Utilização de Técnicas de Mineração de Dados como Auxílio na Detecção de Cartéis em Licitações

Carlos Vinícius Sarmento Silva[1 2], **Célia Ghedini Ralha**[1]

[1]Departamento de Ciência da Computação – Universidade de Brasília (UnB)
Caixa Postal 4466 – CEP 70904-970 – Brasília – DF – Brasil

[2]Controladoria-Geral da União
CEP 70070-905 – Brasília – DF – Brasil

{carlos.vinicius,ghedini}@cic.unb.br, carlos.silva@cgu.gov.br

Abstract. *Government Auditing has been performed by The Office of the Comptroller General (CGU) as the central agency of the Federal Government. Several approaches are being used in an effort to fight and prevent corruption. However, some activities such as government purchasing fraud detection are limited by the difficulty in finding effective ways of implementing it. Researches have been aiming on Data Mining Area due to its achievements on the process of Knowledge Discovery in Database (KDD), and several approaches have been defined, such as classification, clusterization and association rules. This work proposes a solution based on data mining techniques to the problem of cartels acting in public bidding processes.*

Resumo. *O trabalho de auditoria governamental tem sido realizado no âmbito do Poder Executivo Federal pela Controladoria-Geral da União (CGU). Várias estratégias são utilizadas visando a prevenção e o combate à corrupção. No entanto, algumas atividades como a detecção de cartéis em licitações são limitadas pela dificuldade de encontrar soluções efetivas. A área de Mineração de Dados tem sido alvo de várias pesquisas por causa de seus bons resultados no processo de descoberta de conhecimento, e várias técnicas já foram definidas nessa área como classificação, clusterização e regras de associação. Este trabalho prevê uma proposta de utilização da técnica de regras de associação e a combinação desta com a técnica de clusterização para a solução do problema de detecção de cartéis em licitações. O intuito de dar suporte às atividades de auditoria da CGU.*

1. Introdução

Atualmente o volume de dados produzidos e armazenados pelos diversos sistemas de computação tem aumentado expressivamente. A informatização dos diversos setores do mercado e também do governo tem sido a causa primária deste aumento na produção de dados digitais. A Administração Pública atual mantém a maioria de seus processos suportados por sistemas computacionais. O Sistema Integrado de Administração Financeira do Governo Federal (SIAFI) registrou, só no ano passado, 1 bilhão de transações financeiras de 24 mil unidades gestoras. O Portal da Transparência, criado e mantido pela CGU, atualmente mantém quase 900 milhões de registros que totalizam cerca de $6,5$ trilhões de

reais em gastos do Governo Federal tais como transferências de recursos, gastos diretos e cartões corporativos [STN 2009], [CGU/PR 2009].

Os dados provenientes dos sistemas de informação são utilizados pelos órgãos de auditoria governamental para planejamento e execução de suas auditorias e fiscalizações das aplicações dos recursos públicos. 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.

Entretanto, para lidar com grandes volumes de dados, a utilização de técnicas de mineração de dados tem se mostrado de grande valia na obtenção de informações e no processo de descoberta de conhecimento. 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).

Um dos problemas encontrados nos trabalhos de auditoria, especificamente em análises de processos de licitação, é o chamado rodízio de licitação. A prática é ilegal e traz prejuízos significativos ao Erário. No entanto, a detecção de grupos suspeitos de praticar rodízio de licitação é bastante difícil, não havendo formas determinísticas que auxiliem eficazmente nesta tarefa.

Neste artigo, mostraremos a utilização de técnicas de mineração de dados para detecção de grupos suspeitos de praticar rodízio de licitação. A solução deste problema é de grande importância na execução dos trabalhos de auditoria, pois possibilita a detecção automática de alguns pontos críticos na auditoria de processos de licitação, direcionando de forma mais eficaz o trabalho do auditor.

O artigo está estruturado da seguinte forma: A Seção 2 apresenta as definições essenciais para compreensão do problema e da proposta de solução. A Seção 3 apresenta o problema e a proposta de solução, além dos experimentos com a análise dos resultados. Por fim, a Seção 4 traz a conclusão e os trabalhos futuros.

2. Definições

A CGU, órgão central dos Sistemas de Controle Interno e de Correição do Poder Executivo Federal, é responsável por assistir direta e imediatamente o Presidente da República quanto aos assuntos que, no âmbito do Poder Executivo, sejam relativos à defesa do patrimônio público e ao incremento da transparência da gestão, por meio das atividades de controle interno, auditoria pública, correição, prevenção e combate à corrupção e ouvidoria [Brasil 2003].

