# Um Estudo de Caso do Mapeamento de Laudos Endoscópicos para Bases de Dados

**Newton Spolaôr[1], Huei Diana Lee[1,2], Everton Alvares Cherman[1], Daniel De Faveri Honorato[2], João José Fagundes[3], Juvenal Ricardo Navarro Góes[3], Cláudio Sadi Rodrigues Coy[3], Feng Chung Wu[1,2,3]**

[1]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85869-970 – Foz do Iguaçu, PR, Brasil

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

[3]Faculdade de Ciências Médicas – Universidade Estadual de Campinas
Serviço de Coloproctologia
Caixa Postal 6111, 13083-970 – Campinas, SP, Brasil

newtonspolaor@hotmail.com, hueidianalee@gmail.com, evertoncherman@hotmail.com

***Abstract.*** *Technological advance has allowed a considerable increase in the amount of stored data. Processes as Knowledge Discovery in Databases can be used to assist the extraction of knowledge and analysis of these data. However, in order to apply such process, it is necessary to the data to be represented in appropriated formats, as the attribute-value one. In this work, it is presented a case study of a methodology that gives support to mapping of medical findings of High Digestive Endoscopy to structured databases in this format.*

***Resumo.*** *O avanço tecnológico tem permitido um aumento considerável na quantidade de dados armazenados. Processos como o de Descoberta de Conhecimento em Bases de Dados podem ser utilizados para auxiliar na extração de conhecimento a partir desses dados. Para aplicação deste processo, é necessário que os dados estejam dispostos em formatos adequados, como o atributo-valor. Neste trabalho é apresentado um estudo de caso de uma metodologia para o mapeamento de laudos médicos de Endoscopia Digestiva Alta para Bases de Dados nesse formato.*

## 1. Introdução

Na área médica, assim como em outras áreas, a utilização crescente de tecnologia incentiva um incremento na quantidade de dados armazenados. Desse modo, a análise manual se torna inviável, sendo necessário o apoio de métodos computacionais para que análises mais completas possam ser realizadas [Ferro et al. 2002, Lee and Monard 2003, Honorato et al. 2005, Lee 2005]. Processos como o de Descoberta de Conhecimento em Bases de Dados – DCBD, o qual apresenta natureza iterativa e interativa, podem contribuir para a aquisição de conhecimento a partir desses dados [Fayyad et al. 1996]. Esse

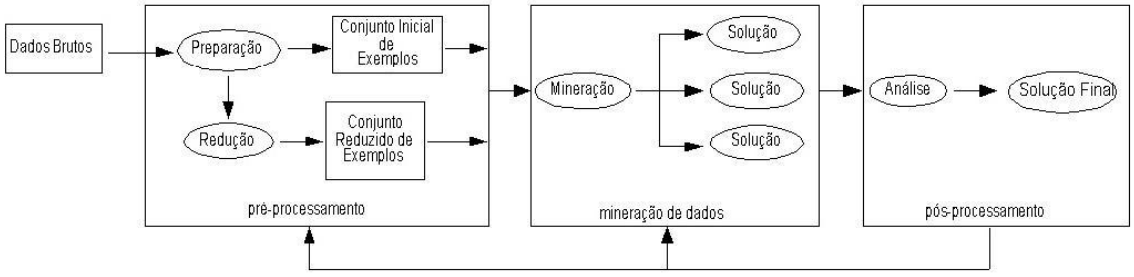


**Figura 1. Processo de Descoberta de Conhecimento em Banco de Dados [Baranauskas 2001].**

processo é constituído pelas etapas de pré-processamento, mineração de dados e pós-processamento e pode ser representado conforme a Figura 1.

O pré-processamento, etapa que consome cerca de 80% do tempo gasto no processo, apresenta como tarefas a transformação, a redução e a preparação dos dados. É também nessa etapa em que os dados são representados em formatos adequados para a mineração de dados, sendo o formato atributo-valor o mais utilizado [Pyle 1999]. Desse modo, o principal objetivo do pré-processamento é assegurar a qualidade dos dados que serão utilizados nas próximas etapas.

