Um modelo de inferência para a classificação de resultados processuais a partir de causas jurídicas oriundas da Justiça Estadual

An inference model for the classification of procedural results from legal causes originating from State Courts

Manoel Alves de Almeida Neto[1] orcid.org/0000-0003-4941-6376
Vinícius Malloni Moura[2] orcid.org/0000-0002-6547-4448
Jonathan da Silva Bandeira[1] orcid.org/0000-0003-1693-2091
Pedro Rudá Cavalcanti Gomes de Freitas[1] orcid.org/0000-0002-8151-1114
Roberta Andrade de Araújo Fagundes[1] orcid.org/0000-0002-7172-4183

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Kurier Tecnologia em Informação, Recife, Pernambuco, Brasil.

E-mail do autor principal: Manoel Alves de A. Neto maan@poli.br

Resumo

Este artigo tem como objetivo apresentar um modelo de inferência para classificar processos jurídicos relacionados à Justiça Estadual utilizando dados documentados de jurisprudência, tais como: comentários dos juízes realizados durante os veredictos, classes jurídicas do processo e UF. Os dados foram extraídos de websites de cortes judiciais, como o Portal do Tribunal de Justiça do Estado de Minas Gerais e o Poder Judiciário do Estado de Alagoas. Em toda a base de dados, foi realizada uma seleção no campo textual da descrição da sentença para extrair as leis que foram consideradas nos veredictos. Para tal seleção, o atributo de publicação e a quantidade de ocorrências da lei na base de dados foram considerados. As técnicas utilizadas para realizar a mineração de dados e classificar os processos como procedentes ou improcedentes foram a árvore de decisão e as redes neurais artificiais. Os testes realizados mostraram resultados satisfatórios e superiores ao valor comum para classificação de dados de jurisprudência, de normalmente 60%.

Palavras-Chave: Mineração de dados; Redes neurais; C4.5; PART; Jurisprudência;

Abstract

This paper aims to develop an inference model for classify the judiciary processes related to State Courts using documental data, such as comments made of the judges in the verdict moment, states and processual juridical class. The data were extracted from Justice courts websites, such as Poder Judiciário do Estados de Alagoas and Portal do Tribunal de Justiça do Estado de Minas Gerais. To extract the laws used in veredicts from database, its amount of ocurrence and the Publication attribute were considered. The techniques used were decision tree and artificial neural network. The tests showed a good shape of the laws and articles to classify appropriate and unfounded cases, which help the decision maker to take a decision. The results obtained were higher than the common rate of 60%.

Key-words: *Data mining; Neural networks; C4.5; PART; Law.*

1. Introdução

O acesso a informações sobre dados Jurídicos é imprescindível para o operador do direito no exercício de suas funções. Contudo, grande parte dessas informações tem fontes distintas e numerosas, como os processos disponibilizados através do sistema Processo Judicial Eletrônico – Pje **[1]**, tornando difícil a correlação destes dados para um processamento estatístico.

Em uma pesquisa de mercado realizada pelo CONJUR **[2]**, é possível chegar à conclusão de que não há uma solução amplamente utilizada no mercado que possua um banco de dados com informações jurídicas relacionadas entre si, de modo a permitir uma abordagem estatística que possibilite entender a relação entre os processos e as variáveis que os levam a classificação de suas sentenças.

Devido à dificuldade de acesso a estas informações, torna-se evidente a importância de um estudo aplicado à área jurídica. Este trabalho tem como objetivo desenvolver um modelo para classificação de resultados processuais a partir de informações extraídas de causas jurídicas, como leis, artigos, fórum e classe processual pertencente ao poder judiciário da Justiça Estadual.

2 Fundamentação Teórica

Para amparar teoricamente a aplicação de conhecimento deste trabalho, se fez necessária a observação e análise de alguns dos fundamentos relacionados à área das ciências jurídicas e da mineração de dados.

2.1 Área de Conhecimento Jurídico

A ciência jurídica ou ciência do direito estuda o fenômeno jurídico em todas as suas manifestações e momentos, tendo como objeto de estudo o conhecimento do direito **[3]** e como meio de expressão a chamada linguagem forense **[4]**.

No Brasil, o direito é dividido em duas principais ramificações: o direito público, que rege os interesses públicos e as relações do estado e o direito privado, que rege os interesses individuais de cada um e as suas respectivas relações particulares **[5]**.

