UM ESTUDO DE CASO DA APLICAÇÃO DE UMA METODOLOGIA DE MAPEAMENTO DE LAUDOS MÉDICOS PARA BASE DE DADOS

EVERTON ALVARES CHERMAN (UNIOESTE) , HUEI DIANA LEE , CARLOS A. FERRERO , DANIEL DE FAVERI HONORATO , WU FENG CHUNG .

evertoncherman@hotmail.com - UNIOESTE

O avanço computacional possibilita um grande crescimento das bases de dados, assim, ferramentas para a análise desses dados se mostram úteis e necessárias. Um dos processos que auxilia nessa análise é o de Descoberta de Conhecimento em Bases de Dados, o qual é composto de três etapas: pré-processamento, mineração de dados e pós-processamento. Neste trabalho é apresentado um estudo de caso da aplicação de uma metodologia de pré-processamento em laudos médicos de Endoscopia Digestiva Alta, mais especificamente informações relacionadas ao estômago, nos quais os dados são mapeados para um formato apropriado para a análise denominado atributo-valor. Observou-se que 87,4% dos registros da base de dados não foram preenchidos devido à grande quantidade de atributos definidos pelo especialista, os quais constavam com pouca freqüência nos laudos.

descoberta de conhecimento; pré-processamento de dados; mapeamento de dados; laudos médicos; base de dados;

Introdução

O avanço tecnológico atual possibilita grande facilidade para o armazenamento de dados. Na área médica, hospitais e clínicas registram um volume cada vez maior de informações, usualmente pouco estruturadas, sobre pacientes e exames laboratoriais. Nesse contexto, torna-se custoso coletar, analisar e extrair informações adicionais que poderiam, por exemplo, auxiliar especialistas no diagnóstico de doenças.

Desse modo, surge a necessidade de ferramentas computacionais para a análise desse grande volume de dados. Um dos processos que auxilia nessa tarefa é o de Descoberta de Conhecimento em Bases de Dados - DCDB - o qual é composto basicamente de três etapas: pré-processamento, mineração de dados e pós-processamento [1,2]. O pré-processamento é etapa que consome o maior tempo computacional no processo, em torno de 80% do total. Nessa etapa são realizadas tarefas como preparação, redução e formatação dos dados. A etapa de mineração de dados tem como objetivo construir modelos capazes de identificar padrões nos dados. Na última etapa os modelos construídos são avaliados e validados pelos especialistas.

A metodologia aplicada neste trabalho tem como objetivo semi-automatizar o mapeamento das informações contidas em Laudos Médicos - LM - de Endoscopia Digestiva Alta - EDA - para Base de Dados - BD - em um formato adequado, denominado atributo-valor, evitando a subjetividade do preenchimento manual, além de possibilitar posterior automatização do mapeando de novos LMs. Neste trabalho a metodologia foi aplicada ao mapeamento de informações de EDA referentes ao estômago.

Material e Métodos

O EDA é um exame importante para o diagnóstico de doenças gastroduodenais, como úlceras e gastrites, os quais têm grande incidência na população [3]. Neste trabalho foram utilizados 610 LMs de EDA, os quais não contêm identificação dos pacientes, fornecidos pelo Serviço de Endoscopia Digestiva do Hospital Municipal de Paulínia.

A metodologia utilizada para o mapeando das informações contidas nos LMs é composta por duas fases: 1) construção de um dicionário do domínio do conhecimento e 2) preenchimento da BD a partir dos LMs [4,5].

Primeira Fase

A primeira fase é constituída por 2 etapas: identificação de padrões nos LMs e construção do dicionário.
A primeira etapa da primeira fase é constituída por quatro tarefas, as quais auxiliam no processo de identificação de padrões: identificação de frases únicas, construção de um Arquivo de Padronização - AP -, Remoção de *Stopwords - RS* - e Aplicação de S*temming - AS*. Na primeira tarefa são coletadas todas as frases contidas nos LMs e apenas um exemplar de cada frase é mantido, formando um primeiro conjunto de frases únicas - CFU1. A partir de CFU1 é possível, com o auxílio dos especialistas, construir o AP para os LMs. Esse arquivo é necessário devido à freqüente utilização de termos distintos que expressam informações semelhantes e também informações descritas de maneira não apropriada para a aplicação da metodologia. Na tarefa RS o objetivo é eliminar dos LMs todas as palavras que não têm relevância no contexto definido para o problema em questão, como conjunções, artigos e preposições [5]. É fundamental, nessa tarefa, uma nova interação com os especialistas, para identificar também *Stopwords* específicas do domínio do conhecimento, ou seja, termos médicos que sendo retirados não modificam a interpretação desejada. Desse modo, ao finalizar a tarefa de RS, um segundo Conjunto de Frases Únicas é construído - CFU2 -, facilitando a identificação de padrões nas informações e o refinamento do AP.
A tarefa de AS permite a eliminação das diferentes inflexões de palavras, mantendo apenas um radical comum a elas. Essa tarefa tem como objetivo auxiliar na remoção de redundâncias e permitir que o AP seja novamente atualizado. Concluída essa tarefa, é construído o terceiro Conjunto de Frases Únicas - CFU3. Na Figura 1 é apresentado um exemplo de uma frase após RS, padronização e AS:

Frase	**- Mucosa de corpo de aspecto friável, diminuição da distensibilidade**
RS	**corpo friavel diminuicao distensibildade**
AP	**corpo friavel anormal**
AS	**corp friav anorm**

Figura 1 - Exemplo de RS, AP e AS.

