Indexação automática de artigos científicos da área da saúde

Fabio Montefuscolo
Faculdade de Ciências Exatas e Tecnologia
Centro Universitário Senac
São Paulo/SP, Brasil
fabio.montefuscolo@gmail.com

Rafael Gomes Câmara
Faculdade de Ciências Exatas e Tecnologia
Centro Universitário Senac
São Paulo/SP, Brasil
rafael.gcamara@gmail.com

Fabrício J. Barth
Faculdade de Ciências Exatas e Tecnologia
Centro Universitário Senac
São Paulo/SP, Brasil
fabricio.jbarth@sp.senac.br

Orlando Rodrigues Jr.
Faculdade de Ciências Exatas e Tecnologia
Centro Universitário Senac
São Paulo/SP, Brasil
orlando.rodrigues@sp.senac.br

RESUMO

Este artigo descreve o desenvolvimento de uma aplicação capaz de analisar um texto científico da área da saúde e sintetizá-lo em palavras-chave, armazenadas em um vocabulário controlado, que representam o conceito transmitido pelo texto. Para isto, foram construídos e utilizados mecanismos capazes de filtrar, inflexionar palavras e identificá-las como possíveis palavras-chave. A intenção é que a aplicação possa ser utilizadano 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 o mais adequado ao documento. Outra forma é recorrer a um conjunto de descritores pré-definidos presentes em um vocabulário controlado, seguindo um conjunto de regras [2].

O vocabulário controlado contém um conjunto de descritores que são agrupados de forma hierárquica. Todo descritor é armazenado de forma que carreguem a descrição do conceito que representa, termos sinônimos e comentários 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 [2].

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 [2].

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 apresenta palavras candidatas a descritores para um artigo, podendo 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, de acordo com 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*) [5], 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*). São analisados três

WEBMEDIA '2009 Fortaleza, Ceará Brasil

tipos de elementos no MeSH para realizar o processo de indexação: descritores, conceitos 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[4].

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 [4].

Para a 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 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* [7], e retirar do texto palavras que não tem função conceitual, através de uma lista de *stop words* [6].

3. DESENVOLVIMENTO DO EXTRATOR

No desenvolvimento deste trabalho, o vocabulário controlado usado foi o MeSH do ano 2008, contendo 24.767 termos em inglês e mais de 97.000 formas sinônimas para os termos autorizados [4]. A lista de *stop-words* utilizada foi a proposta para projeto NLTK [6], composta por 571 termos. Os artigos científicos utilizados são da área da saúde e disponibilizados na biblioteca eletrônica SciELO [10].

Foram extraídos do vocabulário controlado todos os termos dos descritores e dos sinônimos. Os termos retirados dos descritores, que são os termos preferidos para descrever o conceito, são armazenados em um dicionário de descritores. Os outros termos, que são sinônimos ao termo preferido são armazenados em um dicionário de sinônimos. Todo termo presente no dicionário de sinônimos tem uma referência para um termo do dicionário de descritores. Um descritor pode ter vários sinônimos, mas um sinônimo está relacionado somente a um descritor.

Os descritores e sinônimos presentes no vocabulário controlado foram associados a identificadores e armazenados em estruturas do tipo *hashmap*, de forma que fosse possível mapear a relação hierárquica entre os termos do vocabulário controlado.

A aplicação desenvolvida é composta por três módulos: um para leitura do MeSH, outro para leitura dos artigos e outro para validação dos resultados. O objetivo do módulo para leitura do MeSH é de transformá-lo em uma estrutura que será utilizada pelo módulo para leitura dos artigos. O objetivo do módulo para leitura dos artigos é identificar os descritores para cada artigo fornecido. O módulo para validação dos resultados tem como objetivo verificar a eficiência da aplicação desenvolvida neste trabalho.

Os componentes utilizados nos módulos para leitura do MeSH e no módulo para a leitura de artigos são:

- **Alocação do MeSH na Memória**: a leitura do XML do MeSH é feita com a API SAX (*Simple API for XML*) [9], visando poupar recursos computacionais. A medida que um termo é lido pela API SAX, métodos para o armazenamento são executados. Caso o termo encontrado seja um descritor preferido, sua forma normalizada se torna a chave em um dicionário e o termo completo é o valor associado a tal chave. Caso o termo seja um sinônimo, outra estrutura é usada, onde a chave é a forma normalizada do sinônimo, e o valor é a forma normalizada do descritor correspondente. Após formadas as estruturas, elas são serializadas e gravadas em disco através da *API Pickle* [8], para que não seja necessário repetir o procedimento de leitura do XML.

