Programa Institucional de Bolsas de Iniciação Científica
24 a 26 outubro de 2012
Pró-Reitoria de Pesquisa - Pibic/CNPq
Pró-Reitoria de Graduação - SAE/ Unicamp
XX congresso interno de iniciação científica da unicamp 2012

E0541

AVALIAÇÃO DE PESOS NA REPRESENTAÇÃO DE TEXTOS APLICADO À CATEGORIZAÇÃO DE TEXTOS

Diego Peterlevitz Frota (Bolsista PIBIC/CNPq) e Prof. Dr. Jacques Wainer (Orientador), Instituto de Computação - IC, UNICAMP

O rápido crescimento de informação online criou a necessidade de técnicas para organizar documentos de texto. Dentre tais técnicas, podemos citar a categorização de textos, que é a tarefa de automaticamente atribuir documentos não-rotulados a categorias pré-definidas. Cada documento pode ter múltiplas, apenas uma, ou nenhuma categoria. Uma parte vital para realizar tal categorização é definir uma representação do texto no computador, permitindo assim que os documentos possam ser classificados por um classificador. O modelo mais utilizado para isso é o Espaço Vetorial, que representa um documento como um vetor $d = (w_1, \ldots, w_k)$, onde k é o tamanho do conjunto de atributos pré-definidos, e w_i é o peso que representa quanto o i-ésimo termo contribui na semântica do documento. Usualmente cada atributo é relacionado com uma palavra pré-definida por um conjunto enumerado de termos (palavras) denominado *dicionário*. O objeto central do projeto é verificar a qualidade da categorização especificamente quando mudamos a atribuição dos pesos.

Categorização de textos - Inteligência artificial - Data mining