Na mineração de dados, dentre as diversas áreas que provêem suporte, podem ser aplicados algoritmos de inteligência artificial para a extração de padrões, a partir dos conjuntos de dados, possibilitando a construção de diferentes modelos. Existem diversos paradigmas para a representação desses modelos, dentre eles o simbólico, o qual inclui árvores e regras de decisão [Witten and Frank 2005].

Os modelos obtidos são avaliados e validados com o auxílio de especialistas do domínio na etapa de pós-processamento. Posteriormente, o conhecimento obtido pode auxiliar, por exemplo, o usuário no processo de tomada de decisão [Rezende 2003]. A qualquer momento no processo de DCBD é possível retornar as fases anteriores do processo para que, por exemplo, ajustes nos parâmetros dos algoritmos de construção de modelos ou novo pré-processamento nos dados sejam realizados.

Especificamente na área médica, dados como prognósticos e diagnósticos, observados em exames clínicos aplicados a um paciente, são freqüentemente registrados em língua natural, como a língua portuguesa, em documentos semi-estruturados denominados laudos médicos – LM. A organização dos dados que compõem os LMs, na maioria das especialidades médicas, apresenta forma hierárquica. Na Endoscopia Digestiva Alta – EDA, em geral os dados podem ser classificados em três categorias, as quais correspondem às estruturas anatômicas, características e subcaracterísticas relacionadas à um órgão do sistema digestivo. Essa estruturação possibilita a utilização de uma metodologia, baseada no processo de DCBD, de modo que as informações extraídas dos LMs possibilitem auxiliar os especialistas na elaboração de diagnósticos de doenças.

Este trabalho está inserido no projeto de Análise Inteligente de Dados aplicada a Doenças Pépticas [Honorato et al. 2005, Ferrero et al. 2005, Cherman et al. 2006, Honorato et al. 2007, Spolaôr et al. 2007]. É apresentado um estudo de caso em que a metodologia proposta neste projeto para mapeamento de dados de LMs para BDs estru-

turadas é aplicada à LMs de EDA. Nesse tipo de laudo, as informações são descritas em termos do esôfago, do estômago e do duodeno. A metodologia é aplicada a esses órgãos, mapeando os LMs de cada órgão para uma BD distinta.

Este trabalho está organizado do seguinte modo. Na seção 2 são apresentados a metodologia para mapeamento de LMs, bem como o conjunto de laudos de EDA considerado. Na seção 3 são descritos os resultados e a discussão, e as conclusões e trabalhos futuros são apresentados na seção 4.

## 2. Materiais e Métodos

A metodologia apresentada foi aplicada em LMs de EDA, o qual constitui um exame complementar amplamente utilizado devido à alta ocorrência de doenças esofagogastroduodenais na população mundial, como gastrites, lesões esofágicas e úlceras. A aplicação do exame de EDA é indicada para essas enfermidades e também recomendada quando existir sintomas digestórios persistentes, corpos estranhos no sistema digestório e alterações no equilíbrio metabólico [Cordeiro 1994]. As observações resultantes do exame, referentes ao esôfago, estômago, duodeno e conclusões obtidas, são registradas por meio de linguagem natural em LMs, conforme é demonstrado na Figura 2.

```
* ESÔFAGO
- Mucosa de terço distal de coloração esbranquiçada, com extensão mucosa gástrica (Barrett?).
- Calibre e distensibilidade normais.
- Motilidade normal.
- TEG situada a aproximadamente 4,0 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 enantema difuso.
- Motilidade normal.
- Lago mucoso claro.
- Piloro centrado, pérvio.
* DUODENO
- Bulbo amplo, sem lesões.
- Segunda (2ª.) porção normal.

*BIÓPSIA:  ( x )SIM - Esôfago  (  )NÃO
*UREASE: Positiva.
**CONCLUSÃO**: - Hérnia de hiato - 1/11.
- Esôfago Barrett ? - 1/21.
- Esofagite péptica grau C de Los Angeles - 1/4.
- Gastrite enantemática de antro (moderada intensidade) - 2/16.
```

**Figura 2. Representação de um LM de EDA.**

Os dados do LM são dispostos em uma estrutura fixa, na qual os domínios estão separados em blocos textuais. Esses blocos são distribuídos de acordo com a passagem do Endoscópio, ou seja, iniciando pelo esôfago, seguido pelo estômago e terminando na terceira porção do duodeno. Após registrar os dados referentes ao duodeno, o último bloco indica observações importantes, como a conclusão formulada pelo médico e resultados de outros exames complementares como biópsia e teste da urease. Cada uma das frases que compõem um bloco corresponde a fatos considerados importantes pelo médico e verificadas durante a realização do exame.