Dessa forma, a CGU é responsável pelas atividades de auditoria governamental no âmbito do Poder Executivo Federal. Desenvolve, ainda, ações voltadas para a promoção da transparência e a prevenção da corrupção, que se destacam no núcleo essencial da proposta política e do programa de metas fundamentais do Governo Federal. Nesse contexto, a CGU tem se firmado também como uma típica agência anti-corrupção, que privilegia a

elaboração de estratégias e políticas de prevenção e combate a esse mal [CGU 2008].

Além disso, vinculada à Secretaria de Prevenção da Corrupção e Informações Estratégicas da CGU, está a Diretoria de Informações Estratégicas que atua dando suporte a atividades de pesquisa, produção e troca de informações de inteligência, com vistas a colaborar com as atividades das demais unidades da CGU, em especial na detecção de ilicitudes ocultas em atos, contratos e procedimentos administrativos. Essa diretoria colaborou na análise dos resultados desta pesquisa através de especialistas da área.

2.1. Licitações, Contratos e Cartéis

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 e aceitará a mais conveniente para a celebração do contrato. Licitação pode então ser vista como um procedimento democrático para se contratar com o poder público. Este procedimento tem como objetivo garantir a observância do princípio constitucional da isonomia e selecionar a proposta mais vantajosa para a Administração [Brasil 1993].

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 para manobrar a Lei, com finalidades diversas.

Cartel é um acordo explícito ou implícito entre concorrentes que visa, principalmente, a fixação de preços ou quotas de produção, a divisão de clientes e de mercados de atuação. O objetivo 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 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 [Ministério da Justiça 2009].

O rodízio de licitações se dá quando um grupo de empresas forma um cartel com a finalidade de dividir as licitações entre si, elevando o preço de contratação com a Administração Pública, 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, pode também a CGU punir as empresas suspeitas através da Declaração de Inidoneidade, que as impede imediatamente de licitar e contratar com a Administração Pública. 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".

2.2. KDD e Mineração de 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. Esta utilização vai desde a exploração de grandes bases coorporativas, até a análise de comportamento de usuários em um sítio na internet, por exemplo [Fayyad et al. 1996b], [Alam et al. 2009].

O processo é classificado como não trivial porque envolve decisões que estão além da aplicação das técnicas, como a de definir exatamente o problema que se tem para que assim possa se encontrar um caminho de otimização através da aplicação correta de algoritmos para extração da informação.

Nesse sentido, a mineração de dados é a parte desse processo onde se dá a aplicação de algoritmos específicos para extração de padrões (ou modelos) dos dados.

Segundo [Fayyad et al. 1996a], o processo de KDD é interativo e iterativo. Interativo pelo fato de depender de muitas decisões tomadas pelo usuário. E iterativo na medida em que os resultados obtidos por meio da mineração de dados fazem pouco ou nenhum sentindo, exigindo, assim, um recalibramento das funções de mineração de dados até que se tenham dados que sejam de fato úteis ao negócio.

Segundo [Tan et al. 2005], as tarefas de mineração de dados são geralmente divididas em duas categorias principais:

- **Tarefas Preditivas:** têm como objetivo 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 que os atributos usados para fazer a predição são conhecidos como *explanatórios* ou *variáveis independentes*.
- **Tarefas Descritivas:** tem como objetivo derivar padrões como correlações, tendências, grupos, trajetórias e anomalias as quais sumarizam as relações subjacentes nos dados. Tarefas de mineração de dados descritivas são frequentemente exploratórias e frequentemente requerem a utilização de técnicas para validar e explanar o resultado (pós-processamento).

Há diversas técnicas de mineração de dados tais como classificação, clusterização, regras de associação, regras de sequência, regressão, sumarização, entre outras. Destacaremos a seguir duas das principais técnicas: clusterização e regras de associação.

2.2.1. Clusterização

Segundo [Jain and Dubes 1988], clusterização é a tarefa descritiva onde se procura identificar um conjunto finito de categorias ou "clusters" para descrever uma informação. Estas categorias podem ser mutuamente exclusivas ou não.

