Mineração de dados no Portal da Transparência para buscar indícios de fraudes em licitações

Leonardo Smoginski Fernandes[1], Roberto de Carvalho Ferreira[2], Ricardo Costa Pinto e Santos[3]

[1]Instituto Federal de Educação Ciência e Tecnologia do Sudeste de Minas Gerais (IFSudesteMG) – Campus Juiz de Fora

leonardo.smoginski@gmail.com, roberto.ferreira@ifsudestemg.edu.br, ricardo.santos@ifsudestemg.edu.br

Abstract. *This article proposes the using of data mining tools and techniques to find evidence of cartel crimes that are becoming increasingly common in Brazil. In view of the Complementary Law No. 131/2009 (Transparency Law), scripts for data extraction using web crawling techniques were developed, as well as the development of a semi automatic Knowledge Discovery in Databases process (KDD) using computational techniques of visualization, cleaning, processing and analysis of results.*

Resumo. *Este artigo propõe a utilização de ferramentas e técnicas de mineração de dados para buscar indícios de formação de cartel, crime que se torna cada vez mais frequente no Brasil. Tendo em vista que a Lei Complementar Nº 131/2009 (Lei da Transparência) proporciona o livre acesso a uma ampla gama de dados da gestão pública, os dados referentes a licitações foram extraídos de forma automatizada com a aplicação de web crawling. O processo de Descoberta de Conhecimento em Banco de Dados (KDD) se deu com a aplicação do algoritmo de associação apriori, através do qual foi possível obter regras de associação revelando fortes relações estatísticas entre participantes de licitações. As regras de associação obtidas foram submetidas a buscas por processos jurídicos, no intuito de verificar se alguma das relações encontradas foi, em algum momento, julgada pelo Tribunal de Contas da União (TCU). Os resultados revelaram que nenhuma das regras encontradas relacionou empresas julgadas em um mesmo processo, mas algumas das mesmas já estiveram envolvidas em processos por fraude em licitações.*

1. Introdução

A publicidade gerada pela operação Lava Jato (Brasil) demonstrou como esquemas de formação de cartel e pagamento de propinas na concessão de contratos podem afetar a política e a economia de um país. Tal operação revelou não apenas os impactos que podem ocorrer no cenário econômico mundial, mas também que o desvio de dinheiro público soma quantias expressivas.

No que se refere às fraudes cometidas em licitações, apesar dos procedimentos estabelecidos para elaborar os editais de participação, segundo Batista, Neto e Fariello (2015) a lei ainda é falha, portanto é possível aferir que existe a necessidade de elaborar novos trabalhos e mecanismos que contribuam para combate à corrupção sistêmica e

auxiliem nas investigações.

A Lei Complementar 131 de 27 de maio de 2009 determina a transparência orçamentária e financeira da União, dos Estados, do Distrito Federal e dos Municípios. A lei também define que os dados devem ser acessíveis através da internet e que deve ser possível realizar download dos bancos de dados.

Isso possibilita a aplicação de técnicas para Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Database - KDD). Segundo Fayyad et al. (1996b), KDD aborda técnicas computacionais utilizadas para encontrar informações valiosas em grandes volumes de dados e descreve o processo de forma abstrata na Figura 1.

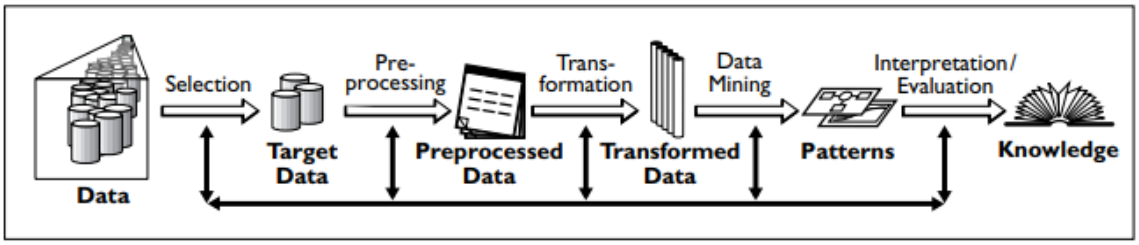

Figura 1. Visão geral dos passos que constituem KDD. (Fayyad et al. 1996b)

Existem diversas aplicações para as técnicas de descoberta de conhecimento em bancos de dados. Uma de suas aplicações, exemplificada por Fayyad et al. (1996a), é com o intuito de detectar fraudes.

Portanto este artigo propõe a seleção, aplicação e avaliação do desempenho de uma técnica de mineração de dados que seja aplicável à base de dados licitatórios fornecida pelo Portal da Transparência, com o intuito de identificar indícios de fraudes cometidas em certames realizados por órgãos públicos.

2. Metodologia

Para melhor direcionamento dos esforços e fundamentação do trabalho foi realizada uma revisão sistemática em busca de obter as técnicas de mineração (Data Mining) mais recomendadas para fins de auditoria em licitações.

