Extração de Informação e Mineração de Dados no Diário Oficial de Pernambuco

Information Extraction and Data Mining in the Official Gazette of Pernambuco

Ricardo Batista das Neves Junior[1] orcid.org/0000-0001-9538-6505
Weverton Fernandes de Medeiros Melo[1] orcid.org/0000-0003-3429-2892
Roberta Andrade de Araújo Fagundes[1] orcid.org/0000-0000-7172-4183
Alexandre Magno Andrade Maciel[1] orcid.org/0000-0003-4348-9291

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: rbnj@ecomp.poli.br

Resumo

O uso de técnicas de mineração de dados tem sido amplamente utilizado para o processamento de uma grande quantidade de dados documentados. No entanto, atualmente, poucos aplicativos mostraram-se efetivos para extrair e minerar dados em diários oficiais. Este trabalho tem como objetivo apresentar um método para construção de uma aplicação que usa um algoritmo para indexar conteúdo da base do Diário Oficial do Estado de Pernambuco, transformando as informações anteriormente disponíveis no texto para o formato estruturado, para aplicar uma Mineração de Dados. Para o desenvolvimento do método, a linguagem Java foi utilizada, com a possibilidade do aplicativo web. O estudo de caso baseou-se em documentos publicados no Diário Oficial de janeiro de 2007 a abril de 2017. Os resultados mostram que é possível indexar e estruturar esses dados, mas ainda há necessidade de uma melhor padronização dos dados.

Palavras-Chave: Mineração de Dados; Diário Oficial; Árvore de Decisão.

Abstract

The use of Data Mining techniques has been widely applied for processing a high amount of documented data. However, to date, there are very few effective applications for extracting and mining data in official journals. This work aims to present a method for the construction of an application that uses an algorithm to index contents of the base of the Official Gazette of the state of Pernambuco, transforming the information previously available in the text to structured format, to apply a Mining of Data. For the development of the method, the Java language was used, with the possibility of the web application. The case study was based on documents published in the Official Gazette from January 2007 to April 2017. The results show that it is possible to index this data and give meaning to it, but there is still a need for a better standardization of the data.

Key-words: *Data Mining; Official Diary; Decision Tree.*

http://dx.doi.org/10.25286/repa.v3i3.892

1 Introdução

O Diário Oficial de Pernambuco (DOE) foi criado no ano de 1924, ano da criação da Companhia Editora de Pernambuco (CEPE) **[1]**. A disponibilização do DOE tem como objetivo manter público aos cidadãos informações pertinentes aos poderes Executivo, Legislativo e Judiciário. Graças ao DOE, todos os atos administrativos do estado tornam-se públicos e alcançáveis por qualquer cidadão aumentando a transparência entre o governo e os indivíduos. Atualmente o DOE é disponibilizado gratuitamente no site da CEPE (*http://cepe.com.br*). As informações disponibilizadas pelo Diário Oficial do Estado podem ser consideradas como dados não estruturados, pois trata-se de um arquivo no formato PDF, com um texto corrido, imagens agregadas ao texto e difícil leitura computacional no contexto da divisão das sessões.

Atualmente, a Controladoria do Estado de Pernambuco, têm a obrigação de utilizar o DOE para obter algumas informações referentes à Inquéritos Administrativos, Processos Administrativos, Sindicâncias Administrativas e entre outras. Para a obtenção destas informações, é necessário efetuar o *download* dos Diários Oficiais, abrir estes documentos, utilizar o atalho *Control Find* (Ctrl + F), pesquisar pelo termo desejado e retirar a informação. Pode-se notar que este é um processo que gastar um tempo que poderia ser investido em outras atividades.

Este trabalho propõe uma solução de extração de informações associada a um mecanismo de mineração de dados para predição dos resultados das sindicâncias, com o objetivo de automatizar e otimizar este processo de acompanhamento. Para isto será desenvolvido um algoritmo capaz de ler a base de dados (Diários Oficiais), pesquisar pelos termos necessários, extrair as informações pertinentes ao conteúdo, estruturar os dados e implementado um motor de inferência baseado em árvore de decisão para predição de resultado das sindicâncias.

