

CONSTRUÇÃO DE UMA ONTOLOGIA PARA O MAPEAMENTO DE LAUDOS MÉDICOS ARTIFICIAIS DE COLOSCOPIA PARA BASES DE DADOS ESTRUTURADAS

Leandro Borges dos Santos (PIBIC/Ações Afirmativas/CNPq/Unioeste), Huei Diana Lee, André Gustavo Maletzke, Carlos Andres Ferrero, Feng Chung Wu (Orientador), e-mail: wufengchung@gmail.com

Universidade Estadual do Oeste do Paraná/Centro de Engenharias e Ciências e Exatas/Campus de Foz do Iguaçu-PR

Ciências da Saúde-Medicina

Palavras-chave: análise inteligente de dados, ontologia médica, processamento de textos.

Resumo:

A tecnologia atualmente permite o armazenamento de grandes volumes de dados em distintos formatos, sendo que na área médica é comum que dados sejam representados por meio de laudos textuais semiestruturados ou não-estruturados. Neste trabalho é apresentada uma ontologia para o mapeamento de laudos artificiais de exames de Coloscopia, para bases de dados estruturadas. A avaliação da ontologia foi realizada juntamente com os especialistas, a qual permitiu uma melhor representatividade dos conceitos do domínio quando comparada com outra técnica.

Introdução

Processos computacionais têm sido propostos para auxiliar no armazenamento, gerenciamento e, principalmente, na análise de dados médicos. Nesse sentido, o processo de Mineração de Dados (MD) [1], tem sido amplamente aplicado com o intuito de extrair conhecimento relevante a partir de grandes conjuntos de dados.

Dados médicos, atualmente, são armazenados em diversos formatos como imagens, vídeos e laudos médicos textuais [2]. Esses laudos geralmente apresentam informações de maneira semiestruturada ou não-estruturada. Para a aplicação de processos computacionais, como o de MD, é necessário que esses dados estejam representados em um formato estruturado, por exemplo, por meio de uma tabela atributo-valor (TAV).

Nesse contexto, este trabalho tem como objetivo construir uma ontologia para auxiliar no mapeamento de laudos médicos artificiais de Coloscopia. Este trabalho constitui parte do projeto de Análise Inteligente de Dados

desenvolvido em parceria entre o Laboratório de Bioinformática da Universidade Estadual do Oeste do Paraná / Foz do Iguaçu, o Laboratório de Inteligência Artificial da Universidade de São Paulo / São Carlos e o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas / Campinas.

Materiais e métodos

O método de mapeamento de laudos médicos para bases de dados utilizado neste trabalho foi proposto em [3] e é executado em duas fases.

A primeira fase é dividida nas seguintes etapas:

- **Identificação de frases únicas:** é realizada a identificação de frases únicas existentes na coleção de laudos, obtendo o conjunto de frases únicas (CFU);
- **Construção de arquivo de padronização:** é construído o Arquivo de Padronização (AP), que contém palavras e expressões que serão substituídas por outras equivalentes;
- **Definição de stopwords:** é definido um conjunto de palavras, como preposições, artigos e conjunções, que quando removidas não alteram o sentido das frases;
- **Aplicação de stemming:** é efetuada a extração do radical das palavras contidas no CFU no intuito de identificar e retirar as frases redundantes;
- **Geração de n-gramas:** é realizada a identificação das unidades terminológicas de maior frequência nos laudos;
- **Construção do Dicionário do Conhecimento (DC):** é construído o DC com auxílio do AP, CFU e dos especialistas do domínio, que consiste em uma estrutura hierárquica de locais, características e subcaracterísticas do domínio;
- **Definição da tabela atributo-valor:** são definidos os atributos que irão compor a TAV e os possíveis valores de cada atributo.

Na segunda fase é construído o Algoritmo de Busca e Preenchimento que utiliza o DC e os laudos padronizados resultantes da primeira fase para o preenchimento da TAV, definida anteriormente.