A partir das técnicas selecionadas na revisão, um fluxo foi elaborado em etapas de extração, pré-processamento, transformação, mineração e análise dos resultados.

Tanto para a extração dos dados iniciais quanto para auxiliar a etapa final de análise dos resultados foram desenvolvidos scripts utilizando técnicas de crawling, definidas por Kausar et al.(2013) como softwares ou scripts que navegam através da World Wide Web de forma sistemática e automatizada.

Uma vez que todos os dados foram coletados, toma início a tarefa de análise e pré-processamento dos dados. Tal etapa utilizou recursos da linguagem de programação Python que foram utilizados com o objetivo de eliminar valores nulos ou inválidos para o propósito do trabalho.

Para a etapa de processamento dos dados, utilizou-se a implementação do algoritmo de associação Apriori contido no pacote “arules”, desenvolvido com a linguagem de programação R, que é amplamente utilizada para fins de análise estatística.

Após a execução e organização das regras, um processo de busca automatizada

foi elaborado para avaliar o desempenho das mesmas em relacionar empresas que eventualmente possam ter se envolvido em esquemas de conluio e formação de cartel. A fonte de busca adotada foi o site Jusbrasil, que concentra documentos oficiais sobre processos ocorridos em todo o país.

2.1. **Revisão Bibliográfica**

A revisão sistemática realizada neste trabalho teve o Google Scholar como principal motor de busca e foi elaborada de forma a utilizar operadores lógicos junto aos termos do domínio de conhecimento, pois assim foi possível obter resultados mais específicos. Tais operadores booleanos foram associados a uma lista de termos de pesquisa divididos em três principais áreas de conteúdo:

1. Material técnico e teórico sobre as tecnologias empregadas em mineração de dados;
2. Tipos de fraudes existentes em eventos licitatórios;
3. Trabalhos empregando "1" em "2".

Após a definição destes itens como foco principal da revisão bibliográfica, foram elaboradas diversas strings de busca combinando os operadores a grupos de palavras para cada uma das áreas definidas previamente, como seguem os exemplos abaixo:

1. "Descoberta de Conhecimento em Bases de Dados" OR "Knowledge Discovery in Database" OR "Mineração de dados" OR "Data Mining" OR "Inteligência Artificial" OR "Artificial intelligence";
2. "Rodízio de Licitação" OR "Collusion" OR "Cartéis" OR "Cartels" OR "Favor exchange" OR "Bidding" OR "Bid-rigging" OR "Corrupção sistêmica" OR "Systemic Corruption" OR "Auditoria de Governo" OR "Government Auditing";
3. Termos do grupo "1" AND termos do grupo "2";

Para cada artigo encontrado com as strings de busca houve uma análise do título e resumo, objetivando descartar resultados considerados irrelevantes.

Após a análise e limpeza dos resultados foram obtidos 11 artigos em português e 7 artigos em inglês, dos quais 5 foram considerados trabalhos de maior relevância para o objetivo final.

Dentre os principais artigos selecionados, encontram-se materiais teóricos sobre técnicas computacionais utilizadas no ramo da Mineração de Dados, conteúdo teórico sobre os tipos de fraude sistêmica e comportamentos dos indivíduos envolvidos, bem como o trabalho realizado por Silva C. V. S. e Ralha, C. G (2010), o qual aborda dentre seus temas, a aplicação do algoritmo Apriori para detecção de cartéis em licitações públicas, trabalho que contribuiu para o direcionamento deste artigo e técnicas necessárias para atingir o objetivo da análise.

2.2. **Extração dos Dados**

O Portal da Transparência do Governo Federal é um site de acesso livre, onde é possível encontrar informações sobre como o dinheiro público é utilizado. Na página "Dados Abertos", que contém uma das principais funcionalidades do portal, foi possível

realizar uma inspeção da rota de download dos dados utilizando a ferramenta de inspeção do navegador Google Chrome. Tal inspeção (demonstrada na Figura 2) possibilitou entender o funcionamento da requisição HTTP, ou seja, quais os metadados de cabeçalho necessários para sua execução e obtenção de resposta bem-sucedida.

Figura 2. Requisição utilizada para download de dados no portal da Transparência.

Dentre os padrões encontrados foi possível verificar que os parâmetros selecionados no formulário de dados abertos eram passados pelo método GET, bastando inserir o ano com 4 dígitos e em seguida o mês com 2 dígitos ao endereço abaixo:

"http://www.portaltransparencia.gov.br/download-de-dados/licitacoes/{ano}{mês}"

Para a obtenção dos dados de forma automatizada, baseando-se nos padrões encontrados e na inexistência de barreiras de dificultam a extração automatizada de dados, um software Web Crawler foi implementado utilizando a linguagem Python e as técnicas para Focused Crawling Kausar et al. (2013), ou seja, o script teve seu escopo definido para realizar as requisições, download e extração das planilhas referentes às licitações de cada mês contido no Portal da Transparência, verifique na Figura 3.