A análise de cluster está relacionada com outras técnicas que são usadas para dividir objetos de dados em grupos. A clusterização pode ser considerada como a forma de classificação em que se cria uma rotularização de objetos com rótulos de classe (que são os clusters). Entretanto, esses rótulos são derivados unicamente dos dados de forma dinâmica.

Em contraste, o processo propriamente dito de classificação, é uma classificação supervisionada, isto é, objetos novos e não rotulados recebem um rótulo de classe usando um modelo desenvolvido a partir de objetos com rótulos de classes já conhecidos. Por essa razão, a análise de clusters é algumas vezes referida como uma espécie de classificação não supervisionada [Tan et al. 2005].

2.2.2. Regras de Associação

Essa técnica de mineração de dados consiste em descobrir relações fortes entre determinados atributos. Tem a capacidade de detectar padrões em forma de regras que associam valores de atributos num determinado conjunto de dados. Essas regras são expressas em forma de conjunções do tipo $atrib_1 = valor_1, atrib_2 = valor_2 \ \ ..., atrib_m = valor_m \rightarrow atrib_{m+1} = valor_{m+1}, atrib_{m+2} = valor_{m+2} \ \ ..., atrib_n$, onde $atrib$ é um atributo do conjunto de dados e $valor$ é o valor do atributo identificado na regra.

Segundo [Witten and Frank 2005], a diferença entre classificação e regras de associação é que estas podem prever padrões com qualquer atributo, e não só da classe selecionada. Diferentes regras de associação expressam diferentes regularidades subjacentes no conjunto de dados, cada uma predizendo coisas diferentes.

A cobertura das regras de associação é medida pela probabilidade da regra se repetir no conjunto de dados, e é chamada também de *suporte*. A acurácia da regra, chamada de confiança, é o percentual de instâncias preditas corretamente pela regra.

Segue a definição formal da técnica segundo [Han and Kamber 2005].

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 $A \cap B = \phi$. 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 transações em D contendo A que também contém B, isto é, a probabilidade condicional $P(B|A)$.

Por exemplo, a seguinte regra: $temperatura = frio \rightarrow umidade = normal$, o suporte será o percentual de instâncias na base de dados em que o atributo $temperatura$ seja $frio$ e o atributo $umidade$ seja $normal$. Já a confiança será a proporção de instâncias com temperatura fria que tenham umidade normal.

Destaca-se como um dos algoritmos mais populares para aplicação dessa técnica o *Apriori*, apresentado em [Agrawal and Srikant 1994].

3. Problema e Proposta de Solução

A identificação de cartéis em licitações, na maioria das vezes, é uma tarefa difícil, pois exige análise de vários processos de licitação, o que normalmente extrapola o escopo de apenas um órgão da Administração Pública. Cartéis podem atuar em vários órgãos, cidades, e até mesmo estados da Federação.

A análise de cruzamentos de dados utilizando linguagens de consultas tais como

SQL (*Structured Query Language*) é também impraticável, pois o espaço de solução é exponencial. Assim, devido às dificuldades inerentes ao processo de detecção de cartéis, as atividades de auditoria que envolvem esse problema se limitam somente à confirmação de suspeitas normalmente levantadas após denúncias.

Assim, não há formas determinísticas que auxiliem eficazmente na identificação de cartéis, pois o espaço de soluções é exponencial quanto ao número de empresas participantes das licitações analisadas.

O problema então se resume em identificar de forma eficaz e eficiente grupos de empresas suspeitos de praticar rodízio em licitações.

3.1. Solução utilizando Regras de Associação

A proposta de utilizar a técnica de regras de associação 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)

$$a_{i,j} = \begin{cases} sim & \text{se fornecedor } j \text{ participou da licitação } i; \\ nao & \text{se fornecedor } j \text{ não participou da licitação } i; \end{cases}$$
$$1 \leq i \leq m; 1 \leq j \leq n;$$

$$a_{i,n+1} = vencedor(i)$$
$$1 \leq i \leq m; vencedor(i) = \text{CNPJ da empresa vencedora da licitação } i$$

Dessa forma, espera-se obter regras do tipo:

$$fornecedor_A = sim, \quad fornecedor_B = sim \rightarrow vencedor = fornecedor_C$$

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

$$fornecedor_A = sim, \quad fornecedor_B = sim \rightarrow fornecedor_C = sim$$

Foram realizados 3 experimentos. No primeiro, utilizou-se apenas a técnica de regras de associação; no segundo, apenas clusterização; e no terceiro, uma combinação das duas técnicas. Passaremos a apresentar a execução dos experimentos.

