**Author:** Luciano Romero Soares de Lima

**Institution:** Department of Computer Science
Federal University of Minas Gerais – Brazil

**Research Area:** Information Retrieval and Medical Informatics

**Thesis Presentation Date:** 2000/06/18

**Orientation:** Dr. Alberto Henrique Frade Laender and Berthier Ribeiro-Neto

**Thesis Title:** Automatic Categorization of Medical Documents

# Abstract

The main objective of this thesis is to propose a categorizing model for medical documents. 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