4º EAICTI
4º Encontro Anual de Iniciação Científica, Tecnológica e Inovação

Expansão de uma ontologia para apoiar o mapeamento de laudos de colonoscopia para bases de dados estruturadas

[1]Silvani Weber da Silva Borges (PIBIC/CNPq/Unioeste), [1,2]Huei Diana Lee (Orientadora), [1]Newton Spolaôr, [1,2]Feng Chung Wu, e-mail: silvani.borges2@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

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

Resumo

Este trabalho teve como objetivo aprimorar uma ontologia de colonoscopia já existente com termos recentes e utilizados mundialmente, bem como utilizar medidas novas, específicas da área de ontologias, para avaliar a qualidade da ontologia. Essa ontologia compõe uma ferramenta computacional desenvolvida para o mapeamento automático de laudos médicos. Para expansão da ontologia foram consultadas duas bases de dados em saúde e uma terminologia internacional no domínio de endoscopia. Como resultado foram identificados 25 novos termos e a partir desses foram criadas 132 novas instâncias. Além disso, as medidas aplicadas para avaliar aspectos da estrutura ontológica, denominadas riqueza de relacionamentos e riqueza de classes, demonstraram que a ontologia apresenta boa diversidade de conexões e instâncias suficientes para representar o conhecimento no domínio de colonoscopia.

Introdução

Diante da importância do exame de videocolonoscopia para detecção de lesões no intestino grosso e reto, este procedimento tem sido cada vez mais indicado por especialistas, e, com isso informações valiosas são armazenadas em Laudos Médicos - LM (Shortliffe & Barnett, 2013). Todavia as observações relativas ao procedimento são registradas em linguagem natural impossibilitando o uso direto de processos computacionais, como a Mineração de Dados – MD, capazes de extrair conhecimento a partir desses laudos (Witten *et al*., 2016).

Nessa perspectiva, o grupo de pesquisa do Laboratório de Bioinformática da Universidade Estadual do Oeste do Paraná, em parceria com o serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de

ISSN 2448-0681

Campinas, desenvolveu um método de mapeamento automático de LM e a ferramenta que o implementa (Wu *et al*., 2010; Coy *et al*., 2015).

Esta ferramenta é compatível com o processo de MD e inclui componentes como uma ontologia, a qual possibilita organizar termos inerentes a exames de colonoscopia ou Endoscopia Digestiva Baixa - EDB de modo estruturado e flexível, por exemplo, no formato de Tabela Atributo-Valor - TAV (Lee *et al*., 2013). Nesta tabela cada coluna refere-se a um atributo/característica enquanto que cada linha constitui um LM com seu respectivo valor.

Vale ressaltar que a ontologia de EDB desenvolvida em Borges *et al*. (2017) foi adaptada da ontologia de Endoscopia Digestiva Alta - EDA apresentada em Oliva (2014) e foi construída somente com termos extraídos de um dicionário do conhecimento elaborado por especialistas do domínio médico e computacional (Lee *et al*. 2013). Embora a ontologia tenha se mostrado funcional, não foi utilizada nenhuma medida que pudesse avaliar a qualidade da ontologia quanto a representação do conhecimento. Assim sendo, este trabalho teve como objetivo aprimorar a ontologia de EDB com termos mais recentes e utilizados mundialmente, bem como aplicar medidas de qualidade apropriadas presentes na literatura (Tartir *et al*., 2005)

Material e Métodos

Para expandir a ontologia foram consultadas duas bases de dados com terminologias na área da saúde, o DeCS e o Snomed (Descritores em Ciências da Saúde, 2018; Snomed International, 2018). Além disso, foi utilizada uma terminologia internacional específica no domínio de endoscopia digestiva denominado *Minimal Standard Terminology* - MST (World Endoscopy Organization, 2016). Esta terminologia foi traduzida manualmente para o Português e contou com o apoio de um especialista do domínio para auxiliar na identificação de termos que não possuem tradução para o Português.

Após a identificação e a inclusão dos novos termos, a ontologia foi avaliada por meio de medidas ontológicas propostas por Tartir *et al*. (2005). Tais medidas possibilitam avaliar determinados aspectos de uma ontologia, como a *riqueza de relacionamentos* e a *riqueza de classes*.

A Riqueza de Relacionamentos (RR) reflete a diversidade de relações e a organização dessas conexões dentro de uma estrutura ontológica. A Riqueza de Classes (RC), por sua vez, está relacionada ao modo como as instâncias estão distribuídas entre as classes. Uma classe descreve um conjunto de objetos com características similares, os quais são denominados de instâncias.

Formalmente, RR representa a proporção entre a quantidade de relacionamentos total (P) definidas no esquema e o número de relacionamentos do tipo classe-subclasse (SC) juntamente com o número de relacionamentos total. RR é definida pela Equação 1.

ISSN 2448-0681

$$RR = \frac{P}{SC+P} \qquad (1)$$

O resultado obtido representa a diversidade de conexões existentes em uma ontologia. Quanto mais próximo de *um* o valor, mais diversas são as conexões, permitindo assim maior comunicação entre objetos de classes diferentes. Por outro lado, um valor próximo à *zero* indica menor conexão entre os objetos de classes diferentes (Tartir *et al*., 2005).

