Metodologia de Mapeamento de Laudos Médicos para Bases de Dados: Aplicação em Laudos Colonoscópicos

Everton Alvares Cherman[1], Huei Diana Lee[1,2], 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, 85856-970 – 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

{evertoncherman,hueidianalee,dfaverih}@gmail.com

Abstract. *In the medical area, as in many other domains, data storage is growing very fast. Computational process, as Knowledge Discovery in Data Bases, may give support to the extraction and analysis of knowledge obtained from these data. Nevertheless, it is necessary that the data is represented in an appropriate format. In this work, it is presented a methodology that assists the mapping of medical findings to a structured database in attribute-value format, as well as the lemmatization and generation of n-grams techniques to aid this process. This methodology was applied to 100 colonoscopy findings.*

Resumo. *Na área da medicina é registrado um volume cada vez maior de dados. Processos computacionais, como o de Descoberta de Conhecimento em Bases de Dados, auxiliam na extração e na análise de conhecimento obtidos a partir desses dados. Para que esse processo seja aplicado é necessário que os dados estejam em um formato adequado. Neste trabalho é apresentada uma metodologia de mapeamento de laudos médicos para uma base de dados estruturada no formato atributo-valor, bem como as técnicas de lematização e geração de n-gramas para auxiliar na primeira etapa da metodologia. A metodologia foi aplicada a 100 laudos médicos de colonoscopia.*

1. Introdução

O registro de um grande volume de dados em diversas áreas do conhecimento tem incentivado a análise de informações e padrões contidos nesses dados. Processos computacionais, como o de Descoberta de Conhecimento em Bases de Dados – DCBD –, podem auxiliar nessa tarefa [Fayyad et al. 1996]. O DCBD é um processo interativo e incremental e é composto por três etapas: pré-processamento, mineração de dados e pós-processamento (Figura 1).

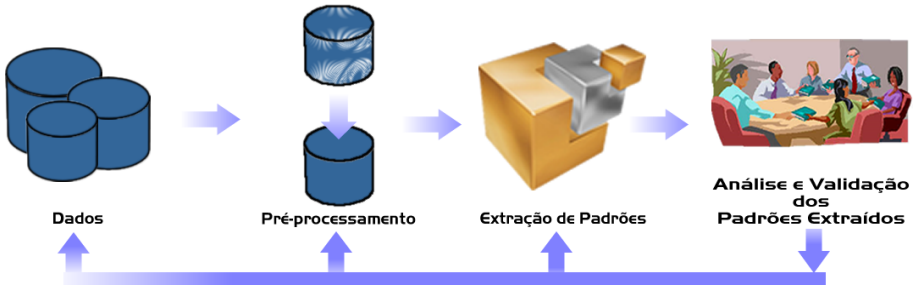

Figura 1. Processo de Descoberta de Conhecimento em Bases de Dados.

A primeira etapa tem como objetivo a preparação, a redução e a transformação dos dados para um formato adequado, em geral o formato atributo-valor, à segunda etapa. Essa etapa usualmente representa o maior custo de tempo no processo, ocupando em torno de 80% do total [Pyle 1999]. Na etapa de mineração de dados são aplicados algoritmos, por exemplo de aprendizado de máquina, para a construção de modelos que representem os padrões contidos nos dados selecionados na etapa de pré-processamento. Diversas iterações podem ser realizadas com o intuito de ajustar os parâmetros dos algoritmos visando melhores resultados nos modelos construídos. Na última etapa, esses modelos são analisados e validados com o apoio de especialistas. Em cada uma dessas etapas é possível retornar a 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 [Rezende 2003, Witten and Frank 2005].

Na área médica, hospitais e clínicas médicas registram um grande volume de dados sobre pacientes e exames laboratoriais. O processo de DCBD pode auxiliar na realização de análises mais completas, as quais podem dar suporte a médicos, por exemplo, no diagnóstico de doenças [Lee 2005, Lee and Monard 2003, Ferro et al. 2002].