- **Normalização de termos**: duas classes foram criadas para fazerem uso de uma lista de *stop-words* e efetuar a normalização dos termos. A primeira é uma implementação em Python do algoritmo de *Stemming* de Martin Porter [1], reponsável pela inflexão de palavras na aplicação. A outra classe identifica as *stop-words*, cruzando cada palavra da entrada com a lista obtida do projeto NLTK.

- **Navegação no MeSH**: para obtenção de novos descritores foi implementado um algoritmo capaz de navegar pela estrutura do MeSH utilizando a identificação hierárquica de cada descritor. Com isso, é possível contar todos os ascendentes de cada descritor. Após esta etapa, é avaliada a relevância de cada ascendente para o texto. Esta relevância é determinada de acordo com o número de filhos do descritor que estão presentes no texto.

- **Janelamento**: existem duas ocasiões em que é empregada esta técnica: a primeira, com tamanho parametrizável, acontece na comparação entre as palavras do artigo e do vocabulário controlado, por existem descritores formados por mais de uma palavra; a segunda ocasião acontece em outra técnica, que é a análise de bigramas.

- **Bigrama**: durante a aplicação do janelamento, foram efetuadas combinações entre a primeira palavra da janela com as demais palavras. Cada uma das combinações encontradas é pesquisada no vocabulário controlado, exceto a primeira, que já aconteceu na aplicação do janelamento sem aplicão dos bigramas.

- **Corte de descritores**: a capacidade de cruzar todas as palavras de um texto com as do vocabulário controlado acarreta no surgimento de descritores irrelevantes para um determinado artigo, prejudicando a qualidade do resultado. Para contornar isto, foram elaboradas duas técnicas para eliminar descritores irrelevantes: a de corte por ocorrência, que elimina o decritores cujo número de aparições seja menor que 20% da quantidade de aparições do descritor mais encontrado e a de corte por quantidade, que considera somente os 10% dos descritores com maior resultado.

Para a validação dos descritores obtidos pelo programa foram utilizados artigos indexados manualmente por um profissional da área. Com isso, foram armazenados artigos e descritores. Para tornar a validação dos resultados coerente, é necessário verificar se os descritores pré-definidos para o artigo realmente estão no texto. Este procedimento é necessário, pois é comum que o profissional utilize descritores que não estão no texto, mas que têm seu uso apropriado. Como a técnica utilizada é apenas a definição de termos encontrados no texto, são comparados apenas os resultados apenas com descritores pré-definidos que realmente aparecem no artigo. O componente corte de descritores pré-definidos que não são existem no texto, verifica se um descritor ou seus sinônimos possuem alguma ocorrência. Quando não não há ocorrências o descritor é ignorado para o processo de avaliação dos resultados obtidos.

4. RESULTADOS

As abordagens propostas e desenvolvidas foram avaliadas a partir dos valores de cobertura, precisão e F-Measure [3] calculados com a execução do protótipo sobre uma coleção de artigos completos e indexados extraídos da SciELO [10].

A avaliação consistiu em coletar descritores de 419 artigos e armazená-los em um arquivo texto, sobre o qual foi foi calculada a média da quantidade de descritores que cada artigo possui, que é de 11,28 descritores com um desvio padrão de 4,57.

O programa teve diferentes técnicas aplicadas, cada uma podendo ser executada de forma independente. Para a execução do programa, foram feitas combinações das técnicas sobre os 419 artigos, visando obter a melhor combinação possível usando as medidas apresentadas anteriormente. Também foram medidos nestes testes a eficiência dos diversos tamanhos de janela de leitura e os tipos de cortes usados. Primeiro foi executado o processo utilizando um tamanho de 3 palavras para a janela de leitura com técnicas que não dependem de nenhum tipo de corte.

Na tabela 1 é possível visualizar os resultados de precisão, cobertura e F-measure encontrados utilizando: apenas o identificador de descritores no artigo; o identificador de descritos junto a navegação na árvore do MeSH; e o identificador de descritos junto a navegação na árvore do MeSH e filtrando os descritores pré-definidos.

