Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico

Maria Fernanda Moura

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 26/08/2009

Assinatura:_____

Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico

Maria Fernanda Moura

Orientadora: *Profª Drª Solange Oliveira Rezende*

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP como parte dos requisitos para obtenção do título de Doutor em Ciências de Computação e Matemática Computacional.

USP - São Carlos
AGOSTO/2009

A Roberto, Yasmin, Sayuri e Janda

"A verdadeira origem da descoberta consiste não em procurar por novas paisagens, mas em ter novos olhos." (Marcel Proust)

Agradecimentos

À minha família pelo apoio e compreensão sempre demonstrados. A meu marido, Roberto, que tantos vezes precisou ser "pai solteiro" durante o período do doutorado, mas cujo amor e apoio seguraram todas as crises; e, também, obrigada pelas dicas e contribuições técnicas a este trabalho. Às minhas filhas, Yasmin e Sayuri, que tão pequenas já ficaram "órfãs" por tantas vezes, mas aprenderam a lidar com o problema e do, jeito delas, apoiar os pais. A minha mãe, Jandira, foi babá de crianças e adultos ultimamente, cujos amor e apoio foram fundamentais durante toda a minha vida. A meu irmão, Carlos, que foi responsável pela minha paixão por livros, ele foi meu primeiro tutor acadêmico. E a meu pai, Fernando (*in memoriam*), que me ensinou o valor do trabalho.

À professora Solange Oliveira Rezende, pela confiança e orientação sem as quais eu não chegaria a esse ponto. Com certeza é difícil ensinar novos truques a velhos cães, mas ela tenta, tenta de novo e consegue. Ela é bem mais que uma orientadora, é uma mentora.

Ao Anandsing Dwarkasing, que foi tradutor de tantos textos, correções de vários deles e responsável por muitas das dicas de escrita, que foram do inglês para o portugês e vice-versa.

Ao pessoal do grupo de pesquisa em Mineração de Textos formado ao longo deste caminho, pelo trabalho, amizade e companheirismo. Ao Ricardo Marcondes Marcacini, nosso Ricardinho, que foi o braço direito e esquerdo em várias partes desta jornada. Ao Bruno Magalhães Nogueira e a Merley da Silva Conrado pelas infinitas discussões, noites de experimentos, validações, revisões e boas risadas. Ao Rafael Geraldelli Rossi companheiro de sofridas buscas a dados e transformações dos mesmos, aprendemos muito, ele é um verdadeiro "caçador de informação". Ao Fabiano Fernandes dos Santos, que já foi o caçula do grupo, mas que cedo passou a intervir com suas contribuições e idéias em tantos pontos importantes. À Tatiane Marcondes Nogueira, a novata de lógica nebulosa, que vai trazer uma pitada diferente aos nossos pratos.

Aos amigos da Embrapa Pecuária Sudeste, Pedro Franklin Barbosa, Sônia Borges de Alencar, Patrícia Menezes e Marcela Vinholis; que foram especialistas do domínio, validaram técnicas, sugeriram inovações e contribuíram muito com este trabalho.

Ao amigos da Embrapa Informática Agropecuária. À Adriana Delfino dos Santos e à Márcia Fugisawa Souza, em especial, pelo apoio anterior e durante todo o programa de doutorado, foram amigas, validadoras, auxiliares, indicaram dados, tiraram dúvidas. À

equipe da BDPA (Bases de Dados da Pesquisa Agropecuária), em especial à Leila Maria Lenk, que me ensinou a usar do modo adequado e interpretar bem os recursos da base; e, ao Isaque Vacari, Marcos Visoli e Maria Goretti Praxedes, que tanto contribuíram na formatação, compreensão e atualização dos dados utilizados. Antes do início do doutorado, o incentivo dos amigos Sílvia M F S Massruhá, Maria Angélica Leite, Sílvio R M Evangelista e Álvaro Seixas Neto foi indispensável à concretização do treinamento. E, não menos importante, à chefia, representada por Eduardo Delgado Assad, José Ruy P Carvalho e Kleber X S Souza, pelo apoio e prestatividades sempre presentes.

Aos amigos, colegas desse período na USP e também embrapianos, Luciana Alvim Romani, Adauto Luiz Mancini e Ana Lúcia Filardi, pelos bons momentos de convivência e apoio constante.

Aos ex-orientadores aos quais devo grande parte de minha formação. Raul Fernando Dada (*in memoriam*), orientador do colégio técnico (UNICAMP), a quem devo grande parte de minha formação como programadora. Ao meu ex-orientador de estágio e iniciação científica, já na Embrapa, Dalton Francisco Andrade, a quem eu devo o fortalecimento de minha formação em estatística experimental e método de pesquisa científica. Meu ex-orientador de mestrado, Mário Jino (FEE-UNICAMP), a quem eu devo o reconhecimento de como escrever uma dissertação, além de grande parte de minha formação como engenheira de software. A Fuad Gattaz Sobrinho, co-orientador extra-oficial de mestrado junto a Embrapa, com quem aprendi a valorizar pequenas tarefas como componentes do todo.

Aos amigos e professores do LABIC, pelo companheirismo e pelos bons momentos de convivência. Em especial às professoras Maria Carolina Monard, Roseli Francelin Romero e Agma Juci Machado Traina, das quais fui aluna e com as quais tive a oportunidade de ampliar minha visão de inteligência artificial e mineração de dados. Também aos companheiros do NILC, em especial a Eloize Rossi Marques Seno e ao Thiago Alexandre Salgueiro Pardo, que tantas vezes nos auxiliaram em validações e nos socorreram em conceitos. À professora Maria das Graças Volpe Nunes que fêz meu *curriculum* chegar às mãos da professora Solange Oliveira Rezende e continuou sempre colaborando com a nossa parceria.

Ao CNPq pelo financiamento dos projetos do Instituto Fábrica do Milênio - IFM e o Modelagem Computacional de Sistemas Complexos utilizando Mineração de Dados, Imagens e Textos (07/2007-Processo: 550963/2007-3) do LABIC.

Às funcionárias do setor de Pós-Graduação do ICMC/USP, Ana Paula, Beth, Laura e Lívia, pelos excelentes serviços prestados à comunidade acadêmica dessa unidade.

À Embrapa em geral, pelo apoio financeiro, mas, principalmente, pela manutenção de políticas de desenvolvimento de pessoas, cumprindo com o seu pessoal também de formadora de pessoas.

A todos que, direta ou indiretamente, contribuíram para a minha formação, profissional e pessoal, e conseqüentemente para a realização deste trabalho.

Resumo

A mineração de textos vem de encontro à realidade atual de se compreender e utilizar grandes massas de dados textuais. Uma forma de auxiliar a compreensão dessas coleções de textos é construir taxonomias de tópicos a partir delas. As taxonomias de tópicos devem organizar esses documentos, preferencialmente em hierarquias, identificando os grupos obtidos por meio de descritores. Construir manual, automática ou semi-automaticamente taxonomias de tópicos de qualidade é uma tarefa nada trivial. Assim, o objetivo deste trabalho é construir taxonomias de tópicos em domínios de conhecimento restrito, por meio de mineração de textos, a fim de auxiliar o especialista no domínio a compreender e organizar os textos. O domínio de conhecimento é restrito para que se possa trabalhar apenas com métodos de aprendizado estatístico não supervisionado sobre representações *bag of words* dos textos. Essas representações independem do contexto das palavras nos textos e, conseqüentemente, nos domínios. Assim, ao se restringir o domínio espera-se diminuir erros de interpretação dos resultados. A metodologia proposta para a construção de taxonomias de tópicos é uma instanciação do processo de mineração de textos. A cada etapa do processo propõem-se soluções adaptadas às necessidades específicas de construção de taxonomias de tópicos, dentre as quais algumas contribuições inovadoras ao estado da arte. Particularmente, este trabalho contribui em três frentes no estado da arte: seleção de atributos n-gramas em tarefas de mineração de textos, dois modelos para rotulação de agrupamento hierárquico de documentos e modelo de validação do processo de rotulação de agrupamento hierárquico de documentos. Além dessas contribuições, ocorrem outras em adaptações e metodologias de escolha de processos de seleção de atributos, forma de geração de atributos, visualização das taxonomias e redução das taxonomias obtidas. Finalmente, a metodologia desenvolvida foi aplicada a problemas reais, tendo obtido bons resultados.

Abstract

Text mining provides powerful techniques to help on the current needs of understanding and organizing huge amounts of textual documents. One way to do this is to build topic taxonomies from these documents. Topic taxonomies can be used to organize the documents, preferably in hierarchies, and to identify groups of related documents and their descriptors. Constructing high quality topic taxonomies, either manually, automatically or semi-automatically, is not a trivial task. This work aims to use text mining techniques to build topic taxonomies for well defined knowledge domains, helping the domain expert to understand and organize document collections. By using well defined knowledge domains, only unsupervised statistical methods are used, with a bag of word representation for textual documents. These representations are independent of the context of the words in the documents as well as in the domain. Thus, if the domain is well defined, a decrease of mistakes of the result interpretation is expected. The proposed methodology for topic taxonomy construction is an instantiation of the text mining process. At each step of the process, some solutions are proposed and adapted to the specific needs of topic taxonomy construction. Among these solutions there are some innovative contributions to the state of the art. Particularly, this work contributes to the state of the art in three different ways: the selection of n-grams attributes in text mining tasks, two models for hierarchical document cluster labeling and a validation model of the hierarchical document cluster labeling. Additional contributions include adaptations and methodologies of attribute selection process choices, attribute representation, taxonomy visualization and obtained taxonomy reduction. Finally, the proposed methodology was also validated by successfully applying it to real problems.

Esta tese foi preparada com o formatador de textos LaTeX. Foi utilizado um estilo (*style*) desenvolvido por Ronaldo Cristiano Prati. O sistema de citações de referências bibliográficas utiliza o padrão *Apalike* do sistema BibTeX.

Algumas palavras utilizadas neste trabalho não foram traduzidas da língua inglesa para a portuguesa por serem amplamente conhecidas e difundidas na comunidade acadêmica.

Sumário

Resumo i
Abstract iii
Sumário ix
Lista de Figuras xi
Lista de Tabelas xiv
Lista de Algoritmos xv
Lista de Abreviaturas xvii

1 Introdução **1**
1.1 Cenário e motivação 1
1.2 Objetivos 4
1.3 Hipóteses 4
1.4 Principais contribuições deste trabalho 5
1.5 Apresentação e resumo dos capítulos 6

2 Mineração de textos **7**
2.1 Introdução 7
2.2 Processo de mineração de textos 8
2.2.1 Identificação do problema 9
2.2.2 Pré-processamento 10
2.2.3 Extração de padrões 15
2.2.4 Pós-processamento e utilização do conhecimento 17
2.3 Considerações finais 18

3 Abordagem para a construção de taxonomias de tópicos em um domínio **19**
3.1 Introdução 19
3.2 Trabalhos relacionados à obtenção de taxonomias de tópicos por meio de mineração de textos 21
3.3 Descrição da metodologia proposta - TopTax 25
3.3.1 Identificação do problema 27
3.3.2 Pré-processamento 28
3.3.3 Extração de padrões 32
3.3.4 Pós-processamento 33

3.3.5 Uso do conhecimento 34
3.4 Estudos de casos com a metodologia TopTax 35
3.4.1 Exemplo no domínio lingüística computacional 35
3.4.2 Exemplo no domínio gado de corte Canchim 37
3.5 Considerações finais 38

4 Seleção de atributos do tipo n-gramas 41
4.1 Introdução 41
4.2 Contextualização do problema 43
4.2.1 Modelo de geração de n-gramas 44
4.2.2 Seleção de atributos unigramas 44
4.2.3 Seleção de atributos n-gramas 46
4.2.4 Atributos n-gramas e redundâncias 50
4.3 Método proposto para a seleção de atributos n-gramas 51
4.3.1 Passo 1 - geração de unigramas 52
4.3.2 Passo 2 - filtro e seleção de *stopwords* da coleção 52
4.3.3 Passo 3 - geração dos n-gramas ($n > 1$) 53
4.3.4 Passo 4 - seleção dos n-gramas e eliminação de redundâncias 54
4.4 Avaliação do método proposto 57
4.5 Experimentos e avaliação dos resultados 58
4.6 Vantagens do método para a seleção de atributos n-gramas 60
4.7 Considerações Finais 60

5 Modelos de rotulação em agrupamentos hierárquicos de documentos 63
5.1 Introdução 63
5.2 Revisão bibliográfica de modelos de rotulação 64
5.3 Métodos desenvolvidos para rotulação hierárquica de agrupamentos 67
5.3.1 Idéia geral de rotulação de agrupamentos 67
5.3.2 Seleção hierárquica de rótulos 69
5.3.3 O método *Robust Labeling Down* - RLDM 71
5.3.4 O método *Robust Labeling Up* - RLUM 75
5.4 Validação do processo de rotulação 79
5.4.1 Avaliação subjetiva da rotulação 80
5.4.2 Avaliação objetiva da rotulação 81
5.5 Resultados experimentais da avaliação subjetiva 86
5.5.1 Delimitação das coleções de textos e pré-processamento 87
5.5.2 Agrupamento hierárquico e rotulação 88
5.5.3 Análise dos resultados 88
5.5.4 Considerações da avaliação subjetiva 91
5.6 Resultados experimentais da avaliação subjetiva integrada a critérios objetivos 92
5.6.1 Obtenção das taxonomias 92
5.6.2 Avaliação subjetiva 93
5.6.3 Avaliação objetiva 94

5.6.4 Considerações sobre a avaliação subjetiva integrada à objetiva . . . 94
5.7 Resultados experimentais da avaliação objetiva 95
5.7.1 Pré-processamento, agrupamento e rotulação 95
5.7.2 Avaliando os rótulos como expressões de busca 96
5.7.3 Avaliando a cobertura do vocabulário 101
5.7.4 Avaliação da representatividade do vocabulário produzido 102
5.8 Considerações finais 103

6 Conclusões e trabalhos futuros 105
6.1 Contribuições do trabalho 105
6.1.1 Contribuições diretas desta tese 105
6.1.2 Contribuições em parcerias 107
6.1.3 Aplicações práticas da TopTax 108
6.2 Principais limitações 108
6.3 Trabalhos futuros 109

Referências 123

A Apêndice - Tabelas de comparação das medidas de interesse de resultados de EMIM e busca 125
A.1 Comparações múltiplas de médias de EMIM 125
A.2 Conjuntos de rótulos como expressões de busca 128

Lista de Figuras

2.1 Etapas do processo de mineração de textos 8

3.1 Metodologia para obtenção de taxonomia de tópicos baseada em mineração de textos 27
3.2 Exemplo de visualização, antes (1) e depois (2) de retiradas *stopwords* específicas da coleção 36
3.3 Visualização da taxonomia inicial para Gado Canchim, termos do nó grifado e texto vinculado 38

4.1 Passos do método proposto para a seleção de atributos n-gramas 52

5.1 Hierarquia inferida sobre a coleção de textos do exemplo ideal 68
5.2 Resultado da aplicação do método de rotulação Mais Freqüentes sobre a coleção de textos do exemplo 69
5.3 Hierarquia genérica de uma taxonomia 70
5.4 Ilustração do método de rotulação RLDM 71
5.5 Exemplo de rotulação com o método de Popescul & Ungar 75
5.6 RLUM aplicado à coleção de textos do exemplo 78
5.7 Hierarquia informática agropecuária: comparação dos resultados para o método Mais Freqüentes e RLDM 88
5.8 Taxonomia parcial obtida a partir da coleção de inteligência artificial . . . 93
5.9 F_{score} do processo de recuperação usando apenas os 10 primeiros elementos das listas de rótulos específicas em cada coleção 94
5.10 Medida F – $L_s(n_{ij})$ 99
5.11 Medida F – $L_g(n_i) \bigcup L_s(n_{ij})$ 99
5.12 Medida *Recall* – $L_g(n_i) \bigcup L_s(n_{ij})$ 99
5.13 Medida *Recall* – $L_s(n_{ij})$ 100
5.14 Precisão – $L_s(n_{ij})$ 100
5.15 Precisão – $L_g(n_i) \bigcup L_s(n_{ij})$ 100

Lista de Tabelas

2.1 Padrão de representação da matriz atributo×valor 12

3.1 Termos para 17 artigos científicos sobre gado de corte Canchim 37

4.1 Coleção de cinco textos fictícios utilizados como exemplo 43
4.2 Exemplo de atributos n-gramas gerados para a coleção de cinco textos . . . 45
4.3 Atributos unigramas da Tabela 4.2 após corte $df \geq 2$ 46
4.4 Bigramas e trigramas da Tabela 4.2 após filtro dos unigramas 47
4.5 Freqüências com w_1 na primeira e w_2 na segunda posição do bigrama w_1w_2 47
4.6 Freqüências para o bigrama "data_transform" 48
4.7 Freqüências para o bigrama "mine_textual" 48
4.8 Exemplo com a freqüência de w_1 e w_2w_3 em $w_1w_2w_3$ 49
4.9 Exemplo com a freqüência de w_1w_2 e w_3 em $w_1w_2w_3$ 49
4.10 Freqüências para o trigrama "data" com "_process_transform" 50
4.11 Freqüências para o trigrama "data_process" com "_transform" 50
4.12 Seleção de n-gramas não redundantes, fonte: Moura et al. (2008d) 51
4.13 Extração de n-gramas 53
4.14 Freqüências de w_1 e w_2w_3 em $w_1w_2w_3$ 54
4.15 w_1w_2 nas primeiras e segundas posições e w_3 nas terceiras posições de todos os trigramas da coleção 55
4.16 Freqüências de w_1 e w_2 no bigrama w_1w_2 56
4.17 Resultados dos primeiro, segundo, terceiro e quarto passos do modelo proposto sobre os dados experimentais 58
4.18 Atributos n-gramas não redundantes obtidos dos dados experimentais . . . 59
4.19 Atributos n-gramas apenas *rankeados* obtidos dos dados experimentais . . 59
4.20 Resultados do classificador C4.5 para testar $H_0 : \overline{acc_r} = \overline{acc_{nr}}$ 59
4.21 Resultados do classificador SVM para testar $H_0 : \overline{acc_r} = \overline{acc_{nr}}$ 59

5.1 Tabela de contingência para $f_i(a_k)$ 70
5.2 Resultados dos passos iniciais dos algoritmos dos métodos RLDM e Popescul & Ungar para o termo *Embrapa* 74

5.3 Resultados dos passos iniciais dos algoritmos do RLUM para o termo *Embrapa* . . . 79
5.4 Tabela de dados para avaliação subjetiva dos métodos de rotulação 80
5.5 Número de recuperações em cada n_i . . . 82
5.6 Número de *unigramas* antes e depois dos filtros utilizados, na análise subjetiva 87
5.7 Comparação múltipla de médias das notas considerando os efeitos dos métodos . . . 90
5.8 Comparação múltipla de notas com o modelo $\hat{g} = \hat{\mu} + \hat{a} + \hat{m(a)} + \hat{c(m)} + \hat{e}$ no domínio Canchim . . . 90
5.9 Comparação múltipla de notas com o modelo $\hat{g} = \hat{\mu} + \hat{a} + \hat{m(a)} + \hat{c(m)} + \hat{e}$ no domínio Informática Agropecuária . . . 91
5.10 Médias da notas ($\overline{g}$) e seus desvios padrões (se) para cada método (m) . . 93
5.11 Número de documentos e atributos das coleções de textos utilizadas 97
5.12 Resultados das medidas de interesse para **WP01** . . . 98
5.13 Cardinalidade dos vocabulários obtidos . . . 102
5.14 Comparação múltipla de médias para cardinalidades dos vocabulários . . . 102
5.15 EMIM - resultados da comparação múltipla de suas médias . . . 103

A.1 EMIM - resultados para o domínio de Ciência da Computação . . . 126
A.2 EMIM - resultados para o domínio de Química . . . 126
A.3 EMIM - resultados para o domínio de Física . . . 127
A.4 EMIM - resultados para as coleções de textos do IFM . . . 127
A.5 Resultados das medidas de interesse para **WP01 e WP02** . . . 131
A.6 Resultados das medidas de interesse para **WP03 e WP04** . . . 132
A.7 Resultados das medidas de interesse para **Hard e IA** . . . 133
A.8 Resultados das medidas de interesse para **HCI e SEC** . . . 134
A.9 Resultados das medidas de interesse para **AnaC e InoC** . . . 135
A.10 Resultados das medidas de interesse para **OrgC e Poly** . . . 136
A.11 Resultados das medidas de interesse para **BioP e GeoP** . . . 137
A.12 Resultados das medidas de interesse para **Mech e Quan** . . . 138

Lista de Algoritmos

1 Seleção de trigramas 55
2 Seleção de bigramas 56
3 RLDM: controle do conjunto de atributos 73
4 RLDM: teste da hipótese de independência 73
5 RLDM: realizar o corte semântico 74
6 RLUM: controlar as visitas aos nós da árvore 76
7 RLUM: selecionar a_k em $L_s(ni)$ ou $L_s(n_{ij})$ 77
8 RLUM: realizar o corte semântico 77

Lista de Abreviaturas

df	*document frequency*
EMIM	*Expected Mutual Information Measure*
idf	*inverse document frequency*
IFM	Instituto Fábrica do Milênio
IESystem	*Information Extraction System*
LABIC	Laboratório de Inteligência Computacional
LSA	*Latent Semantic Analysis*
Luhn-DF	Adaptação do método de Luhn para *document frequency*
MT	Mineração de Textos
NSP	*N-gram Statistical Package*
PCA	*Principal Component Analysis*
PreText	Ferramenta para Pré-Processamento de Textos
RLDM	*Robust Labeling Down Method*
RLUM	*Robust Labeling Up Method*
SVM	*Support Vector Machines*
TC	*Term Contribution*
TopTax	Metodologia *Topic Taxonomy*
tf	*term frequency*
tf-idf	*term frequency-inverse document frequency*
TV	*Term Variance*
TVQ	*Term Variance Quality*
XML	*eXtensible Markup Language*

Capítulo

1

Introdução

Esta tese de doutorado tem como tema principal uma abordagem para a construção de taxonomias de tópicos em domínios de conhecimento restritos. A abordagem é baseada em uma compilação de trabalhos amplamente divulgados na literatura, trazendo como principal novidade a integração de técnicas em etapas muito bem definidas, com várias contribuições inovadoras ao longo das mesmas. As contribuições inovadoras concentram-se em seleção de atributos multi-palavras, rotulação de agrupamento hierárquico de documentos e validação dos modelos de rotulação de agrupamento hierárquico de documentos. Neste capítulo, faz-se uma apresentação geral do cenário, englobando o estado da arte e as principais motivações para o trabalho. Na seqüência são descritos os objetivos e hipóteses que regem esta tese. A seguir resumem-se as principais contribuições obtidas com o trabalho. E, finalmente, apresenta-se a organização desta tese.

1.1 Cenário e motivação

Com o avanço e a facilidade de acesso à tecnologia da informação, cada vez mais diversas organizações dos mais variados setores têm utilizado sistemas digitais de armazenamento de dados. De acordo com estimativas realizadas em 2006, a quantidade de dados no universo digital atingia 161 *hexabytes*, o que correspondia a três milhões de vezes a quantidade de informação contida em todos os livros já escritos. No período de 2006 a 2010, estima-se que o crescimento de informações no universo digital será de 161 *hexabytes* para 988 *hexabytes* (Gantz et al., 2007).

Ante a essa grande quantidade de dados, a capacidade humana de analisá-los, compreendê-los e utilizá-los adequadamente é excedida. Dessa forma, grande parte da informação e do conhecimento potencialmente útil que poderia ser recuperado a partir dessas bases é perdida. Como conseqüência, o desenvolvimento de ferramentas e técnicas de extração de conhecimento de repositórios de dados tem se tornado cada vez mais indispensável, dando

origem aos processos computacionais de extração automática de conhecimento conhecidos como mineração de dados[1] (Witten e Frank, 2005).

Dessa gama de dados no universo digital, grande parte está disponível em formato não-estruturado, principalmente no formato textual, em forma de relatórios, boletins, artigos, *emails* e toda sorte de documentos. Para dados dessa natureza, a utilização de algoritmos de mineração de dados convencional é inviável, uma vez que estes algoritmos requerem dados em uma estrutura prévia para deles extrair conhecimento. Surge, então, uma especialização do processo de mineração de dados, denominada mineração de textos, a qual tem como principal diferença a incorporação de atividades no sentido de estruturar os dados originalmente não-estruturados em um formato adequado para a extração automática de conhecimento.

A mineração de textos vem tornando possível transformar esse grande volume de dados textuais em informação e conhecimento útil, muitas vezes inovador para as empresas. As aplicações de mineração de textos permitem extrair informação e conhecimento, fornecendo elementos de suporte à gestão do conhecimento. Essas aplicações são um modo de reorganizar como o conhecimento é criado, usado, compartilhado, armazenado e avaliado. Tecnologicamente, o apoio de mineração de textos à gestão do conhecimento se dá na transformação do conteúdo de repositórios de dados em informação e conhecimento a ser analisado e compartilhado pela organização. O futuro aproveitamento desse conhecimento gerado pode se dar em várias aplicações que apóiam a tomada de decisões, desde a simples organização e recuperação de informação até sistemas mais complexos de apoio à decisão.

Há muitas linhas de trabalhos na literatura que buscam tópicos entre documentos utilizando mineração de textos. Grande parte deles visa a melhoria de processos de busca, especialmente busca *web*. Nesses casos, por muitas vezes, a perda de detalhes de informação não é tão importante, pois os resultados serão julgados *a posteriore* quando um usuário os recupera. No entanto, outros trabalhos buscam pela prospecção de tópicos entre os dados textuais, tanto com o objetivo de melhorar processos de busca quanto de construir mapas de tópicos. Esses mapas de tópicos, ou taxonomias de tópicos, formam relações conceituais mais genéricas entre os dados, dividindo-os em categorias, por vezes hierárquicas.

Uma organização hierárquica de dados textuais sempre auxilia a compreensão do todo. Um exemplo é a construção de portais para organizar e navegar em publicações e outras informações; por exemplo, a Agência de Informação Embrapa, ou simplesmente Agência. A Agência é um portal *Web* composto por várias agências de produtos, que organizam informação técnica relevante para o agronegócio, especializada por produto e estruturada basicamente sob a óptica da cadeia produtiva do agronegócio. O objetivo é atender a perfis diversificados de consumidores de informação, como produtores rurais, extensionistas, pesquisadores, técnicos, professores e estudantes (Evangelista et al., 2003). A metodologia da Agência traz um diferencial importante na organização da informação, pois implementa uma solução desenvolvida especialmente para ela, que estrutura o conhecimento de uma

[1]Em todo este trabalho, os conceitos de processos de mineração de dados e processos KDD (*Knowledge Discovery in Database*) são usados indistintamente, tal como sugerido por Rezende et al. (2003).

cadeia produtiva de uma forma hierárquica, chamada de árvore do conhecimento. Nos primeiros níveis desta árvore estão os conhecimentos mais genéricos e, nos níveis mais profundos, estão os conhecimentos específicos. Cada nó desta árvore contém um texto sobre um tema que é resultante da compilação do conhecimento produzido por pesquisadores, técnicos extensionistas e agricultores, bem como referências a outras obras que complementam a informação. Mesmo com uma metodologia mais consolidada, construir essas árvores e ainda disponibilizar um portal completando-as com referências a outras fontes de informação, não é um trabalho trivial. Essas tarefas demandam uma equipe bastante especializada que precisa trabalhar com um grande volume de informação disponível em bases de dados textuais, cujo processo de identificação, seleção, classificação e síntese é bastante oneroso. Essa dificuldade de atingir consenso metodológico e a grande quantidade de dados textuais disponíveis, nem sempre de qualidade e de real interesse, cria a demanda por ferramentas, automatizadas ou semi-automatizadas, que analisem os textos originais, de modo a auxiliar a sua organização, filtrando o que é de fato útil para o portal.

A motivação que originou o trabalho aqui apresentado é suprir a necessidade de uma metodologia para a organização de informação a partir de um processo de mineração de textos. Embora houvesse a proliferação de muitas soluções individuais, para problemas intermediários do processo de mineração de textos, havia pouca ou nenhuma metodologia que integrasse as etapas do processo com soluções direcionadas aos objetivos. Assim, tecnologicamente, o objetivo deste trabalho é o desenvolvimento de uma metodologia que auxilie o especialista do domínio a obter uma taxonomia de tópicos de qualidade, por meio de mineração de textos. Busca-se, prioritariamente, organizar a informação de um domínio restrito e fornecer ferramentas que durante o processo de prospecção dos tópicos, permita estudá-los dos mais específicos para os mais genéricos. A fim de garantir a qualidade dos resultados e a utilidade da metodologia para atingir os objetivos, o procedimento todo é realizado com a intervenção de um especialista do domínio de conhecimento.

Ao longo do desenvolvimento deste projeto foram surgindo soluções com objetivos similares às proposições iniciais, por exemplo, o ambiente TaxaMiner (Kashyap et al., 2005a), que realiza a prospecção de uma taxonomia em uma coleção de textos, e a ferramenta WebTopic (Escudeiro e Jorge, 2005) que procura encontrar tópicos em coleções de textos hipermídia dinâmicas. E, em consonância com essas idéias, hoje ainda há muitas proposições surgindo, como o ambiente *ACORN - Automatic Domain Specific Ontology Construction Process* de Bae et al. (2008), no qual a prospecção da taxonomia apóia a construção de ontologias. Embora existam muitas técnicas, desde a conversão dos dados originais em textos sem formatação até a extração de algum mapa de tópicos, integrá-las e adaptá-las para funcionar junto a intervenções humanas e produzir um resultado satisfatório tem sido um desafio.

Do ponto de vista acadêmico, o desenvolvimento da metodologia exercita todo o processo de mineração de textos, em um projeto de pesquisa e desenvolvimento, o que permite investigar técnicas existentes e propor inovações, quando necessário. Como a mineração de textos é uma área relativamente nova, há muitos pontos em aberto. Para a prospecção de mapas de tópicos a partir de coleções de documentos das quais não se tenha informações *a priori*, como um *thesaurus* ou alguma categorização de seus documentos, a técnica mais

utilizada é a clusterização hierárquica de documentos. Em torno da questão clusterização de documentos há vários pontos de atenção, desde a seleção de documentos significativos como a coleção inicial de documentos até a avaliação do conhecimento gerado. E, em termos de mineração de textos há muita pesquisa concentrada nos principais problemas, que correspondem à seleção de atributos, à forma de representação dos atributos, às técnicas e algoritmos de agrupamento, à discriminação dos grupos, à visualização de resultados e às técnicas de validação. Esses pontos foram todos cobertos ao longo da metodologia, procurando pelo estado da arte em sua resolução e propondo adaptações e novos modelos em cada qual.

1.2 Objetivos

O principal objetivo é o desenvolvimento de uma metodologia, utilizando técnicas de mineração de textos, para auxiliar o especialista a organizar hierarquicamente a informação textual do domínio e fornecendo a cada etapa critérios que o auxilie a intervir no processo de modo a melhorar os resultados obtidos.

Como se está tratando com uma área relativamente nova, carente de soluções específicas, e trabalhando-se com dados que por si só carregam interpretação semântica, os objetivos mais específicos são:

1. obter um conjunto de métodos e parâmetros para identificar e delimitar o vocabulário do domínio presente na coleção de textos;

2. obter métodos adequados de filtros de atributos para diminuir a característica esparsa da matriz de dados, procurando não perder representatividade;

3. desenvolver um método eficiente de seleção de termos discriminativos para cada agrupamento obtido, que consiga representá-los de maneira única;

4. desenvolver soluções de avaliação visual para os grupos, permitindo explorá-los e estudá-los; e

5. auxiliar a organização de bases de dados textuais, para as quais precisa-se identificar uma taxonomia e palavras-chaves para cada grupo e documento.

1.3 Hipóteses

A principal hipótese é que a metodologia validada e avaliada atinja o objetivo de auxílio à organização da informação, permitindo realizar essa tarefa com eficácia.

As demais hipóteses decorrem dos objetivos específicos:

- uma melhor estimativa do vocabulário representado na coleção leva a um melhor conjunto de atributos, logo, a uma melhor extração de conhecimento;

- existe alguma representação dos atributos que uma avaliação subjetiva possa julgar como melhor;

- a redução do número de atributos não afetará negativamente o resultado da extração de conhecimento;
- a redução do número de atributos pode melhorar o desempenho dos algoritmos de extração de conhecimento;
- consegue-se representar unicamente cada grupo obtido por meio da seleção dos termos discriminativos;
- coleções textuais podem gerar grandes hierarquias que precisam ser podadas;
- visualizar as hierarquias em diferentes níveis de granularidade auxilia a compreensão da coleção textual;
- obtêm-se, no mínimo, metadados que podem ser usados para classificar os documentos e auxiliar a organização de informação.

1.4 Principais contribuições deste trabalho

As principais contribuições deste doutorado estão relacionadas à originalidade dos métodos propostos e relevância dos resultados computacionais e práticos em três grandes frentes: construção das taxonomias de tópicos, seleção de atributos para as taxonomias de tópicos e rotulação de agrupamentos hierárquicos de documentos.

O primeiro resultado é o desenvolvimento da metodologia para a construção de taxonomias de tópicos em um domínio, denominada TopTax- *Topic Taxonomy*. A metodologia TopTax é uma compilação de trabalhos amplamente divulgados na literatura, trazendo como principal novidade a integração de técnicas em etapas muito bem definidas. Com isso pretendeu-se fornecer um guia de construção das taxonomias, cobrindo desde a seleção dos textos utilizados até a avaliação dos resultados obtidos. Para cada processo envolvido no detalhamento das etapas foram desenvolvidas soluções tecnológicas específicas em sua maioria com contribuições inéditas para o estado da arte.

A segunda contribuição inédita para o estado da arte compete à seleção de atributos n-gramas. Em se tratando de processos de mineração de textos, os atributos carregam algum significado semântico, pois são palavras ou composições de palavras. Os atributos auxiliam os especialistas a interpretarem os resultados intermediários em cada etapa da metodologia e, conseqüentemente, o uso do conhecimento produzido. Assim, é necessário trabalhar com atributos multi-palavras, isto é, compostos por uma ou mais palavras, por exemplo: "tese doutorado", "tese", "doutorado", "tese doutorado inteligência artificial", 'inteligência", "artificial" e "inteligência artificial". Para gerar esses atributos, geram-se outros menos importantes, como "doutorado inteligência", e outros redundantes, por exemplo, apenas "tese" pode não ter peso algum na compreensão geral do assunto em questão. Assim, uma das contribuições é um modelo de seleção de atributos desse tipo utilizados conjuntamente, de modo a remover-lhes redundâncias e atributos de menor importância.

A terceira contribuição para o estado da arte encontra-se na forma de rotular os agrupamentos obtidos. A taxonomia é baseada em um agrupamento hierárquico de documen-

tos. Obter bons descritores para o agrupamento é um desafio que vem sendo tratado por diferentes abordagens. Nesta tese são propostos dois novos modelos para a obtenção dos descritores do agrupamento, sendo o segundo deles o de menor complexidade computacional e de melhor representação dos descritores. Além dos modelos de rotulação de agrupamentos, foram estudados e desenvolvidos modelos para análise subjetiva e objetiva dos resultados, análise subjetiva integrada à análise objetiva e um modelo de análise objetiva que foi o mais satisfatoriamente empregado.

Além dessas, há outras contribuições acadêmicas relacionadas a algumas etapas da metodologia em co-autoria com outros trabalhos, relacionadas no Capítulo 6. E, também a TopTax foi utilizada em alguns problemas reais, o que auxiliou a validação e evolução da mesma.

1.5 Apresentação e resumo dos capítulos

Para avaliar as hipóteses e cumprir os objetivos especificados, apresentando o detalhamento das contribuições deste trabalho, a continuação do texto desta tese divide-se em mais cinco capítulos, os quatro seguintes referentes ao corpo do trabalho e o sexto e último às conclusões e trabalhos futuros.

No Capítulo 2, mineração de textos, procura-se dar uma visão geral da área, processo genérico e quais as técnicas de reconhecimento de padrões mais utilizadas em cada etapa. No Capítulo 3 mostra-se como o processo de mineração de textos foi instanciado para chegar a metodologia de obtenção de taxonomias de tópicos. Nessa instanciação houve a preocupação em se separar objetivos tecnológicos e acadêmicos, gerando contribuições nos dois sentidos. Ainda, no Capítulo 3 também são mostrados alguns dos exemplos de aplicação da metodologia. Nos Capítulos 4 e 5 são descritas as contribuições academicamente inéditas desta tese de doutorado. No Capítulo 4, seleção de atributos do tipo n-gramas, o modelo proposto para a utilização conjunta de atributos multi-palavras é descrito e avaliado. No Capítulo 5 são descritos e avaliados os modelos desenvolvidos para a rotulação hierárquica de agrupamentos de documentos. Finalmente, no Capítulo 6 são resumidas as conclusões do trabalho, as limitações hoje apresentadas e uma possível lista de trabalhos futuros, dentre os quais alguns já em andamento.

Ainda, procurou-se dar independência aos capítulos, de modo que eles possam ser lidos quase que individualmente. Para isso, optou-se muitas vezes por repetir alguns conceitos de um capítulo para outro.

Capítulo

2

Mineração de textos

Mineração de textos (MT) é uma especialização da mineração de dados. É uma área interdisciplinar relativamente nova que engloba: processamento de língua natural, mais particularmente a lingüística computacional; aprendizado de máquina; recuperação de informação; análise estatística de dados; e, visualização de informação. O objetivo deste capítulo é apresentar os principais conceitos envolvidos em MT bem como algumas técnicas utilizadas. As técnicas de reconhecimento de padrões são citadas em cada etapa de MT, conforme a necessidade. Dentre as técnicas utilizadas, dá-se maior atenção a aprendizado de máquina estatístico não supervisionado que é o utilizado no desenvolvimento do trabalho.

2.1 Introdução

Em um contexto em que grande parte dos dados corporativos encontra-se disponível em forma textual, o processo de mineração de textos surge como uma poderosa ferramenta de apoio à gestão de conhecimento. Ebecken et al. (2003) definem a mineração de textos como "um conjunto de técnicas e processos que descobrem conhecimento inovador nos textos". Nesse sentido, o objetivo da prática de mineração de textos centra-se na busca por padrões, tendências e regularidades em documentos escritos em linguagem natural.

Pode-se afirmar que a mineração de textos é uma especialização do processo de mineração de dados. A principal diferença entre os dois processos é que, enquanto a mineração de dados convencional trabalha exclusivamente com dados estruturados, a mineração de textos lida com dados inerentemente não-estruturados (Weiss et al., 2005). Dados estruturados são, por exemplo, tabelas de bancos de dados, nas quais tem-se uma separação definida dos dados, metadados e seus relacionamentos previamente conhecidos. Logo, na MT o primeiro desafio é obter alguma estrutura a partir dos textos e então a partir desta tentar obter conhecimento.

Neste trabalho, os textos que não obedecem a um padrão de formatação são chama-

dos **não-estruturados**; os que seguem algum padrão, como textos científicos que são divididos em tópicos pré-definidos, ou livros divididos em capítulos, etc, são tratados como **semi-estruturados**; e, **textos estruturados** são aqueles representados em uma linguagem de marcação[1]. Nesses textos estruturados, os metadados e seus conteúdos originais estão padronizados no mesmo arquivo de dados e as relações também são representadas pelos marcadores.

O processo de mineração de textos pode ser dividido em uma seqüência de etapas genéricas, formando um ciclo no qual, ao final, obtém-se o conhecimento acerca dos dados analisados. Finalmente, o processo de MT pode ser instanciado de acordo com o conhecimento buscado em cada aplicação, por exemplo, para extrair taxonomias de tópicos a partir de bases textuais. Assim, nas próximas seções e subseções são divididas para dar uma idéia geral do processo de MT, requisitos e cuidados em cada etapa, bem como possíveis técnicas de reconhecimento de padrões a serem utilizadas.

2.2 Processo de mineração de textos

O processo de mineração de textos, tal como o processo de mineração de dados, é formado por cinco grandes etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento. Na Figura 2.1 é ilustrado o ciclo formado por essas etapas, as quais são discutidas nas próximas seções. Também são discutidas as possíveis técnicas de reconhecimento de padrões utilizáveis em cada etapa.

Figura 2.1: Etapas do processo de mineração de textos. Fonte: Rezende et al. (2003)

[1]Em geral, usa-se a linguagem de marcação XML - *eXtended Markup Language*.

2.2.1 Identificação do problema

Nessa etapa, delimita-se o escopo do problema com o qual irá trabalhar-se, definindo o objetivo da aplicação do processo de mineração de textos. Basicamente, o analista define a base de textos com a qual irá trabalhar, o que se espera obter com a análise dos dados e como o resultado da análise pode ser utilizado no contexto de sua organização. Nesse sentido, Rezende et al. (2003) definem quatro questões essenciais a serem respondidas nessa etapa:

- Quais são as principais metas do processo?
- Quais critérios de desempenho são importantes para este problema?
- O conhecimento extraído deve ser compreensível a seres humanos ou um modelo do tipo caixa-preta também é apropriado?
- Qual deve ser a relação entre simplicidade e precisão do modelo extraído?

As respostas a estas questões representam decisões que guiarão os passos consecutivos e terão reflexo no desempenho da aplicação. É importante que, desde essa primeira etapa, as decisões tomadas estejam em concordância com o domínio do problema trabalhado e o objetivo final do processo de extração de conhecimento. Nesse sentido, um ponto importante a ser observado já nessa etapa é se existe necessidade de incorporação do conhecimento do domínio para apoiar a tomada de decisões. Essa incorporação do conhecimento pode ser feita pela constante interação com especialistas no domínio, além de um estudo acerca do domínio do problema a fim de melhor compreender o meio com o qual se lida.

Outra forma de incorporar informação acerca do domínio é utilizar **vocabulário controlado**, **taxonomias**, ***thesaurus*** ou **ontologias**, se algum desses for disponível. Um **vocabulário controlado** é uma coleção de termos de um domínio de conhecimento mantida por alguém que reconhecidamente detém o controle de novos termos e de suas regras de inclusão. Todos os termos dessa coleção devem ter definições não ambíguas, podendo ou não lhes ser associados também significados (Pidcok, 2003). A composição desse vocabulário corresponde a uma relação de termos; na qual, um termo pode ser representado por um conjunto de outros termos que tenham preferência de uso em relação a ele, por exemplo, um sinônimo. Também deve ser verificada a disponibilidade de alguma organização hierárquica mais formalizada de vocabulário controlado, como uma taxonomia, um *thesaurus* ou uma ontologia. Uma **taxonomia** é uma coleção de vocabulários controlados, organizados hierarquicamente, isto é, com relações do tipo pai-filho entre eles (Garshol, 2004; Cimiano, 2006), que podem ser tipificadas como: todo-parte, gênero-espécie, tipo-instância, etc. Um ***thesaurus*** pode ser definido como um vocabulário controlado que representa sinônimos, hierarquias e relacionamentos associativos entre termos (Rychlý e Kilgarriff, 2007). Em um *thesaurus*, além das relações hierárquicas são também apresentadas relações associativas. Ou seja, o *thesaurus* precisa ser representado por um grafo mais complexo, como uma rede, onde as relações pai-filho são tratadas como *broader-term* e *narrower-term*, e as associativas como *related terms* e termo controlado como *use*. Uma

ontologia é uma estruturação hierárquica de conhecimentos sobre coisas que podem ser subcategorizadas de acordo com sua essência, ou ainda, é uma compreensão compartilhada de algum domínio de interesse (Chandrasekaran et al., 1999). Nas ontologias, além de todas as relações esperadas entre os termos, descrevem-se os conceitos que formam sua essência. Existem linguagens formais para se descrever ontologias, com regras gramaticais que restringem o uso dos vocabulários controlados (que são os termos léxicos dessas gramáticas) para expressar algo significativo no domínio de interesse; o quê recebe o nome de *meta model* (Pidcok, 2003).

Outro aspecto de grande importância é a definição da coleção de documentos a ser utilizada. Os documentos utilizados no processo de extração de conhecimento devem representar o domínio do problema tratado de maneira a mais completa possível, a fim de que o conhecimento extraído seja representativo e consiga alcançar o objetivo do processo. Deve-se observar a representatividade desses documentos ao domínio e à aplicação do conhecimento a ser extraído. Esses textos podem ser obtidos de diversas fontes, como livros e artigos disponíveis na forma eletrônica e documentos da *Web*. Essa é uma atividade crítica, uma vez que os textos podem não estar disponíveis no formato adequado, como documentos não-digitalizados ou digitalizados com baixa qualidade. Ainda, se o objetivo é classificar os documentos, deve-se ater à existência ou não de documentos previamente classificados - rotulados.

2.2.2 Pré-processamento

A etapa de pré-processamento é uma das etapas que mais tempo consome durante todo o ciclo do processo de mineração de textos. É nessa etapa que se encontra a principal diferença de processos de mineração de textos para processos de mineração de dados: a estruturação dos textos em um formato adequado para a extração de conhecimento. Em outras palavras, essa etapa é responsável por extrair de textos escritos em língua natural, inerentemente não estruturados, uma representação estruturada, representativa, concisa e manipulável por algoritmos de extração automática de conhecimento. Para tal, essa etapa compreende atividades de tratamento da base de documentos, geração de termos, estruturação dos documentos e redução da dimensionalidade dos termos.

Tratamento da base de documentos

Uma vez que uma coleção de documentos pode ser uma junção de documentos coletados de diversas fontes diferentes, deve-se, nesse ponto, efetuar um tratamento da base de documentos a fim de assegurar a confiabilidade, não-redundância e o balanceamento da coleção de documentos. Para atender a estes aspectos, uma série de ações podem ser tomadas quando necessárias, dentre as quais destacam-se:

- Identificação da qualidade da digitalização dos documentos (quando for o caso);
- Eliminação de repetições de documentos;
- Balanceamento da coleção por reamostragem, quando o objetivo é classificar;

- Redução da quantidade de documentos, quando o objetivo assim permitir;
- Verificação da existência de uma estrutura prévia nos documentos, a fim de utilizar essa informação na estruturação final da coleção;
- Análise do tamanho dos documentos na coleção, verificando a necessidade de uma normalização dos pesos atribuídos aos termos em função do tamanho dos textos.