Um dos principais objetivos do meio jurídico é gerar e documentar os conhecimentos obtidos na área, sejam leis e normativas, informações processuais, etc. Todo dado extraído neste meio é denominado de informação jurídica. Segundo Alonso **[6]**, a informação jurídica pode ser conceituada como qualquer dado ou fato extraído de toda e qualquer forma de conhecimento da área jurídica, obtido por todo e qualquer meio disponibilizado e que pode ser usado, transferido ou comunicado sem a preocupação de estar integrado a um contexto.

A informação jurídica pode ser classificada de três maneiras: legislação, jurisprudência e doutrina. Segundo Passos **[7]**, a informação jurídica também pode ser gerada, registrada e recuperada em três formas distintas: a normativa (legislação), a interpretativa (jurisprudência) e a descritiva (doutrina).

Segundo Passos **[7]** e Barros **[8]**, a atual migração das documentações e extrações de dados jurídicos do meio físico para o digital trouxeram uma série de vantagens, tais como: variedade e quantidade de material disponível, maior acessibilidade e redução de custos. Em contrapartida, há uma série de dificuldades na recuperação de informações jurídicas no meio digital, tais como: a obtenção de toda legislação sobre um determinado assunto, realização de pesquisas de jurisprudência e presença de possíveis deficiências nas bases de dados.

Por fim, quando se fala de bases de dados jurídicas, também é importante saber o que caracteriza uma informação jurídica e se reflete em uma base de dados desta natureza, que segundo Martinho **[9]**, se dá pelos seguintes itens: grandes volumes de informação e rapidez na sua desatualização face a um constante crescimento e criação de novas fontes; público-alvo exigente e diversificado; Necessidade de grande rigor e precisão da sua conservação e rapidez na transmissão de seus dados para assegurar sua correta utilização e aplicação.

2.2 Mineração de Dados

Mineração de dados é uma etapa de um processo maior que envolve várias áreas de

conhecimento diferentes. É comum ao processo o uso de técnicas estatísticas, modelos matemáticos e/ou inteligência artificial para reconhecimento de padrões, proficiência em trabalhar com grandes quantidades de dados, saber lidar com a captura de dados em sistemas de sensores ou sistemas embarcados, conhecimento do problema a ser abordado e capacidade de organizar e apresentar resultados.

Dado que a mineração dos dados é apenas uma etapa, é necessário identificar como todo o processo se desenvolve. O *Knowledge Discovery from Data* (KDD), segundo Fayyad **[10]**, é um processo não trivial, interativo e iterativo, para identificação de padrões que sejam compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. O procedimento de KDD se divide nas fases de agrupamento das informações, pré-processamento, seleção, limpeza, transformação e mineração de dados, além de avaliação do conhecimento extraído durante o procedimento para tomadas de decisões.

3 Materiais e Métodos

Aqui estão descritas as metodologias utilizadas nas etapas de pré-processamento e transformação dos dados.

3.1 Pré-Processamento

Todos os dados utilizados foram extraídos dos portais de cada Tribunal de Justiça, tais como Portal de Serviços de São Paulo **[11]** e Portal do Tribunal de justiça do estado de Minas Gerais **[12]**. Devido cada tribunal possuir sua própria fonte de dados e forma de apresentação, as informações são extraídas de acordo com regras específicas implementadas para cada fonte de dados utilizando recursos como expressões regulares e dicionários de palavras para homogeneizar os termos com a finalidade de agregar valor estatístico.

Na homogeneização dos dados, destacam-se as seguintes transformações: espaços duplos entre palavras para espaços únicos; convenção de todos os caracteres para maiúsculo; remoção de todos os diacríticos, espaços no início e fim de cada frase; e, remoção dos caracteres especiais.

Indicador Êxito é um campo que representa a classe de saída para os algoritmos de classificação, podendo ser Parcialmente Procedente, Procedente ou Improcedente. Para determinar qual indicador de êxito deve ser atribuída a cada processo jurídico, utilizou-se a aplicação de um dicionário de palavras-chave no campo publicação (i.e. todos os comentários dos juízes), a fim de identificar o pensamento dos juízes sobre cada processo julgado.

Após ter feito a atribuição das classes à cada processo jurídico, o próximo passo é realizar a extração das leis e artigos do campo publicação. Para isso, a seguinte expressão regular foi aplicada no campo Publicação:

```
(LEI|ART|ARTIGO)(\W)?(\s+)?(\d{1,10})(\W)?(\d*)(\/)?(\d{1,10})?
```

3.2 Transformação dos Dados

As transformações nos dados foram para utilizar nos algoritmos e classificação em duas etapas.

