Construção Semi-automática de uma Base de Dados a partir de Laudos Colonoscópicos.

Everton Alvares Cherman[1], Huei Diana Lee[1,2], Daniel de Faveri Honorato[2], Cláudio S. R. Coy[3], João J. Fagundes[3], Juvenal R. Navarro Góes[3], Feng Chung Wu[1,3]

[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

O processo de Descoberta de Conhecimento em Bases de Dados – DCBD – tem como objetivo construir modelos que auxiliem especialistas na tomada de decisão. Para que o processo seja aplicado é necessário que os dados estejam em um formato estruturado denominado atributo-valor [1]. Neste trabalho é desenvolvida uma metodologia para mapeamento automático de informações contidas em Laudos Médicos – LM – de colonoscopia para uma Base de Dados – BD – estruturada no formato atributo-valor.

2. Materiais e Métodos

A colonoscopia é um importante exame para o diagnóstico de doenças colorretais. Neste trabalho foram consideradas informações descritas em língua natural contidas em 100 LM de colonoscopia coletados sem identificação dos pacientes no Serviço de Coloproctologia da Faculdade de Ciências Médicas da UNICAMP. A metodologia de mapeamento é composta por duas fases [2]. Na primeira fase é construído um dicionário do domínio do conhecimento por meio de reuniões com especialistas e com o auxílio de um conjunto de frases únicas e de um arquivo de padronização gerados a partir de todas as frases contidas nos LM. O dicionário definido é composto por locais (estruturas anatômicas), características associadas a esses locais e também por atributos e valores da BD relacionados às características. Na segunda fase, com base no dicionário definido na fase anterior, é realizado o mapeamento das informações para a BD por meio de um algoritmo de busca e preenchimento, o qual é aplicado para cada um dos LM. Ao final do processo é construída uma BD estruturada no formato atributo-valor com as informações contidas nos LM.

3. Resultados e Discussão

O dicionário construído na primeira fase foi composto por 34 locais e 84 características distribuídas nesses locais e a BD foi definida com 277 atributos. Na segunda fase foi realizado o processamento dos LM, dos quais foram mapeados 793 valores de atributos dos 967 esperados, representando uma precisão 82%. O mapeamento completo foi inviabilizado devido à forma como as frases foram descritas nos LM, as quais, entre outros, eram dependentes de contextos anteriores. Tal característica demanda um algoritmo de busca e preenchimento muito complexo. Desse modo, os valores não preenchidos foram mapeados manualmente, constituindo 100% de preenchimento da BD. Após uma análise qualitativa do mapeamento, observou-se que todos os valores preenchidos foram mapeados corretamente.

4. Conclusões

Neste trabalho foi construído um dicionário do domínio, o qual permitiu, juntamente com o algoritmo de busca e preenchimento, o mapeamento semi-automático de informações contidas em 100 LM de colonoscopia. Novos conjuntos de LM podem ser mapeados de maneira semi-automática, o que representa uma redução no tempo total do processo. O processo de DCBD poderá ser realizado a partir dessa BD extraindo modelos que podem auxiliar especialistas na tomada de decisão.

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.