UMA METODOLOGIA DE MAPEAMENTO DE LAUDOS ENDOSCÓPICOS PARA BASES DE DADOS ESTRUTURADAS: ESTUDO DE CASO.

Newton Spolaôr (PDTA/FPTI-BR), Huei Diana Lee (Orientadora), Everton Alvares Cherman (PIBIC/CNPq-UNIOESTE), Daniel de Faveri Honorato, João José Fagundes, Juvenal Ricardo Navarro Góes, Feng Chung Wu (Co-orientador), e-mail: newtonspolaor@hotmail.com.

Universidade Estadual do Oeste do Paraná/Centro de Engenharia e Ciências Exatas – Foz do Iguaçu – PR.

Palavras-chave: descoberta de conhecimento, mapeamento de laudos médicos, endoscopia digestiva alta.

Resumo:

A evolução da tecnologia tem permitido um aumento considerável na quantidade de dados armazenados. Processos como o de Descoberta de Conhecimento em Base de Dados podem ser utilizados para auxiliar na extração de conhecimento a partir desses dados. Neste trabalho é apresentado um estudo de caso de uma metodologia para o mapeamento de laudos médicos de Endoscopia Digestiva Alta, referentes ao duodeno, para Bases de Dados.

Introdução

Na área médica, assim como em diversas outras áreas, a utilização cada vez mais freqüente de tecnologia propicia o aumento na quantidade de dados armazenados. Nesse contexto, para que uma análise mais completa desses dados possa ser realizada, torna-se necessária a aplicação de métodos computacionais. Um processo que pode contribuir para essa tarefa é o de Descoberta de Conhecimento em Base de Dados, o qual é constituído de três etapas: pré-processamento, mineração de dados e pós-processamento [1]. O pré-processamento, etapa mais custosa do processo, é responsável por, entre outras tarefas, padronizar os dados em um formato atributo-valor, tornando-os adequados à aplicação posterior da mineração de dados. O objetivo deste trabalho é apresentar um estudo de caso, baseado em uma metodologia automática de mapeamento de Laudos Médicos – LM – de Endoscopia Digestiva Alta – EDA, relacionados ao duodeno, para uma BD estruturada [2].

Materiais e Métodos

A metodologia proposta foi aplicada nos LMs de EDA, exame amplamente praticado devido à alta incidência de doenças gastroduodenais na população mundial [3]. Neste trabalho foram utilizados 609 LMs, sem identificação do

paciente, fornecidos pelo Serviço de Endoscopia Digestiva do Hospital Municipal de Paulínia. A metodologia aplicada é constituída por duas fases: 1) construção de dicionário do domínio e 2) preenchimento da BD.

A primeira fase, é composta por duas etapas: identificação de padrões nos LMs e construção do dicionário. A primeira etapa subdivide-se em quatro tarefas iterativas e interativas: identificação de frases únicas, construção do Arquivo de Padronização – AP, Remoção de *Stopwords* – RS – e Aplicação de Stemming – AS. Inicialmente, são extraídas dos LMs frases únicas, para se obter o primeiro Conjunto de Frases Únicas – CFU1. Posteriormente, são realizadas reuniões com especialistas do domínio para identificar padrões existentes no CFU1, os quais constituem a base para o AP. Esses padrões permitem substituir expressões textuais distintas que possuam semântica similar, ou que estejam em uma disposição inadequada, por expressões pré-definidas pela metodologia. Ao aplicar a RS, são eliminadas dos LMs palavras do domínio dispensáveis segundo o especialista (*stopwords*), conjunções, preposições e artigos. A aplicação do AP em conjunto com a RS reduz o conjunto de frases, gerando o segundo Conjunto de Frases Únicas – CFU2, o que permite refinar o AP com novos padrões. A etapa de AS mantém apenas os radicais das palavras, originando o terceiro Conjunto de Frases Únicas – CFU3.

Na segunda etapa, um dicionário do domínio é construído, baseado na disposição dos LMs de EDA, onde cada frase descreve uma ou mais estruturas anatômicas, características referentes a estas estruturas, e detalhes acerca destas características, o que corresponde, respectivamente, aos locais, características e subcaracterísticas presentes no dicionário [4]. A integração entre o AP, CFU2 e CFU3 com as avaliações dos especialistas possibilita definir os atributos da BD e estruturar o dicionário. Na Figura 1 é apresentada a representação esquemática da primeira fase.

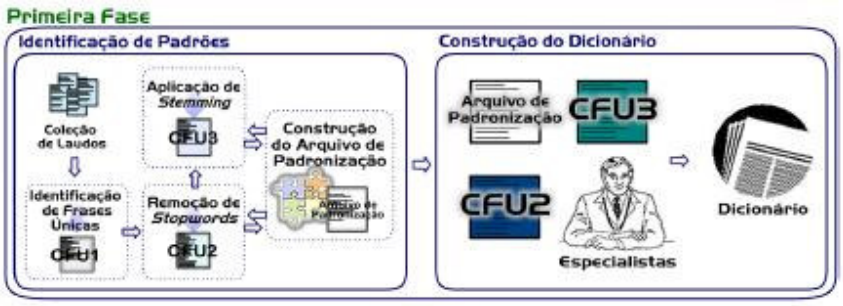

Figura 1 – Esquematização das duas etapas da primeira fase [4].

Na segunda fase, um conjunto de LMs de EDA pode ser processado pela metodologia, a partir do dicionário construído na fase anterior. Um LM é extraído do conjunto e fornece cada uma das suas frases ao mapeamento, conforme as estruturas descritas no dicionário. Se uma palavra da frase analisada for definida como local no dicionário, características relacionadas a

esse local são buscadas na frase. Caso essas associações forem validadas pelo dicionário, é realizada a pesquisa e validação para as possíveis subcaracterísticas dessas características. Ao final do processamento de um LM, as associações encontradas definem os atributos e valores de um registro a ser armazenado na BD, e o processo recomeça até mapear todo o conjunto na BD.

Resultados e Discussão

O conjunto de 609 LMs possuía 1710 frases referentes ao duodeno, que foram reduzidas em 94,74% após a geração do CFU1. Após aplicar AP e RS, o CFU2 era composto por 88 frases, proporcionando uma redução de 94,85% em relação às frases originais e 2,22% em relação ao CFU1. A BD, formada por 51 atributos, apresentou 30,69% de preenchimento, sendo que 94,49% dos dados se concentravam em 16 atributos, apresentando uma alta homogeneidade. Após análise detalhada, observou-se que os dados ausentes na BD também não estavam presentes nos LM. Portanto, a metodologia mapeou 100% dos dados descritos corretamente nos LMs para a BD. O detalhamento com subcaracterísticas tornou o dicionário robusto, permitindo armazenar na BD diagnósticos mais precisos e adequados à mineração de dados.

Conclusões

Neste trabalho foi apresentado um estudo de caso de uma metodologia para mapear LMs de EDA, referentes ao duodeno, para uma BD. Um dicionário foi construído, e poderá ser útil para processar outros conjuntos de LMs. Como trabalho futuro, planeja-se estudar a possibilidade da mineração de dados obter inferências corretas apenas com os atributos com alto preenchimento.

Agradecimentos

Ao Programa de Desenvolvimento Tecnológico Avançado – PDTA/FPTI-BR – pela concessão de bolsa de iniciação científica.

Referências

1. U. Fayyad; G. Piatetsky-Shapiro; P. Smyth. *AI Magazine.* 1996, 37-54.
2. 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.
3. F. Cordeiro; J. S. M. Filho; J. C. Prolla. *Endoscopia Digestiva*, MEDSI, Rio de Janeiro, 1994.
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.