RBBD

ARTIGO

UM MODELO ALGÉBRICO PARA REPRESENTAÇÃO, INDEXAÇÃO E CLASSIFICAÇÃO AUTOMÁTICA DE DOCUMENTOS DIGITAIS *

Elias Oliveira

Patrick Marques Ciarelli

Marcos Hercules Santos

Bruno Oliveira da Costa

Resumo

Apresenta-se idéia da representação, indexação e classificação automática de documentos digitais. A representação de documentos via o modelo vetorial é simples e permite-nos lidar com classificação de uma grande quantidade de documentos os quais estão sendo carregados diariamente nas quase 35 bibliotecas digitais de tese e dissertação no Brasil. A expectativa é de termos outras 20 bibliotecas a mais na lista para o fim deste ano. Comparou-se a metodologia de classificação automática descrita nesse artigo, usando uma amostra de documentos reais, com aquela feita pelo especialista humano. Os resultados mostram que esta metodologia é promissora em se reduzir o esforço dos especialistas na realização dessa tarefa.

Palavras-chave: Indexação automática. Classificação automática. Inteligência Artificial. Modelos Estatísticos.

AN ALGEBRAIC MODEL OF REPRESENTATION, INDEXATION AND AUTOMATIC CLASSIFICATION OF DIGITAL DOCUMENTS

Abstract

In this paper we introduce the idea of representing, indexing and automatically classifying digital documents. The vectorial model of representing documents is simple and allows us to deal with the classification of a great amount of digital documents which were loaded daily in almost 35 Brazilian Digital Library of Thesis and Dissertation. We expect to have another 20 libraries by the end of this year. Using a sample of real documents, we compare this methodology of classification to that done by specialists. The results show that this methodology is promising in reducing the effort of specialists when performing such task.

KEYWORDS: Automatic Indexation. Automatic classification. Artificial Intelligence. Statistical models.