CONSTRUÇÃO DE ONTOLOGIAS PARA O MAPEAMENTO DE INFORMAÇÕES CONTIDAS EM LAUDOS ARTIFICIAIS DE ENDOSCOPIA DIGESTIVA ALTA PARA BASES DE DADOS ESTRUTURADAS

Simone Aparecida Pinto Romero (PIBIC/UNIOESTE/PRPPG), Huei Diana Lee, André Gustavo Maletzke, Cláudio Saddy Rodrigues Coy, João José Fagundes, Wu Feng Chung (Orientador), e-mail: wufengchung@gmail.com

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

Palavras-chave: mineração de dados, mapeamento de laudos médicos, processamento de laudos.

Ciências da Saúde - Medicina

Resumo
Os métodos para a extração de conhecimento auxiliam na análise de grandes volumes de dados. Para tanto, é necessário que esses dados estejam representados e estruturados em um formato adequado. Na área médica, por exemplo, a aplicação de métodos para a extração de conhecimento pode auxiliar especialistas do domínio na tomada de decisão, assim como na identificação e no diagnóstico de doenças. Neste trabalho, é apresentada a construção de uma ontologia para a representação dos termos descritos em laudos textuais, a qual poderá, posteriormente, ser utilizada para o mapeamento de laudos médicos textuais para bases de dados estruturadas.

Introdução
A evolução da tecnologia tem facilitado a aquisição e o armazenamento de dados produzidos nas mais diversas áreas do conhecimento. Nesse sentido, diversos métodos que podem auxiliar na análise e na extração de conhecimento têm sido desenvolvidos, tal como o processo de Mineração de Dados [1]. Para que esses métodos possam ser aplicados, é necessário que os dados estejam estruturados em um formato adequado, como a representação atributo-valor [5].

Na área médica, os dados sobre pacientes são, comumente, representados por meio de laudos médicos textuais (LMt), os quais podem estar descritos de maneira semi-estruturada ou desestruturada. Desse modo, para que processos computacionais possam ser aplicados é necessário que os dados dos LMt sejam representados em um formato estruturado, possibilitando a análise dos dados sob diferentes perspectivas, podendo contribuir para a diminuição do tempo de análise e principalmente

na realização de análises mais completas quando comparados com ferramentas convencionais.

Este trabalho tem por objetivo a construção de uma ontologia para auxiliar no mapeamento de LMt artificiais de Endoscopias Digestivas Altas (EDA) para Bases de Dados (BD) estruturadas e faz parte do projeto Análise Inteligente de Dados, desenvolvido em parceria entre o Laboratório de Bioinformática (LABI) da Universidade Estadual do Oeste do Paraná (UNIOESTE)/Foz do Iguaçu, o Laboratório de Inteligência Computacional (LABIC) da Universidade de São Paulo (USP)/São Carlos, o Grupo Interdisciplinar em Mineração de Dados e Aplicações (GIMDA) da Universidade Federal do ABC (UFABC)/Santo André e o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas (UNICAMP)/Campinas.

Materiais e Métodos

Neste trabalho foram utilizados 609 LMt artificiais de EDA, confeccionados de acordo com os formulários utilizados no Serviço de Endoscopia Digestiva da Gastrocentro da Universidade Estadual de Campinas, referentes à porção anatômica duodeno.

O método para o mapeamento de LMt para BD estruturadas, proposto em [4], é composto por duas fases, sendo a 1º caracterizada pela construção do Dicionário do Conhecimento (DC) e a 2º, pela transformação dos LMt para a construção e preenchimento da Tabela Atributo-Valor (TAV).

A Fase 1 é constituída pelas seguintes etapas:

- **Identificação de frases únicas:** a partir de todas as frases do conjunto de LMt, são removidas as frases repetidas, gerando um Conjunto de Frases Únicas (CFU);
- **Definição da lista de *stopwords*:** definição das palavras a serem removidas do CFU, pois não são relevantes para o mapeamento;
- **Aplicação de *stemming*:** redução das palavras ao seu radical a fim de remover frases redundantes armazenadas no CFU;
- **Construção do Arquivo de Padronização (AP):** arquivo com palavras e expressões usadas para padronizar os termos do domínio;
- **Geração de *n*-gramas:** identificação das unidades terminológicas de maior frequência nos LMt;
- **Construção do DC e definição da TAV:** o DC é construído com base no CFU e no AP, no qual estão descritos hierarquicamente os termos dos LMt, que representam locais, características e subcaracterísticas do domínio. Esse relacionamento entre os termos é utilizado como base para a definição dos atributos e de seus respectivos valores na TAV.

Na Fase 2, o conteúdo do conjunto de LMt é padronizado, para que, com base no DC, seja aplicado o Algoritmo de Busca e Preenchimento, com a finalidade de mapear os LMt e preencher a TAV. Os termos não processados são armazenados para posterior avaliação.