Tabela 1: Técnicas sem corte e seus resultados

Tecnicas	Precisão	Cobertura	F-Measure
Existência dos descritores no Artigo	4,803%	48,809%	8,746%
Navegação na árvore MeSH	4,406%	51,153%	8,114%
Navegação na árvore MeSH+ Filtragem do descritores pré-definidos	4,264%	68,946%	8,031%

É possível observar que o programa conseguiu uma cobertura maior que 60% de descritores dos artigos. Entretanto, a não realização do corte faz com que um grande número de descritores seja retornado, diminuindo então a precisão do programa. A utilização da navegação na estrutura do MeSH e a filtragem dos descritores são técnicas que podem aumentar a cobertura. No caso da filtragem, isso acontece por não comparar descritores que não estejam presentes no texto.

O próximo passo foi avaliar qual o tipo de corte seria mais eficiente. A navegação na estrutura do MeSH foi realizada antes e após o corte, para obter mais descritores e avaliar o melhor momento que ela deveria ser realizada no processo. Na massa de dados utilizada, a navegação após o corte acrescentou uma média de 3,31 descritores novos por texto. Na tabela 2 são apresentados os resultados desta avaliação.

Com os resultados apresentados na tabela 2 afirma-se que o corte por ocorrência é mais eficiente que o corte por quantidade. No entanto, é necessário verificar qual o impacto do uso de bigramas e no uso do componentes que eliminam descritores pré-definidos e não existentes no texto, chamado de filtragem. Na tabela 3 são apresentados os resultados gerados a partir da combinação destes dois componentes.

A tabela 3 revela que tanto o bigrama como o filtro aumentaram entre 7 e 8% a cobertura do programa, mas a precisão cai aproximadamente 0,8%. Entretanto, a medida harmônica F-measure aponta que o melhor resultado fica por conta da não utilização da navegação do MeSH.

Até o presente momento, a melhor abordagem para a obtenção de descritores foi a utilização do corte por ocorrência em conjunto com a técnica de bigramas. No entanto, ainda falta testar o tamanho da janela de leitura. O resultado obtido com janelas de tamanho 3 e 4 pode ser visto na tabela 4.

Tabela 4: Comparações entre janelas

Técnicas	Precisão	Cobertura	F-Measure
Corte por ocorrência+ filtragem+ Bigramas (Janela 3)	18,476%	28,522%	22,425%
Corte por ocorrência+ filtragem+ Bigramas (Janela 4)	18,631%	28,792%	22,623%

Os novos valores para a média harmônica representam uma sensível melhora nos resultados obtidos. Isso porque o número de combinações aumentam para a análise de bigramas, e a janela de leitura pode encontrar um descritor de até quatro palavras no texto. Números menores não são aconselhados por eliminar possíveis descritores existentes no vocabulário controlado, e uma janela maior pode distorcer os resultados do bigrama.

Com isso, é possível afirmar que a melhor configuração para o programa é a utilização de uma janela de leitura igual a quatro palavras, com as técnicas de corte por ocorrência e bigramas. O programa com esta configuração final retornou, para os artigos utilizados na validação, uma média de 35,19 descritores por artigo.

5. CONSIDERAÇÕES FINAIS

Este trabalho apresentou a implementação de um sistema capaz de identificar descritores para um artigo da área da saúde, respeitando um vocabulário controlado. Foram testadas inúmeras configurações para o sistema. A melhor configuração obteve uma precisão de 18,63% e uma cobertura de 28,79%.

Tabela 2: Comparação entre possíveis cortes em diferentes técnicas

Técnicas	Corte por ocorrência			Corte por quantidade		
	Precisão	Cobertura	F-Measure	Precisão	Cobertura	F-Measure
Apenas utilizando o corte	19,184%	20,226%	19,691%	17,993%	19,038%	18,501%
Corte antes da navegação	15,647%	20,532%	17,760%	15,042%	19,240%	16,884%
Navegação do MeSH antes do Corte	16,989%	20,514%	18,586%	15,839%	19,238%	17,374%

Tabela 3: Resultados com as técnicas de filtragem e bigramas

Técnicas	Corte por ocorrência			Corte por quantidade		
	Precisão	Cobertura	F-Measure	Precisão	Cobertura	F-Measure
Corte+Filtragem+Bigrama	18,476%	28,522%	22,425%	17,196%	26,752%	20,935%
Corte antes da navegação+Filtragem	14,989%	28,787%	19,714%	14,134%	27,034%	18,717%

Através dos resultados obtidos pela validação, constatou-se que a média de descritores sugeridos por artigo é alta. Isto reflete a capacidade do computador de cruzar todas as palavras do artigo com as existentes no vocabulário controlado, e assim, retornar um grande número de descritores. Isto seria difícil para um ser humano, pois seria necessária a memorização de todo o vocabulário controlado, ou então procurar uma a uma as palavras do texto no vocabulário controlado.