Geração dos termos

Uma vez definida a coleção de documentos com a qual se irá trabalhar, passa-se à geração dos **termos de domínio**, que serão tomados como atributos que descrevem os documentos. Por **termo** entende-se uma seqüência de *tokens*, simples ou composta, com significado semântico único. Quando um termo é composto por um único *token*, este é conhecido como unigrama ou termo simples. Por sua vez, quando um termo contém mais de um *token*, denomina-se *n-grama* ou termo composto.

De modo geral, as palavras presentes nas coleções textuais são analisadas e tratadas, de maneira a gerar termos correspondentes aos atributos mais consistentes para representar os documentos. A **análise dos termos** pode ser feita segundo duas abordagens: semântica ou estatística. Na **análise semântica**, buscam-se as palavras baseadas no seu contexto, utilizando, para isso, de técnicas de Processamento de Língua Natural (PLN), sendo, portanto, dependente do idioma. Já na **análise estatística**, os termos são avaliados por medidas quantitativas ou qualitativas relativas à presença deles nos textos. As **medidas qualitativas** correspondem à presença ou ausência do termo no texto, são as medidas *booleanas*. As **medidas quantitativas** têm sua base na freqüência do termo no documento e, de variações sobre essa medida, por exemplo, freqüência relativa na coleção ou freqüência normalizada em relação à coleção.

No **tratamento dos termos**, em um esforço inicial, busca-se desconsiderar dos textos as palavras que não representem conhecimento útil, via eliminação de ***stopwords***. Essas palavras não são relevantes para a análise dos textos, sendo, geralmente, constituídas por preposições, pronomes, artigos, interjeições, dentre outras. Para alguns domínios, também é comum existirem listas de palavras que não devem ser consideradas relevantes; essas listas são conhecidas como *stopwords* de domínio.

Posteriormente, busca-se identificar similaridades de significados entre palavras. Em casos de variações morfológicas pode-se, por exemplo, reduzir uma palavra à sua raiz por meio de processos de *stemming*, reduzir as palavras ao seu lema (lematização) ou substantivar os termos (substantivação). Em caso de sinônimos podem-se usar dicionários ou o conteúdo da cláusula "use" em um *thesaurus*, quando disponíveis. Ainda, quando disponível alguma organização hierárquica pode-se utilizar a classe a qual a palavra pertence em lugar do conjunto de palavras daquela classe; por exemplo, utilizar "Gado de Corte" em lugar dos termos mais específicos como "Gado Brahma", "Gado Gir", "Gado Guzera", "Gado Indubrasil", "Gado Nelore" ou "Gado Sindi" [2]. Nesse ponto, obtém-se o conjunto de termos que representam a coleção.

[2]Fonte: THESAGRO, Thesaurus Agrícola Nacional, disponível em: http://www.agricultura.gov.br/portal/page?_pageid=33,959135&_dad=portal&_schema=PORTAL

Estruturação dos documentos

Uma vez gerados os termos que representam a coleção, no Pré-Processamento de textos deve-se buscar a estruturação dos documentos, de maneira a torná-los processáveis por parte dos algoritmos de extração automática de padrões. Embora alguns algoritmos exijam informações mais sofisticadas, como os baseados em conhecimentos lingüísticos, a grande maioria dos algoritmos de extração de padrões necessita apenas que os documentos estejam em um formato de planilha, no qual as linhas correspondam aos documentos e as colunas aos termos presentes na coleção. Esse formato é denominado ***bag of words***, na qual os termos são considerados independentes, formando um conjunto desordenado em que a ordem de ocorrência das palavras não importa.

Uma forma de representar a *bag of words* é por meio de tabelas atributo-valor. Um exemplo dessa representação pode ser observado na Tabela 2.1, na qual d_i corresponde ao i-ésimo documento, t_j representa o j-ésimo atributo, que é o j-ésimo termo[3], a_{ij} é a medida que relaciona o i-ésimo documento com o j-ésimo atributo e y_i representa a classe do i-ésimo documento. É importante ressaltar que esse último elemento não aparece em casos nos quais a coleção de textos seja não-rotulada.

	t_1	t_2	$\dots$	t_M	Y
d_1	a_{11}	a_{12}	$\dots$	a_{1M}	y_1
d_2	a_{21}	a_{22}	$\dots$	a_{2M}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
d_N	a_{N1}	a_{N2}	$\dots$	a_{NM}	y_N

Tabela 2.1: Padrão de representação da matriz atributo×valor

Dessa forma, cada documento pode ser representado como um vetor $\vec{d_i} = (\vec{a}_i, y_i)$, no qual $\vec{a}_i = (a_{i1}, a_{i2}, \dots, a_{iM})$. Segundo Salton e Buckley (1987), o cálculo do valor de a_{ij} depende de três componentes:

- o componente do *documento*, relacionado à estatística de um determinado termo estar presente em um dado documento;
- o componente da *coleção*, relacionado à distribuição da presença do termo ao longo da coleção de documentos;
- o componente da *normalização*, o qual possibilita que documentos de diferentes tamanhos possam ser comparados segundo uma mesma escala.

Diversas medidas foram propostas para calcular os valores de a_{ij}. Essas medidas podem ser classificadas em dois tipos: **binárias** e **baseadas em freqüência**. **Valores binários** indicam a ocorrência ou não de um termo em um determinado documento, podendo ser utilizados, por exemplo, para extrair informações relativas à semelhança de documentos a partir do número de termos em comum. Já os **valores baseados em freqüência** visam contabilizar o número de ocorrências de um determinado termo em um dado documento, servindo como base para a extração de diversas medidas estatísticas na extração de padrões.

[3]Desse ponto em diante, termo e atributo serão usados quase que indistintamente.

Para os valores baseados em medidas de freqüência, podem-se distinguir duas categorias de pesos: **pesos supervisionados**, que dependem de dados com valor de classe conhecido, medindo a importância de um dado atributo na determinação do valor da classe; e **pesos não-supervisionados**, aplicáveis a dados não-rotulados.

Dentre os pesos supervisionados, pode-se citar medidas como a *ConfWeight* (Soucy e Mineau, 2005), a Entropia, a Entropia Conjunta (Haykin, 2001) e a Informação Mútua (Yao, 2003). Como exemplos de pesos não-supervisionados, pode-se citar a *tf* (*term frequency*), a qual considera a freqüência absoluta dos termos nos documentos (Rijsbergen, 1979); a *idf* (*inverse document frequency*) (Salton et al., 1975), que calcula a freqüência inversa do termo, favorecendo termos que aparecem em poucos documentos da coleção; e a *tf-idf* (Salton e Buckley, 1987), que consiste em uma combinação das duas anteriores (*tf* e *idf*).

Redução da dimensionalidade de termos

Gerada a matriz atributo-valor que representa os documentos, pode-se afirmar que essa representação da coleção tem por características a alta dimensionalidade e o fato de ser esparsa. A alta dimensionalidade é causada pelo grande número de termos contidos em uma coleção de textos. Já a natureza esparsa da matriz é ocasionada pelo fato de que grande parte dos termos ocorre somente em uma pequena parte dos documentos, de maneira que a medida de relação com todos os outros documentos da coleção será igual a zero. Ante a esse contexto, vários algoritmos de extração automática de conhecimento têm o seu desempenho negativamente afetado. Yu e Liu (2004) destacam que o excesso de atributos pode causar lentidão no processo de treinamento, bem como decremento na qualidade do conhecimento extraído (*overfitting*, por exemplo). Além disso, nem todos os termos gerados são considerados relevantes para a descrição do domínio retratado pela coleção de documentos.

Dessa forma, é de extrema importância a redução do espaço de atributos para otimização dos resultados do processo de mineração de textos e melhorar a representatividade dos dados, no sentido de determinar os termos mais representativos para a modelagem do domínio do problema, como também para atender a aspectos de compreensibilidade e escalabilidade dos modelos resultantes. A idéia é tornar o conjunto de termos com o qual se trabalha mais conciso, porém não menos representativo em relação ao conjunto original. Em mineração de textos, existem duas abordagens comumente usadas para reduzir o número de atributos: **extração de atributos** e **seleção de atributos**.

O processo de **extração de atributos** está relacionado à criação de um novo conjunto de atributos, usando para tal uma função de mapeamento entre as representações. O novo conjunto gerado é, geralmente, de menor dimensão que o conjunto original, uma vez que os novos atributos são gerados a partir de combinações dos anteriores, substituindo-os. A principal desvantagem desse método é que os atributos gerados não mantêm uma correlação explícita com a configuração real do problema, tornando os modelos gerados de difícil interpretação ou exigindo uma pré-interpretação dos grupos obtidos. Por exemplo, técnicas que trabalham sobre a decomposição espectral da estrutura de covariância, como a *Principal Component Analysis* (PCA) (Jolliffe, 2002), provocam uma rotação

do espaço de características (atributos) e a transformação dos dados, que acaba por dificultar uma interpretação dos mesmos. Outros tipos de análise fatorial, que também requerem rotações dos espaços de características, levam aos mesmos problemas (Mardia et al., 1979). Técnicas que trabalham com a decomposição do espaço de características em valores e vetores singulares, não necessitam de alterações nos vetores de atributos originais, como a *Latent Semantic Analysis* (LSA) (Landauer et al., 1998) e a *Latent Semantic Indexing* (LSI) (Deerwester et al., 1990). Essa técnicas facilitam a interpretação dos resultados, obtendo grupos em diferentes espaços de atributos. Há também técnicas de agrupamento de palavras e para buscar grupos de palavras similares (Manning e Schütze, 1999; Hofmann, 1999; Slonim e Tishby, 2000). Para todas essas técnicas, após encontrar os grupos de palavras similares, pode-se atribuir um termo ou conjunto de termos para representar-lhes, formando assim uma taxonomia de termos da coleção analisada. Por exemplo, se fosse obtido um grupo de atributos com termos como "dentes dentição-característica forma-cérebro tamanho-cérebro postura-ereta" pode-se inferir que se refira a família "hominídeos", ou seja, todos esses vocábulos seriam trocados pela classe "hominídeo" na análise dos textos; essa inferência é subjetiva, realizada por um especialista no domínio do conhecimento.

Selecionar atributos, por sua vez, remete à seleção de um subconjunto de atributos a partir do conjunto original, seguindo alguns critérios. Os atributos selecionados mantêm a interpretação física original, facilitando a compreensão do modelo gerado. Em um conjunto de dados que possui n atributos, existem $2^n - 1$ subconjuntos candidatos (excluindo-se o subconjunto vazio), devendo-se selecionar aquele que apresenta melhor desempenho segundo algum critério de avaliação. Entretanto, a busca exaustiva pelo melhor subconjunto, em contextos de dimensionalidade muito alta, pode tornar-se inviável. Dessa forma, Dash e Liu (1997) sugerem a adoção de um procedimento de busca com um critério de parada bem definido, o qual se constitui de quatro passos básicos:

1. Um procedimento de geração do próximo subconjunto a ser analisado, podendo ser segundo alguma heurística, por meio de busca completa (ou exaustiva) ou aleatória;

2. uma função de avaliação para os subconjuntos gerados;

3. um critério de parada; e

4. um procedimento de validação dos subconjuntos.

Os métodos de busca mais comuns são aqueles que efetuam uma busca gulosa. Esses podem ser divididos em dois tipos: *forward selection* e *backward elimination*. Os métodos *forward selection* iniciam a busca com um subconjunto vazio de atributos, colocando progressivamente um atributo, escolhido segundo algum critério de avaliação, a cada iteração. Já os métodos *backward elimination* iniciam o processo com todos os atributos, eliminando, a cada passo, o menos interessante deles. Essas abordagens podem gerar diferentes subconjuntos e, dependendo da aplicação e dos objetivos, uma pode ser mais indicada que a outra. Existem, ainda, métodos que não efetuam busca para selecionar atributos, os quais, geralmente, fornecem *ranking* de atributos, sem apontar combinações de atributos que formariam subconjuntos "ótimos" locais.

A escolha dos processos de seleção de atributos depende, também, da existência ou não de rótulo nos textos utilizados. Caso existam, podem-se usar métodos supervisionados de seleção de atributos, que avaliam a importância do atributo na determinação do valor de classe, como Ganho de Informação ou χ^2. Por outro lado, métodos não-supervisionados conseguem lidar com casos nos quais trabalha-se com dados não-rotulados. Dentre estes métodos, podem ser citados os métodos de Luhn (Luhn, 1958) e Salton (Salton et al., 1975).

2.2.3 Extração de padrões

Delimitado o problema e devidamente representados os textos, o processo avança, então, para a fase de extração de padrões[4]. As tarefas a serem realizadas nesse ponto são definidas de acordo com o objetivo final do processo de extração de conhecimento. Assim como na mineração de dados, podem-se resumir as principais atividades de extração de padrões em textos em **tarefas preditivas** e **tarefas descritivas**.

Tarefas preditivas consistem na generalização de exemplos ou experiências passadas com respostas conhecidas. Para as atividades **preditivas** são utilizados algoritmos de aprendizado de máquina supervisionados. Esses algoritmos, conhecidos como indutores, exigem um conjunto de exemplos de treinamento para os quais o atributo classe tenha valor conhecido (Mitchell, 1997; Monard e Baranauskas, 2003). Essa classe de algoritmos se divide em suas subclasses principais: os algoritmos de **classificação** e os algoritmos de **regressão**. Um bom exemplo de regressão é predizer o ganho ou a perda de produção em uma cultura, com base no estudo de diferentes adubações, no qual a medida de produtividade é um atributo contínuo. **Classificação** é referente ao processo em que o atributo classe tem valor categórico discreto, enquanto **regressão** remete à predição de variáveis com valores contínuos. A aplicação de tarefas preditivas em mineração de textos tem como principal aplicação a categorização automática de documentos, ou seja, são tarefas de classificação. Por exemplo, a partir da seleção de uma coleção de artigos sobre esportes e outros de uma área diferente, pode-se inferir um modelo de classificação para artigos da área de esporte, que é dito uma **tarefa de classificação ou categorização**. Deve-se observar que a categoria ou classe, nesse caso, é um atributo discreto e que os dados são mapeados em um número finito de categorias, no qual cada categoria corresponde a um **rótulo**.

A classificação de textos sempre ocorre sob categorias pré-definidas, como taxonomias, *thesaurus* ou ontologias. As técnicas utilizadas são muito variadas, mas deve-se escolher entre as que melhor se adaptam a um grande número de atributos (por exemplo, *Suport Vector Machines* (Joachims, 1998)), ou a implementações de algoritmos que lidam bem com grande número de atributos (por exemplo, o **C4.5** para árvores de decisão (Wu et al., 2008; Quinlan, 1993)). Uma aplicação interessante de categorização de documentos é a sua utilização para obter metadados estruturados, como no trabalho de Pierre (2002), em que ele utiliza técnicas de mineração de textos para auxiliar a extração de metadados, para categorizar os textos e então utilizar técnicas de mineração de dados tradicionais

[4]A partir desta etapa o conjunto de dados representa a coleção de documentos/textos.

sobre a base estruturada. Se apenas parte dos exemplos é rotulada, podem ser aplicados modelos de **aprendizado de máquina semi-supervisionados**, em que a informação dos rotulados ajuda a gerar modelos para rotular os não rotulados, ou aplica-se um método não supervisionado para gerar rótulos e então um classificador para tentar melhorar a rotulação, repetindo-se o processo enquanto algum critério de avaliação do mesmo indicar que houve melhora (Matsubara et al., 2005; Brefeld e Scheffer, 2004).

As **tarefas descritivas** consistem na identificação de comportamentos intrínsecos da coleção de dados, sendo que esses dados são **exemplos não rotulados** ou tratados como não rotulados, utiliza-se de modelos de **aprendizado de máquina não-supervisionado**, que por sua vez, produzem conhecimento novo e interpretável. As principais tarefas descritivas em mineração de textos são a obtenção de regras de associação, o agrupamento e a sumarização de documentos.

As **regras de associação** são relações lógicas inferidas entre dados correlacionados de um ou mais conjunto de dados, analisados conjuntamente (Agrawal e Srikant, 1994). Em outras palavras, as regras de associação indicam relações entre dois conjuntos disjuntos de atributos L e R, representando tais relações na forma $L \rightarrow R$, na qual L é conhecido como antecedente e R como o conseqüente da regra. Em mineração de textos, o número de regras de associação obtidas tende a ser muito grande, uma vez que cada termo de um texto pode ser considerado um atributo e, como existem diversos termos diferentes em um documento, várias relações podem ser encontradas. Portanto, as regras obtidas devem, também, ser mineradas até que delas seja possível obter algum conhecimento útil e conclusivo, cabendo essa tarefa à etapa de pós-processamento do processo de mineração de textos.

Tarefas de agrupamento ou *clusterização* buscam uma estrutura de organização em grupos de objetos similares. Especialmente, quando a quantidade de dados é extremamente grande supõe-se que eles tenham uma estrutura de *cluster*, que permite melhor entendê-los e usá-los eficientemente (Duda et al., 2001). Existem diferentes formas e técnicas para agrupamentos, que envolvem modelos probabilísticos (Mardia et al., 1979), com base na distribuição original dos dados. Dentre esses se tem grupos ***flat*** ou **hierárquicos**; isto é, grupos isolados ou com uma característica hierárquica. Nos **agrupamentos probabilísticos** cada exemplo pertence a um grupo com alguma probabilidade entre zero e um; sendo que a soma de todas as probabilidades de um mesmo exemplo para cada grupo deve ser um. Logo, em agrupamentos *flat* a probabilidade de um documento pertencer a um grupo é um ou zero. Em um agrupamento hierárquico *soft assignment* cada elemento tem uma probabilidade de pertencer ao seu grupo e ao grupo superior a ele, com diferentes distribuições de probabilidade. Uma aplicação interessante para *cluster* probabilístico é a desambigüidade entre conceitos, dependendo de cada tópico em que eles são utilizados (Manning e Schütze, 1999); desde que se possa também classificar os textos analisados nos tópicos que se dispõe, os resultados são de grande auxílio à análise como um todo. Ainda, para agrupamentos *soft assignment* tem-se os agrupamentos com base em lógica nebulosa (***fuzzy***), com raciocínio semelhante aos probabilísticos, porém em lugar de estimar as probabilidades em cada grupo, utilizam-se os graus de pertinência. A utilização de agrupamento de dados pode ser interessante para fins de mineração, exploração e sumarização

de coleções de textos (Peltonen et al., 2002).

Ainda, *clusterização* é muito usada em recuperação de informação, partindo-se da hipótese que se um usuário está interessado em um documento específico pertencente a um grupo deve também estar interessado em outros documentos desse grupo (Chakrabarti, 2003; Kobayashi e Aono, 2004). Essa hipótese em concomitância com a derivação de taxonomias de tópicos e suas projeções, levam a buscas mais direcionadas, que procuram por grupos de documentos semelhantes. *Clusterização* tem sido muito utilizada no suporte à geração de taxonomias e aprendizado de ontologias (Maedche, 2002), ou muitas vezes para auxiliar o processo de geração/aprendizado, como no trabalho de Hotho et al. (2002) e, em poucos trabalhos, de forma mais direcionada à obtenção de taxonomias de conceitos, como utilizado por Wives (2004) e Kashyap et al. (2005a).

A **sumarização**, em mineração de textos, é uma área emergente baseada em agrupamentos por tópicos e então extração de informação por tópicos identificados, para um documento apenas que seja representativo do grupo (Weiss et al., 2005). A extração do resumo em geral é feita sentença por sentença, utilizando-se alguma ferramenta de etiquetação para identificá-las e marcá-las no texto. Também se atribui uma ponderação às sentenças, em função da freqüência de um conjunto de interesse de palavras na mesma, bem como se pode distribuir pesos acumulativos a esses de acordo com a seção do texto no qual a sentença ocorre, por exemplo, se o texto for semi-estruturado, diferentes pesos para título, resumo, introdução. Ao final do processo de etiquetação e ponderação chega-se a um conjunto de sentenças no qual é mantida a ordem em que elas ocorreram no documento; e então cada conjunto de sentenças representa o seu tópico. Esse tipo de extração de resumos é puramente estatístico; existem outros métodos com base na estrutura retórica dos textos que podem ser integrados a outras técnicas lingüísticas de identificação de segmentos semânticos (Rino e Pardo, 2003).

2.2.4 Pós-processamento e utilização do conhecimento

Uma vez obtidos os padrões acerca dos documentos, esses devem ser analisados e interpretados. Deve-se avaliar, junto a um especialista, aspectos do conhecimento extraído de acordo com o contexto do problema, como a representatividade do conhecimento obtido, a novidade contida nos resultados e a maneira que o conhecimento deverá ser utilizado.

No que tange à validade do conhecimento, deve-se verificar se os padrões obtidos condizem com a realidade e são aplicáveis ao que se propôs no objetivo inicial. Por exemplo, em aplicações que envolvam tarefas preditivas, pode-se avaliar os modelos com medidas relativas à precisão na predição de dados não vistos, como taxa de erro, precisão e cobertura. Avaliações repetíveis e com medidas comparáveis são ditas **avaliações objetivas**. A avaliação de modelos descritivos é um pouco mais complicada, uma vez que o critério de avaliação varia de acordo com o objetivo do processo. Dessa forma, algumas vezes podem não existir medidas objetivas para avaliar o resultado do processo, ficando a avaliação a cargo de um especialista do domínio, denominada **avaliação subjetiva**.

Outro aspecto importante refere-se à compreensibilidade do conhecimento. Os algoritmos de extração de padrões podem gerar uma quantidade de padrões muito grande,

dificultando a compreensão por parte do usuário e fazendo necessária a aplicação de mecanismos que forneçam aos usuários apenas os padrões mais interessantes (Silberschatz e Tuzhilin, 1995). Em mineração de textos, devido à presença de alta dimensionalidade, esse é um problema recorrente, demandando especial atenção.

Uma tarefa que tem um importante papel no auxílio ao especialista nesse ponto do processo, mas que também pode ser utilizada ao longo das demais etapas, é a **visualização** dos dados. Card et al. (1999) definem essa tarefa como "uma representação visual de dados interativa baseada no computador para ampliar a cognição". Diversos tipos de gráficos e diagramas podem ser utilizados ao longo de todo o processo pelo analista, facilitando a compreensão dos resultados das diversas etapas e auxiliando a tomada de decisão. A análise visual pode, por exemplo, indicar falhas no pré-processamento dos documentos e a necessidade de refazê-lo. Além disso, diversas formas finais de conhecimento são facilmente associáveis a formas de visualização de informação, como é o caso do *cluster* hierárquico de documentos, cujo resultado final é uma árvore que pode ser visualizada, por exemplo, por meio de árvores de diretórios ou árvores hiperbólicas (Lamping et al., 1995).

Após ter percorrido todas as etapas do processo, o conhecimento validado pode ser utilizado pelo usuário. Com o correto desenvolvimento das etapas anteriores, pode-se garantir que o conhecimento, nesse ponto, é válido e útil, podendo ser aplicado no apoio à tomada de decisão para os aspectos estabelecidos nos objetivos do processo.

2.3 Considerações finais

A extração de conhecimento a partir de bases de dados textuais tornou-se um importante aliado às mais diversas corporações, obtendo a partir dos documentos textuais algum conhecimento útil que auxilia o processo de tomada de decisão. A mineração de textos é o agente transformador dos dados textuais em conhecimento, extraindo dos documentos informações novas e potencialmente úteis por meio de padrões detectados ao longo das coleções.

O processo de mineração de textos é constituído de uma seqüência de cinco passos: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento. Cada um desses é constituído de atividades genéricas, as quais podem ser instanciadas de acordo com a especificidade de cada aplicação por meio de, por exemplo, seleção do algoritmo de aprendizado que melhor cumpra o objetivo final, ou de alguma medida que selecione o subconjunto de atributos mais representativo daquela coleção.

Neste capítulo procurou-se cobrir o processo de mineração de textos de um modo geral, para que no próximo capítulo seja mostrada uma instanciação dele como uma metodologia de obtenção de taxonomias de tópicos. A metodologia de obtenção de taxonomias de tópicos é uma das principais contribuições desta tese de doutorado e, dela há mais três contribuições ao estado da arte, representadas pela seleção de atributos n-gramas do Capítulo 4, novos métodos para rotulação de agrupamentos hierárquicos e avaliação objetiva da rotulação no Capítulo 5.

Capítulo

3

Abordagem para a construção de taxonomias de tópicos em um domínio

Neste capítulo é descrita uma proposta metodológica para a construção de taxonomias de tópicos em um domínio de conhecimento. A proposta é uma instanciação do processo de mineração de textos, cobrindo desde a seleção da coleção de documentos até a obtenção automática de uma boa aproximação da taxonomia desejada. Tecnologicamente, a proposta tem como objetivo chegar a um ambiente de software para automatizar o processo, permitindo-se a intervenção dos especialistas do domínio em algumas das etapas, de modo que necessidades de compreensibilidade possam ser satisfeitas. Academicamente, o objetivo é explorar diferentes técnicas em cada uma das etapas da metodologia, de modo que se possam propor novas técnicas, adaptações ou combinações das existentes, que sejam mais adequados aos seus propósitos.

3.1 Introdução

A metodologia desenvolvida e apresentada neste capítulo é denominada TopTax-*Topic Taxonomy*. A TopTax trata-se de uma compilação das iniciativas já amplamente discutidas na literatura, anteriores a sua primeira proposição (Moura, 2006) e podendo incorporar novas soluções. O objetivo é cobrir de modo integrado as etapas de mineração de textos desde a seleção da coleção de textos até a obtenção automática de uma boa aproximação da taxonomia desejada, permitindo a intervenção de especialistas do domínio de conhecimento no processo.

Há muitos trabalhos na literatura sobre hierarquia de tópicos visando à melhoria de processos de busca na *web*. Essa tem sido uma prática utilizada em aplicações voltadas para a recuperação de informações textuais, como a geração de indexadores para as máquinas de busca. Ou ainda, podem-se organizar os resultados da busca em tópicos aproximados, como no caso da organização hierárquica obtida pela ferramenta Vivisimo

(Koshman et al., 2006). Porém, na maioria dos casos, os tópicos são encontrados sob agrupamentos disjuntos não hierárquicos. A organização hierárquica de tópicos, em geral, é realizada manualmente, por meio de um intenso trabalho humano, como para o *site* Yahoo; semi-automática, como na construção de taxonomias ou ontologias (Maedche, 2002; Bloehdorn et al., 2005; Cimiano, 2006), ou na construção de mapas de tópicos (Librelotto et al., 2004).

Um modo de organizar o conhecimento é criar uma estrutura de representação dividida em tópicos hierarquicamente relacionados. Um exemplo, é a organização hierárquica de um domínio de conhecimento como na Agência de Informação Embrapa (Souza et al., 2006) - ou simplesmente Agência. O conhecimento mais genérico está presente nos níveis mais altos da hierarquia enquanto que seus detalhamentos, ou conhecimentos mais específicos, nos níveis mais baixos. Cada nó na hierarquia corresponde a um tópico do domínio de conhecimento e contem um texto relacionado ao mesmo. Esses textos resultam da compilação do conhecimento produzido pelos pesquisadores e algumas referências bibliográficas, dando-se preferência a referências em mídia digital. A criação e atualização dessas hierarquias não é uma tarefa trivial, porque demanda um grande esforço por parte da equipe de especialistas.

Considerando que uma coleção de textos de um domínio de conhecimento organizada em tópicos hierárquicos facilita a compreensão desses documentos, neste capítulo descreve-se a proposta de uma metodologia para tal fim. Para atingir esse objetivo deve-se auxiliar o processo de organização da coleção de textos, produzindo um mapa de tópicos que facilite a análise e compreensão da coleção de textos e que possa ser utilizado junto a ferramentas de recuperação de informação sobre a coleção, fornecendo bons indexadores de busca e permitindo a evolução dos tópicos. A TopTax é composta por etapas completamente automatizadas, porém permite a intervenção dos especialistas do domínio em cada qual, de modo que algumas necessidades de compreensibilidade possam ser satisfeitas, se os especialistas assim o desejarem. Caso contrário, o ferramental pode ser utilizado sem intervenções. Não se pretende, com as soluções automatizáveis, retirar o papel do especialista do domínio na construção dessas taxonomias, mas sim, deixá-lo livre para escolher em que etapas intervir e assim facilitar a sua tarefa. Espera-se chegar a um nível de organização dos dados similar aos mapas de tópicos da Agência.

Desta forma, a TopTax procura integrar essas técnicas em um ambiente de software único, onde se possa trabalhar o desenvolvimento de uma solução tecnológica e no qual se possam incorporar novas soluções. Ainda, academicamente, procura-se explorar diferentes técnicas em cada uma das etapas da metodologia, de modo que se possam propor novas soluções, adaptações ou combinações das existentes, sempre com o objetivo de adequá-las à construção de taxonomias de tópicos.

Na Seção 3.2 são discutidos alguns trabalhos mais diretamente relacionados a este, tanto os que lhe serviram de inspiração, como alguns mais recentes que ajudam a corroborar esta iniciativa. A seguir, Seção 3.3, a metodologia TopTax é apresentada em consonância com sua configuração atual. Na Seção 3.4 são apresentados, resumidamente, alguns exemplos de aplicação da TopTax, com a utilização de algumas ferramentas desenvolvidas. Na Seção 3.5 são resumidas as contribuições da TopTax, suas limitações,

alguns trabalhos futuros e em andamento.

3.2 Trabalhos relacionados à obtenção de taxonomias de tópicos por meio de mineração de textos

Em muitos trabalhos o agrupamento de documentos seguido de alguma descrição do mesmo é tratado como uma sumarização da coleção de textos. Nessa linha há trabalhos que tratam a sumarização em si, encontrando os tópicos de cada grupo de textos, e, então aplicando alguma técnica de resumo aos grupos de documentos. Ainda, há outros que tratam os descritores de cada grupo como descritores de tabelas de bases de dados ou indexadores de busca, diretamente como uma taxonomia de tópicos ou os utilizam para suporte à construção ou evolução de ontologias. A seguir, algumas dessas linhas são colocadas.

Um grupo da IBM iniciou o desenvolvimento do *framework* TaxGen (Müller et al., 1999) no final dos anos 90. Nesse projeto a taxonomia obtida é vista como a representação temática inerente à coleção de textos. O objetivo era criar automaticamente taxonomias comparando os documentos em função de suas características lingüísticas. Para tal usaram algoritmos de agrupamento hierárquico aglomerativos, porém cortavam a taxonomia obtida em, no máximo, sete níveis; de modo que os níveis pudessem ser observados em uma única tela. O agrupamento obtido era antão utilizado para classificar novos exemplos de documentos. Para agilizar os algoritmos de agrupamento, descartavam os termos mais freqüentes e os termos menos freqüentes. O corte em sete níveis da taxonomia é um pouco mais abrupto que em outros trabalhos. Por exemplo, em Aggarwal et al. (1999) os centróides de cada classe são usados como sementes iniciais e então um método de agrupamento hierárquico é aplicado a cada grupo. Durante o processo, os grupos com menor número de documentos são descartados.

No trabalho de Larocca et al. (2000) foi usado o software AutoClass para obter um bom agrupamento de um conjunto de documentos e então cada tópico foi sumarizado. Foi realizado um pré-processamento convencional dos textos: *case-folding*, para não diferenciar maiúsculas de minúsculas; *stemming*; eliminação de *stopwords*; e *n-gram words*, utilizando apenas trigramas. A medida utilizada para eliminar os casos com peso de ocorrência zero foi a *tf-idf*. Obtido o agrupamento, as palavras mais descritivas de cada grupo são consideradas palavras-chaves. Para o documento centróide de cada grupo elaborou-se um resumo dos documentos mais relacionados a ele. No trabalho é proposto que se considerem as palavras-chaves como um ultra-resumo, que ajuda a identificar os tópicos de cada grupo. Então, quando o usuário julga algum grupo mais interessante pode resumir os seus documentos para ter uma melhor idéia do que foi encontrado. Como a abordagem usada é *bag of words*, não se conhece a ordem das palavras, logo para resumir um grupo de documentos foi utilizado um método extrativo para obter as sentenças mais significativas dos textos, usando um peso semelhante a *tf-idf* e estabelecendo um *threshold* para cortes. O agrupamento obtido não é hierárquico e a avaliação do método é muito subjetiva, uma vez que o objetivo é fornecer uma ferramenta de análise exploratória para o especialista

mais voltada à obtenção de resumos e palavras-chaves dos grupos.

Ainda na linha de sumarização, Chuang e Chien (2005) definem um segmento de texto como uma seqüência significativa de palavras que são freqüentemente curtas mas representam um conceito específico em um domínio. Tal seqüência como palavra-chave em um documento corresponde a uma expressão de busca em língua natural. Esses segmentos podem ser palavras, frases, nomes de entidades, expressões de busca em língua natural, novos eventos, nomes de produtos, títulos de artigos ou livros, etc. O objetivo é organizar esses segmentos em hierarquias de tópicos com um método geral e prático que utiliza a web como recurso adicional de conhecimento. Como os resultados não são facilmente interpretáveis, resultando em extensas hierarquias; propuseram um critério de corte que transforma as árvores binárias obtidas em árvores genéricas, conhecido como *min-max*. Para sumarização automática multi-documentos há várias técnicas. Ou et al. (2006) falam do uso de palavras similares entre documentos para indicar tópicos em coleções de artigos para organizar uma biblioteca digital. Porém, concluíram que essa solução não era a ideal, comparando-a com outros tipos de resumos automáticos, extração de frases completas e outras formas de sumarização; que fogem ao foco desta tese. Já a sumarização de agrupamentos por palavras-chaves formadas por frases, como no trabalho de Hammouda e Kamel (2008) divide-se em duas etapas: seleção das chaves, cuja qualidade é dada pela *df* ponderada pelo número de frases na coleção total; e, ranqueamento das frases, as que ficam no topo descrevem o grupo.

Nos trabalhos de descoberta e comparação de hierarquia de tópicos de Lawrie e Croft (2000) e Lawrie et al. (2001) o interesse é gerar taxonomias de domínios específicos, agrupar um conjunto grande de documentos e então obter hierarquias de cada grupo, isto é, hierarquias sob cada tópico. A idéia é que grupos homogêneos de documentos propiciam melhores hierarquias, pois o conjunto de atributos/termos tem um uso mais coeso e melhor delimitado. Nesses trabalhos foram obtidas hierarquias pré-assumidas com características de dependência semântica e hierarquias léxicas; e, ainda foram propostas duas diferentes formas de avaliar as diferenças entre hierarquias, com duas diferentes métricas. A primeira mede a velocidade com que se chega ao conjunto de documentos relevantes, correspondendo a quantas arestas anda-se no grafo, e outra quantifica a similaridade entre duas hierarquias. A questão da similaridade entre duas hierarquias é razoavelmente complicada de ser calculada e a proposta colocada é bastante interessante. O resultado final é um *k-means* modificado e agrupamento hierárquico em um domínio restrito, bastante parecido com o problema de pesquisa deste trabalho.

Procurando-se por atributos mais significativos nos grupos obtidos, tem-se métodos para agrupar e atribuir pesos simultaneamente como proposto por Frigui e Nasraoui (2004). Cada grupo é caracterizado por um conjunto de palavras-chave, então se encontram os pesos das palavras-chave não só para particionar os documentos em grupos semânticos mais significativos, mas também para indicar a relevância dos termos nos grupos. Nesse trabalho também é mostrado como levar essas relevâncias para uma proposta utilizando lógica nebulosa (lógica *fuzzy*). A medida utilizada para o termo no documento, casela da matriz atributo-valor, precisa ser contínua para ser mais rica em informação, dado que o algoritmo proposto é utilizado para resolver, simultaneamente, os agrupamen-

tos ótimos e os pesos dos atributos. O problema é que é necessário estabelecer o número de grupos *a priori* e o modelo não é hierárquico. Assim, embora essa fosse uma alternativa aos métodos que apenas encontram centróides, as medidas de relevância dos atributos são atualizadas junto ao processod e agrupamento e dependem completamente do método; logo, para utilizar as mesmas idéias é necessário adaptá-las a outros métodos de interesse.

Para a melhoria de taxonomias de documentos, alguns autores trabalham com uma classificação pré-existente, como no caso de Punera et al. (2005). Nesse trabalho, a partir de um corpus de documentos rotulados extrai-se uma estrutura hierárquica viável; construída a partir do conhecimento prévio das classes. Para as classes obtidas é ajustado um classificador SVM a fim de refinar a classificação dos documentos; ou seja, o objetivo é melhorar a taxonomia pré-existente.

Em trabalhos em que o objetivo é extrair a taxonomia a partir dos documentos, como em Dupret e Piwowarski (2005), se o domínio não é restrito, a representação *bag of words* não facilita o enfrentamento de problemas de polissemias e sinônimos. Nesse trabalho, por exemplo, procura-se explorar a indução de taxonomias de termos, a partir de uma decomposição espectral do tipo *singular value decomposition* (SVD) e identificar conceitos. Os conceitos devem ser ortogonais, não correlacionados, conseqüentemente menos sensíveis ao vocabulário particular utilizado nos documentos. Esses autores definem conceito a partir dos *ranks* dos termos na decomposição obtida; ou seja, a base matemática é uma decomposição espectral da matriz de similaridades obtida. O modelo obtêm a taxonomia, mas não carrega a informação de quais documentos se encontram sob cada tópico; é obtido um grafo direto e acíclico de termos e conceitos. Um bom contra-exemplo seria a decomposição *LSI - latent semantic indexing* (Deerwester et al., 1990), também baseada em SVD; que permite a identificação de conceitos e documentos relacionados.

Ainda na linha de obter taxonomias a partir dos documentos, Clifton et al. (2004) propuseram identificação dos tópicos freqüentemente discutidos com o objetivo de auxiliar a compreensão dos textos. Assim, precisaram adaptar as técnicas utilizadas para que os tópicos ficassem compreensíveis por humanos. Como eles também desejavam permitir a exploração da coleção, procuraram manter as ligações entre tópicos e textos. Então, utilizando regras de associação, já previamente filtradas como interessantes, agrupavam-nas, procurando regiões de sobreposição, para explorar.

O trabalho de Kashyap et al. (2005a) tem por objetivo reduzir o esforço humano para construir ontologias. O ponto de partida é um agrupamento hierárquico de documentos, obtendo uma taxonomia de conceitos. No trabalho, é definida uma taxonomia como um sistema de organização do conhecimento que representa relações entre tópicos de tal forma que eles arranjem os conceitos dos mais genéricos para os mais específicos - conceitos mais formais foram deixados para trabalhos futuros. Eles também utilizaram técnicas de processamento de língua natural na geração da taxonomia, porém restritas ao pré-processamento, para retirar sintagmas verbais da análise. A técnica de agrupamento foi o *bissecting k-means* e o PDDP - *Principal Direction Divisive Partitioning* (Boley, 1998); extraindo-se a taxonomia a partir da hierarquia e avaliando a qualidade da mesma. Após o agrupamento, são calculados parâmetros que refletem a coesão dos grupos e que são utilizados pelo algoritmo de extração da taxonomia. Esses parâmetros resultam em

pontos de poda na hierarquia inicialmente obtida; ou seja, apenas os nós mais relevantes são mantidos, de modo que se tenha uma árvore mais enxuta e mais coesa. As técnicas de avaliação propostas por eles exigem a comparação com uma taxonomia padrão (*gold taxonomy*) ou a avaliação estrutural. No trabalho do TaxaMiner (Kashyap et al., 2005a) as hierarquias são obtidas de uma coleção de textos controlada, recuperada a partir de um critério de busca que garante a relação de pertinência em uma hierarquia pré-conhecida. Os autores mostram que, de acordo com a presença de termos mais amplos (*broader*) nos nós antecessores e mais específicos (*narrower*) nos nós descendentes de uma taxonomia pré-definida, conseguem avaliar o quanto o método se aproxima da taxonomia padrão. Em um segundo trabalho (Kashyap et al., 2005b), os autores usaram a técnica *Latent Semantic Indexing* (LSI) para reduzir a dimensão do problema e auxiliar a identificação da desambigüidade de termos; e, foram propostas algumas adaptações aos critérios de avaliação da taxonomia, com base na nova forma de indexar (encontrar as medidas da matriz atributo-valor) os termos nos documentos - usando os *scores* da decomposição da matriz de termos na LSI.

Trabalhos que partem da existência de uma taxonomia e procuram melhorá-la contêm idéias interessantes para o crescimento dos agrupamentos. No trabalho de Tang et al. (2006), dada uma hierarquia/taxonomia e um conjunto de treinamento, ajustam-na de acordo com o aprendizado sobre o conjunto de treinamento, considerando-o como novos exemplos a serem classificados e mostram que vão obtendo melhores hierarquias em relação à original. A abordagem de Zhao e Karypis (2005) parte de uma hierarquia de tópicos pré-definidos, por um especialista do domínio; dado que essa hierarquia é incompleta, procuram montar os grupos a partir dela e para completá-la. Também o trabalho de Escudeiro e Jorge (2005) é proposta uma metodologia para assistir editores no processo de compilação de recursos web. Nela permite-se a especificação genérica de um tópico, incluindo sua estrutura ontológica, por uma equipe de editores. A interação do usuário com o ferramental ocorre em dois passos distintos: na fases de definição de tópicos e características (taxonomia inicial) e na fase de exploração dos recursos compilados pelo sistema (de acordo com as preferências dos usuários). Logo, a base é aprendizado semi-supervisionado com algum tipo de *reinforcement learning* não tão explícito, que faz uso das preferências dos usuários.

Na mesma linha de partir de uma taxonomias pré-existentes encontra-se o trabalho de (Tang et al., 2008); porém, com aplicação na identificação dinâmica de perfis de grupos. Os autores propõem adaptações de taxonomias de tópicos para que sejam mais dinâmicas ao apresentar características de grupos em comunidades virtuais. É um caso em que as taxonomia de tópicos de interesse não casam com taxonomias habituais; por exemplo, um furacão está associado à meteorologia, mas também está relacionado a questões políticas. Nesse trabalho, dada uma taxonomia procura-se encontrar um refino realizando uma busca em um espaço de diferentes hierarquias plausíveis. Com o auxílio de uma *edit distance* vão procurando adaptá-las a novos perfis; e, ainda, comparam uma busca gulosa a uma abordagem *top-down* buscando por melhores vizinhança.

Algumas abordagens para a obtenção de taxonomias são muito dependentes dos conceitos de ontologias e processamento de língua natural, especialmente em trabalhos como

os de Hotho et al. (2002) e Jiang e Tan (2005), em que são utilizados analisadores léxicos e sintáticos bastante complexos. Nesses trabalhos também ocorrem consultas a ontologias pré-definidas, para encontrar as similaridades termo a termo e expandir ou retrair as informações encontradas. Nessa linha, encontra-se a ferramenta Text-to-Onto, que está bem longe do escopo original deste trabalho.

Neste trabalho a idéia é usar uma abordagem estatística para o problema, mais próxima a do ambiente TaxaMiner; cobrindo desde a preparação dos dados até a visualização, validação e edição da taxonomia final. Então, a partir da taxonomia obtida poder-se-á adaptar métodos de ajustes de taxonomia. Em todos os passos permite-se que o especialista interfira no processo, pois o primeiro objetivo é facilitar o seu trabalho. Como nenhuma das abordagens vistas resolve o problema proposto por inteiro, a proposta aqui apresentada utiliza várias das idéias da revisão realizada, mediante adaptações, e procurando integrá-las em cada passo. Isso também vai de encontro ao segundo objetivo, que é explorar e propor técnicas adequadas a cada passo da obtenção das taxonomias.

Finalmente, dois trabalhos atuais que corroboram totalmente com a proposta aqui apresentada são o ambiente *ACORN - Automatic Domain Specific Ontology Construction Process* de Bae et al. (2008) e o dos criadores da ferramenta OntoGen, Fortuna et al. (2008). Os dois não trabalham com abordagens híbridas, misturando aprendizado estatístico e técnicas lingüísticas. No ambiente ACORN, extrai-se conceitos do domínio com base em similaridade entre termos de um vocabulário controlado, como a WordNet, bem como co-ocorrência de termos. Depois se identificam relações entre conceitos, explorando agrupamento hierárquico e ordenação dos conceitos em relação à ontologia final. Na preparação de dados removem as *stopwords*, utilizam radicalização inflexional (com o algoritmo de Porter) e utilizam *tf-idf* tanto para os cortes como medida para a matriz atributo-valor. Depois, obtêm um *lattice* das palavras usando uma similaridade de co-ocorrência combinada com a medida de similaridade entre palavras da WordNet, após uma normalização arbitrária (multiplicam por 0.5 a soma). Esse *lattice* conceitual é combinado com os atributos mais discriminativos dos grupos obtidos a partir do *corpus* para melhorar a leitura dos tópicos e os conceitos aprendidos. Esse atributos mais discriminativos em cada grupo são os que possuem melhores ranqueamentos de *tf-idf* no grupo. No trabalho de Fortuna et al. (2008), uma taxonomia de tópicos é tratada como uma ontologia de tópicos, considerando-se cada nó como tendo uma relação de conceito e sub-conceito. Como na TopTax, a idéia desenvolvida no trabalho de Fortuna et al. (2008) procura reduzir o esforço despendido pelo especialista de domínio. A proposta deles consiste na extração de termos do domínio, a partir de abordagens híbridas, extração das relações de hierarquia, população do vocabulário e escolha dos nomes de conceitos comparando-os a um vocabulário controlado.

