Utilização da Indexação Automática para Auxílio à Construção de uma Base de Dados para a Extração de Conhecimento aplicada à Doenças Pépticas

Daniel de F. Honorato[1*], Huei D. Lee[1], Renato B. Machado[1,4], Feng C. Wu[1,2,3], Antonio P. Neto[5]

[1] Laboratório de Bioinformática - LABI
Universidade Estadual do Oeste do Paraná - UNIOESTE
Caixa Postal 961, Foz do Iguaçu, Paraná, 85870-650

[2]Serviço de Coloproctologia da Faculdade de Ciências Médicas
Universidade Estadual de Campinas - UNICAMP

[3]Instituto de Tecnologia em Automação e Informática - ITAI
Campus da Universidade Estadual do Oeste do Paraná - UNIOESTE
Caixa Postal 1511, Foz do Iguaçu, Paraná, 85856-000

[4]Itaipu Binacional

[5]Serviço de Endoscopia Digestiva Alta, Hospital Municipal de Paulínia.

labi@unioeste.com.br

Abstract. *Computational evolution has made possible the generation of an enormous quantity of data. Thus, efficient processes for analysing and understanding the stored data are necessary. One of these processes is the Knowledge Discovery on Databases within it, it's generally necessary to pre-process the data. This work describes part of the pre-processing stage, which was divided into two phases. In the first one, a conversion module and a search algorithm are presented. In the second phase, the stages accomplished to construct a dictionary of words, that will be used as support to the process of construction of the database for the stage of data mining, are presented .*

Resumo. *A evolução da computação possibilitou um enorme crescimento das bases de dados. Assim, processos eficientes para análise e compreensão dos dados armazenados são necessários. Um desses processos é a Extração de Conhecimento em Base de Dados e para que possa ser realizada, é, em geral, necessário que os dados sejam pré-processados. Este trabalho descreve parte da etapa de pré-processamento, a qual foi dividida em duas fases. Na primeira, são apresentados um módulo de conversão e um algoritmo de pesquisa. Na segunda, são apresentadas as etapas realizadas na construção de um dicionário de palavras que será utilizado como apoio ao processo de construção da base de dados para a etapa de mineração de dados.*

---

*Bolsista do ITAI - Instituto de Tecnologia em Automação e Informática / IEL - Instituto Euvaldo Lodi

1. Introdução

Com o enorme crescimento das bases de dados, resultado do avanço tecnológico ocorrido nos últimos anos, tornou-se cada vez mais difícil analisar e extrair, manualmente, informações e padrões a partir dos dados. Assim, surgiu a necessidade de métodos para descoberta de novos conhecimentos e padrões com o auxílio de técnicas computacionais, por exemplo, por meio do processo de Extração de Conhecimento de Base de Dados (ECBD)[1][Fayyad et al., 1996]. A ECBD tem por objetivo descobrir conhecimento, a partir de um conjunto de dados, o qual poderá ser utilizado em um processo decisório [Rezende et al., 2003]. Esse processo, iterativo e interativo, é composto, basicamente, por três etapas: pré-processamento, mineração de dados e pós-processamento (Figura 1).

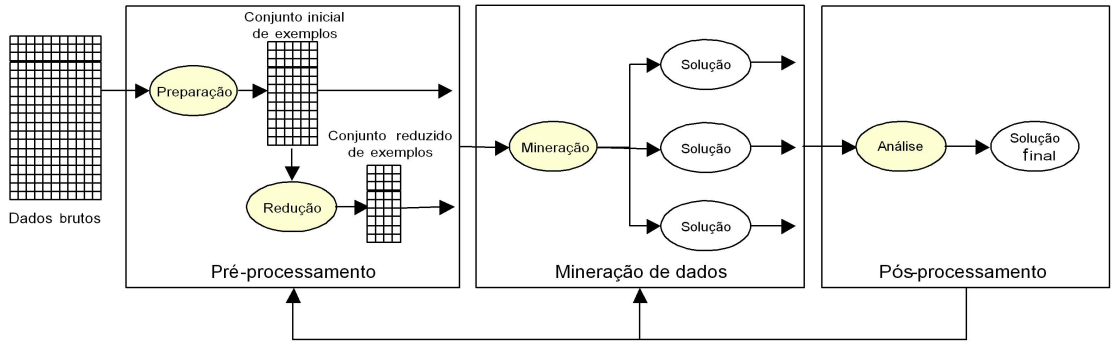


Figura 1: Processo de Extração de Conhecimento em Base de Dados. [Baranauskas, 2001]