Este trabalho está organizado da seguinte maneira: A sessão 2 fala sobre a fundamentação teórica, onde pode-se entender o que é mineração de dados e observar alguns trabalhos que aplicam técnicas de mineração. A sessão 3 elucida sobre materiais e métodos, tais como preparação/transformação dos dados e técnica escolhida. A sessão 4 fala sobre os experimentos realizados bem como os resultados obtidos. A sessão 5 mostra as conclusões e considerações finais sobre o trabalho.

2 Fundamentação Teórica

2.1 Mineração de Dados

A mineração de dados vem atraindo a atenção na indústria da informação e na sociedade como um todo nos últimos anos, devido à grande disponibilidade de enormes quantidades de dados e a iminente necessidade de transformar esses dados em informações e conhecimentos úteis. As informações e os conhecimentos adquiridos podem ser utilizados em aplicações que vão desde análise, detecção de fraude e fidelização de clientes, controle de produção e exploração. A mineração de dados pode ser vista como resultado da evolução natural da informação tecnologia.

A mineração de dados é o processo de extração de conhecimento em grandes quantidades de dados. Ela está inserida em um processo maior denominado Descoberta de conhecimento (KDD – *Knowledge Discovery in Database*) **[3,8]**.

A descoberta de conhecimento como processo consiste numa sequência iterativa de algumas etapas tais como: (*i*) Limpeza de dados - para remover ruídos e dados inconsistentes; (*ii*) Integração de dados - onde várias fontes de dados podem ser combinadas; (*iii*) Seleção de dados - onde os dados relevantes para a tarefa de análise são recuperados da base de dados; (*iv*) Transformação de dados - onde os dados são transformados ou consolidados em formulários para mineração executando operações de resumo ou agregação, por exemplo; (*v*) Mineração de dados - um processo essencial onde são aplicados métodos inteligentes para extrair padrões de dados; (*vi*) Avaliação de padrões - para identificar os padrões verdadeiramente interessantes que representam o conhecimento com base em algumas medidas de interesse e (*vii*) Apresentação do conhecimento - onde técnicas de visualização e de representação do conhecimento são usados para apresentar o conhecimento minado ao usuário **[3]**.

Os passos *i* a *iv* são formas diferentes de pré-processamento de dados, onde os dados são preparados para a mineração. O passo de mineração de dados pode interagir com o usuário ou uma base de conhecimento. Os padrões interessantes são apresentados ao usuário e

podem ser armazenados como novos conhecimentos na base de conhecimento.

Concordamos que a mineração de dados é um passo no processo de descoberta de conhecimento. No entanto, na indústria, na mídia e no ambiente de pesquisa de banco de dados, o termo mineração de dados está se tornando mais popular do que o longo prazo de descoberta de conhecimento a partir de dados **[3]**.

Em princípio, a mineração de dados deve ser aplicável a qualquer tipo de repositório de dados, bem como a dados transitórios, como fluxos de dados. Os sistemas de banco de dados avançados incluem bancos de dados objeto-relacionais e bancos de dados específicos orientados a aplicativos, como bancos de dados espaciais, bancos de dados de séries temporais, bancos de dados de texto e bancos de dados multimídia. Os desafios e técnicas de mineração podem diferir para cada um dos sistemas de repositório **[3]**.

As funcionalidades de mineração de dados são usadas para especificar o tipo de padrões a serem encontrados nas tarefas de mineração de dados. Em geral, as tarefas de mineração de dados podem ser classificadas em duas categorias: descritiva e preditiva. As tarefas de mineração descritivas caracterizam as propriedades gerais dos dados no banco de dados. As tarefas de mineração preditivas realizam inferência nos dados atuais para fazer previsões.

Em alguns casos, os usuários podem não ter ideia sobre quais tipos de padrões em seus dados podem ser interessantes e, portanto, podem gostar de procurar vários tipos diferentes de padrões em paralelo. Assim, é importante ter um sistema de mineração de dados que pode explorar vários tipos de padrões para acomodar diferentes expectativas ou aplicações de usuários. Além disso, os sistemas de mineração de dados devem ser capazes de descobrir padrões em várias granularidades (isto é, diferentes níveis de abstração). Os sistemas de mineração de dados também devem permitir que os usuários especifiquem dicas para orientar ou focalizar a busca por padrões interessantes. Como alguns padrões podem não ser válidos para todos os dados do banco de dados, uma medida de certeza ou "confiabilidade" é geralmente associada a cada padrão descoberto **[3]**.

