METODOLOGIA DE MAPEAMENTO AUTOMÁTICO DE LAUDOS COLONOSCÓPICOS

Everton Alvares Cherman (PIBIC/CNPq-UNIOESTE), Huei Diana Lee (Orientadora), Daniel de Faveri Honorato, Cláudio Sady Rodrigues Coy, João José Fagundes, Juvenal Ricardo Navarro Góes, Feng Chung Wu (Co-orientador), e-mail: evertoncherman@gmail.com.

Universidade Estadual do Oeste do Paraná/Centro de Engenharias e Ciências Exatas/Ciência da Computação – Foz do Iguaçu – PR.

Palavras-chave: descoberta de conhecimento em bases de dados, pré-processamento, laudos médicos.

Resumo:

A Descoberta de Conhecimento em Bases de Dados pode auxiliar na área médica, porém, para que o processo seja aplicado, é necessário que os dados estejam no formato atributo-valor. Neste trabalho é apresentada uma metodologia de mapeamento de laudos médicos descritos em língua natural para bases de dados estruturadas, a qual está sendo aplicada a informações de laudos médicos de colonoscopia. As técnicas de lematização e n-gramas estão sendo estudadas para auxiliar na metodologia.

Introdução

Processos computacionais como o de Descoberta de Conhecimento em Bases de Dados – DCBD – podem auxiliar na extração e análise de informações contidas em Laudos Médicos – LM [1]. O DCBD é composto por três etapas. A primeira corresponde ao pré-processamento de informações, a qual tem como objetivo a transformação dos dados para o formato adequado para etapa seguinte, denominado atributo-valor. O pré-processamento é a etapa que consome o maior tempo computacional no processo, em torno de 80% do total [2]. A próxima etapa é a mineração de dados e é focada em construir modelos que representem padrões contidos nos dados. Na última etapa, denominada pós-processamento, os modelos construídos são avaliados e validados por especialistas.

Em [3,4] é apresentada uma metodologia de mapeamento de LM de Endoscopia Digestiva Alta – EDA – para uma Base de Dados – BD – no formato atributo-valor. Neste trabalho é proposta a utilização de duas técnicas para auxiliar na aplicação da metodologia em LM de colonoscopia.

Materiais e Métodos

A colonoscopia é um importante exame para o diagnóstico de doenças do intestino grosso [5]. Neste trabalho foram utilizados 100 LM de colonoscopia coletados do Serviço de Coloproctologia da Universidade Estadual de

Campinas. Os LM são compostos por dois segmentos: estruturado e semi-estruturado. As informações do segundo segmento são utilizadas para a aplicação da metodologia. Na Figura 1 é apresentado um exemplo de um LM de colonoscopia.

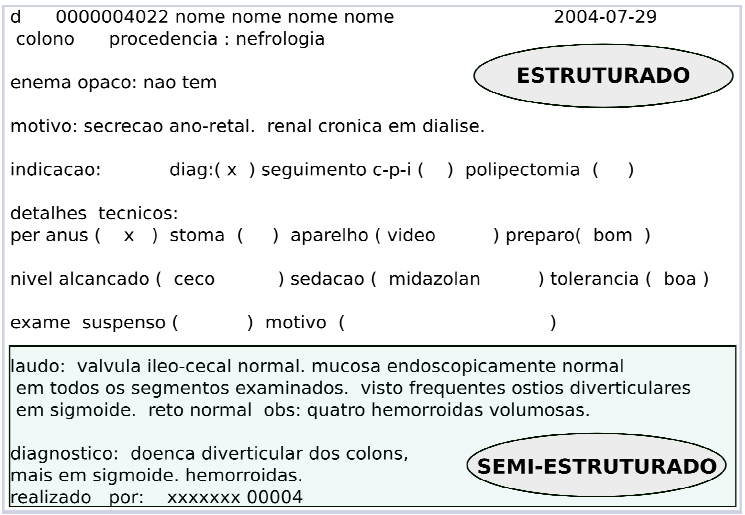

Figura 1 – Exemplo de um LM de colonoscopia.

