Rebecin
REVISTA BRASILEIRA de
EDUCAÇÃO em CIÊNCIA da INFORMAÇÃO

MINERAÇÃO DE DADOS EM BASES JURÍDICAS: UM ESTUDO DE CASO

Talita de Souza Rampão
Universidade Federal do Paraná (UFPR)
Brasil

Denise Fukumi Tsunoda
Universidade Federal do Paraná (UFPR)
Brasil

RESUMO

Estudo de caso sobre mineração de dados aplicada a uma base de dados jurídica contendo processos cíveis de direito do consumidor com enfoque em: tarifa, tarifa e dano moral, revisional, indenizatória e outras. Objetiva a aplicação de técnicas de mineração de dados na área jurídica para verificar a existência de padrões de decisões judiciais de acordo com o Estado em que tramita o processo. Constitui-se de um estudo de caso com pesquisa descritiva, finalidade aplicada e abordagem quantitativa. Realiza a aplicação das tarefas de classificação e associação por meio dos métodos Apriori, PART, *Decision Table*, J48 (C4.5) e REPTree. Demonstra que é possível identificar padrões de decisões judiciais de acordo com o órgão julgador, tipo de ação e região que tramita o processo. Propõe a análise e continuidade do estudo para verificar a aplicação de técnicas de mineração em outras bases de dados jurídicas, a fim de validar a proposta e comparar as variações nos resultados obtidos.

Palavras-Chave: Direito; Gestão da Informação; Descoberta de Conhecimento em Bases de Dados; Mineração de Dados; Tomada de Decisão.

DATA MINING IN LEGAL BASIS: A CASE STUDY

ABSTRACT

Case study on data mining applied to a legal database containing civil cases of consumer law focus on: tariff, tariff and moral damages, revisional, indemnification and others. It aims to apply data mining techniques in the legal area to verify the existence of patterns or tendencies of judicial decisions according to the State in which the process is being processed. It is a case study with descriptive research, applied purpose and quantitative approach. It performs the application of classification and association tasks through the Apriori, PART, *Decision Table*, J48 (C4.5) and REPTree methods. It shows that it is possible to predict trends in judicial decisions according to the adjudicating body, type of action and region that processes the process. It proposes the analysis and continuity of the study to verify the application of mining techniques in other legal databases, in order to validate the proposal and compare the variations in the results obtained.

Keywords: Law; Information Management; Knowledge Discovery in Database; Data Mining; Decision-Making.

1 INTRODUÇÃO

A Era da Informação trouxe mudanças no paradigma da sociedade, facilitando o acesso, uso e compartilhamento instantâneo das informações com o auxílio das Tecnologias de Informação e Comunicação (TIC). Contudo, também trouxe consigo o excesso informacional, ou seja, há informações demais e falta de tempo para analisá-las, tornando cada vez mais complexo o processo de tomada de decisão.

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

De acordo com Sidney (2010), a grande quantidade de dados torna a análise humana onerosa. Por outro lado, métodos tradicionais de recuperação de dados, mesmo que sejam sofisticados, não são eficazes para descoberta de conhecimentos 'ocultos' em massas de dados como *big data*, por exemplo. Neste contexto, a descoberta de conhecimento em bases de dados ou *Knowledge Discovery in Databases* (KDD) surge como alternativa para auxiliar a descoberta automática de conhecimento por meio do processo completo de conversão de dados brutos em informações úteis (TAN; STEINBACH; KUMAR, 2009, p.4).

O KDD possui como propósito realizar a descoberta de informações relevantes a partir de análise de padrões de grandes conjuntos de dados, de modo a apoiar decisões estratégicas. Para isto, conta com as fases de seleção, pré-processamento, transformação, mineração dos dados e interpretação de resultados. Todas as fases são importantes, no entanto, a etapa de mineração de dados recebe maior destaque na literatura, considerando que passou a ser vista como um diferencial competitivo, auxiliando os tomadores de decisão a realizarem escolhas estratégicas.

No campo de atuação jurídico, o avanço tecnológico proporcionou a tramitação dos processos em meio eletrônico, otimizando as atividades dos profissionais da área. No entanto, devido ao grande volume de processos tramitando nos tribunais brasileiros, torna-se complexo extrair padrões entre as decisões proferidas devido à falta da uniformização processual. Torna-se comum encontrar processos com pedidos e alegações semelhantes, mas com julgamentos divergentes, de acordo com o entendimento do juízo em que está tramitando a ação. Além disso, com a advocacia em massa, torna-se um fator de vantagem competitiva tomar decisões com base em informações fundamentadas.

Neste contexto, este artigo descreve a aplicação de técnicas de mineração de dados sobre uma base jurídica cedida por uma organização atuante no ramo, de modo a identificar se existem padrões, conforme o Estado em que tramita o processo. A base de dados analisada é constituída por processos cíveis no direito do consumidor em contratos de financiamento, contendo diferentes tipos de processos: tarifa, tarifa e dano moral, revisional, indenizatória e outras.

Estrada (2015) destaca que a utilização de algoritmos para prever resultados já é utilizada em várias áreas de impacto social, sendo que as previsões orientadas por dados podem fornecer informações adicionais para apoiar a análise dos advogados.

2 KNOWLEDGE DISCOVERY IN DATABASE (KDD)

O *Knowledge Discovery In Database* consiste no processo de descoberta de padrões pela análise de grandes conjuntos de dados, tendo como principal etapa o processo de mineração, consistindo na execução prática de análise e de algoritmos específicos que, sob limitações de eficiência computacionais aceitáveis, produz uma relação particular de padrões a partir de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Fayyad, Piatetsky-Shapiro e Smyth (1996, p.41) afirmam que o processo de KDD é interativo e iterativo, envolvendo vários passos com muitas decisões tomadas pelo usuário. Os autores consideram o processo de KDD dividido em nove etapas, conforme detalhamento apresentado na sequência:

1. o primeiro passo consiste no conhecimento do domínio da aplicação: inclui o conhecimento relevante e as metas do processo KDD para a aplicação;
2. consiste na criação de um banco de dados alvo: inclui selecionar um conjunto de dados ou dar ênfase para um subconjunto de variáveis ou exemplo de dados nos quais o 'descobrimento' será realizado;
3. consiste na limpeza de dados e pré-processamento: inclui operações básicas como remover ruídos, coleta de informação necessária para modelagem,

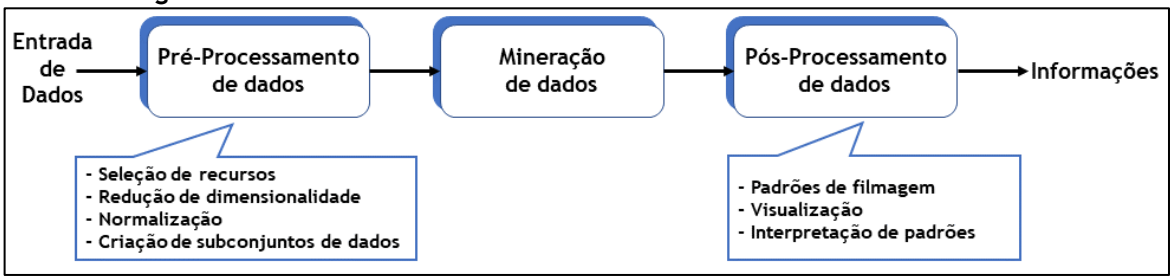
REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

decidir estratégias para manusear (tratar) campos perdidos etc.;

