3º EAICTI
3º Encontro Anual de Iniciação Científica, Tecnológica e Inovação

Desenvolvimento de uma ontologia para doenças do cólon diagnosticáveis por meio de exames de vídeocolonoscopia

[1]Silvani Weber da Silva Borges (PIBIC/CNPq/Unioeste), [12]Huei Diana Lee (Orientadora), [1]Newton Spolaôr, [13]Jefferson Tales Oliva, [12]Feng Chung Wu e-mail: {silvani.borges2,hueidianalee}@gmail.com

[1]Universidade Estadual do Oeste do Paraná/Centro de Engenharias e Ciências Exatas/Laboratório de Bioinformática (LABI)/Foz do Iguaçu, PR
[2]Universidade Estadual de Campinas (UNICAMP)/
Faculdade de Ciências Médicas (FCM)/Campinas, SP
[3]Universidade de São Paulo (USP)/Instituto de Ciências Matemáticas e de Computação (ICMC)/São Carlos, SP

Área/subárea: Ciências Exatas e da Terra/ Ciência da Computação.
Palavras-chave: terminologia, atributo-valor, endoscopia digestiva baixa.

Resumo

As informações colhidas durante o exame de videocolonoscopia são frequentemente descritas em linguagem natural em laudos médicos, os quais constituem uma importante fonte de dados sobre o histórico clínico dos pacientes e podem servir para auxiliar em diagnóstico de casos futuros. O mapeamento desses laudos para uma representação estruturada, útil para a realização de estudos retrospectivos e prospectivos, ocorre geralmente de modo manual, relativamente lento e sujeito à subjetividade. Essa abordagem também inviabiliza o uso de processos computacionais, como a Mineração de Dados, para a extração de conhecimento em dados textuais inerentes aos exames. Para auxiliar na automatização do mapeamento de laudos, uma ontologia que representa termos de um domínio pode ser utilizada com flexibilidade. O objetivo deste trabalho consistiu no desenvolvimento de uma ontologia para apoiar o mapeamento automático dos laudos médicos no domínio de doenças do cólon diagnosticáveis por meio de exames de videocolonoscopia. Essa ontologia foi então incorporada em um sistema computacional para mapeamento de laudos e submetida a uma avaliação experimental. Para isso, foi utilizado um conjunto de 100 laudos artificiais. Na avaliação experimental, a ontologia contribuiu para o mapeamento automático de 88,96% das informações manualmente identificadas no conjunto de laudos. Esse resultado motiva aplicações futuras para a representação estruturada de laudos médicos reais de colonoscopia.

ISSN 2448-0681

Introdução

A vídeocolonoscopia é um exame utilizado no diagnóstico de lesões do reto e do intestino grosso e permite a visibilização de até 90% da parede intestinal (Benevides & Santos, 2016). As observações relevantes do procedimento são registradas como texto em Laudos Médicos (LM), os quais auxiliam a manter um histórico do paciente e podem ser usados para diagnósticos futuros. Em geral, esses laudos são mapeados manualmente para planilhas eletrônicas e podem ser utilizados, por exemplo. No entanto, essa tarefa torna-se inviável para grandes quantidades de laudos devido ao tempo necessário para esse mapeamento manual e à possível subjetividade da tarefa (Oliva, 2014).

Neste sentido, ferramentas e métodos computacionais emergem como um importante instrumento para auxiliar no mapeamento automático de LM (Lee *et al*., 2013). Como resultado, dados não estruturados podem ser transformados para dados estruturados, por exemplo, no formato de Tabela Atributo-Valor (TAV) (Lee *et al*., 2011). Na TAV, cada linha refere-se a um laudo e cada coluna a uma informação mapeada do laudo. Essa estruturação é importante para possibilitar a submissão dos dados para processos computacionais de extração de conhecimento, como a Mineração de Dados (MD) (Han *et al*., 2011; Lee *et al*., 2011).

Para contribuir com o mapeamento automático de laudos para uma Base de Dados (BD) estruturada, compatível com a MD, uma ferramenta computacional foi desenvolvida pelo Laboratório de Bioinformática (LABI), da Universidade Estadual do Oeste do Paraná (UNIOESTE), em parceria com o serviço de Coloproctologia da Faculdade de Ciências Médicas (FCM) da Universidade Estadual de Campinas (UNICAMP) (Wu *et al*., 2010a). Essa ferramenta inclui componentes como uma ontologia, a qual permite organizar termos inerentes a exames de videocolonoscopia ou Endoscopia Digestiva Baixa (EDB), de modo estruturado e flexível (Oliva, 2014).