3.3 Descrição da metodologia proposta - TopTax

Segundo Aristóteles, taxonomia é a ciência de classificar coisas e engloba a noção de categorias e subcategorias, bem como conceitos e sub-conceitos; que são idéias que melhor se aproximam daquelas de sistemas orientados a objetos (veja (Cimiano, 2006)

para detalhes). O principal objetivo, neste trabalho, é chegar a uma taxonomia de tópicos, organizando uma coleção de documentos de tal forma que sua compreensão possa ser melhorada. A coleção de textos será agrupada em categorias e subcategorias, de modo que se obtenha descritores de cada categoria. Para os descritores será utilizada a noção de herança entre os objetos, isto é, uma subclasse ou subcategoria herdará os descritores da sua classe ou categoria.

A coleção de textos será considerada como pertencente a um domínio específico de conhecimento e será hierarquicamente organizada em tópicos e subtópicos, nos quais os documentos se inserem. Embora a idéia seja obter um ambiente de ferramentas bastante automatizado, a intervenção dos especialistas do domínio em algumas de suas etapas é permitida, sendo auxiliada por critérios estatísticos para guiar as escolhas. Para descrever o trabalho desenvolvido, a seguinte terminologia e restrições são aplicadas à coleção de documentos e seus possíveis componentes:

- **coleção de textos**: assume-se que a coleção é de um domínio específico de conhecimento. Especialmente os métodos de extração de conhecimento tratam a coleção de textos como *bag of words*, ou seja, o contexto das palavras nos textos é desconhecido. Desse modo, o ambiente não é capaz de tratar possíveis polissemias. Polissemias são termos com grafia semelhante e significados muito independentes em diferentes contextos, por exemplo, a fruta *manga* e a parte de uma camisa também com o nome *manga*. Logo, restringir o domínio diminui a probabilidade de ocorrência de polissemias na coleção.

- **extração de conhecimento**: sempre serão considerados métodos não supervisionados. Embora o domínio da coleção de textos seja supostamente conhecido, assume-se que não se conhecem os subdomínios de cada documento, ou mesmo de algum subconjunto dos documentos.

- **termo**: é usado no mesmo contexto que para recuperação de informação, pois se considera uma palavra ou composição de palavras, removidas as inflexões ou não. Tratam-se os termos com uma única palavra como *unigramas*, com duas palavras como *bigramas* e com três como *trigramas*; e as coleções sempre como *bag of words*.

- **vocabulário**: o conjunto completo de termos que é usado na construção da taxonomia de tópicos. Logo, o vocabulário corresponde à união de todos os conjuntos de descritores de cada tópico e subtópicos. Esse vocabulário pode ser considerado como uma estimativa do vocabulário do domínio; estimativa que será melhor ou pior dependendo da representatividade da coleção de textos em relação ao domínio do conhecimento.

- **taxonomia de tópicos**: uma coleção de textos, de um dado domínio de conhecimento, hierarquicamente agrupada a partir de uma função de similaridade entre seus documentos, considerado o conjunto de atributos, com cada grupo rotulado de acordo com uma estimativa de vocabulário do domínio.

A base metodológica é uma instanciação do processo de mineração de textos, cujas etapas são ilustradas na Fig. 3.1. Em um primeiro passo, faz-se uma **Identificação do**

Problema, na qual delimita-se o problema a ser trabalhado, bem como as fontes de dados e as ferramentas utilizadas. Logo após, realiza-se um **Pré-processamento** da coleção de documentos, padronizando-a e delimitando o vocabulário que a representa. Passa-se, então, para a etapa de **Extração de Padrões**, por meio da utilização de métodos de agrupamento hierárquico e rotulação das hierarquias. Com a taxonomia extraída, efetua-se o **Pós-processamento** da mesma, procurando validá-la e refiná-la. Por fim, avança-se à exploração e **Uso do Conhecimento**. As etapas são independentes e podem se retroalimentar. Além disso, o objetivo é fornecer uma solução tecnológica a partir da seleção e proposição de técnicas a serem utilizadas em cada etapa. Todas estas etapas estão detalhadas nas seções a seguir.

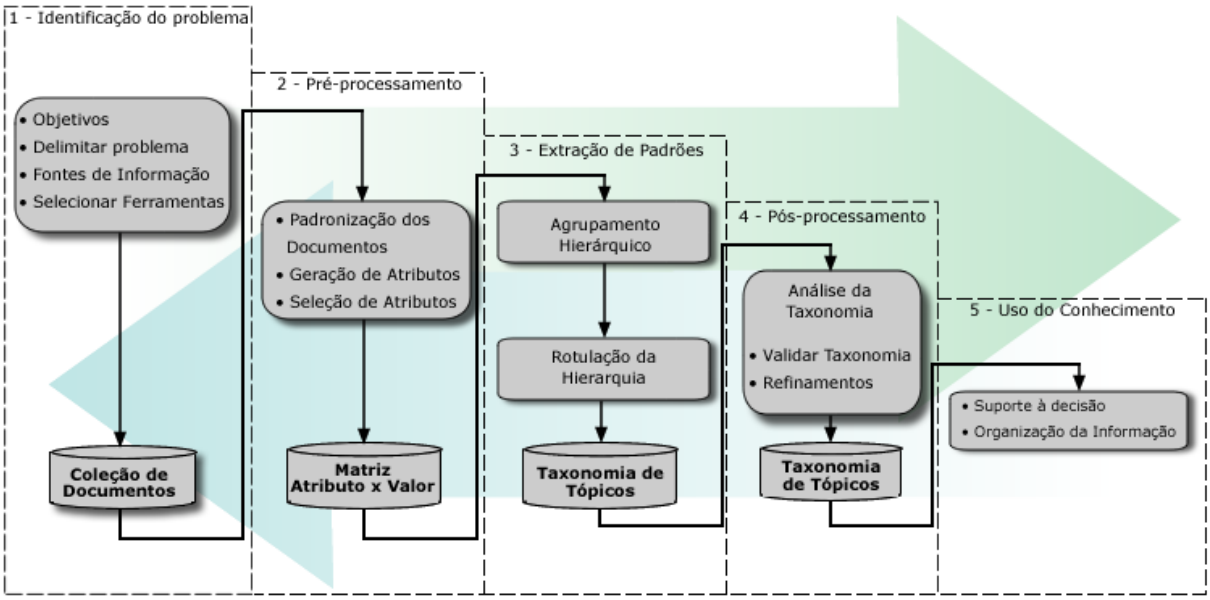

Figura 3.1: Metodologia para obtenção de taxonomia de tópicos baseada em mineração de textos

3.3.1 Identificação do problema

Nessa etapa definem-se os objetivos da aplicação da metodologia, por exemplo, identificar tendências ou simplesmente auxiliar a organização de uma coleção de documentos. A identificação do problema é uma etapa muito importante, dado que não existe descoberta de conhecimento sem demanda pelo mesmo. Nesta etapa o especialista do domínio identifica e delimita o problema, o domínio do problema, a coleção de textos a ser analisada ou sua fonte de busca, se existe algum pré-conhecimento acerca do domínio que possa ser utilizado na análise, o que se espera obter e como os resultados poderão ser utilizados. É uma etapa que demanda muito esforço tanto do especialista do domínio quanto do especialista em mineração de dados, pois esta etapa fornece subsídios a todo o processo, permitindo identificar requisitos e possíveis ferramentas para cada passo.

Identificado o problema, a coleção de textos, que efetivamente corresponde aos dados a serem analisados, deve ser delimitada com o auxílio do especialista do domínio do problema. Neste ponto é conveniente que sejam observadas as seguintes restrições e requisitos:

- a coleção de textos a ser tratada deve estar em meio digital que possa ser convertido para arquivos ASCII comuns; por exemplo: documentos gerados por algum editor de textos ou apresentações e resultados de pesquisa em uma máquina de busca para *web*. Se os documentos a serem analisados não se encontrarem em meio digital, por exemplo, se forem relatórios ou publicações antigas, deve-se também avaliar a relação custo×benefício do processo de digitalização;

- deve-se conhecer a língua ou línguas em que os textos estejam escritos, de modo que se possa pelo menos classificá-los por esse quesito. Boa parte das ferramentas automáticas de mineração de textos são dependentes da língua, dado que no mínimo realizam uma análise léxica dos textos (Martins, 2003). Assim, se os textos em análise estiverem em mais de uma língua é necessário avaliar se essa informação pode ser desprezada ou é objeto da análise;

- dados textuais provenientes de alguma base de dados pré-organizada ou publicações catalogadas em alguma biblioteca possuem metadados. **Metadados** são dados que descrevem os dados originais em um padrão pré-estabelecido e complementam as informações sobre os mesmos (Duval et al., 2002; Souza et al., 2005). Nesses casos, os metadados devem ser considerados, pois existe uma pré-classificação dos dados, que pode ser explícita ou implícita. Mesmo que a classificação seja explícita ela não será utilizada como rótulo ou seja, continuar-se-á trabalhando apenas com aprendizado de máquina não supervisionado, como pressuposto na metodologia.

- outra questão a ser considerada é o número de figuras nos documentos da coleção e a sua importância. Quando os documentos são convertidos para texto simples sem formatação, as figuras são perdidas, restando muitas vezes, apenas as legendas. Logo, se essas figuras pertencerem a alguma base de dados e houver metadados a respeito delas, há que se considerar trocá-las nos textos convertidos por seus metadados; se essa operação for factível.

Determinar o que será feito em termos de mineração dos dados é um processo que se estende desde a delimitação do problema (Ebecken et al., 2003); depende, além da identificação, das necessidades, dos objetivos estabelecidos e do quanto se dispõe de dados e informações sobre eles, bem como do quanto se dispõe de ferramentas adequadas à solução do problema.

Assim, nesta etapa, delimita-se o escopo de tratamento de dados e a fonte de informação, bem como as ferramentas que devem ser utilizadas no processo e, conseqüentemente, a forma de representação dos dados e resultados.

3.3.2 Pré-processamento

No pré-processamento faz-se uma padronização dos documentos, obtém-se a relação de atributos de interesse e reduz-se a dimensão dos atributos obtidos.

Durante a **padronização** analisa-se subjetivamente o número de documentos disponíveis que possam ser utilizados no processo. Nem sempre os documentos recuperados para

serem utilizados estão em formatos que permitem seu uso direto e, muitas vezes, após as conversões, um grande volume de dados é descartado devido a problemas como trocas de caracteres ou perda de partes de palavras. Esta sub-etapa, muitas vezes, é repetida após novas buscas por documentos e ou outras fontes dos mesmos até que se obtenha uma coleção de textos confiável. À coleção final aplicam-se algumas padronizações, como transformar todas as letras em minúsculas, eliminar caracteres especiais, eliminar *stop-words* e inserir *tags* XML. Se necessário, as *tags* XML podem ser utilizadas para delimitar os textos, bem como delimitar metadados.

Esta etapa da metodologia contribuiu na identificação de requisitos para o desenvolvimento de uma ferramenta, denominada **IESystem** - *Information Extraction System* [1], para: converter formatos diversos dos textos em texto plano, separar partes do texto (título, autores, resumos, etc), gravar os textos com *tags* XML adequadas a cada parte. Essa ferramenta também procura identificar os formatos de artigos científicos pré-configurados para ela e, também, pode receber como entrada um formato específico de documento a reconhecer na coleção de textos. Finalmente, essa ferramenta não é dependente da TopTax, pode ser utilizada em vários outros contextos de aplicação de mineração de textos, nos quais seus resultados sejam úteis.

Uma das metas deste passo é identificar e selecionar o método que melhor se adapte às necessidades desta metodologia para a **geração de atributos**[2], que são os termos de interesse. Por exemplo, um modelo de geração de termos identifica palavras simples como "*tokens*" e aplica um processo de radicalização sobre as palavras, removendo-lhes sufixos ("*stemming process*"), com o uso de uma adaptação do algoritmo de Porter (1997) para inglês, português e espanhol, como o de Matsubara et al. (2003). O processo pode gerar apenas palavras simples e/ou multi-palavras; aqui tratados como unigramas e n-gramas. Por exemplo, um termo pode ser "artific", "intelligenc", "conferenc", "artific intelligenc". Entre as técnicas mais utilizadas, encontram-se a radicalização, lematização e substantivação.

A **técnica de *stemming*** ou "stemmização", aqui tratada como radicalização, reduz as palavras às suas formas inflexionáveis e, às vezes, às suas derivações, ou seja, eliminação de prefixos e sufixos das palavras e a utilização dos verbos em suas formas infinitivas (Manning et al., 2008), sendo cada palavra analisada isoladamente. O processo de *stemming* depende da linguagem, pois algoritmos de radicalização que necessitam de conhecimento lingüístico (Silla Jr e Kaestner, 2002). Porém, os mais atuais não utilizam informações do contexto para determinar o sentido correto de cada palavra (Ebecken et al., 2003), devido à maioria das palavras poder ser considerada por um único significado, sendo que nestes casos, o contexto não ajudará no processo de *stemming*. No entanto, deve-se atentar aos possíveis erros resultantes de análise incorreta do sentido das palavras, já que tais algoritmos ignoram o significado dos termos levando, às vezes, a erros. Como algoritmos de radicalização para a língua inglesa, pode-se citar o de Lovins (1968), Stemmer S (Harman, 1991) e de Porter (*Porter Stemming Algorithm*) (Porter, 1997); para o português

[1] Disponível em: http://labic.icmc.usp.br/ferramentas/iesystem

[2] Na literatura, muitas vezes a geração de atributos é chamada de **extração de atributos** e nesta tese os dois termos são indistintamente utilizados.

e espanhol surgiram algumas adaptações a partir desses algoritmos, como a ferramenta PreTexT (Martins et al., 2003) e Stemmer (Orengo e Huyck, 2001).

A **lematização** reduz as palavras a seus lemas. Para as línguas francesa e inglesa, tem-se o software proprietário Sphinx, versão 4 (Sphinx Brasil, 2008). Já para o português, tem-se o Lematizador de Nunes (Nunes, 1996). Há também ferramentas que etiquetam morfologicamente as palavras, como MXPOST (*Maximum entropy pos tagger*) (Ratnaparkhi, 1996), TreeTagger (Schmid, 1994) e o etiquetador de Brill (Brill, 1995), sendo necessário, em seguida, a utilização de ferramenta que lematize tais palavras.

A **substantivação** ou "nominalização" faz com que as palavras exibam um comportamento sintático/semântico semelhante àquele próprio de um nome[3]. Para a língua portuguesa, citam-se as ferramentas FORMA e CHAMA (Gonzalez et al., 2006).

Vários experimentos comparando subjetiva e objetivamente o uso dessas diferentes técnicas na TopTax podem ser encontrados em Conrado et al. (2008) e em Conrado et al. (2009). Essas avaliações costumam ser feitas na etapa de pós-processamento, depois de geradas taxonomias a partir de atributos gerados com cada técnica. Como as avaliações subjetivas sempre apresentam alguma dependência em relação ao avaliador, em trabalhos posteriores incorporaram-se avaliações objetivas. No trabalho de Conrado et al. (2009) foi verificado subjetiva e objetivamente o comportamento de cada técnica em diferentes domínios de conhecimento. Para as comparações objetivas utilizou-se um *thesaurus* expandido. Nesse caso, foi necessário estabelecer um domínio fixo, expandir um *thesaurus* do mesmo e então comparar os resultados de cada técnica (detalhes em Conrado (2009)).

Para os *bigramas* e *trigramas* são realizados alguns testes estatísticos para sua seleção, considerando-se a ordem de ocorrência de cada um de seus componentes em toda a coleção. Essas escolhas tentam identificar potenciais colocações do domínio, que são posteriormente tratadas como termos únicos; além disso, são eliminadas combinações estatisticamente não significantes (Banerjee e Pedersen, 2003; Manning e Schütze, 1999). Por exemplo, eliminadas as *stopwords* e caracteres especiais na frase: "tecnicas inteligencia artificial aplicacoes tem sido utilizadas sistemas apoio decisao"; "tecnicas-inteligencia-artificial", "inteligencia-artificial", "sistemas-apoio-decisao" e "sistemas-decisao" são potenciais colocações e termos do domínio, mas "tecnicas-inteligencia", "artificial-aplicacoes", "aplicacoes-tem-sido" e "aplicacoes-tem" não acrescentam termos relevantes ao domínio; logo, se os testes assim o indicarem essas combinações são eliminadas. Como os testes estatísticos, em geral, resultam em *ranking* de atributos, pode-se optar por um corte pré-definido ou deixar que o especialista do domínio decida o corte. Um bom caso de uso do *ranking* com corte pré-definido é utilizar o *ranking* da razão de máxima verossimilhança (ver (Manning e Schütze, 1999)) e então cortar com um *p-value* predefinido, geralmente 5%.

O número de atributos, mesmo após um cuidadoso processo de limpeza da coleção e geração dos mesmos, é exageradamente grande e, não possui representatividade em cada documento da coleção, levando a representações esparsas de suas ocorrências. Desta forma, outra tarefa de especial interesse no pré-processamento é a **seleção de atributos**. A escolha de um método de seleção de atributos adequado nesta sub-etapa contribuirá muito com a performance dos algoritmos de aprendizado utilizados nas demais etapas.

[3]http://www.dacex.ct.utfpr.edu.br/paulo3.htm

Como a metodologia faz uso de aprendizado não-supervisionado, sem rótulos de classes, existe uma dificuldade a mais ao se delimitar o que é relevante. O método mais comumente utilizado corresponde aos cortes de Luhn (Luhn, 1958). Para realizar esses cortes ordenam-se as freqüências de ocorrência dos termos na coleção e, após obter o gráfico dessa curva, escolhem-se, subjetivamente, os pontos de corte, próximos aos pontos de inflexão, considerando-se que palavras com freqüência de ocorrência muito baixa ou muito alta sejam irrelevantes. Porém, a eliminação de termos de baixa ocorrência não é consenso, sendo inclusive privilegiada pelo indexador *tf-idf* - (*term frequence, inverse document frequency*, que é muito utilizado em recuperação de informação (Salton e McGill, 1986).

Outro método bastante difundido é o de Salton et al. (1975), o qual usa a medida *df* (*document frequency*: número de documentos nos quais um determinado termo aparece) para a seleção dos termos. Nele é sugerido considerar termos que possuam *df* entre 1% e 10% do número total de documentos, sendo considerado um corte bastante agressivo, reduzindo muito o número de termos.

Nogueira et al. (2008a) propõem um método denominado "Luhn-DF" aproveitando as idéias dos cortes de Luhn e Salton. No método proposto, geram-se os histogramas das medidas de *df* dos termos de forma descendente, efetuando os cortes nos pontos de inflexão da curva de tendência, tal qual o primeiro método supra-citado. Este método seleciona, assim como o de Salton, termos cuja *df* não é tão grande, nem tão pequena, adotando pontos de corte mais maleáveis e, por vezes, menos agressivos.

Além destes três métodos, outros métodos de seleção de atributos baseados em variância dos termos e similaridade de documentos são explorados em trabalhos da área, apontando bons resultados. Por exemplo, *Term Contribution* (TC) (Liu et al., 2003) é uma medida que representa o quanto um termo contribui para a similaridade entre documentos na coleção. Outra medida é *Term Variance* (TV) (Liu et al., 2005), a qual calcula a variância de todos os termos da coleção, atribuindo os maiores *scores* àqueles termos que não possuem baixa freqüência em documentos e possuem uma distribuição não-uniforme ao longo da coleção. Por fim, *Term Variance Quality* (TVQ) (Dhillon et al., 2003) é uma outra medida bastante similar à TV, usando a variância no cálculo da qualidade dos termos.

Vários experimentos comparando a eficiência de métodos de seleção não supervisionada de atributos podem ser encontrados em (Nogueira, 2009); dentre os quais, duas novas adaptações de métodos e suas avaliações. E, ainda, esse trabalho é voltado à seleção de atributos unigramas, dado que para n-gramas utilizados em conjunto existem outras considerações a serem feitas.

Quando se trabalha com atributos n-gramas, após selecioná-los por critérios estatísticos como em Banerjee e Pedersen (2003), ainda restam muitas redundâncias. Essa redundâncias referem-se às repetições do mesmo termo em suas formas compostas. Por exemplo da oração "tecnicas inteligencia artificial aplicacoes tem sido utilizadas sistemas apoio decisao", tem-se os seguintes unigramas, bigramas e trigramas seqüencialmente formados: tecnicas, inteligencia, tecnicas-inteligencia, artificial, tecnicas-inteligencia-artificial, inteligencia-artificial, aplicacoes, artifical-aplicacoes, inteligencia-artificial-aplicacoes, ter, aplicacoes-ter, artificial-aplicacoes-ter, ser, ter-ser, aplicacoes-ter-ser, utilizar, ser-utilizar,

ter-ser-utilizar, sistemas, utilizar-sistemas, ser-utilizar-sistemas, apoio, sistemas-apoio, utilizar-sistemas-apoio, decisao, apoio-decisao, sistemas-apoio-decisao. Logo, muito desses n-gramas deveriam ser descartados, uma vez que se decida ficar com algum de mais alta ordem. No trabalho de Moura et al. (2008d) é proposto um modelo automático para a eliminação dessas redundâncias - descrito no Capítulo 4, que integra a geração e seleção de atributos unigramas e inclui a seleção e eliminação de redundâncias entre os atributos n-gramas.

Ao final dessa etapa, gera-se a **matriz atributo-valor** para representar os dados (documentos). Cada documento é representado por um vetor comum de atributos, no qual os valores considerados para cada atributo representam suas freqüências de ocorrência nos documentos. A matriz atributo-valor é utilizada na etapa de **extração de padrões**, na qual realiza-se um agrupamento hierárquico dos documentos e rotula-se automaticamente os agrupamentos obtidos.

3.3.3 Extração de padrões

Esta metodologia se propõe a obter uma taxonomia de tópicos a partir de um agrupamento hierárquico de documentos de um único domínio. Os tópicos são estatisticamente inferidos a partir de listas de termos mais discriminativos de cada agrupamento, esse processo é conhecido como rotulação de agrupamentos (*cluster labeling*).

Tem-se como meta separar o algoritmo de agrupamento utilizado do processo de seleção de termos mais discriminativos em cada grupo para a rotulação dos grupos; lembrando que os termos correspondem aos atributos escolhidos. Dessa forma pode-se utilizar diferentes algoritmos de agrupamento e, conseqüentemente, trabalhar diferentes critérios de corte dos agrupamentos, sem prejuízo das demais escolhas.

Dada a matriz atributo-valor gerada a partir dos critérios de escolha do vocabulário, geração e seleção de atributos, constrói-se um agrupamento. Como todos os métodos encontram-se separados, pode-se testar vários métodos e algoritmos de agrupamento, bem como propor novos. Há muitos algoritmos interessantes para serem utilizados neste passo, tanto aglomerativos quanto divisivos. Para dar continuidade a esta etapa, precisa-se unicamente da hierarquia obtida e dos valores de freqüência de cada atributo em cada grupo obtido.

O agrupamento hierárquico de documentos geralmente fornece muitos grupos e subgrupos, dificultando a análise e interpretação imediata dos resultados. Assim, procurou-se desenvolver um processo para podar o dendrograma em grupos coesos, em um alto nível de divisão da hierarquia obtida, para que possa ser mais facilmente visualizado. Além disso, permite-se que os subgrupos possam ser também visualizados, sem perda de informação. Para isto, foi desenvolvido um método de seleção de grupos significativos baseado em medidas de qualidade de agrupamento. Foram consideradas algumas medidas de qualidade tradicionais da literatura, adaptadas para a alta dimensionalidade encontrada em coleções textuais, de forma a prevalecer as relações hierárquicas entre os grupos. A proposta apresentada por Marcacini et al. (2009) mantém a hierarquia original, mas permite observá-la de um modo reduzido, mais coeso e em tópicos de níveis mais altos, facilitando

a análise e interpretação dos resultados. A seleção é apoiada por medidas de qualidade dos grupos obtidos; foram utilizadas as medidas de similaridade intra-grupo e a inter-grupos, que analisam, respectivamente, a homogeneidade de cada grupo (coesão interna) e a qualidade da relação hierárquica entre pais e filhos.

Dada uma hierarquia qualquer de documentos, para a qual sejam conhecidas as freqüências de ocorrência de cada termo em cada nível da hierarquia (grupo), aplica-se um processo próprio de **rotulação da hierarquia** (Moura et al., 2008c) que depende exclusivamente das freqüências observadas de cada termo em cada grupo de documentos. Essa rotulação permite gerar uma versão da taxonomia na qual não se encontram repetições de termos ao longo do mesmo ramo; dado a suposição que rótulos de nós pais aplicam-se também aos seus filhos. A rotulação é gerada a partir dos termos mais discriminativos de cada nó da hierarquia, funcionando como mais um processo de seleção de atributos. No entanto essa segunda seleção de atributos é um processo supervisionado, pois os diversos grupos e subgrupos são considerados classes. Para cada classe selecionam-se os atributos mais vinculados a ela, decompondo os grupos em tabelas de contingência, veja Capítulo 5 para detalhes. Além da seleção dos termos mais discriminativos para cada nó, também é proposta uma metodologia de avaliação da rotulação obtida, detalhada no Capítulo 5.

Para teste dos métodos de rotulação da hierarquia, vêm sendo utilizados **algoritmos de agrupamento hierárquico** aglomerativos, com a medida de similaridade de cosseno, para obter a relação hierárquica entre os textos da coleção. Os algoritmos utilizados têm sido variados entre o "*single linkage*" para identificar documentos que fogem aos padrões na coleção (discussão sobre "*outliers*" em (Mardia et al., 1979)) e, o "*average*" ou o "*complete linkage*" para obter a hierarquia final.

3.3.4 Pós-processamento

No **pós-processamento** realizam-se as avaliações disponíveis para a taxonomia obtida.

Sempre que possível, realiza-se uma análise subjetiva por parte dos especialistas de domínio. Algumas vezes o objetivo dessa análise é investigar qual a melhor forma de geração dos atributos, por exemplo, se lematizados, radicalizados ou substantivados. Outras vezes, o objetivo é selecionar um método de rotulação de agrupamentos. Enfim, estabelecido o objetivo, o procedimento é gerar diferentes taxonomias, uma para cada técnica analisada. Selecionam-se ramos semelhantes nas diferentes taxonomias, isto é, ramos que correspondem aos mesmos grupos de documentos. Então, pede-se a equipe de especialistas que atribua notas a cada um dos nós desses ramos, devendo consultar cada um dos resultados antes de decidir pelas notas. A seguir, faz-se uma análise de variância a fim de verificar se todos os efeitos foram previstos no modelo (detalhes (Searle, 1971)), isto é, se o avaliador teve um efeito significativo em relação à técnica ou se a técnica foi avaliada independentemente. A seguir faz-se uma comparação múltipla de médias das notas em cada técnica para ver qual saiu estatisticamente melhor.

Isso foi feito durante toda uma fase de desenvolvimento da metodologia focado no processo de rotulação dos agrupamentos (Moura e Rezende, 2007; Moura et al., 2008a,c). Muitas vezes a análise subjetiva indicou direções para a geração e seleção de atributos, e,

até para novas seleções da coleção de textos; mas, não conseguia avaliar apenas a forma de obter os atributos mais discriminativos. Embora os resultados possam ser considerados satisfatórios, quase sempre as questões abertas respondidas pelos especialistas indicavam direções relativas a outras variáveis, não somente à escolha dos termos. Porém, esse tipo de avaliação forneceu subsídios às seleções de técnicas de geração de atributos (Conrado et al., 2008, 2009).

A avaliação objetiva, que vem sendo utilizada na TopTax, é baseada no resultado de um procedimento de recuperação de informações da coleção, utilizando-se os conjuntos de descritores encontrados para cada grupo como expressões de busca. Sempre que os valores de revocação da recuperação de informação apresentam-se próximos a cem por cento, conclui-se que a aproximação gerada como taxonomia foi boa.

A taxonomia é, também, subjetivamente avaliada por especialistas do domínio, apoiados por ferramentas de visualização. Pode-se optar por primeiro realizar uma avaliação objetiva e, conforme o resultado, o especialista decide se avalia subjetivamente. Caso a taxonomia não seja satisfatória, escolhem-se etapas do processo às quais se pode voltar, para que alguma mudança satisfaça critérios aí identificados. Quando a taxonomia é aceita, ela pode ser diretamente utilizada ou, ainda, passar por um processo de refino.

No **refinamento** da taxonomia obtida, os especialistas podem editá-la, auxiliados por medidas estatísticas de validação das edições realizadas. Os especialistas são orientados pelos critérios de corte de ramos dos agrupamentos e de corte do vocabulário utilizado, mas detêm a decisão final, mesmo se contrária às indicações estatísticas. A ferramenta de edição permite que os especialistas criem novos termos, que podem ser resumos de conjuntos de termos; por exemplo, determinar que um ramo corresponde à "família" e que pode ser "filho", "pai" e "mãe". Ou seja, a edição permite que se defina um termo geral, para abrigar os termos identificados discriminativos no ramo, de forma completamente subjetiva.

3.3.5 Uso do conhecimento

Na etapa de **uso do conhecimento**, a taxonomia pode ser utilizada para facilitar processos de organização e recuperação de informação, bem como a própria compreensão da coleção de textos organizada; ou mesmo servir de suporte a sistemas de apoio à decisão.

Para apoiar esta etapa, é importante que sejam desenvolvidas técnicas de auxílio ao especialista do domínio com a finalidade de facilitar o entendimento e a utilização do conhecimento adquirido (Silberschatz e Tuzhilin, 1995). As técnicas e ferramentas para visualização de informação cumprem este objetivo, pois facilitam o reconhecimento de relacionamentos, tendências e padrões do conjunto de dados analisado, potencializando a exploração do conhecimento (Rezende et al., 1998).

Nesta metodologia, é utilizada uma ferramenta específica para exploração da taxonomia (Marcacini, 2008), que permite navegar na taxonomia por meio de diferentes representações gráficas, sumarizar grandes quantidades de dados, realizar buscas, visualizar e editar informações pertencentes a cada tópico, comparar duas ou mais taxonomias obtidas por métodos distintos e, também, fornecer uma interface de avaliação para que seja possível

validar, de forma subjetiva, a taxonomia resultante de um processo de mineração de textos. Essa ferramenta é denominada TaxTools[4].

3.4 Estudos de casos com a metodologia TopTax

Vários conjuntos de dados têm sido utilizados em experimentos com objetivos específicos de produção, como no caso do estudo de tendências em produção científica em gado de corte e leite junto à Embrapa Pecuária Sudeste (Mazzari, 2007; EMBRAPA, 2008), organização de uma biblioteca digital (Marcacini et al., 2007), suporte à gestão de competências (Marcacini et al., 2008) e suporte à gestão do conhecimento na área de engenharia de produção (Rezende et al., 2008).

Adicionalmente, há os estudos de prospecção e validação de métodos e tecnologias utilizados no projeto de construção do ambiente computacional para viabilizar a aplicação da metodologia de forma automatizada. Para esses estudos, têm sido utilizadas bases de dados do domínio de inteligência artificial e afins, devido à disponibilidade de especialistas do domínio em grupos de pesquisa próximos que colaboram nas avaliações.

A seguir serão apresentados dois exemplos. O primeiro, no domínio de lingüística computacional, que fez parte da prospecção e validação de métodos junto aos especialistas em inteligência artificial. O segundo, no domínio de gado de corte Canchim, que fez parte de um trabalho de conclusão de curso de graduação.

3.4.1 Exemplo no domínio lingüística computacional

O uso da metodologia é exemplificado com uma coleção de textos do domínio de lingüística computacional, formada por 51 artigos completos, em português, apresentados nos últimos *workshops* anuais de "Tecnologia da Informação e da Linguagem Humana", TIL, de 2005 a 2007. Ao processo de padronização, aplicado à base de textos do TIL, foi acrescentada a retirada de informações às autorias dos artigos e às referências bibliográficas[5], para que os termos utilizados nesses itens, bem como os nomes próprios, não interferissem nos resultados.

A metodologia foi aplicada aos artigos padronizados, primeiro obtendo-se os atributos na forma de *unigramas* com o uso da ferramenta PreText (Matsubara et al., 2003), que elimina as *stopwords* e inflexões. Para os 51 artigos foram obtidos 7300 *unigramas*. A seguir, obteve-se o gráfico de ocorrência dos *unigramas* na coleção e foram definidos os cortes de Luhn, com valor mínimo 12 e máximo 450; resultando em 1154 *unigramas* (atributos). O agrupamento hierárquico dos documentos foi obtido com o MatLab, escolhendo-se similaridade de cosseno e o algoritmo "*average*". À hierarquia aplicou-se o algoritmo próprio de obtenção dos rótulos dos agrupamentos (Moura et al., 2008c). Então, na primeira avaliação subjetiva visual, realizada pela equipe de mineração de textos, verificou-se que o resultado ainda era muito complexo e não permitia reconhecer eficientemente os grupos.

[4]Disponível em: http://labic.icmc.usp.br/ferramentas/taxtools
[5]http://labic.icmc.usp.br/projects

As observações dos resultados preliminares viabilizaram a identificação de problemas, como o grande valor estatisticamente discriminativo dos verbos e de algumas palavras muito comuns a artigos científicos, como “trabalho”, “artigo”, “figura” e “tabelas”. Assim, realizaram-se escolhas subjetivas de *unigramas* a serem eliminados do processo, que passaram a ser tratados como *stopwords* específicas da coleção de documentos. Ao todo foram identificados 290 *unigramas* a serem eliminados. Após isso, repetiram-se os procedimentos do pré-processamento, aplicando-se os cortes de Luhn, coincidentemente, novamente com um mínimo de 12 e máximo de 450, deixando-se 870 *unigramas* a serem considerados como atributos. E, novamente, foi aplicado o algoritmo de agrupamento *“average”* com medida de similaridade de cosseno e, a seguir, o método próprio de rotulação de agrupamentos hierárquicos, ordenando-se os rótulos por freqüência e mantendo um máximo de 30 rótulos por grupo, isto é, os 30 *unigramas* mais discriminativos de cada grupo.

Para exemplificar o resultado, na Fig. 3.2, pode-se observar um ramo da taxonomia, cujo comportamento é muito semelhante antes e depois do corte dos termos subjetivamente considerados *stopwords*. Alguns dos *unigramas* cortados estão hachurados na figura, em sua parte 1. Deve-se notar que a sua retirada não alterou a interpretação dos ramos. Na parte 2 da figura, nota-se claramente que os artigos sob esse ramo tratam de modelos de extração de conceitos e termos de domínio, que manipulam ontologias, por meio de várias ferramentas de construção de ontologias.

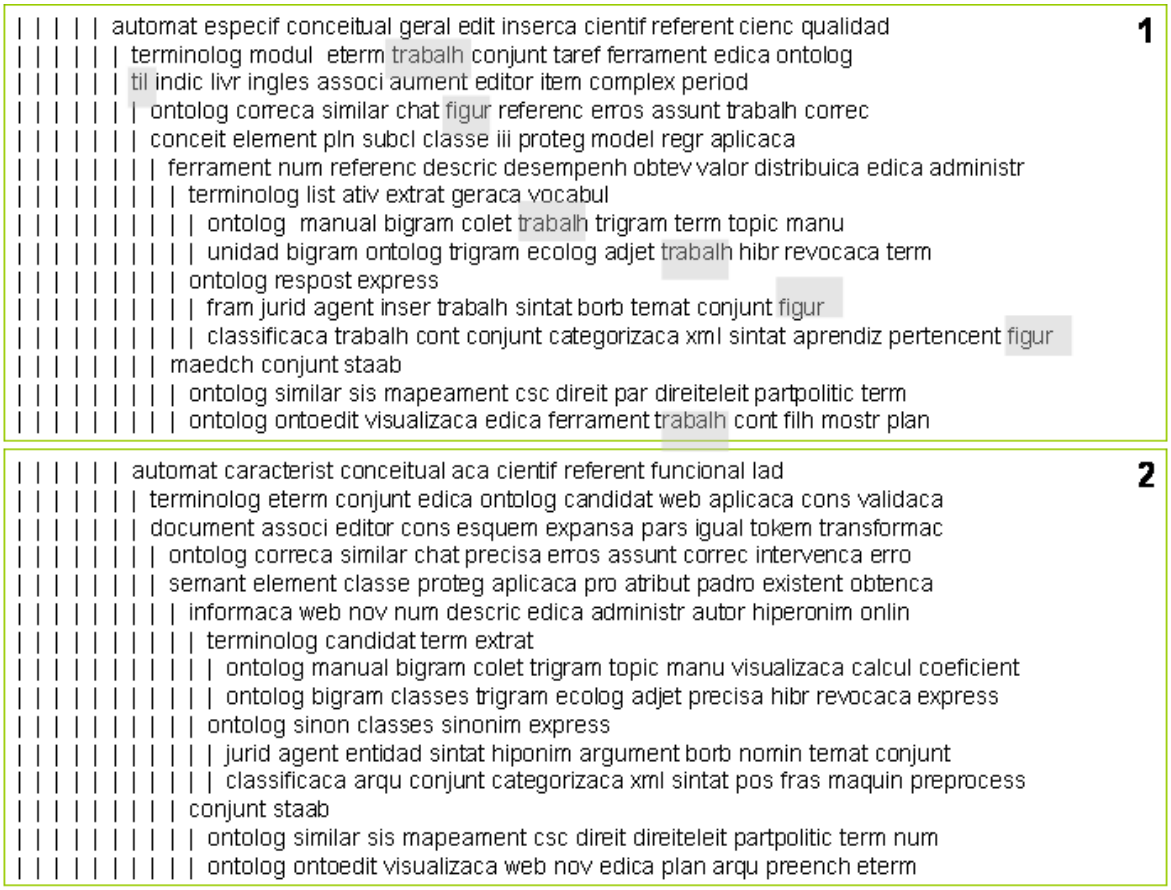

Figura 3.2: Exemplo de visualização, antes (1) e depois (2) de retiradas *stopwords* específicas da coleção

Em alguns outros experimentos, nos quais as coleções correspondiam a resumos dos

textos, e não a textos completos, o peso dos verbos e termos específicos, como tabela e figura, não tiveram tanta importância. Dessa experiência, incorporou-se aos requisitos do ambiente computacional de suporte à metodologia, também a necessidade de uma ferramenta para eliminar verbos na obtenção dos atributos. Esse tipo de ferramenta é incorporado ao pré-processamento, na etapa de geração de atributos e, hoje, representada pela integração da ferramenta de substantivação à TopTax.

3.4.2 Exemplo no domínio gado de corte Canchim

Esse exemplo foi originalmente trabalhado visando-se a construção de uma Agência de Informação (Souza et al., 2006) exclusivamente sobre gado de corte da raça Canchim, auxiliado por um grupo especialista na área da Embrapa Pecuária Sudeste. O auxílio desse grupo de especialistas foi de grande valia na fundamentação dos conceitos, bem como a seleção de técnicas e ferramentas para a TopTax. A seguir, tem-se um exemplo com um conjunto de 17 textos sobre a raça Canchim. Esse exemplo permite que se tenha uma boa idéia da eficiência da metodologia em encontrar tópicos e subtópicos de interesse.

Na Tabela 3.1 são apresentadas as seqüências de unigramas, bigramas e trigramas em termos numéricos. Note-se que, na primeira coluna são apresentados os números originalmente obtidos, apenas com a eliminação das *stopwords*. Na segunda coluna estão os números após a eliminação das *stopwords da coleção*, que são definidas e explicadas no Capítulo 4. Na terceira coluna encontra-se o resultado da aplicação de um filtro muito simples, que consiste em ficar apenas com os termos que ocorrem em pelo menos dois arquivos da coleção, isto é, termos que apresentam $df \geq 2$. Esse filtro é interessante ao considerar a aplicação de um agrupamento aglomerativo, pois nesse, um termo só consegue definir um grupo se pertencer a pelo menos dois arquivos. Finalmente na última coluna encontramos o número de atributos utilizado, pois a eliminação de redundâncias foi aplicada a todos eles, e esta também é detalhada no Capítulo 4.

termos	originais	removendo *stopwords* da coleção	filtro:$2 \geq df \leq 17$	após remover redundâncias
unigramas	725	329	329	242
bigramas	2112	1551	410	49
trigramas	2585	2002	317	147

Tabela 3.1: Termos para 17 artigos científicos sobre gado de corte Canchim

Na Figura 3.3 pode-se observar a taxonomia obtida. Foi aplicado um algoritmo aglomerativo do tipo "average", utilizando distância de cosseno e o procedimento de rotulação descrito em Moura et al. (2008c). No tema "manejo de acasalamento" (dois níveis acima do grifado na figura) e "monta" (um nível acima) tem-se o "comportamento sexual dos touros" e sua influência na fertilidade do rebanho, que é justamente o texto correspondente a essa folha da árvore. Pode-se notar, na figura, a relação de termos mais discriminativos obtidos para cada um dos nós observados e para o grifado o detalhe na tabela entre o nó e o texto, na figura.

Ainda, deve-se notar que o tratamento aplicado a esses dezessete artigos foi bem simples. A geração e seleção de atributos foi a mais trivial executada sob a metodologia.

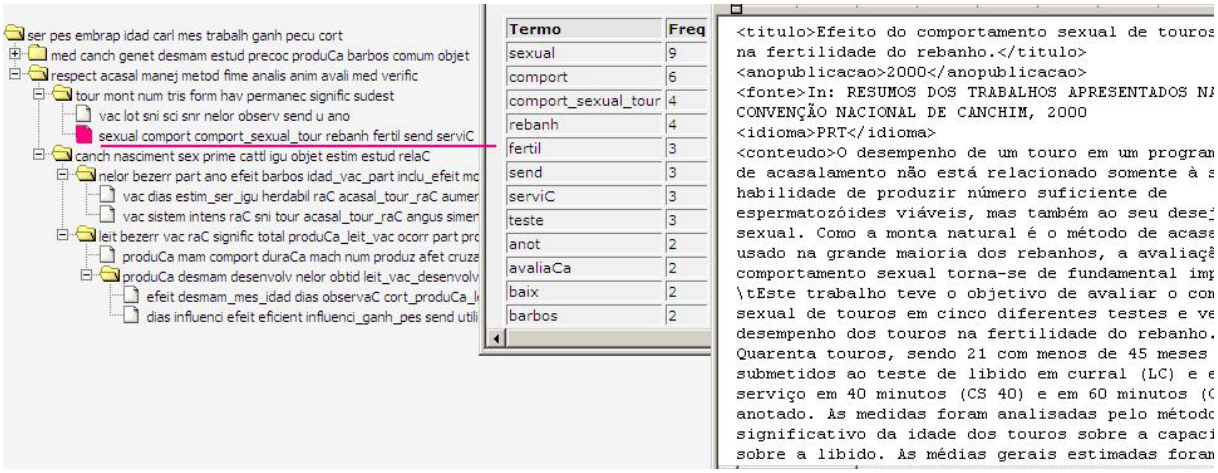

Figura 3.3: Visualização da taxonomia inicial para Gado Canchim, termos do nó grifado e texto vinculado

Poder-se-ia tentar o uso de outras formas de agrupamento e também aplicar os procedimentos de corte propostos por Marcacini et al. (2009). Logo, mesmo utilizando a TopTax em uma configuração muito simples, o resultado obtido já é bastante interessante.

3.5 Considerações finais

Neste capítulo é apresentada a proposta de metodologia para a construção de taxonomias de tópicos, instanciada a partir de um processo de mineração de textos, utilizando agrupamento de documentos e rotulação automática dos mesmos. A metodologia vem sendo utilizada em processos de organização de coleções de documentos e na identificação de tendências de tópicos e sub-tópicos nas coleções.

A configuração atual do ambiente de suporte computacional à metodologia permite utilizá-la e produzir bons resultados, como visto nos exemplos. Adicionalmente, prevê-se a avaliação e validação de outras técnicas em cada parte do processo, com o objetivo de avaliar técnicas que melhor se adaptem a cada tipo de coleção de textos. Com isso, espera-se melhor auxiliar os especialistas do domínio, interessados em construir as taxonomias. O auxílio deve ser na forma de ferramental para executar diferentes tarefas sob diferentes condições e no fornecimento de parâmetros que norteiem a tomada de decisões.

Sobre a avaliação das taxonomias, além dos métodos já empregados na TopTax, vem-se trabalhando um modelo de validação contra taxonomias *gold*. Uma taxonomia *gold* é aquela que pode ser considerada bem aceita por uma comunidade que detém o conhecimento do domínio. Nesse trabalho, ainda em desenvolvimento, são utilizadas as medidas de avaliação propostas pelo grupo do Taxaminer (Kashyap et al., 2005a). As taxonomias consideradas *golds* foram emprestadas da Agência Embrapa. Para realizar a comparação utiliza-se o Thesagro (*Thesaurus* Agrícola Nacional) para encontrar termos similares a serem considerados nas comparações. A vantagem esperada com esse trabalho é que ele se aproxime do que se gostaria de obter em um julgamento subjetivo; pois, checa-se o que se deveria ter contra o que se obtém, tentando-se obter uma medida que melhor reflita a aproximação ou separação dos dois padrões.

Como contribuições acadêmicas a metodologia corresponde a uma tese de doutorado

com duas contribuições inéditas, o modelo de eliminação de redundâncias em atributos n-gramas (Capítulo 4) e dois novos métodos de rotulação de agrupamento hierárquico de documentos (Capítulo 5). Além disso, dois mestrados derivaram-se da escolha de estratégias de geração e seleção de atributos (Conrado, 2009; Nogueira, 2009). Em um terceiro mestrado, em andamento, trabalham-se regras de associação entre os atributos para encontrar os rótulos dos agrupamentos. Ainda, no pré-processamento tem-se uma iniciação científica que resultou na ferramenta IESystem. De outra iniciação científica e posterior conclusão de curso resultou a TaxTools, um ambiente que integra agrupamento de documentos, os algoritmos de corte (Marcacini et al., 2009), os algoritmos de rotulação e visualização das taxonomias obtidas (Marcacini, 2008).