A segunda medida utilizada neste trabalho consiste na RC. Formalmente, esta medida representa a proporção de classes que apresentam instâncias (c') pelo total de classes definidas no esquema ontológico (c). RC é definida pela Equação 2.

$$RC = \frac{c\prime}{c} \qquad (2)$$

Uma porcentagem baixa de RC sugere que a ontologia não possui instâncias suficientes para representar o conhecimento de um domínio. Por outro lado, um resultado próximo a 100% indica uma boa representação do conhecimento (Tartir *et al*., 2005).

A ontologia expandida foi avaliada por meio do mapeamento de 92 laudos de EDB.

Resultados e Discussão

Após a busca nas bases de dados DeCS e Snomed, bem como na tradução da terminologia MST, foram identificados 25 novos termos para incorporar à ontologia de EDB. Assim, 132 novas instâncias foram geradas a partir desses termos e distribuídas entre classes da seguinte maneira:

- 25 instâncias para a classe *Observação*, a qual descreve características, anormalidades e outras situações, por exemplo, *ulcera*;
- 107 instâncias para a classe *Atributo_Base_de_Dados*, a qual representa um atributo sentença descrita nos LM, por exemplo, *sigmoide_ulcera*.

Com a inclusão dos novos termos, a ontologia de EDB originalmente com 442 instâncias, passou a ficou constituída por 574 instâncias. No entanto, não houve necessidade de aumentar a quantidade de classes.

Quanto às medidas para avaliação da ontologia, foram contabilizados 1.031 relacionamentos, sendo que quatorze relações são do tipo classe-subclasse e 1.017 são do tipo classe-instância. Como resultado, obteve-se RR=0,99, sugerindo uma boa diversidade de conexões entre objetos de classes distintas e consequentemente melhor representação do conhecimento.

ISSN 2448-0681

A medida de RC, por sua vez, permitiu observar que 60% das classes estão instanciadas. No entanto não se pode sugerir que a ontologia não apresente instâncias suficientes para representar o conhecimento, visto que algumas classes, como *Atributo_Base_de_Dados* e *Característica*, apresentam um alto número de instâncias, 442 e 88 respectivamente.

Vale ressaltar que no domínio de EDB há uma alta variedade de termos em relação ao domínio de EDA. Por exemplo, as mesmas classes *Atributo_Base_de_Dados* e *Característica* no domínio de EDA apresentam dezesseis e seis instâncias respectivamente.

Essa variedade de termos também influencia na taxa de mapeamento dos laudos, visto que a ontologia de EDB auxiliou no mapeamento de 70,11% dos 174 termos identificados em laudos, enquanto que a ontologia de EDA alcançou 100%. No entanto ambas as ontologias foram avaliadas por especialistas e foram consideradas aptas para representar o conhecimento do domínio em questão.

Conclusões

Neste trabalho a ontologia de EDB foi aprimorada com termos recentes e utilizados mundialmente considerando a terminologia MST e bases de dados em saúde. Além disso, foram aplicadas medidas de qualidade, as quais indicaram que a ontologia é capaz de representar o conhecimento no domínio de colonoscopia.

Trabalhos futuros incluem melhorar a taxa de mapeamento por meio de aprimoramento de estruturas adicionais da ferramenta computacional, como o Arquivo de Padronização.

Agradecimentos

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

Referências

Borges, S.W.S, Lee, H.D., Spolaôr, N., Oliva, J.T. & Wu, F.C. (2017). Desenvolvimento de uma ontologia para doenças do colón diagnosticáveis por meio de exames de videocolonoscopia. In Anais do 3º Encontro Anual de Iniciação Cientifica, Tecnológico e Inovação, Cascavel, Paraná, Brasil.

Coy, C.S.R., Oliva, J.T., Lee, H.D., Wu, F.C., Fagundes, J.J., Machado, R.B., Spolaôr, N., Fonteque Junior, M., Leal, R.F., Ayrizono, M.L.S. (2015) BR Patente 12015000342-9.

ISSN 2448-0681

Descritores em Ciências da Saúde (2018). Biblioteca virtual em saúde. http://decs.bvs.br/. Acesso em 22 de agosto de 2018

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.

Minimal standard terminology for gastrointestinal endoscopy – MST 3.0. (2009). http://www.worldendo.org/assets/downloads/pdf/resources/mst/mst30.pdf. Acesso em 10 de agosto de 2018.

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.

Shortliffe, E.H. & Barnett, G.O (2013). Biomedical data: their acquisition, storage, and use. In Shortliffe, E.H. & Cimino, J.J. (Eds.), *Biomedical Informatics: computer apllications in heath care and biomedicine* (pp. 39-66). New York: Springer.

Snomed International (2018). The Global Language of Healthcare. https://www.snomed.org/. Acesso em 22 de agosto de 2018.

Tartir, S., Arpinar, I.B., Moore, M., Sheth, A.P. & Aleman-Meza, B. (2005). OntoQA: metric-based ontology quality analysis. In Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed Autonomous, Semantically Heterogeneous Data and Knowledge Sources, Nova Orleans, Estados Unidos.

Witten, I., Frank, E., Hall, M. & Pal, C. (2016). *Data mining: practical machine learning tools and techniques. 4th ed*. Burlington: Elsevier.

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. (2010) BR Patente 01810036941.

ISSN 2448-0681