava uma forte ligação entre três fornecedores atuando nas licitações de um *dataset*, o que poderia ser indício de

um rodízio de licitações. Quando verificamos na base real, em quase nenhuma das ocasiões em que esses três fornecedores participaram juntos de licitações, algum deles conseguiu fechar um contrato com a APF, suas participações nos mesmos processos licitatórios foram uma mera coincidência.

Ressaltamos que o uso de suporte e confiança como itens de mensuração na qualidade das regras resultantes não são suficientes. Desta forma, os especialistas inseriram a função V(F) como fator da função de avaliação. A função V(F) terá como valor máximo de seu resultado o próprio suporte da regra avaliada multiplicado pelo número de instâncias do *dataset*. Em outras palavras, a função de avaliação retornará a probabilidade do grupo identificado na regra vencer as licitações de que participa, retornando um valor entre 0 e 100.

Todas as regras foram avaliadas por meio da função V(F), sendo que para análise dos resultados foram selecionadas as dez melhores regras. Note-se na Figura 2 que as melhores regras considerando todas as Unidades da Federação que tiveram na média os melhores números de ocorrência (suporte multiplicado pelo número de instâncias). As dez melhores regras obtidas pela aplicação de regras de associação nas regiões definidas pelos *clusters* apresentaram um aumento de cerca de 100% no valor de avaliação em relação às primeiras.

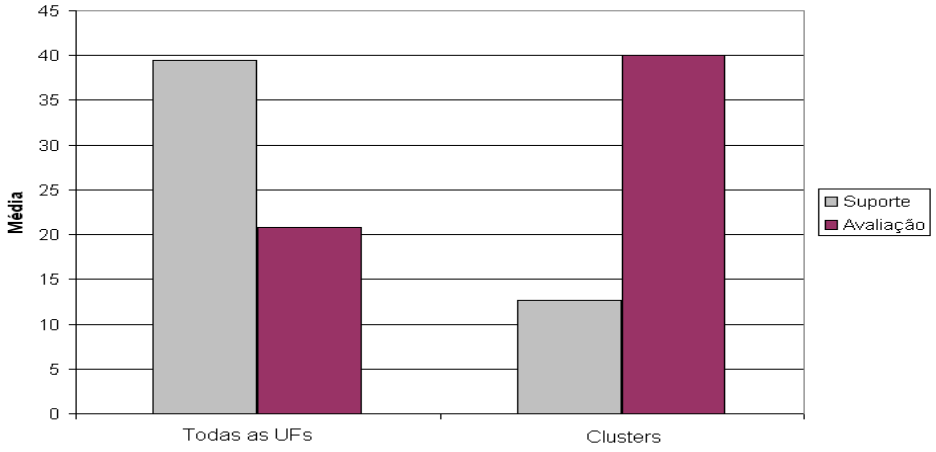

Figura 2. Média de suporte e avaliação das 10 melhores regras

Fonte: Silva e Ralha (2010).

Esse resultado mostra que, segundo a avaliação adotada, as melhores regras na nossa base tendem a aparecer quando o suporte é baixo e quando há uma melhor definição do espaço de soluções - nesse caso, definido pelos *clusters* encontrados. Por isso, as regras que abrangem o Brasil todo não foram tão boas quanto àquelas encontradas em regiões do país.

O gráfico da Figura 3 apresenta a comparação entre as dez melhores regras obtidas nos modelos dos *clusters* (Tabela 4). Note-se que as melho-

res regras foram obtidas no Cluster 6, conforme avaliação das regras aplicando a Equação 1, com valores de 70 a 90 de um total de 100.