Na primeira etapa, utilizou-se os seguintes campos:

- Dispositivos (leis e artigos) como entradas;
- Indicador de êxito como saída.

A Tabela I mostra um exemplo da base de dados utilizada nos primeiros experimentos após a aplicação das transformações.

Neste exemplo, o processo 01 foi considerado procedente utilizando a lei 6.830/80 e o artigo 33. O processo 04 teve seu indicador de êxito classificado como parcialmente procedente com a utilização do artigo 33. Por fim, o processo 07 foi classificado como improcedente através do artigo 535.Já segunda etapa, foram utilizados os seguintes dados:

Tabela I – Primeira transformação dos dados

N. Proc.	Ind. de Êxito	LEI. 6.830/86	ART. 535
1	Procedente	1	0
4	Parcialmente Procedente	0	0
7	Improcedente	0	1

Após o resultado dos primeiros experimentos utilizando as transformações citadas anteriormente, os seguintes processamentos foram adicionados com a finalidade de agregar mais informações à base de dados final:

http://dx.doi.org/10.25286/repa.v3i3.907

- Adição das colunas Classe, UF e Fórum ao conjunto de dados;
- Transformação dos valores destas colunas para binário, de forma a representar a existência do valor deste atributo no processo determinado.

A Tabela II mostra um exemplo da base de dados após estas transformações aplicada nesta segunda abordagem. Neste exemplo, o processo 02 foi classificado como procedente para os dados de entrada: PE, Fórum Joana Bezerra, Lei 6.830/80 e artigo 535. O processo 08 foi classificado como procedente tendo como dados de entrada Acre, Trabalhista e o artigo 535. Por último, o processo 10 foi classificado como improcedente para o conjunto de entrada PE, Fórum Joana Bezerra e lei 6.858/80.

Tabela II – Segunda transformação dos dados

N. Proc.	Ind. de Êxito	PE	AC	Joana Bezerra	Trabalhista	LEI. 6.830/80
2	Procedente	1	0	1	0	1
8	Procedente	0	1	0	1	0
10	Improcedente	1	0	1	0	1

4 Parametrização das Técnicas

Para este trabalho, as técnicas utilizadas foram árvore de decisão e Rede Neural Artificial. A utilização da árvore de decisão teve como objetivo mostrar, de forma estrutura e sequência, os dados de entrada, UF, Fórum, Dispositivos (leis e artigos) e Classe Processual para explicar os resultados. Já o propósito da utilização da Rede Neural Artificial foi realizar um estudo comparativo com a árvore de decisão para aferir, por meio de experimentações, qual das duas técnicas oferece maior precisão em seus resultados, levando em consideração o volume e desbalanceamento das bases de dados jurídicas analisadas.

4.1 Árvores de Decisão

Os algoritmos de árvores de decisão utilizados foram, C4.5 e *Recursive Partitioning and Regression Trees* (PART). O algoritmo C4.5 utiliza o cálculo da entropia, equação (1), para montar a estrutura da árvore. Já o algoritmo PART é um método estatístico para análises de multi-variáveis, e utiliza o particionamento recursivo para montar a árvore de decisão com apenas as variáveis de entrada que contém maiores índices de correlação entre si. Para ambas técnicas, a abordagem pós-poda foi adotada para recalcular os nós que possuem baixa relevância na árvore.

$$H(T) = -\sum_{j=1}^{k} \frac{freq(c_j, T)}{|T|} x log_2 \frac{freq(c_j, T)}{|T|} \quad \textbf{(1)}$$

Onde:

- freq(cj , T) quantidade de registros da classe cj em T;
- |T| número total de registros do conjunto T;
- k número de classes distintas que ocorrem em registros de T.

4.2 Redes Neurais

A rede neural utilizada foi a MLP (*Multi-Layer Perceptron*) com o algoritmo de *backpropagation*, funções de ativação não lineares (sigmóde logística), uma camada de entrada com 2900 neurônios, uma camada intermediária com 30 neurônios e uma camada de saída com um neurônio. O algoritmo de treinamento é clássico e tem duas fases: *Feedfoward* e *Backward* ou *Backpropagation*.

