Universidade Federal de Pernambuco

Centro de Informática

Disciplina: Aprendizagem de máquina

Professor: Ricardo Prudêncio

Grupo:

Adônis Tavares

Nivan Roberto

Vinícius Cantarelli

Descrição do projeto

O projeto consistirá na implementação da técnica de *clustering hierárquico* proposta nos artigo Uma Experiencia em Mineração de Textos Utilizando Clustering Probabilıstico e Clustering Hierarquico.

Esse artigo aborda a mineração de textos que é um processo de conhecer de que se trata o texto através da freqüência de suas palavras.

Desejamos realizar a tarefa de categorização automática de textos. Esta consiste em, dado um conjunto de documentos *D={d1,d2,...,dn}* e um conjunto de categorias C={c1,c*2*,...,c*m*} decidir se um documento d*i* pertence a categoria c*j*. Este processo é formado por 4 etapas principais: a coleta de documentos, pré-processamento desses documentos, extração de conhecimento e avaliação e interpretação dos resultados.

Na etapa de pré-processamento, serão realizadas atividades de pré-processamento, a fim de reduzir a dimensionalidade dos dados a serem analisados e de construir uma representação do documento que possa ser utilizada pelo algoritmo de aprendizagem que será aplicado posteriormente. Serão aplicadas técnicas como o algoritmo *stemming* e a lei de Zipf. Na etapa de extração de conhecimento, usaremos a representação obtida no passo anterior para aplicar o algoritmo WebIHCA (*Web Incremental Hierarchical Clustering Algorithm*), este é o algoritmo utilizado no artigo, e outros algoritmos de *clustering hierárquico*, a fim de comparar os desempenhos dos mesmos.