4. consiste na redução de dados e projeção: inclui encontrar formas práticas para se representar dados;
5. consiste na escolha da tarefa de mineração de dados: inclui a decisão do propósito do modelo derivado do algoritmo de mineração de dados (Ex. classificação, regressão, regras de associação e agrupamento);
6. consiste em encontrar o algoritmo de mineração de dados: inclui selecionar métodos para serem usados para procurar por modelos nos dados, como decidir quais modelos e parâmetros podem ser apropriados e determinar um método de mineração de dados particular como modelo global do processo KDD;
7. consiste na interpretação: inclui a interpretação do modelo descoberto e possível retorno a algum passo anterior como também uma possível visualização do modelo extraído, removendo modelos redundantes ou irrelevantes e traduzindo os úteis em termos compreendidos pelos usuários;
8. consiste na utilização do descobrimento obtido: inclui incorporar este conhecimento no desempenho do sistema, tomando ações baseadas no conhecimento, ou simplesmente documentando e reportando para grupos interessados;
9. consiste em agir sobre o conhecimento descoberto: inclui usar o conhecimento diretamente, incorporando-o em outro sistema de novas ações, ou simplesmente documentá-lo e denunciá-lo às partes interessadas.

Calil *et al*. (2008) complementam que as fases do KDD podem ser agrupadas em três grandes grupos: pré-processamento, mineração de dados e pós-processamento (Figura 1).

Figura 1: Processo de Descoberta de Conhecimento em Bancos de Dados.



Fonte: Tan, Steinbach e Kumar – 2009 - p.4.

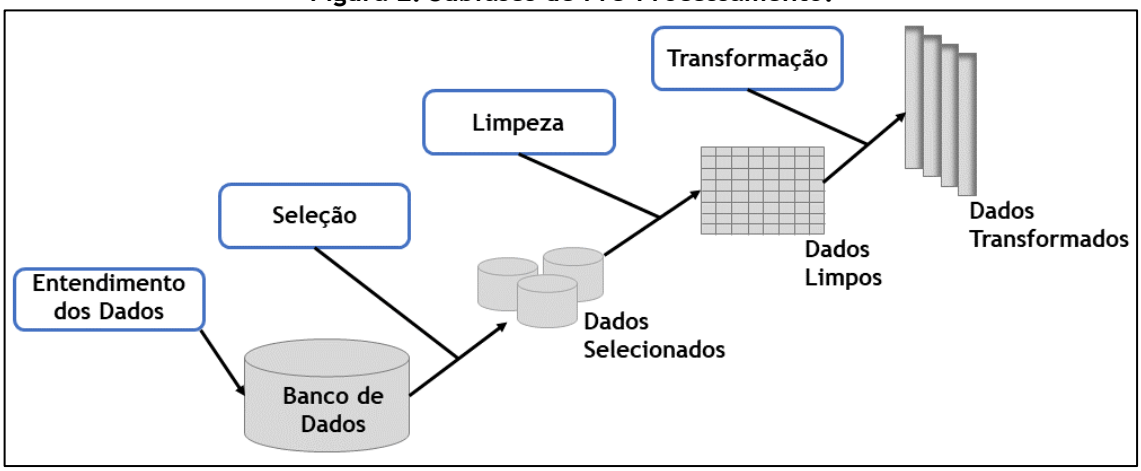
A seguir são abordados estes três grandes grupos.

2.1 Pré-Processamento

De acordo com Neves (2003), a fase de pré-processamento é composta pelas subfases: entendimento, seleção, limpeza e transformação de dados, conforme demonstra a Figura 2.

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

Figura 2: Subfases de Pré-Processamento.



Fonte: Neves - 2003.

Conforme define a autora, o entendimento dos dados consiste em analisar os dados fornecidos pelos especialistas, entendendo do que se tratam as tabelas envolvidas, o significado, relevância, formato, tamanho e tipo de dado dos atributos; identificando os atributos chaves; realizando levantamentos estatísticos e verificando a qualidade dos dados.

A seleção de dados envolve a escolha da (s) tabela (s), atributos e instâncias da (s) mesma (s) em relação aos objetivos do usuário.

A limpeza de dados refere-se a garantia da qualidade dos dados que pode ser obtida por meio de algumas operações, tais como: padronização de dados, tratamento de valores ausentes, eliminação de dados errôneos e de duplicatas.

A transformação de dados corresponde a operações que tornem a apresentação dos dados apropriada a técnica de mineração de dados a ser utilizada. Assim encontram-se descritas operações do tipo normalização de dados, conversões de valores simbólicos para valores numéricos, discretização e composição de atributos.

2.2 Mineração de Dados

A mineração de dados (*Data Mining*, em inglês) é um dos principais passos no processo de KDD, tendo sido utilizada para melhorar sistemas de recuperação de informações. À ela corresponde parte da descoberta de conhecimento em bases de dados (KDD), tendo surgido a partir da necessidade de desenvolver ferramentas mais eficientes e escaláveis que pudessem lidar com diversos tipos de dados. Assim, os trabalhos que culminaram na área de mineração de dados constituíram-se sobre a metodologia e algoritmos que pesquisadores já haviam utilizado anteriormente, como a (1) amostragem, estimativa e teste de hipótese a partir de estatísticas e (2) algoritmos de busca, técnicas de modelagem e teorias de aprendizagem da inteligência artificial, reconhecimento de padrões e aprendizagem de máquina. A proposta, por sua vez, também foi rápida em adotar ideias de outras áreas, incluindo a otimização, computação evolutiva, teoria da informação, processamento de sinais, visualização e recuperação de informações (TAN; STEINBACH; KUMAR, 2009, p.7).