A etapa de pré-processamento é, em geral, a etapa que representa maior custo de tempo dentro do processo. Tem como objetivo realizar tarefas tais como a preparação, redução e transformação dos dados. Ainda nessa etapa, é necessário que os dados estejam representados no formato apropriado para a próxima etapa. Um dos formatos mais comumente utilizados é o atributo-valor. A mineração de dados tem como característica a configuração, escolha e execução de um ou mais algoritmos de extração de padrões sobre os dados selecionados na fase de pré-processamento. Essa etapa é realizada de maneira iterativa, sendo portanto necessários vários ajustes nos parâmetros durante o processo de mineração de dados visando melhores resultados nos modelos construídos. Após a execução da etapa de extração de padrões, acontece a etapa de pós-processamento, na qual os modelos construídos são avaliados e validados. Em cada uma dessas etapas é possível retornar à anterior. Depois de concluído o processo, o conhecimento extraído é disponibilizado ao usuário, o qual pode ser utilizado como auxílio no processo de tomada de decisões.

No cenário atual de desenvolvimento tecnológico, hospitais e clínicas médicas estão registrando cada vez mais informações sobre pacientes e processos laboratoriais. Com esse acúmulo de informações, pode tornar-se difícil aos profissionais da saúde coletar, analisar e extrair informações que possam ajudá-los, por exemplo, no diagnóstico de doenças. Desse modo, torna-se necessária a aplicação de métodos computacionais que possam auxiliar a análise, de maneira mais completa, dessa grande quantidade de dados [Monard and Lee, 2003, Ferro et al., 2002, Lee et al., 2000].

[1]KDD - *Knowledge Discovery on Databases*

Atualmente, as doenças pépticas gastroduodenais representam uma das entidades patológicas de maior incidência na população despertando cada vez mais interesse na pesquisa dessa área. Este trabalho está inserido dentro do projeto de Análise Inteligente de Dados aplicado a Doenças Pépticas, o qual está sendo desenvolvido no Laboratório de Bioinformática da Unioeste em parceria com o Hospital Municipal de Paulínia e o Laboratório de Inteligência Computacional do ICMC, Universidade de São Paulo - São Carlos.

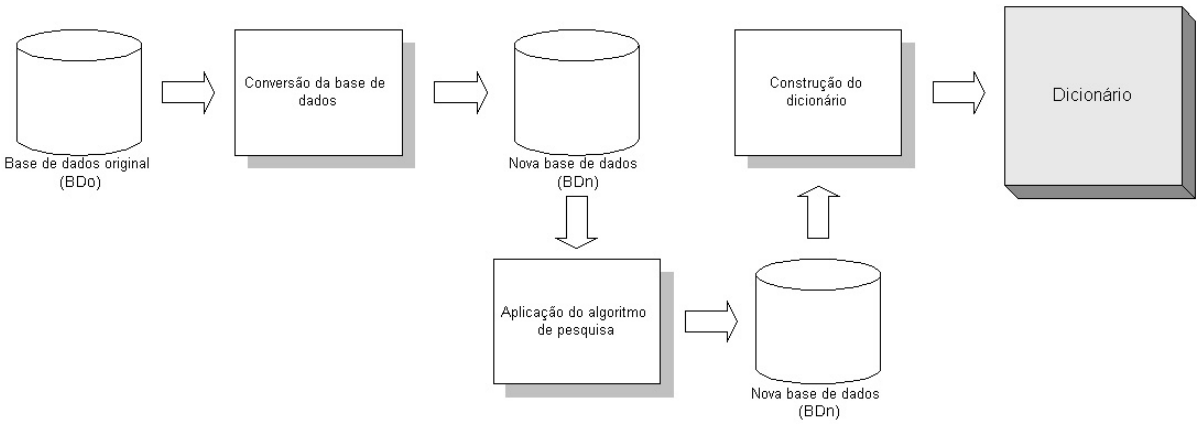


Figura 2: Processo realizado neste trabalho.

