Informática Pública vol. 5 (1): 133-134, 2003

## Tese de Doutorado

# Recuperação Vertical de Informação: um estudo de caso na área jurídica

**MARIA DE LOURDES DA SILVEIRA**[1]

***PALAVRAS-CHAVE***

*Recuperação vertical de Informação – Tesauros – Redes Bayesianas – Área jurídica.*

***RESUMO***

Os sistemas de Recuperação de Informação (RI) foram criados para facilitar o acesso à informação em bibliotecas digitais modernas. Esses sistemas permitem organizar, indexar e recuperar informação sobre os documentos de uma coleção. Nesse processo, a tarefa de indexação tem importância central. Para indexar os documentos é necessário identificar termos de indexação que reflitam o conteúdo semântico dos documentos e que possam ser lembrados pelos usuários no momento de efetuar as consultas. Para normatizar o processo de seleção de termos de indexação, foram criados tesauros. Tesauros são ferramentas desenvolvidas tanto para controlar e padronizar a linguagem de indexação, quanto para facilitar a construção e reformulação de consultas pelos usuários. Tesauros são, em geral, específicos para uma dada área do saber, tais como a médica e a jurídica, dentre outras. Dado um tesauro para uma área específica, pergunta-se: é possível construir um sistema de RI especializado para aquela área do saber que permita a obtenção de resultados de qualidade superior? O objetivo deste trabalho é produzir uma resposta para essa questão. Nossa proposta se baseia no projeto de um sistema de Recuperação Vertical de Informação que, utilizando-se do conhecimento codificado em um tesauro, facilite ao usuário encontrar informação de seu interesse. Para combinar informação evidencial proveniente da consulta do usuário (isto é, as palavras-chave especificadas pelo usuário) com informação evidencial proveniente do tesauro, adotamos o arcabouço teórico das redes Bayesianas. O resultado é um modelo de rede de crenças que leva a uma nova fórmula de ordenação dos documentos na resposta. Para validar nosso sistema, realizamos um estudo de caso voltado para a área jurídica. Utilizando uma coleção de referência composta por jurisprudências dos tribunais federais brasileiros e o Tesauro da Justiça Federal, avaliamos nosso sistema de recuperação vertical de informação. Nossos resultados indicam que a combinação de informação proveniente das consultas com informação proveniente do tesauro possibilita ganhos médios em precisão da ordem de 31%.

***KEYWORDS***

*Vertical Information Retrieval – Thesaurus – Bayesian Networks – Juridical Area*

[1] E-mail: dilu@pbh.gov.br

***ABSTRACT***

*Information Retrieval (IR) systems were created to facilitate the access to information in modern digital libraries. These systems organize and index the documents of the collection, so that information about them can be retrieved. In this process, the indexing task is of central importance. To index the documents, it is necessary to identify indexing terms that reflect the semantic content of the documents and that are likely to be remembered by the users at querying time. To standardize the process of selecting the index terms, thesauri were created. A thesaurus is a tool that allows standardizing the indexing language and facilitating the construction and reformulation of queries by the users. Thesauri are, in general, specific to a given knowledge area, such as the medical and juridical areas, among others. Given a thesaurus to a specific area, one can ask: is it possible to build an IR system, specialized to that area of knowledge, that will generate higher quality results? The objective of this work is to produce a response to this question. Our proposal is based on the design of a Vertical Information Retrieval System that, using the knowledge encoded in a thesaurus, facilitates the task of finding information of interest. In order to combine evidential information in the user query (i.e., the keywords specified by the user) with evidential information in the thesaurus, we adopt the theoretical framework of Bayesian networks. The result is a belief network model that yields a new equation for ranking the documents in the response set. To validate our system, we carried out a case study in the juridical area. Using a reference collection composed of jurisprudences of Brazilian federal courts and the Federal Justice Thesaurus, we evaluated our vertical information retrieval system. Our results indicate that the combination of information from the user queries with information from the thesaurus leads to gains in average precision of the order of 31%.*

***SOBRE A AUTORA***

**MARIA DE LOURDES DA SILVEIRA**[1]
*Doutora em Ciência da Computação*
*Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais*
*Data de defesa: 10/04/2003*
*Local: ICEX/UFMG*