O objetivo da mineração de dados é a extração de conhecimento implícito por meio da descoberta de padrões e regras significativas, a partir de grande quantidade de dados armazenados, de forma automática ou semiautomática, utilizando modelos computacionais construídos para descobrir novos fatos e relacionamentos entre dados, de forma repetida e interativa (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

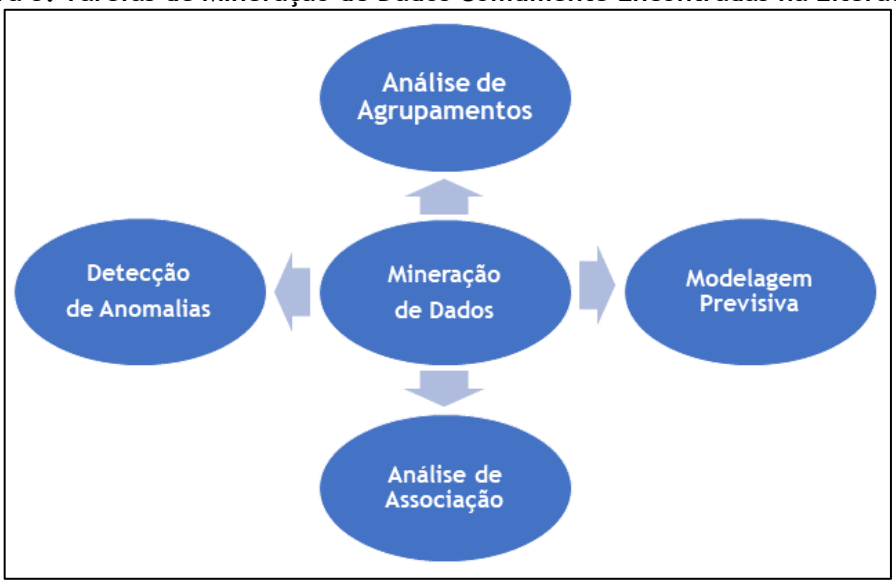
São típicas aplicações da mineração de dados para análise e predição de crédito, detecção de fraudes, predição do

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

mercado financeiro, relacionamento com clientes, predição de falência corporativa, entre muitas outras. Exemplos de segmentos de aplicação incluem setor financeiro; planejamento estratégico empresarial; planejamento do setor portuário; setores de energia (petróleo, gás, energia elétrica, biocombustíveis etc.); educação; logística; planejamento das cadeias de produção, distribuição e suprimentos; meio ambiente; e Internet (portais, redes sociais, comércio eletrônico etc.) (CASTRO; FERRARI, 2016, p.17).

A Figura 3 ilustra quatro das tarefas centrais da mineração de dados de acordo com a abordagem de Tan, Steinbach e Kumar (2009, p.9).

Figura 3: Tarefas de Mineração de Dados Comumente Encontradas na Literatura.



Fonte: Adaptado de Tan, Steinbach e Kumar - 2009 - p.9.

A modelagem de previsão se refere à atividade de construir um modelo para a variável alvo (também conhecida por meta ou objetivo) como uma função das atividades explicativas. Há dois tipos de tarefas de modelagem de previsão: classificação, a qual é usada para variáveis-alvo discretas, e regressão, que é usada para variáveis-alvo contínuas. (TAN; STEINBACH; KUMAR, 2009, p.9).

A análise de associação é usada para descobrir padrões que descrevam características altamente associadas dentro dos dados. Os padrões descobertos são normalmente representados na forma de regras de implicação ou subconjuntos de características. Devido ao tamanho exponencial de seu espaço de busca, o objetivo da análise de associação é extrair os padrões mais interessantes de forma mais eficiente. (TAN; STEINBACH; KUMAR, 2009, p.11).

A análise de grupo procura encontrar grupos de observações intimamente relacionadas de modo que observações que pertençam ao mesmo grupo sejam mais semelhantes entre si do que com as que pertençam a outros grupos. (TAN; STEINBACH; KUMAR, 2009, p.12)

Por fim, a detecção de anomalias é a tarefa de identificar observações cujas características sejam significativamente diferentes do resto dos dados. Tais observações são conhecidas como anomalias ou fatores estranhos. O objetivo de um algoritmo de detecção de anomalias é descobrir as anomalias verdadeiras e evitar rotular erroneamente objetos normais como anômalos. Em outras palavras, um bom detector de anomalias deve ter uma alta taxa de detecção e uma baixa taxa de alarme falso. (TAN; STEINBACH; KUMAR, 2009, p.13).

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

2.3 Pós-Processamento

De acordo com Milani e Carvalho (2013), o pós-processamento tem como principal objetivo apoiar na verificação de até que ponto estes padrões contribuem na solução do problema inicialmente identificado. As autoras destacam que para operacionalizar o pós-processamento existem várias estratégias propostas, entre elas eliminar a redundância, generalizar, identificar no conjunto aqueles com maior potencial de serem interessantes etc.

Tan, Steinbach e Kumar (2009, p.5) destacam que o pós-processamento deve assegurar que somente resultados válidos e úteis sejam incorporados ao sistema de apoio a decisão. Os autores exemplificam a visualização, a qual permite que os analistas explorem os dados e os resultados da mineração dos mesmos a partir de uma diversidade de pontos de vista. Também de acordo com os autores, as medições estatísticas ou métodos de teste de hipóteses podem ser igualmente aplicados durante o pós-processamento para eliminar os resultados não legítimos da mineração de dados.

3 METODOLOGIA

A pesquisa apresenta finalidade aplicada e abordagem quantitativa, pois objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. Quanto aos objetivos, é caracterizada como descritiva e quanto aos procedimentos como um estudo de caso, considerando que descreve uma determinada situação do contexto em que está sendo desenvolvida determinada investigação.

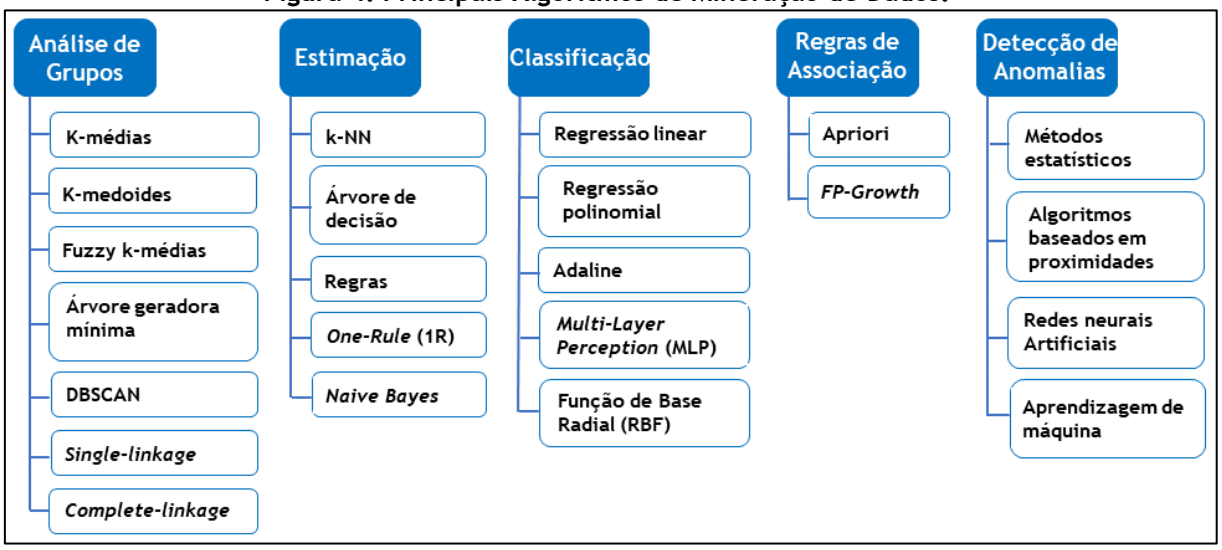
A pesquisa é realizada em uma organização privada de Curitiba atuante no segmento jurídico. A base de dados analisada contém aproximadamente mil processos cíveis de direito do consumidor.