Nesse sentido, este trabalho está inserido no projeto de Análise Inteligente de Dados, o qual é desenvolvido em uma parceria entre o Laboratório de Bioinformática – LABI – da Universidade Estadual do Oeste do Paraná – UNIOESTE/Foz do Iguaçu –, o Laboratório de Inteligência Computacional – LABIC – da Universidade de São Paulo – USP/São Carlos – e o Serviço de Coloproctologia da Universidade Estadual de Campinas – UNICAMP.

Neste projeto foi proposta uma metodologia de pré-processamento de informações contidas em Laudos Médicos – LM –, mapeando-as para uma Base de Dados – BD – estruturada no formato atributo-valor [Honorato et al. 2005, Lee 2005]. Essa metodologia foi aplicada em informações referentes à LM de endoscopia digestiva alta e de processamento de sêmen possibilitando uma redução do tempo de mapeamento e eliminando a subjetividade contida no possível mapeamento manual dessas informações para a BD.

Neste trabalho é apresentado um estudo de caso da aplicação dessa metodologia à LM de exames de colonoscopia e é proposta a aplicação de outras técnicas para dar suporte à primeira etapa dessa metodologia.

Este trabalho está organizado da seguinte maneira. Na Seção 2 são apresentados o formato dos LM de colonoscopia bem como a metodologia para o mapeamento desses LM para a BD. Na Seção 3, os resultados são apresentados e discutidos e as conclusões e

trabalhos futuros são apresentados na Seção 4.

2. Materiais e Métodos

Uma das doenças de maior ocorrência no Brasil é o câncer colorretal que, segundo o Instituto Nacional do Câncer, constitui a quarta maior incidência entre todos os tumores malignos, independentemente do sexo [Quilici 2000]. Nesse contexto, os exames de colonoscopia são indispensáveis para o diagnóstico de doenças do intestino grosso [Quilici 2000, Cotran et al. 2000].

Neste trabalho são considerados 100 exames de colonoscopia, os quais não possuem identificação do paciente, coletados no Serviço de Coloproctologia da Faculdade de Ciências Médicas da UNICAMP. Os exames são compostos por um segmento estruturado e um semi-estruturado conforme ilustrado em um exemplo na Figura 2. Neste trabalho são utilizadas as informações referentes aos laudos contidas no segmento semi-estruturado.