Como trabalhos futuros, há dois mestrados em andamento diretamente inspirados na TopTax. O primeiro trata do processo de crescimento das taxonomias de tópicos, isto é, da atualização das taxonomias por meio da inserção e remoção de documentos. Nesse trabalho serão avaliadas técnicas existentes e provavelmente proposto algum ajuste ou integração das mesmas. O segundo mestrado mescla um pouco de geração e seleção de atributos, trabalhando com regras de associação para gerar novos atributos. Nesse trabalho será avaliada a redução do espaço de características e se há melhora significativa na interpretação dos atributos. Mais dois doutorados também estão inspirados diretamente na metodologia. O primeiro, na linha do primeiro mestrado, também investigará técnicas de crescimento das taxonomias, devendo produzir um novo modelo adaptado às necessidades da TopTax e melhorado em relação aos já existentes. O segundo doutorado, segue a linha de geração de atributos, dando continuidade ao trabalho de mestrado de Conrado (2009). Nesse trabalho, serão explorados modelos híbridos de geração de atributos. Os modelos híbridos misturam técnicas puramente estatísticas de extração de atributos a técnicas de lingüística computacional; esses modelos têm sido bem avaliados, produzindo bons resultados (Lima et al., 2007; Conrado, 2009). Assim, espera-se incrementar o uso de modelos híbridos de extração de atributos na TopTax. Ainda, há um terceiro doutorado, em andamento, trabalhando numa linha um pouco diferente do paradigma de aprendizado de maquina atual, que é o estatístico. Nesse terceiro doutorado, vem-se trabalhando com lógica fuzzy na obtenção e rotulação dos agrupamentos. Pretende-se com isso avaliar o uso de *soft assignment* aos agrupamentos na metodologia, porém dentro do paradigma de aprendizado nebuloso.

Assim, nos dois próximos capítulos são tratadas apenas as contribuições inéditas desta tese de doutorado. O Capítulo 4 apresenta o modelo proposto para a seleção de atributos n-gramas. O Capítulo 5 descreve os modelos propostos para a rotulação de agrupamentos hierárquicos de documentos e a metodologia de validação da rotulação.

Capítulo

4

Seleção de atributos do tipo n-gramas

Uma proposta completa para resolver o problema de selecionar automaticamente atributos não redundantes do tipo n-gramas é apresentada neste capítulo. Geralmente, o uso de n-gramas é um requisito para melhorar a interpretação subjetiva dos resultados das tarefas de mineração de textos, considerando-se o $n \geq 2$. Nesses casos, os termos compostos, *multi words* ou n-gramas, são estatisticamente gerados e selecionados. Após a seleção, em geral, há a presença de redundâncias, por exemplo, o termo "inteligência artificial" e seus componentes "inteligência" e "artificial". Isso pode ser nitidamente observado pelos resultados experimentais com o uso de um classificador obtido com os dados originais e um outro com os dados sem as redundâncias. Nota-se nos resultados que, como esperado, não há perda de efetividade na categorização após a eliminação das redundâncias. Além disso, os classificadores que lidam com os dados não redundantes recebem conjuntos menores de atributos para trabalhar, o que pode ser interessante para algumas tarefas de extração de padrões. Deve-se salientar que o método proposto é útil a qualquer algoritmo de aprendizado de máquina aplicado a uma tarefa de mineração de textos.

4.1 Introdução

Relembrando o processo de mineração de textos é dividido em cinco passos: identificação do problema, pré-processamento, extração de padrões, pós-processamento e uso do conhecimento. Freqüentemente, o pré-processamento tende a ser considerado como um passo de menor importância, ou de menor interesse que os demais, devido à ausência de *glamour* técnico e excesso de tarefas manuais. Basicamente, este passo é responsável pela transformação da coleção de textos em uma forma útil aos algoritmos de aprendizado de máquina. Logo, envolve operações como o tratamento e limpeza dos dados e redução de atributos. Neste passo, as características mais representativas da coleção devem ser consideradas, procurando-se eliminar as características irrelevantes. Das dificuldades e desafios

impostos neste passo nada trivial resultam a boa ou má qualidade dos dados analisados e, conseqüentemente, a eficiência dos algoritmos de aprendizado de máquina utilizados e grande parte da confiança nos resultados obtidos. Além disso, é importante considerar o conhecimento explícito carregado pelos atributos de um processo de mineração de textos, pois os resultados finais ou intermediários sempre podem ser subjetivamente analisados enriquecendo todo o processo (Weiss et al., 2005).

A probabilidade de considerar um atributo como um termo de domínio em uma coleção de textos pode ser entendida como uma propriedade qualitativa e discriminativa. Essa propriedade deve ser buscada tanto para palavras simples como compostas, como "inteligência" e "inteligência artificial", na mesma coleção de textos. Embora muitos autores considerem o aumento de dimensionalidade mais prejudicial que benéfico, o aumento do número de atributos, quando se usam palavras simples e compostas parecem facilitar a interpretação de resultados e, conseqüentemente, ser mais adequado aos usuários dos resultados em mineração de textos (Miao et al., 2005; Moura e Rezende, 2007). Em geral, quando se trabalha com palavras simples e compostas, ocorrem algumas redundâncias. Nesses casos, poder-se-ia descartar alguns desses atributos por meio de um filtro adequado a eles. Há poucas iniciativas na literatura para filtrar atributos entre n-gramas, exceto pelos trabalhos de Zhang e Zhu (2007) e de Dejun e Maosong (2004), resumidos a seguir.

De acordo com Dejun e Maosong (2004) há um tipo de redundância resultante de características sobrepostas para caracteres n-gramas em chinês. Caracteres n-gramas são as composições de radicais ou partes de palavras, que para línguas orientais, como o chinês e japonês, muitas vezes correspondem a uma palavra ou conceito. Então, eles propuseram um modelo para evitar as redundâncias, que eles acreditam ser independente de linguagem, nas quais a sobreposição é estimada e absorvida pelo modelo de atributos. Além disso, o modelo não utiliza rótulos na seleção dos n-gramas, logo, não se restringe a aprendizado supervisionado. Na proposta de Zhang e Zhu (2007), um comprimento de janela entre palavras que formam um n-grama é estimado em um modelo empírico que procura um valor discriminativo para esse tamanho, baseado na freqüência dos n-gramas nos documentos e no rótulo da classe. Conseqüentemente o modelo é mais adequado a aprendizado supervisionado, porém, é uma solução interessante para remover palavras entre os n-gramas que efetivamente interessam.

Neste capítulo é apresentada uma proposta para resolver o problema de selecionar automaticamente n-gramas bons e não redundantes como atributos. O modelo é útil tanto para aprendizado não supervisionado como supervisionado, pois depende exclusivamente de testes estatísticos sobre os atributos candidatos. Na próxima seção faz-se a contextualização do problema, definindo os conceitos utilizados e explicando as redundâncias consideradas. Então, na seção seguinte, a metodologia para selecionar os n-gramas sem redundâncias é descrita, bem como seus algoritmos são apresentados. Na seção de experimentos pode-se observar que a redução do espaço de atributos não sacrifica os resultados obtidos; o que é um bom indício de que os objetivos foram satisfeitos. Finalmente, algumas considerações gerais e trabalhos futuros são discutidos na subseção 4.7.

4.2 Contextualização do problema

Em um processo de mineração de textos escolhe-se a forma de representação dos atributos, após selecionar a coleção de textos com a qual se vai trabalhar. Supondo que se vá trabalhar exclusivamente com a coleção de textos como uma *bag of words*, desconsiderando o contexto semântico, utilizam-se as palavras presentes nos textos ou suas combinações. Essas palavras ou suas combinações são chamadas de n-gramas. Por exemplo, um n-grama pode ser o unigrama "inteligência", ou o bigrama "inteligência artificial", ou o trigrama "técnicas inteligência artificial", ou outros. Ainda, podem-se reduzir as palavras a radicais, lemas ou outras formas simplificadas. Seja qual a forma escolhida, tem-se um enorme conjunto de atributos, do qual devem ser descartados os que sejam pouco discriminativos e/ou redundantes, ou seja, aqueles que não sejam indispensáveis ao processo.

Para acompanhar o desenvolvimento das idéias apresentadas, será utilizada a seguinte coleção de cinco textos, completamente fictícios (Tabela 4.1):

texto 1: Textual data mining is data mining over textual data. Textual data are textual collections in several formats. Data mining has several definitions, but data are always data that are aimed at a special meaning and use.
texto 2: Textual data mining is used in business process. It is very useful for knowledge managment.
texto 3: Textual data mining is a special kind of data mining. It works with textual data in textual collections. Data has to be transformed in useful knowledge.
texto 4: Textual data mining is a powerful tool. Many applications use textual data to discover market tendencies. Textual data can be magazine articles, internet chats, etc, every textual collection in a digital media. Data mining over textual data is the process which transforms data in useful information.
texto 5: Textual data mining is the subject of this text. We like to talk about data mining over textual data. Textual data are in textual collections. Data mining is the process of transforming data in useful information.

Tabela 4.1: Coleção de cinco textos fictícios utilizados como exemplo

Embora essa coleção seja pequena e, aparentemente, gere poucos atributos, ela ilustra bem os problemas aqui tratados. Partindo desses textos, provavelmente nos seria útil obter o seguinte conjunto de atributos: "textual data mining", "data mining", "textual data", "textual collections", "special meaning", "useful information", "useful knowledge", "data" e "use". Outras combinações de palavras, ou mesmo apenas palavras simples, parecem ser não interessantes para discriminar o conjunto de textos. O problema é como chegar automaticamente a um conjunto de atributos enxuto e discriminativo.

Na próxima subseção discorre-se sobre modelos de geração de n-gramas, ou seja, como identificá-los nas coleções de textos, codificá-los e apresentá-los. A seguir, Subseção 4.2.2, mostra-se como selecionar os atributos, que técnicas podem ser usadas para decidir se o atributo é ou não mantido no processo de MT. Na seqüência, Subseção 4.2.3, as técnicas e testes mais comumente utilizados para a seleção de palavras compostas são apresentados. Supondo-se que se tenham atributos unigramas e n-gramas, $n \geq 2$, na Subseção 4.2.4 discute-se como selecionar os atributos a serem efetivamente utilizados. Finalmente, na

Subseção 4.2.4, mostra-se o que está sendo tratado como redundância neste ponto do pré-processamento das coleções de textos.

4.2.1 Modelo de geração de n-gramas

Ferramentas que geram as combinações de palavras presentes nos textos, geralmente, descartam as *stopwords* especificadas e depois aplicam algum processo de identificação das palavras de interesse, dependendo de como se definem os *tokens* para seus analisadores léxicos. As *stopwords* e os *tokens* são definidas pelo usuário da ferramenta. Por exemplo, para a ferramenta *N-gram Statistical Package*, **NSP** (Banerjee e Pedersen, 2003), define-se um conjunto de expressões regulares que delimitam os *tokens* de interesse; já para a ferramenta PreText (Matsubara et al., 2003) define-se qual a língua na qual os textos estão escritos, pois ela trabalha com radicalização inflexional.

Assim, para a coleção de textos do exemplo, considerando-se a geração de unigramas, bigramas e trigramas com a ferramenta PreText, têm-se: 32 unigramas, 51 bigramas e 63 trigramas. Todos esses n-gramas encontram-se na Tabela 4.2. Cada valor apresentado nas caselas deve ser interpretado como $tf : (df/\#d) : termo$. O valor tf (*term frequency*) corresponde à freqüência total de ocorrências do termo na coleção. O valor df (*document frequency*) corresponde ao número de documentos nos quais o termo é presente. O valor $\#d$ corresponde ao número total de documentos na coleção. E, finalmente *termo* corresponde ao n-grama gerado.

Observando-se os n-gramas apresentados na Tabela 4.2 é evidente que nem todas as combinações levam a termos de interesse. Logo, grande parte dessas combinações deve ser cortada do conjunto de atributos.

4.2.2 Seleção de atributos unigramas

Na TopTax (Capítulo 3) há um passo sobre seleção de atributos para o qual foram desenvolvidos e testados modelos de filtros para unigramas (Nogueira, 2009). Resumidamente, os filtros são baseados em freqüências observadas ou variâncias dos atributos na coleção.

Relembrando, o filtro mais comumente utilizado é o método de Luhn (Luhn, 1958), em que se escolhem dois pontos de corte após apresentar graficamente os atributos ordenados pela sua freqüência de ocorrência na coleção. Os pontos escolhidos costumam estar próximos aos pontos de inflexão da curva descrita pelo gráfico, considerando-se que atributos de baixíssima e altíssima freqüência sejam irrelevantes. Também, bastante comum é o método de Salton que usa os termos cujas freqüências de ocorrência nos documentos da coleção estejam no intervalo de um a dez por cento do total de documentos (Salton et al., 1975). No trabalho de Nogueira et al. (2008b) é proposta uma variação do método de Luhn adaptada à idéia desse método de Salton, chamada de Luhn-DF; mas, que carrega os mesmos problemas do método de Luhn, pois implica na escolha empírica dos pontos de corte. Há também alguns filtros com base em representatividade da variância, citados anteriormente no Capítulo 3 e completamente descritos em Nogueira (2009).

32 unigramas			
1:(1/5):market	1:(1/5):articl	1:(1/5):kind	2:(2/5):knowledg
1:(1/5):busi	1:(1/5):digit	1:(1/5):applic	2:(2/5):inform
1:(1/5):media	1:(1/5):discov	1:(1/5):text	3:(3/5):process
1:(1/5):tool	1:(1/5):magazin	1:(1/5):chat	3:(3/5):transform
1:(1/5):power	1:(1/5):definit	1:(1/5):tendenc	4:(4/5):collect
1:(1/5):internet	1:(1/5):work	1:(1/5):format	11:(5/5):mine
1:(1/5):subject	1:(1/5):mean	1:(1/5):talk	17:(5/5):textual
1:(1/5):manag	1:(1/5):aim	2:(2/5):special	24:(5/5):data

51 bigramas		
1:(1/5):process_knowledg	1:(1/5):data_data	1:(1/5):mine_subject
1:(1/5):discov_market	1:(1/5):digit_media	1:(1/5):data_process
1:(1/5):transform_knowledg	1:(1/5):mine_power	1:(1/5):special_mean
1:(1/5):definit_data	1:(1/5):mine_work	1:(1/5):articl_internet
1:(1/5):chat_textual	1:(1/5):mine_data	1:(1/5):data_discov
1:(1/5):internet_chat	1:(1/5):data_transform	1:(1/5):data_magazin
1:(1/5):text_talk	1:(1/5):subject_text	1:(1/5):data_aim
1:(1/5):collect_digit	1:(1/5):tendenc_textual	1:(1/5):collect_format
1:(1/5):market_tendenc	1:(1/5):format_data	2:(2/5):data_inform
1:(1/5):mine_busi	1:(1/5):applic_textual	2:(2/5):process_transform
1:(1/5):mine_definit	1:(1/5):work_textual	2:(2/5):transform_data
1:(1/5):busi_process	1:(1/5):magazin_articl	2:(2/5):collect_data
1:(1/5):media_data	1:(1/5):mine_special	3:(3/5):mine_textual
1:(1/5):power_tool	1:(1/5):tool_applic	4:(4/5):textual_collect
1:(1/5):aim_special	1:(1/5):knowledg_manag	5:(3/5):data_textual
1:(1/5):kind_data	1:(1/5):mine_process	11:(5/5):data_mine
1:(1/5):special_kind	1:(1/5):talk_data	13:(5/5):textual_data

63 trigramas		
1:(1/5):textual_data_discov	1:(1/5):data_mine_special	1:(1/5):collect_data_mine
1:(1/5):mine_process_transform	1:(1/5):collect_digit_media	1:(1/5):articl_internet_chat
1:(1/5):data_mine_subject	1:(1/5):power_tool_applic	1:(1/5):data_mine_process
1:(1/5):magazin_articl_internet	1:(1/5):aim_special_mean	1:(1/5):work_textual_data
1:(1/5):data_data_aim	1:(1/5):mine_work_textual	1:(1/5):data_mine_power
1:(1/5):internet_chat_textual	1:(1/5):mine_subject_text	1:(1/5):text_talk_data
1:(1/5):data_aim_special	1:(1/5):subject_text_talk	1:(1/5):media_data_mine
1:(1/5):process_knowledg_manag	1:(1/5):talk_data_mine	1:(1/5):tendenc_textual_data
1:(1/5):collect_format_data	1:(1/5):digit_media_data	1:(1/5):definit_data_data
1:(1/5):chat_textual_collect	1:(1/5):data_mine_work	1:(1/5):data_process_transform
1:(1/5):data_mine_data	1:(1/5):discov_market_tendenc	1:(1/5):mine_busi_process
1:(1/5):data_mine_definit	1:(1/5):textual_collect_format	1:(1/5):mine_power_tool
1:(1/5):data_transform_knowledg	1:(1/5):special_kind_data	2:(2/5):data_textual_data
1:(1/5):kind_data_mine	1:(1/5):textual_data_magazin	2:(2/5):transform_data_inform
1:(1/5):mine_data_mine	1:(1/5):applic_textual_data	2:(2/5):process_transform_data
1:(1/5):collect_data_transform	1:(1/5):format_data_mine	2:(2/5):textual_collect_data
1:(1/5):data_magazin_articl	1:(1/5):data_mine_busi	3:(3/5):mine_textual_data
1:(1/5):market_tendenc_textual	1:(1/5):mine_definit_data	3:(3/5):data_mine_textual
1:(1/5):busi_process_knowledg	1:(1/5):textual_data_process	3:(3/5):data_textual_collect
1:(1/5):tool_applic_textual	1:(1/5):textual_collect_digit	5:(3/5):textual_data_textual
1:(1/5):mine_special_kind	1:(1/5):data_discov_market	5:(5/5):textual_data_mine

Tabela 4.2: Exemplo de atributos n-gramas gerados para a coleção de cinco textos

Filtros baseados em freqüência ou contribuição para a variância são mais confiáveis quando aplicados a conjuntos de atributos do tipo unigrama, pois tendem a favorecer termos com valores de freqüências de ocorrências nem tão raras nem tão comuns. Experimentalmente, é fácil observar que para palavras compostas essa regra não vale, basta observar a Tabela 4.2. Normalmente, as palavras compostas com alta freqüência de ocorrência é que são os melhores candidatos a atributos, embora essa observação não garanta que eles sejam candidatos a termos de domínio ou colocações (Manning e Schütze, 1999).

No entanto, sempre o bom senso deve prevalecer. Por exemplo, para os unigramas

apresentados na Tabela 4.2 obtidos de apenas cinco documentos, se um desses filtros for aleatoriamente escolhido, pode-se perder atributos valiosos para a análise. No entanto, apenas a título de exemplo, serão escolhidos apenas os unigramas, ou todos os n-gramas, que aparecem em pelo menos dois dos textos da coleção, ou seja, com $df \geq 2$. A medida **df**, *document frequency*, corresponde ao número de documentos da coleção nos quais o termo é presente. Dessa forma, se esses unigramas forem utilizados em um agrupamento hierárquico aglomerativo, com $df \geq 2$ eles terão maior chance de contribuir para a formação de um grupo. Com esse primeiro filtro, $df \geq 2$, obtêm-se os n-gramas da Tabela 4.3.

9 unigramas - após a aplicação do filtro $df \geq 2$	
2:(2/5):special	4:(4/5):collect
2:(2/5):knowledg	11:(5/5):mine
2:(2/5):inform	17:(5/5):textual
3:(3/5):process	24:(5/5):data
3:(3/5):transform	

Tabela 4.3: Atributos unigramas da Tabela 4.2 após corte $df \geq 2$

4.2.3 Seleção de atributos n-gramas

Para selecionar n-gramas com $n \geq 2$ a regra de alta ou baixa freqüência nos textos não pode ser diretamente aplicada. Faz-se necessário verificar se as combinações encontradas são de fato importantes ou puramente casuais. Qualquer teste que seja empregado estará sempre se referindo à coleção de textos disponível, logo, uma combinação de palavras (gramas) será considerada casual ou importante na coleção de textos.

Candidatos a colocações, expressões ou frases interessantes em uma coleção de textos podem ser estatisticamente indicados sem uma análise lingüística rigorosa para, então, serem selecionados por especialistas do domínio (Manning e Schütze, 1999). Esses termos candidatos serão automaticamente selecionados e considerados os atributos mais interessantes no processo de Mineração de Textos, caso não se possa contar com mais análises (Conrado et al., 2008).

De acordo com Banerjee e Pedersen (2003) e Manning e Schütze (1999), pode-se estabelecer hipóteses de dependência entre as combinações de n-gramas geradas ou *rankings* de dependência. Se as combinações são independentes, então o n-grama é desconsiderado na análise, caso contrário pode-se fazer um *ranking*. Quando se obtém um *ranking* é necessário decidir empiricamente um ponto de corte. Nesse caso, pode-se utilizar valores tabelados como os valores críticos de aceitação da hipótese de independência entre os gramas que formam o n-grama, estabelecendo-se níveis de significância. Ou, pode-se estabelecer um intervalo de confiança para a aceitação dos valores obtidos. Seja qual for o estimador escolhido, corte ou valor crítico utilizado, deve-se verificar se está adequado ao problema e à distribuição da freqüência observada dos dados. Há várias possibilidades para os testes e a ferramenta NSP (Banerjee e Pedersen, 2003) possui vários desses testes estatísticos implementados.

Neste trabalho, todos os testes foram ilustrados com o estimador Q de Yule e um intervalo de confiança de 95% para a estimativa obtida. Por exemplo, um bigrama genérico

19 bigramas		
1:(1/5):process_knowledg	1:(1/5):data_transform	4:(4/5):textual_collect
1:(1/5):transform_knowledg	2:(2/5):process_transform	4:(4/5):collect_data
1:(1/5):mine_special	2:(2/5):transform_data	5:(4/5):mine_textual
1:(1/5):data_special	2:(2/5):mine_process	7:(4/5):data_textual
1:(1/5):data_data	2:(2/5):data_inform	11:(5/5):data_mine
1:(1/5):data_process	3:(2/5):mine_data	13:(5/5):textual_data
1:(1/5):special_data	.	.

24 trigramas		
1:(1/5):mine_process_transform	1:(1/5):mine_data_data	3:(3/5):data_textual_data
1:(1/5):mine_special_data	1:(1/5):data_mine_special	3:(2/5):data_mine_data
1:(1/5):textual_data_process	1:(1/5):mine_process_knowledg	4:(4/5):data_textual_collect
1:(1/5):special_data_mine	2:(2/5):transform_data_inform	4:(4/5):textual_collect_data
1:(1/5):data_transform_knowledg	2:(2/5):process_transform_data	5:(4/5):mine_textual_data
1:(1/5):data_data_special	2:(2/5):data_mine_process	5:(4/5):data_mine_textual
1:(1/5):collect_data_transform	2:(2/5):mine_data_mine	5:(5/5):textual_data_mine
1:(1/5):data_process_transform	3:(3/5):collect_data_mine	7:(4/5):textual_data_textual

Tabela 4.4: Bigramas e trigramas da Tabela 4.2 após filtro dos unigramas

é w_1w_2, sendo w_1 a primeira palavra (grama) e w_2 a segunda palavra (grama) na sua composição. A freqüência observada de w_1w_2 na coleção é f_{11}, a freqüência observada de w_1 na primeira posição em todos os bigramas obtidos da coleção é $f_{1.}$ e, a freqüência de w_2 na segunda posição entre todos os bigramas obtidos da coleção é $f_{.1}$. A freqüência de todos os bigramas na coleção é $f_{..}$. Para cada bigrama w_1w_2 tabulam-se todos esses valores como na Tabela 4.5.

	w_2	$!(w_2)$	
w_1	f_{11}	$f_{1.} - f_{11}$	$f_{1.}$
$!(w_1)$	$f_{.1} - f_{11}$	$(f_{..} - f_{1.}) - (f_{.1} - f_{11})$	$f_{..} - f_{1.}$
	$f_{.1}$	$f_{...} - f_{..3}$	$f_{..}$

Tabela 4.5: Freqüências com w_1 na primeira e w_2 na segunda posição do bigrama w_1w_2

Para testar a hipótese de independência entre w_1 ocorrendo na primeira posição e w_2 ocorrendo na segunda posição do bigrama w_1w_2, usando o Q de Yule, os seguintes valores devem ser obtidos (para detalhes veja Bishop et al. (1975)):

$$d = ((f_{..} - f_{1.}) - (f_{.1} - f_{11}))$$

$$\alpha = \frac{f_{11} * d}{(f_{1.} - f_{11}) * (f_{.1} - f_{11})}$$

$$Q_1 = (\alpha - 1)/(\alpha + 1)$$

$$v = \sqrt{\frac{1}{f_{11}} + \frac{1}{f_{1.} - f_{11}} + \frac{1}{f_{.1} - f_{11}} + \frac{1}{d}}$$

$$\sigma_{Q_1} = \frac{1}{2} * (1 - Q_1^2) * v$$

$$IC_{Q_1} = [Q_1 - 2 * \sigma_{Q_1}, Q_1 + 2 * \sigma_{Q_1}]$$

Para que $w1$ e $w2$ nessas posições sejam independentes, o valor α deve se aproximar de 1, logo Q_1 deve se aproximar de 0. O intervalo de confiança da estimativa de Q é construído para cobrir 95% dos valores, dado que Q tem uma distribuição Normal (veja Bishop et al. (1975)). Assim, se $0 \notin IC_{Q_1}$ então w_1w_2 é candidato à colocação de domínio,

ou termo ou frase interessante; e, o bigrama w_1w_2 é candidato a atributo do processo de MT.

Por exemplo, na Tabela 4.4 estão os bigramas e trigramas gerados após o corte de $df \geq 2$ para os unigramas do exemplo anterior e mais algumas *stopwords* especialmente definidas (cujas corretas definições são colocadas com detalhes na próxima seção).

Para esses bigramas e trigramas, da Tabela 4.4, pode-se calcular os mais estatisticamente significantes e depois obter seus *rankings*. Por exemplo, para o bigrama "data_transform" na Tabela 4.6, tem-se os seguintes valores:

	$transform$	$!(transform)$	
$data$	1	23	24
$!(data)$	2	37	39
	3	60	63

Tabela 4.6: Freqüências para o bigrama "data_transform"

$$\alpha = \frac{1 \times 37}{2 \times 23} = 0.8043, \; Q = \frac{\alpha - 1}{\alpha + 1} = -0.1084$$

$$v = \sqrt{\frac{1}{1} + \frac{1}{37} + \frac{1}{2} + \frac{1}{23}} = 1.2532$$

$$\sigma_Q = \frac{1}{2} * (1 - Q^2) * v = 1.2385$$

$$IC_Q = [-1.3469, 1.1301] \Rightarrow 0 \in IC_Q$$

Logo, "data_transform" pode ser desconsiderado do conjunto de atributos, dado que a combinação w_1w_2 é casual segundo o teste realizado. Já para o bigrama "mine_textual", representado na Tabela 4.7, tem-se os seguintes valores:

	$textual$	$!(textual)$	
$mine$	5	6	11
$!(mine)$	7	45	52
	12	51	63

Tabela 4.7: Freqüências para o bigrama "mine_textual"

$$\alpha = \frac{5 \times 45}{6 \times 7} = 5.3571, \; Q = \frac{\alpha - 1}{\alpha + 1} = 0.6854$$

$$v = \sqrt{\frac{1}{5} + \frac{1}{7} + \frac{1}{6} + \frac{1}{45}} = 0.7293$$

$$\sigma_Q = \frac{1}{2} * (1 - Q^2) * v = 0.1933$$

$$IC_Q = [0.2987, 1.0721] \Rightarrow 0 \notin IC_Q$$

Assim, o bigrama "mine_textual" é incluído na lista de atributos.

Para testar a validade de um trigrama, como $w_1w_2w_3$, precisa-se de sua freqüência observada f_{123} e, também da freqüência observada de w_1w_2, $f_{12.}$, sempre que esse bigrama

é presente na primeira posição de todos os trigramas da coleção. É necessária a freqüência observada do bigrama w_2w_3, $f_{.23}$, na segunda e terceira posições de todos os n-gramas da coleção. E, ainda, a freqüência observada de w_1, ($f_{1..}$), na primeira posição de todos os trigramas da coleção e de w_3, ($f_{..3}$), na terceira posição de todos os trigramas da coleção. Finalmente, precisa-se da freqüência observada de todos os trigramas da coleção ($f_{...}$). Isto porque, é preciso realizar o teste de w_1 com w_2w_3 e w_1w_2 com w_3.

Assim, na Tabela 4.8, a freqüência de w_1 na primeira posição ($f_{1..}$) e w_2w_3 na segunda e terceira posições ($f_{.23}$) do trigrama $w_1w_2w_3$ (f_{123}) são ilustradas, bem como seus complementos.

	w_2w_3	$!(w_2w_3)$	
w_1	f_{123}	$f_{1..} - f_{123}$	$f_{1..}$
$!(w_1)$	$f_{.23} - f_{123}$	$(f_{...} - f_{1..}) - (f_{.23} - f_{123})$	$f_{...} - f_{1..}$
	$f_{.23}$	$f_{...} - f_{.23}$	$f_{...}$

Tabela 4.8: Exemplo com a freqüência de w_1 e w_2w_3 em $w_1w_2w_3$

Para testar a hipótese de independência entre as ocorrências de w_1 na primeira posição e w_2w_3 na segunda e terceira posições, usa-se o estimador Q de Yule após o cálculo dos seguintes valores:

$$d_1 = ((f_{...} - f_{1..}) - (f_{.23} - f_{123}))$$

$$\alpha_1 = \frac{f_{123} * d_1}{(f_{.23} - f_{123}) * (f_{1..} - f_{123})}$$

$$Q_1 = (\alpha_1 - 1)/(\alpha_1 + 1)$$

$$v_1 = \sqrt{\frac{1}{f_{123}} + \frac{1}{f_{1..} - f_{123}} + \frac{1}{f_{.23} - f_{123}} + \frac{1}{d_1}}$$

$$\sigma_{Q_1} = \frac{1}{2} * (1 - Q_1^2) * v_1$$

$$IC_{Q_1} = [Q_1 - 2 * \sigma_{Q_1}, Q_1 + 2 * \sigma_{Q_1}]$$

E, na Tabela 4.9, os valores de freqüência de w_1w_2 na primeira e segunda posições ($f_{12.}$) e w_3 na terceira posição ($f_{..3}$) do trigrama $w_1w_2w_3$ (f_{123}) são organizados.

	w_3	$!(w_3)$	
w_1w_2	f_{123}	$f_{12.} - f_{123}$	$f_{12.}$
$!(w_1w_2)$	$f_{..3} - f_{123}$	$(f_{...} - f_{12.}) - (f_{..3} - f_{123})$	$f_{...} - f_{12.}$
	$f_{..3}$	$f_{...} - f_{..3}$	$f_{...}$

Tabela 4.9: Exemplo com a freqüência de w_1w_2 e w_3 em $w_1w_2w_3$

E, para testar a hipótese de independência entre as ocorrências de w_3 na terceira posição e w_1w_2 na primeira e segunda posições, usando o Q de Yule, obtém-se os seguintes valores:

$$d_2 = ((f_{...} - f_{..3}) - (f_{12.} - f_{123}))$$

$$\alpha_2 = \frac{f_{123} * d_2}{(f_{12.} - f_{123}) * (f_{..3} - f_{123})}$$

$$Q_2 = (\alpha_2 - 1)/(\alpha_2 + 1)$$

$$v_2 = \sqrt{\frac{1}{f_{123}} + \frac{1}{f_{..3} - f_{123}} + \frac{1}{f_{12.} - f_{123}} + \frac{1}{d_2}}$$

$$\sigma_{Q_2} = \frac{1}{2} * (1 - Q_2^2) * v_2$$

$$IC_{Q_2} = [Q_2 - 2 * \sigma_{Q_2}, Q_2 + 2 * \sigma_{Q_2}]$$

Se $0 \notin IC_{Q_1}$ e $0 \notin IC_{Q_2}$ então $w_1 w_2 w_3$ é selecionado como atributo.

Por exemplo, para o trigrama "data_process_transform" da Tabela 4.4, as freqüências dos n-gramas que o compõe estão ilustradas nas Tabelas 4.10 e 4.11.

	$process_transform$	$!(process_transform)$	
$data$	1	20	21
$!(data)$	1	36	37
	2	56	58

Tabela 4.10: Freqüências para o trigrama "data" com "_process_transform"

	$transform$	$!(transform)$	
$data_process$	1	0	1
$!(data_process)$	2	55	57
	3	55	58

Tabela 4.11: Freqüências para o trigrama "data_process" com "_transform"

A partir das Tabelas 4.10 e 4.11 calculam-se os seguintes valores:

$$\alpha = \frac{1 \times 36}{1 \times 20} = 1.8, \; Q_1 = \frac{\alpha - 1}{\alpha + 1} = 0.2857$$

$$v = \sqrt{\frac{1}{1} + \frac{1}{36} + \frac{1}{1} + \frac{1}{20}} = 1.4415$$

$$\sigma_{Q_1} = \frac{1}{2} * (1 - Q_1^2) * v = 0.6619$$

$$IC_{Q_1} = [-1.0381, 1.6095] \Rightarrow 0 \in IC_{Q_1}$$

$$\alpha \; com \; divisor \;\; zero \;\; \Rightarrow \;\; Q_2 = indefinido$$

Logo, esse trigrama não é considerado importante como atributo, pois $0 \in IC_{Q_1}$, embora no segundo teste tenha sido constatada associação completa. A associação completa deve-se à freqüência zerada de *data_process* com outro unigrama na terceira posição. Para detalhes de como identificar associação completa veja Bishop et al. (1975).

4.2.4 Atributos n-gramas e redundâncias

Quando se deseja trabalhar com atributos unigramas e n-gramas, $n \geq 2$, faz-se necessário combinar as técnicas de seleção dos dois tipos de atributos. Pode-se simplesmente utilizar uma técnica para selecionar unigramas e depois algum critério para selecionar os demais n-gramas. O problema é que também ao utilizar mais de um grama, novamente tem-se um grande aumento no número de atributos e talvez mais redundância, pois os n-gramas são formados pelas combinações dos unigramas apresentados e são menos freqüentes que esses.

n-gramas escolhidos subjetivamente	freqüência observada	subtrai ocorrência trigrama	subtrai ocorrência bigrama
textual data mining	1		
data mining	3	3-1=2	
textual data	3	3-1=2	
textual collections	1	1-0=1	
several formats	1	1-0=1	
several definitions	1	1-0=1	
several meaning	1	1-0=1	
special meaning	1	1-0=1	
data	7	7-1=6	6-2-2-1=1
mining	3	3-1=2	2-2=0
textual	4	4-1=3	3-2-1=0
collections	1	1-0=1	1-1=0
several	2	2-0=2	2-1-1=0
formats	1	1-0=1	1-1=0
definitions	1	1-0=1	1-1=0
aimed	1	1-0=1	1-0=1
special	1	1-0=1	1-1=0
meaning	1	1-0=1	1-1=0
use	1	1-0=1	1-0=1

Tabela 4.12: Seleção de n-gramas não redundantes, fonte: Moura et al. (2008d)

Para facilitar a ilustração do que está sendo considerado **redundância** separaram-se apenas os atributos do primeiro texto do exemplo da Tabela 4.1. Na Tabela 4.12 são mostrados apenas os trigramas, bigramas e unigramas presentes nesse primeiro texto, que subjetivamente foram considerados úteis.

Na **segunda coluna** da Tabela 4.12 estão as freqüências observadas para cada n-grama. Os n-gramas foram ordenados da mais alta ordem para a mais baixa e pela freqüência de ocorrência dentro dessa primeira categoria. Na **terceira coluna** dessa tabela, o número de ocorrências foi subtraído do n-grama de mais alta ordem, trigrama, que contem o n-grama da linha. Finalmente, na **quarta coluna**, as freqüências de cada unigrama foram subtraídas das dos bigramas, não removidos na primeira subtração, nos quais eles ocorrem. Logo, os n-gramas cujas freqüências foram zeradas nesse processo são descartados do conjunto de atributos; pois, suas freqüências são combinações lineares das anteriores. Ou seja, esses n-gramas removidos representam redundâncias no conjunto de atributos inicial. Conseqüentemente, os n-gramas cujas freqüências não foram zeradas são não redundantes e devem ser utilizados. Devido aos n-gramas de mais alta ordem serem menos freqüentes que os de mais baixa ordem e formados pelos mesmos, opta-se por primeiro selecionar os de mais alta ordem e então eliminar as redundâncias.

Assim, neste capítulo é proposto um modelo de seleção de n-gramas. Nesse modelo faz-se a seleção de atributos unigramas e, então, adaptam-se algumas regras de seleção de atributos para selecionar também os n-gramas com $n \geq 2$ e eliminar as redundâncias presentes entre eles.

4.3 Método proposto para a seleção de atributos n-gramas

No método proposto e desenvolvido neste trabalho, supõe-se que a coleção de textos pertence a um domínio de conhecimento restrito e que será representada como uma *bag*

of words.

O método proposto é baseado nas idéias apresentadas na seção anterior. Ele é dividido em quatro passos, como ilustrado na Figura 4.1. As subseções seguintes correspondem a cada um desses passos e os descrevem.

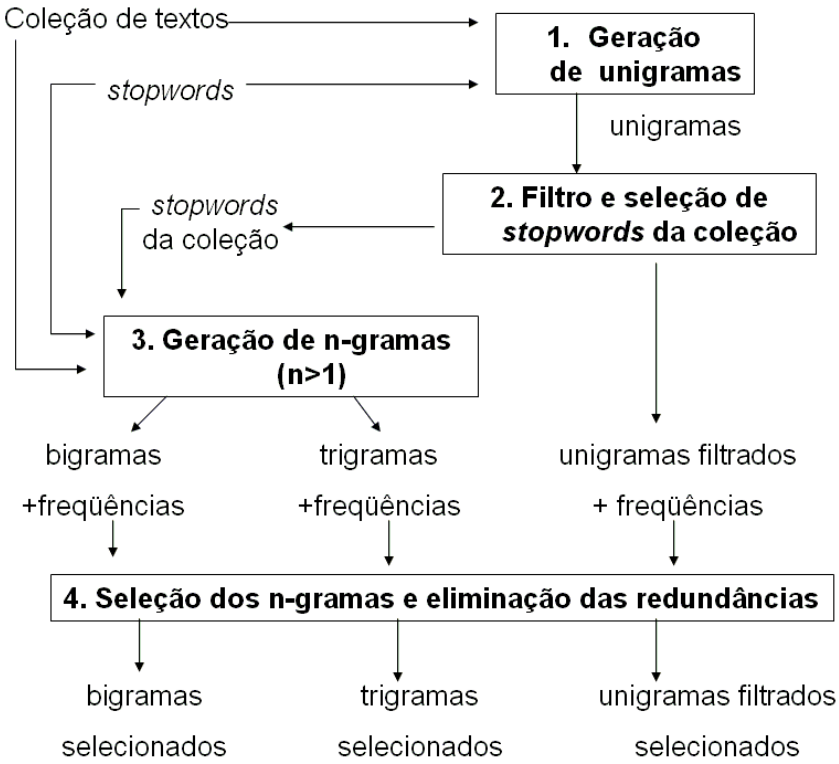

Figura 4.1: Passos do método proposto para a seleção de atributos n-gramas

4.3.1 Passo 1 - geração de unigramas

O primeiro passo é gerar todos os possíveis unigramas da coleção de textos. Pode-se usar qualquer ferramenta que permita gerar os unigramas e contabilizar suas freqüências, por exemplo, a PreText ou a NSP. Neste passo, as entradas do processo são a coleção de textos e a lista de *stopwords* a serem eliminadas do processo. Logo, a saída deste processo são os unigramas gerados e contabilizados, pela sua freqüência de ocorrência na coleção de textos, medida tf.

4.3.2 Passo 2 - filtro e seleção de *stopwords* da coleção

No segundo passo, escolhe-se um filtro para os unigramas gerados, de acordo com as propostas apresentadas por Nogueira (2009) e as freqüências fornecidas.

Aplicado o filtro tem-se os unigramas filtrados e as *stopwords* da coleção. Os **unigramas filtrados** estão sendo tratados como aqueles que foram selecionados como atributos após a aplicação do filtro. Neste método proposto para a seleção de n-gramas, as ***stopwords* da coleção** correspondem aos unigramas que foram descartados pelo filtro. Esses unigramas que foram descartados são utilizados como *stopwords* no próximo passo.

Ainda, quando se utiliza uma ferramenta que promove a radicalização, como é o caso da PreText, há mais um item que merece atenção. Não basta relacionar os "radicais" cortados como *stopwords* da coleção, é preciso relacionar todas as suas inflexões presentes na coleção de textos; pois, todas as inflexões foram cortadas.

4.3.3 Passo 3 - geração dos n-gramas ($n > 1$)

As ***stopwords* da coleção** precisam ser descartadas da construção dos n-gramas de mais alta ordem, porque podem vir a interferir na sua construção e seleção. Por exemplo, para um n-grama como "pastagem pesquisada melhorada Embrapa", se as inflexões do verbo "pesquisar" tiverem sido descartadas, bem como o nome próprio "Embrapa", o correto seria gerar apenas o bigrama "pastagem melhorada".

Com base no exemplo utilizado na seção anterior, o trigrama "subject_text_talk" da Tabela 4.2, simplesmente deixa de existir após o corte de unigramas com $df < 2$, pois ficam como unigramas filtrados apenas os que possuem $df \geq 2$. E, por exemplo, o trigrama "textual_data_magazin" também deixa de existir após esse corte, mas o bigrama "textual_data" continua existindo.

Resumidamente, retirando-se as *stopwords* da coleção, geram-se os n-gramas de mais alta ordem formados apenas pelos unigramas estatisticamente mais discriminativos. Assim, o processo de geração de n-gramas neste passo, recebe como entrada a coleção de textos, a lista de *stopwords* e a lista de *stopwords* da coleção. Com isso, espera-se gerar apenas os n-gramas, $n > 1$, que não contenham os unigramas descartados pelo filtro.

No exemplo anterior, especificam-se as *stopwords* da coleção como os unigramas descartados após o corte $df \geq 2$ e todas as suas inflexões. Então, reaplicando a ferramenta PreText à coleção de documentos são gerados apenas os trigramas e bigramas relacionados na Tabela 4.4.

Tomando-se um exemplo completo, supor os passos de extração dos n-gramas na frase da Tabela 4.13. Na primeira linha, a frase está completa, nenhum tratamento lhe foi aplicado. Na segunda linha, as *stopwords* são removidas; essas *stopwords* são as geralmente removidas, como artigos, interjeições, etc; que, neste caso foram: "'à"', "'de"' e "'seus"'. Na terceira linha, algum processo de seleção de unigramas foi aplicado Nogueira et al. (2008b). Com a seleção dos unigramas, a frase ficou curta; dado que lhe são descontadas as *stopwords* da coleção, que foram: "'marrom"', "'acinzentada"', "'enfatizando"' e "'fortes"'. Conseqüentemente, para formar os n-gramas com $2 \geq n \geq 3$, ter-se-iam os resultados apresentados na quarta linha.

frase original	bateu à porta de madeira marrom acinzentada enfatizando seus fortes toques
removidas as stopwords	bateu porta madeira marrom acinzentada enfatizando fortes toques
escolhidos unigramas	bateu porta madeira toques
possíveis n-gramas	bateu-porta porta-madeira madeira-toques bateu-porta-madeira porta-mdeira-toques

Tabela 4.13: Extração de n-gramas

4.3.4 Passo 4 - seleção dos n-gramas e eliminação de redundâncias

Neste passo, os n-gramas com $n = 3, 2$ devem ser estatisticamente selecionados como candidatos a atributos. Isto é feito, realizando-se um conjunto de testes de dependência entre os n-gramas de mais baixa ordem que os formam, de acordo com suas freqüências de ocorrência na coleção de textos (Manning e Schütze, 1999; Banerjee e Pedersen, 2003). De modo a simplificar a escolha dos testes de independência e sua aderência ao problema, selecionou-se o estimador Q de Yule, que não depende da distribuição dos dados, aplicado a decomposições das tabelas de freqüências dos n-gramas[1].

Seja $w_1w_2w_3$ um trigrama com freqüência observada f_{123}; no qual, w_1 corresponde à palavra da primeira posição do trigrama, w_2 à palavra da segunda posição e w_3 à palavra da terceira posição. Deve ser obtida a freqüência observada de w_1w_2, $f_{12.}$, sempre que esse bigrama é presente na primeira posição de todos os trigramas gerados para a coleção. Também é necessária a freqüência observada do bigrama w_2w_3, $f_{.23}$, na segunda e terceira posições de todos os trigramas da coleção. E, ainda, a freqüência observada de w_1, ($f_{1..}$), na primeira posição de todos os trigramas da coleção. Finalmente, precisa-se da freqüência observada de todos os trigramas da coleção ($f_{...}$).

Na Tabela 4.14, a freqüência de w_1 na primeira posição ($f_{1..}$) e w_2w_3 na segunda e terceira posições ($f_{.23}$) do trigrama $w_1w_2w_3$ (f_{123}) são ilustradas, bem como seus complementos.

	w_2w_3	$!(w_2w_3)$	
w_1	f_{123}	$f_{1..} - f_{123}$	$f_{1..}$
$!(w_1)$	$f_{.23} - f_{123}$	$(f_{...} - f_{1..}) - (f_{.23} - f_{123})$	$f_{...} - f_{1..}$
	$f_{.23}$	$f_{...} - f_{.23}$	$f_{...}$

Tabela 4.14: Freqüências de w_1 e w_2w_3 em $w_1w_2w_3$

Para testar a hipótese de independência entre as ocorrências de w_1 na primeira posição e w_2w_3 na segunda e terceira posições, pode-se usar o estimador Q de Yule após o cálculo dos seguintes valores:

$$d_1 = ((f_{...} - f_{1..}) - (f_{.23} - f_{123}))$$

$$\alpha_1 = \frac{f_{123} * d_1}{(f_{.23} - f_{123}) * (f_{1..} - f_{123})}$$

$$Q_1 = (\alpha_1 - 1)/(\alpha_1 + 1)$$

$$v_1 = \sqrt{\frac{1}{f_{123}} + \frac{1}{f_{1..} - f_{123}} + \frac{1}{f_{.23} - f_{123}} + \frac{1}{d_1}}$$

$$\sigma_{Q_1} = \frac{1}{2} * (1 - Q_1^2) * v_1$$

$$IC_{Q_1} = [Q_1 - 2 * \sigma_{Q_1}, Q_1 + 2 * \sigma_{Q_1}]$$

Na tabela 4.15, são tabelados os valores de freqüência de w_1w_2 na primeira e segunda posições ($f_{12.}$) e w_3 na terceira posição ($f_{..3}$) do trigrama $w_1w_2w_3$ (f_{123}).