Primeiramente, os pesos de cada neurônio foram inicializados com valores pequenos e randômicos e foram definidas as condições de parada da execução do algoritmo (atingir um número máximo de ciclos ou repetir vinte ciclos sem ganhos significativos nos resultados e sem mudanças nos pesos dos neurônios). Em seguida, na fase *Foward*, os padrões de treinamento foram passados pelas unidades da camada de entrada, intermediária e saída. As unidades da camada de entrada receberam os dados e os dissiparam para as unidades da camada seguinte (intermediária). As unidades da camada intermediária ponderaram os sinais recebidos por meio dos seus pesos, aplicaram sua função de ativação para computar suas saídas e as enviaram para as unidades da camada de saída. As unidades da camada de saída

ponderaram seus sinais de entrada e aplicaram sua função de ativação para computarem seus resultados.

Na fase seguinte, a *Backward*, os erros foram calculados na camada de saída e retropropagados para as camadas anteriores que com base nestes, também calcularam seus respectivos erros. Após essas duas fases do algoritmo, os pesos e bias foram ajustados conforme a necessidade.

5 Experimentos Realizados

Os experimentos foram realizados em duas etapas para testar as técnicas de árvore de decisão C4.5 e *Recursive Partitioning and Regression Trees* (PART), e a Rede Neural Artificial. A primeira etapa utilizou-se apenas as leis e artigos citados nas sentenças dos processos (i.e campo publicação - tabela I). Na segunda etapa foram utilizados os campos UF, Fórum, Classe Processual, leis e artigos. O intuito da primeira abordagem foi analisar quais são as influências que as leis e artigos têm no resultado de cada causa jurídica. A intenção da segunda abordagem foi verificar o comportamento das leis e artigos para cada UF, Fórum e Classe Processual.

Foram realizadas 30 execuções para cada técnica. O método *cross validation*, *K-folds* foi adotado para realizar o processo de treinamento, teste e validação, e o valor de k foi igual a 3. Todas as configurações utilizadas para cada técnica serão descritas a seguir.

Para a primeira abordagem, a base de dados utilizada continha 4.157 instâncias com 140 colunas. A distribuição dos dados estava estruturada em 139 colunas representando artigos e leis, e uma coluna representando a saída da classificação, podendo ser avaliada como Procedente, Parcialmente Procedente ou Improcedente.

- C4.5:
  - Limitação de até 7 nós;
  - Pós-poda;
- PART:
  - Aceitação das folhas que contêm pelo menos 25% de relacionamento com os dados;
  - Pós-poda;
- RNA:
  - Máximo de ciclos de treinamento = 600 ciclos;
  - Neurônios na camada de saída = 40 neurônios;
  - Funções de ativação (Na camada escondida e na camada de saída): Sigmóide logística;

Para a segunda abordagem, a base de dados utilizada continha 8.050 instâncias com 2.900 colunas. O mapeamento dos dados estava estruturado nas seguintes colunas: 1.798 leis e artigos, 5 classes, 21 UFs e 1.076 Fóruns. A saída única apresentou os resultados de classificação, sendo avaliada como Procedente ou Improcedente.

- C4.5:
  - Limitação de até 7 nós;
  - Pós-poda;
- PART:
  - Aceitação das folhas que contém acima de 35% de relacionamento com os dados;
  - Pós-poda;
- RNA:
  - Alpha (taxa de aprendizado) = 0,85 e beta (momentum) = 0,25;
  - 100 ciclos máximo de treinamento;
  - 30 Neurônios na camada de saída;
  - Funções de ativação (Na camada escondida e na camada de saída): *Sigmóide* logística.

6 Resultados

Os resultados obtidos através dos experimentos foram organizados em formato visual utilizando gráficos no formato boxplot e tabelas contendo as informações pertinentes à cada técnica.

Todos os resultados obtidos na primeira abordagem estão descritos na Tabela III. Como é possível notar, ao aplicar apenas as leis e artigos como entradas e utilizar três possíveis valores de saída: Parcialmente Procedente, Procedente ou Improcedente, os resultados obtidos das três técnicas não foram muito favoráveis, pois a média das taxas de acerto foram de 65.23%, 65.92% e 59.65% para as técnicas C4.5, PART e Rede Neural.

http://dx.doi.org/10.25286/repa.v3i3.907

O valor do índice Kappa foi 17.68% para C4.5, 19.5% para a técnica PART, e 16.40% para a RNA. Os valores da média do erro absoluto de todas as execuções foram, 35.27%, 34.65% e 42.87% para as técnicas C4.5, PART e Rede Neural. Nesta primeira abordagem, a técnica que obteve o melhor resultado foi a *Recursive Partitioning and Regression Trees* (PART).