Neste trabalho, é apresentada parte da etapa de pré-processamento, a qual está sendo realizada em duas fases. A primeira constitui na seleção de atributos de interesse da base de dados original (BDo) e conversão para uma nova base de dados (BDn). Também nessa fase, é realizada a aplicação de um algoritmo de pesquisa sobre a BDn, especificamente sobre o atributo laudo, na qual são selecionados os casos em que foi realizada a biópsia do tecido gastroduodenal. A segunda fase compreende a construção de um dicionário, que irá auxiliar na semi-automatização da construção da base de dados que será usada na etapa de mineração de dados (Figura 2).

2. Descrição da Base de Dados Original

Neste trabalho, foi utilizada uma base de dados relacional obtida no Serviço de Endoscopia Digestiva Alta do Hospital Municipal de Paulínia. Os dados armazenados nessa base são provenientes de exames realizados no período de março a novembro de 2001, num total de 1950 casos descritos com 58 atributos cada um. Nesses atributos são armazenadas informações sobre o exame de Endoscopia Digestiva Alta (EDA) tais como data, médico que atendeu, idade, entre outros, assim como o laudo, que armazena o resultado do exame de EDA em formato texto.

3. Conversão da Base de Dados

A conversão da base de dados foi realizada por meio de um módulo implementado na linguagem Delphi 6.0 apoiada pela linguagem de consulta SQL. Esse módulo tem como objetivo selecionar apenas um subconjunto de atributos da BDo e inserí-los na BDn. A

BDn foi criada no Sistema Gerenciador de Banco de Dados (SGBD) mySQL que é gratuito e possui todos os recursos necessários para o desenvolvimento deste trabalho. O subconjunto de atributos selecionados são descritos na Tabela 1.

Tabela 1: Atributos selecionados.

Atributo	Descrição
codpaciente	código do paciente
laudo	armazena o resultado do exame de EDA
usuário	médico que realizou a consulta
solicitante	médico que fez a solicitação da consulta
codtipoexame	código do tipo de exame realizado
codconvenio	código do convênio
data_exame	data em que ocorreu o exame
registro	número de registro do paciente
idade_exame	idade do paciente quando realizou o exame

4. Algoritmo de Pesquisa

O algoritmo de pesquisa tem por objetivo localizar, dentre todos exames realizados, os casos cujos pacientes realizaram biópsia do tecido gastroduodenal. Esse algoritmo é necessário pois a linguagem SQL não suporta pesquisa em atributos do tipo texto. A implementação foi realizada na linguagem Perl [Sheppard, 2000], a qual tem como principal característica a facilidade de extração de informações em textos. Neste trabalho foi utilizada a versão ActivePerl v. 5.8.0.

Na Figura 3 são ilustrados exemplos com fragmentos do atributo laudo.

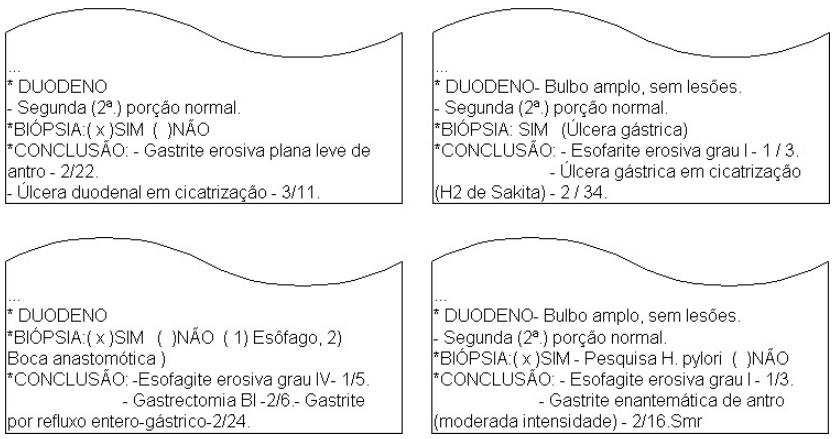


Figura 3: Fragmentos de atributos laudos.