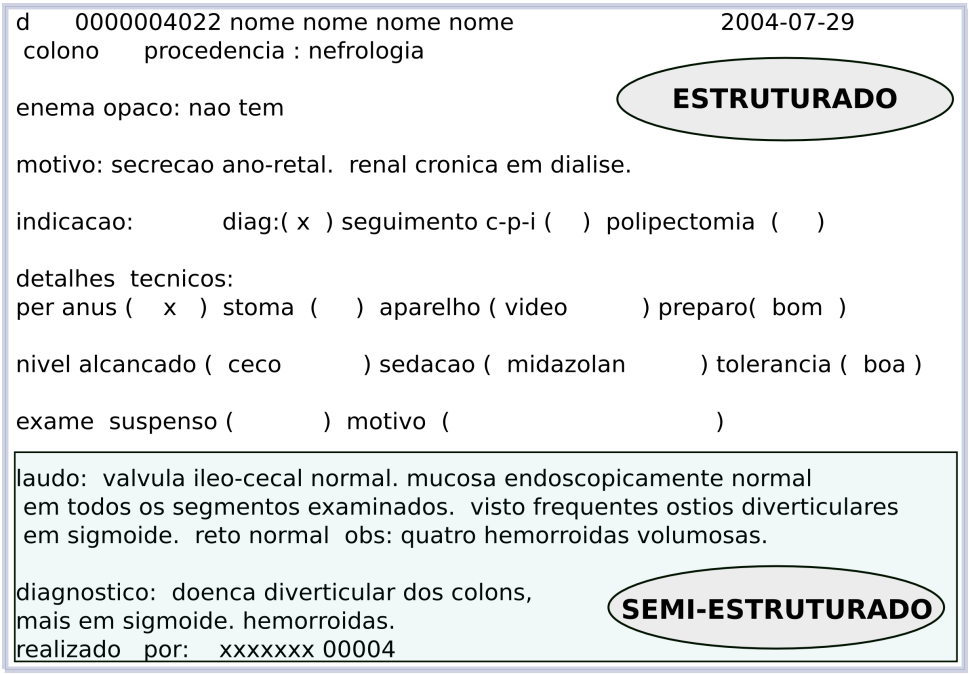

Figura 2. Ilustração de um exame de Colonoscopia.

A metodologia para mapeamento das informações é composta por duas fases: 1) construção de um dicionário do domínio do conhecimento e 2) elaboração e aplicação de uma lógica de mapeamento das informações. Na Figura 3 está ilustrada a metodologia proposta.

2.1. Primeira Fase

A primeira fase, como mencionado anteriormente, tem como objetivo construir, conjuntamente com especilistas, um dicionário do domínio do conhecimento para auxiliar na segunda fase. Esse dicionário consiste em registrar a classificação de todos os termos médicos contidos no domínio dos LM, bem como a especificação dos atributos que irão compor a BD, os quais são associados aos termos médicos na descrição do dicionário.

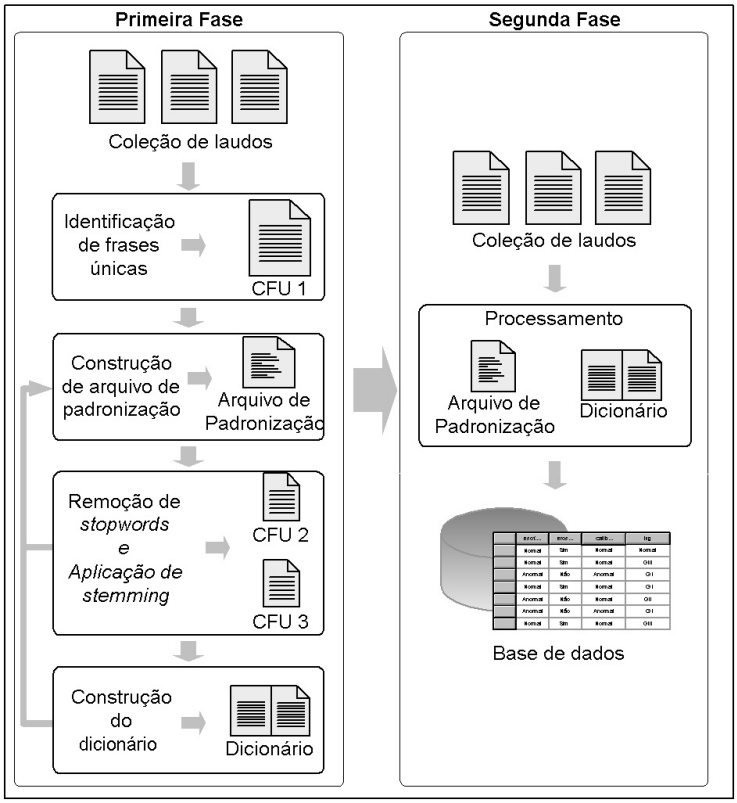

Figura 3. Representação esquemática da metodologia [Honorato et al. 2005].

Os LM de diversas especialidades, bem como a de colonoscopia, possuem informações descritas em estruturas anatômicas e características referentes a essas estruturas. Baseada nessa propriedade, a estrutura de classificação de termos do dicionário é composta por três classes: locais (estruturas anatômicas), características e opcionalmente subcaracterísticas, as quais são características associadas a uma característica [Honorato et al. 2007, Honorato et al. 2005].

Desse modo, as combinações de local com característica e de local com característica e subcaracterística constituem valores de atributos a serem preenchidos na BD, os quais, como os atributos, também são especificados no dicionário. Na Figura 4 está representada esquematicamente a estrutura do dicionário.

