Categorizador de ementas trabalhista

Thiago Ferauche, Maurício Amaral de Almeida, Márcia Ito
Laboratório de Pesquisa em Ciência de Serviços – Programa de Mestrado -
Centro Estadual de Educação Tecnológica Paula Souza – SP – Brasil
thiago.ferauche@gmail.com, madealmeida@gmail.com,
marcia.ito@centropaulasouza.sp.gov.br

Abstract – The summaries that build the Tribunal Regional do Trabalho da 2a. Região – SP jurisprudence have their own classification, devised and performed by specialists of the Labor Law. This proposed research intends to perform the same classification using an automatic system, using techniques of text mining, and verify its effectiveness by the validation of these specialists Labor Law.

Resumo – As ementas que formam a jurisprudência do Tribunal Regional do Trabalho da 2ª. Região – SP possuem uma classificação própria, idealizada e executada por especialistas do Direito do Trabalho. Esta proposta de pesquisa pretende executar a mesma classificação através de um sistema automático, utilizando-se de técnicas de mineração de texto, e verificar a sua eficácia com a validação dos mesmos especialistas do Direito do Trabalho.

Palavras-chave: Categorização de textos, mineração de textos, textos jurídicos, jurisprudência, sistema de serviço.

Introdução

Existe uma forte frente política para a Modernização Tecnológica do Judiciário Brasileiro, não apenas para a compra de novos equipamentos, mas também para o desenvolvimento de sistemas capazes de trabalhar com documentos eletrônicos, viabilizando assim o chamado “Processo Eletrônico”, instaurado através da Lei Federal Nº 11.416, de 19 de dezembro de 2006 [1], em seu capítulo III, e regulamentado no poder judiciário trabalhista através da Instrução Normativa Nº 30 de 2007 do Tribunal Superior do Trabalho [2].

Entretanto, documentos jurídicos são de difícil acesso devido ao grande número de decisões existentes e a forma que elas são armazenadas (livros, resenhas, banco de dados, etc.). Esta situação impele os profissionais jurídicos a despender mais tempo na busca por uma decisão jurídica adequada ao seu problema. Trabalhos de recuperação de informação já foram realizados com documentos da jurisprudência do Tribunal de Justiça de Santa Catarina, utilizando a abordagem “Baseada em Casos” [3][4].

Nesta proposta de pesquisa pretende-se construir, utilizando técnicas de Mineração de Texto, um protótipo de um sistema especialista que seja capaz de aprender e categorizar os documentos da jurisprudência trabalhista utilizando os documentos da Ementa da Jurisprudência do Tribunal Regional do Trabalho da 2ª Região - SP, previamente categorizados por especialistas no domínio do Direito Trabalhista.

Jurisprudência

A jurisprudência, no domínio do Direito, possui um importante papel como fonte do direito, sendo que o seu conteúdo auxilia a interpretação da lei e sua aplicação na solução de um problema jurídico. Os documentos da jurisprudência são documentos gerados em formato de hipertexto a partir de informações do banco de dados, conforme a Figura 1.

TIPO: RECURSO ORDINÁRIO
DATA DE JULGAMENTO: 16/11/2004
RELATOR(A): RICARDO ARTUR COSTA E TRIGUEIROS
REVISOR(A): CARLOS ROBERTO HUSEK
ACÓRDÃO Nº: 20040643829
PROCESSO Nº: 01152-1998-445-02-00-5 **ANO:** 2004 **TURMA:** 4ª
DATA DE PUBLICAÇÃO: 26/11/2004
PARTES:

RECORRENTE(S):
INSTITUTO NACIONAL DO SEGURO SOCIAL INSS

RECORRIDO(S):
RODRIMAR S/A TRANSP EQUIPS INDS ARM GER
GENILSON ALMEIDA GOIS
EMENTA:

