III Jornada Científica da Unioeste

CONSTRUÇÃO DE UMA BASE DE DADOS ESTRUTURADA A PARTIR DE LAUDOS MÉDICOS DE ENDOSCOPIA DIGESTIVA ALTA[1]

DANIEL DE FAVERI HONORATO[2], HUEI DIANA LEE[3], FENG CHUNG WU[4], RENATO BOBSIN MACHADO[5], MARIA CAROLINA MONARD[6], CARLOS ANDRÉS FERRERO[7]

Escrito para apresentação na III JORNADA CIENTÍFICA DA UNIOESTE
15 a 17 de junho de 2005 - Unioeste - PRPPG - Campus de Marechal Cândido Rondon - PR

RESUMO: Com o crescente desenvolvimento tecnológico, as bases de dados estão tornando-se cada vez maiores. Para auxiliar na análise e na compreensão dessa grande quantidade de dados presentes nas bases de dados, processos apoiados pela área da computação, tal qual a de Descoberta de Conhecimento em Bases de Dados, tem sido desenvolvidos. Para que esse processo possa ser realizado, é necessário que os dados estejam representados no formato atributo-valor. Neste trabalho, é apresentada uma metodologia para auxiliar no processo de semi-automatização de construção de uma base de dados nesse formato a partir de informações presentes em laudos médicos descritas em linguagem natural. As informações contidas nesses laudos estão relacionadas à Endoscopia Digestiva Alta.

PALAVRAS-CHAVE: laudos médicos semi-estruturados, dicionário, metodologia.

ABSTRACT: With the increasing development of technology, databases are growing very fast. To give assistance during the analysis and understanding of data stored in these databases, processes like Knowledge Discovery in Databases have been proposed. In order to perform this process, data usually needs to be represented in the attribute-value format. In this work, it is presented a methodology to help in the semi-automatization of building a database in the attribute-value format from information contained in medical findings described in natural language. These medical findings are related to High Digestive Endoscopies.

KEY WORDS: semi-structured medical findings, dictionary, methodology.

INTRODUÇÃO: O processo de Descoberta de Conhecimento em Base de Dados (DCBD)[8] (FAYYAD et al., 1996) auxilia na análise e compreensão de informações armazenadas em bases de dados e, para que possa ser aplicado, é necessário que as informações estejam mapeadas, geralmente, no formato atributo-valor. Nesse contexto, hospitais e clínicas médicas registram grande quantidade de informações de pacientes e processos laboratoriais, as quais, freqüentemente, são armazenadas de maneira semi-estruturada e descritas em linguagem natural. Desse modo, para que possa se aplicado o processo de

---

[1] Projeto de pesquisa desenvolvido no Laboratório de Bioinformática – LABI, Centro de Engenharias e Ciências Exatas, UNIOESTE, Foz do Iguaçu-PR, Avenida Tarquínio Joslin dos Santos, 1300, Caixa Postal 961 CEP 85870-900, Foz do Iguaçu, PR Tel: 45 3576-8114.

[2] Bolsista de Iniciação Científica do Parque Tecnológico Itaipu (PTI) e pesquisador estagiário do LABI; E-mail: dfaverih@hotmail.com.

[3] Profa. Mestre do CECE da UNIOESTE, Foz do Iguaçu-PR; Coord. geral do LABI; Doutoranda, ICMC-USP, São Carlos-SP.

[4] Prof. Doutor da Coloproctologia da UNICAMP, Campinas-SP; Coord. da área médica do LABI.

[5] Prof. Mestre do CECE da UNIOESTE, Foz do Iguaçu-PR; Coord. da área computacional do LABI.

[6] Profa. Titular do Instituto de Ciências Matemáticas e de Computação da USP, São Carlos-SP.

[7] Bolsista de Iniciação Científica do Instituto de Tecnologia e Automação em Informática-ITAI e estagiário do LABI.