Figura 1 – boxplot da primeira abordagem

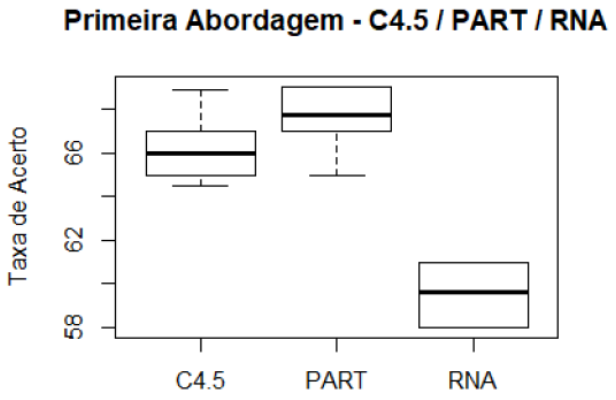


Tabela III – Segunda transformação dos dados

Métrica	C4.5	PART	Rede Neural
Média da taxa de acerto	66.20%	67.59%	59.65%
Índice Kappa	17.68%	19.57%	16.48%
Média do erro absoluto	35.27%	34.65%	42.87%

A Figura 3 mostra o gráfico boxplot dos resultados obtidos na primeira abordagem. Na Tabela IV estão os valores das análises descritiva.

Tabela IV – Resultados estatísticos da primeira abordagem

Métrica	C4.5	PART	RNA
Mínimo	64.50%	65%	58%
Máximo	68.90%	69.04%	61%
Médiana	65.90%	67.75%	59.65%
Média	66.20%	67.59%	59.55%
Desvio Padrão	1.64%	1.56%	1.40%

Com relação a segunda abordagem, os resultados obtidos estão descritos na Tabela V. Nessa segunda abordagem, ao adicionar os demais dados, UF, Fórum e Classe Processual junto com as leis e artigos, e considerando apenas duas possíveis respostas no resultado dos algoritmos, procedente ou improcedente, fica evidente que os resultados obtidos foram muito superiores em comparação com a primeira abordagem. Nessa segunda abordagem, a técnica que obteve maior ganho e melhores resultados em comparação com as demais, foi a Rede Neural Artificial. Na primeira abordagem, a RNA obteve 59.65% para a média da taxa de acerto, porém na segunda abordagem, esse resultado passou para 76.06%. Ou seja, um ganho de 16.41% na média geral da taxa de acerto. Já para as técnicas de árvores de decisão, o ganho na média da taxa de acerto não foi tão significativo, 8.05% para C4.5 e 6.72% para PART.

Tabela V – Resultados obtidos na segunda abordagem

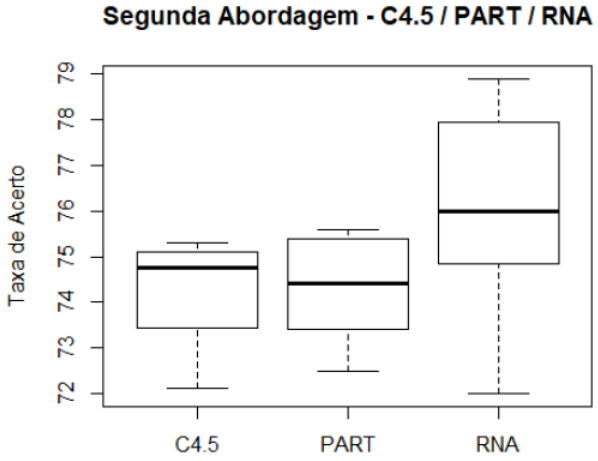
Métrica	C4.5	PART	Rede Neural
Média da taxa de acerto	74.25%	74.31%	76.06%
Índice Kappa	49.14%	50.15%	53.00%
Média do erro absoluto	28.01%	27.43%	27.72%

A Figura 2 mostra o gráfico boxplot dos resultados obtidos na segunda abordagem.