E, para testar a hipótese de independência entre as ocorrências de w_3 na terceira posição e w_1w_2 na terceira e segunda posições, usando o Q de Yule, obtém-se os seguintes

[1]Para detalhes sobre o estimador e decomposição das tabelas veja Bishop et al. (1975).

	w_3	$!(w_3)$	
w_1w_2	f_{123}	$f_{12.} - f_{123}$	$f_{12.}$
$!(w_1w_2)$	$f_{..3} - f_{123}$	$(f_{...} - f_{12.}) - (f_{..3} - f_{123})$	$f_{...} - f_{12.}$
	$f_{..3}$	$f_{...} - f_{..3}$	$f_{...}$

Tabela 4.15: w_1w_2 nas primeiras e segundas posições e w_3 nas terceiras posições de todos os trigramas da coleção

valores:

$$d_2 = ((f_{...} - f_{..3}) - (f_{12.} - f_{123}))$$

$$\alpha_2 = \frac{f_{123} * d_2}{(f_{12.} - f_{123}) * (f_{..3} - f_{123})}$$

$$Q_2 = (\alpha_2 - 1)/(\alpha_2 + 1)$$

$$v_2 = \sqrt{\frac{1}{f_{123}} + \frac{1}{f_{..3} - f_{123}} + \frac{1}{f_{12.} - f_{123}} + \frac{1}{d_2}}$$

$$\sigma_{Q_2} = \frac{1}{2} * (1 - Q_2^2) * v_2$$

$$IC_{Q_2} = [Q_2 - 2 * \sigma_{Q_2}, Q_2 + 2 * \sigma_{Q_2}]$$

Se $0 \notin IC_{Q_1}$ e $0 \notin IC_{Q_2}$ então $w_1w_2w_3$ é selecionado como atributo. Todos os trigramas selecionados como atributos são *rankeados* de acordo com suas freqüências de ocorrência na coleção de textos, isto é, pelos valores obtidos como f_{123}.

Tão logo todos os trigramas tenham sido testados, selecionados e *rankeados*, todos os bigramas e unigramas cujas freqüências observadas formam os trigramas por combinações lineares, são eliminados com o Algoritmo 1. Para isso, utilizam-se as freqüências observadas, notadas por O_f neste trabalho. Logo, tem-se: $O_{f_{ij}}$ com $i, j = 1, 2, 3$ para os bigramas e O_{f_i} com $i = 1, 2, 3$ para os unigramas. Todos os n-gramas são tratados pelo algoritmo na ordem em que foram *rankeados* pelas suas freqüências. A base do algoritmo é que para cada trigrama, há algum(ns) bigrama(s) e unigrama(s) com freqüência maior ou igual à deles. Nesse caso, a redundância funciona como um critério de seleção de atributos e é usada para eliminar essas ocorrências de n-gramas de mais baixa ordem. Deve-se salientar que, apenas os melhores candidatos a atributos entre os trigramas ficam na lista de atributos finais.

Input: *TriGramList*: lista de trigramas ordenados pelo valor de seus *ranks*
Output: *FinalAttributesList*: lista de atributos finais com freqüências atualizadas
for *todos os trigramas em TriGramList* **do**
  selecionar o trigrama melhor *rankeado*, $w_1w_2w_3$;
  procurar por w_1w_2, com $O_{f_{12}} \geq O_{f_{123}}$;
  procurar por w_2w_3, com $O_{f_{23}} \geq O_{f_{123}}$;
  procurar por w_1, com $O_{f_1} \geq O_{f_{123}}$;
  procurar por w_2, com $O_{f_2} \geq O_{f_{123}}$;
  procurar por w_3, com $O_{f_3} \geq O_{f_{123}}$;
  subtrair $O_{f_{123}}$ de todos os n-gramas acima, de cada $O_{f_{ij}}$ e O_{f_i};
  adicionar $w_1w_2w_3$ em *FinalAttributesList*;
end

Algoritmo 1: Seleção de trigramas

Após a seleção dos trigramas e eliminação de redundâncias, o mesmo processo e idéias são aplicados aos bigramas, que não foram anteriormente eliminados. Deste modo, os can-

didatos a bigramas são selecionados com o teste de independência e cálculo do estimador Q de Yule; e a seguir, *rankeados* pela sua freqüência de ocorrência.

Seja w_1w_2 um bigrama com freqüência observada f_{11}. Também é necessário calcular as freqüências observadas do unigrama w_1 na primeira posição entre todos os bigramas como $f_{1.}$ e a freqüência observada de w_2 na segunda posição entre todos os bigramas como $f_{.1}$. A freqüência de todos os bigramas na coleção, após o a eliminação de redundâncias e seleção dos n-gramas de mais alta ordem, corresponde a $f_{..}$. Então esses valores são tabulados como na Tabela 4.16.

	w_2	$!(w_2)$	
w_1	f_{11}	$f_{1.} - f_{11}$	$f_{1.}$
$!(w_1)$	$f_{.1} - f_{11}$	$(f_{..} - f_{1.}) - (f_{.1} - f_{11})$	$f_{..} - f_{1.}$
	$f_{.1}$	$f_{...} - f_{..3}$	$f_{..}$

Tabela 4.16: Freqüências de w_1 e w_2 no bigrama w_1w_2

Para testar a hipótese de independência entre a ocorrência de w_1 na primeira posição e de w_2 na segunda posição, usando o estimador Q de Yule, os seguintes valores precisam ser obtidos:

$$d = ((f_{..} - f_{1.}) - (f_{.1} - f_{11}))$$

$$\alpha = \frac{f_{11} * d}{(f_{1.} - f_{11}) * (f_{.1} - f_{11})}$$

$$Q = (\alpha - 1)/(\alpha + 1)$$

$$v = \sqrt{\frac{1}{f_{11}} + \frac{1}{f_{1.} - f_{11}} + \frac{1}{f_{.1} - f_{11}} + \frac{1}{d}}$$

$$\sigma_Q = \frac{1}{2} * (1 - Q^2) * v$$

$$IC_Q = [Q - 2 * \sigma_Q, Q_1 + 2 * \sigma_Q]$$

Se $0 \notin IC_Q$ então w_1w_2 é candidato a atributo. Todos os bigramas selecionados são *rankeados* pela sua freqüência de ocorrência f_{11}. Tão logo, todos eles tenham sido selecionados e *rankeados*, *rankeiam-se* os unigramas remanescentes do processo anterior de eliminação de redundâncias, pela sua freqüência de ocorrência. Obtém-se as freqüências O_{f_i}, $i = 1, 2$ e aplica-se o Algoritmo 2.

Input: *BiGramList*: lista dos bigramas remanescentes ordenados pelos valores de seus *ranks*
Output: *FinalAttributesList*: lista de atributos finais com freqüências atualizadas
for *para todos os bigramas em BiGramList* **do**
  selecionar o bigrama melhor *rankeado*, w_1w_2 com sua freqüência remanescente $O_{f_{11}}$;
  procurar por w_1, com $O_{f_1} \geq O_{f_{11}}$;
  procurar por w_2, com $O_{f_2} \geq O_{f_{11}}$;
  subtrair $O_{f_{11}}$ de todos os bigramas acima, isto é de O_{f_1} e O_{f_2};
  adicionar w_1w_2 em *FinalAttributesList*;
end

Algoritmo 2: Seleção de bigramas

Logo, apenas os melhores bigramas candidatos a atributos são selecionados no final do processo. Também restam apenas os unigramas que não representavam combinações lineares com os n-gramas de mais alta ordem. Deste modo, novamente a redundância serviu como um critério de seleção de atributos, ou melhor, de eliminação de atributos.

Embora o processo tenha sido aplicado apenas a trigramas, bigramas e unigramas, é fácil notar que ele pode ser expandido para qualquer n, desde que se apliquem as devidas partições às tabelas de contingência dos n-gramas.

4.4 Avaliação do método proposto

Para testar o método proposto para seleção de atributos n-gramas foram implementados alguns protótipos para realizar os testes, *ranking* e seleção dos atributos. Deste modo é possível obter as listas de candidatos a atributos e as listas finais resultantes do método proposto, para compará-los. A melhor comparação deveria ser uma análise subjetiva, dado que se tratam de atributos que também são candidatos a termos de domínio. Logo, seria interessante que um grupo de especialistas do domínio, preferencialmente, terminólogos, avaliassem os resultados. No entanto, qualquer que seja o tamanho da coleção de textos, para que se possa representar pelo menos parte de um domínio de conhecimento, é gerado um número muito grande de atributos. Logo, uma análise subjetiva seria uma tarefa nada trivial e muito onerosa.

Uma análise objetiva simples seria utilizar coleções de textos rotuladas e verificar o desempenho de um ou mais classificadores mediante diferentes conjuntos de atributos. Porém, a seleção de atributos é realizada como se a coleção fosse não rotulada, ou seja, em nenhum momento inclui-se a informação do rótulo na seleção de atributos. Assim, constroem-se dois conjuntos de atributos para cada coleção de textos considerada:

1. **atributos *rankeados***: geram-se os unigramas e os demais n-gramas. Filtram-se os unigramas e encontram-se as *stopwords* da coleção. E, então todos os n-gramas são selecionados e *rankeados* de acordo com o teste de independência anteriormente explicado.

2. **atributos não redundantes**: aplica-se o método proposto para obter o conjunto de atributos.

O teste de seleção dos n-gramas é o mesmo aplicado nos dois conjuntos de atributos, para que não se insiram outros efeitos à comparação, além da eliminação de redundâncias e *stopwords* da coleção.

Então, uma matriz atributo×valor é construída para cada conjunto de atributos; ou seja, uma matriz atributo×valor para o conjunto de dados com atributos apenas *rankeados* e outra matriz atributo×valor para o mesmo conjunto de dados porém removidas as *stopwords* da coleção e as redundâncias.

Em cada matriz atributo×valor, as células correspondem à freqüência de ocorrência do atributo no documento e, a última coluna corresponde ao rótulo do documento. Para validar os resultados, para cada conjunto de atributos, dois modelos de classificação foram construídos utilizando dois algoritmos de classificação amplamente conhecidos: C4.5 para árvores de decisão e *Support Vector Machines* (SVM). Ambos escolhidos devido ao fato de suportarem bem matrizes esparsas; especialmente, SVM comporta-se bem e encontra um modelo mesmo na presença de muitos dados redundantes (Joachims, 1998). Adicionalmente, para estimar a taxa de acurácia dos classificadores, utilizou-se o modelo de

amostragem cruzada de 10 pastas (*10-fold cross validation*). Finalmente, após obter as taxas de acurácia de cada classificador, calculam-se suas médias e comparam-nas usando-se o estimador *t Student*.

4.5 Experimentos e avaliação dos resultados

Três conjuntos de dados de diferentes domínios e línguas foram selecionados. O primeiro é uma coleção de textos de artigos completos do Instituto Fábrica do Milênio (**IFM**). Esse conjunto de dados é composto por 614 documentos em português, divididos em quatro classes, com 291 documentos na classe majoritária. A segunda coleção de documentos é o conjunto de artigos *Case Based Reasoning- Inductive Logic Programming - Information Retrieval - Sonification (CBR-ILP-IR-SON)*[2] composto por 681 resumos de documentos em inglês, classificados em quatro classes e com 276 documentos na classe majoritária, tratando-se exclusivamente de resumos de artigos científicos e referenciada neste trabalho por **CIIS**. A terceira coleção são artigos completos em inglês sobre inteligência artificial (**IA**), divididos em cinco categorias: agentes e multi-agentes, lógica *fuzzy*, aprendizado de máquina, planejamento e robótica[3].

Um processo completamente aleatório foi utilizado para balancear as coleções de textos, escolhendo cinqüenta documentos por classe. Exceto pela coleção **IFM**; nessa, três classes sofreram o processo de escolha aletória de 50 de seus documentos, porém, uma das classses tinha apenas 43 documentos e ficou com todos os originais. Deste modo, procurou-se minimizar efeitos causados pelo desbalanceamento das classes. Na Tabela 4.17 encontram-se o número de documentos em cada coleção, o número total de unigramas antes e depois do filtro. Utilizou-se o filtro TC (Liu et al., 2003), tomando-se apenas 10% dos unigramas *rankeados*, que na verdade era resultado de experimentos prévios. Nesse experimentos foi observado que filtros baseados na contribuição do termo para a variância total explicada resulta no menor conjunto de atributos mais representativo da coleção (Nogueira et al., 2008b; Liu et al., 2003) e alguns outros experimentos sugerem que somente 10% dos atributos *rankeados* são tão confiáveis quanto maiores conjuntos de atributos (Nogueira et al., 2008a). Na mesma tabela, as duas últimas colunas referem-se ao número de bigramas e trigramas após o filtro dos unigramas e a remoção de *stopwords* da coleção.

Coleção de Textos	**# de docs**	**# de unigramas**	**# de unigramas pós-filtro**	**# de bigramas**	**# de trigramas**
IA	250	42753	4276	87297	30326
CIIS	200	10810	1081	17940	17944
IFM	193	20795	2080	82325	84200

Tabela 4.17: Resultados dos primeiro, segundo, terceiro e quarto passos do modelo proposto sobre os dados experimentais

Assim, as três últimas colunas da Tabela 4.17 correspondem aos primeiros conjuntos de candidatos a atributos. Para encontrar o conjunto não redundante aplica-se o quinto passo do modelo proposto. Deste modo, para cada coleção tem-se os dois conjuntos de atributos

[2]Disponível em http://infoserver.lcad.icmc.usp.br/infovis2/PEx.
[3]Disponível em http://labic.icmc.usp.br/projects.

só *rankeados* (R) e não redundantes (NR). O número de atributos não redundantes (NR) é resumido na Tabela 4.18 e os *rankeados* na Tabela 4.19.

Coleções de textos	**# unigramas NR**	**# bigramas NR**	**# trigramas NR**
IA	4264	26270	24303
CIIS	1065	1379	1831
IFM	1984	7043	42741

Tabela 4.18: Atributos n-gramas não redundantes obtidos dos dados experimentais

Coleções de textos	**# unigramas R**	**# bigramas R**	**# trigramas R**
IA	4276	53595	30060
CIIS	1081	4284	2811
IFM	2080	48775	82354

Tabela 4.19: Atributos n-gramas apenas *rankeados* obtidos dos dados experimentais

Para cada coleção e conjunto de atributos, os classificadores C4.5 e SVM foram aplicados e suas taxas de acurácia foram estimadas usando validação *10-fold cross*. É muito importante ressaltar que os classificadores não estão sendo comparados, o foco é observar o comportamento de mais de um classificador com diferentes conjuntos de atributos.

Coleções de textos	*Rankeados* $acc \pm se$	Não Red. $acc \pm se$	teste t $p - value$
IA	90 ± 5.42	89.6 ± 7.11	0.3838
CIIS	87 ± 9.19	87 ± 9.19	0.3880
IFM	70.45 ± 8.28	75.63 ± 7.88	0.1385

Tabela 4.20: Resultados do classificador C4.5 para testar $H_0 : \overline{acc_r} = \overline{acc_{nr}}$

Coleções de textos	*Rankeados* $acc \pm se$	Não Red. $acc \pm se$	teste t $p - value$
IA	$79.6 \pm 8,1$	82 ± 7.12	0.2968
CIIS	90 ± 8.82	90.5 ± 6.85	0.3837
IFM	59.97 ± 10.24	68.84 ± 11.74	0.0835

Tabela 4.21: Resultados do classificador SVM para testar $H_0 : \overline{acc_r} = \overline{acc_{nr}}$

Nas Tabelas 4.20 e 4.21, notam-se os resultados das avaliações das acurácias e seus respectivos desvios padrões para cada grupo de atributos, coleção e classificador; os grupos são *rankeados* (R) e não redundantes (NR). Na última coluna estão só resultados do teste *t Student* para a hipótese de igualdade das médias das acurácias, isto é, $H_0 : \overline{acc_r} = \overline{acc_{nr}}$). Tem-se que $\overline{acc_r}$ e $(se)_r$ correspondem respectivamente ap valor médio e ao desvio padrão de cada acurácia, mediante o *10-fold cross validation*, para o conjunto de *rankeados*. E, que $\overline{acc_{nr}}$ e $(se)_{nr}$ correspondem respectivamente ao valor médio e ao desvio padrão de cada acurácia, mediante o *10-fold cross validation*, para o conjunto de não redundantes. A Equação 4.1 foi utilizada na obtenção de cada valor de t e então calculada a probabilidade acumulada desse valor sob uma distribuição t-Student. Se a probabilidade acumulada ficar no intervalo $[0.025, 0.975]$, a hipótese de igualdade das acurácias é aceita. Esse intervalo corresponde a valores com nove graus de liberdade. Logo, o valor de p-value na última coluna corresponde ao complemento dessa probabilidade em relação a um. Como todos

os valores obtidos são maiores que 0.025 então a hipótese de igualdade é aceita em todos os casos.

$$t_{value} = \frac{\overline{acc_r} - \overline{acc_{nr}}}{\sqrt{\frac{(se)_r^2}{10} + \frac{(se)nr^2}{10}}} \tag{4.1}$$

Num primeiro momento, focando apenas na performance de cada classificador, parece que retirar a redundância foi melhor, especialmente devido ao intervalo de valores dos desvios padrões. Mas, ao aplicar o teste t vê-se que estatisticamente as acurácias são iguais a um nível de significância de 95%. Deste modo, podemos observar que a remoção da redundância não sacrificou a efetividade da categorização, como esperado.

4.6 Vantagens do método para a seleção de atributos n-gramas

Como o método proposto não depende do conhecimento prévio de rótulos, pode ser utilizado em aprendizado supervisionado ou não supervisionado. No entanto, para aprendizado supervisionado seria interessante trocar o processo de seleção dos unigramas, no primeiro filtro do método. Para ter uma melhor seleção dos unigramas, poder-se-ia acrescentar o conhecimento dos rótulos.

O método proposto depende somente dos testes estatísticos sobre uma coleção desordenada de palavras (ou radicais, ou lemas, etc) n-gramas em lugar de caracteres n-gramas contíguos, como nos métodos de Dejun e Maosong (2004). Além disso, Dejun e Maosong (2004) acreditavam que seus métodos eram independentes de língua, e o método aqui proposto o é para a remoção de redundâncias. Deve-se lembrar que a identificação dos n-gramas depende da língua, após esse processo, a remoção de redundâncias é independente. Como evidência dessa afirmação pode-se considerar as coleções de textos utilizadas no experimento, pois uma está em português e as duas outras em inglês.

Além disso, o método proposto define e utiliza *stopwords* da coleção de textos para remover combinações indesejadas de n-gramas. Essa remoção de *stopwords* tanto funciona como um filtro para que não ocorram n-grams formados pelos unigramas cortados, quanto substitui janelas de formação de n-gramas. Por exemplo, no trabalho de Zhang e Zhu (2007) há uma proposta de estimação de tamanhos de janelas entre os gramas, de modo que sejam eliminadas algumas combinações indesejadas. Com a remoção das stopwords da coleção acredita-se que o método proposto se aproximadas idéias de Zhang e Zhu (2007), causando uma melhora da informação semântica similar à estimativa de janelas em um processo muito mais simples.

4.7 Considerações Finais

Um método para resolver o problema de automaticamente selecionar bons atributos n-gramas não redundantes é proposto. O método baseia-se no pressuposto que a redundância é sempre presente quando se usam n-gramas, $n \geq 2$, devido a sua natureza estocástica.

O método proposto alcançou seus objetivos, selecionando bons candidatos a atributos n-gramas e eliminando as redundâncias apresentadas. Nos experimentos foram obtidas

boas estimativas de acurácia, o que evidencia uma boa qualidade dos atributos. Além disso, a efetividade dos classificadores não foi diminuída após a remoção das redundâncias entre os atributos, o que indica que apenas as redundâncias foram realmente eliminadas.

Como trabalho futuro, deve ser realizada uma análise subjetiva sobre os atributos gerados, junto a especialistas do domínio, para tentar avaliar melhorias semânticas na interpretação. Também seria interessante comparar os resultados a indicação de janelas entre os n-gramas gerados, como feito por Zhang e Zhu (2007).

Por fim, é necessário expandir os algoritmos e implementações para n qualquer, dado que o modelo apresentado limitou-se a $n \leq 3$.

No Capítulo 5, a seguir, trata-se de um quesito da etapa de extração de conhecimento da TopTax. O assunto é a seleção de atributos como rótulos discriminativos de agrupamentos hierárquicos de documentos, ao quê dá-se o nome de rotulação de agrupamentos. Para a rotulação de agrupamentos hierárquicos de documentos foram propostos dois modelos inéditos, bem como a metodologia de avaliação dos resultados desses modelos.

Capítulo

5

Modelos de rotulação em agrupamentos hierárquicos de documentos

Um novo método de rotulação para agrupamentos hierárquicos de documentos e uma proposta de validação do método são apresentados neste capítulo. O objetivo do método é produzir rótulos discriminativos dos grupos, sem repetição ao longo dos ramos, com um vocabulário variado e de qualidade e, independente do algoritmo de agrupamento utilizado, bem como independente da língua. Para isso o método foi baseado em uma definição formal de uma taxonomia de tópicos, a partir das definições de seus elementos. O método foi implementado e validado contra três outros métodos da literatura. A validação passa pelos critérios de complexidade do algoritmo, melhor discriminação dos grupos, melhor variabilidade e qualidade do vocabulário gerado; em todos esses quesitos o novo método supera estatisticamente os outros três.

5.1 Introdução

Segundo Konchady (2006), um número considerável de propostas explora o agrupamento aglomerativo de documentos sob a perspectiva da recuperação de informação, mas um aspecto igualmente importante é o processo de rotular os grupos obtidos. O usuário final pode decidir explorar um agrupamento geralmente guiado pelo rótulo do mesmo; porém, resumir um agrupamento com base em um pequeno conjunto de termos é bastante difícil. Pois, a questão de abstração de dados em problemas de agrupamento envolve obter descrições concisas e significativas para os grupos e essas descrições chegam a ser tão importante quanto o próprio agrupamento (Feldman e Sanger, 2007).

Há vários trabalhos que exploram especialmente a geração de rótulos para o agrupamento de documentos textuais, que visam à identificação de palavras-chaves para indicar os possíveis tópicos aos quais os documentos agrupados se refiram. Em geral, esses métodos dependem diretamente da forma como o agrupamento é obtido, isto é, do método

utilizado para agrupar os documentos. Os agrupamentos são calculados a partir da representação matricial da coleção de textos, considerando-se cada texto um elemento a ser agrupado representado por um vetor de atributos. Cada atributo corresponde a uma palavra ou composição de palavras (por exemplo, "inteligência artificial"), para as quais é obtida uma medida de freqüência, que pode ser qualitativa, representando a pertinência ou não da palavra no documento, ou quantitativa, como a freqüência ou a freqüência inversa de ocorrência da palavra. Os agrupamentos obtidos refletem tópicos ou sub-tópicos aos quais os documentos se referem, logo, são representados por conjuntos de atributos mais significativos no grupo, que podem ser interpretados como um conceito associado ao grupo.

Um dos principais pontos de investigação está relacionado a como rotular os tópicos das coleções de textos hierarquicamente organizadas. A organização hierárquica de uma coleção de textos seguida de um procedimento de rotulação tem como objetivo organizar a informação, atribuir metadados únicos a cada grupo da hierarquia e auxiliar o usuário em navegação e busca, de modo que o usuário possa melhor compreender a coleção de dados. Assim, no escopo do projeto TopTax- *Topic Taxonomy* (Moura et al., 2008a), descrito no Capítulo 3, a rotulação de grupos foi separada da obtenção dos mesmos. Desta forma, pode-se avaliar diferentes métodos de obtenção da hierarquia e de rotulação separadamente, procurando não perder informação ou compreensibilidade. Além disso, com o processo separado da rotulação, a técnica pode ser aplicada a qualquer hierarquia, tenha ela sido obtida de qualquer procedimento de agrupamento hierárquico, inclusive manual. Neste capítulo trata-se exclusivamente o problema de rotulação dos grupos.

Na próxima seção é discutida a revisão bibliográfica realizada sobre métodos de rotulação de agrupamentos que se aproximam dos requisitos do projeto TopTax. A seguir, descrevem-se as idéias que inspiraram os dois modelos propostos para rotulação de agrupamento hierárquico; para ilustrar essas idéias utiliza-se um exemplo propositadamente criado. Depois os modelos, algoritmos e características são descritos. Na Seção 5.4 discute-se uma forma de avaliação objetiva do modelo proposto comparando-o a outros pré-existentes. Na Seção 5.7 discorre-se sobre os experimentos realizados, apresentando dados e resultados. E, finalmente as considerações finais.

5.2 Revisão bibliográfica de modelos de rotulação

Obtidos os agrupamentos, a tarefa de identificar as palavras mais discriminativas em cada qual, também conhecida como rotulação de agrupamentos, pode ser vista como um problema de seleção de atributos (Weiss et al., 2005). Para seleção supervisionada, considera-se cada grupo como uma classe, podendo-se utilizar informação mútua ou ganho de informação; e, para abordagens não supervisionadas utiliza-se a *tf-idf (term frequency - inverse document frequency)* ou alguma análise fatorial. Um bom exemplo de seleção não supervisionada é a utilização da *tf-idf* no módulo *k-means* do ambiente TMSK - *Text Mining Software Kit* (Weiss et al., 2005). A *tf-idf* corresponde à freqüência observada da palavra na coleção, ou no agrupamento como utilizada nesse caso, ponderada pelo logaritmo da sua freqüência inversa na coleção; a ponderação visa eliminar palavras muito

freqüentes e privilegiar as menos freqüentes. No ambiente TMSK, para selecionar as palavras mais discriminativas de cada agrupamento encontrado, o método empregado considera aquelas que possuem as maiores *tf-idfs* médias.

Dentre as idéias mais difundidas e utilizadas para rotular os grupos de documentos estão os métodos baseados em centróides, para agrupamentos obtidos com *k-means* ou *k-medóide*, ou método deles derivado. Esses métodos geram agrupamentos disjuntos e em alguns casos hierárquicos - aplicando-se métodos semelhantes ao *bissecting k-means* - que divide o conjunto de dados em duas partes, depois cada parte em duas partes e, assim, sucessivamente. Para obter o conjunto de palavras-chaves que melhor discriminem os grupos, em cada grupo calculam-se os centróides, para a base *k-means*, ou as medianas, para os similares ao *k-medóide*; e, então se utilizam os valores de cada célula dos vetores como pesos para cada atributo considerado. Os atributos com maior peso em cada grupo são considerados os mais discriminativos do agrupamento (Larsen e Aone, 1999). Recentemente, Zhang et al. (2009) propuseram um método simples cuja base é o vetor central de um agrupamento, tomando-lhe as cinco palavras de maior peso, seguido de uma escolha heurística das melhores palavras; o problema é que a heurística é um tanto quanto subjetiva. A validação dessa proposta é realizada sobre um agrupamento manualmente anotado e usando diferentes classificadores.

Trabalhando-se com modelos probabilísticos de agrupamentos, há o modelo CAM - *Cluster Abstraction Model*, desenvolvido por Hofmann (1999). Nesse modelo restringe-se a presença do termo no nó da árvore como ascendente ao agrupamento no qual o termo é significativo no documento somado à distribuição das palavras condicionadas aos documentos, $P(w/d) = \sum_c P(c_d = c/w_d, \theta) \sum_a P(a/c)P(w/a)$, com w correspondendo a cada palavra, d a cada documento e c a uma classe/grupo. Nesse caso, se for considerado *hard assigment* para a distribuição dos grupos dadas as palavras, captura-se a idéia de propagação de termos em uma taxonomia, em que se verifica que um documento é composto por uma mistura de termos que vão dos mais genéricos para os mais específicos. O resultado é muito bom, embora a forma de se chegar ao mesmo seja muito complexa e envolva a estimação dos critérios de convergência dos métodos empregados. Em consonância com este trabalho, as abstrações probabilísticas de hierarquias de Segal et al. (2001) utilizaram um modelo com classes específicas e probabilísticas para cada nó. Nesse trabalho, a divergência de Kullback-Leibler (Duda et al., 2001) define a distância entre as categorias; permitindo minimizar a distância total e maximizar a verossimilhança, complementando o CAM em relação a encontrar um melhor agrupamento.

Uma solução não probabilística e mais simples é a utilizada no ambiente TaxaMiner (Kashyap et al., 2005a), que busca identificar taxonomias de tópicos a partir de *clusterização* hierárquica; em que os documentos são agrupados mediante o cálculo de um fator de coesão dos grupos e a taxa de coesão de cada grupo indica também os atributos mais significativos em cada qual. Nesse método é realizado um procedimento de *pruning* da árvore de *cluster* com base na idéia de propagação de termos em uma taxonomia. A base do procedimento de *pruning* é que níveis mais específicos da árvore que nada acrescentam aos seus antecedentes podem ser podados; para tomar essa decisão o método utiliza a medida de coesão, nele definida, para maximizar a relação *intracluster* e também estabelece

empiricamente um *threshold*.

Outra solução recente, independente de método de agrupamento e de linguagem, foi proposta por Ienco e Meo (2008). Eles usam o algoritmo *PageRank* do Google (Mihalcea, 2005), sobre o grafo obtido após aplicar um algoritmo hierárquico de agrupamento a uma coleção de documentos. O algoritmo *PageRank* é baseado em passeios aleatórios através do grafo para determinar a importância dos vértices; é uma algoritmo iterativo, usando as probabilidades de transição entre os vértices do grafo para produzir pesos e ranquear as palavras-chaves. Dessa forma a seleção dos descritores, ou seja, dos rótulos dos grupos, é induzida pelo PageRank.

Geralmente, quando estamos interessados na tarefa de rotular os agrupamentos de forma independente de como o *cluster* foi obtido, usa-se a distribuição de freqüências das palavras ao longo dos grupos. Glover et al. (2002) propuseram um modelo baseado exclusivamente nas freqüências observadas para cada palavra em cada grupo, isto é, nas estimativas de máxima verossimilhança de suas probabilidades: $p(w/c)$ e $p(w)$, com w correspondente a palavra e c ao grupo (considerado como uma classe). A hipótese assumida é que, se $p(w/c)$ é muito comum e $p(w)$ é raro, então a palavra é um bom discriminador para a classe c; caso contrário, se $p(w/c)$ e $p(w)$ são comuns então a palavra discrimina a *classe ascendente de* c e, finalmente, se $p(w/c)$ é muito comum e $p(w)$ é relativamente rara na coleção, então a palavra é um melhor discriminador para a *classe descendente de* c. Os valores de *muito comum* e *raro* são experimentalmente determinados. Foi proposta uma modificação a este método, na qual foi estabelecido um compromisso entre uma única palavra como rótulo e uma lista de palavras como rótulo, estabelecendo-se uma pontuação descritiva ponderada pela *tf-idf* (Treeratpituk e Callan, 2006). Embora os resultados sejam bons, o problema de determinação experimental dos critérios de convergência foram entendidos também aos novos pontos de corte dessas pontuações descritivas.

A abordagem de Lamirel et al. (2008) é uma combinação da construção de uma árvore hiperbólica e técnicas de visualização com técnicas de rotulação de agrupamentos; porém a forma de atribuição dos rótulos não depende do algoritmo de agrupamento, logo, a solução pode ser utilizada inclusive para agrupamento não hierárquico. O método de rotulação é baseado no critério de avaliação do conteúdo dos grupos por meio da *medida* F. Os autores redefiniram as medidas *recall* (cobertura), *precision* (precisão) e F, usadas em recuperação de informação, para refletir a qualidade dos conjuntos de rótulos. Para isso, baseados na distribuição de freqüências de todos os termos na coleção de textos e em cada grupo, alguns pesos relacionados às freqüências dos termos foram considerados. Deste modo, a precisão e a cobertura em cada grupo podem ser vistas como probabilidades condicionais, da ocorrência do termo dado o grupo. Adicionalmente, eles propuseram uma medida de penalidade e de ganho para cada escolha dependendo se discriminavam simultaneamente cada nó pai e filho. Com essas pontuações, as escolhas foram boas e o método teve um resultado significativamente melhor que aqueles com base apenas nas maiores freqüências e em seleção por X^2 (qui-quadrado).

Um método pouco mais antigo, que trabalha sobre uma distribuição multinomial em um agrupamento hierárquico foi desenvolvido por Popescul e Ungar (2000). Nessa proposta, um atributo é discriminativo ou não em um dado nó da hierarquia se ele depende

estatisticamente do nó. A seleção é realizada da raiz para a folhas da árvore, testando-se cada palavra em relação a sua dependência dos filhos do nó. Se a hipótese de independência é aceita, então a palavra tem distribuição homogênea nos filhos e está associada apenas ao nó pai, caso contrário, a palavra deve estar associada a algum dos filhos ou seus descendentes. Um problema nesse método é garantir a convergência das diferenças entre os valores observados e esperados para o estimador X^2, usado no teste de hipótese, quando os valores esperados ou observados são muito pequenos - próximos a zero. De acordo com o critério estabelecido para os valores mínimos aceitáveis e os intervalos de confiança das hipóteses validadas, os resultados obtidos podem não ser confiáveis. Mas, o método tem soluções interessantes e poderia ser mais bem explorado para evitar os riscos inerentes ao processo de decisão.

Primeiramente, foi proposta uma modificação ao método de Popescul & Ungar na qual alguns problemas foram solucionados (Moura et al., 2008c). As modificações incluem o tratamento de baixas freqüências de ocorrência observadas e esperadas, isto é, quando próximas a zero, e a troca do estimador por um mais robusto, não dependente da distribuição das diferenças entre valores esperados e observados. Com isso, os algoritmos sempre conseguem tomar uma decisão, produzindo conjuntos de rótulos pequenos e sem repetição ao longo da hierarquia. Embora essas modificações tenham tido uma boa avaliação, os algoritmos do método têm alta complexidade computacional e existe uma alta taxa de propagação de erros nos resultados. Em se tratando de método estatístico o erro é previsto, mas poderia ser evitado com uma reespecificação dos algoritmos. Esse método é descrito na Seção 5.3.3, RLDM - *Robust Labeling Down Method*.

Assim, foi proposto um novo método com base nas idéias dos dois últimos. Esse método também é independente da língua e produz conjuntos de rótulos pequenos e sem repetição ao longo da hierarquia. Comparando-o aos dois últimos, ele tem uma complexidade computacional muito menor, diminui a propagação de erro, produz um vocabulário de qualidade superior e mais variado que os demais; conseqüentemente, melhora a representatividade do vocabulário. Esse método é descrito na Seção 5.3.4, RLUM - *Robust Labeling Up Method*.

5.3 Métodos desenvolvidos para rotulação hierárquica de agrupamentos

Esta seção tem início com a idéia geral empregada aos métodos de rotulação desenvolvidos neste trabalho, bem como seu embasamento na literatura. Os algoritmos pré-existentes e os novos, desenvolvidos neste trabalho, são apresentados na seqüência, bem como o método utilizado para a validação objetiva.

5.3.1 Idéia geral de rotulação de agrupamentos

Para ilustrar as idéias aqui apresentadas, será utilizada uma coleção de seis textos em inglês. Os textos foram proposital e cuidadosamente construídos para cobrir tópicos relacionados à piscicultura (*pisciculture*), com duas espécies de peixes do Pantanal, e

tópicos sobre produção e reprodução de gado de corte (*beef catle*). Os textos foram criados para ter em comum o termo *Pantanal* e apenas termos do domínio foram utilizados nos textos; isto é, termos que foram obtidos de um *thesaurus*. Adicionalmente, procurou-se distribuir as freqüências dos termos mais ou menos uniformemente ao longo dos textos. Conseqüentemente, essa coleção de textos é um exemplo ideal, pois após obter a estrutura hierárquica deve-se chegar a uma hierarquia de tópicos basicamente sem erros; pois, os termos selecionados supostamente são os melhores, dado que são termos de domínio e estão balanceadamente distribuídos ao longo dos textos.

Após aplicar um método de agrupamento hierárquico aglomerativo, usando similaridade de cosseno e o algoritmo *average linkage*, espera-se que os textos sejam separados por tópicos e sub-tópicos de acordo com os respectivos assuntos e todos agrupados sob um tópico raiz que contém o termo *Pantanal* em seus descritores; como ilustrado na Figura 5.1. Nessa figura, para todos os nós é representada a freqüência acumulada de cada termo em cada grupo de documentos.

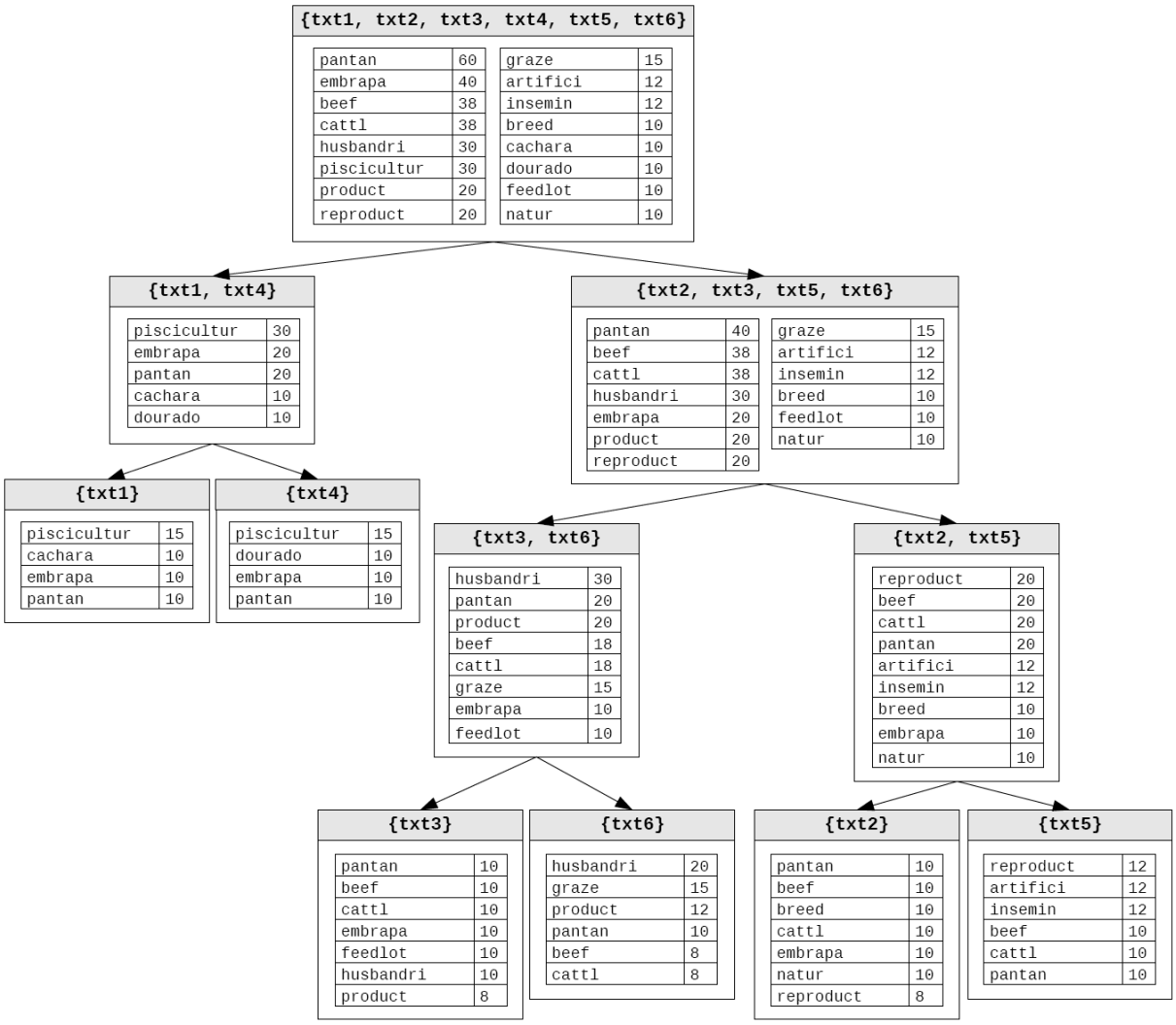

Figura 5.1: Hierarquia inferida sobre a coleção de textos do exemplo ideal

Dada a hierarquia da Figura 5.1, se os rótulos de cada grupo forem obtidos com o método de seleção dos termos mais freqüentes em cada grupo, os termos com maior freqüência acumulada nos agrupamentos são mantidos como rótulos após a determinação de quantos termos deverão compor cada conjunto de rótulos. Assim, usando, no máxi-

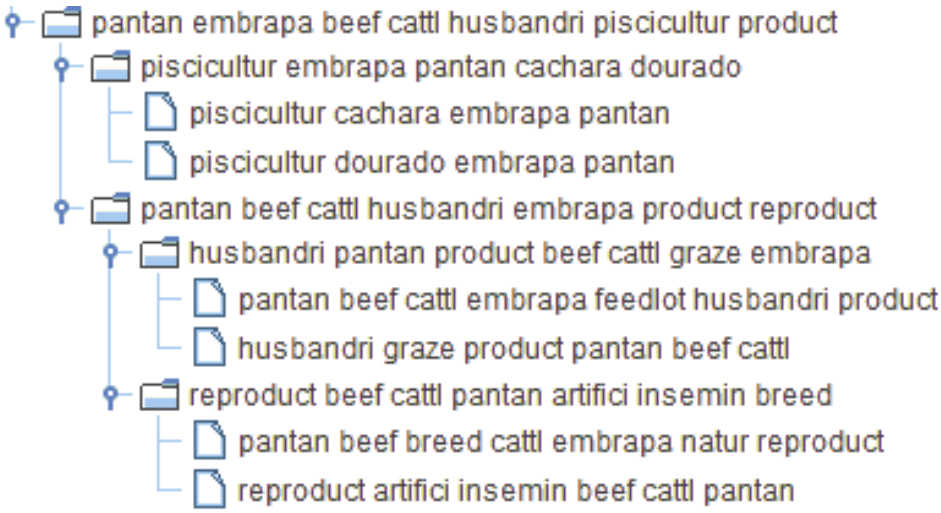

Figura 5.2: Resultado da aplicação do método de rotulação Mais Freqüentes sobre a coleção de textos do exemplo

mo, sete dos termos em cada conjunto, tem-se o resultado apresentado na Figura 5.2. Esse método geralmente é chamado de **Mais Freqüentes**, denominação seguida neste trabalho. Deve-se ressaltar que esse método é bastante usado em agrupamentos não hierárquicos e disjuntos (*flat clusters*), para os quais apresenta bons resultados. Porém, em uma hierarquia esse método introduz uma série de repetições desnecessárias de termos ao longo dos ramos; dado que os rótulos dos nós ascendentes também se aplicam aos seus descendentes, de acordo com as idéias de propagação de termos em taxonomias. Embora, tenha sido utilizado em alguns trabalhos junto a agrupamentos hierárquicos, como no TaxGen (Müller et al., 1999). No TaxGen foram utilizados no máximo sete níveis de hierarquia e os cinco termos mais freqüentes como rótulos em cada grupo.

Finalmente, uma **taxonomia de tópicos** é definida, neste trabalho, como uma coleção de um domínio de conhecimento específico, agrupada de acordo com algum critério de similaridade entre os documentos e, tendo cada grupo associado a um conjunto de termos, que é o **rótulo** de cada agrupamento. A união de todos os conjuntos de termos forma uma estimativa do **vocabulário** do domínio no qual a coleção se insere. A primeira estimativa do vocabulário é o conjunto inicial de **atributos**, após a aplicação de algum filtro.

5.3.2 Seleção hierárquica de rótulos

Com o exemplo anterior, pode-se observar que, o método Mais Freqüentes não executa uma seleção adequada de rótulos para agrupamentos hierárquicos; uma boa idéia é procurar não repetir os termos dos rótulos ao longo dos mesmos ramos da hierarquia. Para explicar as idéias e definir os métodos desenvolvidos será utilizada a hierarquia genérica representada na Figura 5.3 e utilizadas as notações descritas a seguir. Supõe-se que a hierarquia foi obtida de um método de agrupamento hierárquico sobre uma coleção de documentos, $D = \{d_1, ..., d_x\}$, com x documentos, representados por medidas sobre um conjunto de atributos, $A = \{a_1, ..a_m\}$, com m atributos. Assim, cada nó n_i na hierarquia corresponde a um grupo com c filhos e é formado por uma subcoleção de D. Deste modo, $f_i(a_k)$ é definida como a freqüência acumulada do k-ésimo atributo de A em n_i; ou simi-

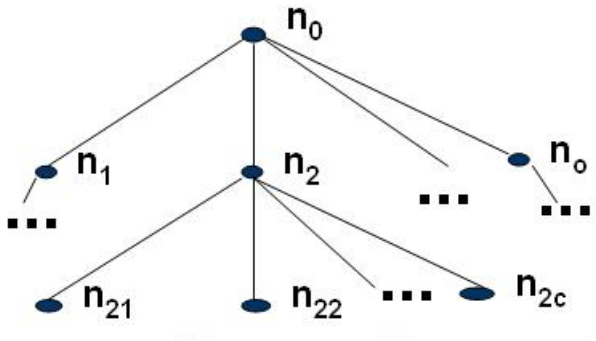

Figura 5.3: Hierarquia genérica de uma taxonomia

larmente, essa quantidade corresponde à freqüência acumulada de a_k nos documentos de n_i. Para tomar uma decisão sobre quais atributos melhor discriminam o i-ésimo grupo, para cada a_k, $k = 1, ..m$, em um determinado n_i, constrói-se uma tabela de contingência das freqüências dos termos em cada filho de n_i, ilustrada na Tabela 5.1.