Para realizar a mineração de dados na base de dados é de suma importância identificar o tipo de atributo, ou seja, nominal, ordinal, intervalar ou proporcional, sendo estas características decisivas para a escolha do método de mineração de dados

Na etapa de mineração são escolhidos os algoritmos que melhor atendam aos requisitos da base de dados. A escolha é realizada com base nos tipos de atributos previamente descritos no pré-processamento, também levando em consideração a complexidade para análise e interpretação dos resultados. Os métodos e algoritmos definidos na revisão da literatura foram sintetizados na Figura 4, de modo a facilitar a visualização e escolha do algoritmo que melhor extraia conhecimento da base de dados jurídica.

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

Figura 4: Principais Algoritmos de Mineração de Dados.



Fonte: Elaborado pelas autoras - 2017.

Esta etapa também envolve a escolha do software de mineração de dados a ser utilizado. Como critérios, primeiramente foram eliminados os softwares pagos. Por fim, optou-se pela utilização do *Weka*, considerando os seguintes critérios: já ter sido utilizado em disciplina acadêmica, ser livremente distribuído e apresentar compatibilidade com diversos sistemas operacionais.

A validação da proposta corresponde a etapa de pós-processamento, na qual os resultados são analisados para verificar se houve efetivamente alguma descoberta de conhecimento. Para isto, os resultados foram apresentados a um profissional da área jurídica (entrevista) para verificar se o conhecimento extraído contribuiu de alguma forma para a tomada de decisão ou se foram obtidos apenas resultados que já eram evidentes para a área. Além de avaliar as eventuais sugestões de melhorias.

4 RESULTADOS

A base de dados analisada contém 1.169 processos ativos em diversas fases processuais. Para análise, primeiramente definiu-se como atributo meta a coluna 'Motivo Arquivamento', a qual possui as categorias: condenação; acordo, improcedência; extinção sem mérito; e outros.

Para análise foram filtrados todos os processos da base de dados arquivados - aqueles que já foram julgados e encaminhados ao arquivo - entre o período de abril de 2014 e janeiro de 2016, totalizando 701 processos. Primeiramente foram retiradas as colunas da base de dados que não exerceriam influência sobre o atributo meta analisado, restando 16 colunas. Os atributos que continham letra e número foram divididos em colunas distintas e os atributos numéricos foram discretizados em intervalos.

A partir da análise da base de dados foi possível identificar que a distribuição dos processos estava desproporcional, optando-se pela distribuição dos processos por região: Sul, Sudeste, Centro-oeste, Norte e Nordeste.

Para proceder a análise da base é necessário escolher os métodos que serão utilizados para dar suporte a análise dos dados. Primeiramente optou-se pelas heurísticas de classificação, pois são as mais conhecidas e utilizadas e consistem em associar objetos a um conjunto pré-definido de classes de acordo com as suas características. Na tarefa de classificação foram utilizadas as heurísticas de regras e árvores, pois apresentam maior facilidade para compreensão dos resultados. Na tarefa de associação foi utilizado apenas o algoritmo Apriori, considerando ser o

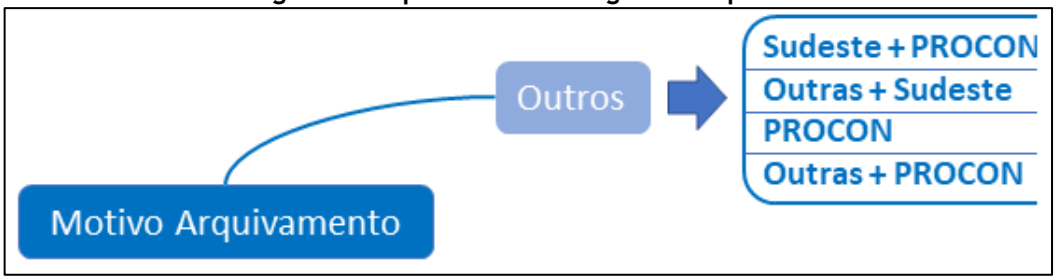
método mais conhecido para mineração de regras de associação.

4.1 Apriori

No primeiro experimento foram considerados os parâmetros *default*, com exceção do número de linhas que foi alterado para apresentar o máximo de resultados possíveis. O algoritmo apresentou 2.047 regras, gerando dificuldade para efetuar a análise. Desta forma, foi realizado um segundo experimento retirando alguns atributos que não seriam interessantes para a análise, mantendo apenas: acao; regiao; OJ; motivo_arquivamento. Como resultado, foram apresentadas 14 regras.

Como o foco da pesquisa consiste somente em verificar se há relações entre o Estado e o motivo do arquivamento, as relações que não atendiam este requisito foram descartadas, restando apenas seis regras. Com base nestas regras é possível analisar que ações classificadas como 'outras' tramitando na região sudeste e pelo PROCON apresentam motivo de arquivamento 'Outros' (Figura 5).

Figura 5: Mapa Conceitual Algoritmo Apriori.



Fonte: Elaborado pelas autoras - 2017.

4.2 Part

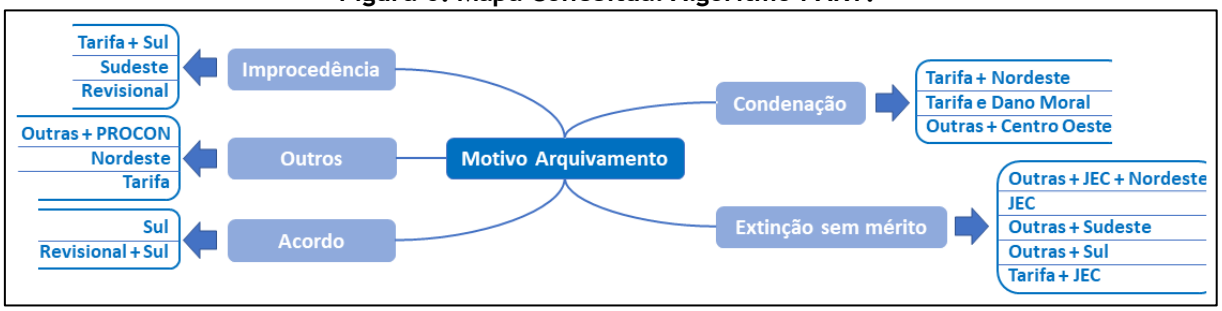
Na heurística de regras foi executado o algoritmo PART, o qual é definido pelo software *Weka* como: 'classe para gerar uma lista de decisão' PART. Constrói uma árvore de decisão C4.5 parcial em cada iteração e faz a 'melhor' folha em uma regra'.

