UM ESTUDO SOBRE AS TÉCNICAS DE MINERAÇÃO DE DADOS CLASSIFICAÇÃO E CLUSTERIZAÇÃO

LIMA, Luiz Gole[1]; MOZZAQUATRO, Patrícia Mariotto[2]

Palavras-Chave: Mineração de Dados. Técnica de Classificação. Técnica de Clusterização

Introdução

O excedente informacional produzido nesses últimos anos, em particular, na Internet, trouxe consigo uma nova dificuldade aos usuários da informação eletrônica (MARCONDES; SAYÃO, 2002). Em conseqüência disso, vê-se que, cada vez mais, o problema de identificação da informação relevante para um usuário algo crítico, formando assim, o fenômeno conhecido como "sobrecarga de informação". O que torna questionável é a estrutura organizacional, ou seja, grande parte das informações apresenta pouca organização e muitas vezes não atendem as necessidades dos usuários. A maioria dos sistemas e ferramentas computacionais não possibilitam mecanismos que proporcionem buscas dentro de um contexto específico almejado, exigindo que usuários explorem cada informação individualmente, não proporcionando o processamento automático das mesmas, ou melhor, não fornecendo mecanismos para classificação e filtragem na extração de informações (DIAS, 2004). Neste sentido, pesquisas vêm sendo desenvolvidas com o intuito de incrementar o processo de classificação da informação e melhorar o processo de extração das mesmas em base dos dados. A Mineração de Dados (MD) pode ser vista como uma técnica para auxiliar nos processos de extração e busca das informações, permitindo um tratamento individualizado, tornando possível o conhecimento de características e preferências dos usuários.

A mineração de dados é dividida em métodos e técnicas, as quais são responsáveis pela análise e a extração das informações que estão armazenados em um banco de dados. Na pesquisa proposta serão estudadas as técnicas de classificação e segmentação ou *clustering* a serem aplicadas na extração de informações em base de dados. A escolha das técnicas justifica-se pela sua abordagem nos trabalhos de (MACHADO, 2002), (HAN, 2001),(MOZZAQUATRO,2006).

[1] Acadêmico do Curso de Ciência da Computação – Universidade de Cruz Alta (email: lhlima85@gmail.com)
[2] Professora orientadora do Curso de Ciência da Computação – Universidade de Cruz Alta (email: patriciamozzaquatro@gmail.com

Segundo (HAN, 2001), a técnica de classificação consiste em descobrir conhecimento capaz de prever situações ou eventos futuros. Ela se enquadra no aprendizado supervisionado, pois executa um método que separa ou divide um item em uma ou várias classes. Já a técnica de Segmentação ou *Clustering* constitui-se em um método descritivo que identifica um conjunto finito de classes ou clusters para descrever os dados. É também conhecida como identificação de classes, segmentação, ou agrupamento automático.

Técnicas de Mineração de Dados

Mineração de Dados nada mais é que extrair ou minerar dados de um grande conjunto, estocados em uma base de dados. Segundo (DIAS, 2004). as técnicas mais utilizadas são classificação e *Clustering*, sendo que cada uma delas pode envolver um ou mais algoritmos. A seguir serão apresentadas as técnica de Mineração de dados Classificação e Clusterização.

A técnica de MD Classificação tem como objetivo classificar objetos de itens em uma entre diversas classes previamente definidas, com base em propriedades comuns entre um conjunto de objetos no banco de dados e seu método é supervisionado. Na técnica de classificação, possui-se um conjunto de dados pré-determinado para a classificação, isto caracteriza um método de aprendizado supervisionado, onde o algoritmo é controlado por parâmetros que são passados ao sistema (MACHADO, 2002). O objetivo da classificação é examinar os dados em treinamento assim enviando uma relação especifica ou um modelo para cada classe utilizando atributos disponíveis nos dados que foi feita a descrição. (MOZZAQUATRO, 2006). A seguir é apresentado o algoritmo de classificação *Classification and Regression Trees – CART.*

O algoritmo *Classification and Regression Trees* (*CART*) baseia-se no modelo de regressão não-paramétrico que estabelece uma ligação entre as variáveis independentes (x), com apenas uma única variável dependente, resposta (*target*) ou alvo. O modelo é encaixado mediante sucessivas divisões binárias no grupo de dados, assim criando novos subgrupos de dados da variável resposta cada vez mais homogêneos (RODRIGUES, 2004). Os componentes básicos do algoritmo CART são os "nós e as regras de decisões". Os nós estão associados aos subconjuntos resultantes da aplicação de uma regra de divisão a determinado conjunto de dados. Segundo (DIAS, 2004). o algoritmo de CART apresenta os seguintes passos: 1- Dado um seguinte nó, o algoritmo coloca em prática todas as regras que possui para separar o conjunto de dados associados ao mesmo. Cada valor que uma variável