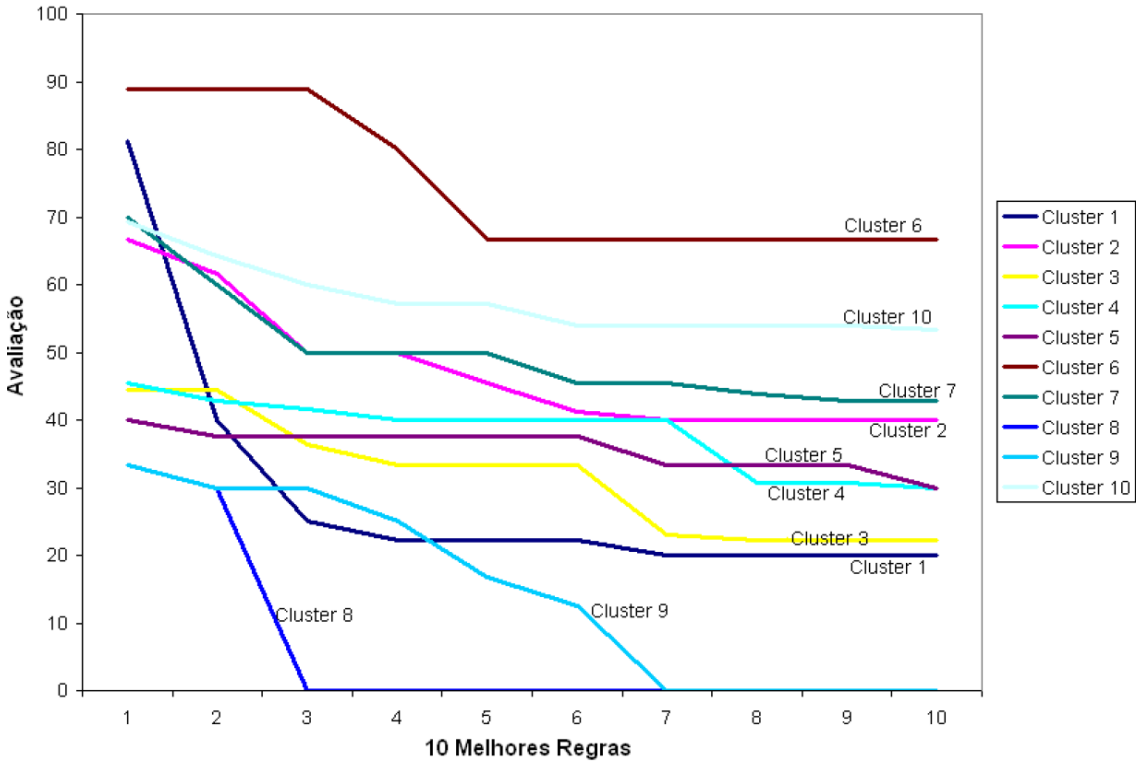

Figura 3. Comparação das melhores regras dos modelos de *cluster* (SILVA e RALHA, 2010).

Fonte: dados da pesquisa.

7.1 CONHECIMENTO DESCOBERTO

Através da realização dos diversos experimentos, chegamos à conclusão que a técnica de MD que gerou maior interesse por parte dos especialistas auditores da CGU foi a de clusterização, pois segundo os especialistas as atividades de rodízio de licitações são tipicamente regionais. Isso significa que mesmo que uma empresa tenha atuação em âmbito nacional e pratique rodízio de licitações com um grupo, é improvável que esse grupo atue em todo o país. Desta forma, a regra que apresenta uma associação de fornecedores em provável conluio teria maior suporte em apenas uma região, que seria a região de atuação do cartel.

Verificamos que o Cluster 1, apresentado na Figura 3, trouxe um resultado interessante por fugir do padrão de regionalização geográfica. Os estados de São Paulo, Mato Grosso e Mato Grosso do Sul se agruparam com os estados de Alagoas, Ceará, Paraíba, Pernambuco, Piauí e Rio Grande do Norte. Esse resultado trouxe outras propostas de pesquisa no intuito de levantar, dentre as empresas que atuaram nesses estados, quais delas contribuíram para essa distribuição atípica nas participações em licitações.

Um rápido levantamento mostrou 76 empresas que atuaram na sub-região formada pelos estados de Alagoas, Ceará, Paraíba, Pernambuco, Piauí, Rio Grande do Norte e na sub-região formada pelos estados de São Paulo, Mato Grosso e Mato Grosso do Sul. Dessas empresas, 8 participaram em mais de 15 licitações, tanto numa sub-região quanto na outra. Dessas 8 empresas, nenhuma é da sub-região composta por São Paulo, Mato Grosso ou Mato Grosso do Sul.

Considerando os resultados favoráveis desta pesquisa imaginamos que novas bases de dados poderão ser utilizadas em experimentações futuras, com o intuito de detectar outros *clusters* de interesse para investigações; como por exemplo, *clusters* envolvendo órgãos superiores da APF (e.g., os gabinetes, as secretarias-gerais, as inspetorias-gerais, as procuradorias administrativas e judiciais).