2.2 Árvore de Decisão

A Árvore de Decisão é uma técnica de classificação de dados dentro da Mineração de Dados (*Data Mining*). Podem ser usadas em conjunto com outras tecnologias de regras, mas são as únicas a apresentar os resultados hierarquicamente (com priorização). Nela, o atributo mais importante é apresentado na árvore como o primeiro nó, e os atributos menos relevantes são mostradas nos nós subsequente. A vantagem principal das Árvores de Decisão é a tomada de decisões levando em consideração os atributos mais relevantes, além de compreensíveis para a maioria das pessoas. Ao escolher e apresentar os atributos em ordem de importância, as Árvores de Decisão permitem aos usuários conhecer quais fatores mais influenciam os seus trabalhos.

Está técnica é uma representação simples das informações e um caminho eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados **[9]**. Uma Árvore de Decisão utiliza a estratégia chamada dividir-para-conquistar, ou seja, um problema complexo é decomposto em subproblemas mais simples. Repetidamente, a mesma estratégia é aplicada a cada subproblema **[10]**. A capacidade de discriminação de uma Árvore de Decisão vem das características de divisão do espaço definido pelos atributos em subespaços e da associação de uma classe a cada subespaço.

Segundo **[9]**, as Árvores de Decisão são compostas de: nós, que representam os atributos, e de ramos, oriundos desses nós e que recebem os valores possíveis para esses atributos (cada ramo descendente corresponde a um possível valor desse atributo). Nas árvores existem nós folha, que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está associada a uma classe. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

3 Materiais e Métodos

3.1 Preparação dos Dados

Durante a preparação dos dados para a realização da mineração de dados foi considerado

http://dx.doi.org/10.25286/repa.v3i3.892

algumas etapas descritas em **[2]**. Inicialmente, foram utilizados Diários Oficiais do Estado referentes ao Poder Executivo publicados no período de janeiro de 2007 a abril de 2017. Essa base de dados foi lida e pré-processada. A fase de pré-processamento dos dados foi dividida nas seguintes etapas: Tokenização, Remoção de stop *word* e *Stemming*.

No processo de tokenização foi definida algumas palavras-chave tais como: "Sindicância Administrativa", "Sindicância Investigativa", "Sindicância Disciplinar" e "Sindicância Administrativa Disciplinar". O objetivo do algoritmo é buscar por essas palavras-chave e entregar informações referentes aos termos solicitados. Dentre o conteúdo retornado pelo algoritmo estão inclusos os termos: "data", "palavra-chave", "informação", "número portaria", "tipo comissão", "secretaria", "lei número", "vencimento" e "resultado". Onde "data" é a data de publicação do DOE, "palavra-chave" é qual palavra-chave é correspondente à informação retornada, "informação" é todo o parágrafo o qual a palavra chave está inserida, "número portaria" é o número da portaria envolvida na sindicância, "tipo comissão" é o tipo da comissão da sindicância que pode ser comissão específica ou comissão permanente, "secretaria" é qual secretaria do Estado está envolvida na sindicância retornada, "lei número" é qual o número da lei é referente à sindicância, "vencimento" é a data de vencimento da sindicância e "resultado" é em que resultou a sindicância (i.e. processo arquivado, funcionário repreendido ou abertura de um inquérito administrativo).

Não houve dificuldades para remover as palavras desprezíveis, dado que é necessário extrair uma baixa quantidade de informações, em comparação com à quantidade de informação existente no DOE. Então, o algoritmo automaticamente despreza tudo que não há relação com as palavras-chave.