Ao executar o algoritmo PART com os parâmetros *default* foram geradas 57 regras em 0,2 segundos. Mantendo apenas os quatro atributos principais (acao; regiao; OJ; motivo_arquivamento) foram geradas 17 regras em 0,08 segundos.

Para facilitar a visualização das informações os resultados foram sintetizados em um mapa conceitual disposto na Figura 6. Nele é possível identificar a tendência de realização de acordos em ações revisionais que tramitam na região Sul. Também é possível analisar que são comuns condenações em ações de cobrança de tarifa e dano moral, ações de tarifa tramitando na região Nordeste e ações outras tramitando na região Centro-oeste. Desta forma, estes seriam os casos mais críticos que deveriam ser analisados pelo escritório de advocacia para conseguir reduzir a quantidade de condenações. Os casos de 'Extinção Sem Mérito' apresentaram maior variação, sendo que geralmente correspondem a reclamações de tarifa tramitando no JEC, ações outras tramitando no JEC da região Nordeste e ações outras tramitando na região Sudeste e Sul. Foram identificados padrões de improcedência na região Sudeste, em ações de tarifa tramitando na região Sul e em ações revisionais. Por fim, as ações arquivadas por motivo Outros correspondem as ações 'outras' tramitando no PROCON, na região Nordeste e ações discutindo tarifas.

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

Figura 6: Mapa Conceitual Algoritmo PART.



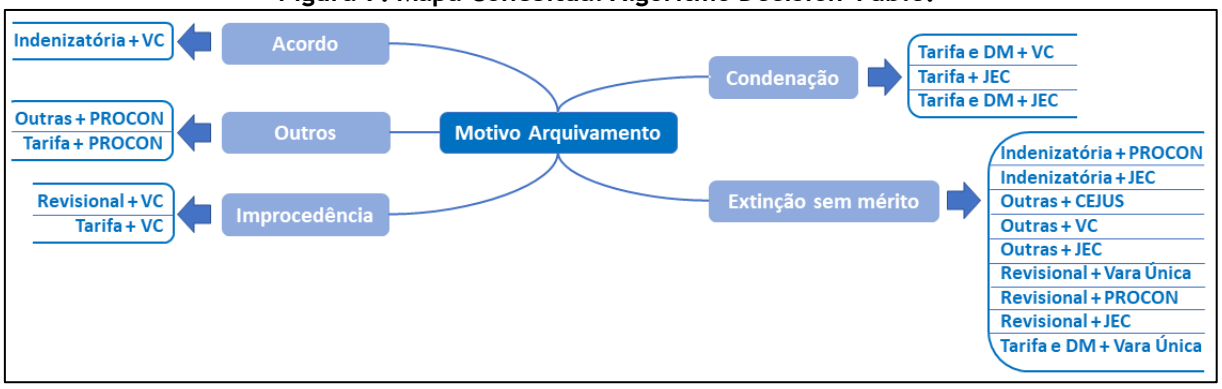
Fonte: Elaborado pelas autoras - 2017.

Na heurística de regras também foi utilizado o algoritmo *Decision Table*. O mesmo é definido pelo software *Weka* como: 'classe para a construção e utilização de uma simples tabela de decisão pela classificação da maioria'.

O primeiro experimento foi realizado alterando apenas o parâmetro 'displayRules'para *True* para que nos resultados fosse apresentada a tabela de regras. O algoritmo demorou 0,17 segundos para ser executado e apresentou que os atributos ação, OJ, risco_atual são influenciadores para determinar o motivo_arquivamento. No total foram geradas 41 regras. Contudo, como o atributo risco_atual não era interessante para a análise, foi realizado um novo experimento removendo todas as colunas que não exerceriam influência sobre o resultado almejado.

O segundo experimento foi realizado considerando apenas os atributos principais. Com isto, o tempo de execução foi reduzido para 0,01 segundos e foram encontradas 17 regras. Com base nos resultados apresentados foi criado um mapa conceitual (Figura 7), visando otimizar a visualização dos resultados.

Figura 7: Mapa Conceitual Algoritmo *Decision Table*.



Fonte: Elaborado pelas autoras - 2017.

Conforme demonstram os resultados, a classificação dos processos com 'Extinção Sem Mérito' ainda são os que apresentam maior dificuldade no reconhecimento de padrões, encontrando quase todos os tipos de ações e órgãos julgadores. Contudo, ainda assim os resultados mostraram-se satisfatórios, pois permitiram a identificação de padrões. No entanto, seria mais interessante se fosse apresentada na tabela a relação entre a região, o que não foi realizado automaticamente pela determinação da importância dos atributos pelo algoritmo.

4.3 J48

Na heurística de árvores foi executado o algoritmo J48, o qual é definido pelo software *Weka* como: 'classe para gerar uma árvore de decisão C4.5 podada ou não podada'. Ao executar executá-lo com os parâmetros *default* foi

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

gerada uma árvore com 69 folhas e tamanho 94 em 0 segundos. Porém, devido ao seu tamanho, as folhas ficaram sobrepostas, impossibilitando a análise.

Para facilitar a análise e compreensão dos dados, então, foi realizado um segundo experimento considerando apenas os quatro atributos principais. A nova execução o algoritmo também demorou 0 segundos, porém reduziu o tamanho da árvore para 31 com 25 folhas.

Na leitura da árvore de decisão foi identificado que a raiz da árvore corresponde ao Órgão Julgador, sendo este, portanto, o atributo com maior influência. O segundo atributo com maior influência corresponde ao tipo de ação, seguido de região. Para PROCON, Vara Única e Cejusc o resultado foi simplificado, mostrando diretamente o motivo do arquivamento. Já pra VC e JEC existem outros atributos que exercem influência sobre o motivo do arquivamento.

No PROCON foi identificado o motivo de arquivamento 'Outros', enquanto na Vara Única Improcedência e no Cejusc 'Extinção Sem Mérito'. No JEC o tipo de ação exerce influência sobre o resultado, sendo para ações revisionais, indenizatórias e outras obtido 'Extinção Sem Mérito' e para tarifas e tarifa e dano moral 'Condenação'. Por fim, na Vara Cível (VC), além do tipo de ação, a região também exerce influência sobre o resultado das ações revisionais, tarifas e outras. Em ações indenizatórias foi obtido como resultado 'Acordo' e em ações de tarifa e dano moral 'Condenação'. Em ações revisionais que tramitam no Sul foi obtido 'Acordo'; no Nordeste, 'Extinção Sem Mérito'; no Sudeste, Centro-oeste e Norte, 'Improcedência'. Em ações de tarifa que tramitam no Sul foi obtido 'Improcedência'; no Nordeste, 'Condenação'; no Sudeste, 'Extinção Sem Mérito'; no Centro-oeste, 'Outros'; e, no Norte, 'Condenação'.

Também na heurística de árvores foi executado o algoritmo REPTree, o qual é definido pelo software *Weka* como: 'classe para a construção de uma árvore que considera atributos K escolhidos aleatoriamente em cada nó. Não executa nenhuma poda. Também tem uma opção para permitir a estimativa de probabilidades de classe'.