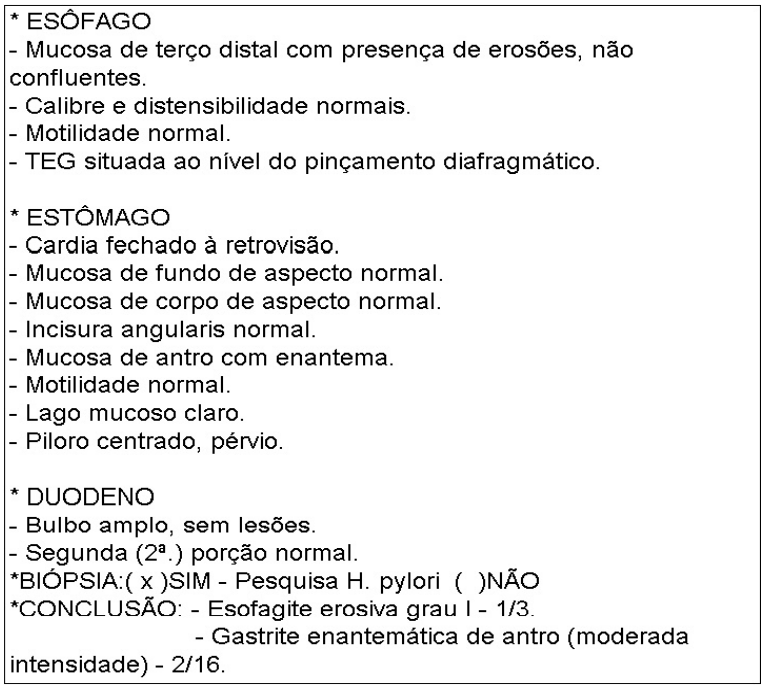
Nos exemplos apresentados, pode-se observar que não existe um padrão quando colocada a situação da biópsia, isto é, se foi realizada ou não. Apenas é possível identificar que as palavras BIÓPSIA, SIM e NÃO estão presentes na mesma linha. Portanto, o algoritmo de pesquisa elaborado tem como objetivo extrair a situação da biópsia descrita no laudo, por meio da verificação das possíveis combinações que podem ocorrer. O algoritmo é apresentado a seguir em pseudocódigo:

```
rotina realizouBiopsia(laudo)

começo
   se laudo contêm "biópsia = SIM"
      retorna verdadeiro
     senão
      se laudo contêm "biópsia = NÃO" ou
         se não existe "biópsia = SIM" ou "biópsia = NÃO" em laudo
            retorna falso
fim
```

5. Construção do Dicionário

O atributo laudo armazena informações sobre o exame de EDA, especificamente sobre esôfago, estômago e duodeno. Na Figura 4 é ilustrado um exemplo da apresentação de informações no atributo laudo, as quais estão sendo utilizadas para auxiliar a construção do dicionário de palavras.


* ESÔFAGO
- Mucosa de terço distal com presença de erosões, não confluentes.
- Calibre e distensibilidade normais.
- Motilidade normal.
- TEG situada ao nível do pinçamento diafragmático.

* ESTÔMAGO
- Cardia fechado à retrovisão.
- Mucosa de fundo de aspecto normal.
- Mucosa de corpo de aspecto normal.
- Incisura angularis normal.
- Mucosa de antro com 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 I - 1/3.
- Gastrite enantemática de antro (moderada intensidade) - 2/16.

Figura 4: Laudo de biópsia do paciente.

Optou-se por aplicar o método de indexação automática, o qual tem como característica armazenar, em uma estrutura de índices, os termos de importância encontrados no atributo laudo. Portanto, para que seja possível extrair somente esses termos, é necessário preparar o atributo laudo, isto é, deixá-lo em um formato ideal para extração de atributos. O método de indexação automática é composto por algumas fases que são, normalmente, definidas como: identificação de termos (simples e compostos), remoção de *stopwords*[2], normalização morfológica (*stemming*) e seleção de termos [Wives, 2002].

