FERRAMENTA PARA AVALIAÇÃO DE PROCESSOS LICITATÓRIOS

Gabriel Batista Vieira[1], João Baptista Cardia Neto[2]

[1]Faculdade de Tecnologia de Garça "Dep. Julio Julinho Marcondes de Moura"

gabrielbvieira08@gmail.com, joao.cardia@fatec.sp.gov.br

Abstract. *Brazil lives a great awareness and fight against corruption, this has been evidenced by the popular movements and greater rigor in the investigations. Although there is a legislation requiring public administrations to publish their account installments, these data are not usually easily understood and, in this way, it is difficult for the society to discover the irregularities. The present work aims to use Data Mining and Machine Learning concepts in the bidding data of the city of Marília and to create a tool that can identify the existence of irregularities in these processes.*

Resumo. *O país vive uma grande conscientização e luta contra a corrupção, isso tem se evidenciado pelas movimentações populares e maior rigor nas investigações. Apesar de existir legislação que obriga as administrações públicas a publicarem suas prestações de conta, esses dados não são normalmente de fácil entendimento e, desta forma, dificulta a apuração da sociedade e descoberta de irregularidade. O presente trabalho visa empregar conceitos de Data Mining e Machine Learning nos dados de licitações da cidade de Marília e criar uma ferramenta que consiga identificar a existência de irregularidades nesses processos.*

1. Introdução

O momento atual da política Brasileira se dá, principalmente, pelo clamor popular do combate à corrupção e adoção de medidas com maior transparência nos gastos públicos. No Brasil a corrupção chega a desviar R$ 200 bi por ano (ESTADÃO, 2015), dinheiro esse que poderia ser aplicado em áreas deficientes, como saúde, educação, segurança e previdência.

Segundo (CARDOSO, 1986) em um estado de direito, a administração pública estatal é submetida ao estabelecido na Carta Magna. Estão contidos nela, os princípios da legalidade, moralidade, impessoalidade, eficiência e publicidade, conforme previsto no art. 37 da Constituição Federal do Brasil.

A licitação é o procedimento previsto no ordenamento jurídico para a Administração Pública eleger a proposta mais vantajosa financeiramente. Este instituto está previsto na Constituição Federal e foi regulamentado pela Lei 8.666/93.

O princípio basilar do direito administrativo é o princípio da legalidade, que vincula a administração às leis existentes e a submete ao controle jurisdicional para exame da observância das leis no exercício da sua competência (MAURER, 2006). Segundo (SPECK, 2000) os diversos escândalos no Brasil aumentaram o debate sobre corrupção. Esse tema obteve densidade a partir da divulgação dos escândalos recentes e levou à avaliação mais aprofundada de possíveis falhas processuais e estruturais que possibilitam

esquemas inidôneos. Desta forma, o autor observa os custos e as consequências sociais da corrupção.

Segundo dados de (TRANSPARÊNCIA INTERNACIONAL, 2010), atos de corrupção podem elevar os custos de uma obra em até 50%. Acrescenta que o primeiro passo na tentativa de manter o controle sob estes atos, é reconhecer como as fraudes ocorrem. Desta forma, as possibilidades de desvios, ou fraudes, são apresentadas segundo as fases da contratação de obras Públicas.

Argumenta (Tanzi, 1997) que o grau de complexidade de projetos de obras públicas, apresentam maiores relações para com atos de corrupção em um país. Exemplifica com dados da cidade de Milão na Itália, que teve redução de 52% nos custos de construção de metrô e 59% de redução na construção do terminal do aeroporto local, após a descoberta de um grande esquema de corrupção. A redução nos custos foi provocada pela queda no custo das obras.

2. Proposta

O presente trabalho objetiva localizar na cidade de Marília-SP, possíveis fraudes em editais públicos para contratação de serviços diversos.

O estudo do caso foi realizado com auxílio da MATRA "Marília Transparente", uma Organização Não Governamental com a missão de fiscalizar os poderes municipais e, assim manter a transparência na gestão pública municipal. Com esse auxílio foram obtidas informações relacionadas a identificação destes agentes fraudulentos em diversos domínios.

Considerando que os dados fornecidos nos portais oficiais do município, apresentam-se de maneira desagregada e pouco semânticos, é necessária uma abordagem para construção de uma base de dados. Após a construção desta base, é desejado aplicar algoritmos de aprendizado de máquina para realizar verificações em dados de maneira eficiente, otimizando o processo de verificação manual.

3. Construção da Base de Dados

Visando a obtenção de dados diversos foi construída uma base, utilizando arquivos disponibilizados nos portais públicos.