O processo de *Stemming* consiste em reduzir ao radical algumas palavras que se deseja buscar com o objetivo de identificar a palavra independente do tempo verbal ou se a palavra está no gerúndio, infinitivo ou particípio. Ao encontrar uma palavra-chave no DOE, o algoritmo precisa verificar se a sindicância encontrada tem um resultado conclusivo (i.e. funcionário repreendido, processo arquivado ou abertura de um inquérito administrativo), se houver as palavras-chave podem aparecer no algoritmo de formas diferentes (e.g. "arquivado", "arquivou-se", "arquivando", "repreendido", "repreensão" e etc.), então, para encontrar o resultado em qualquer terminação da palavra, o resultado foi buscado por "arquiv", "repreen", "inquérito". Na busca pelos resultados, além de utilizar a técnica *Stemming*, foi realizada algumas variações nas palavras com todas as letras minúsculas, a primeira letra maiúscula e todas as letras maiúsculas.

3.2 Transformação dos Dados

O algoritmo de extração de informação, através das regras estabelecidas, conseguiu formar uma base de dados com 339 registros. Infelizmente, a falta de consistência e padrão nas publicações do Diário Oficial afetou na formação da base e gerou alguns registros com campos vazios, necessitando realizar transformações nos dados para prepara-los para a mineração.

Primeiramente, foram filtrados os resultados, selecionando apenas os que tinham sindicância fechada (inquérito administrativo, arquivado e repreensão). Em seguida, foram eliminados os registros que estavam com o campo "secretaria" vazio. Por último, os atributos que não seriam utilizadas para a classificação foram retirados, restando apenas os campos "Secretaria", "Tipo de Comissão" e "Resultado".

Após a transformação, a base contou com 40 registros, onde o campo "Tipo de Comissão" foi alterado para a forma binária (comissão permanente = 0 e comissão especifica = 1) e "Secretaria" em forma numérica (ADAGRO = 1, DETRAN = 2, HSE = 3, IRH = 4, JUCEPE = 5, SAD = 6, SASSEPE = 7, SCGE = 8, SDS = 9, SES = 10, SUAPE = 11 e UPE = 12).

3.3 Técnica Utilizada

A base de dados utilizada foram os Diários Oficiais de Pernambuco, após realizar a extração de informação e estruturação dos dados, restou uma base de dados com apenas 339 registros, destes registros, a minoria dos campos estavam preenchidos valor do "resultado" (que informa o resultado final da sindicância). Como o objetivo do trabalho é prever qual será o resultado de uma sindicância, existiam apenas 40 registros disponíveis para executar a previsão.

Diante deste contexto, a base de dados foi preparada com intuito de relacionar as secretarias e comissões com os resultados das sindicâncias, para que fosse possível encontrar uma relação com essa classificação. Para isso, foi aplicado a técnica de Árvore de Decisão utilizando o software Weka, que já possui uma biblioteca para a mesma e é muito simples para implementar. Os tipos de árvore utilizados foram a *J48* e *RandomTree*. O *J48* é uma implementação do algoritmo C4.5 dentro do programa Weka, que gera uma árvore de classificação, onde, a cada nó, o algoritmo escolhe um atributo que irá subdividir de forma mais eficiente o conjunto de dados em subconjuntos homogêneos e qualificados por sua classe. Já o *RandomTree* constrói a árvore de classificação escolhendo K atributos de forma aleatória em cada nó, não realizando a poda, o que acaba plotando uma árvore grande.

4 Experimentos

A técnica aplicada foi capaz de relacionar as variáveis e classificar as possibilidades, onde na Árvore *J48* o tipo de comissão é o atributo mais importante, enquanto na *RandomTree* a secretaria é o atributo mais importante. Como a base de dados foi transformada, transformando campos texto em valores numéricos, as árvores fazem a verificação do valorcontido no campo, por exemplo: $Secretaria >= 8$ (SCGE = 8, SDS = 9, SES = 10, SUAPE = 11 e UPE = 12) e $Tipo\ Comissao > 0$ (comissão especifica = 1).

Dada o tipo de comissão e a secretaria, as Árvores de Decisão realizam a previsão do resultado de uma sindicância, essa informação pode ser muito valiosa para o tomador de decisão

Da Controladoria do Estado de Pernambuco, pois, de uma determinada secretaria sempre arquivar as sindicâncias, ou qual tipo de comissão é mais propensa a arquivar um processo, abrir inquérito administrativo ou repreender um funcionário.