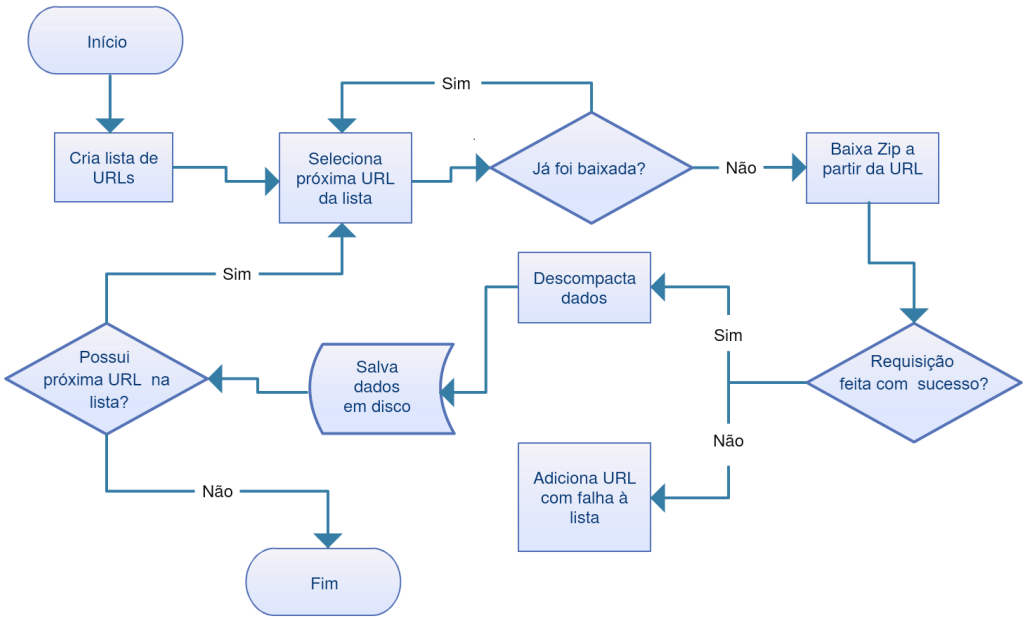

Figura 3. Processo elaborado para extração de dados no Portal da Transparência.

No momento de execução dos scripts o portal continha os registros referentes às licitações federais realizadas entre janeiro de 2013 e maio de 2019. Tais dados foram obtidos de forma rápida e padronizada, minimizando a possibilidade de falhas humanas durante possível gerenciamento manual de 76 downloads. Os dados referentes a cada mês encontram-se divididos em três planilhas no formato CSV (Comma Separated Values), sendo elas categorizadas como “Licitação”, “ItemLicitação” e “ParticipantesLicitação”.

2.3. Análise e Pré-Processamento

Pode-se considerar que algumas das etapas mais cruciais para a Mineração de Dados sejam a análise exploratória e o pré-processamento dos dados, pois, quando não são executadas com precisão, os resultados obtidos após a mineração propriamente dita podem ser inconclusivos ou até mesmo interpretados de forma incorreta, o que gera grande retrabalho em diversas etapas do processo de descoberta de conhecimento.

2.3.1. Análise Exploratória

Para obter maior entendimento sobre o significado, formato e valores possivelmente inválidos ou desnecessários dentre os atributos da base de dados, foram aplicadas técnicas de Análise Descritiva, descrita por Cortês et al.(2002) como um processo composto pelas etapas de Análise Prévia e Descobrimento.

A Análise Prévia possui o objetivo de identificar valores discrepantes e anomalias que possam atrapalhar o processamento realizado pelos algoritmos de mineração nas etapas posteriores.

O Descobrimento tem como principal intuito aplicar técnicas que auxiliem na obtenção de entendimento a respeito dos dados, mesmo que sem nenhuma hipótese prévia sobre o relacionamento dos campos ou significado intrínseco dos dados.

As principais ferramentas utilizadas para ambas as etapas da Análise Descritiva foram o ambiente de desenvolvimento Jupyter, que permite a execução de scripts de

forma interativa (Figura 4), e a biblioteca Pandas que favorece o processo exploratório dos dados com diversas implementações performáticas para a manipulação e visualização dos dados.