Material e Métodos

A ferramenta computacional utilizada neste trabalho implementa as duas fases do método de mapeamento automático de LM proposto em (Wu *et al*., 2010a; Wu *et al*., 2010b). A Fase 1 conta com apoio de especialistas para identificar padrões em laudos e definir os atributos que irão compor uma BD no formato TAV. A Fase 2 utiliza esses padrões para mapear automaticamente o conteúdo de LM para uma BD. Em particular, a Fase 1 é constituída pelas seguintes tarefas (Oliva, 2014):

- **Identificação de frases únicas**: geração de um Conjunto de Frases Únicas (CFU) a partir dos LM, o qual permite remover frases redundantes para obter apenas uma única instância de cada frase;
- **Normalização dos termos**: remoção de acentos e substituição de caracteres maiúsculos por minúsculos no CFU;

ISSN 2448-0681

- **Remoção de *stopwords***: eliminação de termos considerados irrelevantes pelos especialistas para compreensão das frases do CFU. As *stopwords* são registradas em uma lista (*stoplist*);
- **Lematização:** transformação de termos do CFU para determinar seu lema, por exemplo, as formas *ulceração* e *ulcerado* provém do mesmo lema, *úlcera*;
- **Construção do Arquivo de Padronização (AP)**: definição de palavras e expressões para serem substituídas por termos equivalentes.

Após o pré-processamento, os especialistas analisam o CFU resultante, definindo os atributos e valores que irão compor a TAV e a ontologia, com objetivo de representar padrões relevantes embutidos nos laudos. As Regras de Mapeamento (RM) inerentes à ontologia, que possibilitam mapear esses padrões para uma BD na Fase 2, são também elaboradas com o auxílio dos especialistas. Em seguida, os atributos e as RM são utilizados para a construção de uma ontologia constituída por classes e instâncias associadas a termos presentes em LM de um domínio. As propriedades de cada classe são instanciadas por indivíduos que representam termos mapeáveis por meio de RM, atributos e valores relacionados. Por exemplo, o padrão *reto úlcera* é associado com: (1) a instância *reto* de uma classe referente a uma região anatômica, (2) a instância *ulcera* de uma classe relacionada a uma característica de região anatômica, e (3) a instância *reto_ulcera* referente a uma classe para um atributo. A ontologia de EDB foi inspirada na estrutura que foi desenvolvida para Endoscopia Digestiva Alta (EDA) (Oliva, 2014).

Muitos termos que integram a ontologia do presente trabalho foram extraídos de uma estrutura com termos do domínio, denominada dicionário do conhecimento, proposto em Lee *et al*. (2013). Esse dicionário foi construído com base nas sentenças descritas em linguagem natural, utilizando 100 laudos textuais artificiais referentes às informações consideradas relevantes encontradas em exames de EDB. A partir dos termos desta estrutura, a ontologia foi gerada por meio do framework Protégé (http://protege.stanford.edu/) e descrita em linguagem *Ontology Web Language* (https://www.w3.org/OWL/).

Na Fase 2, o conjunto de laudos a ser mapeado é padronizado com apoio das tarefas citadas, da *stoplist* e do AP. Após, um algoritmo de mapeamento é aplicado em cada laudo para identificar valores que, quando associados a RM baseados em componentes da ontologia, são usados para preencher uma TAV (Oliva, 2014).

Resultados e Discussão

O uso da ontologia possibilita a representação de atributos da TAV por meio de esquemas flexíveis, como ilustrado na Figura 1. Nesse esquema, proposto em Wu *et al*. (2010a) e Oliva (2014), *Thing* é a classe principal que abrange as demais classes. *Atributo* e *Valor_do_atributo* correspondem aos atributos e valores da TAV, respectivamente. *Termo* representa os termos presentes nos laudos, sendo esta

ISSN 2448-0681

especializada nas classes *Região*, com descrição de porções anatômicas do corpo humano, e *Observação*, a qual generaliza *Característica* e *Anormalidade*.

A ontologia construída neste trabalho, pioneiro no domínio de EDB, especifica o esquema da figura com 443 instâncias, sendo 35 da classe *Região*, 63 da classe *Característica* e 335 instâncias referentes a 489 atributos.