[8] *KDD - Knowledge Discovery in Databases.*

DCBD, são necessárias a interpretação e a transformação dessas informações para formato atributo-valor. Esse processo, além de ser custoso, está sujeito à interpretação subjetiva de quem o está realizando (LEE, 2000). Sendo assim, processos para auxiliar na semi-automatização dessa tarefa, poderiam prover ganho em tempo, além de promover a padronização no tratamento das informações contidas em laudos médicos. Neste trabalho é apresentada uma metodologia para dar suporte à construção de bases de dados estruturadas a partir de laudos médicos de Endoscopia Digestiva Alta – EDA, semi-estruturados, descritos em linguagem natural (HONORATO et al., 2004). Nessa metodologia é construído um dicionário, com o auxílio de especialistas do domínio, a partir da identificação de padrões que ocorrem nos laudos. Esse dicionário é então utilizado com o intuito de mapear os laudos médicos, por meio de casamento de padrões, para conjuntos de dados no formato atributo-valor.

MATERIAIS E MÉTODOS: Doenças pépticas gastroduodenais representam uma das entidades patológicas de maior incidência na população, despertando cada vez mais interesse na pesquisa dessa área (PELLICANO et al., 2004). Conforme citado, neste trabalho[9] é apresentada a construção de uma base de dados a partir de uma coleção de 100 laudos de EDA, na qual não consta a identificação dos pacientes. Esses laudos armazenam informações relacionadas ao exame de EDA realizados no período de março a novembro de 2001 no Serviço de Endoscopia Digestiva do Hospital Municipal de Paulínia. Os laudos armazenam informações descritas em linguagem natural relacionadas ao esôfago, estômago, duodeno e conclusões do exame (Figura 1). Neste trabalho, foram consideradas apenas informações relacionadas ao esôfago.

```
* ESÔFAGO
- Mucosa de terço distal com presença de erosões lineares, confluentes.
- Calibre e distensibilidade normais.
- Motilidade normal.
- TEG situada a aproximadamente 2,0 cm acima do pinçamento diafragmático.
* ESTÔMAGO
- Cardia aberto à retrovisão.
- Mucosa de fundo de aspecto normal.
- Mucosa de corpo de aspecto normal.
- Incisura angularis normal.
- Mucosa de antro com áreas de enantema.
- Motilidade normal.
- Lago mucoso claro.
- Piloro centrado, pérvio.
* DUODENO
- Bulbo amplo, sem lesões.
- Segunda (2ª.) porção normal.
*BIÓPSIA:( x )SIM - (pesquisa H. Pylori)  ( )NÃO
**CONCLUSÃO**: - Esofagite Erosiva Grau II  -  1/4.
                - Gastrite enantemática de Antro (Leve Intensidade) .
```

Figura 1: Exemplo de laudo.

A metodologia proposta é composta por duas fases (Figura 2). A primeira, caracteriza-se pela construção de um dicionário do domínio do conhecimento considerado, na qual, o auxílio do especialista é de fundamental importância. Na segunda fase, esse dicionário é utilizado para a transformação de laudos médicos desse domínio por meio de casamento de padrões na construção da base de dados no formato atributo-valor. A construção do dicionário é feita por meio de quatro etapas iterativas e interativas: 1 - identificação de frases

[9] Neste trabalho, os algoritmos foram implementados utilizando o paradigma de orientação a objetos em Perl (SCHWARTZ et al., 1997).