A segunda fase do pré-processamento apresentado neste trabalho, caracteriza-se pela construção de um dicionário de palavras, o qual servirá como apoio para preparação

[2]palavras irrelevantes encontradas no texto.

para a mineração de dados. O objetivo do dicionário é armazenar, em uma estrutura de dados, palavras de interesse extraídas dos atributos laudos da BDn, juntamente com a característica histológica da mesma. A construção do dicionário está sendo realizada, com auxílio do especialista, considerando-se as quatro fases citadas anteriormente: identificação de frases únicas nos laudos, remoção de *stopwords*, normalização das palavras por meio da aplicação de *stemming* e finalmente a inserção das palavras no dicionário. A Figura 5 apresenta o esquema do processo de construção do dicionário.

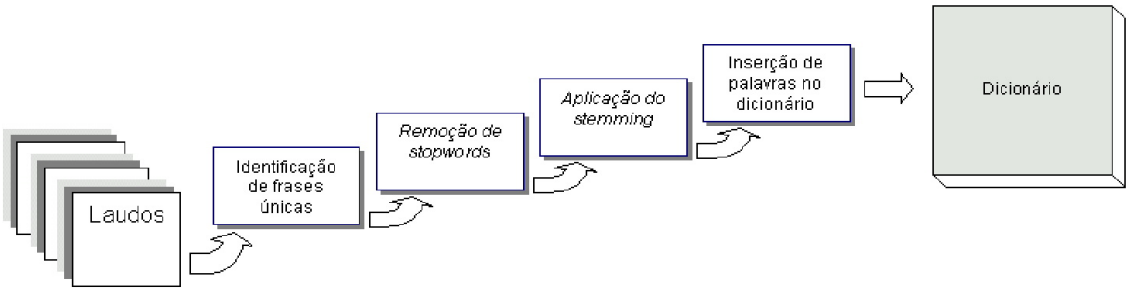


Figura 5: Processo de construção do dicionário.

5.1. Identificação de Frases Únicas

A primeira etapa executada para a construção do dicionário foi identificar as frases únicas existentes nos laudos contidos na nova base de dados (BDn). Assim, foi implementado um algoritmo com o objetivo de armazenar em uma estrutura de dados do tipo lista o conteúdo de cada laudo contido na BDn, sendo cada posição da lista correspondente a uma linha do atributo laudo. Depois de concluída a construção da lista, foi aplicado um outro algoritmo para ordenar a mesma. Essa ordenação teve como objetivo reunir frases repetidas permitindo que apenas um exemplar de cada frase fosse mantido. Sendo assim, ao final do processo, tem-se como resultado um arquivo de frases (ARQf). A Figura 6 ilustra o processo de identificação de frases únicas.

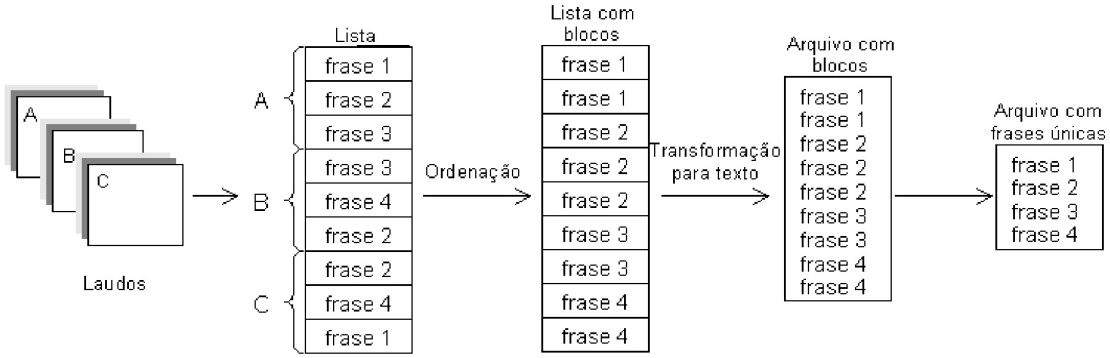


Figura 6: Processo de identificação de frases únicas.