4.1 Árvore J48

A árvore J48 é uma técnica simples, logo, apresenta como vantagem um menor custo computacional. Entretanto, em seu resultado, além de alcançar uma menor taxa de acerto em relação à árvore *RandomTree*, apresenta também um menor índice Kappa, que é o índice que reflete a confiabilidade do modelo, ou seja, quanto maior o índice Kappa, mais confiável é o modelo. A Figura 1 mostra a árvore gerada pelo algoritmo J48, pode-se notar a baixa complexidade da árvore gerada obtendo uma taxa de acerto de 70% com um índice Kappa de 0.4217. O valor do Erro médio absoluto alcançado foi de 0.1759 e o Erro médio quadrático igual a 0.2966.

4.2 Árvore *RandomTree*

A árvore *RandomTree* traz como desvantagem um maior custo computacional em relação à J48, mas, a desvantagem é compensada pelo seu resultado significativamente superior em termos de Taxa de acerto e índice Kappa. Na Figura 4 pode-se visualizar que o algoritmo *RandomTree* gerou uma árvore maior, mais complexa e com maior riqueza de detalhes em relação à J48. A aplicação de um grande conjunto de dados nessa técnica pode resultar em um alto custo computacional. O resultado obtido é superior, com uma taxa de acerto de 80% com um índice Kappa de 0.624. O valor do Erro médio absoluto foi de 0.09 e o valor do Erro médio quadrático alcançado igual a 0.2121.

http://dx.doi.org/10.25286/repa.v3i3.892

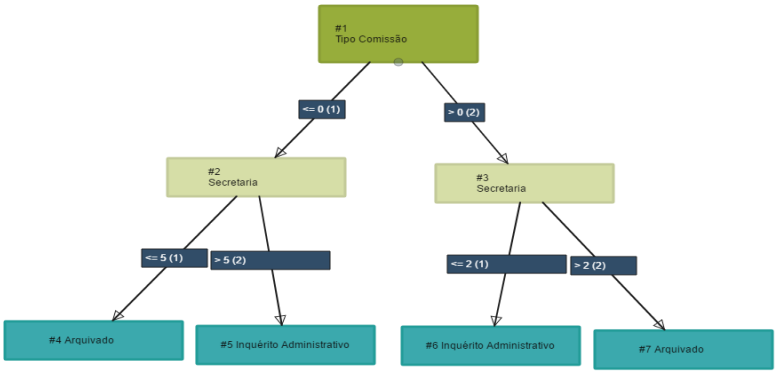

Figura 1:Árvore de Decisão - J48.