Neste trabalho foram utilizados 609 LMs, sem identificação dos pacientes e fornecidos pelo Serviço de Endoscopia Digestiva do Hospital Municipal de Paulínia, dos quais foram extraídos os dados referentes ao esôfago, estômago e duodeno. A metodologia aplicada, constituída por duas fases, possibilitou mapear esses LMs para as respectivas

BDs de cada domínio do conhecimento considerado. As duas fases dessa metodologia, representadas na Figura 3, são descritas a seguir.

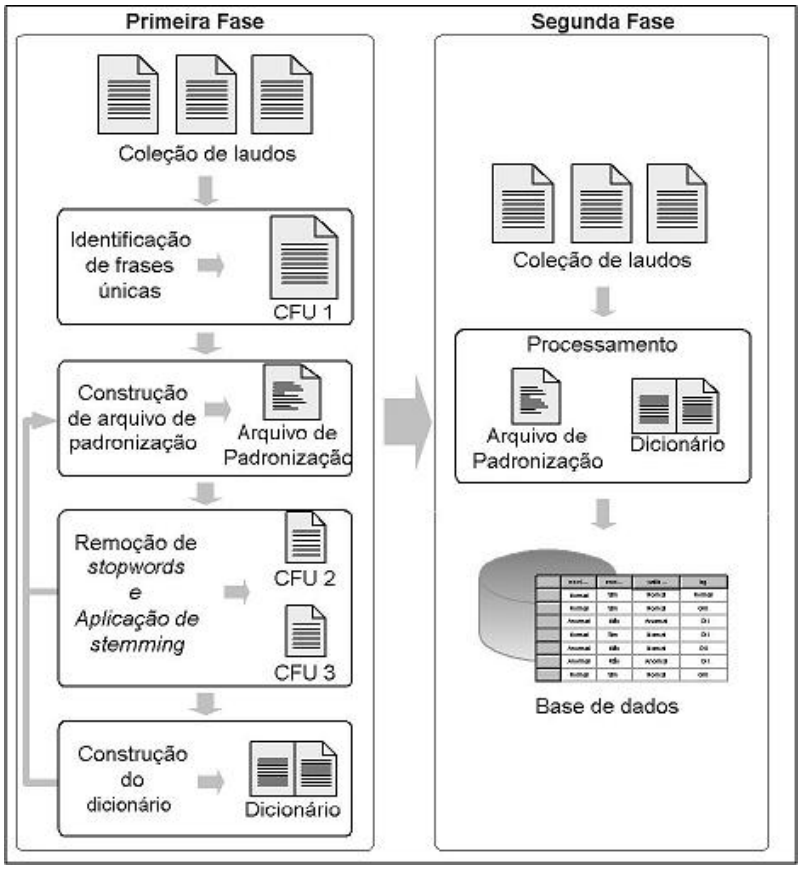


**Figura 3. Representação das duas fases da metodologia [Honorato et al. 2005].**

## 2.1. Primeira fase

A primeira fase é constituída por duas etapas. O objetivo principal dessa fase é a construção de um dicionário, o qual contém o conhecimento do domínio e que permitirá, juntamente com outros componentes da metodologia, o mapeamento dos LMs para a BD.

### 2.1.1. Primeira etapa

Essa etapa é constituída pelas tarefas de identificação de frases únicas, construção de arquivo de padronização – AP, remoção de *stopwords* – RS e aplicação de *stemming* – AS, as quais são realizadas de modo iterativo e interativo. A característica iterativa se deve ao fato da primeira etapa ser constituída por tarefas, as quais realizam a redução da quantidade de frases únicas e o refinamento das tarefas anteriores, a medida que são aplicadas. A natureza interativa se justifica pela necessidade de contribuições dos especialistas do domínio e do usuário com a finalidade de identificar corretamente as *stopwords*, *stemmings* e padrões presentes nos LMs. As tarefas constituintes dessa etapa são abordadas a seguir.