assume dentro de uma base de dados é uma possível regra. As regras mais prováveis para dividir um nó seriam apresentadas: X >= 0.1? - X >= 0.7? - X >=3.4? - Y é baixo? - Y é médio? - Y é alto? Assim a regra se divide em duas gerando dois nós filhos. No caso os que correspondem ao sim para uma regra vão para nó filho da esquerda e os que respondem ao não, vão para a direita; 2- o algoritmo CART aplica em cada nó-filho um critério de partição com todas as regras possíveis. Assim é definido como o grau de impureza de um nó t que é mostrado como 1 – FI, onde FI é a função de impureza: $FI = -\sum p2\ (j|t)$ para j = 1, 2.., k; 3- o algoritmo escolhe a regra que alcançou a redução de impureza da árvore; 4- ocorre uma divisão em dois grupos de dados atribuindo a regra selecionada; 5- cada nó-filho é então classificado dentro de uma possível classe do grupo de treinamento; No último passo o CART permanece dividindo a árvore aplicando os passos que foram mostrados de recursiva aos nós-filhos gerando até que só exista nós-filhos 100% puros, ou com grau de impureza considerado aceitável.

Técnica de Mineração de Dados Clustering

A técnica de Clusterização é a classificação não supervisionada de padrões em conjuntos de dados, onde o objetivo é localizar uma possível igualdade entre os objetos e formar uma classificação de acordo com suas propriedades em comum, diferente de uma classificação supervisionada, em que os padrões já têm uma classificação referente a ele (DIAS, 2004). Os Algoritmos de Clusterização dividem os dados em grupos úteis ou significativos, chamados clusters, nos quais a similaridade intracluster é maximizada e a similaridade inter-cluster é minimizada (BELTRAME; FONSECA, 2010, p.15). A técnica de aprendizado não supervisionado ou *clustering* (Agrupamento) tem como prioridade, buscar a extração de informação relevante de dados que não estão rotulados. Para conseguir medidas de similaridade entre dois clusters assim como um critério global são utilizados atualmente muitos algoritmos que realizam o trabalho de agrupamento. A seguir é apresentado o algoritmo de *clusterização: K-means.* O algoritmo de *clusterizacao, K-Means* (também chamado de *K-Médias*) fornece uma classificação de informações de acordo com os próprios dados. Este algoritmo tem como base a análise e comparações entre os valores numéricos dos dados. Com isto, o algoritmo automaticamente vai enviar uma classificação automática sem a necessidade de nenhuma supervisão humana, sem nenhuma pré-classificação existente. O algoritmo de *K-means* pode ser visto através dos passos apresentados: 1- Atribuem-se valores iniciais para os protótipos seguindo algum critério, por exemplo, sorteio aleatório desses valores dentro dos limites de domínio de cada atributo; 2- Atribui-se

cada objeto ao grupo cujo protótipo possua maior similaridade com o objeto; 3- Recalcula-se o valor do centróide (protótipo) de cada grupo, como sendo a média dos objetos atuais do grupo; 4- Repete-se os passos 2 e 3 até que os grupos se estabilizem.

Conclusão

Através deste artigo buscou-se apresentar um estudo sobre as técnicas de Mineração de dados Classificação e Clusterização. As técnicas analisadas auxiliam nos processos de extração e busca das informações, permitindo um tratamento individualizado dos dados, tornando possível o conhecimento de características e preferências dos usuários. Este artigo é parte integrante de um trabalho de conclusão de curso que objetiva implementar os algoritmos integrante das técnicas de clusterização (K-means) e Classificação (*Classification and Regression Trees (CART)* comparando e medindo sua eficiência na extração de informações em banco de dados.

Referências

BELTRAME, Walber Antônio Ramos; FONSECA, Felipe Cesar Stanzani, 2010. **Aplicações Práticas dos Algoritmos de Clusterização Kmeans e Bisecting K-means.** Departamento de Informática – Universidade Federal do Espírito Santo (UFES) Av. Fernando Ferrari, 514 – Vitória – ES – Brasil

DIAS, Maria Abadia Lacerda**; Extração Automática de Palavras-Chave na Língua Portuguesa Aplicada a Dissertações e Teses da Área das Engenharias.** Dissertação de Mestrado em Engenharia Elétrica, FEEC-UNICAMP, SP 2004

HAN, J.; KAMBER, M, 2001. **Mining: Concepts and Techniques**. San Francisco: Morgan Kaufmann. 550 p.

MACHADO, Letícia Santos. **Mineração do Uso da Web na Educação a Distância**: Propostas para a Condução de um Processo a partir de um Estudo de Caso.Porto Alegre: Pontifícia Universidade Católica do Rio Grande do Sul, 2002. Dissertação (Pós-graduação em Ciência da Computação), Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, 2002.

MARCONDES, C. H.; SAYÃO, L. F. **Documentos Digitais e Novas Formas de Cooperação entre Sistemas de Informação em C&T**. Ci. Inf., Brasília, v. 37, n. 3, p. 42–54, 2002.

MOZZAQUATRO, Patrícia Mariotto. **Estudo da Aquisição e Modelos de Usuários da Biblioteca Digital Acadêmica**. Trabalho de Conclusão de Curso em Sistemas de Informação. Universidade Luterana do Brasil - ULBRA, 2006.

RODRIGUES, Marco Antonio dos Santos, 2004. **Árvore de classificação**. Universidade dos Açores – departamento de Matemática – Monografia.