Construção de Bases de Dados a partir de Laudos Endoscópicos: Estudo de Caso

Newton Spolaôr[1], Huei Diana Lee[1,2], Everton A. Cherman[1], Daniel de F. Honorato[2], João J. Fagundes[3], Juvenal R. N. Góes[3], Cláudio S. R. Coy[3], Feng Chung Wu[1,2,3]

[1]Laboratório de Bioinformática (LABI),
Universidade Estadual do Oeste do Paraná (UNIOESTE), Parque Tecnológico Itaipu (PTI)
[2]Laboratório de Inteligência Computacional (LABIC), Universidade de São Paulo (USP)
[3]Universidade Estadual de Campinas (UNICAMP)

1. Objetivo

Endoscopia Digestiva Alta – EDA – é um exame amplamente realizado, pois auxilia na detecção de doenças gastroduodenais. As observações verificadas nesse exame são descritas em Laudos Médicos – LM, a partir dos quais é possível adquirir conhecimento. A extração de padrões pode ser realizada por meio de processos computacionais, como a Descoberta de Conhecimento em Bases de Dados [1]. No entanto, é necessário que os dados estejam dispostos no formato atributo-valor. Neste trabalho é apresentado um estudo de caso que realiza a construção de Bases de Dados – BD, a partir de informações do esôfago, estômago e duodeno, descritas em LMs de EDA.

2. Materiais e Métodos

Os LMs de EDA são compostos por blocos textuais, cada um correspondendo a um dos três órgãos verificados durante esse exame. Os dados que compõem um bloco são categorizados, freqüentemente, como locais, características e subcaracterísticas. A partir dos 609 LMs utilizados neste trabalho, foram extraídos os dados referentes ao estômago, esôfago e duodeno. A metodologia [2], composta por duas fases, foi aplicada no processamento dos dados de cada órgão, possibilitando a construção de um dicionário e uma BD para cada órgão considerado. Na primeira fase foram elaborados, com o apoio de técnicas de inteligência artificial, padronizações textuais e conjuntos de frases únicas – CFU, os quais foram analisados com o auxílio de especialistas para definir um dicionário e os atributos que constituem a BD de um domínio. A estrutura hierárquica que o dicionário apresentou refletiu a disposição dos dados em um LM. A segunda fase compreendeu o mapeamento dos dados do conjunto de LMs para a BD, utilizando o dicionário construído na fase anterior. Quando foram encontradas relações entre locais, características e subcaracterísticas, os atributos correspondentes foram preenchidos, possibilitando a construção das BDs de cada domínio.

3. Resultados e Discussão

Os dados corretamente descritos foram mapeados na íntegra, comprovando o sucesso da metodologia. O CFU correspondente ao esôfago reduziu em 97,86% a quantidade inicial de 3044 frases. A BD construída para esse domínio era composta por 16 atributos e apresentou um preenchimento de 88,99%. O CFU construído para o estômago apresentava 95,04% das frases originais referentes a esse órgão e a BD foi composta por 22 atributos, alcançando 25,17% de preenchimento. Foram eliminadas 94,85% das 1710 frases relativas ao duodeno e a BD correspondente era composta por 51 atributos com preenchimento de 30,92%. As BDs do estômago e duodeno apresentaram baixo preenchimento devido à escassa presença de dados referentes a alguns atributos na BD, o que provocou a concentração dos dados em poucos atributos.

4. Conclusões

Neste trabalho foi apresentada a aplicação da metodologia proposta no processamento de três domínios. Os dicionários construídos podem auxiliar no mapeamento de outros conjuntos de LMs para posterior aplicação da DCBD.

5. Referências

[1] Fayyad U. M., Piatetsky-Shapiro G., Smyth P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine. vol. 17:37–54.

[2] Honorato D. D. F., et al. (2007). Construção de uma Representação Atributo-valor para Extração de Conhecimento a partir de Informações Semi-estruturadas de Laudos Médicos. In: Anais do XXXIII Conferencia Latinoamericana de Informática (a ser publicado). San José - Costa Rica.