Também, constatou-se que a utilização da navegação na estrutura do MeSH para identificar descritores prejudicou a qualidade dos resultados. O baixo desempenho dos testes que utilizaram a navegação na estrutura do MeSH pode ser atribuído a dois fatores: (*i*) a heurística utilizada para definir novos descritores está imprecisa. Isto acarreta em trazer descritores diferentes dos escolhidos por um indexador manual, e diminuir a cobertura e a precisão da validação; e, (*ii*) os profissionais que definiram os descritores para os artigos avaliados preferem utilizar aqueles que são de assuntos mais específicos, evitando a utilização de um descritor mais genérico que consiga representar uma parte significativa destes assuntos.

Existem algumas alterações possíveis ao sistema para que este consiga obter resultados melhores:

- A criação de um módulo que utiliza técnicas de aprendizagem de máquina poderia personalizar o programa com as características de um único indexador manual, aprendendo com os artigos já indexados. Assim o resultado será aproximado ao de um profissional da área.

- Atribuir pesos para algumas seções do texto, como por exemplo, o título, resumo e conclusões, já que estes são trechos que provêem uma síntese do conteúdo tratado no documento. Os descritores encontrados nesta parte do texto teriam um peso maior que os descritores encontrados em outras partes do texto.

- Levar em consideração o contexto onde o sistema está inserido. Dependendo da biblioteca onde os artigos estão armazenados, podem ser escolhidos descritores mais abrangentes ou que tratam de um assunto mais específico. Por exemplo, um mesmo texto sobre uso de medicamentos na gravidez poderia ter um conjunto de descritores abrangentes em uma biblioteca universitária, usando termos mais prováveis de serem pesquisados por alunos. Mas se fosse uma biblioteca de algum laboratório, os descritores poderiam ser mais específicos. Os termos poderiam ser nomes de compostos químicos e suas reações no organismo de gestantes.

Com base nos resultados obtidos, a aplicação desenvolvida neste trabalho pode ser utilizada para a sugestão de palavras-chave aos profissionais que exercem a atividade de indexação em grandes centros de informação e documentação. Assim, um profissional com uma grande carga de trabalho pode reduzir o seu tempo de análise por artigo.

6. REFERÊNCIAS

[1] V. Gupta. *Python Implementation of Porter Stemming Algorithm*. Obtido em: http://tartarus.org/ martin/PorterStemmer/python.txt, Acessado em 5 de Março de 2009, 2 edition, 2008.

[2] F. W. LANCASTER. *Indexação e Resumos: Teoria e Prática*. Briquet de Lemos, Brasília, 2004.

[3] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.

[4] NLM. *Concept Structure in XML MeSH*. Obtido em http://www.nlm.nih.gov/mesh/concept_structure.html, Acessado em 5 de Março de 2009., 2006.

[5] NLM. *Medical Subject Headings*. Obtido em: ftp://nlmpubs.nlm.nih.gov/online/mesh/.xmlmesh/desc2008.gz, Acesso em 5 de Març de 2009., 2008.

[6] NLTK. *Natural Language Toolkit*. http://www.nltk.org, Acessado em 4 de Março de 2009, 2008.

[7] M. F. Porter. An algorithm for suffix stripping. *Readings in information retrieval*, pages 313–316, 1997.

[8] Python. *Python object serialization*. Obtido em: http://docs.python.org/library/pickle.html?highlight=pickle, Acessado em 4 de Março de 2009., 2008.

[9] Python. *Support for SAX2 Parsers*. Obtido em: http://docs.python.org/library/xml.sax.html, Acessado em 4 de Março de 2009, 2008.

[10] SciELO. *Scientific Eletronic Library*. http://www.scielo.org, Acessado em 4 de Março de 2009., 2008.