A metodologia é constituída por duas fases. O objetivo da primeira é construir um dicionário do domínio do conhecimento, no qual estão descritos e classificados todos os termos do domínio extraídos de um conjunto de LM, bem como seus respectivos atributos a serem preenchidos na BD. A construção do dicionário é realizada em quatro etapas, nas quais a interação com especialistas é importante e freqüente. A primeira tem como objetivo a definição de um primeiro Conjunto de Frases Únicas – CFU1 –, reunindo em um arquivo todas as frases contidas no conjunto de LM e eliminando as frases redundantes. Na próxima etapa é realizada a definição de um Arquivo de Padronização – AP –, devido ao uso freqüente na área médica de sinônimos para representar as informações, as quais, com o AP, são convergidas para um mesmo termo. A terceira etapa consiste na técnica de remoção de *stopwords*, a qual é definida como a remoção de palavras do texto que ao serem retiradas não modificam o sentido original da frase. Ao aplicar a remoção de *stopwords* e padronizações no CFU1 é constituído o CFU2. Na última etapa é aplicada sobre o CFU2 a técnica de *stemming*, a qual consiste em manter apenas o radical de cada palavra, constituindo o terceiro conjunto de frases únicas, denominado CFU3.

Desse modo, em reunião com especialistas e com auxílio dos CFUs e do AP, são identificados os termos contidos nos LM e seus relacionamentos, definindo o dicionário do domínio do conhecimento e os atributos da BD.

A segunda fase tem como objetivo definir uma lógica de mapeando das informações para a BD, a qual deve ser capaz de mapear qualquer frase dos LM, levando em consideração a disposição dos termos na frase, de modo que respeite a descrição contida no dicionário.

Neste trabalho é proposta a utilização de duas técnicas para auxiliar a metodologia na identificação de termos e padrões dos LM. A primeira é

definida como lematização e consiste em substituir as palavras por suas formas canônicas, como o singular de um substantivo ou o infinitivo de um verbo. A segunda técnica denomina-se n-gramas e tem como objetivo fornecer a freqüência com que toda composição de n palavras consecutivas são descritas em um conjunto de textos. Por exemplo, unigrama é a freqüência de toda palavra isolada, bem como bigrama é a freqüência de toda composição de duas palavras consecutivas presentes nos textos.

Resultados e Discussão

Foram coletadas um total de 474 frases do conjunto de 100 LM de colonoscopia, das quais 412, 396 e 393 compuseram o CFU1, CFU2 e CFU3 respectivamente, promovendo uma redução de 13,08%, 16,45% e 17,08% em relação ao total de frases. Ao aplicar a técnica de lematização no CFU2, foi gerado outro CFU contendo 393 frases, observando uma redução de 17,08%, isto é, mesma percentagem do CFU3, porém contendo frases mais legíveis. Do total de 1812 bigramas e 2321 trigramas, 129 e 52 respectivamente apresentaram freqüência maior ou igual a cinco.

Conclusões

Em uma análise conjunta com especialistas da área e comparativa em relação aos estudos realizados em LM de EDA, foi possível constatar que as informações de colonoscopia apresentam baixa uniformidade de descrição, dificultando identificação dos termos que compõem o domínio e a definição de uma lógica de mapeamento. Portanto, justifica-se a utilização de outras técnicas como lematização e n-gramas para auxiliar na construção do dicionário e posteriormente mapear as informações para a BD.

Agradecimentos

Agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq – pela concessão da bolsa de iniciação científica e à Fundação Parque Tecnológico Itaipu – FPTI-BR.

Referências

1. U. Fayyad; G. Piatetsky-Shapiro; P. Smyth. *AI Magazine.* 1996, 37-54.
2. D. Pyle. *Data Preparation for Data Mining*, Morgan Kaufmann, Califórnia, 1999.
3. D. D. F. Honorato; H. D. Lee; M. C. Monard; F. C. Wu; R. B. Machado; A. P. Neto; C. A. Ferrero in Anais do V Encontro Nacional de Inteligência Artificial, XXV CSBC, Porto Alegre, 2005, 593-601.
4. E. A. Cherman; H. D. Lee; C. A. Ferrero; D. D. F. Honorato; R. B. Machado; F. C. Wu in Anais do XIV Simpósio Internacional de Iniciação Científica da USP, São Paulo, 2006.
5. F. A. Quilici; E. C. Grecco. *Colonoscopia*, Lemos-Editorial, São Paulo, 2000.