5
Algoritmos de Geração de Categorias

Algoritmos de geração de categorias têm por objetivo gerar, da forma mais automatizada possível, categorias em que documentos possam ser encaixados segundo algum critério. Para isso, é necessário que seja fornecido inicialmente um conjunto de "documentos teste", que servirá de amostra para a geração. Esses documentos devem ser escolhidos de tal maneira a representar da melhor forma possível a natureza do conjunto total de documentos cujas categorias devem ser capazes de abranger. O algoritmo trabalha procurando similaridades entre os documentos, que podem ser de diversas naturezas, desde as mais simples, como freqüência da ocorrência de palavras ou correlação entre palavras, até as mais complexas, como análise de semântica de frases ou análise de propriedades específicas dos documentos, como as meta-informações presentes nos documentos em formato PDF e PS (autor, assunto, palavras-chave, etc.), ou os *hyperlinks* no caso de HTML. Essas similaridades são utilizadas para a geração de regras. Uma categoria é vinculada a um conjunto de regras que documentos devem satisfazer para serem considerados pertencentes a ela. Na maioria dos casos, é altamente desejável uma intervenção humana após a execução de um algoritmo dessa natureza, a fim de refinar os resultados.

A literatura está repleta de algoritmos com esse propósito, com todo o tipo de precisão e complexidade. Alguns deles são mais direcionados a determinados tipos de documentos.

Em [43] e [44] são apresentados algoritmos especialmente desenvolvidos para a geração de categorias de documentos que possuem *hyperlinks*, como documentos HTML, que levam em conta não somente o conteúdo dos documentos (*categorization by content*) mas também o conteúdo dos seus *hyperlinks* (*categorization by context*).

Em [45] e [46], são apresentados algoritmos baseados em hierarquias de categorias, que podem ser vistas como ontologias.

Em [47] são apresentados mais dois algoritmos para geração de categorias baseados em técnicas de *data mining*. Em [48] é apresentado um algoritmo cujo tempo de processamento varia linearmente com o número de documentos processados, sendo portanto indicado para a geração de categorias de grandes coleções.