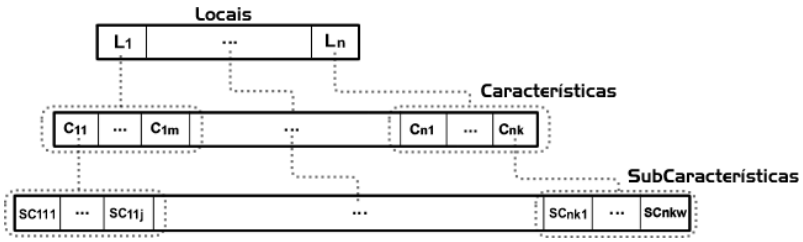

Figura 4. Representação da estrutura do dicionário.

Para auxiliar na construção do dicionário e na identificação dos padrões contidos nas descrições dos LM, são realizadas quatro tarefas: "Identificação de Frases Únicas", "Construção de um Arquivo de Padronização", "Remoção de *Stopwords*" e "Aplicação de

Stemming". Neste trabalho é proposta a aplicação de outras duas tarefas: "Lematização" e "Geração de n-gramas", as quais podem auxiliar nesse processo.

2.1.1. Identificação de Frases Únicas

A identificação de frases únicas consiste em:

- Concatenar em um arquivo todas as frases de um conjunto de LM;
- Ordenar as frases por ordem alfabética;
- Retirar as frases redundantes mantendo apenas um exemplar de cada.

A realização desse processo no conjunto de LM originais constitui o primeiro Conjunto de Frases Únicas – CFU1. Na Figura 5 é ilustrado esse processo.

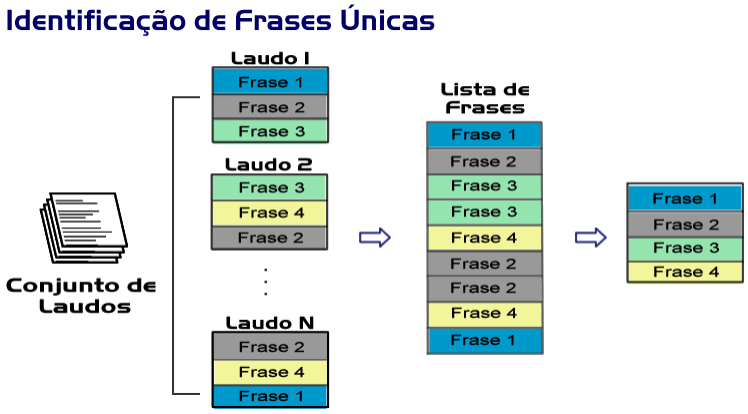

Figura 5. Representação do processo de identificação de frases únicas.

2.1.2. Construção de um Arquivo de Padronização

A freqüente utilização de sinônimos para a descrição de informações semelhantes presentes nos LM ou a presença de frases que expressam informações de uma maneira diferente da que será utilizada pelo dicionário, faz com que a padronização das informações contidas nos laudos seja necessária.

Após a obtenção do CFU1 é possível identificar parte das informações que poderão ser padronizadas. A construção do Arquivo de Padronização - AP - é realizada à medida que as informações que podem ser padronizadas são identificadas e continua até a completa construção do dicionário.

2.1.3. Remoção de *Stopwords*

A próxima tarefa consiste em aplicar a técnica de remoção de *stopwords* no CFU1, com o objetivo de reduzir a quantidade de frases. Essa técnica é definida como a remoção de palavras do texto que ao serem retiradas não modificam o significado original da frase, como artigos, pronomes, preposições e algumas palavras do domínio indicadas por especialistas. Após a aplicação dessa técnica é constituído o segundo Conjunto de Frases Únicas - CFU2.

2.1.4. Aplicação de *Stemming*

O *stemming* tem como objetivo manter no texto apenas o radical de cada palavra, eliminando as diferentes inflexões existentes da mesma palavra [Orengo and Huyck 2001]. A aplicação de *stemming* no CFU2 possibilita a geração do um terceiro Conjunto de Frases Únicas - CFU3.