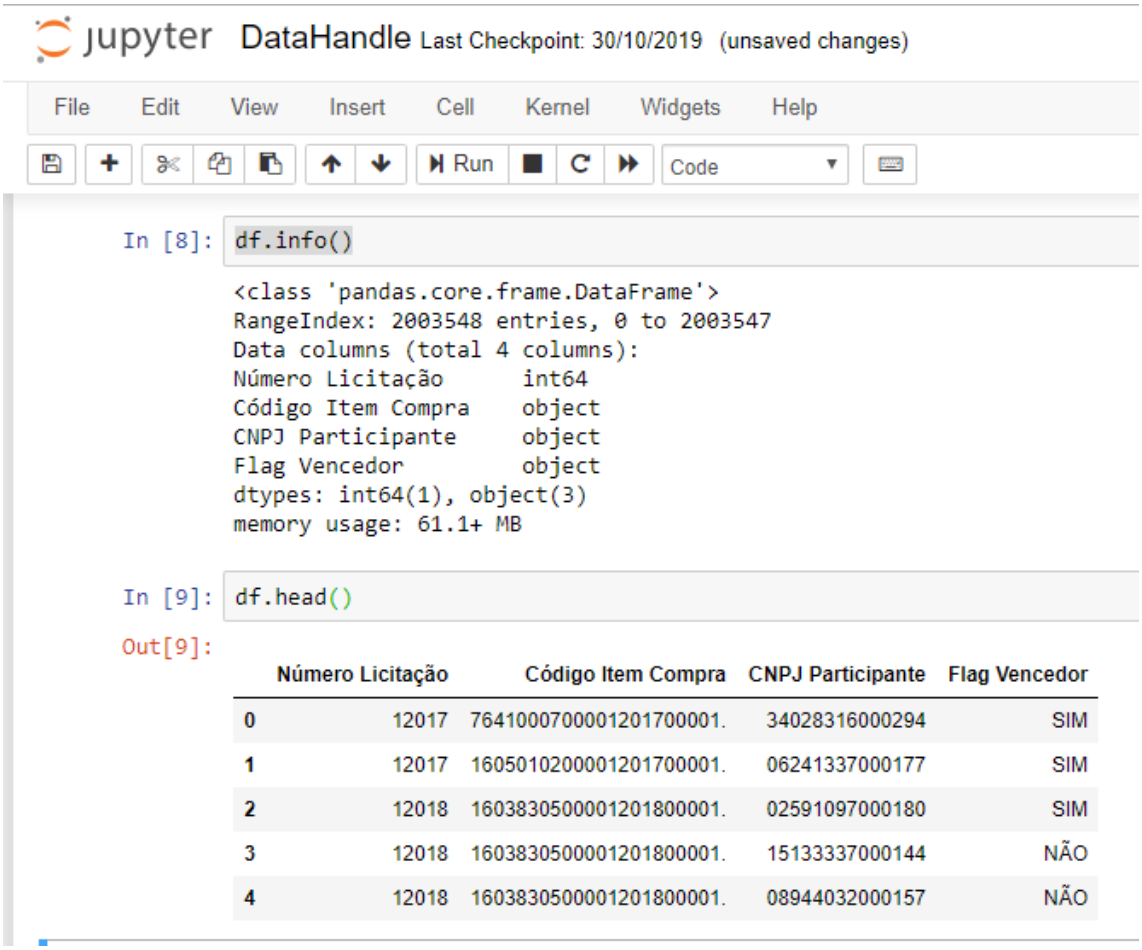

Figura 4. Utilização da interface Jupyter para análise interativa dos dados.

As análises exploratórias executadas revelaram que, dentre as três categorias de planilha citadas anteriormente(“Licitação”, “ItemLicitação” e “ParticipanteLicitação”), os dados necessários para a aplicação das técnicas de mineração podiam ser encontrados diretamente na planilha “ParticipanteLicitação”, bem como a possibilidade de descartar diversos campos da mesma, havendo necessidade de permanecer apenas “Número Licitação”, “Código Item Compra”, “CNPJ Participante” e “Flag Vencedor”.

Dentre outras informações obtidas, uma das quais considera-se importante citar é a necessidade de realizar o agrupamento dos participantes de acordo com a chave primária dos respectivos eventos licitatórios.

2.3.2. **Pré-Processamento**

Conforme informações obtidas durante a etapa de análise exploratória, todos os dados foram agrupados de acordo com a chave primária do evento licitatório, porém foi necessário estabelecer filtros para eliminar dados que viriam a afetar o funcionamento do algoritmo de mineração selecionado. Para que os dados estivessem limpos e dentro da conformidade necessária, foram utilizados os seguintes critérios demonstrados na Tabela 1.

Tabela 1. Critérios de aceitação para registros de licitação.

Atributo	Filtro	Motivação
CnpjParticipante	!= -1	Licitações para Órgãos de segurança pública ou que possuam caráter sigiloso adicionam o CNPJ dos participantes com o valor "-1" para que não possam ser identificados, fazendo com que a singularidade das empresas sejam omitidas entre outras. Portanto eventos com tais características foram eliminados.
CodItemCompra	\d{21,23}\.	A expressão regular encontra o padrão de 21 a 23 números seguidos por um ponto, tal filtro foi elaborado com o intuito de eliminar eventos onde o Item de Compra possua valores inválidos, uma vez que o campo influência na obtenção da chave primária para o evento licitatório. Irregularidades neste campo podem comprometer sua singularidade.
FlagVencedor	"SIM" == 1 && "NÃO" >= 1	Apesar da possibilidade de filtrar os eventos por categorias de licitações específicas, uma vez que os participantes foram agrupados por evento, o filtro adotado para selecionar os eventos a serem processados levou em consideração qualquer evento, independente da categoria, onde exista o perfil de concorrência, ou seja, com a contagem indicando apenas um vencedor e, pelo menos, um perdedor no certame.