Neste trabalho, é construída uma ontologia para substituir o DC, a qual será utilizada para representar o conjunto de termos descritos nos LMt.

Na área de Inteligência Artificial, o termo ontologia foi adotado para se referir aos termos e conceitos que podem ser utilizados para descrever uma área do conhecimento ou construir uma representação da mesma [3]. Portanto, uma ontologia é capaz de representar os conceitos dentro de um determinado domínio, assim como a semântica e o relacionamento entre eles, fornecendo uma estrutura básica, a qual serve de alicerce para a construção de uma base de conhecimento [2, 3].

Para a construção da ontologia foi utilizado o método a seguir [6]:

- **Definição de escopo e objetivo:** definição do domínio e dos requisitos a serem atendidos, assim como escolha das ferramentas e da linguagem a ser utilizada para representar o conhecimento;
- **Construção:** construção da ontologia com base nos conceitos do escopo definido e nos relacionamentos existentes entre eles;
- **Avaliação:** a partir da especificação inicial dos requisitos, é avaliada a consistência e a expressividade da ontologia, de maneira iterativa e interativa até que todos os requisitos sejam atendidos.

Para a construção da ontologia foi utilizado o *software* Protégé, que utiliza Web Ontology Language (OWL) como linguagem de representação.

Resultados e Discussão

Os principais conceitos representados no DC foram utilizados como base para determinar as classes da ontologia. Na Figura 1 é apresentada parte da ontologia construída, na qual *Thing* é a classe principal que contém todas as outras classes. A classe *Atributo_BD* corresponde a todos os atributos da TAV e a classe *Valor_Atributo* a todos os valores possíveis para um atributo, portanto ambas as classes possuem subclasses não apresentadas na Figura 1, por questões de espaço. A classe *Termo* representa, em dois tipos principais, os termos presentes nos LMt padronizados: *Regiao*, referente às regiões anatômicas e *Observacao,* que representa as características (classe *Caracteristica*) e as situações (*classe Situacao* e as subclasses *Anormalidade* e *Outra_Situacao*) descritas nos LMt.

Com base na avaliação de especialistas do domínio, constatou-se que a ontologia possibilitou uma melhor modelagem das informações contidas nos LMt, classificando os conceitos do domínio de maneira mais específica, permitindo que as relações hierárquicas entre local, característica e subcaracterística fossem representadas com maior

completude, uma vez que, no DC essas relações são representadas parcialmente.

O entendimento dos termos descritos nos laudos e a compreensão de como se relacionam, aliado às características da ontologia, foram os principais aspectos para a substituição do DC por uma ontologia.

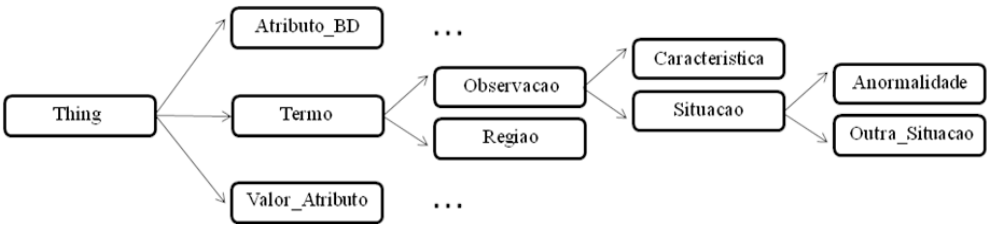

Figura 1 – Representação parcial dos principais conceitos da ontologia.

Conclusões

A utilização de ontologias mostrou-se eficiente, pois a ontologia construída representou adequadamente os termos descritos nos LMt, tendo um nível de eficácia muito similar em comparação com o DC. Nesse sentido, trabalhos futuros incluem a utilização da ontologia para o mapeamento dos LMt e preenchimento da TAV, para posterior aplicação a laudos reais.

Agradecimentos

À UNIOESTE pelo auxílio por meio do financiamento de bolsas PIBIC.

Referências

1. FAYYAD, U. M. et. al. From data mining to knowledge discovery in databases. **AI Magazine**. USA, v. 17, p. 37-54, 1996.
2. GRUBER, T. R. **Toward principles for the design of ontologies used for knowledge sharing**. Formal Ontology in Conceptual Analysis and Knowledge Representation, 1993.
3. GUIMARÃES, F. J. Z. **Utilização de ontologias no domínio B2C**. 2002. Dissertação (Mestrado em Informática), PUC-RJ, 2002. Disponível em: http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0024134_02_cap_04.pdf
4. HONORATO, D. F. et. al. Construction of an attribute-value representation for semi-structured medical findings knowledge extraction. **CLEI Eletronic Journal,** 2008, v. 11, p. 1-12.
5. PYLE, D. **Data preparation for data mining**. California: Morgan Kaufmann, 1999.
6. USCHOLD, M. et. al. Ontologies: principles, methods and applications. **Knowledge Engineering Review**. UK, v. 11, p. 93–155, 1996.