Foi desenvolvido um *crawler* para extração dos dados relevantes à licitação. Para seu desenvolvimento foi utilizada a linguagem de programação Python, e a biblioteca Beatiful Soup.

Um sistema de obtenção de informação da Web, conhecido como *crawler*, é um componente de software, utilizado desde o início da Internet (MCBryan, 1994). A sua complexidade possui diversas variáveis, dependente dos requisitos do sistema onde estará integrado, utilizado comercialmente, ou em aplicações acadêmicas.

Foram obtidos aproximadamente oito mil arquivos, entre editais abertos, concluídos, licitações, e arquivos anexos destes documentos. Tudo isso soma aproximadamente seis gigabytes de arquivos.

4. Extração de Informação

Considerando a grande quantidade de dados a qual o sistema está sujeito a processar, torna-se necessária uma abordagem para recuperação destes dados desagregados, torná-los semânticos, e persisti-los em formato que possibilite ao algoritmo um acesso rápido e eficiente.

Um sistema de extração de informação visa extrair com especificidade tipos de informações de textos (Scarinci, 1997). O particionamento do texto requisitado permite que as partes consideradas não pertinentes ao domínio sejam ignoradas efetivamente.

Extração de Informação é uma abordagem consideravelmente mais robusta e prática em comparativo ao Processamento da Linguagem Natural tradicional e tem obtido sucesso (Riloff, 1994).

Processamento de linguagem natural (NLP), pode ser definido como uma abordagem que visa tratar com auxílio computacional aspectos diversos da comunicação humana, sons, palavras, estruturas, significados, contextos, formatos e referências. Desta forma, NLP é conhecido como o processo que possibilita ao dispositivo eletrônico, não necessariamente em todos os níveis, mas a possibilidade de comunicar-se em linguagem humana.

Em comparação a NLP, extração de informação é computacionalmente mais simples, exigindo menos recurso computacional. Isso se deve ao fato de que muitas frases ou trechos inteiros podem ser ignorados, caso não sejam considerados pertencentes ao domínio de interesse (Riloff, 1994). Devido ao fato do sistema estar somente direcionado a trechos em específico do texto, pertencente ao domínio, alguns problemas pertencentes a NLP são resolvidos com facilidade. (por exemplo, resolução de ambiguidades).

Segundo (Cowie, 1996), Extração de Informação pode ser definida como qualquer processo, que visa combinar e obter estruturas que são recuperadas em um ou vários textos.

Para extração de informações dos dados obtidos e posicionados na base de dados previamente construída foi utilizada a biblioteca Python PDFMiner. A utilização desta biblioteca possibilita a retirada dos dados persistidos em formato PDF e a conversão para diversos formatos ou salvá-los em novos arquivos após processamento.

Neste trabalho os dados, após sua extração, são persistidos em formato de texto puro (txt), dando maior flexibilidade de leitura e visando a utilização destas informações em algoritmos de *Machine Learning*.

5. *Bag of Words*

O modelo *Bag of words* é um dos modelos para representação de dados textuais conhecido como espaço-vetorial, onde cada documento é um vetor em um espaço multidimensional e cada dimensão representa um termo da coleção (Feldman, 2006). Para isto, é possível inserir os dados em uma estrutura *bag of words*, e os textos são tratados de modo independente.

Bag of Words, é uma forma de representação de documentos que possibilita a utilização de algoritmos de aprendizado em seus dados. Cada elemento desta representação é equivalente a cada atributo (Palavra) pertencente ao documento. Podem ser localizados um ou mais atributos (a_x, x = 1,2,3, ... n), para um ou mais documentos

(d_y = 1,2, 3...m). Esses atributos são de diversos tipos, podendo ser um booleano que indica a aparição do atributo no documento.

Neste trabalho, o *bag of words* é utilizado para auxílio da tarefa de classificação de documentos. Após a recuperação dos documentos requiridos na base de dados, e a extração das informações, os dados obtidos são inseridos na estrutura e são identificados os atributos chave que possibilitam categorizar estes documentos de acordo com o tipo de material tratado neste documento (Editais).

Visando identificar a classificação do documento em questão, são visualizados os termos mais comuns presentes, por exemplo: Um edital que visa adquirir materiais escolares, mais especificamente cadeiras, terá em seu corpo o termo cadeiras diversas vezes, desta forma, entende-se que o documento trata da aquisição do objeto cadeiras, pela incidência do termo. Ao adquirir esta informação, é possível localizar junto a ela a possível quantidade, e alguma especificidade para aquisição deste objeto, após construção do *bag of words*, o treinamento de um classificador é facilitada.