INSS. RECURSO ORDINÁRIO. NÃO CONHECIMENTO. INADEQUAÇÃO, AUSÊNCIA DE INTERESSE E IRREGULARIDADE DA REPRESENTAÇÃO.Recurso do INSS que não se conhece em razão de: (1) inadequação, vez que é notória a impropriedade do recurso ordinário (art. 895, CLT), cabível apenas na fase cognitiva, para atacar decisão terminativa em sede de execução, para a qual o recurso específico é o agravo de petição (art. 897, CLT), sendo inaplicável à espécie o princípio da fungibilidade; (2) ausência de interesse porquanto o valor previdenciário já foi quitado, configurando sanha arrecadatória a pretensão do Instituto de receber o que já lhe foi pago; (3) irregularidade da representação, em vista da subscrição do apelo por advogado particular e não por procurador autárquico.

ÍNDICE: PREVIDÊNCIA SOCIAL, Recurso do INSS

Serviço de Jurisprudência e Divulgação

Figura 1 - Jurisprudência retirada do site do Tribunal Regional do Trabalho da 2ª Região

O documento segue uma estrutura, sendo possível identificar uma espécie de cabeçalho, com informações que identificam a origem da jurisprudência, como: Tipo do processo, Data de julgamento, Juiz Relator e Revisor do acórdão, Número do acórdão, Ano do acórdão, Turma do acórdão, Data de publicação, Número do processo e Partes envolvidas. É possível ainda identificarmos mais duas partes da estrutura: a ementa e o índice. A ementa é onde se encontra a síntese do que foi decidido no acórdão, suas premissas e justificativas, nela está concentrada todo o conhecimento da jurisprudência, e é digitada por servidores públicos das Secretarias das Turmas. Por último, existe o índice que é a classificação da jurisprudência, utilizada para organizar e facilitar a busca da jurisprudência, preenchido pelos servidores públicos do Serviço de Jurisprudência e Divulgação, que podem ser identificados como os especialistas do sistema, pois são eles que lêem a ementa, identificam relações na área do Direito, e depois classificam a jurisprudência.

Metodologia

O trabalho será dividido em 4 etapas: Extração das informações da jurisprudência, Pré-processamento do texto das ementas, Classificação da jurisprudência e Verificação da eficiência do classificador por especialistas.

As informações da jurisprudência do TRT da 2ª região encontram-se espalhadas em seu vasto banco de dados, e são extraídas através de um sistema já desenvolvido para gerar documentos no formato HTML, conforme apresentado na Figura 1, com o intuito de serem utilizados no sistema de busca por localização de palavras disponível através do endereço "*http://www.trt02.gov.br/cgi-bin/db2www/geral/consulta/macros/ementas_palavra.mac/input"*.

A extração das informações será feita diretamente do banco de dados do TRT da 2ª região para a geração de arquivos, um para cada ementa, em formato de texto puro, o que servirá de entrada para a próxima etapa.

Em um processo de Mineração de Textos, no qual os dados estão em um formato não estruturado textual, é necessário que esses dados sejam pré-processados, de tal forma que possam ser submetidos a algoritmos de aprendizado.

Para executar a etapa de pré-processamento da ementa, os textos das Ementas serão extraídos dos arquivos textos, e serão pré-processados pela ferramenta PRE-TEXT II, desenvolvida pelos pesquisadores do Laboratório de Inteligência Computacional da USP (LABIC-USP) [5]. Uma de suas principais funcionalidades é transformar palavras em *stem*s (uma espécie de radical das palavras), utilizando a técnica conhecida como *bag-of-words*, em conjunto com uma lista de palavras a serem descartadas, *stop-list*. A saída desta etapa é um arquivo contendo os documentos representados na forma de uma tabela atributo-valor e um arquivo contendo informações sobre os atributos.

Na etapa de categorização da Jurisprudência serão utilizados algoritmos de aprendizado de máquina com o objetivo de extrair a partir dos documentos pré-processados, conhecimento na forma de regras de associação, relações, segmentação, classificação de textos, entre outros. Será utilizado o algoritmo Support Vector Machine (SVM) para a extração do conhecimento e para efetuar a classificação das ementas.