Ao executar o algoritmo REPTree com os parâmetros *default* foi gerada uma árvore com tamanho 43 em 0,4 segundos. Para facilitar a análise e compreensão dos dados foi realizado um novo experimento considerando apenas os quatro principais atributos. Na nova execução, o algoritmo demorou 0,01 segundos e reduziu o tamanho da árvore para 31.

Pela leitura da árvore de decisão foi possível identificar que a raiz da árvore corresponde ao Órgão Julgador, sendo este, portanto, o atributo com maior influência. O segundo atributo com maior influência corresponde ao tipo de ação, seguido de região. Para PROCON, Vara Única e Cejusc o resultado foi simplificado, mostrando diretamente o motivo do arquivamento. Já pra VC e JEC existem outros atributos que exercem influência sobre o motivo do arquivamento.

No PROCON foi identificado o motivo de arquivamento 'Outros', enquanto na Vara Única 'Improcedência' e no Cejusc 'Extinção Sem Mérito'. No JEC o tipo de ação e a região exercem influência sobre o resultado, sendo para ações revisionais e outras obtido 'Extinção Sem Mérito'; para tarifas e tarifa e dano moral 'Condenação'; e para indenizatórias que tramitam no Sul 'Condenação'; no Nordeste, Centro-oeste e Norte, 'Extinção Sem Mérito'; e no Sudeste, 'Condenação'. Por fim, na Vara Cível (VC) também recebeu influência do tipo de ação e da região nas ações revisionais e de tarifas. Em ações indenizatórias foi obtido como resultado 'Acordo'; em ações de tarifa e dano moral, 'Condenação'; e, em Outras, 'Extinção Sem Mérito'. Em ações revisionais que tramitam no Sul foi obtido 'Acordo'; no Nordeste, 'Extinção Sem Mérito'; no Sudeste, Centro-oeste e Norte, 'Improcedência'. Em ações de tarifa que tramitam no Sul foi obtido 'Improcedência'; no Nordeste, 'Condenação'; no Sudeste, 'Extinção Sem Mérito'; no Centro-oeste, Outros; e, no Norte, 'Condenação'.

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

4.4 Análise dos Resultados

A partir dos testes realizados com os algoritmos Apriori, PART, *Decision Table*, J48 e REPTree foi possível identificar que, de maneira geral, o experimento 1 apresentou melhores resultados quanto a acurácia da base de dados. No entanto, quanto a análise dos resultados, o experimento 2 mostrou-se mais eficiente, considerando a apresentação mais simplificada e compreensível os resultados.

A Tabela 1 apresenta uma comparação entre o desempenho de classificação dos algoritmos no primeiro experimento. Nele é possível identificar o tempo de execução e os parâmetros de validação cruzada estratificada apresentados pelo *Weka*.

Tabela 1: Desempenho de Classificação - Experimento 1.

Parâmetros	*Decision Table*	PART	J48	REPTree	Média	Desvio Padrão
Correto	439,0000	444,0000	465,0000	461,0000	452,2500	12,6853
Incorreto	262,0000	257,0000	236,0000	240,0000	248,7500	12,6853
Tempo de Execução	0,5300	0,2000	0,0000	0,0800	0,20025	0,2333
Kappa Statistic	0,5173	0,5285	0,5678	0,5587	0,5431	0,0240
Mean absolute error	0,1989	0,1704	0,1714	0,1776	0,1796	0,0133
Root mean squared error	0,3101	0,3278	0,3118	0,3102	0,3150	0,0086
Relative absolute error (%)	63,6192	54,4967	54,8111	56,7864	57,4284	4,2498
Root relative squared error (%)	78,4263	82,9222	78,8722	78,4613	79,6705	2,1772

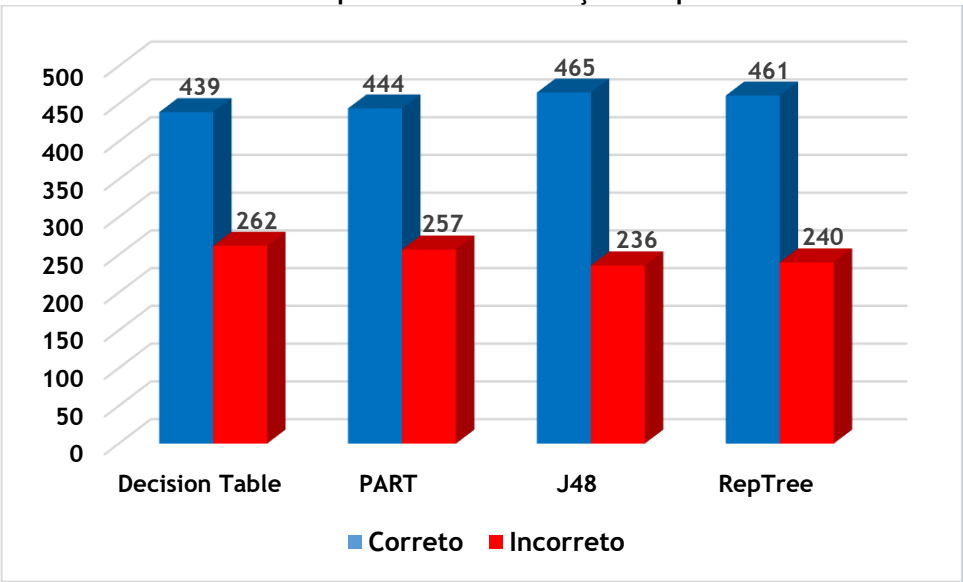
Fonte: Elaborado pelas autoras – 2017.

O tempo de execução foi bom para todos os experimentos, tendo em vista que a base de dados contém apenas 701 instâncias. Contudo, é possível identificar que o *Decision Table* foi o que demorou mais tempo para ser executado. Quanto ao *Kappa Statistic* (Estatística Kappa), o resultado foi inconclusivo, considerando que os valores ficaram na metade do intervalo entre 0 e 1, não permitindo afirmar que existe correlação. Quanto ao *Mean absolute error* (Erro Absoluto Médio), o algoritmo PART foi o que apresentou melhor resultados, gerando menor erro na classificação dos atributos. Quanto ao Root *Mean Squared Error* (Erro Quadrado Médio), o algoritmo *Decision Table* apresentou menor erro entre os valores atuais e os valores preditos, porém ambos os experimentos tiveram valores aproximados. Quanto ao *Relative Absolute Error* (Erro Absoluto Relativo), o algoritmo PART obteve mais precisão para previsão numérica do que os demais experimentos. Por fim, quanto ao *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo), o algoritmo *Decision Table* obteve melhores resultados, pois apresentou menor erro que os demais experimentos.

O Gráfico 1 apresenta a comparação entre os resultados dos algoritmos com base nas instâncias classificadas correta e incorretamente. Analisando os dados é possível identificar que o algoritmo J48 realizou uma classificação mais eficiente dos atributos, seguido de REPTree, PART e *Decision Table* .