Quanto às regras de associação obtidas neste trabalho de pesquisa, algumas das melhores regras foram apresentadas aos especialistas para verificação. Grupos de empresas foram detectados em que a média de participações juntas e as vitórias em licitações levavam a indícios de conluio. Passaremos a apresentar três regras encontradas com a intenção de ilustrar os resultados alcançados:

- Uma regra envolvendo três empresas somava 14 certames de participação conjunta, onde o grupo celebrou oito contratos com a APF. Cada uma delas tinha uma média de participação individual relativamente baixa na base de dados (30 licitações);
- Duas empresas de um mesmo estado, com o total de participações individuais em licitações de 75 e 78 respectivamente, sendo que entre essas, em 68 licitações, elas participaram juntas e ganharam 14 contratos entre os anos de 2005 e 2007;
- No ano de 2008, uma empresa ganhou nove licitações em um mesmo órgão, concorrendo com outra empresa que não ganhou nenhuma das licitações em que ambas participaram. O detalhe é que as nove licitações perdidas pela segunda empresa foram exatamente as únicas licitações da base de dados em que ela participou. O total de vitórias da primeira empresa na base de dados era de apenas 12, mostrando que não se tratava de um grande fornecedor. Segundo o especialista, essa regra aponta fortes características de cartelização e simulação de concorrência.
- Ressaltamos que muitas regras encontradas, inclusive dentre as melhores, trouxeram grupos de fornecedores que, diante da comparação do número de participações no grupo com o número de participações individuais nas licitações, não eram considerados suspeitos de praticar cartel. Por exemplo, uma regra trazia três fornecedores que participaram conjuntamente de 22 licitações, mas cada um dos fornecedores tinha participado de mais de 100 licitações. Ou seja, tratavam-se apenas de grandes fornecedores que, coincidentemente, participaram de algumas licitações juntos.

Isso leva a crer que a fórmula de avaliação criada ainda necessita de aprimoramento para filtrar melhor as regras (Equação 1).

8 CONCLUSÃO

Neste trabalho, foi apresentada a aplicação de agentes de mineração de dados para detecção de cartéis em licitações públicas, a qual se mostrou útil às atividades de auditoria governamental realizadas pela CGU.

O resultado experimental desta pesquisa mostrou que a combinação de técnicas de MD, como a clusterização e as regras de associação, possibilitou claramente o enriquecimento do conhecimento descoberto, trazendo boas expectativas quanto às futuras combinações de outras técnicas de MD associadas aos agentes de mineração de dados. Em relação ao tempo necessário para a realização dos experimentos, incluindo a preparação dos dados e processamento dos *datasets*, a proposta de agentes de mineração mostrou-se claramente com bom potencial quando comparada ao processo tradicional de MD.

A análise dos *clusters* descobertos apresentou fortes indícios de cartelização, o que pôde ser reforçado posteriormente, com a aplicação das regras de associação. As regras descobertas neste trabalho estão sob validação final dos auditores da CGU para apuração real de irregularidades cometidas, com o intuito de utilizar os resultados na prevenção da corrupção e na aplicação de penalidades cabíveis em parceria com o CADE.

Em trabalhos futuros pretendemos avançar nossas pesquisas na integração entre as áreas de MD e de SMA como suporte ao trabalho de auditoria governamental, seguindo a linha de pesquisa de *AMII* (RALHA, 2009). A continuação do trabalho levará à definição de um modelo arquitetural de SMA, com o desenvolvimento de um protótipo para aplicação em problemas tanto de detecção de formação de cartéis em licitações públicas, como em outras possibilidades de auditoria no âmbito da CGU (SILVA e RALHA, 2011).

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: International Conference on Very Large Data Bases (VLDB´94), 20., San Francisco. *Proceedings...*, Morgan Kaufmann Publishers Inc., p. 487-499, 1994.

BELLIFEMINE, F. L.; CAIRE, G.; GREENWOOD, D. Developing Multi-Agent Systems with JADE (Wiley Series in Agent Technology). Wiley, April 2007.

BRASIL. Lei n. 8.666, de 21 de junho de 1993. D.O.U. de 22/06/1993. Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/L8666cons.htm. Acesso em: 04/09/10.