O SVM é um dos principais algoritmos para categorização de textos que mais vêm sendo utilizado e, por conseguinte, tem mostrado ser um dos mais eficientes nessa área. SVM é um método de aprendizado de máquina supervisionado oriundo da Estatística clássica que foi introduzido como uma técnica para reconhecimento de padrões para a solução de uma variedade de problemas de aprendizado [6]. SVM é um algoritmo de aprendizado e se baseia na teoria de aprendizado estatístico, combinando controle de generalização com uma técnica para tratar o problema da dimensionalidade [7].

Será utilizada como base a ferramenta WEKA – Waikato Environment for Knowledge Analysis [8], desenvolvida pela Universidade de Waikato, totalmente desenvolvida em linguagem JAVA e possui seu código aberto, o que permitiu variações da ferramenta para vários tipos de problemas. Pretende-se neste trabalho analisar todas as variações e adaptá-las para utilizar o algoritmo SVM no contexto jurídico, a que se destina o trabalho aqui presente.

Após a classificação dos textos da jurisprudência, para averiguar a eficiência do classificador automático, será desenvolvido um sistema de busca baseado em um *"search engine"*, utilizando-se os índices de identificação da jurisprudência,

compondo assim a quarta e última etapa, a verificação da eficiência do classificador por especialistas.

Desta maneira, especialistas em jurisprudência trabalhista, poderão averiguar a eficiência do classificador, e a eficácia da gestão do conhecimento, através de questionários [9] após a utilização do sistema de busca.

Conclusão

Na Jurisprudência encontra-se uma grande quantidade de conhecimento armazenado. O auxílio computacional para organizar, gerenciar e utilizar tal conhecimento é pouco explorado.

Existem técnicas de mineração de texto, como o pré-processamento com *bag-of-words*, e de análise de dados como o *Support Vector Machine*, que poderão ser utilizadas para a criação de um protótipo com a finalidade de executar a categorização automática das ementas da jurisprudência do Tribunal Regional do Trabalho da 2ª. Região – SP.

A eficácia de tal categorização automática é incerta, e somente será possível aferi-la após o aval dos especialistas que realizam a categorização manualmente.

Referências

[1] Presidência da República, Casa Civil, Sub-chefia para Assuntos Jurídicos. “Lei Nº 11.419, de 19 de dezembro de 2006”. http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11419.htm. Acessado em 2 de dezembro de 2010.

[2] Tribunal Superior do Trabalho. “Instrução Normativa Nº 30 de 2007”. http://www.tst.gov.br/DGCJ/instrnorm/30.htm. Acessado em 2 de dezembro de 2010.

[3] Bueno, T. D. (1999). “Recuperação da Informação Jurídica – Uma abordagem Baseada em Casos”. Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Brasil.

[4] Instituto de Governo Eletrônico, Inteligência Jurídica e Sistemas. “Uso da teoria jurídica para recuperação em amplas bases de textos jurídicos”. http://www.i3g.org.br/producaotc/direito_digital/inteligencia/enia99b.htm. Acessado em 11 de janeiro de 2009.

[5] Soares, M.V.B., Prati, R.C., Monard, M.C. (2008). “PreText II: Descrição da Reestruturação da Ferramenta de Pré-processamento de Textos”. Relatório Técnico nº 333, Instituto de Ciências Matemáticas e de Computação, USP, Brasil.

[6] Huang, Y. (2000). “Support Vector Machines for Text Categorization Based on Latent Semantic Indexing”. Eletrical and Computer Department, The Johns Hopkins University.

[7] Joachims, T. (1998). “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”. Computer Science Department, University of Dortmund.

[8] Universuty of Waikato. “WEKA 3: Data Mining Software in Java”. http://www.cs.waikato.ac.nz/ml/weka/. Acessado em 4 de setembro de 2009.

[9] Gonçalves, L. S. M.; Rezende, S. O. (2001). “Categorização em Text Mining”. Escola de Engenharia, Instituto de Ciências Matemáticas e de Computação, USP, Brasil.