Indexação automática de artigos científicos da área da saúde

1

Abstract. *This article describes the development of an application capable of analyzing a scientific text and synthesize it in keywords that represent the concept forwarded by text. To this, were constructed and used mechanisms based on natural language processing. These mechanisms are capable of filtering, inflect words and identify them as potential keywords. The intention is that implementation can be used to help professionals engaged in this activity in large centers of information and documentation.*

Resumo. *Este artigo descreve o desenvolvimento de uma aplicação capaz de analisar um texto científico e sintetizá-lo em palavras-chave que representam o conceito transmitido pelo texto. Para isto, foram construídos e utilizados mecanismos com base em processamento de linguagem natural. Estes mecanismos são capazes de filtrar, inflexionar palavras e identificá-las como possíveis palavras-chave. A intenção é que a aplicação possa ser utilizada no auxílio de profissionais que exercem esta atividade em grandes centros de informação e documentação.*

1. Introdução

Em bibliotecas físicas e digitais, a organização da informação é baseada em descritores, que são palavras-chave cujos conceitos podem representar de forma resumida o documento onde estão contidas. Os descritores de um texto podem ser escolhidos livremente, de acordo com que o autor ou profissional de indexação imagina ser mais adequada ao documento. Outra forma é recorrer a um conjunto de descritores pré-definidos presentes em um vocabulário controlado, seguindo um conjunto de regras [LANCASTER 2004].

O vocabulário controlado contém um conjunto de descritores que são agrupados de forma hierárquica. Todo descritor é armazenado de forma que consigo estejam presentes a descrição de seu conceito, termos sinônimos e comentários deixados para ajudar o indexador em sua tarefa. Com uso do vocabulário controlado é possível organizar de forma eficiente uma coleção de documentos, pois além do conceito do descritor que sintetiza o assunto do documento, a hierarquia do vocabulário classifica este documento em tópicos mais genéricos [LANCASTER 2004].

O processo de indexação é feito por um profissional da informação, comumente chamado de indexador, que seguindo um conjunto de regras e critérios constrói representações de documentos para que sejam incluídos em uma base de dados. Este processo é uma atividade complexa, lenta e custosa, geralmente realizada por poucas pessoas [LANCASTER 2004].

O objetivo deste trabalho reside na construção de uma aplicação capaz de auxiliar e melhorar a qualidade do trabalho desse profissional no reconhecimento de descritores

para artigos científicos da área da saúde. A aplicação construída deverá apresentar palavras candidatas a descritores para um artigo, o que pode aumentar a quantidade de artigos indexados em um determinado intervalo de tempo, fazendo com que o indexador complete seu trabalho de forma mais eficiente e menos custosa.

Este artigo está estruturado da seguinte maneira: na seção 2 é descrita uma breve fundamentação teórica sobre o tema e sobre os dados utilizados neste trabalho; na seção 3 é descrita a implementação do extrator de descritores; na seção 4 são apresentados os resultados obtidos e, por fim, a seção 5 contém as considerações finais deste trabalho.

2. Fundamentação teórica

Um vocabulário controlado é uma ferramenta que agrupa conceitos e termos de forma hierárquica que representam áreas temáticas. Os termos existentes no vocabulário controlado são usados, combinados ou não, para descrever de forma sucinta um documento. O vocabulário controlado utilizado neste trabalho foi o MeSH (Medical Subjects Heading) [NLM 2008], distribuído pela NLM (National Library of Medicine).

O MeSH pode ser encontrado em arquivos de vários padrões, mas para este trabalho foi usado o arquivo XML (Extensible Markup Language). Este trabalho analisa no MeSH três tipos de elementos para realizar o processo de indexação: descritores, conceito e termos. Os descritores são representados por termos, cuja função é expor de forma objetiva os conceitos existentes no vocabulário controlado. Descritores são representados pelo elemento descriptor no MeSH, que aninha uma lista de conceitos[NLM 2006].

Os conceitos, representados no arquivo pelo elemento *concept*, guardam um termo e uma definição conceitual para este termo. Um conceito é preferido quando seu uso é indicado para o descritor ao qual está associado. O conceito preferido guarda um termo idêntico ao termo do descritor a que pertence. Já os conceitos não preferidos têm termos diferentes do termo do descritor, mas são considerados sinônimos do conceito preferido [NLM 2006].

Para e extração de descritores de documentos eletrônicos é necessário que sejam encontradas palavras cujo conceito, além de descrever o conteúdo de texto também estejam presentes no vocabulário controlado utilizado para a indexação.

É necessário que haja um tratamento dos textos de entrada e do vocabulário controlado para que as buscas por descritores nos textos aconteçam da forma mais eficaz possível. Este tratamento implica em reduzir as palavras aos seus radicais, através de um algoritmo de *stemming* [Porter 1997], e retirar do texto palavras que não tem função conceitual, através de uma lista de *stop words* [NLTK 2008].

3. Desenvolvimento

No desenvolvimento deste trabalho, o vocabulário controlado usado foi o MeSH do ano 2008, que conta com 24.767 termos em inglês e mais de 97.000 formas sinônimas para os termos autorizados [NLM 2006]. A lista de *stop-words* utilizada é a lista proposta pelo projeto NLTK [NLTK 2008], composta por 571 termos. Os artigos científicos utilizados são da área da saúde e disponibilizados na biblioteca eletrônica SciELO [SciELO 2008].