únicas; 2 - construção de arquivo de padronização; 3 - remoção de *stopwords* e aplicação de *stemming* e 4 - construção da base de conhecimento do dicionário. As três primeiras etapas têm como objetivo auxiliar no processo de identificação dos padrões contidos nos laudos para que esses possam ser mapeados para o dicionário. Na primeira etapa são identificadas as frases únicas existentes na coleção de laudos. As informações presentes nos laudos são mapeadas por meio de frases, onde cada frase refere-se a um diagnóstico, um prognóstico ou uma observação do médico sobre o exame realizado. As frases contidas em cada laudo são coletadas em um único documento e organizadas em ordem alfabética. Esse processo permite reunir frases repetidas, uma vez que a mesma frase está, freqüentemente, presente em diversos laudos. Essas frases repetidas são removidas e apenas um exemplar de cada frase é mantido. Ao final dessa etapa, obtém-se como resultado um primeiro conjunto de frases únicas – CFU1 – relacionado à coleção de laudos. Após a construção de CFU1, inicia-se a segunda etapa por meio da construção do arquivo de padronização, a qual é realizada a medida são identificadas as informações que podem ser padronizadas. Esse processo continua até o final da primeira fase da metodologia proposta. A padronização das informações é necessária, pois é freqüente a utilização de sinônimos na descrição de informações semelhantes presentes nos laudos médicos ou a presença de frases que expressam informações de maneira diferente da que será utilizada pelo dicionário.

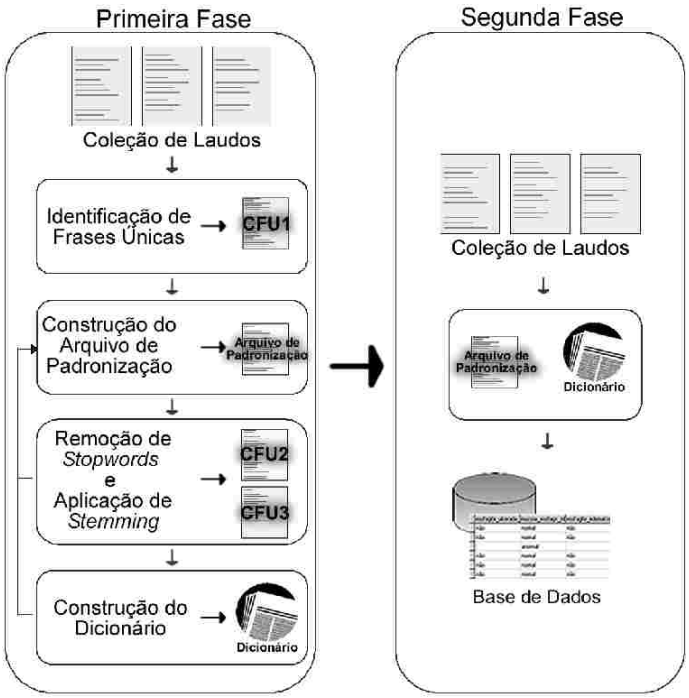

Figura 2: Metodologia proposta.

Na Tabela 1 são apresentados dois exemplos de padronização, no contexto deste trabalho.

Antes da padronização	Depois da padronização
coloração esbranquiçada	anormal
calibre e distensibilidade normais	calibre normal distensibilidade normal

Tabela 1: Exemplo de padronização.

Na segunda linha é apresentada uma palavra composta sem padronização e a respectiva palavra padronizada pelo especialista. Na terceira linha é apresentada uma frase, a qual