2.4. **Transformação**

A transformação dos dados foi realizada conforme o necessário para a execução do algoritmo apriori. Portanto um artefato de código foi elaborado para realizar a transformação, além disso os dados totais (licitações entre janeiro de 2013 e maio de 2019), foram subdivididos em três níveis de granularidade.

A técnica de agregação dos dados, utilizada para separar níveis de granularidade, segundo Cortês et al.(2002) é realizada através da aplicação de operações de sumarização. O mesmo também cita como exemplo vendas diárias que podem ser agregadas em vendas semanais, quinzenais e mensais. Assim, os dados foram agregados em diferentes conjuntos, de acordo com as respectivas datas dos eventos licitatórios:

1. Baixa granularidade - Conjunto de dados foi gerado de forma única, concentrando todos os dados históricos para um único processamento.
2. Média granularidade - Os dados históricos das licitações foram divididos em conjuntos anuais, de forma a fornecer resultados a respeito das licitações realizadas em cada ano.
3. Alta granularidade - Os dados foram divididos em conjuntos mensais, para

permitir execuções e resultados com escopo reduzido.

Os dados equivalentes a cada um dos níveis citados acima foram exportados no formato CSV e verificados manualmente para assegurar a conformidade com os parâmetros necessários antes de prosseguir com a próxima etapa do processo mineração de dados.

2.5. **Mineração de regras de associação**

A abordagem apriori apresentada por Agrawal e Srikant (1994) é amplamente adotada atualmente na mineração de regras de associação, ou seja, na mineração de relações estatísticas entre itens presentes em diferentes transações.

Como exemplo, podemos citar a aplicação do algoritmo de associação apriori para relacionar produtos adquiridos em diferentes compras de supermercado. Oonde, definindo A como pão, B como leite e C como geléia e analisando a lista de compras realizadas em uma série temporal encontram-se regras estatísticas no formato A,B => C, ou seja, se "A" e "B" então "C". A regra indica uma relação forte de compra, revelando que quando se compra o pão "A" e leite "B" existe grande probabilidade de também se comprar a geleia "C".

Em comparação ao exemplo fornecido anteriormente, o algoritmo apriori foi utilizado visando a obtenção de regras de associação entre empresas que participem de eventos licitatórios, ou seja, os eventos licitatórios fariam o papel das compras realizadas e o CNPJ das empresas o papel dos produtos.

Mesmo com a aplicação de filtros e restrições nas etapas de pré-processamento, a quantidade de dados para processamento permanece massiva, e como o algoritmo de associação utiliza sua parametrização para selecionar os requisitos de obtenção para as regras, foi necessário realizar diferentes execuções do mesmo em busca dos valores mais adequados para obter regras fortes, consideradas de maior validade para o propósito de localização de indícios de fraude.

Ainda seguindo a linha de raciocínio exemplificada, as regras de associação obtidas revelaram relações fortes entre empresas que participaram de licitações federais. Tais regras podem ser exemplificadas pelo no formato CNPJ A, CNPJ B => CNPJ C. Ou seja, quando determinadas empresas participam juntas de um evento provavelmente uma terceira empresa também participa.

Tal padrão de participação em licitações pode, como demonstrado por Silva C. V. S. e Ralha C. G. (2010), ser resultante das participações de fortes concorrentes dentro de um mesmo ramo de licitação. Porém, ainda podem revelar comportamentos que ocorrem em casos de prática de cartel, descrita pelo artigo 4° da Lei Nº 8.137 de 27 de dezembro de 1990 como:

I. Abusar do poder econômico, dominando o mercado ou eliminando, total ou parcialmente, a concorrência mediante qualquer forma de ajuste ou acordo de empresas;
II. Formar acordo, convênio, ajuste ou aliança entre ofertantes, visando:
    A. À fixação artificial de preços ou quantidades vendidas ou produzidas;
    B. Ao controle regionalizado do mercado por empresa ou grupo de empresas;

C. Ao controle, em detrimento da concorrência, de rede de distribuição ou de fornecedores.

Quanto aos parâmetros utilizados para obtenção das regras, podemos citar de forma breve que são Suporte, Confiança e Levantamento, detalhados nos tópicos posteriores.

2.5.1. **Suporte (Support)**

A métrica de suporte é a primeira métrica a ser verificada para a eliminação de regras que não se se enquadrem no valor que foi parametrizado para execução. O suporte representa a frequência com que ocorre uma relação de itens. Por exemplo, A e B no total de transações N que está sendo analisado.