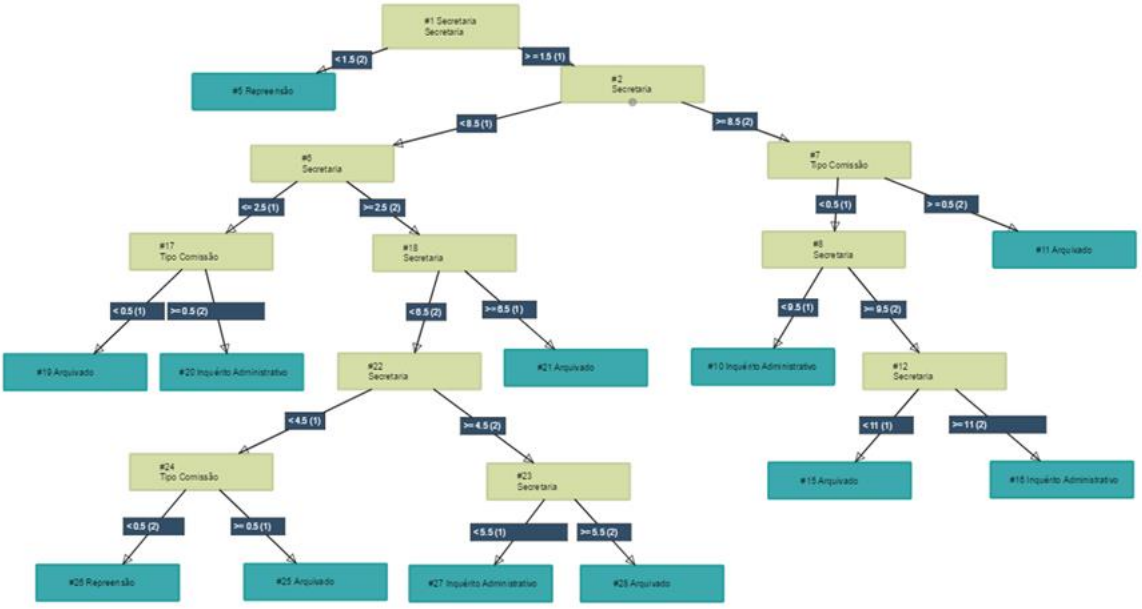

Figura 2:Árvore de Decisão – *RandomTree.*

5 Conclusões

Este trabalho entrega um projeto piloto que poderá ser expandido para Diários Oficiais de outros estados. Desenvolveu-se um motor de inferência para extração de informações de texto corrido e aplicou-se a técnica de mineração de dados conhecida como Árvore de Decisão a fim de facilitar a visualização, organizar os documentos publicados no Diário Oficial do Estado de Pernambuco e retirar informações relevantes a

partir dos dados. Através das etapas de pré-processamento e extração propostas pelo trabalho, pôde-se recuperar informação, antes apresentada em formato de linguagem natural para posteriormente transforma-la em uma base de dados possível de ser persistida. A etapa de mineração de dados demonstrou que é possível classificar e relacionar os dados, gerando informações que podem ser de extrema relevância para os tomadores de decisão.

Caso os órgãos desejem um algoritmo com melhores resultados, é necessário realizar uma padronização mínima dos Diário Oficiais para que a extração seja mais eficiente. Além disso, existe uma necessidade de um maior estudo das regras de padrões nos termos exibidos no Diário Oficial.

Referências

[1] COMPANHIA EDITORA DE PERNAMBUCO. A Cepe.**CEPE**. Disponível em: <https://www.cepe.com.br/index.php/cepe.html>. Acesso em: 24 abr. 2017.

[2] PATEL, Falguni N.; SONI, Neha R. Text mining: A Brief survey. **International Journal of Advanced Computer Research**, v. 2, n. 4, p. 243-248, 2012.

[3]JIAWEI, Han;PEI, Jian; KAMBER, Micheline.**Data mining: concepts and techniques**. Amsterdã: Elsevier, 2011.

[4]IBM Knowledge Center. Disponível em: <https://www.ibm.com/support/knowledgecenter/pt-br/SS3RA7-17.1.0 /modeler-crispdm-ddita/clementine/crisp-help/crisp-overview.html>. Acesso em: 24 abr. 2017.

[5]BHARANIPRIYA, V.; PRASAD, V. Kamakshi. Web content mining tools: a comparative study. **International Journal of Information Technology and Knowledge Management**, v. 4, n. 1, p. 211-215, 2011.
.

[6] Revista de Ciências Exatas e Tecnologia, v. 3 n. 3, 2008.

[7]MORAIS, Edison Andrade M.; AMBRÓSIO, Ana Paula L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

[8]FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory;SMYTH, Padhraic. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: International Conference on Knowledge Discovery and Data Mining, 2., 1996, Portland. **Proceedings**... Portland: KDD, 1996. p. 82-88.

[9]GARCIA, Simone C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. In: SEMANA ACADÊMICA, 2000,Rio Grande do Sul. **Anais**.... Rio Grande do Sul: UFRGS, 2000.

[10]Gama, J. **Árvores de decisão**. 2000. Disponível em:<http://www.liacc.up.pt/~jgama/Mestrado/ECD1/Arvores.html>. Acesso em: 14 ago. 2002.

[11]QUINLAN, J.Ross. **C4.5**: programs for machine learning. Sydney, Australia: Morgan Kaufmann Publishers, 1993. p. 302.

http://dx.doi.org/10.25286/repa.v3i3.892