ISBN 857644020-2

depois de padronizada transforma-se em outras duas frases. A partir de CFU1 é possível identificar parte das informações que poderão ser padronizadas. Posteriormente, a aplicação da padronização permitirá que as informações contidas nos laudos estejam mapeadas em um formato padrão para ser utilizado pelo dicionário e pelo processo de preenchimento da base de dados durante a segunda fase da metodologia proposta. Paralelamente à construção do arquivo de padronização, ocorre a identificação de padrões nas informações contidas em CFU1. Para auxiliar no desenvolvimento dessa tarefa, é realizada a remoção de *stopwords* (REZENDE, 2003) e a aplicação de *stemming* (SEBASTIANI, 2002), as quais têm como objetivo auxiliar, no processo de identificação dos padrões utilizados pelos especialistas nos laudos. Para tanto, é realizada a remoção de *stopwords* sobre o CFU1, gerando CFU2. Depois dessas etapas, realiza-se a aplicação de *stemming*, o qual faz a remoção de variações morfológicas das palavras sinalizando as frases redundantes e possibilitando a redução da dimensão de CFU2. Com isso, o CFU3 é originado. Esse último, construído a partir de CFU2, é utilizado em dois momentos: primeiro - ajudar o especialista, durante a análise das frases únicas na identificação de padrões e segundo - auxiliar na decisão de como as informações serão organizadas (padronizadas) na construção do dicionário. Como mencionado anteriormente, o dicionário auxilia na construção da base de dados, isto é, no preenchimento dos valores dos atributos da base de dados utilizando as informações contidas nos laudos. Desse modo, antes de iniciar a construção do dicionário, é necessário, com a ajuda dos especialistas, definir quais atributos vão compor a base de dados. Depois de identificados os atributos, cria-se a estrutura da base de dados que receberá informações a partir do processamento dos laudos utilizando o dicionário. Assim, a construção do dicionário é realizada, conjuntamente com o especialista do domínio, com base nas informações existentes em CFU2 e CFU3 e no arquivo de padronização. Laudos médicos de diversas especialidades possuem informações organizadas na forma de estrutura anatômica e característica associadas a essa estrutura. Desse modo, segundo a metodologia proposta, a estrutura base do dicionário é composta pela localização anatômica (locais) e as suas respectivas características. Por exemplo, a partir da frase “terço distal com erosões”, é mapeado o local (terço distal) e na seqüência a característica (com erosões). Esse procedimento é repetido até que todas as informações relacionadas ao exame tenham sido mapeadas para a estrutura do dicionário. Na Figura 3 é ilustrada a estrutura base do dicionário.

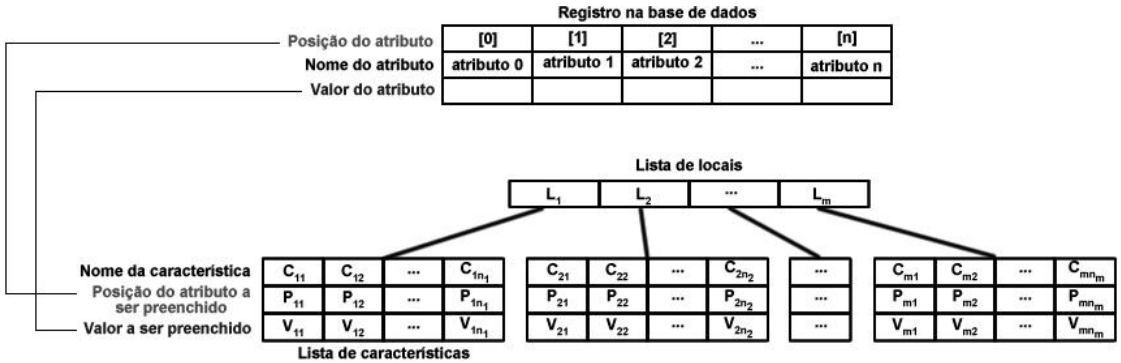

Figura 3: Estrutura base do dicionário.

Conforme é apresentada na Figura 3, a lista de locais armazena o nome de um determinado local e, cada local, possui uma lista de uma ou mais características associadas. A lista de características armazena, além do nome da característica, o número correspondente à posição do atributo no registro na base de dados – RBD – e o valor que deverá ser armazenado no atributo correspondente do RBD. Na segunda fase, a coleção de laudos é processada, com base nas informações mapeadas na estrutura do dicionário (locais e características) e os valores dos atributos presentes na estrutura do RBD são preenchidos.