5.2. Remoção de *Stopwords*

Essa fase tem como objetivo remover *stopwords*, as quais são palavras consideradas não relevantes para uso na análise do texto. Normalmente, são compostas por preposições,

conjunções e artigos. A remoção de *stopwords* foi realizada de modo automático. Primeiramente, criou-se uma estrutura de dados do tipo lista, denominada *stoplist*, na qual foram definidas todas as palavras que poderiam ser removidas de ARQf. Posteriormente, aplicou-se um algoritmo cujo objetivo foi realizar uma pesquisa em ARQf e remover as palavras que fossem iguais à alguma presente na *stoplist*. A Figura 7 ilustra um exemplo de remoção de *stopwords*.

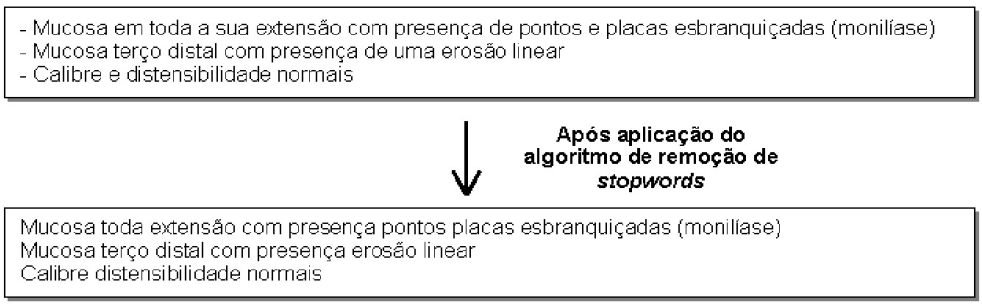


Figura 7: Exemplo de remoção de *stopwords*.

5.3. Aplicação do Algoritmo de *Stemming*

Após a remoção de *stopwords*, foi realizada a normalização das palavras que restaram em ARQf. Esse procedimento permite que seja possível trabalhar com um vocabulário controlado. Neste trabalho, a normalização foi realizada por meio da aplicação do *stemming* utilizando o Método de Porter [Orengo and Huyck, 2001]. A aplicação desse método consiste na identificacão das diferentes inflexões referentes à mesma palavra e sua substituição por um radical comum [Ebecken et al., 2003]. Por exemplo, o radical **consider** pode ter vários sufixos diferentes como: **considerar**, **considerado**, **consideração** entre outros. Neste trabalho, o processo foi realizado por meio da extração de cada palavra de ARQf e uma redução a sua provável palavra raiz. Ao final da aplicação desse método as informações estão normalizadas conforme ilustra a Figura 8.

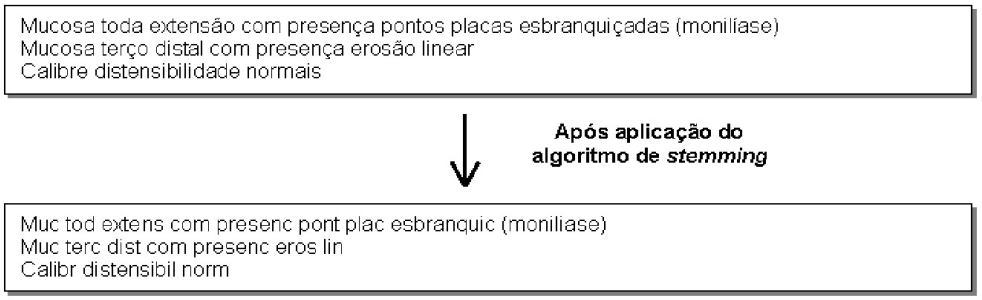


Figura 8: Exemplo de aplicação de *stemming*.

5.4. Inserção de Palavras no Dicionário

O dicionário que está sendo construído tem por objetivo armazenar palavras de interesse existentes em ARQf normalizado e a característica histológica de cada uma delas. A unidade de armazenamento do dicionário é constituída por uma estrutura de dados do tipo *hash*, a qual tem como característica armazenar uma lista dinâmica, na qual cada

posição da lista é composta por um atributo do tipo {chave, valor}. A construção do dicionário está sendo realizada com o auxílio do especialista de domínio. Nesse processo, todas as palavras contidas em ARQf estão sendo analisadas. As palavras de interesse são inseridas na estrutura *hash* como chave, juntamente com a característica histológica da mesma como valor. Na Figura 9 é ilustrada a estrutura do dicionário construído.

