Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais

SOUZA, Renato Rocha. *Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais*. 2005. 214f. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação da UFMG, Belo Horizonte.

Desde que se tornaram inviáveis em alguns contextos os processos manuais de indexação de documentos, buscam-se alternativas eficazes que possibilitem a representação automática dos assuntos principais desses documentos. Os processos mais comuns de indexação automática descrevem os documentos através de uma lógica simplista advinda da análise de freqüência das palavras que neles ocorrem. Buscando propor processo de indexação mais eficaz, que analise as palavras e expressões no âmbito de seus contextos lingüísticos, três pressupostos são definidos: a) a utilização de sintagmas nominais como descritores apresenta vantagens em relação ao uso de palavras-chave; b) a extração de sintagmas nominais de textos de documentos digitalizados é possível e viável com ferramentas tecnológicas atualmente disponíveis e c) é possível estabelecer processo automatizado e eficaz para escolha de descritores significativos para documentos digitalizados, utilizando sintagmas nominais. O objetivo da pesquisa é apresentar uma metodologia para viabilizar o processo de atribuição de descritores a textos digitalizados – indexação – através da extração de sintagmas nominais e da análise de fatores como a freqüência de ocorrência desses sintagmas nominais nos textos dos documentos, no conjunto dos documentos; a estrutura dos sintagmas nominais; o nível dos sintagmas nominais e a ocorrência desses em tesauro de um campo de conhecimento específico. Para atingir esse objetivo são analisados (a) um corpus de 15 documentos dos quais foram extraídos os sintagmas nominais manualmente, para testar o processo de extração automática e (b) um corpus de 60 documentos provenientes de publicações eletrônicas da área de ciência da informação. A metodologia proposta foi aplicada inicialmente a parte do corpus para validação e parametrização das variáveis do algoritmo, e, então, novamente aplicada, com alterações, à totalidade do corpus. Os resultados apresentados demonstraram grande pertinência dos descritores atribuídos aos documentos e permitiram concluir que a metodologia obtém sucesso inequívoco nas condições estudadas.