Na segunda etapa da primeira fase é realizada a construção de um dicionário do domínio do conhecimento. A estrutura do dicionário é baseada na disposição das informações dos LMs, onde cada frase dos LMs é composta por uma ou mais estruturas anatômicas (locais) e suas características. Desse modo, a estrutura do dicionário é composta por todos os possíveis locais e suas características presentes nos LMs. Para a construção do dicionário é necessário que, com o auxílio dos especialistas, sejam identificados os atributos que irão compor a BD. Os arquivos CFU2, CFU3 e AP auxiliam nesse processo interativo. Na Figura 2 é ilustrada a primeira fase da metodologia.

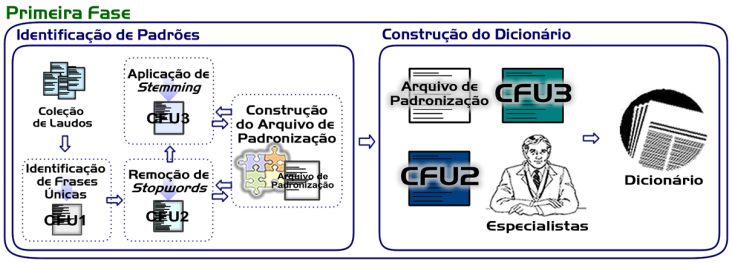

Figura 2 - Ilustração da Primeira Fase.

Segunda Fase

Na segunda fase é realizado o processamento dos LMs com base na estrutura do dicionário definido e construído na fase anterior. Desse modo, é preenchida a BD, na qual cada LM corresponde a um exemplo desse conjunto de dados.
O processamento é executado em ciclos iterativos, nos quais é obtido um LM e uma frase é retirada. Cada palavra contida nessa frase é classificada conforme as características descritas no dicionário. Se essa palavra pertencer a um local definido no dicionário, inicia-se,

na frase, a busca pelas características referentes àquele local. Se o relacionamento entre local e característica for confirmado pelo dicionário, então os atributos da BD são preenchidos com os valores correspondentes. O processo é repetido para todas a frases dos LMs.

Resultados e Discussão

Foram coletadas do conjunto de 610 LMs de EDA um total de 5213 frases relacionadas ao estômago. Esse conjunto foi processado e 348 frases únicas, isto é, 6,7% do total de frases, compuseram CFU1. Após RS e AP, deu-se origem ao CFU3, que por sua vez, representou 5,4% do total de frases, e uma diminuição de 19,2% em relação ao CFU1. Para efeito de comparação com estatísticas de trabalhos anteriores, o CFU3 com informações relacionadas ao esôfago continha 24,14% menos frases que o seu CFU1.
Todas as informações contidas nos LMs descritas corretamente foram mapeadas integralmente para a BD. Nesse trabalho, constatou-se, porém, que 87,4% dos atributos na BD não foram preenchidos devido ao grande número de atributos definidos pelo especialista para BD de estômago, onde mesmo atributos que constam com pouca freqüência nos LMs foram definidos para o mapeamento. Observou-se também a necessidade de correções ortográficas em algumas frases, o que permitiria o mapeamento mais completo das informações para a BD.

Conclusões

Neste trabalho foi apresentado um estudo de caso da aplicação de uma metodologia para mapeamento de LMs de EDA com informações referentes ao estômago. Os resultados mostram que, apesar das informações contidas nos LMs não terem sido mapeadas completamente para BD devido ao formato inadequado da descrição de algumas informações nesses laudos, foi criado um dicionário do domínio do conhecimento capaz de mapear novos conjuntos de LMs, diminuindo o tempo gasto no processo de preenchimento da BD e evitando a subjetividade do preenchimento manual.
Como trabalhos futuros pretende-se construir o dicionário para mapeamento das informações de LMs de EDA relacionadas ao duodeno, aplicar a metodologia para LMs de colonoscopia e definir um maior número de atributos para o dicionário de estômago, provendo maior detalhamento das informações. Pretende-se também desenvolver ferramentas de correção ortográfica, para auxiliar no mapeamento completo dos LMs.

Agradecimentos

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

Referências Bibliográficas

1. U. Fayyad; G. Piatetsky-Shapiro; P. Smyth in Second International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA; 1996. p. 82-88.

2. S. O Rezende. *Sistemas Inteligentes: Fundamentos e Aplicações*. Editora Manole, Barueri, 2003.

3. R. Pellicano; S. Fagoonee; G. Palestro; M. Rizzetto; N. Figura; A. Ponzetto. The diagnosis of *Helicobacter* pylori infection: guidelines from the maastricht 2-2000 consensus report. Minerva Gastroenterol Dietol, 2004, vol. 50(2):125-33.

4. 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 Congresso da Sociedade Brasileira de Computação, Porto Alegre, 2005, pages 593-601.

5. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys. 2002*, 34(1):1-47.