A tarefa de classificação torna-se imprescindível, para identificar possíveis fraudes, pois é necessário conhecer o domínio específico do documento, para identificar de acordo com editais pertencentes a domínio semelhante presentes na base, dados incomuns para a categoria em questão.

6. Aprendizado Supervisionado

Segundo (Monard e Baranauskas, 2003), aprendizado de Máquina é uma área da inteligência artificial que objetiva o desenvolvimento de técnicas computacionais sobre aprendizado e a construção de sistemas capazes de obter conhecimento de maneira individual e automática. Define também que um sistema de aprendizado é um algoritmo capaz de tomar decisões (ou auxiliar na tomada de decisão), baseando-se em experiências oriundas de tentativas bem-sucedidas anteriormente.

A obtenção de conclusões genéricas sobre um conjunto específico de documentos pode ser definida como indução, podendo ser definida como o raciocínio que é originado de um conceito específico, sendo generalizado, da parte para o todo. Através da inferência indutiva, um conceito pode ser aprendido utilizando as informações obtidas sobre exemplos apresentados anteriormente. (Monard e Baranauskas, 2003).

Aprendizado indutivo permite a divisão em: Aprendizado supervisionado e não-supervisionado. Neste trabalho, serão utilizados algoritmos de aprendizado supervisionado.

No aprendizado supervisionado são necessários exemplos de treinamento para quais o rótulo da classe associada é conhecido, estes exemplos são apresentados ao algoritmo de aprendizado. Normalmente, um exemplo é constituído por um vetor de características, atributos e o rótulo da classe associada.

O objetivo do algoritmo de indução é a construção de um classificador que possibilite determinar a classe de novos exemplos, ainda não rotulados.

6. Resultados e trabalhos futuros

Até o presente momento, foram determinados os algoritmos de extração de informação, e construída a base de dados com documentos públicos obtidos nos portais municipais, totalizando aproximadamente oito mil arquivos.

Após a extração dos documentos, estão sendo implementados os algoritmos para o bag of words, porém, foi visualizada a necessidade de construção de expressões regulares para organização dos dados obtidos, de modo a corrigir a desestruturação dos documentos.

Com a conclusão desta etapa, serão utilizados algoritmos de *Machine Learning*, visando a classificação destes documentos, possibilitando a avaliação de superfaturamento de obras (Primeiro método de avaliação), utilizando documentos comuns e com aspectos semelhantes, desta forma, possibilitando que sejam encontrados dados financeiramente incomuns nestes editais, podendo ser, ou não, possíveis fraudes.

Referências

Boulic, R. and Renault, O. (1991) "3D Hierarchies for Animation", In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons ltd., England.

CARDOSO, Antonio; Manoel Bandeira. A Magna Carta – Conceituação e Antecedentes. Brasília, 1986.

COWIE, Jim; LEHNERT, Wendy. Information Extraction. Communications of the ACM, New York, 1996.

ESTADÃO (2015). Corrupção desvia R$ 200 bi, por ano, no Brasil, diz coordenador da Lava Jato Disponível em:http://politica.estadao.com.br/blogs/fausto-macedo/corrupcao-desvia-r-200-bi-por-ano-no-brasil-diz-coordenador-da-lava-jato/. 04/10/2017.

FELDMAN, R. and SANGER, J, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2006.

MAURER, Hartmut. Direito administrativo geral. (Tradução de Luís Afonso Heck). Barueri, 2006.

MCBRYAN, O. A. GENVL and WWWW: Tools for taming the Web. In Proceedings of the First International World Wide Web Conference, 1994.

Monard, M. C. e Baranauskas, J. A. (2003). Sistemas Inteligentes: Fundamentos e Aplicações, Capítulo 4: Conceitos sobre Aprendizado de Máquina.

PINKERTON, B. Finding What People Want: Experiences with the WebCrawler. In Proceedings of the Second International World Wide Web.

RILOFF, Ellen; LEHNERT, Wendy. Information Extraction as a Basis for High-Precision Text Classification, 1994.

SCARINCI, Rui G; PALAZZO, José M. SES – Sistema de Extração Semântica de Informações. Porto Alegre: CPGCC da UFRGS, 1997.

SPECK, BRUNO WILHELM. Mensurando a corrupção: uma revisão de dados provenientes de pesquisas empíricas. Cadernos Adenauer no 10. São Paulo: Fundação Konrad Adenauer, 2000.

TANZI, V. Corruption Around the World: Causes, Consequences, Scope, and Cures. Staff Papers. International Monetary Fund, 1997.

TRANSPARÊNCIA INTERNACIONAL. Public procurement. Disponível em: https://www.transparency.org/topic/detail/public_procurement. 04/10/2017.