3.2. Experimentações

Foram realizadas atividades de mineração de dados em uma base de licitações extraída do sistema ComprasNet, onde são realizados os pregões eletrônicos do Governo Federal. 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.

Os testes foram executados utilizando a ferramenta Weka, em sua versão 3.6.1. A ferramenta foi desenvolvida pela Universidade de Waikato na Nova Zelândia, e foi escolhida por ser software livre e implementar vários algoritmos de mineração de dados, além de disponibilizar bibliotecas para acesso a esses algoritmos através de aplicações Java [Waikato 2009].

A Tabela 1 mostra alguns dados da base utilizada nos experimentos. Cada registro da base de dados representa a participação de uma empresa em uma determinada licitação.

Table 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

3.2.1. Experimento 1

Dois *datasets* foram preparados no intuito de aplicar as técnicas de regras de associação para detecção de grupos suspeitos de fazer rodízio de licitações. O algoritmo utilizado neste experimento foi o *Apriori*. Este algoritmo foi escolhido por ser um algoritmo seminal da técnica de regras de associação, além de ser apontando na comunidade científica como um dos melhores algoritmos de mineração de dados, 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 serão também frequentes. Dessa forma, *Apriori* inicia sua execução procurando os conjuntos menores de itens e prossegue analisando os conjuntos em ordem crescente, ignorando os superconjuntos que contêm conjuntos já descartados nos processos anteriores.

O primeiro *dataset* 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). Essa escolha se deu pelo fato de estarmos procurando grupos de empresas atuando em cartéis. Portanto, não faz sentido procurar entre aquelas que participaram de apenas uma licitação.

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. Isso porque as regras que indicam a não participação de fornecedores não traz, a princípio, nenhum resultado de interesse para o

problema de rodízio de licitações (ex: $fornecedor_A = nao, \quad fornecedor_B = sim \rightarrow forncedor_C = nao$...).

A Tabela 2 mostra o resultado da execução do algoritmo *Apriori* nos *datasets* preparados.

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

	Instâncias	Atributos	Sup. Mín.	Conf. Mín.	Nº de Regras
Dataset 1	2701	3051	1%	70%	294
Dataset 2	2370	1086	1%	80%	145

A escolha de valores altos na configuração do suporte mínimo para execução do algoritmo não nos garante boas regras para identificação de cartéis. Uma regra que associa alguns fornecedores e que tem suporte alto provavelmente indica a presença de grandes fornecedores participando de várias licitações. Dessa forma, a configuração de um suporte mínimo alto para execução do algoritmo pode suprimir a aparição de diversas regras boas, com reais características de cartéis. Valores altos de confiança, por sua vez, garantem a seleção de regras boas. Assim, foi definida uma função de avaliação de regras para poder classificar melhor as regras obtidas. Isso porque, com a redução do suporte mínimo, muitas regras foram obtidas na execução do algoritmo, como mostra a Tabela 2.

3.2.2. Experimento 2

A tentativa de aplicar a técnica de regras de associação em dados de todo o país deixou o espaço de soluções bastante esparso. O estudo do negócio possibilitou verificar que muitas vezes os fornecedores não se restringem necessariamente às regiões macroeconômicas. Um exemplo típico é a situação 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, embora Goiás e Mato Grosso do Sul estejam na região Centro Oeste e Tocantins esteja na região Norte. Por isso, foi necessário aplicar técnicas de clusterização para mapear os grupos comuns de atuação dos fornecedores.

Foi aplicada a técnica de clusterização nos dados de toda a base com o objetivo de definir as regiões geográficas comuns de participação de empresas em licitações. O algoritmo utilizado para a descoberta não supervisionada de clusters foi o EM (*Expectation-Maximization*).

