Tese de Doutorado
Categorização Automática de Documentos Médicos

Luciano Romero Soares de Lima
Doutor em Ciência da Computação da Universidade Federal de Minas Gerais
Belo Horizonte
18 /07/2000

PALAVRAS-CHAVE

Base de documentos médicos - Categorização automática - Codificação automática - Ferramenta interativa Web - Grafo de assinalamento - Hierarquia de termos - HiMeD - Informática médica - MedCode - Modelo hierárquico para categorização de documentos médicos - Modelo vetorial de recuperação de informação - Pesquisa aproximada em texto - Recuperação de informação - Vocabulário controlado

RESUMO

Este trabalho tem como objetivo principal propor um modelo para categorização de documentos médicos, chamado *HiMeD*. O modelo é baseado no princípio que denominamos *correlação hierárquica de termos especializados*, no qual um conceito médico utilizado num processo de categorização pode ser representado por termos vinculados hierarquicamente entre si, e esta vinculação pode conter um conjunto de componentes que permite a determinação dessa categorização ordenada pelo grau de relevância do conceito adotado. O uso desse princípio nos permite isolar a tarefa de categorização da influência desnecessária de termos não pertencentes ao vocabulário médico controlado de referência e da linearidade do cálculo do peso de um termo na recuperação de informação proporcionada pelos modelos tradicionais. Os conceitos aqui desenvolvidos foram utilizados em vários experimentos de categorização automática de documentos médicos que comprovaram a qualidade do modelo proposto, sendo os experimentos realizados uma contribuição relevante adicional do trabalho. Finalizando, foi implementada uma ferramenta para

[1] E-mail: luciano@bhz.sarah.br

codificação automática de documentos médicos baseado nos componentes do nosso modelo, que demonstrou a sua capacidade tecnológica em construir ferramentas de apoio à categorização de informação médica. A ferramenta, chamada MedCode, foi utilizada em experimentos realizados com a participação de especialistas em codificação médica e sua utilização melhorou a precisão da codificação automática de documentos médicos. Essa melhoria ocorreu em grande parte devido às características visuais e interativas que a ferramenta possui, que permitiu aos especialistas alterar o ambiente de categorização, o tipo de algoritmo de processamento a realizar e outras opções de processamento dos documentos a categorizar.

ABSTRACT

The main objective of this thesis is to propose a categorizing model for medical documents, called HiMeD. *The model is based on the principle that we denominated* hierarchical correlation of specialized terms*, in which a medical concept, to be used in an automatic categorization process, can always be represented by terms, where these terms are linked up in a hierarchical path. This hierarchical linking can contain components that allow the determination of these categories ordered by the degree of relevance of the adopted concept. The use of this principle allows us to isolate the categorization tasks from the unnecessary influence of terms not belonging to the medical vocabulary of reference and of the straight calculation of the term-weight in the information retrieval process used by the classic models. The concepts developed here were used in several experiments that demonstrated the quality of the proposed model. These experiments are another important contribution of this work. Finally, a tool for automatic coding of medical documents was implemented based on the components of our model, thus demonstrating its technological capacity in building automatic categorization tools. This tool, called MedCode, was used in experiments carried out with the help of medical coding specialists, and its use improved the precision of the automatic coding of medical documents. This improvement is largely due to the interactive and visual characteristics of the prototype, which allowed the specialists to modify the coding environment, to select the type of processing algorithm, and to modify other document processing options.*

KEYWORDS

Approximate string matching - Assignment graph - Automatic - Categorization - Automatic coding - Controlled vocabulary - Hierarchical model for categorization of medical documents - Hierarchical terms - Information retrieval - MedCode - HiMeD model - Medical document databases - Medical informatics - Vector space model