$filho$	a_k	$!a_k$	total
n_{i1}	$f_{i1}(a_k)$	$\sum_{t=1,t\neq k}^{m} f_{i1}(a_t)$	$\sum_{t=1}^{m} f_{i1}(a_t)$
...	...	...	...
n_{ic}	$f_{ic}(a_k)$	$\sum_{t=1,t\neq k}^{m} f_{ic}(a_t)$	$\sum_{t=1}^{m} f_{ic}(a_t)$
$total$	$f_i(a_k)$	$\sum_{t=1,t\neq k}^{m} f_i(a_t)$	$\sum_{t=1}^{m} f_i(a_t)$

Tabela 5.1: Tabela de contingência para $f_i(a_k)$

Para decidir se o termo a_k discrimina somente o nó ascendente, nó pai n_i, ou somente um dos filhos ou todo o conjunto de filhos, $\{n_{i1}, .., n_{ic}\}$, um teste de independência é aplicado sobre a distribuição do termo nos filhos. Sob a hipótese de independência, isto é, o termo fixado a_k não depende dos nós filhos, cada $f_{ij}(a_k)$ depende exclusivamente das freqüências marginais, isto é, o valor esperado para cada $f_{ij}(a_k)$ deve ser:

$$E(f_{ij}(a_k)) = f_i(a_k) \times \sum_{t=1}^{m} f_{ij}(a_t) \times \frac{1}{\sum_{t=1}^{m} f_i(a_t))}$$

Se o termo independe de todos os nós filhos, considera-se que ele pertence ao conjunto de rótulos genéricos do nó pai n_i, caso contrário, ele é considerado como pertencente aos conjuntos de rótulos específicos de cada nó filho com o qual a relação de dependência tenha sido verificada. Esse teste é aplicado ao longo de toda a hierarquia, para cada atributo, iniciando-se na raiz e descendo-se até as folhas.

O método de Moura et al. (2008c), a partir daqui tratado por RLDM - ***Robust Labeling Down Method***, e o de Popescul e Ungar (2000) implementam essas idéias, tendo como principais diferenças os testes utilizados para verificar a hipótese de independência e o tratamento das restrições de aplicação dos testes. Popescul e Ungar usaram o estimador χ^2 com a restrição de todos $E(f_{ij}(a_k)) \geq 5$ bem como todas as freqüências observadas na tabela ≥ 5. Como em mineração de textos as matrizes atributo$\times$valor são muito esparsas, também após o agrupamento muitos dos atributos apresentam freqüências baixas. Assim, com uma restrição como $E(f_{ij}(a_k)) \geq 5$ ou $f_{ij}(a_k) geq 5$, esse método acaba se aproximando muito dos resultados obtidos com o método dos Mais Freqüentes, comentado e ilustrado anteriormente.

Quanto ao uso do chi-quadrado e de medidas de associação, verificou-se que haviam recomendações diferentes. Por exemplo, segundo Kachigan (1986) a regra de considerar o valor cinco como limite inferior para as freqüências esperadas e observadas é muito conservadora. Fienberg (1985) e Bishop et al. (1975) consideram que se o tamanho total da amostra é grande, nesse caso as caselas de total $\sum_{t=1}^{m} f_i(a_t)$, pode-se utilizar a estimativa de qui-quadrado sem fatores de correção de continuidade. Nesse caso, considera-se a regra geral de freqüências e valores esperados ≥ 1, ou seja, basta que a freqüência do termo não seja zerada em cada casela. Fienberg (1985) e Bishop et al. (1975) também afirmam que, nesses casos, pode-se considerar medidas de associação invariantes para as diferentes combinações de linhas e colunas, como a *cross-product ratio*. Assim, no RLDM decidiu-se pelo uso da *cross-product ratio* junto a medidas de associação total.

5.3.3 O método *Robust Labeling Down* - RLDM

Nesse método, casos nos quais algum $f_{ij}(a_k) \approx 0$ são tratados separadamente, como casos de associação ou dissociação completa do $k-$ésimo termo a algum grupo. Deste modo o método permite que seu algoritmo sempre tome uma decisão em relação a algum a_k em algum n_i e, conseqüentemente, os algoritmos sempre produzem um conjunto pequeno de rótulos em cada grupo e não replicam os termos ao longo da hierarquia; como pode ser observado na Figura 5.4. Adicionalmente, são utilizados os estimadores Q *de* $Yule$ e o U^2 para testar as hipóteses de independência (para maiores detalhes veja Bishop et al. (1975)).

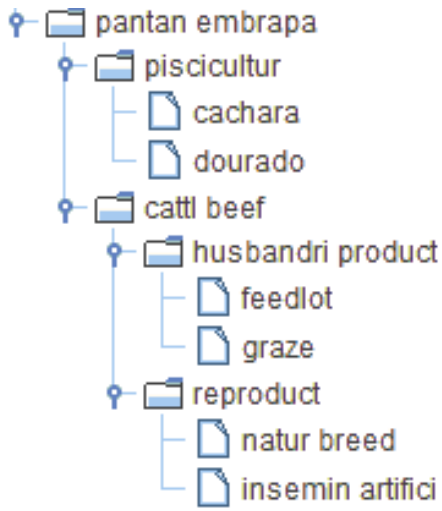

Figura 5.4: Ilustração do método de rotulação RLDM

Assim, em tabelas 2×2, isto é, quando algum n_i tem somente dois filhos, $c = 1, 2$; o estimador escolhido é o Q *de* $Yule$. Para testar a hipótese de associação usando o Q *de* $Yule$, a razão do produto cruzado é calculada:

$$\alpha = (f_{i1}(a_k) * \sum_{t=1, t \neq k}^{m} f_{i2}(a_t)) \ / \ (f_{i2}(a_k) * \sum_{t=1, t \neq k}^{m} f_{i1}(a_t))$$

E, então, estima-se Q (Bishop et al., 1975)[1]:

$$\widehat{Q} = \frac{\alpha - 1}{\alpha + 1},$$

$$\widehat{\sigma}_Q = \frac{1}{2} * (1 - \widehat{Q}^2) * \sqrt{\frac{1}{f_{i1}(a_k)} + \frac{1}{\sum_{t=1,t\neq k}^{m} f_{i2}(a_t)} + \frac{1}{f_{i2}(a_k)} + \frac{1}{\sum_{t=1,t\neq k}^{m} f_{i1}(a_t)}}$$

$$\widehat{Q} \approx N(\widehat{Q}, \widehat{\sigma}_Q) \Rightarrow \widehat{Q} \pm 2 * \widehat{\sigma}_Q$$

O valor máximo da função é alcançado quando $\widehat{\alpha} = 1$ e $\widehat{Q} = 0$ e, $\widehat{Q} = 1$ ou $\widehat{Q} = -1$ ocorrem quando algum $f_{ij} = 0$; assim, se o valor $0 \in \left[\widehat{Q} - 2 * \widehat{\sigma}, \widehat{Q} + 2 * \widehat{\sigma}\right]$ então a hipótese de independência ou desassociação é verdadeira.

A expansão do teste mais robusto para tabelas de contingência $c \times 2$, isto é, n_i pode ter qualquer número de filhos, $(c \geq 3)$, usa o estimador U^2:

$$U^2 = \frac{(c-1) * BSS}{TSS},$$

$$BSS = TSS - WSS$$

$$TSS = (c/2) - (1/2c) \sum_{j=1}^{c} (f_{ij}(a_k))^2$$

$$WSS = (c/2) - (1/2) \left(f_i(a_k) + \sum_{t=1,t\neq k}^{m} f_i(a_t) \right) \times \sum_{j=1}^{c} (f_{ij}(a_k))^2$$

O valor TSS é interpretado como a variância total na tabela, ou a dispersão total entre os valores. O valor WSS é a variação dos filhos dentro da classe, sendo que a classe é positiva quando o termo está presente no nó filho. O valor BSS é a variância entre as classes. Assim, U^2 é um estimador da redução da proporção da variância explicada pelos dados, isto é, a variação da freqüência do termo e, assintoticamente, aproxima-se de uma distribuição χ^2 com $(c-1)$ graus de liberdade (veja Bishop et al. (1975) para detalhes), embora não dependa da probabilidade de distribuição das freqüências.

Algoritmos para o RLDM

Os Algoritmos 3 e 4 implementam a idéia básica de percorrer a árvore da raiz para as folhas, para cada atributo a_k, $k = 1, ..., m$, realizando os testes de hipótese e decidindo para quais nós, ou grupos, o atributo é ou não discriminativo. Deve-se notar que com esses passos sendo seguidos no processo decisório, não sobram replicações de termos ao longo de um mesmo ramo da hierarquia.

Ainda, como para cada atributo sempre existe uma decisão, em alguns casos, o conjunto de rótulos de um grupo resulta vazio. Nesses casos, não foi encontrado um atributo especificamente discriminativo para o grupo, isto é, ele apenas herda os rótulos discriminativos de seus antecessores. Isso ocorre porque a coleção de textos é limitada, bem como o conjunto de atributos. Logo, nem sempre há atributos que representam bem cada grupo, independentemente. Então, esses grupos são cortados da hierarquia e, os descendentes

[1] Toda estimativa é notada com um chapéu.

desses grupos passam a ser parte de seu primeiro antecessor rotulado. Para realizar esse corte semântico da hierarquia, após a aplicação dos algoritmos anteriores a toda a árvore, aplica-se o Algoritmo 5.

```
Input: n_i, o nó raiz precisa ser o primeiro
Output: hierarquia atualizada, com seus conjuntos de rótulos
for all a_k, k=1,..,m do
 | h executa testeRLDM(n_i,a_k);
end
execute CombinaSimplifica(n_i);
```

Algoritmo 3: RLDM: controle do conjunto de atributos

```
Input: n_i: i-ésimo grupo; a_k: k-ésimo atributo
for all filhos de n_i do
 | verifica quantos filhos possuem a_k;
end
if quantos filhos == 1 then
 | executa testeRLDM(esse filho,a_k);
end
if quantos filhos == 2 then
 | calcula Q de Yule e testa a hipótese;
 | if independe dos filhos then
 | | fica no pai; remove dos filhos;
 | end
 | else
 | | if depende de um filho then
 | | | fica nesse filho; remove do pai e do irmão;
 | | end
 | | else
 | | | remove do pai; executa testeRLDM(primeiro filho,a_k); executa testeRLDM(segundo
 | | | filho,a_k);
 | | end
 | end
end
if quantos filhos ≥ 3 then
 | calcula U^2 e testa a hipótese;
 | if independe dos filhos then
 | | fica no pai; remove dos filhos;
 | end
 | else
 | | verifica de quais filhos independe; remove do pai e dos irmãos;
 | | for all filhos onde há dependência do
 | | | executa testeRLDM(filho,a_k);
 | | end
 | end
end
```

Algoritmo 4: RLDM: teste da hipótese de independência

Complexidade computacional do RLDM

Os algoritmos que descrevem o RLDM, bem como os do método de Popescul e Ungar [2000], tem alta complexidade computacional, da ordem de $O(m * z^2)$, considerando-se os m atributos de A e que a árvore tenha z nós. Primeiramente, esses métodos precisam acumular as freqüências de cada termo ao longo da hierarquia. Depois, percorrem a hierarquia da raiz para as folhas para cada termo. Em cada nó eles testam a discriminação do termo e decidem se o termo discrimina o nó ou ramos de seus filhos. Se o termo discrimina apenas o nó n_i ou algum n_{ij}, ele é removido de todas as listas de termos dos

```
Input: hierarquia
Output: hierarquia atualizada
for das últimas folhas até a raiz do
    if nó tem conjunto de rótulos vazio then
        if nó tem filhos then
            seu pai recebe os seus filhos; remove o nó;
        end
    end
end
```

Algoritmo 5: RLDM: realizar o corte semântico

descendentes deste, antes de se passar ao teste de um outro termo. Além disso, o método RLDM ainda propõe o corte semântico sobre a hierarquia, e esse processo é aplicado sobre toda a hierarquia novamente, o que incrementa um pouco a sua complexidade passando a $O(m * z^2 + z)$.

Propagação de erros

Ainda, ao usar as freqüências acumuladas em cada nó e a aplicação dos testes da raiz para as folhas, um erro pode ser inserido diversas vezes no processo de decisão. Observando a Figura 5.4, o termo *Embrapa* foi considerado discriminativo na raiz da hierarquia, mas de fato ele não é presente em todos os nós da hierarquia, o que pode ser facilmente verificado observando-se a Figura 5.1.

Para ilustrar esse problema, os valores na Tabela 5.2 representam os três primeiros passos nos dois métodos, aplicados ao termo *Embrapa*, para a coleção de textos do exemplo (Figura 5.1). Os valores das estimativas do Q *de Yule* e do χ^2 são dados para cada passo; e a numeração do nó segue a notação geral de acordo com a Figura 5.3.

1° Passo - nó raiz - n_0			
filho	*Embrapa*	!*Embrapa*	total
n_1	20	80	100
n_2	20	255	275
total	40	335	375
$\alpha = 3.1875$, $\hat{Q} = 0.5224$, $\hat{\sigma}_Q = 0.1240$			
$0 \notin [0.2743, 0.5893]$ and $\hat{\chi}^2 = 11.1362 > 3.84$			
2° Passo - nó n_1			
filho	*Embrapa*	!*Embrapa*	total
n_{11}	10	35	45
n_{12}	10	35	45
total	20	70	90
$\alpha = 1$, $\hat{Q} = 0$			
$0 \in [interval]$ and $\hat{\chi}^2 = 0 < 3.84$			
3° Passo - nó n_2			
filho	*Embrapa*	!*Embrapa*	total
n_{21}	10	131	141
n_{22}	10	124	134
total	20	255	275
$\alpha = 0.9466$, $\hat{Q} = 0.03$			
$0 \in [interval]$ and $\hat{\chi}^2 = 0.0129 < 3.84$			

Tabela 5.2: Resultados dos passos iniciais dos algoritmos dos métodos RLDM e Popescul & Ungar para o termo *Embrapa*

Os resultados para o RLDM podem ser observados na Figura 5.4. No resultado apresentado nessa figura, o termo *Embrapa* foi incluído no conjunto de rótulos da raiz, porque

o termo resultou dependente dos dois filhos de n_0, como se pode observar pelos resultados do primeiro e segundo passos. Os resultados do método de Popescul e Ungar (2000) estão ilustrados na Figura 5.5, na qual o mesmo vê-se que praticamente o mesmo resultado foi observado para o termo *Embrapa*; porém, no método o termo foi incluído nos dois conjuntos de rótulos de n_1 e n_2.

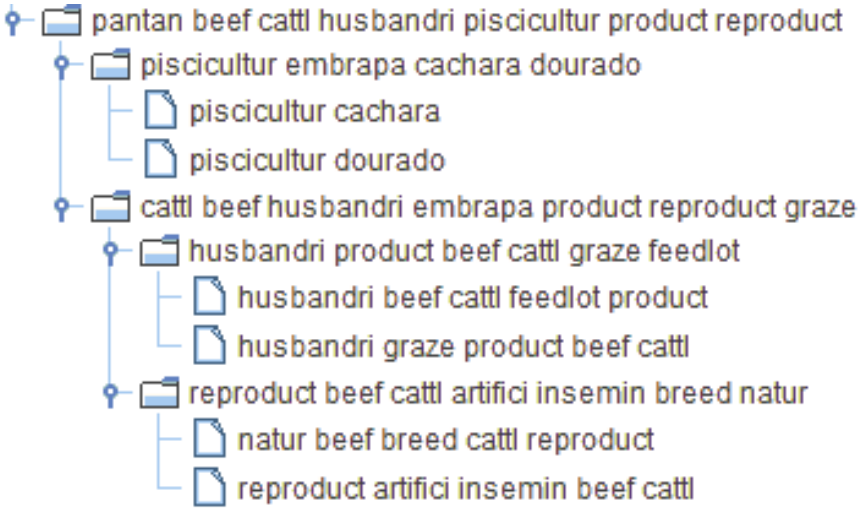

Figura 5.5: Exemplo de rotulação com o método de Popescul & Ungar

Deve ser ressaltado que, os dois resultados estão estatisticamente corretos. Entretanto, o termo *Embrapa* não está presente em todos os nós da hierarquia, nem ao menos em todos os descendentes de n_2, o que é facilmente observado na Figura 5.1. Assim, esse erro é parte dos métodos, e quanto maior a coleção de documentos e, conseqüentemente, a hierarquia, maior a propagação desse tipo de erro ao longo da taxonomia.

5.3.4 O método *Robust Labeling Up* - RLUM

O RLUM - ***Robust Labeling up Method***, utiliza as idéias de Popescul e Ungar e do RLDM, mas foca na redução da complexidade computacional e em uma boa definição dos rótulos para construção da taxonomia de tópicos, procurando com isso minimizar a propagação de erros. Mais precisamente, o foco é uma bem definida estimativa dos rótulos da taxonomia de tópicos; pois, os rótulos identificados são conjuntos de possíveis termos de domínio, ou seja, são aproximações ou estimativas desses termos.

Assim, o conjunto de termos que define o **rótulo específico** L_s de um nó n_i é definido como:

$$L_s(n_i) = \{\forall\ a_k\ /\ I\ é\ V\ \wedge\ II\ é\ V\}$$

$$I : f_{ij}(a_k)\ >\ 0,\ \forall\ j = 1, ..., c$$

$$II : p(a_k|n_{ij}) = E(f_{ij}(a_k)) = f_i(a_k) \times \sum_{t=1}^{m} f_{ij}(a_t) \times \frac{1}{\sum_{t=1}^{m} f_i(a_t))}$$

$$\forall\ j = 1, ..., c$$

Ou seja, se algum atributo a_k tem freqüência não zerada em todos os filhos de um nó

n_i e, se sua freqüência esperada, sob a hipótese de homogeneidade de distribuição entre os filhos de n_i, estiver de acordo com a premissa II, então a_k é um descritor específico do nó n_i. Caso contrário, a_k é descritor específico de cada nó filho de n_i:

$$L_s(n_{ij}) = \{\forall\ a_k\ /\ I\ é\ V\ \wedge\ II\ é\ F\}$$

Conseqüentemente, todo $L_s(n_i)$ é aplicado a n_i e a todos os nós que dele descendem, funcionando como um rótulo genérico de todos os descendentes de n_i. Logo, um **rótulo genérico** pode ser recursivamente definido, a partir do nó raiz n_o, como:

$$L_s(n_0)\ =\ L_g(n_0)$$

e,

$$L_g(n_{ij})\ =\ L_s(n_{ij})\ \bigcup\ L_g(n_i)$$

Dadas essas definições, uma hierarquia de documentos (representada como uma árvore genérica), o conjunto D de documentos, o conjunto A de atributos e uma matriz de incidência das freqüências dos atributos em cada documento (também chamada de modelo espaço-vetorial ou de matriz atributo-valor), definem-se os algoritmos desse método.

Algoritmos para o RLUM

Os Algoritmos **6**, **7** e **8** são aplicados para construir os conjuntos de rótulos de cada nó e, conseqüentemente, estimar os termos da taxonomia de tópicos dessa coleção de documentos.

Input: n_i, a raiz precisa ser o primeiro nó
Output: *Hierarquia atualizada*, com os conjuntos de rótulos em cada nó
if *n_i tem pelo menos um filho* **then**
  recursivamente chama percArvore(filho de n_i);
end
if *se pai de n_i tem próximo filho* **then**
  if *esse filho não é herdado* **then**
    recursivamente chama percArvore(próximo filho do pai de n_i);
  end
end
else
  chama Seleciona(n_i);
end

Algoritmo 6: RLUM: controlar as visitas aos nós da árvore

Deve-se notar que, cada $L_s(n_{ij})$ produzido pelo Algoritmo **7** é uma atualização dos rótulos já existentes para cada n_{ij}. Os que não são folhas inicialmente têm seus $L_s(n_{ij}) = \phi$. Os que são folhas têm seus $L_s(n_{ij})$, no primeiro instante, correspondentes a todos os a_k com freqüência maior que zero nos documentos que formam o grupo n_i. Logo, neste método não é necessário totalizar as freqüências de a_k nos grupos, pois os a_k são selecionados das folhas para a raiz e suas freqüências nos grupos vão sendo atualizadas a cada passo. Ainda, os estimadores usados para verificar as hipóteses de independência de a_k nos n_{ij} são os mesmos utilizados no RLDM.

Input: n_i: nó corrente
Output: $L_s(n_i)$ *e* $L_s(n_{ij})$: atualizados
for *all* a_k, *k=1,..,m, e all* $f_{ij}(a_k) > 0$, *j=1,...,c* **do**
  if $c = 2$ **then**
    if $fi1 \approx 0 \ \vee \ fi2 \approx 0$ **then**
      $L_s(n_i) \leftarrow L_s(n_i) \bigcup \{a_k\}$; $L_s(n_{i1}) \leftarrow L_s(n_{i1}) - \{a_k\}$; $L_s(n_{i2}) \leftarrow L_s(n_{i2}) - \{a_k\}$;
    end
    else
      if a_k *independente de acordo com Q de Yule* **then**
        $L_s(n_i) \leftarrow L_s(n_i) \bigcup \{a_k\}$; $L_s(n_{i1}) \leftarrow L_s(n_{i1}) - \{a_k\}$; $L_s(n_{i2}) \leftarrow L_s(n_{i2}) - \{a_k\}$;
      end
    end
  end
  else
    if $c \geq 3$ **then**
      if a_k *independente de acordo com* U^2 **then**
        for *all filhos de* n_i, $j = 1, ..., c$ **do**
          $L_s(n_{ij}) \leftarrow L_s(n_{ij}) - \{a_k\}$;
        end
        $L_s(n_i) \leftarrow L_s(n_i) \bigcup \{a_k\}$;
      end
    end
    else
      $L_s(n_i) \leftarrow L_s(n_i) \bigcup \{a_k\}$; $L_s(n_{i1}) \leftarrow L_s(n_{i1}) - \{a_k\}$;
    end
  end
end

Algoritmo 7: RLUM: selecionar a_k em $L_s(ni)$ ou $L_s(n_{ij})$

Input: n_i, o nó corrente
Output: n_{ij}, com seu *status* atualizado
for *all filhos de* n_i **do**
  if *all* $f_{ij}(a_k) = 0$,$k = 1, ..., m$ **then**
    $n_i \leftarrow (children\ of\ this\ n_{ij})$; all desses filhos têm seu *status* mudado para "herdado"; n_{ij} é removido da hierarquia;
  end
end
if *all* $f_i(a_k) = 0$, $k = 1, ..., m$ **then**
  (seu nó pai) $\leftarrow$ (all n_i filhos); all desses filhos têm seu *status* mudado para "herdado"; n_i é removido da hierarquia;
end

Algoritmo 8: RLUM: realizar o corte semântico

O Algoritmo **6** percorre a árvore e controla se o nó atual é uma herança. Um nó é herdado quando seu nó pai é cortado da árvore; então ele é herdado pelo seu avô. Um nó n_i é cortado quando não há termos discriminativos em seu conjunto de rótulos, ou seja, o conjunto está vazio após a aplicação do algoritmo **7** a esse nó. Isso corresponde ao corte semântico realizado no RLDM. Conseqüentemente, no Algoritmo **6**, identificar um nó herdado significa identificar que esse nó já foi tratado. O Algoritmo **8** é responsável pelo corte da árvore a cada passo; e, não mais no final do processo, como no RLDM. Assim, também esse processo de corte tem sua complexidade reduzida em relação ao método anterior, que percorre novamente toda a árvore.

Vantagens do método

Este método tem uma complexidade linear $O(mz)$, considerando-se m atributos e z nós para a árvore, porque as decisões são tomadas das classes mais profundas para as mais genéricas. Além disso, esse modo de visitar os nós da árvore diminui a propagação de erros presente nos dois métodos anteriormente mencionados.

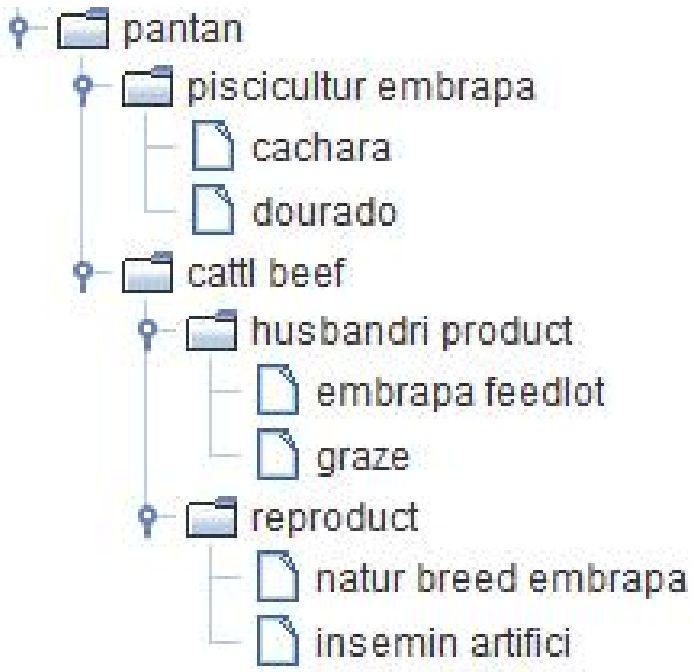

Figura 5.6: RLUM aplicado à coleção de textos do exemplo

O resultado fornecido por este novo método sobre o exemplo é ilustrado na Figura 5.6; onde se observa que o termo *Embrapa* é presente somente nos nós para os quais sua freqüência de ocorrência é realmente significante. Os resultados dos três primeiros passos dos algoritmos em relação ao termo *Embrapa* encontram-se na Tabela 5.3. Para facilitar a interpretação da tabela, deve-se visualizar os nós pelos textos da Figura 5.1. No primeiro passo, o nó n_{11} corresponde ao texto "txt1" e o nó n_{22} corresponde ao texto "txt4"; note que ambos possuem o termo "Embrapa", isto é, o termo tem freqüência não zerada nos dois textos. Logo, nesse primeiro passo, o rótulo específico do nó pai n_1 vai ficar com o termo "Embrapa", que será removido dos rótulos específicos dos nós filhos n_{11} e n_{12}. Seguindo o mesmo raciocínio, os nós $n_{(21)1}$ e $n_{(21)2}$ correspondem respectivamente

aos textos “txt3” e “txt6”; e os nós $n_{(22)1}$ e $n_{(22)2}$ correspondem respectivamente aos textos “txt2” e “txt5”. Assim, $L_s(n_{21})$ não conterá “Embrapa” porque o termo não está na intersecção dos rótulos específicos de seus filhos e, da mesma forma, $L_s(n_{22})$ não conterá “Embrapa” porque o termo não está na intersecção dos rótulos específicos de seus filhos. Conseqüentemente, $L_s(n_2)$ também não conterá “Embrapa”, conforme pode ser observado na Figura 5.6. Deve-se notar, nesses passos, que os valores esperados não chegaram a ser calculados, bem como os testes não chegaram a ser aplicados, pois a primeira condição $(I : f_{ij}(a_k) > 0, \forall\ j = 1, ..., c)$ foi falsa. Deste modo, o termo *Embrapa* pertence a $L_s(n_1)$, isto é, ele pertence a $L_g(n_{11})$ e $L_g(n_{12})$.

1° **Passo - nó** n_1 $“Embrapa'' \in L_s(n_{11}) \bigcap L_s(n_{12})$ $\Rightarrow$ $L_s(n_1) \leftarrow “Embrapa''$ $L_s(n_{11}) \leftarrow L_s(n_{11}) -'' Embrapa''$ $L_s(n_{12}) \leftarrow L_s(n_{12}) -'' Embrapa''$
2° **Passo - nó** n_{21} $“Embrapa'' \notin L_s(n_{(21)1}) \bigcap L_s(n_{(21)2})$ $\Rightarrow$ $“Embrapa'' \notin L_s(n_{21})$
3° **Passo - nó** n_{22} $“Embrapa'' \notin L_s(n_{(22)1}) \bigcap L_s(n_{(22)2})$ $\Rightarrow$ $“Embrapa'' \notin L_s(n_{22})$

Tabela 5.3: Resultados dos passos iniciais dos algoritmos do RLUM para o termo *Embrapa*

5.4 Validação do processo de rotulação

Embora a melhor forma de comparação deva ser uma análise subjetiva junto a especialistas do domínio de conhecimento das coleções de textos utilizadas no processo, devido à natureza dos dados e atributos, essa tarefa não é nada fácil mediante coleções de textos um pouco extensas, muito menos ainda sobre grandes coleções de textos. Mesmo em uma pequena coleção, freqüentemente resultam grandes conjuntos de rótulos e uma grande hierarquia, se nenhum processo de poda de agrupamento for aplicado; e isso dificulta um julgamento subjetivo por parte dos especialistas.

Mesmo assim, o primeiro critério utilizado nas validações foi subjetivo. Procurou-se reduzir as taxonomias apresentadas aos especialistas e avaliar os resultados mediante a comparação de notas. Essa avaliação é detalhada na Subseção 5.4.1.

Ainda, é difícil para o especialista do domínio observar apenas um quesito de julgamento numa análise subjetiva. Por exemplo, ao julgar a forma de atribuir os rótulos a cada grupo, a forma como esses atributos foram extraídos da coleção de textos influencia o julgamento, bem como algum conhecimento anteriormente presente no julgamento do especialista. Então, propõe-se realizar uma análise objetiva dos resultados, detalhada na Subseção 5.4.2.

5.4.1 Avaliação subjetiva da rotulação

Para avaliar a aplicabilidade de cada método de construção do conjunto de rótulos dos agrupamentos realiza-se uma análise subjetiva junto a especialistas do domínio de conhecimento.

Assim, para uma dada coleção de textos de um domínio de conhecimento geram-se tantas taxonomias de tópicos quantos sejam os métodos de rotulação a comparar. Os cuidados que se tomam para conseguir comparações justas é aplicar um único procedimento de pré-processamento à coleção de textos. Então, a partir da matriz atributo×valor geram-se as diferentes taxonomias. Cada taxonomia é levada a uma ferramenta de visualização que permite gravar notas para cada nó. Pede-se então a equipe de avaliadores, que são especialistas do domínio, que atribuam notas a cada lista de atributos em cada nó visualizado, tentando considerar a hierarquia como um todo. Foi escolhido um número par de notas para evitar que, em dúvida, o avaliador decida pela nota média. Dessa forma o avaliador, em cada grupo, deve atribuir uma das seguintes notas a cada lista de rótulos:

- 1: o conjunto de palavras indicado atrapalha a identificação do tópico;
- 2: o conjunto de palavras ajuda pouco na identificação do tópico;
- 3: o conjunto de palavras realmente ajuda a identificação do tópico;
- 4: o conjunto de palavras aproxima-se do ideal na identificação do tópico.

Ao final desse processo de avaliação tem-se uma relação de avaliadores, métodos, grupos e notas, em um delineamento hierárquico (veja (Searle, 1971) para detalhes), pois as notas são dadas dentro de cada grupo e os métodos considerados para cada avaliador. Os dados resultantes são tabelados como na Tabela 5.4, onde:

	avaliador	*método*	*grupo*	*nota*
1	a_1	m_1	c_1	g_{111}
2	a_1	m_1	c_2	g_{112}
...	...	...	...	...

Tabela 5.4: Tabela de dados para avaliação subjetiva dos métodos de rotulação

- a_i: i-ésimo avaliador, $i = 1, 2, ..., A$, A o número de avaliadores;
- m_j: j-ésimo método, com $j = 1, 2, ..., M$, M o número de métodos avaliados;
- c_k: k-ésimo grupo da hierarquia, sendo que k depende do maior número de grupos internos encontrados para cada coleção de textos, isto é, o número total de textos menos 1 (tomando-se todos os grupos mais internos de uma *clusterização* hierárquica *bottom-up*);
- g_{ijk}: nota do i-ésimo avaliador, para o j-ésimo método no k-ésimo grupo.

O modelo para a análise de variância é um modelo linear generalizado hierárquico (Searle, 1971), que pode ser especificado de duas formas:

$$\hat{g} = \hat{\mu} + \hat{a} + \hat{m} + c(\hat{m}) + \hat{e} \tag{5.1}$$

$$\hat{g} = \hat{\mu} + \hat{a} + \hat{m(a)} + \hat{c(m)} + \hat{e} \tag{5.2}$$

nos quais, após o ajuste do modelo tem-se:

- $\hat{g}$: a estimativa da nota;
- $\hat{\mu}$: estimativa do efeito da média geral do modelo sobre a nota;
- $\hat{a}$: estimativa do efeito do avaliador sobre a nota;
- $\hat{m}$: estimativa do efeito do método sobre a nota;
- $\hat{m(a)}$: estimativa do efeito do método aninhado ao avaliador sobre a nota;
- $\hat{c(m)}$: estimativa do efeito do grupo (*cluster*) aninhado ao método sobre a nota;
- $\hat{e}$: erro aleatório da estimativa da nota.

O interesse é verificar se os efeitos do avaliador e método, bem como o efeito método aninhado em avaliador, influenciam significativamente a nota. Após essa verificação, realizam-se as comparações múltiplas de médias das notas dados esses efeitos do modelo (detalhes veja (Snedecor e Cochran, 1967)). Com as comparações múltiplas das médias dos efeitos espera-se inferir como os métodos se agrupam; e, se atendem às expectativas dos avaliadores e às expectativas de seu desenvolvimento.

5.4.2 Avaliação objetiva da rotulação

Na avaliação objetiva, o RLUM é comparado aos outros três métodos descritos neste Capítulo. Assim os quatro métodos foram implementados, por meio de protótipos (na linguagem de programação C): Popescul e Ungar [2000], os Mais Freqüentes (apenas ordenam-se as freqüências em cada nó e, estabelecendo-se um ponto de corte), o RLDM e o RLUM. Deste modo foi possível compará-los em relação aos conjuntos de rótulos para as mesmas hierarquias das mesmas coleções de documentos.

Primeiro, obtém-se o agrupamento hierárquico da coleção de textos, então uma cópia dessa hierarquia e da matriz de incidência é fornecida a cada implementação dos métodos. Como o RLDM e o RLUM podem realizar cortes na hierarquia, eles podem alterar a hierarquia final. De modo a comparar corretamente os métodos, não se devem comparar os conjuntos de rótulos obtidos para a hierarquia toda a partir dos quatro métodos, mas sim apenas os conjuntos de rótulos dos grupos (nós) mantido nas hierarquias após a aplicação dos métodos. Deste modo, as comparações são realizadas nó a nó, de acordo com a presença dos mesmos na hierarquia. Ainda, para que as comparações sejam justas, o ideal é que os conjuntos de rótulos sejam de tamanhos semelhantes. Pode-se forçar essa situação, colocando-se um corte de número máximo de rótulos em cada conjunto, como costuma ser feito para o método dos Mais Freqüentes. Utilizando essa idéia, basta aplicar o método dos Mais Freqüentes a cada resultado apresentado pelos demais métodos; ou seja, toma-se cada $L_s(n_i)$, ordenam-se os elementos por freqüência e então se limita o conjunto aos f elementos mais freqüentes. Assim, neste trabalho, em todas as comparações, serão

utilizados os mesmos grupos, que são os nós mantidos pelos diferentes métodos, e conjuntos de rótulos de tamanhos semelhantes.

Decorre das definições de taxonomia de tópicos, neste trabalho, que os termos não são replicados ao longo dos mesmos ramos na hierarquia. Deste modo, o RLUM deve produzir realmente os conjuntos de rótulos mais específicos em cada nó e os mais gerais para os descendentes desse nó com melhor índice de acertos que os três outros métodos. Para testar essa primeira hipótese, a proposta é utilizar os conjuntos de rótulos selecionados como expressões de busca em um processo de recuperação de informações, como explicado na próxima sub-seção. Uma segunda hipótese é que o RLUM produz um vocabulário maior, ou no mínimo equivalente, comparado aos demais métodos. Se um vocabulário maior é obtido, espera-se que esse vocabulário tenha uma melhor cobertura do domínio em que a coleção se insere. Finalmente, se o vocabulário tem uma maior cobertura, sua qualidade deve ser medida e comparada à qualidade dos vocabulários obtidos com os três outros métodos, conforme explicado na última sub-seção.

Conjuntos de rótulos como expressões de busca

Todos os conjuntos de rótulos são tratados como expressões de busca, nos quais o operador "*and*" é usado entre os termos de um mesmo conjunto. Por exemplo, o conjunto de rótulos $L_s(n_{21}) = \{husbandri, product\}$ é usado como uma expressão de busca da seguinte forma: "$husbandri \wedge product$". Deste modo, espera-se que os documentos do nó n_{21} sejam recuperados, isto é, $txt3$ e $txt6$ (veja Figura 5.1); se o processo e busca for aplicado aos textos da coleção do exemplo.

O processo de recuperação de informações foi implementado (outro protótipo em C) sobre a matriz atributo-valor utilizada no agrupamento de documentos. Na representação utilizada para essa matriz temos em suas linhas os documentos e em suas colunas os atributos utilizados, sendo que cada célula corresponde à freqüência de ocorrência do termo no documento. Para decidir se um documento foi ou não recuperado, pela expressão, todos os termos da busca devem ter freqüência maior que zero no documento. Após a aplicação do processo de recuperação, para cada grupo (nó da hierarquia) calculam-se os valores da Tabela 5.5.

	recuperados	*!recuperados*	*total*
n_i	t_p	f_n	d_c
$!(n_i)$	f_p	t_n	n_c
	r_d	n_r	o

Tabela 5.5: Número de recuperações em cada n_i

As medidas de interesse sobre os valores da Tabela 5.5 que serão utilizadas são:

- precisão: a proporção de documentos dentre os recuperados que está na classe correta, $p = t_p/r_d$;

- "recall": a proporção de documentos dentre os do grupo, que foram corretamente recuperados, $r = t_p/d_c$, também chamada de **cobertura de recuperação**;

- "false alarm": a proporção dentre os documentos recuperados que, de fato, não pertence ao grupo, $f_a = f_p/n_c$;

- "F measure": a média harmônica entre precisão e "recall", $F = 2 * p * r/(p + r)$. O valor ideal de F é igual a um, porque o ideal é ter $p = 1$ e $r = 1$.

Essas medidas podem ser diretamente usadas para comparar os resultados, porque eles estão sendo computados para os mesmos grupos apenas, isto é, grupos com os mesmos documentos; a única diferença são os conjuntos de rótulos selecionados por um mesmo método, sendo estes limitados a tamanhos semelhantes, com no máximo f elementos cada.

Particularmente, a medida de maior interesse é a cobertura de recuperação ("recall"), porque ela representa bem a qualidade do conjunto de rótulos, quando usado como expressão de busca. Quando os rótulos são usados como palavras-chaves de busca, ligados pelo operador "*and*", se forem realmente representativos do grupo, o valor de "recall" deverá se aproximar de um, isto é, ter-se-ia um número muito baixo de falsos negativos (f_n) e um número de positivos verdadeiros (t_p) bastante alto, próximo ao total de documentos no grupo.

Para verificar experimentalmente uma hipótese sobre o comportamento de uma dessas medidas para cada um dos métodos de rotulação, calculam-se as medidas para cada um dos nós mantidos nas quatro taxonomias. Tabulam-se essas medidas junto ao número do nó e ao método que as gerou, e, então se ajusta um modelo linear generalizado (ver (Searle, 1971)) para analisar a variância e obter estimativas para cada medida. O modelo considera os efeitos do nó e do método de rotulação, além da média geral:

$$\hat{m_e} = \hat{\mu} + \hat{n_{ij}} + \hat{l_m} + \hat{\epsilon} \tag{5.3}$$

na qual,

- $\hat{m_e}$: estimativa da medida de avaliação após ajuste do modelo;

- $\hat{\mu}$: média geral do modelo, sem qualquer outro efeito;

- $\hat{n_{ij}}$: estimativa do efeito do nó n_{ij}, que é o quanto a medida $\hat{m_e}$ desvia-se da média geral devido a esse efeito;

- $\hat{l_m}$: estimativa do efeito do método de rotulação,que é o quanto a medida $\hat{m_e}$ desvia-se da média geral devido à aplicação de um dado método;

- $\hat{\epsilon}$: efeito aleatório associado a cada estimativa.

Com o uso dessas estimativas para cada medida, $\hat{m_e}$, para cada método de rotulação $\hat{l_m}$ espera-se obter comparações mais estatisticamente confiáveis. Então, aplicam-se testes de comparação múltiplas de médias a cada modelo, e, também se utilizam essas estimativas para construir gráficos das medidas para cada coleção de textos e método de rotulação.

No exemplo usado para mostrar o desenvolvimento das idéias de rotulação, foi utilizado um conjunto de atributos retirado de um vocabulário do domínio pré-existente, no caso, do Thesagro (*Thesaurus* Agrícola Nacional (BINAGRI, 2009)). Quando isso ocorre, espera-se que os vocabulários finais tenham cardinalidades equivalentes para qualquer método de rotulação. Pois, se os termos estão nos textos e passaram pelo filtro de atributos aplicado espera-se que farão diferença significativa nos agrupamentos e servirão para discriminar os grupos na rotulação. Quando se utiliza um conjunto de atributos com base apenas na geração de termos a partir das coleções com a ferramenta PreText (Matsubara et al., 2003), não há garantia de que todos os termos selecionados para o conjunto de atributos façam parte de algum *thesaurus*, ou mesmo que todos eles estejam presentes no vocabulário final obtido.

Assim, se os atributos foram selecionados entre termos gerais da coleção e textos, retiradas as *stopwords* e aplicado algum filtro estatístico (por exemplo, cortes de Luhn), espera-se que o número de diferentes termos produzidos pelo RLUM seja maior que por todos os outros. Como mencionado, anteriormente, se o vocabulário tiver essa propriedade, é esperado que ele seja capaz de ter uma melhor cobertura do domínio no qual a coleção de textos se insere, ou seja, seria uma melhor estimativa do vocabulário de domínio. Espera-se, nesse caso, que o RLUM produza um vocabulário maior, ou no mínimo igual ou equivalente, em cardinalidade ao dos outros métodos.

Para verificar essa hipótese, as cardinalidades dos vocabulários obtidos por cada método devem ser comparadas, para diferentes coleções de textos. Experimentalmente, é aconselhável ter coleções de diferentes tamanhos, diferentes idiomas ou domínios; e, então, aplica-se um teste de comparação múltipla de médias (veja (Searle, 1971; Ramsey, 1993; Snedecor e Cochran, 1967), para maiores detalhes), para agrupar comparativamente as cardinalidades.

Primeiramente, vamos definir o vocabulário V como a união de todos os conjuntos de rótulos da taxonomia de tópicos obtida:

$$V = \bigcup_{i=1}^{z} L_s(n_i),$$

$$z: \ \textit{número total de nós na hierarquia}$$

Então, elabora-se a hipótese sobre as diferentes cardinalidades:

$$H_0 : card(V_R) \geq card(V_F) \geq card(V_P) \geq card(V_M)$$

$$V_R: \ \textit{vocabulário para RLUM}$$

$$V_F: \ \textit{vocabulário para "Most Frequent"}$$

$$V_P: \ \textit{vocabulário para Popescul e Ungar}$$

$$V_M: \ \textit{vocabulário para RLDM}$$

Para testar essa hipótese, calculam-se os vocabulários para algumas diferentes bases, ajusta-se um modelo linear generalizado para calcular as variações de cada efeito dos métodos e cardinalidade do conjunto e atributos iniciais sobre as cardinalidades finais do vocabulário, da seguinte forma[2]:

$$\hat{card}_v = \hat{\mu} + < \hat{card}_A > + \hat{m} + \hat{\epsilon} \tag{5.4}$$

no qual,

- $\hat{card}_v$: cardinalidade final inferida para cada vocabulário em cada método e cada base;
- $\hat{\mu}$: a média geral de dispersão estimada para inferir as cardinalidades finais;
- $< \hat{card}_A >$: estimativa do efeito da cardinalidade do conjunto de atributos iniciais, em cada caso, para inferência da cardinalidade final. Utilizando-se a cardinalidade de A como covariável é esperado que o efeito de diferentes número de atributos seja devidamente tratado e considerado nas inferências finais, permitindo que se possa comparar apenas os efeitos relativos a diferentes métodos de rotulação sobre a cardinalidade do vocabulário final;
- $\hat{m}$: estimativa do efeito do método de rotulação aplicado, isto é, quanto a estimativa final desvia-se da média geral de dispersão em relação a esse efeito;
- $\hat{\epsilon}$: erro aleatório de ajuste do modelo linear.