- Os LMs de EDA são compostos por frases, as quais contém diagnósticos, prognósticos e observações do médico sobre o exame realizado. Na primeira tarefa, as frases de todos os LMs do conjunto são concatenadas em um arquivo, o qual é ordenado alfabeticamente, facilitando a detecção de frases repetidas. São coletadas desse arquivo apenas um exemplar de cada frase, originando o primeiro conjunto de frases únicas – CFU1;
- A partir da criação do CFU1, são realizadas reuniões com especialistas para a identificação de padrões existentes nesse conjunto, os quais permitem padronizar os dados do conjunto de LMs em um mesmo formato, compatível com o dicionário a ser construído e com o processo de preenchimento da BD na segunda fase da metodologia. A detecção de padrões é fundamental para novas reduções na quantidade de frases únicas, pois possibilita substituir expressões textuais distintas que apresentam semântica similar, ou que estejam em uma disposição inadequada, por expressões pré-definidas pela metodologia proposta. A construção do AP prossegue durante a aplicação das próximas tarefas da primeira etapa, a medida que são encontrados dados nos quais a identificação de padrões seja possível;
- Ao aplicar a tarefa de RS, surgem novas redundâncias, o que possibilita a redução do CFU1 e ocasiona a geração do segundo conjunto de frases únicas – CFU2. Para alcançar essa redução, a primeira versão do AP deve estar criada e as expressões não representativas, denominadas *stopwords*, devem ser removidas quando presentes no CFU1. Essas expressões correspondem às conjunções, preposições, artigos e outras palavras, as quais são termos do domínio considerados dispensáveis por especialistas;
- A última tarefa realizada é a AS nas frases contidas no CFU2, substituindo variações morfológicas, como gênero, grau e número, de cada palavra por um radical comum, denominado *stemming*. Freqüentemente, as diferentes morfologias de uma palavra não alteram o significado desta, permitindo a adoção de apenas uma expressão, como o *stemming*, para representar essas variações. Interações com especialistas são necessárias, pois nem todas as palavras com mesmo *stemming* contém significado semelhante. A AS torna novas redundâncias evidentes, o que possibilita a criação do terceiro conjunto de frases únicas – CFU3, reduzindo ainda mais a quantidade de frases únicas identificadas nos LMs.

### 2.1.2. Segunda etapa

Os dados contidos no CFU2 e CFU3, assim como os padrões dispostos no arquivo de padronização, são analisados em conjunto com os especialistas para a elaboração do dicionário, o qual é um artefato criado pela metodologia, com o intuito de auxiliar o preenchimento da BD. Para construí-lo, são necessárias interações com os especialistas para definição dos atributos que integram a BD, os quais podem representar os dados presentes nos LMs.

Em seguida, é definida a estrutura do dicionário, conforme a disposição dos dados nos LMs. A maioria das especialidades médicas preenche os respectivos LMs com os dados referentes às estruturas anatômicas examinadas e as características associadas a essas estruturas. Nos LMs de EDA observa-se também, freqüentemente, detalhes relacionados às características conforme demonstrado na Figura 4. Neste exemplo, o LM

feito pelo especialista apresenta três estruturas inter-relacionadas, que correspondem a um local (exemplo: bulbo parede anterior), uma característica (exemplo: úlcera) e uma subcaracterística (exemplo: fibrina).


**Figura 4. Frase de um LM.**

Para comportar essa disposição dos dados, o dicionário assume forma hierárquica, sendo constituído por locais, características e subcaracterísticas, os quais correspondem, respectivamente, às estruturas anatômicas, características e detalhes presentes nos LMs. O dicionário compreende uma lista de locais, na qual cada local está relacionado com a sua lista de características. A representação de cada característica obedece um formato, que indica o nome da característica, nome do atributo referente à esta característica e o valor que esse atributo deve conter na BD. Este fato ocorre quando essa característica estiver presente em um LM. Uma característica apresenta uma lista de subcaracterísticas, a partir da qual cada elemento se apresenta no dicionário no mesmo formato. A Figura 5 representa a hierarquia do dicionário e o suporte oferecido por essa disposição no preenchimento da BD, na qual observa-se o aspecto atributo-valor, de acordo com a metodologia proposta.