Para melhor entendimento da métrica é possível exemplificar um caso hipotético, onde um conjunto de dados que possua 100 transações seja analisado e o suporte mínimo seja parametrizado como 30%. O algoritmo irá retornar as relações que ocorreram em pelo menos 30 transações. Veja a definição matemática a seguir.

$$Suporte = \frac{Frequência(A, B)}{N}$$

Para o caso de aplicação deste trabalho, onde o algoritmo de associação foi utilizado em eventos licitatórios, foi necessário realizar diversas execuções até que o suporte fosse ajustado entre 1% e 3%, pois de acordo com o conjunto utilizado, valores de suporte mínimos superiores a 3% não retornam regra alguma.

2.5.2. **Comprimento Mínimo (Min. Length)**

Apesar de nenhum dos artigos encontrados na etapa de revisão bibliográfica citar este parâmetro, diversas bibliotecas de código adotam a métrica de comprimento mínimo para selecionar as relações que possam exportar regras com o perfil desejado.

Este parâmetro tem a funcionalidade de eliminar regras que não possuam a quantidade definida de elementos, ou seja, para casos onde se deseja encontrar relações com pelo menos 3 itens, basta definir o valor da métrica de acordo com o desejado.

Para o artigo em questão, foram realizadas execuções definindo o comprimento mínimo como 2, 3 e 4, porém, a quantidade de resultados obtidos nos casos com valor 3 foi surpreendentemente baixa, assim como as execuções com valor 4, onde apenas algumas regras em conjuntos de dados específicos foram encontradas.

Portanto, considerou-se a utilização do comprimento mínimo 2, devido ao fato de que nenhuma relação seria perdida. Pois o algoritmo utiliza a métrica apenas como forma de poda para valores abaixo do desejado, mantendo as regras com 3 e 4 elementos dentro do retorno fornecido.

2.5.3. **Confiança (Confidence)**

A métrica de confiança indica a frequência dos itens A e B da regra A => B, onde lê-se "A" implica em "B", dentro da quantidade total de transações onde o item A esteve presente. Vale ressaltar que as relações entre 2 itens geralmente exportam a regra oposta,

ou seja, entre os itens B e A. Tal relação será considerada como uma nova regra, que terá sua confiança representada como a frequência com que os itens A e B ocorrem dentro da quantidade total de transações do item B.

$$Confiança = \frac{Frequência(A, B)}{Frequência(A)}$$

No que se refere à utilização da métrica de confiança neste trabalho, realizaram-se execuções com valores progressivos iniciando em 20% até atingir 100% de confiança mínima. Observou-se que grande parte das regras obtidas já apresentavam confiança mínima acima de 40%. Portanto, adotou-se o valor 30% como métrica para execuções finais, de forma a obter uma margem de segurança de 10%, evitando excluir as relações consideradas fracas pois as mesmas podem revelar indícios de casos onde empresas participam juntas apenas para fraudar licitações em específico.

2.5.4. **Levantamento (Lift)**

Utilizado para medir a força da regra, ou seja, avalia o grau de aleatoriedade de uma relação, conforme a divisão do suporte encontrado para os itens A e B simultaneamente pelo produto do suporte individual de cada um dos elementos.

$$Levantamento = \frac{Suporte(A, B)}{Suporte(A) \times Suporte(B)}$$

Apesar de ser citado como uma das métricas utilizadas para associação, a biblioteca utilizada não permite a parametrização de valores mínimos, mas retorna os valores de levantamento obtidos para cada respectiva regra do resultado, podendo ser utilizado na análise das regras obtidas.

Para este trabalho os valores de lift foram considerados posteriormente à análise dos valores de confiança obtidos. Como exemplo, é possível citar casos onde obteve-se 100% de acoplamento entre empresas mas que podem possuir alto nível de aleatoriedade na participação. Isso pode revelar um indícios para verificar empresas criadas exclusivamente para fraudar licitações em específico.

2.6. **Análise das Regras**

Para permitir a análise das regras de associação geradas pelo algoritmo, inicialmente foi realizada uma busca por casos de empresas já condenadas por fraudes em licitações. O intuito desta busca foi compreender quais são os órgãos judiciais responsáveis por julgar tais casos e encontrar uma fonte para pesquisa de casos judiciais onde seja possível realizar buscas por CNPJ.

Através da análise, foi possível identificar que boa parte dos casos de fraude em licitações federais são julgados pelo Tribunal de Contas da União (TCU) e verificou-se também que o site Jusbrasil seria uma potencial fonte para busca pois retornou a maior quantidade documentos processuais. Além de possuir os dados processuais necessários, o site permite a busca de processos utilizando quaisquer termos que possam estar contidos

no texto do documento judicial.

Portando, uma vez selecionada a fonte para busca por CNPJ e identificado o órgão responsável pelo julgamento dos casos convenientes à pesquisa, um novo script para web crawling foi desenvolvido com o objetivo de buscar por processos das empresas encontradas em regras que obtiveram 100% de confiança.