CAO, L.; GORODETSKY, V.; MITKAS, P. A. Agent mining: the synergy of agents and data mining. *IEEE Intelligent Systems*, v. 24, n. 3, p. 64-72, May/June, 2009. doi:10.1109/MIS.2009.45.

CGU/PR. Controle interno, auditoria pública, correição, prevenção e combate à corrupção e ouvidoria. Disponível em: http://www.cgu.gov.br. Acesso em 09/09/2011.

CGU/PR. Portal da transparência do governo federal. Disponível em: http://www.portaltransparencia.gov.br. Acesso em: 09/09/2011.

DI PIETRO, M. S. Z. *Direito administrativo*. São Paulo: Atlas, 2009. ISBN: 9788522453801.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, v. 39, n. 11, p. 27-34, November, 1999.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. In: *Advances in knowledge discovery and data mining*. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.) American Association for Artificial Intelligence, Menlo Park, CA, USA, p. 1-34, 1999. doi:10.1145/240455.240464

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: an overview. AI *Magazine*, v. 13, p. 57-70, 1992.

HAN, J.; KAMBER, M. *Data mining:* concepts and techniques. San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 2nd edition, 2006.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. San Francisco, CA, USA. Prentice-Hall, Inc. 1988.

JUSTIÇA, MINISTÉRIO DA. Direito da concorrência: cartel. 2010. Disponível em: http://portal.mj.gov.br/data/Pages/MJ9F537202ITEMIDDEB1A9D4FCE04052A5D948E2F2FA2BD5PTBRNN.htm. Acesso em: 05/09/10.

MPOG. Portal de Compras do Governo Federal. Ministério do Planejamento, Orçamento e Gestão. Disponível em: http://www.comprasnet.gov.br/. Acesso em 04/09/2010.

RALHA, C. G. Towards the integration of multiagent applications and data mining. In: Longbing Cao (Org.). *Introduction to agent mining interaction and integration,* p. 37-46. Springer Science + Business Media, 2009. ISBN: 978-1-4419-0521-5. doi:10.1007/978-1-4419-0522-2_2.

RUSSEL, S.; NORVIG, P. *Artificial intelligence: a modern approach.* 3rd edition. Prentice-Hall Inc., 2010.

SILVA, C. V. S.; RALHA, C. G., Utilização de técnicas de mineração de dados como auxílio na detecção de cartéis em licitações. In: Workshop de Computação Aplicada a Governo Eletrônico, Belo Horizonte. *Anais...*, julho de 2010.

SILVA, C. V. S.; RALHA, C. G., AGMI - An agent-mining tool and its application to Brazilian government auditing. In: 7th International

Conference on Web Information Systems and Technologies (WEBIST), Noordwijkerhout, Netherlands. *Proceedings...*, p. 535-538, May 2011.

STN. Portal SIAFI. Disponível em: http://www.tesouro.fazenda.gov.br/siafi/. Acesso em: 05/09/2010.

STN. Estatística de uso do SIAFI. Ministério da Fazenda. Disponível em: http://consulta.tesouro.fazenda.gov.br/estatisticas/index_estatistica_menu.asp. Acesso em: 09/09/2011.

TAN, P; STEINBACH, M.; KUMAR, V. *Introduction to data mining*. Addison-Wesley Longman Publishing Co. Inc., 2005.

WAIKATO, University. *Weka machine learning project*. 2010. Disponível em: http://www.cs.waikato.ac.nz/ml/index.html. Acesso em: 05/09/2010.

WITTEN, I. H.; FRANK, E. *Data mining: practical machine learning tools and techniques*. 2nd edition. Morgan Kaufmann Series in Data Management Systems, 2005.

WOOLDRIDGE, M.; JENNINGS, N. R. Intelligent agents: theory and practice. *Knowledge Engineering Review*, v. 10, p. 115-152. 1995. doi:10.1017/S0269888900008122

WOOLDRIDGE, M. J.. *Introduction to multiagent systems.* 2nd edition, John Wiley & Sons Inc., 2009.

WU, X.; KUMAR, V.; ROSS QUINLAN, J.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. *Knowledge Information Systems*, v. 14, n. 1, p. 1-37, 2007. doi:10.1007/s10115-007-0114-2