Tabela VI – Resultados estatísticos da primeira abordagem

Métrica	C4.5	PART	RNA
Mínimo	72.10%	72.50%	72%
Máximo	75.29%	75.60%	78.90%
Médiana	74.75%	74.40%	76.10%
Média	74.25%	74.31%	76.60%
Desvio Padrão	1.16%	1.17%	2.28%

Figura 2 – boxplot da segunda abordagem



7 Conclusões e Trabalhos futuros

Com o término deste trabalho, pôde-se observar que a maior dificuldade esteve em lidar com o tratamento de uma base de dados de natureza jurídica, de modo que os resultados finais pudessem convergir para algo dentro do confiável e aceitável aos padrões técnicos e necessidades corporativas e acadêmicas. Além disto, a subjetividade do tipo de informação jurídica (jurisprudência, que gera dados interpretativos)

reduz consideravelmente a eficácia da aplicação de técnicas de aprendizado de máquina, que realizam classificações objetivas.

Como trabalhos derivados e relacionados, visando o aprimoramento dos resultados obtidos e modelos desenvolvidos neste trabalho, pode-se destacar a importância da realização de:

- Análises preditivas utilizando modelos estatísticos e/ou computacionais para estimar percentuais de confiabilidade na utilização de determinadas leis e artigos para dados processos judiciais, tendo como base os resultados já obtidos e modelos desenvolvidos neste trabalho;
- Com o objetivo de auxiliar os juristas, desenvolver aplicação para realização de classificação de documentação normativa segundo conteúdo do assunto. Isto colaboraria para otimizar pesquisas referentes a processos de determinados casos que se encontrem vinculados aos assuntos em questão. A organização desse tipo de informação já proporcionaria aos profissionais da área e pesquisadores que desejam realizar aplicações na mesma, bases específicas para busca e extração do conhecimento desejado;
- Avaliação de envolvimento de súmulas unificadas nas análises de jurisprudência, pois essas informações fornecem maior confiabilidade quanto ao comportamento dos magistrados com relação às análises dos processos;
- Aplicar técnicas de mineração textual afim de analisar todo o contexto geral da publicação, não apenas às leis e artigos.

Referências

[1] Portal CNJ - Processo Judicial Eletrônico (PJe). Disponível em <http://www.cnj.jus.br/tecnologia-da-informacao/processo-judicialeletronico-pje>. Acesso em: 30 de junho 2017.

[2] Portal Conjur. Disponível em <http://www.conjur.com.br/2015-jun 24/conheca-softwares-juridicos-usados advogados>. Acesso em: 30 de junho 2017.

[3] REALE, Miguel. **Lições preliminares de direito**. São Paulo: Saraiva, 2003, p. 321.

[4] REOLON, Suzana Minuzzi. A linguagem jurídica e a comunicação entre o advogado e seu cliente na atualidade. Direito & Justiça, v. 36, n. 2, p. 180-191, jul./dez. 2010. Disponível em: <http://revistaseletronicas.pucrs.br/ojs/index.php/fadir/article/viewFile/9101/6347>.

[5] SILVA, Andréia Gonçalves. **Fontes de informação jurídica: conceitos e técnicas da leitura para o profissional da informação**. Rio de Janeiro: Interciência, 2010.

[6] ALONSO, Cecília Andreotti Atienza. A informação jurídica face às comunidades da área do direito e a dos fornecedores da informação jurídica. In: CIBERNÉTIA, SIMPÓSIO INTERNACIONAL DE PROPRIEDADE INTELECTUAL, INFORMÁTICA E ÉTICA, 1998, Florianópolis. **Anais**... Florianópolis, 1998.

[7] PASSOS, Edilenice (Org.). **Informação Jurídica: teoria e prática**. Brasília, DF: Thesaurus, 2004.

[8] BARROS, Lucivaldo Vasconcelos. Avaliação de Fontes de informação para busca de documentos jurídicos na Internet: uma reflexão à luz das cinco leis de Ranganathan e dos critérios de acessibilidade. In: SEMINÁRIO NACIONAL DE DOCUMENTAÇÃO E INFORMAÇÃO JURÍDICAS, 2., 2010, Brasília. **Anais**... Brasília:

[9] MARTINHO, A. M. O bibliotecário jurídico: identidade e competências profissionais. In: ENCONTRO NACIONAL DE BIBLIOTECAS JURÍDICAS, 1., 2004, Lisboa. **Anais**... Lisboa: Faculdade de Direito da Universidade de Lisboa, 2006.

[10] FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

[11] Portal de Serviços de São Paulo. Disponível em <https://esaj.tjsp.jus.br/cpopg/open.do>. Acesso em: 28 abr. 2017.

http://dx.doi.org/10.25286/repa.v3i3.907

[12] Portal do Tribunal de Justiça do Estado de Minas Gerais. Disponível em <http://www.tjmg.jus.br/portal/>. Acesso em: 28 abr. 2017.