

243

**CATEGORIZAÇÃO AUTOMÁTICA DE DOCUMENTOS DE TEXTO UTILIZANDO LÓGICA DIFUSA.** *Thais Silva Galho, Silvia Maria Wanderely Moraes (orient.)* (Informática, IA, ULBRA).

Com o advento dos computadores, muita informação, hoje, é armazenada e disponibilizada em meio magnético na forma de documentos-texto. Constantemente, tais informações precisam ser recuperadas, entretanto, à medida que o volume de documentos cresce mais dispendioso computacionalmente torna-se o processo de busca e recuperação dessas informações. Devido a isso, a recuperação de informações é um tema muito pesquisado. Embora o homem ainda seja o melhor classificador, a quantidade de documentos torna a classificação manual impraticável, deixando evidente a necessidade do uso do computador. Para viabilizar tal tarefa, no Processamento de Linguagem Natural, tem-se usado a técnica de Categorização Automática de Documentos de Texto (CADT). Segundo Rizzi et al (2000), “a categorização de textos é uma técnica utilizada para classificar um conjunto de documentos em uma ou mais categorias existentes. Ela é geralmente utilizada para classificar mensagens, notícias, resumos e publicações. A categorização também pode ser utilizada para organizar e filtrar informações. Essa capacidade faz com que esta técnica possa ser aplicada em empresas, contribuindo no processo de coleta, análise e distribuição de informações e, conseqüentemente, na gestão e na estratégia competitiva de uma empresa.” Para se construir um sistema de CADT são utilizados desde métodos puramente estatísticos a técnicas de Inteligência Artificial como Redes Neurais e Lógica Difusa. O trabalho em desenvolvimento objetiva estudar o processo de CADT, analisando os diferentes métodos utilizados para isto, suas vantagens e aplicações e, ainda, projetar e desenvolver um protótipo na forma de um estudo de caso que valide a pesquisa realizada. Atualmente, o trabalho está na fase de desenvolvimento do protótipo para categorização automática de textos. Escolheu-se uma coleção de aproximadamente trezentos documentos, em português, extraídos de jornais e revistas, para treinamento e testes. A definição automática das categorias está sendo implementada através da técnica *Score de Relevância* como em Rizzi (2000a), e a classificação de novos documentos através da similaridade baseada em funções difusas apresentada em Wives (1999).