Anais do Conic-Semesp. Volume 1, 2013 - Faculdade Anhanguera de Campinas - Unidade 3. ISSN 2357-8904

TÍTULO: DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DO PROCON UTILIZANDO ALGORITMOS E TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

CATEGORIA: EM ANDAMENTO

ÁREA: ENGENHARIAS E TECNOLOGIAS

SUBÁREA: COMPUTAÇÃO E INFORMÁTICA

INSTITUIÇÃO: UNIVERSIDADE SANTA CECÍLIA

AUTOR(ES): JÉSSICA MANAIA CASSIMIRO DA SILVA, EDSON PEREIRA DO NASCIMENTO, EDUARDO FRANCIS DOS SANTOS, FELIPE BRANDÃO FRISCHEISEN

ORIENTADOR(ES): LUIZ ANTONIO FERRARO MATHIAS

Realização: SEMESP Apoio: Anhanguera

1. RESUMO

Em média, a cada vinte meses dobram o número de informações nas bases de dados do mundo inteiro. Nestas, existe um volume de conhecimento muito grande que normalmente, não pode ser extraído através de métodos computacionais tradicionais e triviais de busca. Esta grande disponibilidade de dados permitiu o surgimento de uma nova área de pesquisa e desenvolvimento em Computação: a descoberta de conhecimento em base de dados (Knowledge Discovery in Databases), que conta com uma etapa fundamental intitulada de mineração de dados (Data Mining), a qual aplica técnicas e paradigmas baseados em Inteligência Artificial para a identificação de conhecimento válido em grandes volumes de dados.

2. INTRODUÇÃO

O Cadastro Nacional de Reclamações Fundamentadas é formado pelo PROCON (Programa de Proteção e Defesa do Consumidor) junto ao SINDEC (Sistema Nacional de Informações de Defesa do Consumidor), no período de 12 meses. Esse cadastro representa uma importante referência para órgãos de defesa do consumidor, imprensa, e para os próprios fornecedores. No último cadastro divulgado, por exemplo, as Reclamações Fundamentadas representaram 15,3% de todos os atendimentos.

De um modo geral, são registradas na forma de Reclamação demandas de consumidores que necessitam da realização de audiência para tentativa de resolução, o que pode acontecer tanto pela complexidade da demanda quanto pelo tipo de postura adotada pelo fornecedor no enfrentamento do conflito.

O objetivo deste projeto de pesquisa acadêmica é a aplicação dos conceitos pertinentes ao processo de descoberta do conhecimento em bases de dados, utilizando técnicas e algoritmos de Mineração de Dados baseadas em Inteligência Artificial e software de mineração de dados Weka (Waikato Enviroment for Knowledge Analisys – Ambiente para Análise de Conhecimento de Waikato), visando assim, a identificação de padrões e correlações dos registros realizados na base do PROCON.

3. OBJETIVOS

Aplicar técnicas de mineração de dados e conceitos da inteligência artificial na base de dados do PROCON, com o apoio do software de mineração de dados

Weka, a fim de descobrir associações e padrões nos registros encontrados na base de dados.

4. METODOLOGIA

A metodologia empregada no projeto considerou uma pesquisa bibliográfica a respeito do processo de descoberta do conhecimento em base de dados, as etapas que compõem tal processo, um estudo aprofundado do funcionamento da ferramenta de mineração de dados WEKA, com foco em seus algoritmos de busca baseados em inteligência artificial, uma análise crítica dos dados registrados na base de dados do PROCON, seus atributos, domínios e abrangência, aplicação deste banco de dados na ferramenta após etapas de limpeza e transformação de dados.

5. DESENVOLVIMENTO

Após a seleção e agregação dos dados disponibilizados na base de dados do PROCON, foi necessária uma limpeza de tais dados que continham aproximadamente 150 mil registros, muitos dos quais com nomes duplicados, escritos de forma incorreta ou com falhas de formatação. Para auxiliar na conversão dos dados, foi desenvolvido um software em linguagem C# que converteu os registros armazenados em planilha eletrônica para formato arff (extensão utilizada pela ferramenta Weka). Durante a conversão do formato dos dados, o software desenvolvido realizou uma limpeza vertical dos dados, com a eliminação dos atributos não selecionados, e uma limpeza horizontal, excluindo registros nulos, para o correto processamento dos algoritmos baseados em inteligência artificial.

A próxima etapa do desenvolvimento prevê o emprego de alguns algoritmos das tarefas de associação e clusterização - dentre eles Tertius, Apriori e Simple K-Means -, na ferramenta Weka, na etapa de mineração de dados, para a identificação de padrões e correlações entre os dados.

6. RESULTADOS PRELIMINARES

Após as pesquisas teóricas sobre KDD e Data Mining, iniciamos a parte prática, que reflete no objetivo do projeto.

A base de dados utilizada foi baixada do site “dados.gov.br”, do governo federal, onde são encontradas diversas informações de órgãos públicos, com base na lei de acesso à informação criada em 2011.

Depois de baixarmos a base de dados, tivemos que trata-la, removendo erros de formatação e digitação existentes no arquivo original - que encontra-se em arquivo CSV -, e, através do software de conversão desenvolvido pelo grupo, foi criado o arquivo ARFF, considerando-se apenas os seguintes atributos: Região, UF, RazãoSocial, DescCNAEPrincipal, Atendida, DescricaoAssunto, DescricaoProblema, SexoConsumidor, FaixaEtariaConsumidor.

Com o arquivo ARFF criado, importamos o mesmo na ferramenta WEKA, através da interface Explorer, e testamos o desempenho de diversos algoritmos disponíveis na mineração destas informações.

Preliminarmente, descartamos utilizar a tarefa de classificação, devido a seus algoritmos trazer melhores resultados em bases de dados com valores numéricos, o que não é o caso, onde temos todos os valores nominais.

Sendo assim, até o presente momento, estamos estudando a ferramenta Weka, juntamente com os algoritmos Tertius e Apriori, relativos à tarefa de Associação, e, Simple K-Means e Farthest-First, pertinentes à tarefa de Clusterização, que retornaram como resultados alguns padrões relevantes que também estão sendo analisados. Porém, ainda poderá haver troca destes algoritmos, no caso de encontrarmos outros que apresentem melhores resultados.

7. FONTES CONSULTADAS

FAYYAD, Usama et al. From Data Mining to Knowledge Discovery in Databases. AI Magazine, California, v. 17, n. 3. P. 37-54, 1996.

FRAWLEY, William J. et al. Knowledge Discovery in Databases: An Overview. AI Magazine, Califórnia, v. 13, n. 3. P. 57-70, 1992.

GOLDSCHIMIDT, Ronaldo; PASSOS, Emmanuel. Data Mining um guia prático. Rio de Janeiro: Elsevier, 2005.

HAN, Jiawei; KEMBER, Micheline. Data Mining: Concepts and Techniques 2nd Edition. São Francisco, Califórnia: Elsevier, 2006.

WILLIAMS, Graham J.; HUANG, Zhexue. Modelling the KDD Process. Data Mining Portfolio, Australia, 1996.