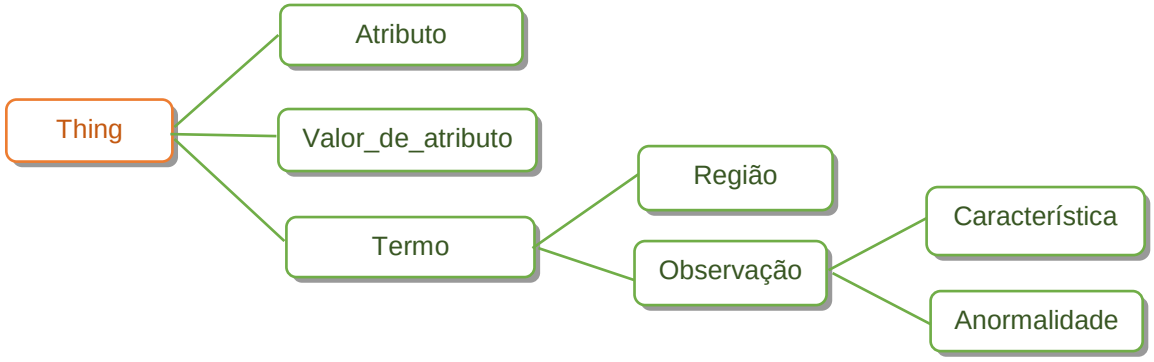

Figura 1 – Estrutura geral da ontologia proposta por Wu *et al*. (2010a) e Oliva (2014).

Os resultados gerados com o mapeamento dos laudos permitem observar que 88,96% dos atributos identificados por meio da Fase 1 no conjunto de 100 laudos foram preenchidos com os respectivos valores, sendo que, em média, 4,35 atributos foram preenchidos para cada laudo (desvio padrão de 14,47).

É importante ressaltar que a complexidade do algoritmo de mapeamento é proporcional à variabilidade de termos a serem processados em cada laudo. No domínio de EDA se nota relativamente pouca variação nos termos, resultando em um AP com poucos padrões, 81 ao todo, e em uma ontologia com relativamente poucas classes e instâncias, 15 e 51 respectivamente. A aplicação da ontologia correspondente levou a uma taxa de mapeamento de 100% dos valores esperados (Oliva, 2014). Por outro lado, no domínio de EDB, os LM são constituídos por termos que tem muitas variações, demandando a inclusão de 417 padrões no AP. Nesse cenário, diante das circunstâncias específicas do domínio, é possível considerar que o desempenho atingido no presente trabalho é satisfatório, podendo ser melhorado com a definição de novas regras de mapeamento e de padronização.

Conclusões

A ontologia de EDB desenvolvida neste trabalho auxiliou no mapeamento automático de um conjunto de 100 LM artificiais para uma BD estruturada. Essa base é compatível com o processo de MD, o qual, quando aplicado, possibilita a extração de padrões que representam o conhecimento embutido nos dados.

ISSN 2448-0681

Trabalhos futuros incluem aumentar a porcentagem de termos mapeados nos laudos ao aprimorar a ontologia e o arquivo de padronização, bem como incorporar termos do domínio mais recentes, publicados na literatura nacional e internacional.

Agradecimentos

À UNIOESTE/CNPq pela concessão de bolsa de iniciação científica.

Referências

Benevides, I.B.S. & Santos, C.H.M. (2016). Colonoscopy in the diagnosis of acute lower gastrointestinal bleeding. *Journal of Coloproctology* **36**, 185-188.

Han, J., Kamber, M. & Pei, J. (2011). *Data mining*: *Concepts and techniques*. Waltham: Elsevier.

Lee, H.D., Monard, M.C., Honorato, D.F., Lorena, A.C., Ferrero, C.A., Maletzke, A.G., Zalewski, W., Coy, C.S.R., Fagundes, J.J. & Wu, F.C. (2011). Mapping unstructured data in digital and printed documents into attribute-value tables. In: Rafael, A., Andrade, E., Gómez, J.M., Valdés, A.R. (Org.). *Towards a trans-disciplinary technology for business intelligence, gathering knowledge discovery, knowledge management and decision making* (pp. 198-209). Verlag: Shaker.

Lee, H.D., Oliva, J.T., Maletzke, A.G., Machado, R.B., Voltolini, R.F., Coy, C.S.R., Fagundes, J.J. & Wu, F.C. (2013). Sistema computacional para automatização do processo de mapeamento de laudos médicos por ontologias. In Anais do 62º Congresso Brasileiro de Coloproctologia, São Paulo, São Paulo, Brasil.

Oliva, J.T. (2014). *Automatização do processo de mapeamento de laudos médicos para uma representação estruturada*. Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia de Sistemas Dinâmicos e Energéticos, Universidade Estadual do Oeste do Paraná.

Wu, F.C., Lee, H.D., Coy, C.S.R., Fagundes, J.J., Ferrero, C.A., Machado, R.B., Maletzke, A.G., Zalewski, W., Leal, R.F., Ayrizono, M.L.S. & Costa, L.H.D. (2010a) BR Patente 01810036941.

Wu, F.C., Lee, H.D., Ferrero, C.A., Coy, C.S.R., Fagundes, J.J., Machado, R.B & Costa, L.H.D. (2010b). Development of an Ontology-based Approach for Mapping High Digestive Endoscopy Medical Reports into Structured Databases. In Anais ALIO-INFORMS Joint International Meeting, Buenos Aires, Argentina.

ISSN 2448-0681