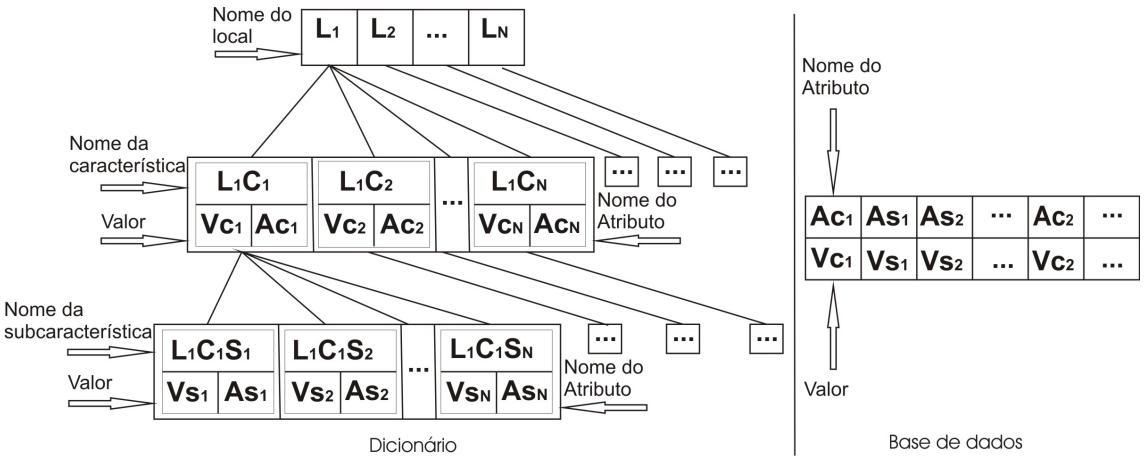


**Figura 5. Estrutura do dicionário e Base de Dados.**

Para exemplificar a estrutura do dicionário, são utilizados o local denominado $L_1$ e as suas relações, apresentadas a seguir. O local $L_1$ está atrelado à uma lista de N características, a partir da qual se analisa o elemento designado em $L_1C_1$. Quando essa característica estiver presente em um LM, processado pela metodologia apresentada, o valor indicado em $Vc_1$ corresponde ao dado a ser armazenado na BD, no atributo descrito por $Ac_1$. O mapeamento prossegue, verificando os termos constituintes do conjunto de N subcaracterísticas correspondente à $L_1C_1$ que ocorrem nos LMs. No momento em que a verificação for efetivada, a subcaracterística simbolizada por $L_1C_1S_1$ registra no atributo $As_1$ o dado contido em $Vs_1$. Durante a aplicação da metodologia, de acordo com os dados

que compõem o LM, podem ser encontrados vários locais, cada um apresentando diversas características, as quais também podem se relacionar com mais de uma subcaracterística.

### 2.2. Segunda fase

Após a construção do dicionário na etapa anterior, é possível o mapeamento dos LMs. A partir de um conjunto de LMs de EDA, um laudo é extraído, do qual se processam, frase por frase, os dados encontrados. Cada local do dicionário é pesquisado na frase analisada e, caso esteja presente, é realizada a busca por características, relacionadas à este local, ao longo da frase. Para cada característica encontrada, é realizada a pesquisa por possíveis subcaracterísticas relacionadas à característica, presentes ainda na mesma frase. Após a identificação das relações entre locais, características e subcaracterísticas existentes na frase, armazenam-se os valores dos atributos correspondentes a essas conexões em um Registro de Base de Dados – RBD, o qual contém, ao final do processo, os dados presentes no LM analisado. O processamento de todas as frases de um LM possibilita o armazenamento do RBD na BD, e outro LM do conjunto é selecionado para o mapeamento pela metodologia.

## 3. Resultados e Discussão

A metodologia apresentada foi aplicada ao processamento de 609 LMs, os quais continham dados do esôfago, do estômago e do duodeno. O mapeamento dos dados resultou na construção de dicionários e no preenchimento de BDs, os quais foram distintos para cada órgão. Todos os dicionários obtidos apresentam a estrutura hierárquica mencionada, com locais, características e subcaracterísticas. Um resumo dos resultados é apresentado na Tabela 1.