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

Gráfico 1: Desempenho de Classificação - Experimento 1.



Fonte: Elaborado pelas autoras - 2017.

A Tabela 2 apresenta uma comparação entre o desempenho de classificação dos algoritmos no segundo experimento. Nele é possível identificar o tempo de execução e os parâmetros de validação cruzada estratificada apresentados pelo *Weka*.

Tabela 2: Desempenho de Classificação - Experimento 2.

Parâmetros	*Decision Table*	PART	J48	REPTree	Média	Desvio Padrão
Correto	361,0000	349,0000	354,0000	350,0000	353,0000	5,4467
Incorreto	340,0000	352,0000	347,0000	351,0000	347,5000	5,4467
Tempo de Execução	0,0100	0,0800	0,0000	0,01	0,0300	0,0436
Kappa Statistic	0,3744	0,3539	0,3649	0,3564	0,3624	0,0093
Mean absolute error	0,2427	0,2338	0,2330	0,2349	0,2361	0,0045
Root mean squared error	0,3472	0,3485	0,3511	0,3509	0,3494	0,0019
Relative absolute error (%)	77,6157	74,7582	88,8055	88,7495	88,3772	1,4293
Root relative squared error (%)	87,8149	88,1387	88,8055	88,7495	88,3772	04,814

Fonte: Elaborado pelas autoras - 2017.

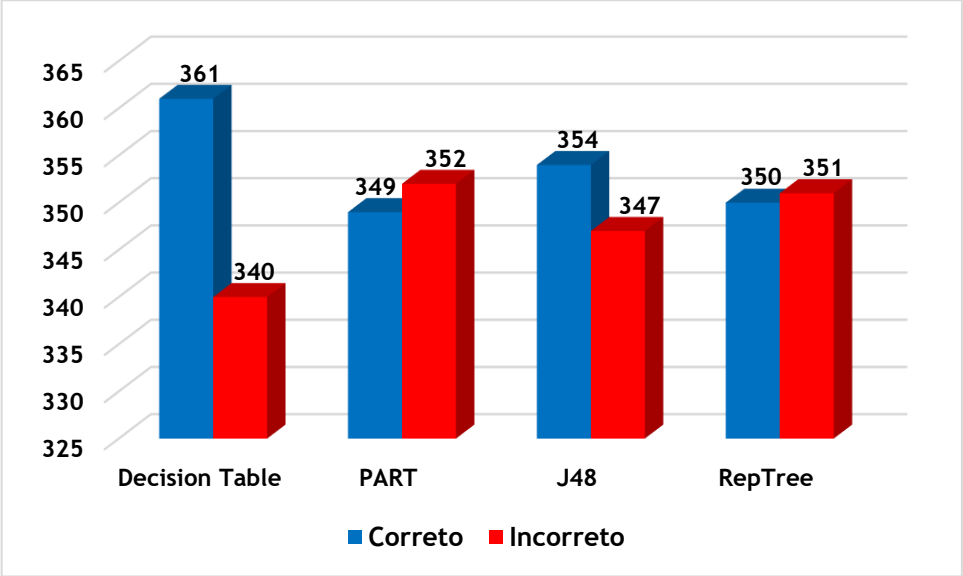
O tempo de execução foi bom para todos os experimentos, tendo o mais lento demorado 0,08 (PART) segundos e o mais rápido 0. (J48). Houve também uma redução no tempo de execução quando comparado ao primeiro experimento, considerando a redução do número de atributos de 10 para 4. Quanto ao *Kappa Statistic* (Estatística Kappa), o resultado foi inconclusivo, considerando que os valores ficaram na metade do intervalo entre 0 e 1, não permitindo afirmar que existe correlação. Quanto ao *Mean absolute error* (Erro Absoluto Médio), o algoritmo J48 foi o que apresentou melhor resultados, gerando menor erro na classificação dos atributos. Quanto ao *Root Mean Squared Error* (Erro Quadrado Médio), o algoritmo *Decision Table* apresentou menor erro entre os valores atuais e os valores preditos. Quanto ao *Relative Absolute Error* (Erro Absoluto Relativo), o algoritmo J48 obteve mais precisão para previsão numérica do que os demais experimentos. Por fim, quanto ao *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo), o

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

algoritmo *Decision Table* obteve melhores resultados, pois apresentou menor erro que os demais experimentos.

A Gráfico 2 apresenta a comparação entre os resultados dos algoritmos com base nas instâncias classificadas correta e incorretamente. Analisando os dados é possível identificar que o algoritmo *Decision Table* realizou uma classificação mais eficiente dos atributos, seguido de J48, REPTree e Part.

Gráfico 2: Desempenho de Classificação - Experimento 2.



Fonte: Elaborado pelas autoras - 2017.

Analisando os resultados e desempenhos obtidos pelos experimentos, os algoritmos J48 e *Decision Table* atenderam melhor as características da base de dados. Eles conseguiram apresentar resultados satisfatórios para análise, mantendo a acurácia da base de dados e tornando os resultados relevantes para a análise. O J48 apresenta a vantagem de gerar a árvore de decisão que facilita a análise para a tomada de decisão. O *Decision Table*, por sua vez, gera uma tabela de decisão que também permite analisar condições, contudo, torna a análise mais demorada por não gerar uma representação gráfica. Além disso, o algoritmo gera resultados mais simplificados, tendo em vista que não considera todas as hipóteses possíveis, enquanto na árvore de decisão podem ser analisados todos os caminhos possíveis.

Para validação dos resultados foi realizada uma entrevista com duração de aproximadamente uma hora com dois profissionais da área jurídica. Os resultados alcançados foram submetidos a análise, a fim de verificar se foram satisfatórios para identificação de padrões de decisões judiciais, bem como questionar melhorias.

Para facilitar a análise foram apresentadas as árvores de decisão geradas pelos algoritmos J48 e REPTree, pois tornam mais fácil a compreensão dos resultados através da representação gráfica. Após a apresentação e discussão dos resultados foi questionada qual árvore de decisão atende melhor a realidade da área jurídica, com base no conhecimento e experiência que eles possuem sobre o assunto. Ambos concluíram que a árvore gerada pelo REPTree foi mais satisfatória, pois além de estar mais simplificada, apresentou resultados mais próximos da realidade. Outra observação realizada pelos entrevistados ocorreu em relação a Vara Única, que apresentou como resultado Improcedência nas duas árvores de decisões. Segundo os entrevistados, estas ações correspondem à Vara Cível, porém

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

foram classificadas incorretamente na base de dados e poderiam ser desconsideradas.

Os entrevistados mostraram-se satisfeitos com os resultados apresentados e destacaram a relevância da pesquisa para a área, pois declararam não ter conhecimento de uma proposta no mercado que realize esse tipo de análise. Além disso, os entrevistados explicaram que tomavam as decisões com base no conhecimento tácito e experiências inerentes a cada um, sem realizar um efetivo estudo. Porém, com a falta de uniformização das decisões judiciais, a identificação de padrões é de suma importância para estabelecimento das estratégias a serem adotadas.

