Artigo 2: Uma Experiência em Mineração de Textos Utilizando Clustering Probabilístico e Clustering Hierárquico

Este artigo aborda a mineração de textos que é um processo de conhecer de que se trata o texto através da freqüência de suas palavras, é elaborada uma tabela com os valores da freqüência para palavras chaves, a quantidade das palavras é comparada por algoritmos não supervisionados de aprendizagem de máquina tais como Clustering Probabilistico e Hierárquico.

O objetivo do artigo é descrever os experimentos realizados com mineração de textos na língua inglesa usando os algoritmos de Clustering Hierárquico e Clustering Probabilística. A vantagem desse estudo é o agrupamento de textos sem precisar saber de que ele se trata.

A grande diferença entre Mineração de Dados e a Mineração de Textos é que em MD temos dados estruturados enquanto que em MT temos dados não estruturados, o que é um problema, para resolvê-lo precisamos de uma etapa de pré-processamentos de dados e para esta situação apresenta-se uma abordagem simples mas muito custosa é a criação de uma tabela no formato atributo X valor onde cada texto é um exemplo nessa tabela e as palavras de cada texto determinam os atributos. Após o pré-processamento é realizado os algoritmos de agrupamento (Clustering) para agrupar textos de mesmo tema sem a necessidade de conhecê-los.

O pré-processamento dos textos foi realizado com a ferramenta PreTex que atua em textos em português, espanhol e inglês da seguinte maneira.

* Eliminação dos elementos que não são radicais das palavras, por exemplo sufixos e prefixos.
* Eliminação de palavras insignificantes para o texto como um todo, tais como conjunções e artigos.
* Calcular medidas de atribuição à valores.
* Lei de Zipf, lei que calcula através de potencias as relações entre os elementos do texto.
* Cortes de Luhn elimina-se palavras com baixa freqüência nos textos.
* Escolhem-se palavras pela sua complexidade, palavras simples ou compostas.
* Geração de gráficos
* Geração da tabela atributo X valor

Ao fim do pré-processamento cada texto foi representado por 430 atributos mais a sua classificação e seu registro.

Os exemplos foram submetidos ao método de Clustering Probabilístico través do algoritmo AutoClass que é baseada em técnicas probabilísticas de Bayes que determina o número de clusters automaticamente através de atributos discretos e contínuos, manipula *missing values*. Ou seja, ele descreve clusters a partir da distribuição probabilística sobre os atributos de cada exemplo. Para rotular os exemplos dos clusters encontrados pelo AutoClass foi utilizada a ferramenta InClass. Foi utilizado também o software See5 o qual é um software de Aprendizagem de Máquina supervisionado que manipula atributos induzindo regras como árvores de decisão.

Para formar Cluster Hierárquico foi utilizado o algoritmo WebIHCA (Web Incremental Hierarchical Clustering Algorithm) que funciona da seguinte forma, é criado um conjunto de categorias, o conjunto é incrementado a cada descoberta de uma nova categoria pelo algoritmo. Para cada categoria é criado um conjunto de chaves que o descreve e cada chave é associada a um peso que descrimina sua importância para a categoria a qual faz parte. Estas categorias são organizadas em uma hierarquia crescente que pode ser vista como uma árvore. Cada nó interno sumariza palavras chaves que são compartilhadas com categorias filhas. Assim temos uma generalização das folhas para a raiz e uma especialização da raiz às folhas. Desta forma o algoritmo tenta enquadrar os textos comparando-os a partir da raiz em direção as folhas. Se nenhuma folha enquadra o exemplo de maneira satisfatória então é criado uma nova categoria.

Foi identificado que a utilização de modelos mais complexos perde em desempenho em relação a modelos mais simples. No algoritmo AutoClass os agrupamentos encontrados não apresentam relações diretas com as categorias apresentadas pelo especialista, enquanto que no algoritmo WebIHCA eles estão diretamente relacionados a essas categoria. Entretanto não foi possível concluir que um método é melhor que o outro, pois na utilização de vários especialista eles diferem entre si devido aos seus focos pessoais. Assim se o foco do especialista estiver voltado para uma abordagem probabilística o AutoClass apresentará um melhor desempenho. Se o foco for voltados a hierarquias o WebIHCA apresentará melhor desempenho.