Segundo [Han and Kamber 2005], EM é um algoritmo de refinamento iterativo que pode ser usado para encontrar estimativas de parâmetro. Pode ser visto como uma extensão do paradigma *k-means*, que associa um objeto ao cluster que lhe é mais similar, baseado na média encontrada. Ao invés de associar cada objeto a um único cluster, o EM pode associar objetos a mais de um cluster e definir um peso a cada associação. Esse peso representa a probabilidade daquele objeto pertencer ao cluster. Ou seja, não há fronteiras bem definidas entre os clusters, e isso é interessante na análise de regionalização de mercados de licitações. Pode ser que um estado tenha 55% de chance de pertencer a um cluster e 45% de chance de pertencer a um outro cluster. Esta segunda associação não

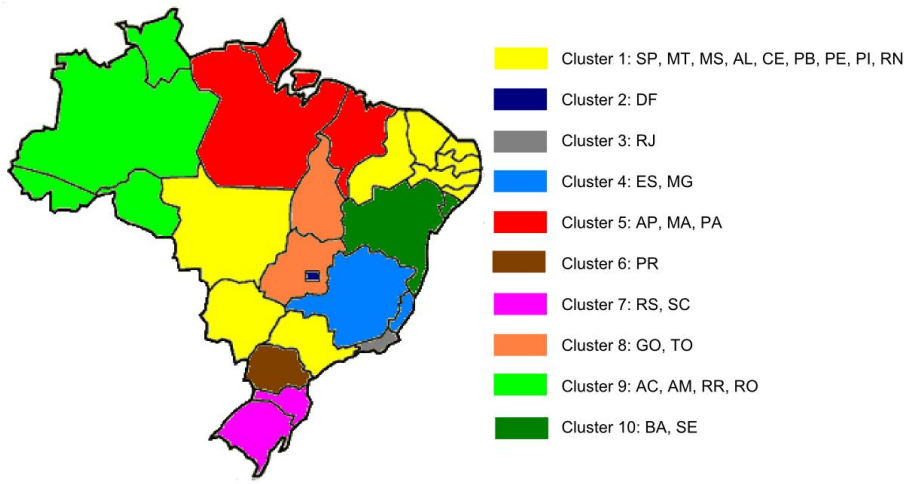

Figure 1. Classificação de UF nos Clusters

pode ser ignorada quando se trata de mercado de licitações públicas. Por esse motivo, o algoritmo EM pareceu mais propício na descoberta de regiões geográficas formadas pelos fornecedores de licitações.

O teste foi executado nas 26.615 instâncias da base, tendo como atributos o Fornecedor e a UF onde participou da licitação, e trouxe como resultado 10 clusters. Na Figura 1, pode-se notar que a maioria dos clusters encontrados tem como característica a proximidade geográfica. A Tabela 3 mostra a distribuição das instâncias nos clusters encontrados.

Table 3. Distribuição das Instâncias por Cluster

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

3.2.3. Experimento 3

A partir das regiões obtidas no Experimento 2, foi aplicada a técnica de regras de associação em cada cluster na tentativa de identificar grupos de empresas associadas atuando especificamente na região. Os resultados deste experimento podem ser vistos na Tabela 4.

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

Cluster	Inst.	Atrib.	Sup.	Conf.	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

O valor do suporte mínimo foi adaptado em cada cluster para que as regras obtidas na execução do algoritmo sejam válidas para pelo menos 9 instâncias do *dataset*, isto é, para que o grupo de fornecedores apontado por uma regra de associação tenha atuado em pelo menos 9 licitações.

3.2.4. Avaliação dos Resultados Obtidos

Com ajuda de especialistas, definimos um método de avaliação das regras obtidas através do processo de mineração de dados. A fórmula de avaliação definida foi:

$$M = 100.\frac{V(F)}{Sup. \times Inst.} \tag{1}$$

Onde:

$Sup.$ = valor do suporte da regra;
$Inst.$ = número de instâncias do $dataset$;
F = conjunto de fornecedores que figuram na regra;
L = licitações em que todos os fornecedores de F participaram conjuntamente;
$V(F)$ = número de vitórias nas licitações L por algum fornecedor de F.

Normalmente, taxas altas de suporte e confiança determinam o quão boa uma regra de associação é. Suponha que seja encontrada uma regra de associação que mostre uma forte ligação entre três fornecedores atuando nas licitações de um *dataset*. Isso poderia ser indício de um rodízio de licitações. No entanto, quando verificamos na base, em quase nenhuma das ocasiões em que esses três fornecedores participaram juntos de licitações, algum deles conseguiu fechar um contrato com a Administração Pública. Portanto, suas participações nos mesmos processos licitatórios foram uma mera coincidência.

É notório que apenas suporte e confiança não são capazes de mensurar a qualidade de nossas regras. Por isso, os especialistas inseriram a função $V(F)$ na 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*. De forma mais simples, 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.

