INDEXAÇÃO AUTOMÁTICA E INFOMETRIA COMO FERRAMENTAS NO ESTUDO DA LINGUAGEM NATURAL

Jaime Robredo
Universidade de Brasília

A indexação automática de textos[1] conheceu um importante desenvolvimento a partir do momento em que começou a ganhar espaço o uso da linguagem natural, tanto no processo de indexação dos documentos como no de busca e recuperação da informação.

Os processos de indexação automática são hoje uma realidade que permite acelerar consideravelmente a entrada dos registros bibliográficos e seus resumos nas bases de dados referenciais. Como subproduto da indexação automática, é fácil gerar também, automaticamente, um dicionário de termos significativos, com suas respectivas frequências. Abre-se assim caminho a importantes aplicações dentre as quais cabe destacar a criação de listas de associações binárias entre termos de frequência elevada com outros termos e indicação do número de co-ocorrências. A utilização das listagens de frequências dos termos significativos e das ocorrências das associações binárias entre eles - para estabelecer redes ou identificar conjuntos de associações mais prováveis *(clusters)*, utilizando-se de abordagens infométricas bem estabelecidas -, abre uma ampla frente de pesquisa no campo da linguagem natural e mais particularmente no desenvolvimento de vocabulários ou dicionários especializados, os quais, num processo de iteração associado à indexação automática, se constituem em elementos cada vez mais indispensáveis para aliar a rapidez, a qualidade e a facilidade aos processos de indexação e de recuperação no mundo da documentação virtual, que tanto estão faltando no momento atual. O sistema InfoDoc®, idealizado pelo autor e desenvolvido com auxílio parcial do CNPq, utiliza o princípio de filtragem de termos não significativos, com importantes aprimoramentos. A associação do InfoDoc® a outros programas, também desenvolvidos pelo autor para facilitar diversos cálculos infométricos, fazem desse sistema uma poderosa ferramenta para o estudo de diversos aspectos da linguagem natural.