Representatividade do vocabulário

Uma maior cardinalidade do vocabulário não garante a qualidade do mesmo, pois verificar essa propriedade pode ser apenas reflexo de uma série de más escolhas. Para avaliar as escolhas, que podem ser vistas como quão bem os rótulos selecionados predizem o conteúdo dos documentos em cada grupo, além do uso da medida *recall*, pode-se utilizar a esperança da informação mútua entre o conjunto de atributos e o vocabulário final. Para a coleção completa de documentos após a rotulação, a Esperança da Informação Mútua (EMIM) ((Manning e Schütze, 1999) e (Lawrie et al., 2001)) é definida como:

$$EMIM(V, A) = \sum_{u=1}^{card(V)} \sum_{k=1}^{m} p(v_u, a_k) * \log(\frac{p(v_u, a_k)}{p(v_u) \times p(a_k)})$$

Considerando-se cada $p(v_u, a_k)$ como a probabilidade de ocorrerem simultaneamente o vocábulo $v_u \in V$ e o atributo $a_k \in A$, e, esta estimativa da probabilidade ser o estimador de máxima verossimilhança dado pelo número de ocorrências de v_u e a_k em cada documento da coleção. Assim, como definida aqui, a EMIM reflete estatisticamente a qualidade do vocabulário selecionado. Para comparar essa qualidade entre os quatro métodos, a proposta é considerar os diferentes conjuntos de rótulos produzidos por cada

[2]Toda estimativa é notada com chapéu.

método para cada grupo de documentos e calcular:

$$EMIMn_i(L_s(n_i), A) = \sum_{u=1}^{card(L_s(n_i))} \sum_{k=1}^{m} p(l_u, a_k) * \log(\frac{p(l_u, a_k)}{p(l_u) \times p(a_k)})$$

Considerando-se cada l_u como um elemento do conjunto de rótulos $L_s(n_i)$. Como esses valores são calculados para os mesmos grupos de documentos, os valores obtidos para a EMIM podem ser comparados utilizando-se comparação múltipla de médias. E, ainda, como a EMIM não tem um intervalo de valores definidos ou mesmo um máximo, os maiores valores obtidos refletem as melhores predições. Porém, a EMIM é sensível ao tamanho dos conjuntos envolvidos; mas esse problema já foi evitado ao se considerar apenas os nós (grupos) comuns às hierarquias e limitar o tamanho de cada conjunto de rótulos a, no máximo, f elementos cada. Assim, para testar a hipótese de que a EMIM em cada nó, para cada método, é maior para o RLUM, podendo, no mínimo, compararse aos demais métodos, ajusta-se o seguinte modelo linear para a análise de variância e comparação múltipla de médias:

$$\hat{emim}(n_i) \;=\; \hat{\mu} \;+\; \hat{n_i} \;+\; \hat{m} \;+\; \hat{\epsilon} \tag{5.5}$$

no qual,

- $\hat{emim}(n_i)$: EMIM final inferida para cada nó em cada método e cada base;
- $\hat{\mu}$: a média geral de dispersão estimada para inferir a $emim(n_i)$ final;
- $\hat{n_i}$: estimativa do efeito de cada nó n_i em cada base na inferência da emim final. O nó funciona como efeito de categoria para cada EMIM, fazendo com que as comparações de médias dos efeitos dos métodos sejam justas por estarem sendo realizadas nó a nó;
- $\hat{m}$: estimativa do efeito do método de rotulação aplicado, isto é, quanto a média geral de dispersão desvia-se da EMIM final em relação a esse efeito;
- $\hat{\epsilon}$: erro aleatório de ajuste do modelo linear.

5.5 Resultados experimentais da avaliação subjetiva

Em um primeiro momento foram realizados experimentos com bases de dados relativamente grandes, contendo mais de mil textos de um único domínio de conhecimento. Os domínios foram de Gado de Corte, Informática Agropecuária, Manufatura e Melhoramento de Gado, especificamente. Observou-se que seria impossível apresentar extensas taxonomias a especialistas do domínio, para que realizassem uma análise subjetiva minuciosa; a tarefa não seria trivial e comprometeria a qualidade da avaliação.

Para tentar viabilizar uma avaliação de qualidade, que permitisse comparar os métodos implementados, optou-se em primeiro lugar por uma base de dados menor, que pudesse ser julgada significativa para os domínios em questão. Primeiramente, acertou-se quais

seriam os especialistas que julgariam os métodos e em função desses fixou-se o domínio de conhecimento e, conseqüentemente, a amostra de textos a ser utilizada. Então se realizou o pré-processamento dos dados, foram escolhidos cortes e representações; extraíram-se os agrupamentos e geraram-se os rótulos. Cada taxonomia gerada foi levada a uma ferramenta de avaliação, que permite visualizá-las e atribuir notas aos nós. Assim, foram realizadas as avaliações por parte dos especialistas; e a seguir tabulados e analisados os resultados.

Nessa avaliação subjetiva, o RLUM ainda não havia sido desenvolvido, e o quarto método julgado é uma modificação à restrição do método de Popescul & Ungar (Moura e Rezende, 2007). Esse método foi chamado no experimento de qui-quadrado adaptado, pois em lugar de utilizar a restrição de freqüências e valores esperados ≥ 5 ele usa freqüências e valores esperados ≥ 1 (veja Seção 5.3.2). Esse método serviu para avaliar a modificação da restrição do Popescul & Ungar, mas deixou de ser utilizado, pois ele não se mostrou vantajoso em relação ao RLDM.

5.5.1 Delimitação das coleções de textos e pré-processamento

Foram escolhidas coleções de textos para o domínio gado de corte, especificamente gado da raça Canchim, e para Informática Agropecuária. Para o domínio Canchim houve a disponibilidade de dois avaliadores especialistas do domínio, um avaliador especialista em informação[3] para o domínio e 48 textos completos - digitalizados em sua íntegra e em português. Para o domínio de Informática Agropecuária, muito amplo, foram selecionados 31 resumos de assuntos variados e atuais, em português, havia um especialista em informação para o domínio e três especialistas em subdomínios e com bom conhecimento do domínio como um todo.

	$CANCHIM$	$InformáticaAgropecuária$
$documentos$	48	31
$unigramas$	7025	604
$unigramas - Luhn$	2095	125
$unigramas - Salton$	5868	514

Tabela 5.6: Número de *unigramas* antes e depois dos filtros utilizados, na análise subjetiva

Para o pré-processamento das coleções de textos foram utilizadas as mesmas listas de *stopwords* e o mesmo procedimento de *stemmização*, gerando-se atributos *unigramas* e contabilizando-se a freqüência absoluta de ocorrência desses nos textos. Para realizar os filtros foram utilizados cortes de Luhn, como mostrado na Tabela 5.6. Porém, como a coleção de textos de Informática Agropecuária escolhida é muito pequena e composta por resumos, o número de atributos ficou muito reduzido e obteve-se, na fase seguinte, um agrupamento quase linear dos textos. Assim, fez-se uma nova escolha dos filtros, optando-se por usar os unigramas que ocorreram entre um a dez por cento da coleção de textos, que é uma recomendação de Salton et al. (1975). Essa escolha aumentou o número de atributos e permitiu obter uma hierarquia mais representativa para os 31 resumos.

[3]Os especialistas em informação para o domínio foram bibliotecários experientes em catalogação de publicações do domínio específico.

5.5.2 Agrupamento hierárquico e rotulação

Para o agrupamento hierárquico aglomerativo utilizou-se a medida de dissimilaridade baseada em distância euclidiana e o algoritmo *complete linkage* do MatLab (versão 7). Na Figura 5.7 é ilustrada a hierarquia para a amostra de textos de informática agropecuária, já com os rótulos calculados para os métodos Mais Freqüentes e RLDM; os dois outros métodos não apresentam diferenças significativas, pois de Popescul & Ungar aproxima-se muito do Mais Freqüentes.

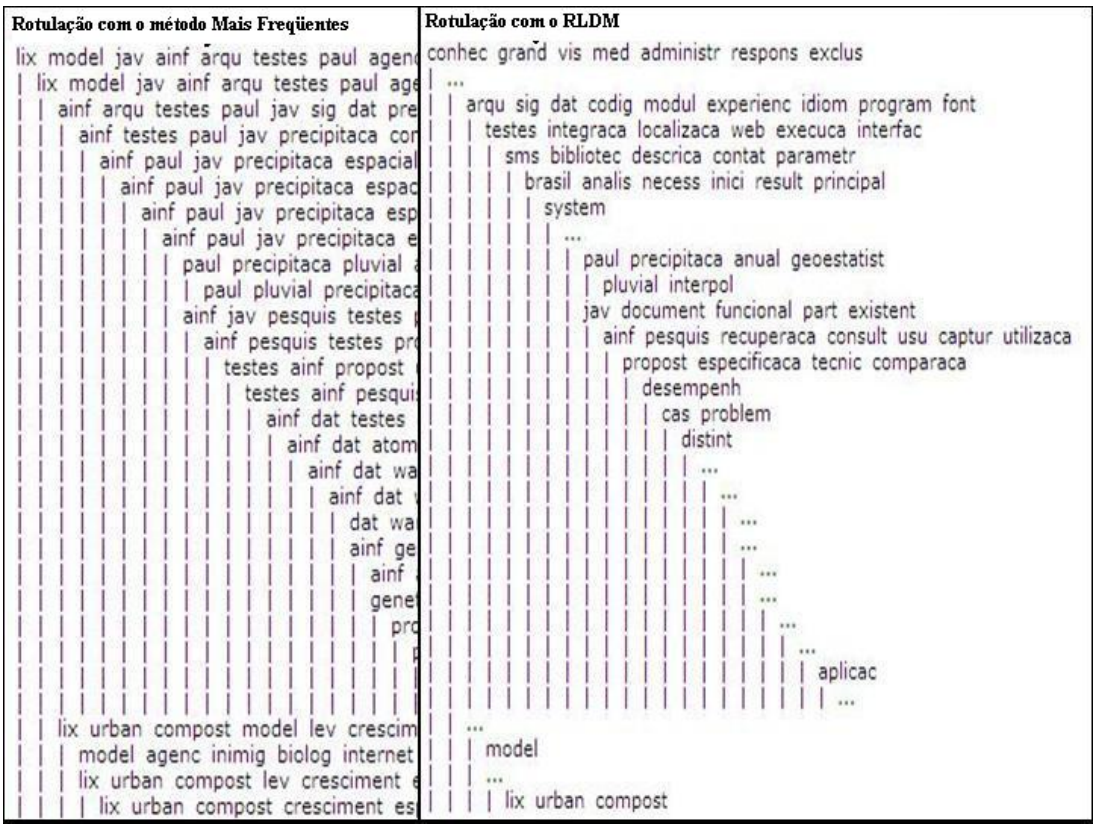

Figura 5.7: Hierarquia informática agropecuária: comparação dos resultados para o método Mais Freqüentes e RLDM

Deve-se notar que no RLDM, cujo resultado é ilustrado na Figura 5.7, quando nenhuma das palavras de um grupo é significativamente associada ao grupo, a lista de palavras discriminativas fica vazia. Isso ocorre quando se usa a opção de não cortar a hierarquia nesses casos. Nesta avaliação essa opção foi usada para manter as mesmas estruturas hierárquicas nas taxonomias. Os avaliadores foram devidamente informados do porquê de manter esses rótulos vazios. Na visualização criada essa lista vazia é representada por "...". A idéia é que a lista ficou vazia porque não havia sequer um termo, pertencente à coleção e selecionado como atributo, que fosse significativo para discriminar o grupo em questão.

5.5.3 Análise dos resultados

A partir das taxonomias geradas para cada domínio e método foram realizadas as avaliações, tabulados e analisados os resultados. Além das notas atribuídas a cada nó e método pelos avaliadores, eles também enviaram críticas e sugestões. Dentre esses

comentários os mais relevantes referiram-se à forma de gerar os atributos. Especialmente entre os especialistas em informação para o domínio, habituados ao uso de *thesaurus*, há a expectativa de encontrar representações n-gramas, ou melhor, termos de fato compostos, como por exemplo "composto de lixo urbano" em lugar de uma lista com "lix urban compost cresciment ...", como nas últimas folhas das taxonomias mostradas na Figura 5.7.

Outro ponto de confusão foram os nós em que a lista palavras discriminativas eram vazias e representadas pelos "...”; para alguns avaliadores o significado disso era extremamente claro e para outros completamente obscuro. Ou seja, ter mantido as taxonomias com a mesma configuração hierárquica acabou sendo um fator negativo na avaliação.

Devido à confusão relativa aos nós rotulados como "...”, na avaliação dos efeitos do método e do método aninhado a avaliador nos diferentes modelos, foram sempre consideradas dois conjuntos de dados, para cada domínio separadamente: **completo**, isto é, todas as avaliações para todos os nós, em todos os métodos e domínios; e, **sem os "...”**, isto é, apenas as observações relativas àqueles nós que não apresentaram os "...” em nenhum dos métodos de rotulação.

Para cada tabela de resultados, são apresentadas as comparações múltiplas de médias das notas para cada grupo de efeitos estudado. Nessas comparações foi utilizado o teste SNK (Student-Newman-Keuls), com a raiz do erro quadrático do modelo ajustado, os graus de liberdade desse erro e um nível de significância de 5%. Esse teste foi escolhido devido a sua força, pois diferenças reais são mais freqüentemente apontadas por ele (Snedecor e Cochran, 1967). Nas colunas de comparação das tabelas, letras iguais significam que as notas comparadas pertencem aos mesmos grupos; sendo que as letras estão em ordem ascendente de agrupamento de notas.

Na Tabela 5.7 estão tabuladas as comparações múltiplas da variável nota ajustada pelos efeitos de média geral, avaliador, método e nó aninhado a método. Nota-se que no domínio Canchim houve uma clara preferência pelo método dos Mais Freqüentes, seguido pelo de Popescul & Ungar, dado que eles sempre provêm uma lista de palavras interpretável para cada agrupamento, mesmo que com palavras em excesso. No entanto, retirando-se os nós sem rótulos e reajustando-se o modelo, os métodos de Popescul & Ungar e RLDM têm suas notas mais próximas. Em contrapartida, para o domínio de Informática Agropecuária, os métodos estão claramente divididos em ordem de preferência: Popescul & Ungar, Mais Freqüentes, RLDM e qui-quadrado adaptado; e, quando se retiram os nós sem rótulos, então se agrupam os que provêm maiores listas de palavras (Popescul & Ungar e Mais Freqüentes) e os que mais eliminam palavras da hierarquia (RLDM e qui-quadrado adaptado). Orientando-se somente por essa primeira análise, parece que retirar as palavras menos discriminativas dos rótulos dos nós não ajuda o especialista a identificá-los, sem o auxílio das listas de rótulos dos nós ascendentes.

Na análise seguinte, procurou-se observar o efeito do método aninhado a cada avaliador. Para isso, ajustou-se o modelo de nota dependente de média geral, avaliador, método aninhado a avaliador e nó aninhado a método, de acordo com a Equação 5.2. A Tabela 5.8 corresponde à comparação múltipla de médias usando esse modelo para o domínio Canchim e a Tabela 5.9 para o domínio de Informática Agropecuária. Os efeitos de cada avaliador aninhado ao m-ésimo método são representados por $i_k(m)$, o k-ésimo especialista

$CANCHIM$					
$Completo$			$Sem-...$		
$\sqrt{e}=0.29, df=290$			$\sqrt{e}=0.27, df=122$		
m	g	$grupo$	m	g	$comp.$
$MaisFrequentes$	2.58	a	$MaisFrequentes$	2.51	a
$Popescul\&Ungar$	2.20	b	$Popescul\&Ungar$	2.19	b
$RLDM$	1.61	c	$RLDM$	2.13	b
$qui-quadrado$	1.48	c	$qui-quadrado$	1.60	c
Informática Agropecuária					
$Completo$			$Sem-...$		
$\sqrt{e}=0.34, df=316$			$\sqrt{e}=0.39, df=133$		
m	g	$comp.$	m	g	$comp.$
$Popescul\&Ungar$	2.83	a	$Popescul\&Ungar$	2.65	a
$MaisFrequentes$	2.67	b	$MaisFrequentes$	2.44	a
$RLDM$	1.62	c	$RLDM$	1.79	b
$qui-quadrado$	1.35	d	$qui-quadrado$	1.75	b

Tabela 5.7: Comparação múltipla de médias das notas considerando os efeitos dos métodos

em informação para o domínio, e $d_j(m)$, o j-ésimo especialista no domínio. Os métodos são: 1-Popescul & Ungar, 2- qui-quadrado adaptado, 3- RLDM e 4-Mais Freqüentes.

A principal intenção dessa análise foi verificar se as avaliações dos especialistas em informação para o domínio diferiam muito dos especialistas no domínio. No entanto, as maiores diferenças apresentadas foram entre diferentes especialistas do domínio; ou seja, ficou difícil avaliar o que seria melhor para esse grupo de especialistas. Assim, focou-se a atenção nas avaliações dos especialistas em informação para o domínio, dado que estes correspondem a bibliotecários experientes na catalogação de informação para o domínio.

$Completo$			$Sem-...$		
$\sqrt{e}=0.29, df=284$			$\sqrt{e}=0.26, df=116$		
$m(a)$	g	$comp.$	$m(a)$	g	$comp.$
$d_1(4)$	3.02	a	$d_1(4)$	3.06	a
$i_1(4)$	2.66	b	$d_1(1)$	2.81	a
$d_1(1)$	2.51	bc	$d_1(3)$	2.69	a
$i_1(1)$	2.30	c	$i_1(4)$	2.69	a
$d_2(4)$	1.92	d	$i_1(3)$	2.25	b
$d_1(3)$	1.72	de	$i_1(1)$	2.19	b
$d_2(1)$	1.67	def	$d_1(2)$	1.94	bc
$d_1(2)$	1.66	def	$d_2(4)$	1.73	cd
$i_1(3)$	1.65	def	$i_1(2)$	1.63	cde
$i_1(2)$	1.40	efg	$d_2(1)$	1.53	cde
$d_2(3)$	1.36	fg	$d_2(3)$	1.40	de
$d_2(2)$	1.21	g	$d_2(2)$	1.20	e

Tabela 5.8: Comparação múltipla de notas com o modelo $\hat{g}=\hat{\mu}+\hat{a}+\hat{m(a)}+\hat{c(m)}+\hat{e}$ no domínio Canchim

Para o domínio de Canchim, Tabela 5.8, considerando-se as avaliações do especialista em informação, nota-se que no modelo ajustado para todos os nós há uma nítida preferência pelo método dos Mais Freqüentes seguido pelo de Popescul & Ungar. Nesse caso, os métodos qui-quadrado adaptado e RLDM ficam agrupados em uma terceira colocação. Descontando-se os nós que não possuem rótulos e reajustando-se o modelo, o método Mais Freqüentes continua na liderança, porém o método RLDM aproxima-se do Popescul & Ungar, ficando no mesmo grupo, e o método qui-quadrado adaptado é empurrado sozinho para a terceira colocação. Isso pode sugerir que o método RLDM atende bem

Completo $\sqrt{e} = 0.30, df = 307$			*Sem − ...* $\sqrt{e} = 0.31, df = 124$		
$m(a)$	g	*cmp*	$m(a)$	g	*comp.*
$d_1(1)$	3.10	a	$d_1(1)$	3.00	a
$d_2(1)$	2.97	a	$d_2(1)$	2.83	a
$i_1(1)$	2.88	a	$i_1(4)$	2.75	ab
$d_1(4)$	2.87	a	$d_1(3)$	2.67	abc
$i_1(4)$	2.83	a	$d_1(4)$	2.67	abc
$d_2(4)$	2.80	a	$i_1(1)$	2.67	$abcd$
$d_3(1)$	2.40	b	$d_2(4)$	2.50	$abcde$
$d_1(3)$	2.27	b	$d_1(2)$	2.33	$abcdef$
$d_3(4)$	2.20	b	$d_3(1)$	2.08	$bcdefg$
$i_1(2)$	1.67	c	$i_1(2)$	2.00	$c\ efg$
$d_1(2)$	1.67	c	$d_3(4)$	1.83	fg
$i_1(3)$	1.52	c	$d_3(2)$	1.75	fg
$d_3(3)$	1.37	cd	$i_1(3)$	1.58	gh
$d_2(3)$	1.30	cd	$d_3(3)$	1.50	gh
$d_3(2)$	1.30	cd	$d_2(3)$	1.42	gh
$d_2(2)$	1.03	cd	$d_2(2)$	1.08	gh

Tabela 5.9: Comparação múltipla de notas com o modelo $\hat{g} = \hat{\mu} + \hat{a} + \hat{m(a)} + \hat{c(m)} + \hat{e}$ no domínio Informática Agropecuária

às expectativas desse especialista, quando se refere exclusivamente a apontar palavras realmente discriminativas do grupo.

Na Tabela 5.9 encontram-se os resultados para o domínio de informática agropecuária. O especialista em informação, nesse domínio claramente ficou com os métodos Popescul & Ungar e Mais Freqüentes na primeira colocação e os qui-quadrado adaptado e RLDM em outro grupo, quase único; considerando os dois grupos de observações dos dados. Porém, devido à grande discrepância entre os especialistas do domínio, quando se retiram as observações para as quais alguns nós apresentam lista de rótulos vazia, não há diferença significativa entre as notas dos métodos para os especialistas 1 e 2, enquanto que, para os de número 2 e 3 o método RLDM seja o pior de todos.

5.5.4 Considerações da avaliação subjetiva

Como vimos, não foi uma boa escolha deixar as hierarquias iguais nas taxonomias avaliadas. A manutenção de listas de rótulos vazias prejudicou a avaliação do RLDM. Além disso, os avaliadores deram preferência ao método que sempre produz um conjunto maior de termos, mesmo que os repita ao longo da hierarquia. E, ainda, os especialistas deixam-se influenciar por várias outras variáveis, desde a forma de geração dos atributos. Ainda, nas críticas e sugestões enviadas vinham observações sobre a hierarquia e questionamentos variados sobre o porquê de uma árvore e não um grafo genérico, etc. O perfil do especialista também parece influenciar o resultado, pois os de informação para o domínio acabam tendo uma melhor compreensão dos objetivos. Ou seja, mesmo preparando anteriormente os especialistas, explicando o objetivo da análise e fornecendo um exemplo, outros fatores subjetivos prejudicavam a leitura dos resultados.

Assim, dessas primeiras avaliações, concluiu-se que a avaliação puramente subjetiva acaba sendo inconclusiva. Por isso, para os próximos experimentos procurou-se integrar o uso de critérios objetivos à avaliação.

5.6 Resultados experimentais da avaliação subjetiva integrada a critérios objetivos

Nesse experimento, usou-se a avaliação subjetiva proposta e a avaliação objetiva utilizando os rótulos gerados como expressão de busca como definidos na Subseção 5.7.2. Apenas os conjuntos de rótulos específicos foram utilizados como expressão de busca neste experimento. A idéia era tentar repetir o que os avaliadores julgavam importante, isto é, que cada conjunto de rótulos específicos fosse capaz de representar bem o conjunto de documentos sob o seu grupo e, supostamente, recuperar esse grupo ou a maioria de seus documentos.

Primeiramente os avaliadores foram escolhidos entre alguns especialistas para cada coleção de textos. Então, tomou-se uma pequena e bem estabelecida amostra de documentos de cada domínio. Novamente, as amostras são pequenas em função da dificuldade em avaliar subjetivamente grandes taxonomias.

A primeira coleção de documentos foi escolhida entre publicações científicas em português sobre inteligência artificial, correspondendo a um total de 47 artigos completos. A segunda coleção veio do domínio de lingüística computacional, correspondendo a um conjunto de 51 artigos dos eventos dos seminários do TIL (TIL - Tecnologia da Informação e da Linguagem Humana), ocorridos entre 2005 a 2007.

5.6.1 Obtenção das taxonomias

Na etapa de pré-processamento, as mesmas listas de *stopwords* e o mesmo processo de radicalização foram aplicados às duas coleções de textos separadamente, usando a ferramenta PreText (Matsubara et al., 2003). Foram extraídos atributos unigramas e criadas as matrizes atributo×valor de cada coleção. O processo de filtro foi baseado em cortes de Luhn após o corte de $df \geq 2$, devido ao fato de se estar utilizando *cluster* aglomerativo. As matrizes foram processadas com o MatLab, usou-se dissimilaridade de cosseno e o algoritmo *average linkage* para construir as hierarquias. Para rotular as hierarquias foram usados apenas os métodos de Popescul & Ungar e o RLDM, dado que o interesse era avaliar o RLDM em relação a seu método inspirador (Moura et al., 2008c).

Os resultados de um ramo da hierarquia obtida a partir da coleção de textos de inteligência artificial, rotulada com os dois métodos, pode ser observada na Figura 5.8. Na figura, para o método de Popescul & Ungar, alguns termos são repetidos ao longo dos nós filhos, por exemplo *"document, sentenc, ontolog, reticul..."*. Por outro lado, nos resultados obtidos com o RLDM os termos não se repetem e são aparentemente mais discriminativos. Um bom exemplo é comparar os resultados dos dois métodos para os últimos filhos das hierarquias mostradas na Figura 5.8, nos quais *"document, ontolo, domin..."* correspondem ao rótulo *"sinonimia, ..."* e *"text, sentence, reticul, extract..."* correspondem à lista de rótulos *"summariz, estrutur, corresponden, ..."*. Esse exemplo é uma evidência de que no RLDM, de fato, os termos mais genéricos ficam nos nós mais altos e os mais específicos mais próximos às folhas, sem apresentar repetições.

Figura 5.8: Taxonomia parcial obtida a partir da coleção de inteligência artificial

5.6.2 Avaliação subjetiva

A avaliação subjetiva foi realizada para cada método e coleção, junto aos especialistas do domínio. Para comparar os métodos os avaliadores foram divididos em pares. Cada par de avaliadores atribuía uma nota a cada lista de rótulos após observar os resultados dos dois métodos para o mesmo nó, junto ao ramo onde o nó se encontra na taxonomia. Na Tabela 5.10 encontram-se as médias finais das notas ($\overline{g}$) e seus desvios padrões (se) para cada método. Como são apenas dois métodos, em lugar de utilizar um teste de comparação múltipla de médias, será utilizado um teste t *Student* bicaudal (Bussab e Morettin, 2006). Os valores calculados para os $p-values$ foram 0.0819 e 0.0713 para inteligência artificial e a coleção do TIL respectivamente. Deste modo, pode-se concluir que os especialistas não encontraram uma diferença estatisticamente significante na interpretação da rotulação com os dois diferentes métodos ao nível de 5% de significância. Isto é, pode-se aceitar a igualdade estatística de interpretação dos métodos com 95% de certeza.

Inteligência Artificial			*TIL*		
74 *notas*			98 *notas*		
m	$\overline{g}$	se	m	$\overline{g}$	se
P&U	2.7703	0.7860	*P&U*	2.7449	0.7365
RLDM	2.5811	0.9364	*RLDM*	2.6020	0.6996

Tabela 5.10: Médias da notas ($\overline{g}$) e seus desvios padrões (se) para cada método (m)

Ainda, alguns dos especialistas sugeriram que a presença de termos repetidos ao longo da taxonomia, como ocorre no método de Popescul & Ungar, facilita a interpretação, ampliando o raio de alcance da mesma. Então, os especialistas achavam que, provavelmente, essa seria a melhor escolha em um processo de recuperação de informação. Aparentemente, parece ser mesmo uma boa escolha, mas quando se analisa o grupo isoladamente, isto é, como se tratasse do resultado de um *flat cluster*. Se os grupos são completamente independentes, o resultado de uma expressão de busca como "document or text or sentence" seria capaz de recuperar os documentos de interesse, embora provavelmente viriam muitos documentos a mais.

5.6.3 Avaliação objetiva

Em uma hierarquia os termos devem ser mais específicos, ou todos os documentos da coleção seriam recuperados, isto é, ter-se-ia um *recall* ≈ 1 com uma precisão muitíssimo baixa. Para testar a hipótese de que gerar termos mais específicos é mais interessante para recuperar os documentos corretos, calcularam-se os pontos de F_{score} para as duas coleções de documentos, como mostrado na Figura 5.9. Para calcular os F_{score}, a precisão e a cobertura (*recall*), foram considerados apenas os dez rótulos mais freqüentes em cada lista de rótulos específicos. Com isso, tentou-se repetir como o especialista vê as listas de rótulos, os sete ou dez mais importantes; e, também utilizando-se o mesmo número de rótulos as comparações tendem as ser mais justas.

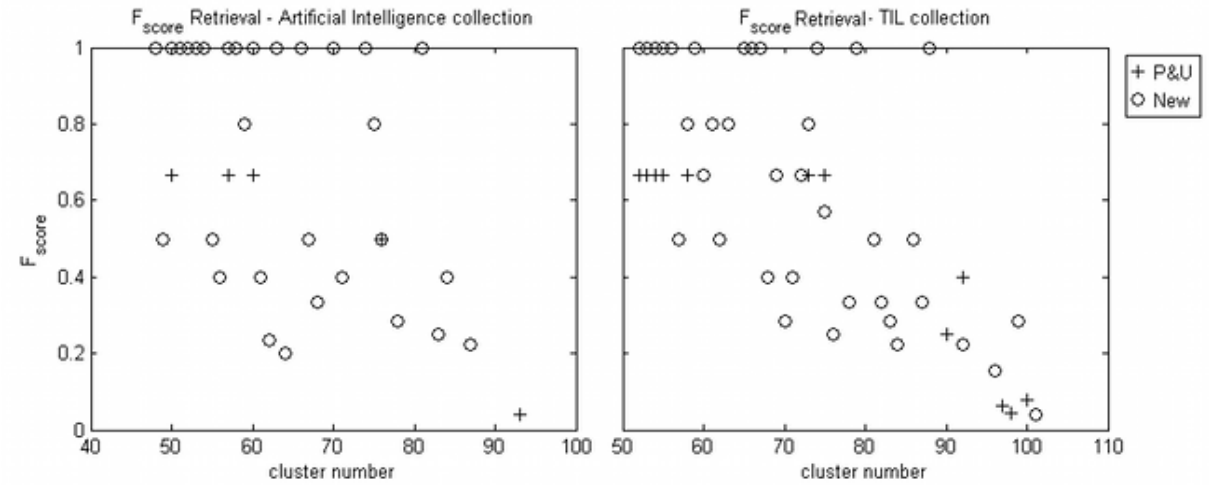

Figura 5.9: F_{score} do processo de recuperação usando apenas os 10 primeiros elementos das listas de rótulos específicas em cada coleção

Alguns pontos do F_{score} não puderam ser definidos, porque resultaram em divisão por zero. Isso ocorre quando recall e *precision* tendem a zero. Nos gráficos pode-se observar que as listas de rótulos específicas do RLDM provêem melhores resultados de busca que as listas específicas produzidas pelo método de Popescul & Ungar. O RLDM é responsável por 30 (65%) do total de pontos de F_{score} na coleção de inteligência artificial contra 5 (11%) dos pontos produzidos pelo método de Popescul & Ungar. Para a coleção do TIL, o RLDM foi capaz de representar 37 (74%) dos pontos contra 12 (24%) dos pontos produzidos pelo método de Popescul & Ungar. Observando o gráfico, o método proposto encontra as mais discriminativas listas de rótulos, pois consegue responder satisfatoriamente às expressões de busca em um número bem maior de vezes.

5.6.4 Considerações sobre a avaliação subjetiva integrada à objetiva

De acordo com essa avaliação, o RLDM alcançou os melhores resultados, pois foi comparável ao de Popescul & Ungar na avaliação subjetiva e obteve melhores resultados na avaliação objetiva. Por estar comparável ao método que nas avaliações anteriores esteve em primeiro lugar, parece que o RLDM conseguiu não perder em compreensibilidade, mantendo seu objetivos de não repetir termos ao longo da taxonomia.

Porém, apesar das boa avaliação subjetiva, ainda encontraram-se variáveis extras rotulação nos comentários dos especialistas. Novamente houve o comentário sobre a ausência do uso de termos compostos, n-gramas. Como provavelmente o uso de n-gramas adiciona informação semântica à abordagem *bag of words*, esse comentário tornou-se um requisito. Devido a esse requisito foi desenvolvido o trabalho de Moura et al. (2008d) para escolher bons atributos n-gramas e sem redundâncias entre os mesmos.

Dessa forma, as avaliações subjetivas sempre apontaram melhorias ao processo de obtenção das taxonomias de tópicos, mas não conseguiam medir de fato apenas a rotulação. Como se desejava saber se, além de recuperar melhor com a lista de termos mais específicos, também se gerariam melhores vocabulários, passou-se a trabalhar mais a questão de avaliação objetiva, e então se chegou ao modelo apresentado na Seção 5.4.2.

5.7 Resultados experimentais da avaliação objetiva

Quatro coleções de textos foram selecionadas para os experimentos, cada qual dividida em quatro categorias, de tamanhos diferentes, que foram utilizados como bases isoladas, logo, dezesseis coleções de textos foram utilizadas nos testes. As coleções passaram pelo mesmo processo de pré-processamento, seguido por processos idênticos de agrupamento hierárquico e depois pelos quatro métodos de rotulação. As análises dos resultados foram conduzidas em relação a cada um dos fatores considerados: (i) cobertura para a recuperação de informações a partir dos conjuntos de rótulos, (ii) cobertura de vocabulário de cada método e (iii) qualidade de vocabulário para cada método. As sub-seções a seguir, apresentam cada passo dos experimentos.

Deve-se salientar que, os resultados de comparações múltiplas de médias, apresentados nesta seção, foram realizados com o teste SNK (*Student-Newman-Keuls*), usando o quadrado médio do erro do modelo linear utilizado na análise de variância, com o grau de liberdade desse erro e o nível de significância de 5%. O teste foi escolhido devido a sua força e sua característica de sempre mostrar a real diferença entre as médias (Snedecor e Cochran, 1967). Nas tabelas apresentadas, diferentes letras na coluna de grupos correspondem a diferentes grupos de médias. Em cada sub-título das tabelas tem-se os graus de liberdade, a variância do erro utilizada e o número de nós envolvidos nos testes. Pois, os testes são realizados comparando-se as medidas em cada nó, ou seja, em cada grupo, mantido nas hierarquias rotuladas por diferentes métodos. Ainda, quando algum modelo de análise de variância apresentou uma matriz singular, impossibilitando a aplicação da comparação múltipla de médias a todos os métodos, isso foi indicado na tabela e, calculada a média da medida para o valor fora das comparações; como ocorrido para a medida F no sub-item 5.7.2.

5.7.1 Pré-processamento, agrupamento e rotulação

Na Tabela 5.11 é apresentado, de forma resumida, o resultado do pré-processamento das coleções utilizadas no experimento. As quatro coleções de textos podem ser acessadas

no *site* do projeto TopTax[4]. A primeira coleção de textos corresponde a artigos completos, em português, da segunda assembléia do Instituto Fábrica do Milênio (**IFM**[5]). No primeiro momento essa coleção tinha 614 documentos, dividida em quatro categorias, com 198 documentos na maior classe; as classes correspondem a projetos de trabalho do instituto (**WP01**, **WP02**, **WP03** e **WP04**). A segunda coleção de textos corresponde a artigos em inglês, 408 documentos, sobre computação, divididos nos temas: inteligência artificial (**AI**), interface homem máquina (**HCI**), *hardware* (**Hard**) e segurança e criptografia (**Sec**). A terceira coleção corresponde a 396 artigos sobre química, divididos nos temas: orgânica (**OrgC**), inorgânica (**InoC**), analítica (**AnaC**) e ciência dos polímeros (**PolyC**). A quarta coleção são artigos sobre física, divididos em biofísica (**BioP**), geofísica (**GeoP**), mecânica (**Mech**) e física quântica (**Quan**).

Na segunda coluna da Tabela 5.11 encontram-se os números de documentos utilizados em cada coleção após a conversão dos textos do formato original (PDF) para texto plano. Quando esse processo é aplicado, algumas vezes, alguns documentos são descartados, devido à conversão apresentar algum tipo de problema; fica-se apenas com os textos cujo conteúdo possa ser utilizado no processo. A seguir, aplicou-se *stemmização* às coleções, obtendo-se unigramas (número total em cada coleção na terceira coluna da tabela), com a ferramenta PreText (Matsubara et al., 2003). Para simplificar e padronizar o procedimento de filtro foi utilizado aproximadamente a indicação de Salton (Salton et al., 1975) para palavras discriminativas, que é usar as palavras que possuem freqüência de ocorrência nos documentos (*df: document frequency*) entre um a dez por cento da coleção; isto é, com $D = \{d_1, ..., d_x\}$, com x documentos, então cada termo (*stem* na tabela) que ficará no conjunto A precisa ter sua *df* no intervalo: $0.01x \leq df(termo) \leq 0.10x$. No entanto, como as coleções serão submetidas a um processo de agrupamento aglomerativo, usou-se no mínimo $df = 2$ e, para compensar a retirada de $df = 1$, a maior *df* foi acrescida de um ou dois pontos. Após a aplicação desse filtro, resumido na quarta coluna da tabela, tem-se a cardinalidade do conjunto A de atributos.

Ainda, para manter uma padronização de resultados, um agrupamento aglomerativo é aplicado a cada coleção, utilizando-se similaridade de cosseno e o algoritmo *average linkage*. Após esse procedimento, são aplicados os métodos de rotulação. Como eles geram diferentes conjuntos específicos de rótulos, em cada nó, de diferentes cardinalidades, fixou-se um número máximo de rótulos por nó, em catorze (14), uma escolha completamente aleatória. Para fazer isso, os rótulos obtidos em cada nó, para cada método, foram ordenados por freqüência de ocorrência, como no método dos Mais Freqüentes. Desse modo, cada método de rotulação gera, no máximo, catorze rótulos por $L_s(n_i)$ ordenados por freqüência.

5.7.2 Avaliando os rótulos como expressões de busca

Para essa avaliação consideraram-se os conjuntos de rótulos genéricos e os conjuntos de rótulos específicos separadamente, em diferentes experimentos. A idéia é avaliar a

[4]In: http://www.labic.icmc.usp.br/projects

[5]O IFM é uma organização brasileira cujas ações são focadas na busca por soluções de necessidades da manufatura nas indústrias; veja: http://www.ifm.org.br/

TC	#docs	#stems	filtro com DF	$card(A)$
WP01	80	10268	$2 \geq df \leq 10$	4799
WP02	65	11630	$2 \geq df \leq 9$	5000
WP03	198	19904	$2 \geq df \leq 21$	9302
WP04	48	13870	$2 \geq df \leq 7$	5252
Hard	89	11014	$2 \geq df \leq 10$	3453
AI	94	10128	$2 \geq df \leq 11$	4445
HCI	98	8915	$2 \geq df \leq 12$	3971
Sec	127	18780	$2 \geq df \leq 14$	6253
AnaC	100	18408	$2 \geq df \leq 12$	6645
InoC	97	18436	$2 \geq df \leq 12$	8864
OrgC	99	20116	$2 \geq df \leq 12$	5238
PolyC	100	11628	$2 \geq df \leq 12$	5634
BioP	95	13179	$2 \geq df \leq 12$	6187
GeoP	97	12011	$2 \geq df \leq 12$	5100
Mech	117	12660	$2 \geq df \leq 14$	4378
Quan	82	9586	$2 \geq df \leq 10$	3798

Tabela 5.11: Número de documentos e atributos das coleções de textos utilizadas

utilização dos conjuntos específicos como expressão de busca, isto é, cada $L_s(n_{ij})$ como se o grupo fosse resultado de um "*flat cluster*" e, então comparar os resultados obtidos por cada método de rotulação. A seguir, utilizaram-se os rótulos genéricos $L_g(n_i)$ junto ao rótulo específico em cada filho de n_i, isto é, utilizou-se a expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$ para cada n_{ij}; essa segunda expressão permite que se avalie a rotulação hierárquica propriamente dita, pois reflete o fato de que os rótulos genéricos dos antecessores discriminam o nó atual.

Na Tabela 5.12 é apresentado o resumo dos resultados desse primeiro experimento para a coleção de textos **WP01**. Nessa tabela estão presentes os resultados das comparações múltiplas de médias para cada medida de interesse: F, média harmônica de precisão e revocação; *prec*, precisão; r, revocação ou cobertura (*recall*); e, F_a, falso alarme. Para obter as tabelas foi utilizado o modelo linear generalizado especificado na equação 5.3. Para cada medida ajustou-se um modelo linear generalizado dessa forma. Ainda, os mesmos cálculos e avaliações foram realizados para as dezesseis coleções de textos, relacionadas no Apêndice A; e, o comportamento observado para as medidas nas dezesseis coleções foi semelhante.

Deve-se notar na Tabela 5.12 que, para algumas medidas, como a F e a precisão *prec* não há o mesmo número de pontos calculados, para cada nó (n, na tabela); o que ocorre devido a alguma divisão por zero, por exemplo. Ainda, na tabela, para a expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$, a medida F do RLUM não pode ser comparada às demais, por insuficiência de pontos nos resultados dos demais métodos. Porém, nota-se que o resultado de F para RLUM foi 1, nos trinta pontos totais, ou seja, o resultado atingiu o valor ideal. Nota-se, na tabela, que o RLUM só perde para o RLDM na precisão, quando utilizada a expressão de busca $L_s(n_{ij})$, mas em todos os demais casos seu desempenho é melhor que dos outros três. Quanto ao falso alarme, o comportamento do RLUM é sempre aceitável, comparável ao outros; porém o RLDM erra mais.

Outros resultados interessantes referem-se, por exemplo, aos métodos dos Mais Freqüentes e de Popescul e Ungar, que sempre apresentam um comportamento semelhante. O método RLDM tem um melhor desempenho, no que tange à precisão e "*F measure*",

modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 1$				$L_s(n_{ij})$ $\sqrt{e} = 0.0282, gl = 60$			
método	n	F	*grupo*	método	n	F	*grupo*
MostFreq	1	0.697248	*a*	*RLDM*	30	0.807975	*a*
PopeUngar	1	0.697248	*a*	*RLUM*	30	0.701492	*b*
RLDM	1	0.367816	*b*	*PopeUngar*	17	0.401681	*c*
$\overline{F}_{RLUM}$	30	1.000000	*singular*	*MostFreq*	17	0.401681	*c*
modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0179, gl = 64$				$L_s(n_{ij})$ $\sqrt{e} = 0.0371, gl = 89$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
RLUM	30	0.590959	*a*	*RLDM*	31	0.801459	*a*
RLDM	6	0.166667	*b*	*RLUM*	30	0.579107	*b*
PopeUngar	31	0.032258	*c*	*PopeUngar*	31	0.197366	*c*
MostFreq	31	0.032258	*c*	*MostFreq*	31	0.197366	*c*
modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0007, df = 89$				$L_s(n_{ij})$ $\sqrt{e} = 0.0541, df = 89$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
RLUM	30	1.000000	*a*	*RLUM*	30	1.000000	*a*
PopeUngar	31	0.017265	*b*	*RLDM*	31	0.848667	*b*
MostFreq	31	0.017265	*b*	*PopeUngar*	31	0.326405	*c*
RLDM	31	0.007269	*b*	*MostFreq*	31	0.326405	*c*
modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0042, df = 89$				$L_s(n_{ij})$ $\sqrt{e} = 0.0041, df = 89$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
MostFreq	31	0.064665	*a*	*MostFreq*	31	0.065600	*a*
PopeUngar	31	0.064665	*a*	*PopeUngar*	31	0.065600	*a*
RLUM	30	0.056270	*a*	*RLUM*	30	0.059169	*a*
RLDM	31	0.003314	*b*	*RLDM*	31	0.009922	*b*

Tabela 5.12: Resultados das medidas de interesse para **WP01**

quando se usam os conjuntos de rótulos específicos como expressão de busca, obtendo-se resultados melhores que os demais métodos. Já o método RLUM sempre tem um excelente desempenho no valor de *recall*, próximo a um, um bom índice da "*F measure*" e um falso alarme dentro do esperado. Seu desempenho é sempre superior ou, no mínimo, equivalente aos demais, quando essas medidas são boas para os demais e a expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$ é usada; logo, estatisticamente, ele obtém os conjuntos de rótulos mais representativos quando toda a hierarquia é considerada.

Reforçando essas observações, os resultados das medidas *recall*, *F-measure* e *precision* para as dezesseis coleções de textos estão resumidos nos gráficos a seguir. Primeiramente, observe os Gráficos 5.10 e 5.11. Usando-se os conjuntos de rótulos específicos de n_{ij} ($L_s(n_{ij})$) como a expressão de busca, isto é, considerando cada grupo como um "*flat cluster*", o RLDM e o RLUM tem uma boa performance. Os dois métodos tem seus valores de F próximos a um. Porém os métodos de Popescul & Ungar e o Mais Freqüentes apresentam uma grande variação da F para algumas coleções de textos. Por outro lado, usando a expressão de busca como os conjuntos de rótulos genéricos $L_s(n_{ij}) \bigcup L_g(n_i)$, o RLUM apresenta o mesmo comportamento para a F, o que era esperado. O RLDM, por sua vez, apresenta um grande decréscimo da F. O Popescul & Ungar e o Mais Freqüentes novamente apresentam comportamento similar, embora também apresentem um decréscimo da F. Tomando apenas essa medida de interesse (a F), tem-se uma boa evidência de que

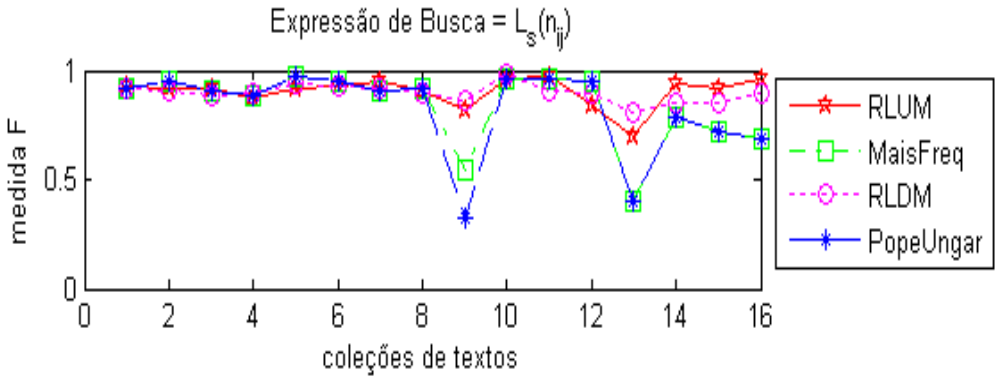

Figura 5.10: Medida F – $L_s(n_{ij})$

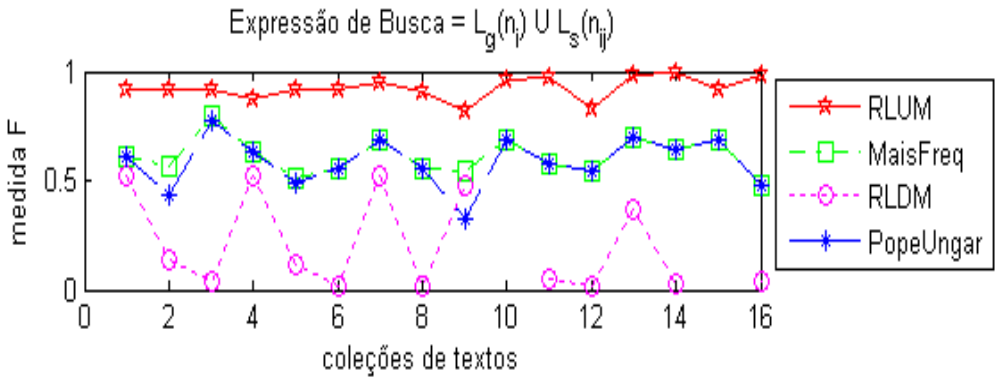

Figura 5.11: Medida F – $L_g(n_i) \bigcup L_s(n_{ij})$

o RLUM produza as melhores taxonomias.