As regras foram avaliadas por meio da Equação 1. Para análise dos resultados, foram selecionadas as 10 melhores regras segundo a função de avaliação. As melhores regras obtidas no Experimento 1 tiveram, na média, melhores números de ocorrências (suporte multiplicado pelo número de instâncias). Isso comparado com as melhores regras obtidas pelos modelos gerados no Experimento 3. No entanto, as regras obtidas no Experimento 3 tiveram um aumento de cerca de 100% no valor de avaliação. O gráfico da Figura 2 mostra, para cada cluster, a avaliação obtida nas suas 10 melhores regras.

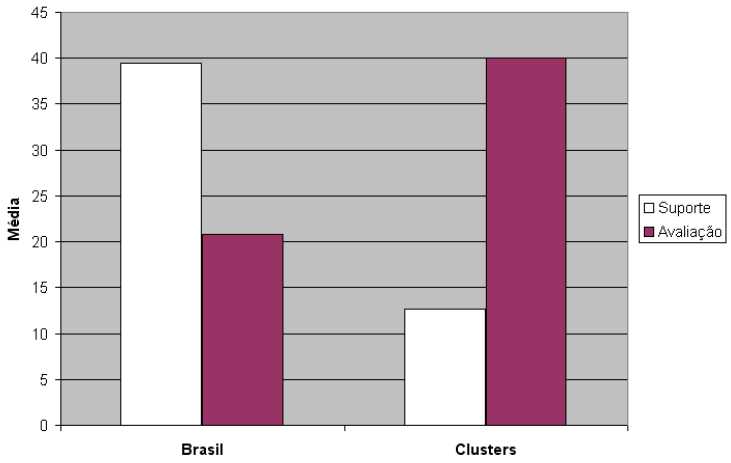

Figure 2. Média de Suporte e Avaliação das 10 melhores regras

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 abrangem o Brasil todo não foram tão boas quanto àquelas encontradas em regiões do país.

Entre os modelos gerados a partir dos clusters (Tabela 4), as melhores regras foram obtidas no Cluster 6. A comparação entre as 10 melhores regras obtidas nesses modelos pode ser vista no gráfico da Figura 3.

3.2.5. Conhecimento Descoberto

O modelo de cluster gerou interesse por parte do especialista, que explicou que 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. Assim, 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.

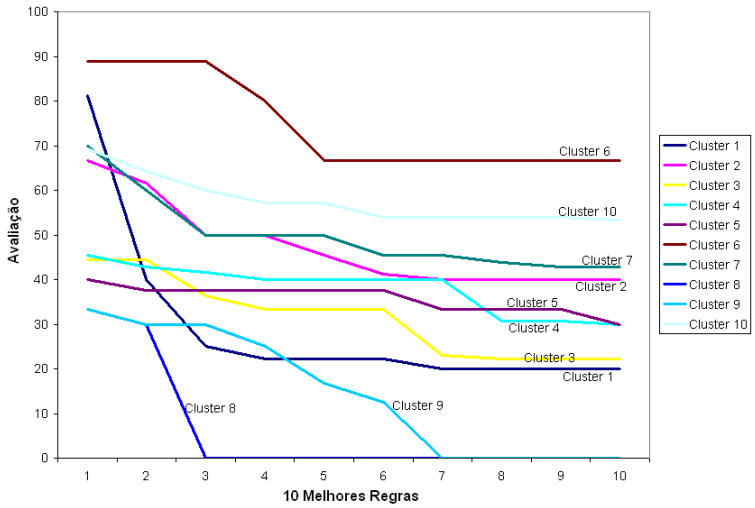

Figure 3. Comparação das 10 melhores regras dos Modelos de Cluster

O Cluster 1 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, 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 de mais de 15 licitações tanto numa sub-região quanto na outra. Dessas 8 empresas, nenhuma é da sub-região composta por São Paulo ou Mato Grosso ou Mato Grosso do Sul.

As próximas atividades serão no sentido de experimentar novas bases de dados na tentativa de detectar outros clusters de interesse para investigações, como por exemplo, clusters envolvendo órgãos superiores.