O fluxo do script baseia-se nas seguintes etapas:

1. O script desenvolvido seleciona os CNPJs contidos em cada regra de associação;
2. Elabora strings de busca compostas pelo CNPJ, órgão responsável pelo processo e termos jurídicos que foram observados com frequência em casos de julgamento de fraude, resultando no padrão de texto "{CNPJ}+{TCU}+{Termo}";
3. A lista de termos elaborados é submetida de forma automatizada através da reprodução da requisição GET do sistema Jusbrasil;
4. Os resultados contidos na página resultante da busca são extraídos utilizando a linguagem de caminhos XML (XML Path Language - XPath), que define regras de sintaxe para permitir a seleção de partes do documento HTML conforme a hierarquia, atributos e informações das tags;
5. Exportação em formato CSV das URLs resultantes obtidas através das buscas realizada para cada regra de associação, bem como os termos que levaram ao resultado.

Devido à grande quantidade de regras obtidas no processamento, tal script favorece a verificação dos padrões de associação encontrados, já que a relação entre os participantes pode ter sido exposta em casos de fraude já julgados pelo Tribunal de Contas da União.

3. Resultados Obtidos

Para a abordagem utilizando o conjunto de dados com o total de licitações ocorridas no período estudado (2013 a 2019) nenhuma regra de associação foi obtida, pois não foi possível executar o algoritmo. Mesmo utilizando um servidor com 160GB de memória RAM observou-se que em conjuntos com milhões de transações, o algoritmo precisou alocar valores superiores a 256GB para processar os dados.

Para a abordagem separando as licitações em conjuntos de dados anuais a distribuição das regras ocorreu conforme apresentado na Tabela 2.

Tabela 2. Descrição das regras encontradas nos conjuntos de dados anuais

Ano	Nº de Regras	Suporte Médio	Confiança Média	Levantamento Médio	Ocorrência Média
2013	28	1,2%	72%	38,37	192
2014	144	1,4%	69%	17,70	464
2015	159	1,3%	66%	15,32	369
2016	57	1,3%	67%	23,55	363
2017	184	1,3%	57%	14,69	343
2018	14	1,3%	73%	37,42	449
2019	191	1,5%	50%	7,75	171

Dentre as regras extraídas, apenas 7 regras obtiveram 100% de confiança, até então adotado como principal métrica para medir a força da regra, porém das 7 empresas apenas 2 resultados foram obtidos pelas buscas de documentos, ambos utilizando o CNPJ de duas empresas concatenado aos termos "TCU" e "fraude". Porém, verificou-se que os documentos encontrados tratavam-se de falso positivos, pois continham casos de fraude cometido por outras empresas sendo relatados no mesmo documento em que os CNPJs contidos nas regras eram citados em casos não relacionados a fraude.

Para as regras com valores diversos na métrica de confiança foram obtidos os resultados exibidos na Tabela 3.

Tabela 3. Quantidade de documentos jurídicos encontrados em relação aos termos utilizados na busca

Termo	Documentos Encontrados
CNPJ + TCU + Conluio	5
CNPJ + TCU + Cartel	2
CNPJ + TCU + Fraude	4
CNPJ + TCU + Crime	1
CNPJ + TCU + Lavagem	1
CNPJ + TCU + Ilícito	10

Para verificação dos resultados, cada um dos documentos jurídicos foi analisado manualmente, objetivando verificar a existência de alguma relação entre as regras de associação e os documentos gerados por casos oficiais de fraudes já detectadas.

Dentre os resultados encontrados, os termos "Cartel", "Crime" e "Ilícito" retornaram apenas falso positivos, devido ao fato de que os termos fazem parte do texto que constitui o edital de grande parte dos processos licitatórios.

O resultado obtido utilizando o termo "Lavagem" foi também um falso positivo pois o termo estava presente como a descrição da prestação de serviço a ser realizada.

Dos resultados obtidos utilizando os termos "Conluio" e "Fraude" foram encontrados 5 processos que realmente se tratavam de casos onde ocorreram fraudes no processo licitatório, envolvendo 3 empresas diferentes que apesar de não terem sido relacionadas nos mesmos processos, estiveram relacionadas em uma mesma regra.

Para as regras de associação obtidas nos conjuntos de dados mensais foram possíveis localizar as mesmas regras dos conjuntos anuais e outras 2538 regras que obtiveram 100% de confiança, retornando 740 documentos judiciais, quantidade bastante expressiva para que análises manuais sejam realizadas.

Outra consideração a ser feita é que dentre as regras onde nenhum documento relacionado a fraude foi encontrado, houve uma empresa que foi foco de diversas relações com alto nível de confiança. Relações das quais foram obtidas através do conjunto de dados de granularidade anual, referente a 2019.

Para este caso em particular uma pesquisa manual foi realizada com a finalidade de obter mais informações a respeito da empresa em questão, como a existência de algum site, identidade visual, áreas de atuação e endereço.