Neste trabalho foi aplicada somente a primeira fase do método proposto em [3], para a construção do DC. Para isso, foram utilizados 100 laudos médicos referentes a exames de Coloscopia, gerados artificialmente seguindo protocolos utilizados pelo Serviço de Coloproctologia da Faculdade de

Ciências Médicas da Universidade de Campinas. Para construção do DC utilizou-se a linguagem de marcação *Extensible Markup Language*¹ (XML). Posteriormente, a partir dos mesmos laudos médicos foi realizada a construção de uma ontologia, utilizando o método proposto em [4]:

1. **Definição do escopo e objetivo:** são definidos o domínio e o grau de especificidade da ontologia, os requisitos devem ser atendidos, as ferramentas e a linguagem para representar o conhecimento;
2. **Construção:** é realizada a construção da ontologia com base nos conceitos obtidos na fase anterior e seus relacionamentos;
3. **Avaliação:** é avaliada a especificação inicial da ontologia, a sua expressividade e consistência. Caso algum requisito não seja satisfeito retorna-se à construção da ontologia, de maneira iterativa e interativa até que todos os requisitos sejam atendidos.

Para a construção da ontologia foi utilizado o *software* Protégé² e a linguagem *Web Ontologie Language*³ (OWL).

Após, foi realizada uma comparação do conhecimento representado entre o DC e a ontologia. Essa comparação foi efetuada mediante reuniões com especialistas da área médica, visando apontar itens como correteza das representações dos conceitos, interpretabilidade e reutilização.

Resultados e Discussão

Na Figura 1 é apresentada parte da ontologia de domínio desenvolvida utilizando o *software* Protégé, na qual *Thing* é a classe principal da ontologia. A classe *Atributo* contém todos os atributos que irão compor a TAV e a classe *Valor_de_Atributo* representa os possíveis valores desses atributos. A classe *Termo* contém as subclasses: *Regiao* que representa as regiões anatômicas presentes nos laudos e *Observacao* que representa *Caracteristica* e *Situacao* que contém dentre outras a subclasse *Anormalidade*.

A partir da comparação entre a ontologia e o DC observou-se que a ontologia representou com maior completude e consistência o conhecimento do domínio, pois possibilitou expressar relacionamentos entre os conceitos, ao contrário do DC que permitiu representar, unicamente, o conhecimento de maneira hierárquica.

¹ <http://www.w3.org/XML>

² <http://protege.stanford.edu>

³ <http://www.w3.org/TR/owl-features>

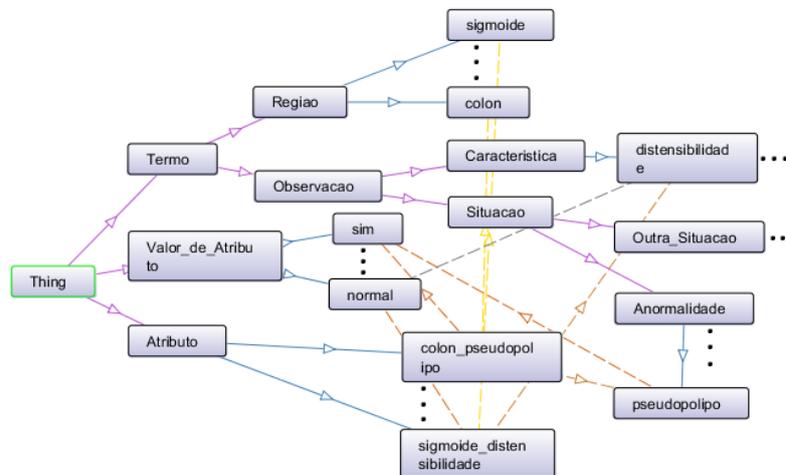


Figura 1 – Ontologia utilizando o Protégé.

Conclusões

Neste trabalho foi realizada a construção de uma ontologia a partir de laudos médicos artificiais de Coloscopia. De acordo com os especialistas a ontologia possibilitou expressar adequadamente o conhecimento do domínio e preservar o relacionamento entre os conceitos. Trabalhos futuros incluem o desenvolvimento do algoritmo de busca e preenchimento e a avaliação da ontologia no mapeamento de laudos médicos textuais para bases de dados.

Agradecimentos

Ao programa PIBIC-AF-CNPq.

Referência

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 1. ed. California, USA: Morgan Kaufmann, 2006.

SHORTLIFFE, E. H., CIMINO, J. J. **Biomedical informatics**. 3. ed. New York, USA: Springer, 2006.

HONORATO, D. F.; et al. Uma metodologia para auxiliar no processo de construção de bases de dados estruturados a partir de laudos médicos. In: ENIA, 5., 2005. **Anais do Congresso da Sociedade Brasileira de Computação**. São Leopoldo: ENIA, 2005. p. 593-601.

USCHOLD, M.; GRUNINGER, M. Ontologies: principles, methods and applications. **Knowledge Engineering Review**. UK, v.11, p.93-155, 1996.