**Tabela 1. Resultados da aplicação da metodologia nos domínios considerados.**

| Domínio | Quantidade de Frases | | | Número de atributos | % Preenchimento BD | |
|---|---|---|---|---|---|---|
| | Inicial | CFU1 | CFU2 | | Atributos | Registros |
| Esôfago | 3044 | 81 | 65 | 16 | 88,99 | 89,9 |
| Estômago | 5226 | 352 | 259 | 22 | 25,17 | 2,19 |
| Duodeno | 1710 | 90 | 88 | 51 | 30,92 | 30,69 |

O esôfago foi o primeiro domínio processado, contabilizando 3044 frases no arquivo que continha concatenados todos os dados do domínio, extraídos do conjunto de LMs. A aplicação da identificação de frases únicas possibilitou a geração do CFU1, composto por 81 frases. Esse processo reduziu em 97,34% a quantidade de frases originais. A criação do arquivo de padronização e a posterior remoção de *stopwords* resultou na elaboração de um CFU2 com 65 frases, diminuindo em 97,86% as frases iniciais e 19,75% o conteúdo do CFU1. Após a aplicação de *stemming*, o último conjunto obtido foi o CFU3, o qual indicava 64 frases, permitindo concluir que, após a aplicação dessas tarefas, foi possível uma redução de 97,9% na quantia de frases presentes no conjunto de LMs. Em seguida, o dicionário foi construído adequadamente com os padrões encontrados. A BD obtida foi composta por 16 atributos, e após o término da aplicação da metodologia proposta, alcançou-se um preenchimento de 88,99% em relação aos atributos e 89,9% quando consideradas a ocupação dos registros. Os dados apresentaram

boa distribuição na BD, sendo que os atributos com menos valores contém 85,39% das lacunas preenchidas.

O processamento dos dados do estômago ocorreu em seguida, e verificou-se que a união de todas as frases dos LMs resultava em 5226 frases. O CFU1 construído com a identificação de frases únicas era composto por 352 frases, de modo que se alcançou uma redução nas frases iniciais de 93,27%. Nas duas tarefas seguintes, a quantia de frases foi diminuída em 95,04%, de modo que o CFU2 apresentou 259 frases. Dando sequência ao mapeamento dos dados de estômago, os 22 atributos da BD desse domínio alcançaram um preenchimento de 25,17%, mas os registros apresentavam 2,19% de ocupação. É fundamental salientar que o baixo suprimento da BD se deve à presença escassa de dados referentes a certos atributos nos LMs. Em outras palavras, muitos dos atributos encontravam-se presentes apenas em uma pequena fração dos LMs. Outro fato que contribuiu para que a ocupação da BD fosse mínima foi a disposição dos dados, a qual foi homogênea, com uma concentração de 81,41% dos dados em 6 atributos. Os dados descritos corretamente foram mapeados em sua totalidade, o que comprova a eficiência da metodologia.

O conjunto de dados referentes ao duodeno continha inicialmente 1710 frases, as quais 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 30,92% de ocupação dos atributos e 30,69% de preenchimento dos registros. Assim como no domínio do estômago, o baixo suprimento da BD do duodeno se explica pela característica dos dados, os quais apresentam homogeneidade, fato esse observado pela concentração de 94,49% das células preenchidas em 16 atributos.

Uma análise das BDs permitiu concluir que, apesar das características próprias dos dados de cada domínio, a metodologia auxilia o mapeamento adequado de todos os LMs. A BD do esôfago apresentou a maior taxa de preenchimento, por não apresentar as deficiências, em termos de informações descritas nos laudos, presentes nos dados de outros domínios. Os domínios do estômago e do duodeno, embora demonstrem uma taxa de preenchimento baixa no contexto geral, atingiram um alto suprimento em alguns atributos. A adoção de subcaracterísticas torna os dicionários dos domínios eficientes, possibilitando o armazenamento de diagnósticos mais precisos e adequados à mineração de dados. As particularidades, intrínsecas à uma característica, proporcionam um conhecimento mais eficaz sobre a anormalidade que está afetando a estrutura anatômica correspondente. Ao obter um mapeamento mais completo, as informações extraídas podem proporcionar a realização de uma predição mais próxima da realidade, auxiliando o usuário na tomada de decisão.