Com os resultados obtidos foi possível observar que a empresa trabalha com mais de 20 áreas de atuação, não tratando de nenhum ramo de mercado específico. Possui capital inicial de R$150.000,00, porém não possui nenhum site ou qualquer identidade

visual como logomarca e está localizada em um imóvel abandonado, verificado através de imagens fornecidas pelo Google Maps.

4. Conclusão

A primeira conclusão obtida a respeito da aplicação do algoritmo de associação em licitações foi sobre a inviabilidade em conjuntos de dados possuindo milhões de transações, não apenas pela alta alocação de recursos de hardware, mas por dificultar a parametrização da métrica "Suporte". Uma vez que a mesma se baseia na frequência de uma transação em relação à quantidade total, até mesmo valores como 1% tornam-se inviáveis para a aplicação pois entre as 1.166.304 licitações que ocorreram entre 2013 e 2019, seria necessário que uma relação ocorra em ao menos 11.663 transações para que seja considerada uma regra.

Dentre as regras de associação analisadas, foi possível observar que boa parte das mesmas agruparam empresas concorrentes na mesma área de atuação, como já foi observado no trabalho de SILVA e RALHA(2010), regras para as quais apenas alguns documentos revelam histórico de fraudes. Este efeito de agrupamento ocorre porque os conjuntos de dados não foram estruturados por áreas de atuação portanto, dentre a ampla gama de eventos licitatórios, é natural que empresas pertencentes a um mesmo ramo de trabalho sejam relacionadas. Isso faz com que sejam obtidos grandes volumes de regras de associação, porém nem todas podem ser consideradas assertivas ou utilizadas como indícios para investigação.

Apesar da ocorrência de tais agrupamentos, também não é possível afirmar que as regras obtidas são totalmente inválidas, pois como foi demonstrado no tópico de resultados, haviam empresas relacionadas em uma mesma regra que já estiveram envolvidas em fraudes de licitações. Para este caso, mesmo que as empresas não tenham sido investigadas em um mesmo processo, o fato de que atualmente estão fortemente relacionadas em diversas licitações poderia ser considerado como indício para investigações nos certames arrematados pelas mesmas.

5. Trabalhos Futuros

Foi possível observar que, para o processo de descoberta de conhecimento elaborado, existem diversas oportunidades promissoras para melhorias e evolução do trabalho, das quais são possíveis citar:

1. Utilização de técnicas de classificação que permitam separar os conjuntos de dados por ramo de atuação. Desta forma, torna-se possível obter regras de associação fortes dentre os concorrentes que já atuam em um mesmo ramo. Tal combinação de técnicas pode se mostrar eficaz para detectar cartéis de empresas em nichos específicos como, por exemplo, o nicho de engenharia e obras de infraestrutura;
2. Execução e análise da técnica de associação em conjuntos de dados orientados a históricos de empresas específicas. Tal aplicação pode ser útil em mostrar relações fortes mesmo entre as empresas que são consideradas grandes fornecedoras com alto nível de concorrência;
3. Aprimoramento dos termos jurídicos utilizados na investigação do histórico de fraudes relacionadas às regras. Seria interessante agregar o ponto de vista de um profissional de direito para a elaboração de novas

listas de termos e órgãos julgadores;

4. Melhoria dos processos automatizados visando a criação de uma arquitetura integrada para total automatização do processo de descoberta de conhecimento, culminando em uma base de dados rica em indícios que possam ser investigados posteriormente;
5. Aplicação de técnicas de aprendizado de máquina que permitam a análise automática dos textos e documentos jurídicos obtidos, de forma a classificar os casos onde as empresas associadas pelas regras tenham sido consideradas culpadas por crimes em licitações. Permitindo assim, melhores ajustes dos parâmetros de associação e a aplicação de novas técnicas para classificar as regras, propriamente ditas, em tipos de padrões de comportamento em licitações que melhor descrevam as fraudes cometidas.

6. Referências

BATISTA, H. G.; NETO, J. S.; FARIELLO, D. (2015) "Nas licitações, a lei ainda é falha", https://oglobo.globo.com/economia/infraestrutura/nas-licitacoes-lei-ainda-falha-16152630, O Globo.

CORTÊS, S. C.; PORCARO, R. M.; LIFSCHITZ, S. (2002) "Mineração de Dados - Funcionalidades, Técnicas e Abordagens", PUC-Rio.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. (1996a) "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. (1996b) "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM.

KAUSAR, M. A.; DHAKA, V. S.; SINGH, S. K. (2013) "Web Crawler: A Review", International Journal of Computer Applications.

Portal da Transparência. (2004) "O que é e como funciona", http://www.portaltransparencia.gov.br/sobre/o-que-e-e-como-funciona.

SILVA, C. V. S.; RALHA, C. G. (2010) "Detecção de Cartéis em Licitações Públicas com Agentes de Mineração de Dados", Revista Eletrônica de Sistemas de Informação.