Outras duas técnicas que podem auxiliar na construção do dicionário e na identificação de padrões contidos nos LM são descritas a seguir.

2.1.5. Lematização

A técnica de lematização consiste em substituir as palavras por suas formas canônicas, como o singular de um substantivo ou o infinitivo de um verbo. Do mesmo modo que a técnica de *stemming*, a lematização elimina as diferentes inflexões das palavras, porém mantendo-as de forma mais legível [B.Sc. 2002].

2.1.6. Geração de n-gramas

A geração de n-gramas tem como objetivo fornecer a freqüência com que toda composição de n palavras consecutivas são descritas em um conjunto de textos. Por exemplo, unigrama é a freqüência de toda palavra isolada, bem como bigrama é a freqüência de toda composição de duas palavras consecutivas presentes nos textos. Os n-gramas são analisados em conjunto com especialistas e a aplicação desse técnica no conjunto de LM possibilita a identificação de possíveis termos do domínio, os quais podem compor o dicionário.

2.2. Segunda Fase

Com base no dicionário definido na fase anterior e nos padrões identificados na descrição dos LM, é elaborada uma lógica de mapeando das informações denominada de Algoritmo de Busca e Preenchimento – ABP. O algoritmo deve contemplar a lógica de distribuição dos termos em cada frase, mapeando para a BD as informações contidas nessas frases. Desse modo, o ABP é executado para todas as frases de cada LM do conjunto, identificando a classificação e a combinação das palavras, bem como os respectivos atributos e seus valores a serem preenchidos na BD. Na Figura 6 está ilustrado o processo de mapeando dos LM.

3. Resultados e Discussão

A metodologia foi aplicada considerando um conjunto de 100 LM de colonoscopia, os quais contêm um total de 474 frases.

Na primeira fase aplicou-se as tarefas para identificação de frases únicas e foram construídos o AP e o dicionário conjuntamente com especialistas. O CFU1 foi composto por 412 frases, diminuindo 13,08% em relação ao total. A remoção de *stopwords* e a aplicação de *stemming* sobre o CFU1 resultaram no CFU2 e no CFU3 contendo 396 e 393 frases respectivamente, reduções de 16,45% e 17,08% em relação ao total. A técnica

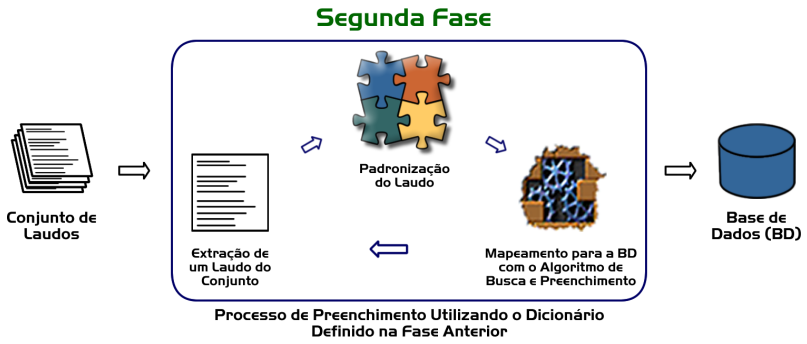

Figura 6. Ilustração do processo de mapeando dos LM.

de lematização foi aplicada ao CFU2 gerando a mesma quantidade de frases constatadas no CFU3, porém contendo frases com maior legibilidade. A geração de n-gramas sobre os laudos resultou em 1812 bigramas e 2321 trigramas, dos quais 129 e 52 respectivamente foram identificados em cinco ou mais LM.