O processo é ilustrado na Figura 4. Cada laudo corresponde a um exemplo na base de dados no formato atributo-valor. O processo de armazenamento – PA – recebe como entrada um laudo, no qual previamente foi aplicado o processo de padronização, e uma frase é extraída. A execução do PA é realizada por meio de ciclos de interações de pesquisa entre a estrutura do dicionário e a frase extraída do laudo. Primeiramente, é verificada a existência do primeiro local da lista de locais do dicionário na frase extraída. Se estiver presente, cada uma das características associadas a esse local é pesquisada na frase sob análise e as informações associadas às características encontradas são armazenadas no RBD por meio da verificação, na estrutura do dicionário, da posição do atributo no qual deverá ser armazenado. O mesmo procedimento é repetido para todos os locais e suas respectivas características até que os locais descritos no dicionário tenham sido pesquisados integralmente na frase corrente. Esse processo é repetido até que todas as frases do laudo tenham sido completamente processadas. Ao final, o RBD, preenchido com as informações desse laudo, é inserido na base de dados e uma nova iteração é iniciada com o processamento do próximo laudo.

RESULTADOS E DISCUSSÃO: Neste trabalho foram utilizadas informações sobre o esôfago de uma coleção de 100 laudos de EDA. A partir dessa coleção de laudos, foi construída uma base de dados com informações no formato atributo-valor, utilizando-se da metodologia proposta. Primeiramente foi realizada a identificação de frases únicas existentes na coleção de laudos, a qual resultou no CFU1 preenchido com apenas um exemplar de cada frase totalizando 23 frases.

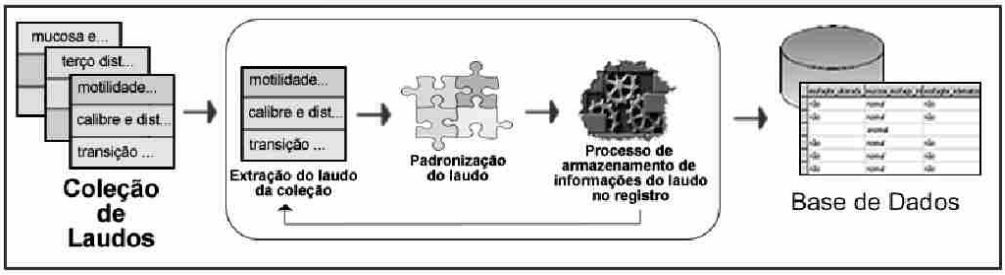

Figura 4: Construção da base de dados.

Em seguida, iniciou-se a construção do arquivo de padronização, utilizando como base as informações contidas no CFU1. Depois da construção de uma primeira versão do arquivo de padronização, foi realizada a remoção de stopwords, juntamente com o especialista, o qual indicou, além das palavras usuais como preposições, artigos e conjunções, algumas palavras do domínio que poderiam ser adicionadas à lista de stopwords mantendo o significado do texto presente no laudo. Após, foi realizada a aplicação de *stemming* sobre o CFU2, o qual possibilitou que fossem visibilizadas frases redundantes que poderiam ser removidas, por exemplo, as frases “terço distal erosões” e “terço distal erosão” foram transformadas em “terc dist eros” e “terc dist eros”, respectivamente. Desse modo, após aplicação de *stemming* foram removidas as frases redundantes e obteve-se o CFU3 com 18 frases. O algoritmo de *stemming* utilizado neste trabalho é baseado no algoritmo de Porter adaptado para a língua portuguesa (ORENGO e HUYCH, 2001). Uma avaliação do Conjunto de Frases Únicas, antes e depois da remoção de *stopwords* e da aplicação de *stemming*, mostrou que houve uma redução de 23 frases nesse conjunto inicial para 18 frases após a realização dessas tarefas, representando uma redução de 21,7%. A existência de 23 frases diferentes em um conjunto de 100 laudos mostra que existe certa uniformidade na maneira como os laudos são descritos pelos especialistas, especialmente para o domínio em questão. Esse fato é confirmado depois da realização das tarefas citadas anteriormente, na qual apenas cinco frases foram consideradas redundantes. Posteriormente, o arquivo de