Os entrevistados também comentaram sobre a importância da análise das decisões ao longo do tempo, sendo relevante realizar um planejamento anual para acompanhamento dos novos resultados a partir da mudança nas estratégias. Pela mineração de dados é possível realizar esse tipo de análise, pois ela demonstra o histórico dos processos da organização ao longo do tempo, podendo aumentar ou reduzir a quantidade de condenações com base no posicionamento adotado pelo escritório.

5 CONSIDERAÇÕES FINAIS

Os avanços tecnológicos proporcionaram um crescimento exponencial no volume de dados devido ao aumento de usuários na Internet e de conteúdos publicadas diariamente. Neste contexto, tornou-se um desafio gerenciar as informações de forma a extrair conhecimento para a tomada de decisão. Este crescimento afetou também o ambiente organizacional, no qual as bases de dados cresceram tanto que dificultaram a análise manual, sendo necessárias novas técnicas e ferramentas capazes de analisar grandes volumes de dados de forma inteligente, visando gerar vantagem competitiva.

Nesse cenário, a mineração de dados tem sido uma ferramenta de apoio com papel fundamental na gestão da informação dentro das organizações. No entanto, a escolha do método de mineração de dados não é uma tarefa fácil, pois não existe um padrão para a escolha, variando de acordo com os tipos de atributos da base de dados. Assim, é destacada a importância do pré-processamento, considerando que as atividades de limpeza, integração, redução, transformação e discretização são essenciais para realizar uma efetiva mineração de dados. A escolha do software de mineração de dados é outra atividade importante, considerando que o mesmo deve atender ao método de mineração anteriormente definido.

Neste contexto e para atingir o objetivo da pesquisa, foi realizado um levantamento bibliográfico, a fim de nortear a escolha das tarefas e heurísticas a serem utilizadas nesta pesquisa. Destaca-se, nesta etapa, a dificuldade no levantamento de material bibliográfico, pois no Brasil ainda não existem muitas pesquisas desenvolvidas na área e os principais estudos são na área de Ciência da Computação, apresentando, portanto, uma linguagem mais técnica e de difícil compreensão.

Na tarefa de associação, os resultados não foram muito satisfatórios, pois o algoritmo não permite escolher o atributo meta, gerando poucas regras com importância para o enfoque da pesquisa.

Na tarefa de classificação, optou-se pela utilização das heurísticas árvores de decisão e regras. Para cada algoritmo foi realizado dois experimentos, pois após a primeira execução verificou-se a necessidade de simplificar os resultados para facilitar a análise. Além disso, houve apenas o retrabalho de gerar as árvores de decisão manualmente por deficiência do software que não permite o redimensionamento das folhas.

Devido à falta de conhecimento sobre a área jurídica, foi elaborado um glossário de termos para facilitar a compreensão de alguns conceitos recorrentes. Além disso, os resultados foram submetidos a um profissional da área para validação da proposta. Em uma entrevista com duração de aproximadamente uma hora, dois advogados do escritório que cedeu a base de dados para o estudo analisaram a

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193

veracidade dos resultados apresentados. Na entrevista foi possível identificar a importância da criação manual das árvores de decisão, pois a apresentação gráfica permitiu aos advogados compreenderem cada situação, percorrendo os caminhos possíveis da árvore para identificar possíveis inconsistências. O *feedback* foi muito relevante e contribuiu para o estudo, pois demonstrou a importância crescente da aplicação de tecnologias da informação na área jurídica, sendo um ramo interessante para pesquisas aprofundadas. Além disso, conforme a pesquisa já apontava, não existem muitos estudos desenvolvidos que auxiliem esses profissionais a realizarem a tomada de decisão com base em informações fundamentadas. Conforme apontado pelos entrevistados e afirmado por Surden (2014), estes profissionais utilizam uma mistura de experiências e habilidades para fazerem as avaliações e conseguirem possíveis resultados. Neste contexto, a mineração de dados permite apresentar os resultados com base no histórico identificado nos processos, apoiando a análise dos advogados. A partir disso, os mesmos podem estudar a estratégia de atuação, propondo mudanças para os casos em que são identificadas sentenças desfavoráveis, a fim de melhorar o resultado a seu favor.

Para trabalhos futuros sugere-se a aplicação do estudo em outras bases de dados jurídicas, de forma a validar a proposta e comparar as mudanças nos resultados obtidos. Além disso, é recomendada a aplicação das técnicas em outras áreas do Direito, a fim de verificar se também ocorre a falta de uniformização das decisões jurídicas. Por fim, é sugerido testar a aplicação de outras técnicas de mineração de dados para verificar a descoberta de novos conhecimentos.

REFERÊNCIAS

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação da informação:** conceitos e tecnologia das máquinas de busca. 2.ed. Porto Alegre: Bookman, 2013. 590p.

CALIL, L. A. A. *et al*. **Mineração de dados e pós-processamento em padrões descobertos.** PUBLICATIO UEPG Ciências Exatas e da Terra, Ciências Agrárias e Engenharias, Ponta Grossa (PR), v.14, n.3, p.207-215, dez. 2008. Disponível em: <http://www.revistas2.uepg.br/index.php/exatas/article/view/946>. Acesso em: 27 maio 2016.

CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

ESTRADA, M. M. P. **A criação do direito pela inteligência artificial**. Disponível em: <http://direitoeti.com.br/artigos/a-criacao-do-direito-pela-inteligencia-artificial/>. Acesso em: 25 fev. 2016.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From *data mining* to knowledge discovery in databases. **AI magazine**, v.17, n.3, p.37, 1996. Disponível em: <http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>. Acesso em: 3 mar. 2016.

NEVES, R. C. D. **Pré-processamento no processo de descoberta de conhecimento em banco de dados.** 2003. 137f. Dissertação (Mestrado) - Programa de Pós-graduação em Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, 2003.

SIDNEY, Christiane Faleiro. **Aplicação de mineração de dados no banco de dados do zoneamento ecológico econômico de minas gerais**. 2010. 60f. TCC (Graduação) – Sistemas de Informação, Departamento de Ciência da Computação, Universidade Federal de Lavras, Lavras, 2010. Disponível em: <http://goo.gl/zZk0ds>. Acesso em: 8 mar. 2016.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao *data mining***: mineração de dados. Rio de Janeiro: Ciência Moderna, 2009. 900p.

Talita de Souza Rampão
Universidade Federal do Paraná (UFPR)
Bacharel em Gestão da Informação
E-Mail:
Brasil

Denise Fukumi Tsunoda
Universidade Federal do Paraná (UFPR)
Docente do Programa de Pós-Graduação
em Gestão da Informação
E-Mail: dtsunoda@ufpr.br
Brasil

REBECIN, v.6, n. esp., p.61-76, 1. sem. 2019. Disponível em: <http://abecin.org.br/portalderevistas/index.php/rebecin>.
ISSN: 2358-3193