Em [Honorato et al. 2005] foram utilizados 100 LM de endoscopia digestiva alta e contatou-se CFU de menor magnitude. Naquele trabalho o CFU3 foi composto por 18 frases, o que representa uma grande uniformidade na descrição dos LM. Essa característica não é observada nos LM de colonoscopia, os quais contêm freqüentemente frases distintas, e não apenas termos, expressando um mesmo significado, isto é, atributos e valores iguais. Desse modo, para transformar as descrições em um formato adequado ao dicionário e ao ABP, foi necessário constituir o AP contendo 274 padronizações. Juntamente com especialistas foi definido que a construção do dicionário fosse realizada considerando apenas as classificações de local e característica. Desse modo, o dicionário foi constituído por 84 características distintas distribuídas em 34 locais e a BD foi composta por 277 atributos.

Finalizada a construção do dicionário e da BD iniciou-se o processamento dos LM na segunda fase. Para que as informações contidas nos 100 LM sejam integralmente mapeadas para os atributos da BD é necessário que o ABP identifique corretamente 967 valores distribuídos nos LM. Após o processamento, realizou-se uma análise no BD e foi constatado o mapeando de 793 valores, isto é, 82,00% do total. Essa análise permitiu observar que 57 LM tiveram uma percentagem de mapeamento igual ou superior a 80% e 34 foram mapeados integralmente para a BD. O mapeamento do conjunto de LM obteve uma média de 76,17% de preenchimento. Não foram preenchidos 174 valores, 18,00% do total. Destes, 134, o que representa 77,01%, se devem à forma de descrição das frases nos LM, as quais eram muito dependentes do contexto, isto é, para identificar os atributos de uma determinada frase era necessário memorizar e analisar as frases anteriores, o que demanda um algoritmo de busca e preenchimento muito complexo. Desse modo, para a completude da BD, os valores não mapeados automaticamente foram preenchidos manualmente, completando 100% do preenchimento esperado para a BD. Realizando uma análise do mapeando, observou-se que todos os valores preenchidos foram mapeados corretamente.

4. Conclusões

Neste trabalho foram apresentados uma metodologia de mapeamento de LM e um estudo de caso de laudos de colonoscopia. Foram propostas as técnicas de lematização e geração de n-gramas, as quais auxiliaram na identificação de frases únicas e de padrões contidos no conjunto de LM. Foi construído um dicionário do domínio capaz de mapear informações para uma BD composta por 277 atributos. Com o dicionário e a BD definidos foram mapeados de forma semi-automática os LM para a BD, o que representa uma redução no tempo de preenchimento da BD se comparado ao preenchimento manual bem como a possibilidade de um preenchimento padronizado e não subjetivo das informações contidas nos LM para a BD. Ao final, foi construída uma BD com as informações de 100 LM de colonoscopia, a qual pode ser disponibilizada para etapa de mineração de dados. Como trabalho futuro pretende-se adicionar a classificação de subcaracterísticas ao dicionário, o que possibilita o mapeamento de um maior número de atributos.

Agradecimentos

Agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq – e ao Programa de Desenvolvimento Tecnológico Avançado – PDTA/FPTI-BR.

Referências

B.Sc., P. E. R. (2002). *Matrix: A statical method and software tool for linguistic analysis through corpus comparison*. PhD thesis, Cumputing Department, Lancaster University, Lancaster.

Cotran, R. S., Kumar, V., and Collins, T. (2000). *Patologia Estrututural e Funcional*. Guanabara Koogan, USA.

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

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 - Brasil.

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 *Anais do XXXIII Conferencia Latinoamericana de Informática - CLEI (a ser publicado)*, San José - Costa Rica.

Honorato, 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 base de dados estruturadas a partir de laudos médicos. In *Encontro Internacional de Inteligência Artificial*, pages 593–601, São Leopoldo - Brasil.

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.

Orengo, M. and Huyck, C. (2001). A stemming algorithm for the portuguese language. In *SPIRE 2001*, IEEE Computer Society.

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann, San Diego - USA.

Quilici, F. A. (2000). *Colonoscopia*. Lemos-Editorial, São Paulo - Brasil.

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

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Diego - USA, 2 edition.