ISBN 85-7644-020-2

padronização foi incrementado com novas informações obtidas após essa etapa. Com o auxílio dos especialistas e a utilização de CFU2 e CFU3, foi realizada a definição de quais atributos fariam parte da base de dados e quais locais e características estariam presentes na estrutura do dicionário. Foi decidido que as informações referentes ao esôfago poderiam ser mapeadas por 16 atributos. Com o dicionário estruturado, iniciou-se a segunda fase da metodologia proposta: a construção da base de dados, no formato atributo-valor, por meio do mapeamento das informações contidas nos laudos para os registros da base dados. É importante ressaltar que cada laudo correspondia a um registro na base de dados. Uma análise dos resultados após o processamento dos laudos para a construção da base de dados, mostrou que do total de 100 laudos mapeados em 100 registros, apenas 14 deles não tiveram todos os atributos preenchidos. A avaliação dos laudos relacionados a esses 14 registros, juntamente com os especialistas, demostrou que as informações que não haviam sido preenchidas também não tinham sido informadas nos laudos. No contexto deste trabalho, o resultado foi considerado muito bom. Após análise junto ao especialista do domínio, constatou-se que a metodologia desenvolvida atendeu aos requisitos estabelecidos.

CONCLUSÕES: Neste trabalho foi apresentada uma metodologia para a semi-automatização do processo de mapeamento de laudos médicos em bases de dados apropriadas para a extração automática de conhecimento. Foi apresentado também um estudo de caso aplicando a metodologia desenvolvida a uma coleção de laudos de exames de Endoscopia Digestiva Alta. Depois de construída a base de dados, os resultados foram analisados e a metodologia considerada adequada de acordo com o objetivo proposto. A construção do dicionário proporcionou uma diminuição do custo de tempo usado na fase de preparação dos dados, uma vez que, manualmente, seria necessário um maior envolvimento do especialista, e a padronização do mapeamento dos laudos para a base de dados estaria sujeita a uma excessiva interpretação subjetiva. Além disso, a metodologia proposta poderá ser utilizada na construção de outros dicionários para o mapeamento de informações em outras bases de dados. Outro aspecto importante é que, após a construção do dicionário, novos laudos podem ser facilmente mapeados para a base de dados no formato atributo-valor. Como trabalho em andamento, essa metodologia está sendo utilizada para extrair informações relacionadas ao estômago e ao duodeno, as quais estão contidas nos laudos utilizados neste trabalho. Além disso, a metodologia proposta está sendo aplicada em laudos de processamento de sêmen, nos quais temos enfocado na extração de conhecimento com auxílio de especialistas do domínio.

REFERÊNCIAS BIBLIOGRÁFICAS

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, páginas 37—54 vol. 17, 1996.

HONORATO, D. F.; LEE, H. D.; MACHADO, R. B.; WU, F. C.; NETO, A. P. Utilização da Indexação Automática para Auxílio à Construção de uma Base de Dados para a Extração de Conhecimento aplicada à Doenças Pépticas. I Workcomp Sul, Palhoça, 2004.

LEE, H. D.; MONARD, M. C.; ESTEVES, S. C. Indução Construtiva guiada pelo Conhecimento: um Estudo de Caso do Processamento de Sêmen Diagnóstico. **Open Track Discussion Proceedings**, páginas 157—166, 2000.

ORENGO, M.; HUYCH, C. A Stemming Algorithm for the Portuguese Language. **Spire**, 2001.

ISBN 85-7644-020-2

PELLICANO, R.; FAGOONEE, S.; PALESTRO, G.; RIZZETTO, M.; FIGURA, N.; PONZETTO, A. The diagnosis of helicobacter pylori: guidelines from the maastricht 2-2000 consensus report. **Minerva Gastroenterol Dietol***,* vol. 50(2): 125-33, 2004.

REZENDE, S. O. Sistemas Inteligentes: Fundamentos e Aplicações. Editora Malone, Barueri, SP, Brasil, 2003.

SCHWARTZ, R.; CHRISTIANSEN, T.; PYLE, L. W. Learning Perl, 1997.

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