Quanto às regras de associação, algumas das melhores regras foram apresentadas ao especialista para verificação. Grupos de empresas foram detectados onde a média de participações juntas e as vitórias em licitações levavam a indícios de conluio. Alguns exemplos de regras encontradas são detalhadas a seguir:

- Duas empresas de um mesmo estado, sendo que o total de licitações de que cada uma participou individualmente foi de 75 e 78 licitações. Dentre essas, em 68 licitações, participaram juntas e ganharam 14 contratos entre os anos de 2005 a 2007;
- Outra regra, envolvendo 3 empresas, somava 14 certames de participação conjunta. O grupo celebrou 8 contratos com a Administração. Cada uma delas tinha uma média de participação individual relativamente baixa na base de dados (média de 30 licitações);
- No ano de 2008, uma empresa ganhou 9 licitações em um mesmo órgão, concorrendo com outra empresa que não ganhou nenhuma das licitações em que ambas participaram. O detalhe é que as 9 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.

No entanto, 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 que trazia 3 fornecedores que participaram conjuntamente de 22 licitações, mas que cada um dos fornecedores tinha participado de mais de 100 licitações. Ou seja, tratava-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.

4. Conclusões e Trabalhos Futuros

Neste trabalho, foi apresentada uma proposta de aplicação de técnicas de mineração de dados para solução do problema de rodízio de licitações, para auxílio às atividades de auditoria governamental.

Os resultados dos experimentos executados mostraram claramente o potencial da aplicação de técnicas de mineração de dados como auxílio na detecção de cartéis em licitações. A análise dos clusters descobertos apresentou fortes indícios de cartelização, o que pôde ser confirmado, posteriormente, com a aplicação das regras de associação. Além disso, comprovou-se a utilidade da aplicação das técnicas como suporte no trabalho de auditoria governamental, possibilitando a atuação imediata da CGU na prevenção da corrupção e na aplicação de penalidades cabíveis em parceria com o CADE.

Os experimentos mostraram também que a combinação das técnicas de clusterização e regras de associação possibilitou claramente o enriquecimento do conhecimento descoberto, trazendo boas expectativas quanto às futuras combinações de outras técnicas de mineração de dados.

Em trabalhos futuros, será feita uma integração entre as áreas de Mineração de Dados e de Sistemas Multagentes (SMA) para suporte ao trabalho de auditoria governamental, seguindo a linha de pesquisa de *Agents and Data Mining Interaction and Integration* - AMII [Ralha 2009].

A intenção de integrar as duas áreas de conhecimento tem como meta o enriquecimento no processo de mineração através da aplicação de agentes inteligentes, autônomos e independentes, além de possibilitar a exploração de outras características de SMA tais como distribuição de recursos e controle, descentralização dos dados, comunicação assíncrona, entre outras [Wooldridge and Jennings 1995].

A continuação do trabalho levará ao modelo arquitetural de integração e ao desenvolvimento do protótipo para aplicação no problema de detecção de formação de cartéis em licitações públicas. Outras possibilidades de auditoria serão consideradas no âmbito da CGU.

References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *VLDB '94: Proceedings of the 20th International Conference on Very*

Large Data Bases, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alam, S., Dobbie, G., and Riddle, P. (2009). Exploiting swarm behaviour of simple agents for clustering web users session data. In Cao, L., editor, *Data Mining and Multi-agent Integration*. Springer US.

Brasil (1993). Lei n. 8.666, de 21 de junho de 1993. D.O.U. de 22/06/1993.

Brasil (2003). Lei n. 10.683, de 28 de maio de 2003. D.O.U. de 29/05/2003.

CGU (2008). Controle interno, prevenção e combate à corrupção - ações da cgu em 2008. http://www.cgu.gov.br.

CGU/PR (2009). Portal da transparência do governo federal. http://www.portaltransparencia.gov.br.

Di Pietro, M. S. Z. (2009). *Direito Administrativo*. Atlas, 22 edition.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11):27–34.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996b). From data mining to knowledge discovery: an overview. In *Advances in knowledge discovery and data mining*, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Frawley, W. J., Shapiro, P. G., and Matheus, C. J. (1992). Knowledge discovery in databases - an overview. *Ai Magazine*, 13:57–70.

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

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Ministério da Justiça, S. B. d. D. d. C. (2009). Cartel. http://www.mj.gov.br.

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.

STN (2009). Portal siafi. http://www.tesouro.fazenda.gov.br/siafi/.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley, us ed edition.

Waikato, U. (2009). Weka machine learning project. http://www.cs.waikato.ac.nz/ ml/index.html.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition.

Wooldridge, M. and Jennings, N. R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10:115–152.

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.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37.