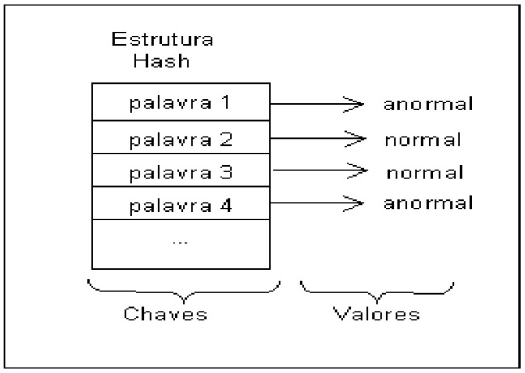


Figura 9: Estrutura do dicionário.

6. Considerações Finais

Neste artigo foi apresentada parte da etapa de pré-processamento referente ao projeto de Análise Inteligente de Dados aplicada à Doenças Pépticas, a qual foi dividida em duas fases. A primeira fase teve como objetivo converter a base de dados original (BDo) para uma nova base de dados (BDn), bem como selecionar apenas os casos nos quais os pacientes sofreram biópsia do tecido gastroduodenal. Na primeira fase do pré-processamento, após a conversão da BDo, dos 58 atributos iniciais foram definidos 8 atributos para compor a BDn. Após a aplicação do algoritmo de pesquisa, dos 1950 casos iniciais, foram selecionados 614 casos para compor a BDn. Na segunda fase foi realizada a construção do dicionário de palavras, na qual foi utilizada como base os dados contidos nos atributos laudos da BDn.

Após o término da construção do dicionário, será iniciada uma terceira fase da etapa de pré-processamento. Nessa fase o dicionário será usado para auxiliar na transformação do atributo laudo em uma base de dados, que complementada com outros atributos já selecionados, poderá ser utilizada para a próxima etapa do processo de ECBD, a etapa de mineração de dados.

A construção do dicionário irá proporcionar a diminuição do custo de tempo usado na fase de preparação dos dados, já que, manualmente, seria necessário um maior envolvimento do especialista. Além disso, a metodologia desenvolvida poderá ser também utilizada para a construção de outros dicionários para extração de informações em outras bases de dados.

Referências

Baranauskas, J. A. (2001). Extração automática de conhecimento por múltiplos indutores. Tese de Doutorado, ICMC-USP, `http://www.teses.usp.br/teses/`

disponiveis/55/55134/tde-08102001-112806.

Ebecken, N. F. F., Lopes, M. C. S., and de Aragão Costa, M. C. (2003). *Mineração de Textos*, chapter 12. In [Rezende, 2003].

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 82–88.

Ferro, M., Lee, H. D., and Esteves, S. C. (2002). Intelligent data analysis: A case study of the diagnostic sperm processing. In *Proceedings of ACIS - CSITeAt'02*, pages 352–356.

Lee, H. D., Monard, M. C., and Esteves, S. C. (2000). Indução construtiva guiada pelo conhecimento: Um estudo de caso do processamento sêmen diagnóstico. In *Proceedings of the IBERAMIA/SBIA*, pages 157–166, Atibaia, SP.

Monard, M. C. and Lee, H. D. (2003). *Processamento de Sêmen Diagnóstico*, pages 461–463. Volume 1 of [Rezende, 2003], 1 edition. Parte II, Aplicação V, ISBN 85-204-1683-7.

Orengo, V. M. and Huyck, C. R. (2001). A stemming algorithm for the portuguese language. In *Proceedings of SPIREŠ2001 Symposium on String Processing and Information Retrieval*, Laguna de San Raphael, Chile.

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

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., and de Paula, M. F. (2003). *Mineração de dados*, chapter 12, pages 307–335. In [Rezende, 2003].

Sheppard, D. (2000). Beginner's introduction to perl. http://www.perl.com/pub/a/2000/10/begperl1.html. acesso em 20/10/2003.

Wives, L. K. (2002). Tecnologia de descoberta de conhecimento em textos aplicadas à inteligência competitiva. Master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.