## 4. Conclusões

Neste trabalho foi apresentado um estudo de caso, por meio de uma metodologia para mapeamento de LMs de EDA para BDs referentes aos domínios do esôfago, do estômago e do duodeno. A metodologia possui como principal artefato um dicionário, o qual foi construído para cada órgão, podendo ser útil no processamento de outros conjuntos de LMs referentes a outros órgãos. A aplicação da metodologia garante o mapeamento dos LMs, reduzindo o tempo necessário e a subjetividade no preenchimento da BD. Como trabalho futuro, planeja-se estudar se a mineração de dados pode auxiliar na obtenção

de conhecimento apenas com os atributos com preenchimento relevante. Esse estudo se justifica pela homogeneidade dos LMs nos domínios do estômago e do duodeno, os quais contém a maioria dos seus dados concentrados em alguns atributos e com valores iguais.

**Agradecimentos**

Ao Programa de Desenvolvimento Tecnológico Avançado — PDTA/FPTI-BR — pelo auxílio por meio da linha de financiamento de bolsas.

## Referências

Baranauskas, J. A. (2001). *Extração automática de conhecimento por múltiplos indutores*. PhD thesis, USP.

Cherman, E. A., Lee, H. D., Ferrero, C. A., Honorato, D. D. F., and Chung, W. F. (2006). Um estudo de caso da aplicação de uma metodologia de mapeamento de laudos médicos para base de dados. In *Anais do XV Encontro Anual de Iniciação Científica - EAIC*, pages 1–3, Ponta Grossa - PR.

Cordeiro, F. (1994). *Endoscopia Digestiva*. MEDSI.

Fayyad, U., Piatetsky, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54.

Ferrero, C. A., Lee, H. D., Wu, F. C., Machado, R. B., Honorato, D. D. F., Fagundes, J. J., and Góes, J. R. N. (2005). Um estudo de caso de construção de base de dados a partir de laudos médicos. In *13º Simpósio Internacional de Iniciação Científica da USP (SIICUSP)*, São Carlos - SP.

Ferro, M., Lee, H. D., and Esteves, S. C. (2002). Intelligent data analysis: A case study of the diagnostic sperm processing. In *International Conference on Computer Science, Software Engineering, Information Technology, e-Business and Applications - CSITeA'02*, pages 352–356, Foz do Iguaçu - PR.

Honorato, D. D. F., Cherman, E. A., Lee, H. D., Monard, M. C., and Wu, F. C. (2007). Construção de uma representação atributo-valor para extração de conhecimento a partir de informações semi-estruturadas de laudos médicos. In *XXXIII Conferência Latinoamericana de Informática - CLEI (a ser publicado)*, San José - Costa Rica.

Honorato, D. D. F., Lee, H. D., Monard, M. C., Wu, F. C., Machado, R. B., Neto, A. P., and Ferrero, C. A. (2005). Uma metodologia para auxiliar no processo de construção de bases de dados estruturadas a partir de laudos médicos. In *Anais do Encontro Nacional de Inteligência Artificial*, pages 593–601, São Leopoldo - RS.

Lee, H. D. (2005). *Seleção de atributos importantes para a extração de conhecimento de bases de dados.* PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos - Brasil.

Lee, H. D. and Monard, M. C. (2003). Seleção de atributos para algoritmos de aprendizado de máquina supervisionado utilizando como filtro a dimensão fractal. *Revista de La Sociedad Chilena de Ciencia de La Computación*, pages 1–8.

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann.

Rezende, S. O. (2003). *Sistemas Inteligentes: Fundamentos e Aplicações*. Manole.

Spolaôr, N., Lee, H. D., Cherman, E. A., Honorato, D. D. F., Fagundes, J. J., Góes, J. R. N., and Wu, F. C. (2007). Uma metodologia de mapeamento de laudos endoscópicos para bases de dados estruturadas: Estudo de caso. In *Anais do XVI Encontro Anual de Iniciação Científica - EAIC (a ser publicado)*, Maringá - PR.

Witten, I. H. and Frank, E. (2005). *Data Mining : Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.