A segunda medida de interesse para esse tipo de comparação é a *recall*. Se os conjuntos de rótulos construídos são bons, eles devem ser representativos do nó. Deste modo, espera-se que os documentos recuperados por cada conjunto de rótulos sejam o mais próximo do conjunto de documentos do grupo que eles representam; isto é, que a *recall* seja próxima a um. No Gráfico 5.12 os resultados das médias para a medida *recall* usando-se a expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$ podem ser observados. Usando esse conjunto genérico de rótulos, a categoria do documento na hierarquia deveria ser refletida nos resultados obtidos. Como esperado o RLUM obteve os melhores resultados, porém os demais métodos apresentaram resultados bastante ruins. O Popescul & Ungar e os Mais Freqüentes estão sempre com resultados similares, porém ruins. Para o RLDM esperavam-

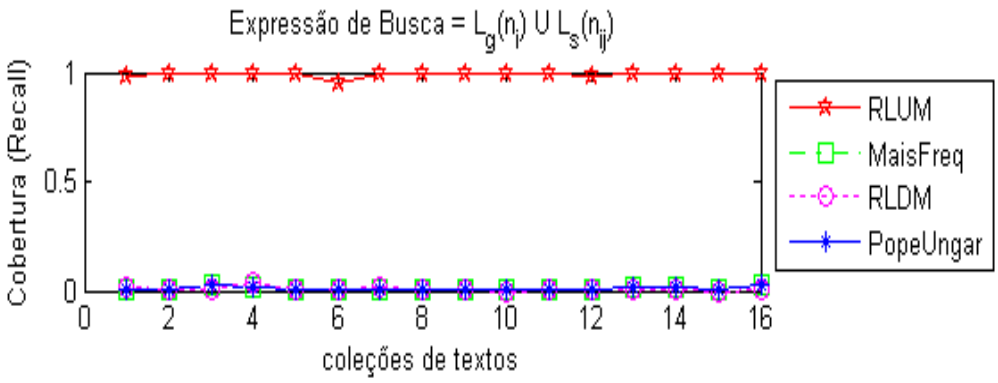

Figura 5.12: Medida *Recall* – $L_g(n_i) \bigcup L_s(n_{ij})$

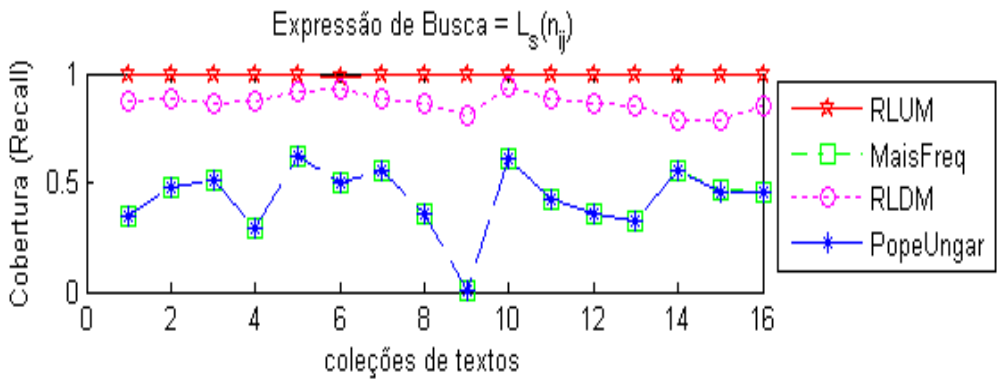

Figura 5.13: Medida *Recall* – $L_s(n_{ij})$

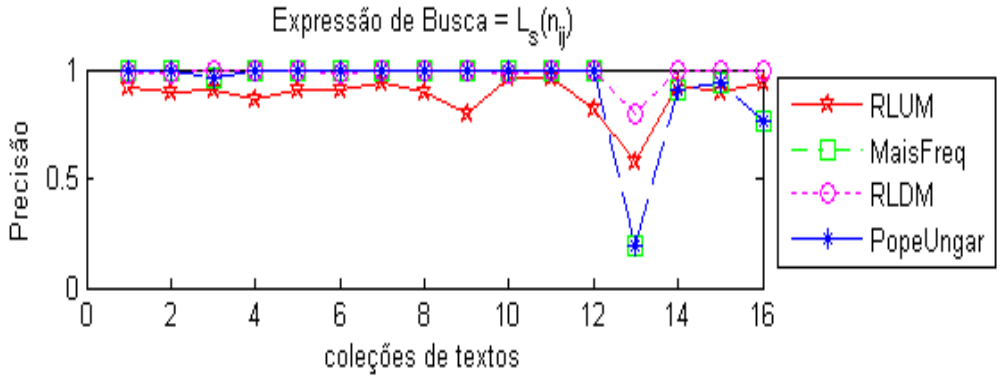

Figura 5.14: Precisão – $L_s(n_{ij})$

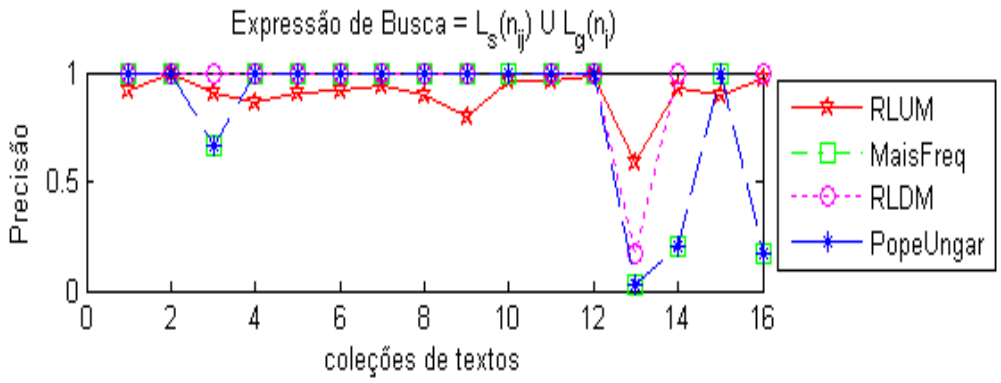

Figura 5.15: Precisão – $L_g(n_i) \bigcup L_s(n_{ij})$

se melhores resultados, mas provavelmente a propagação de erros nele presente prejudicou seu desempenho. Por outro lado, usando apenas os conjuntos de rótulos específicos, pode-se observar uma melhora significativa nos resultados desses três últimos métodos, dado que o RLUM sempre mantém seus resultados altos; como pode ser observado na Figura 5.13. Ou seja, quando a expressão de busca refere-se a um grupo isolado, os quatro métodos tem bons resultados. Mas, também se deve notar que houve aumento da variância, ou seja, os gráficos apresentam maior dispersão entre os pontos. Como esperado o Mais Freqüentes é bom para grupos isolados, porém o Popescul & Ungar acabou apresentando comportamento similar. O RLDM apresentou bons resultados, provavelmente porque a propagação de erros influenciou menos os grupos isoladamente. Salienta-se o resultado de que o RLUM produz os melhores rótulos na hierarquia, dado que seus resultados de "*recall*" para a expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$ são estatisticamente iguais a um, e, também para os grupos isoladamente, com a expressão $L_s(n_{ij})$.

Os outros três métodos não apresentaram resultados tão ruins para a medida F devido ao comportamento da precisão, de acordo com os resultados apresentados nas Figuras 5.14 e 5.15. Embora a precisão para os outros três métodos pareça ser melhor que para o RLUM, exceto em algumas das coleções de textos, isso não mostra que produzam conjuntos de rótulos mais significativos. A precisão pode ser boa mesmo quando a cobertura (recall) é ruim, pois recuperar um único documento correto significa ter uma precisão de cem por cento, porém, se o total de documentos no grupo é de 100, recuperou-se apenas um por cento do grupo.

5.7.3 Avaliando a cobertura do vocabulário

Como há uma indicação de que o método RLUM produz um vocabulário maior que o dos outros métodos, espera-se que este método consiga melhor cobrir um vocabulário e domínio, já que se o número é maior espera-se que as chances dos termos de domínio estarem nele contidos também sejam maiores. Assim, nesta sub-seção verifica-se a hipótese:

$$H_0 : card(V_U) \geq card(V_P) \geq card(V_M) \geq card(V_D)$$

Pois aparentemente o RLDM gera o menor vocabulário e os outros dois geram vocabulários de tamanhos semelhantes, como observado na Tabela 5.13. Como também se deve considerar o fato de que os conjuntos de atributos iniciais têm diferentes tamanhos, como observado na Tabela 5.11, faz-se necessário retirar o efeito dessa covariável do modelo, quando definida a equação 5.4.

Após o ajuste do modelo linear generalizado (equação 5.4) e então realizada a comparação múltipla de médias das cardinalidades finais obteve-se o resultado apresentado na Tabela 5.14. Pode-se notar que houve o mesmo número de observações para cada método, isto é, a cardinalidade foi observada e considerada para as dezesseis bases. A cardinalidade média para cada método, $\overline{card(V_m)}$, foi maior para o RLUM ($card(V_U)$), com 0.95 de certeza. Como esperado as cardinalidades do Popescul e Ungar ($card(V_P)$) e dos Mais Freqüentes $card(V_M)$ são estatisticamente iguais. Porém, estatisticamente, já não se observa que o RLDM ($card(V_D)$) apresente uma cardinalidade mais baixa que

TC	$card(V_P)$	$card(V_M)$	$card(V_D)$	$card(V_U)$
WP01	645	645	488	927
WP02	566	566	690	867
WP03	1383	1384	1595	2245
WP04	495	495	558	739
Hard	836	824	478	1000
AI	747	746	665	928
HCI	808	801	597	937
Sec	1116	1116	766	1359
AnaC	915	914	838	1244
InoC	803	798	974	1120
OrgC	1051	1030	512	1309
PolyC	786	784	757	967
BioP	759	754	799	977
GeoP	825	824	793	1032
Mech	937	937	691	1179
Quan	719	716	607	912

Tabela 5.13: Cardinalidade dos vocabulários obtidos

Teste SNK para a cardinalidade final, com rejeição=0.05			
Com 59 graus de liberdade, variância média= 42482.6158			
método	observações	$\overline{card(V_m)}$	grupo
RLUM	16	1108.875000	*a*
PopeUngar	16	836.937500	*b*
MaisFreq	16	833.375000	*b*
RLDM	16	738.000000	*b*

Tabela 5.14: Comparação múltipla de médias para cardinalidades dos vocabulários

os demais, ou seja, ele pode ser considerado como obtendo um vocabulário tão variado quanto os métodos de Popescul e Ungar e o dos Mais Freqüentes.

5.7.4 Avaliação da representatividade do vocabulário produzido

Como dito anteriormente, apenas a cardinalidade do método ser superior à dos demais não garante que a qualidade do vocabulário produzido seja boa, pois ele pode ser resultado de uma série de más escolhas. Embora a medida "*recall*" já indique se as escolhas foram ou não boas para a formação dos conjuntos de rótulos de cada grupo. Assim, nesta fase do experimento vamos utilizar a EMIM para testar a hipótese de que a qualidade do vocabulário produzido pelo RLUM é superior a dos outros três métodos, ou seja:

$$H_0 : EMIM(RLUM) \geq EMIM(P) \approx EMIM(M) \geq EMIM(D)$$

Para verificar essa hipótese, calcula-se a EMIM em cada nó para cada taxonomia de tópicos resultante de cada método e, então, compara-se a EMIM dos nós correspondentes em cada resultado utilizando o modelo linear generalizado, definido pela equação 5.5; então se realiza a análise de variância e a comparação múltipla das médias com o teste SNK, com 0.95 de certeza. A Tabela 5.15 foi obtida para essa comparação a partir das bases do domínio de computação. Os demais resultados para as outras bases encontram-se no Anexo A e repetem um comportamento muito similar a estes.

Deve-se observar, na Tabela 5.15 que todos os valores são calculados, tem-se sempre o mesmo número de grupos (ou nós), n, para cada média calculada. Percebe-se também que

Col. de Textos AI				*Col. de Textos HCI*			
$gl = 147, var.media = 160140.0445$				$gl = 108, var.media = 286839.4677$			
método	*n*	*emim*	*grupo*	*método*	*n*	*emim*	*grupo*
RLUM	50	1089.899965	*a*	*RLUM*	37	885.627493	*a*
PopeUngar	50	893.817625	*b*	*PopeUngar*	37	765.692200	*a*
MostFreq	50	893.742870	*b*	*MostFreq*	37	762.961843	*a*
RLDM	50	557.042183	*c*	*RLDM*	37	373.175243	*b*
Col. de Textos Hard				*Col. de Textos Sec*			
$gl = 36, var.media = 185914.7144$				$gl = 159, var.media = 487370.7896$			
método	*n*	*emim*	*grupo*	*método*	*n*	*emim*	*grupo*
RLUM	13	1104.943006	*a*	*RLUM*	54	1613.500940	*a*
PopeUngar	13	909.675011	*a*	*PopeUngar*	54	1337.548719	*a*
MostFreq	13	906.329243	*a*	*MostFreq*	54	1337.383964	*a*
RLDM	13	77.376233	*b*	*RLDM*	54	636.411828	*b*

Tabela 5.15: EMIM - resultados da comparação múltipla de suas médias

o agrupamento esperado de médias se confirma nas quatro bases apresentadas (bem como se confirma nos resultados apresentados no Apêndice A), ou seja, a hipótese de qualidade dos vocabulários, de acordo com a EMIM, é confirmada com 0.95 de certeza. Assim, além do RLUM apresentar a maior variabilidade de termos no vocabulário produzido, ele também garante a melhor qualidade desses termos, em relação aos três métodos comparados. Também se deve observar que os métodos de Popescul e Ungar e dos Mais Freqüentes sempre se encontram nos mesmos grupos, o que mostra que seus resultados são sempre muito próximos devido a lacuna do processo decisório no método de Popescul e Ungar, discutido anteriormente. Ainda, o RLDM tende a produzir um vocabulário mais pobre, em termos de cardinalidade, e de menor qualidade.

5.8 Considerações finais

Neste capítulo foi apresentada a evolução do método de rotulação proposto para agrupamentos hierárquicos de documentos e uma proposta de validação objetiva do método. Desde a primeira proposta, com o RLDM procurou-se produzir rótulos discriminativos dos grupos, sem repetição ao longo dos ramos, com um vocabulário variado e de qualidade. Porém, apesar de ter sido bem avaliado (Moura et al., 2008c), o RLDM apresentava uma complexidade computacional muito alta e uma propagação de erros que poderia ser minimizada.

Em um primeiro momento, o novo método era uma proposta de diminuição da complexidade do RLDM, objetivo imediatamente atingido, pois o RLDM apresenta complexidade da ordem de $O(mz^2)$ e o RLUM apresenta complexidade linear, $O(mz)$, considerando-se m o número de atributos e z o número de nós da hierarquia. O RLUM diminuiu a propagação de erros ao longo da tomada de decisão sobre os atributos selecionados como rótulos, porque constrói os conjuntos de rótulos das folhas para a raiz, enquanto lhes atualiza.

A melhor discriminação dos conjuntos de rótulos foi verificada utilizando-se os mesmos como expressões de busca sobre as coleções de textos e verificando a cobertura ("*recall*") dos resultados. O RLUM apresentou a melhor cobertura de recuperação, estatisticamente mais significante, com 0.95 de certeza, que os demais em todos os casos testados, para as

dezesseis coleções de textos.

A ordem de grandeza dos vocabulários produzidos foi testada comparando-se a cardinalidade dos vocabulários produzidos pelos quatro métodos de rotulação, em dezesseis diferentes coleções de textos. O RLUM obteve uma maior variabilidade de vocabulário em todos os testes. E, quando testada a qualidade desse vocabulário, contra a dos obtidos pelos demais métodos, também o RLUM superou os demais, ou no mínimo igualou-se a eles.

Assim, todas as hipóteses iniciais sobre o RLUM e suas relações com os métodos de Popescul e Ungar, dos Mais Freqüentes e do RLDM foram validadas como verdadeiras, ou seja, o método atingiu todos os objetivos inicialmente propostos.

Como comentado no Capítulo 3 e visto na Seção 5.6, em outros experimentos realizou-se uma validação subjetiva de métodos de rotulação junto a especialistas de domínio (Moura e Rezende, 2007; Moura et al., 2008c), com o intuito de avaliar o resultado da taxonomia obtida nó a nó. Uma série de fatores interfere nesse processo de validação, desde a forma de obtenção dos termos até o tamanho final das taxonomias, o que acaba gerando limitações e não permite que se tenha grande credibilidade nos resultados das avaliações. Isto é, as avaliações não refletem exclusivamente o método de rotulação, acabam refletindo outras variáveis subjetivas, que não são fáceis de identificar. Assim, como trabalhos futuros, pretende-se utilizar o RLUM em um processo de construção de taxonomias de tópicos e validar o processo contra taxonomias *gold*, isto é, taxonomias pré-existentes e já bem aceitas, como realizado por outros autores ((Kashyap et al., 2005a),(Krishnapuram e Kummamuru, 2003)). Espera-se com esse tipo de validação conseguir simular uma validação subjetiva exclusivamente do método de rotulação.

Capítulo

6

Conclusões e trabalhos futuros

O objetivo principal desta tese de doutorado é a construção de taxonomias de tópicos em domínios restritos. A metodologia desenvolvida recebeu o nome de TopTax e é uma instanciação do processo de mineração de textos. A partir dessa instanciação foram propostas soluções específicas direcionadas ao objetivo de construir taxonomias. Essas propostas contribuem para o estado da arte em algumas tarefas de mineração de textos, bem como satisfazem aos objetivos específicos e às avaliações das hipóteses inicialmente elencadas. Além disso, a metodologia foi empregada em problemas práticos, tendo atingido um bom nível de aceitação e resultados. Assim, neste capítulo resumem-se todas as contribuições desta tese por meio dos resultados alcançados para cada objetivo específico e hipóteses validadas. Também relacionam-se as principais limitações atuais ao uso da TopTax e suas ferramentas e, finalmente, apresenta-se uma lista de trabalhos futuros, dentre os quais alguns já em andamento.

6.1 Contribuições do trabalho

As contribuições desta tese dividem-se em três grupos. O primeiro grupo, trata das contribuições diretas desta tese de doutorado, que vão de encontro aos objetivos mais amplos inicialmente traçados e alguns dos objetivos específicos diretamente ligados a estes. No segundo grupo estão as contribuições em parceria no grupo de pesquisa, que focam vários dos objetivos específicos iniciais. Por último, o terceiro grupo corresponde às aplicações reais da metodologia TopTax durante esse período, que serviram para validar o seu projeto e verificar sua utilidade prática.

6.1.1 Contribuições diretas desta tese

A primeira grande contribuição é o desenvolvimento da metodologia para a construção de taxonomias de tópicos em um domínio (Moura et al., 2008a,b), denominada TopTax-

Topic Taxonomy. A metodologia é um guia de construção de taxonomias, cobrindo desde a seleção dos textos utilizados até a avaliação dos resultados obtidos, fornecendo soluções automatizadas e permitindo a intervenção do especialista de domínio em cada etapa. Assim, o primeiro objetivo do trabalho foi satisfeito, que corresponde à construção de uma metodologia, utilizando técnicas de mineração de textos, para auxiliar um especialista a organizar hierarquicamente a informação textual do domínio e fornecendo a cada etapa critérios que o auxilie a intervir no processo de modo a melhorar os resultados obtidos.

A segunda contribuição remete ao estado da arte em seleção de atributos, correspondendo ao modelo de seleção de atributos n-gramas (Moura et al., 2008d). Embora existam recomendações de seleção de n-gramas como candidatos a colocações de domínio há poucos trabalhos tratando esse problema como seleção de atributos e, conseqüentemente, poucos trabalhos deram importância à remoção de redundâncias. Com o modelo proposto, selecionam-se n-gramas com maior importância semântica e removem-se as redundâncias entre todos eles. Os experimentos realizados mostraram que o espaço de atributos não perde em discriminação com a redução das redundâncias. Essa constatação valida uma das hipóteses inicialmente colocadas de que a redução do número de atributos não afeta negativamente o resultado da extração de conhecimento. Ainda, essa contribuição vai de encontro aos dois primeiros objetivos específicos inicialmente traçados, que são: obter um conjunto de métodos e parâmetros para identificar e delimitar o vocabulário do domínio presente na coleção de textos; e, obter métodos adequados de filtro de atributos para diminuir a característica esparsa da matriz de dados, procurando não perder representatividade. Lembrando que o conjunto de atributos selecionados corresponde à primeira estimativa do vocabulário do domínio presente na coleção de textos.

A terceira contribuição para o estado da arte encontra-se na forma de rotular os grupos obtidos. Embora haja muito estudo em métodos e algoritmos de agrupamentos de documentos há poucas soluções focadas em rotulação dos grupos. Os rótulos auxiliam a interpretação dos grupos e seus conjuntos dão o embasamento à construção de uma taxonomia de tópicos. Obter bons descritores para o agrupamento é um desafio e, nesta tese, são propostos dois modelos para esse fim. O primeiro modelo apresenta bons resultados embora seja computacionalmente bastante complexo (Moura et al., 2008c). O segundo modelo proposto tem baixa complexidade computacional e produz um melhor conjunto de rótulos, tanto do ponto de vista de recuperação de informação quanto da variabilidade do vocabulário utilizado (Moura e Rezende, 2009b,a). Com esses dois modelos conseguiu-se desenvolver mais de um método eficiente de seleção de termos discriminativos para cada agrupamento obtido, representando-os de maneira única. Assim, validou-se a hipótese de que se consegue representar unicamente cada grupo obtido por meio da seleção dos termos discriminativos.

Ainda, do modelo de rotulação decorre uma outra contribuição, que é a forma de avaliar os resultados. Foram propostas duas formas de avaliação subjetiva, objetiva e o uso das duas conjuntamente. Iniciou-se pelo uso da avaliação subjetiva, que obteve resultados bastante interessantes. Porém esses resultados não refletiram apenas a rotulação, sofreram a influência de outras variáveis difíceis de serem isoladas (Moura e Rezende, 2007). Assim, numa segunda proposição de avaliação integraram-se a avaliação subjetiva

e objetiva pelos critérios de recuperação de informação (Moura et al., 2008c). Essa segunda forma de avaliação produziu resultados mais facilmente interpretáveis, mas ainda não isolou completamente apenas a análise dos métodos de rotulação. Assim, decidiu-se por investir em um modelo mais completo de avaliação objetiva, utilizando critérios de recuperação de informação e qualidade do vocabulário produzido (Moura e Rezende, 2009c). Esse último modelo de avaliação foi o mais satisfatoriamente empregado para validar a hipótese de representar unicamente cada grupo.

6.1.2 Contribuições em parcerias

Alguns dos objetivos específicos foram atingidos em parceria com mestrandos e alunos de graduação em projetos de conclusão de curso e de iniciação científica.

Um dos objetivos específicos era desenvolver soluções de avaliação visual para os grupos, permitindo explorá-los e estudá-los. Apenas para visualizá-los e explorá-los foi desenvolvida uma ferramenta denominada TaxTools em um projeto de conclusão de curso de graduação (Marcacini, 2008). Outra contribuição importante refere-se à exploração visual das taxonomias obtidas em diferentes níveis de granularidade, com base na hipótese inicial de que coleções textuais geram grandes hierarquias que precisam ser cortadas. Para isso foram adaptadas técnicas de corte de cluster ao processo de visualização no trabalho de Marcacini et al. (2009). Com isso, permite-se que o especialista decida se deseja reduzir a taxonomia obtida, ou apenas navegá-la em diferentes níveis, satisfazendo a hipótese de que visualizar as hierarquias em diferentes níveis de granularidade auxilia a compreensão da coleção textual.

O objetivo específico de obter métodos adequados de filtro de atributos para diminuir a característica esparsa da matriz de dados, procurando não perder representatividade, também foi foco de um trabalho de mestrado (Nogueira, 2009). Esse mestrado foi desenvolvido tendo como um de seus objetivos contribuir com a TopTax. Em Nogueira et al. (2008a), foram propostas algumas adaptações a métodos de filtro comumente utilizados, comparados seus desempenhos e proposto o uso de alguns deles. Ocorreu também um primeiro estudo dos filtros, frente a outros problemas de pré-processamento, no trabalho de Nogueira et al. (2008b). Nesses dois trabalhos a seleção não supervisionada de atributos concentra-se em atributos compostos por uma única palavra (unigrama). Conseguiu-se observar que a redução do número de atributos não prejudica o desempenho dos algoritmos de extração de conhecimento, mas não se validou a hipótese de que tenha havido melhora de desempenho.

Outro objetivo específico é obter um conjunto de métodos e parâmetros para identificar e delimitar o vocabulário do domínio presente na coleção de textos. Para tal, tem-se como a hipótese que existe alguma representação dos atributos que uma avaliação subjetiva possa julgar como melhor. Atingir esse objetivo e tentar validar essa hipótese foi o tema de um outro mestrado (Conrado, 2009), no qual os estudos e desenvolvimentos tinham como um de seus objetivos contribuir com a TopTax. Primeiramente, é necessário buscar por uma forma de representação dos atributos, que seja adequada ao problema e, então, selecionar os atributos. Em relação à forma de representação dos atributos,

foram realizados alguns estudos (Conrado et al., 2008, 2009). Desses estudos, o que se pode observar é que o uso de métodos híbridos para a geração de atributos parece indicar um caminho de trabalho futuro mais promissor, pois, subjetivamente, os resultados de métodos híbridos satisfazem melhor aos especialistas.

E, ainda, um trabalho de iniciação científica foi inspirado nas necessidades de preparação dos dados textuais observadas durante as aplicações da TopTax. Desse trabalho resultou uma ferramenta de extração de informação, a IESystem (Rossi et al., 2009), capaz de trabalhar com dados textuais de diferentes fontes de informação, isto é, em vários formatos e em diferentes línguas. Com a ferramenta constrói-se um modelo de dados e ela transforma os arquivos de entrada em arquivos textos com marcação de metadados em XML.

6.1.3 Aplicações práticas da TopTax

A metodologia TopTax foi utilizada em alguns projetos práticos, o que auxiliou a validação e evolução da mesma. A organização da biblioteca digital do Instituto Fábrica do Milênio - IFM foi auxiliada pelo uso da metodologia Marcacini et al. (2007), tanto para a organização quanto para melhoria dos processos de recuperação de informação. A TopTax também foi empregada em um projeto de gestão de competências junto à Embrapa Pecuária Sudeste (Mazzari, 2007), no qual ela foi utilizada como ferramenta de análise exploratória para estudar as tendências em pesquisa e desenvolvimento da empresa no tema gado de corte e leite. A TopTax também foi utilizada em mais um projeto de gestão de competências e conhecimento (Marcacini et al., 2008) e depois em suporte à gestão do conhecimento na área de engenharia de produção (Rezende et al., 2008).

Essas aplicações práticas foram muito úteis para observar o quanto as soluções empregadas aproximaram-se dos objetivos específicos iniciais de auxílio à organização de bases de dados textuais, identificando-se tópicos e palavras-chaves para cada grupo e documento. Assim, a hipótese inicial de que a metodologia validada e avaliada atingisse o objetivo de auxílio à organização da informação, permitindo realizar essa tarefa com eficácia não chegou a ser provada, mas foi experimentada. E a hipótese de que, pelo menos, metadados são obtidos e podem ser usados para classificar os documentos e, conseqüentemente auxiliar o especialista em organização de informação, foi confirmada na prática.

6.2 Principais limitações

A configuração atual do ambiente de suporte computacional à metodologia permite utilizá-la e produzir bons resultados, como visto nos capítulos anteriores. Entretanto, o objetivo do desenvolvimento dessas ferramentas é testar os modelos propostos; logo, elas ainda são protótipos. Assim, é necessário investir na transformação desses protótipos em ferramentas robustas, para trabalhar com grandes bases de dados e, efetivamente, cumprir o objetivo de auxílio ao especialista de domínio.

Além disso, o método de seleção de n-gramas precisa ser expandido para lidar com n-gramas de maior ordem, $n > 3$, e, conseqüentemente, também todos os seus algoritmos e implementações.

6.3 Trabalhos futuros

Precisa-ses obter uma avaliação formal para a metodologia TopTax. Embora, na prática, tenha-se confirmado sua utilidade, falta-lhe uma avaliação formal. Cada uma das etapas foi avaliada independentemente, com o objetivo de avaliar cada nova técnica colocada à disposição no ambiente da TopTax. Assim, é necessário investir esforços em realizar experimentos com as várias combinações de técnicas em vários domínios. Esses experimentos precisam contar apenas com avaliações objetivas, pois seria muito oneroso envolver especialistas na avaliação de um elevado número de taxonomias. Para viabilizar uma análise objetiva detalhada, quase similar a uma subjetiva, vem-se desenvolvendo uma metodologia de avaliação por meio de taxonomias *gold*. Uma taxonomia *gold* é uma taxonomia já bem aceita em uma comunidade. Para isso, vem-se coletando taxonomias *gold* de diferentes domínios vinculadas a coleções de documentos e implementando alguns algoritmos de comparação para executar esses experimentos.

Outros trabalhos futuros derivam-se diretamente dos trabalhos em desenvolvimento e geraram dois projetos de mestrado, já em andamento. O primeiro trata do processo de crescimento das taxonomias de tópicos, isto é, da atualização das taxonomias por meio da inserção e remoção de documentos. Nesse trabalho serão avaliadas técnicas existentes e provavelmente proposto algum ajuste ou integração das mesmas. O segundo mescla um pouco de geração e seleção de atributos, trabalhando com regras de associação para gerar novos atributos. Nesse trabalho será avaliada a redução do espaço de características e se há melhora significativa na interpretação dos atributos.

Mais dois trabalhos futuros em direção ao estado da arte estão inspirados diretamente na metodologia e já em andamento no formato de dois projetos de doutorado. O primeiro, também na linha de crescimento das taxonomias de tópicos, investigará técnicas de crescimento das taxonomias, devendo produzir um novo modelo adaptado às necessidades da TopTax e melhorado em relação aos já existentes. O segundo, segue a linha de geração de atributos, dando continuidade ao trabalho de mestrado de Conrado (2009). Nesse trabalho, serão explorados modelos híbridos de geração de atributos. Os modelos híbridos misturam técnicas puramente estatísticas de extração de atributos a técnicas de lingüística computacional. Assim, espera-se incrementar o uso de modelos híbridos de extração de atributos na TopTax.

Ainda em andamento, há um trabalho numa linha um pouco diferente do paradigma de aprendizado de máquina atual, que é o estatístico. Nesse vem-se trabalhando com lógica *fuzzy* na organização e recuperação de informação. Pretende-se com isso avaliar o uso de *soft cluster* na metodologia, porém dentro do paradigma de aprendizado nebuloso, especialmente na etapa de extração de conhecimento por meio de agrupamento *fuzzy* e no pós-processamento avaliando-se as melhorias do processo de busca. Para validar o trabalho, será também interessante utilizar cluster hierárquico probabilístico para gerar as taxonomias, embutindo-lhes a tarefa de rotulação, como idealizado por Hofmann (1999), e adaptar o processo de busca a utilizar as probabilidades estimadas cada cada grupo em lugar dos graus de pertinência dos agrupamentos *fuzzy*.

Na linha de experimentos sobre as técnicas já desenvolvidas, pretende-se expandir o

modelo de seleção de n-gramas para trabalhar com $n > 3$. O novo modelo será experimentado na seleção inicial de atributos, na etapa de pré-processamento, e, na etapa de pós-processamento, após a seleção dos rótulos de cada agrupamento. Com esses experimentos, pretende-se avaliar em qual etapa é mais vantajoso utilizá-lo. E, ainda, como trabalho futuro sobre esse modelo, seria interessante comparar os resultados com a indicação de janelas entre os n-gramas gerados, como feito por Zhang e Zhu (2007).

Especificamente, sobre a rotulação de agrupamentos hierárquicos, pretende-se avaliar o uso de diferentes testes estatísticos nos algoritmos de seleção de rótulos do RLUM. Pretende-se avaliar a cada caso, isto é, a cada tabela de contingência para cada atributo, qual o melhor estimador a ser utilizado nos testes. O objetivo é avaliar se há melhorias significativas nos conjuntos de rótulos identificados. Esse experimento implica em alterar a forma de testes de independência nos algoritmos e reimplementá-los. Para avaliar as diferenças obtidas entre os resultados com os novos algoritmos e os anteriores, pretende-se utilizar as avaliações contra taxonomias *gold*.

No que tange às aplicações práticas, a TopTax será utilizada em um projeto de pesquisa e desenvolvimento junto ao projeto de manutenção da Base de Dados da Pesquisa Agropecuária - BDPA [1]. Pretende-se utilizar a TopTax na identificação e extração de metadados, bem como identificação de possíveis taxonomias de tópicos em subdomínios contidos na BDPA, e, na seqüência desenvolver um classificador de documentos a partir dos resultados encontrados. A TopTax também será parcialmente utilizada em um projeto de organização de informação sobre recursos hídricos, também da Embrapa.

[1] Disponível em http://bdpa.cnptia.embrapa.br/.

Referências Bibliográficas

Aggarwal, C. C., Gates, S. C., e Yu, P. S. (1999). On the merits of building categorization systems by supervised clustering. Em *Proceedings of the* 5th *International Conference on Knowledge Discovery and Data Mining - SIGKDD - ACM*, páginas 352–356. Citado na página 21.

Agrawal, R. e Srikant, R. (1994). Fast algorithms for mining association rules in large databases. Em *Proceedings of XX International Conference on Very Large Data Bases - VLDB*, páginas 487–499. Citado na página 16.

Bae, E., Vasudevan, B. G., e Balakrishnan, R. (2008). ACORN: Towards automating domain specific ontology construction process. Em *Progress in WWW Research and Development*, volume 4976 of *Lecture Notes in Computer Science*, páginas 487–498, Verlag - Berlin. Springer. Citado nas páginas 3 e 25.

Banerjee, S. e Pedersen, T. (2003). The design, implementation, and use of the ngram statistics package. Em *Proceedings of* 4th *Conference on Intelligent Text Processing and Computational Linguistics - CICLing*, páginas 370–381. Citado nas páginas 30, 31, 44, 46, e 54.

BINAGRI (2009). Thesaurus agrícola nacional – Thesagro. Disponível em http:www.agricultura.gov.brportal, [Maio, 2009]. Citado na página 84.

Bishop, Y. M. M., Fienberg, S. E., e Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice,*. MIT Press, London - UK. Citado nas páginas 47, 50, 54, 71, e 72.

Bloehdorn, S., Cimiano, P., Hotho, A., e Staab, S. (2005). An ontology-based framework for text mining. *Journal of Computacional Linguistics - LDV Forum*, 20(1):87–112. Citado na página 20.

Boley, D. (1998). Principal direction divisive partitioning. *Data Mining Knowledge Discovery*, 2(4):325–344. Citado na página 23.

Brefeld, U. e Scheffer, T. (2004). Co-em support vector learning. Em *Proceedings of the Twenty-first International Conference on Machine Learning - ICML*. Citado na página 16.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Journal of Computational Linguistics*, 21(4):543–565. Citado na página 30.

Bussab, W. O. e Morettin, P. A. (2006). *Estatística Básica*. Saraiva, São Paulo - SP. Citado na página 93.

Card, S., Mackinlay, J., e Schneiderman, B. (1999). Information visualization. Em Card, S., Mackinlay, J., e Schneiderman, B., editores, *Readings in Information Visualization: Using Vision to Think*, capítulo 1, páginas 1–34. Morgan Kaufmann Publishers, 1 edição. Citado na página 18.

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann Publishers. Citado na página 17.

Chandrasekaran, B., Josephson, J. R., e Benjamins, V. R. (1999). What are ontologies, and why do we need them? *Intelligent Systems and their Applications - IEEE*, 14(1):20–26. Citado na página 10.

Chuang, S.-L. e Chien, L.-F. (2005). Taxonomy generation for text segments: A practical web-based approach. *Transactions on Information Systems - ACM*, 23(4):363–396. Citado na página 22.

Cimiano, P. (2006). *Ontology Learning Population from Text: Algorithms, Evaluation and Applications*. Springer Science+BusinessnMedia, LLC. Citado nas páginas 9, 20, e 25.

Clifton, C., Cooley, R., e Rennie, J. (2004). Topcat: Data mining for topic identification in a text corpus. *Transactions on Knowledge and Data Engineerings - IEEE*, 16(8):949–964. Citado na página 23.

Conrado, M. S. (2009). O efeito do uso de diferentes formas de extração de termos na compreensibilidade e representatividade dos termos em coleções textuais na língua portuguesa. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos - SP. Dissertação depositada em [Julho-2009]. Citado nas páginas 30, 39, 107, e 109.

Conrado, M. S., Marcacini, R. M., Moura, M. F., e Rezende, S. O. (2009). O efeito do uso de diferentes formas de geração de termos na compreensibilidade e representatividade dos termos em coleções textuais na língua portuguesa. Em *Proceedings of II Workshop on Web and Text Intelligence - WTI*. No prelo. Citado nas páginas 30, 34, e 108.

Conrado, M. S., Moura, M. F., Marcacini, R. M., e Rezende, S. O. (2008). Avaliando diferentes formas de geração de termos a partir de coleções textuais. Relatório técnico 334, Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos - SP. Citado nas páginas 30, 34, 46, e 108.

Dash, M. e Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(3):131–156. Citado na página 14.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., e Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407. Citado nas páginas 14 e 23.

Dejun, X. e Maosong, S. (2004). Raising high-degree overlapped character bigrams into trigrams for dimensionality reduction in chinese text categorization. Em *Proceedings of* 5th *Conference on Intelligent Text Processing and Computational Linguistics - CICLing*, páginas 584–595. Citado nas páginas 42 e 60.

Dhillon, I., Kogan, J., e Nicholas, C. (2003). Feature selection and document clustering. Em Berry, M. W., editor, *Survey of Text Mining*, páginas 73–100. Springer. Citado na página 31.

Duda, R. O., Hart, P. E., e Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, Inc. Citado nas páginas 16 e 65.

Dupret, G. e Piwowarski, B. (2005). Deducing a term taxonomy from term similarities. Em *Proceedings of Second International Workshop on Knowledge Discovery and Ontologies - KDO*, páginas 11–22. Citado na página 23.

Duval, E., Hodgins, W., Sutton, S., e Weibel, S. L. (2002). Metadata principles and practicabilities. *D-Lib Magazine*, 8(4). Disponível em http://www.dlib.org/dlib/april02/weibrl/04-weibel.html, [Agosto, 2006]. Citado na página 28.

Ebecken, N. F. F., Lopes, M. C. S., e Aragão, M. C. d. (2003). Mineração de textos. Em Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, capítulo 13, páginas 337–364. Manole, 1 edição. Citado nas páginas 7, 28, e 29.

EMBRAPA (2008). Embrapa pecuária sudeste relatório da oficina de trabalho - Embrapa do futuro. Documentos - 47p 83, Embrapa Pecuária Sudeste, São Carlos - SP. Disponível em http://www.repdigital.cnptia.embrapa.br/bitstream/CPPSE/18030/1/Documentos83.pdf, [Agosto, 2009]. Citado na página 35.

Escudeiro, N. F. e Jorge, A. M. (2005). Semi-automatic creation and maintenance of web resources with webtopic. Em *Proceedings of European Web Mining Forum - EWMF*, volume 4289 of *Lecture Notes in Artificial Intelligence*, páginas 82–102, Verlag - Berlin. Springer. Citado nas páginas 3 e 24.

Evangelista, S. R. M., Souza, K. X. S., Souza, M. I. F., Cruz, S. A. B., Leite, M. A. A., Santos, A. D., e Moura, M. F. (2003). Gerenciador de conteúdos da agência Embrapa de informação. Em *International Symposium on Knowledge Managment - ISKM*, volume CD-ROM, páginas 1–12. PUC-Curitiba. Citado na página 2.

Feldman, R. e Sanger, J. (2007). *The Text Mining Hand Book - Advanced Approaches in Analysing Unstructured Data*. Cambridge University Press. Citado na página 63.

Fienberg, S. E. (1985). *The analysis of cross-classified categorical data*. MIT, Cambridge - UK. Citado na página 71.

Fortuna, B., Lavrac, N., e Velardi, P. (2008). Advancing topic ontology learning through term extraction. Em *Proceedings of* 10th *Pacific Rim International Conference on Artificial Intelligence - PRICAI*, volume 5351 of *Lecture Notes in Computer Science*, páginas 626–635. Springer. Citado na página 25.

Frigui, H. e Nasraoui, O. (2004). Simultaneous clustering and dynamic keyword weighting for text documents. Em Berry, M. W., editor, *Survey of Text Mining Clustering, Classification and Retrieval*, capítulo 1, páginas 45–72. Springer, Verlag - Berlim, 1 edição. Citado na página 22.

Gantz, J. F., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Xheneti, I., Toncheva, A., e Manfrediz, A. (2007). The expanding digital universe: A forecast of worldwide information growth through 2010. *External Publication of IDC (Analyse the Future) Information and Data*, páginas 1–21. Citado na página 1.

Garshol, L. M. (2004). Metadata? thesauri? taxonomies? topic maps! *Journal of Information Science*, 30(4):378–391. Citado na página 9.

Glover, E., Lawrence, D. M. P. S., e Krovetz, R. (2002). Inferring hierarchical descriptions. Em *Proceedings of the* 11th *International Conference on Information and Knowledge Management - CIKM*, páginas 507–514, New York - EUA. ACM. Citado na página 66.

Gonzalez, M. A. I., de Lima, V. L. S., e de Lima, J. V. (2006). Tools for nominalization: An alternative for lexical normalization. *Proceedings of the* 7th *Workshop on Computational Processing of Written and Spoken Portuguese - PROPOR*, 3960:100–109. Citado na página 30.

Hammouda, K. e Kamel, M. (2008). Distributed collaborative web document clustering using cluster keyphrase summaries. *Information Fusion*, 9(4):465–480. Citado na página 22.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information*, 42:7–15. Citado na página 29.

Haykin, S. (2001). *Redes Neurais: Princípios e Práticas*. Bookman, Porto Alegre - RS. Citado na página 13.

Hofmann, T. (1999). The cluster abstraction model: Unsupervised learning of topic hierarquies from text data. Em *Proceedings of International Joint Conference on Artificial Intelligence - IJCAI*, páginas 682–687. Citado nas páginas 14, 65, e 109.

Hotho, A., Maedche, A., e Staab, S. (2002). Ontology-based text document clustering. *Künstliche Intelligenz*, 16(4):48–54. Citado nas páginas 17 e 25.

Ienco, D. e Meo, R. (2008). Towards the automatic construction of conceptual taxonomies. Em *Data Warehousing and Knowledge Discovery*, volume 5182 of *Lecture Notes in Computer Science*, páginas 327–336. Springer. Citado na página 66.

Jiang, X. e Tan, A.-H. (2005). Mining ontological knowledge from domain-specific text documents. Em *Proceedings of the* 5th *International Conference on Data Mining - ICDM*, páginas 665–668. Citado na página 25.

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. Em *Proceedings of* 10th *European Conference on Machine Learning - ECML*, páginas 137–142. Citado nas páginas 15 e 57.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, New York - EUA. Citado na página 13.

Kachigan, S. K. (1986). *Statistical Analysis*. Radius Press, New York - EUA. Citado na página 71.

Kashyap, V., Ramakrishnan, C., Thomas, C., e Sheth, A. (2005a). Taxaminer: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, 1(2):240–266. Citado nas páginas 3, 17, 23, 24, 38, 65, e 104.

Kashyap, V., Ramakrishnan, C., Thomas, C., e Sheth, A. (2005b). Taxaminer: Improving taxonomy label quality using latent semantic indexing. Relatório técnico, Computer Science Department - University of Georgia. Disponível em http://lsdis.cs.uga.edu/ cthomas/resources/taxaminer.pdf, [Julho, 2005]. Citado na página 24.

Kobayashi, M. e Aono, M. (2004). Vector space models for search and cluster mining. Em Berry, M. W., editor, *Survey of Text Mining: Clustering, Classification and Retrieval*, capítulo 5, páginas 103–120. Springer, Verlag - Berlim, 1 edição. Citado na página 17.

Konchady, M. (2006). *Text mining application programming*. Charles River Media, Boston - EUA. Citado na página 63.

Koshman, S., Spink, A., e Jansen, B. J. (2006). Web searching on the vivisimo search engine. *Journal of the American Society for Information Science and Technology*, 57(14):1875–1887. Citado na página 20.

Krishnapuram, R. e Kummamuru, K. (2003). Automatic taxonomy generation: Issues and possibilities. Em *Proceedings of International Fuzzy Sets and Systems – IFSA*, volume 2715 of *Lecture Notes in Computer Science*, páginas 52–63. Springer. Citado na página 104.

Lamirel, J.-C., Ta, A. P., e Attik, M. (2008). Novel labeling strategies for hierarchical representation of multidimensional data analysis. Em *Proceedings of the* 26th *International Conference on Artificial Intelligence and Applications - IASTED*, páginas 401–408, New York - EUA. ACM PressAddison-Wesley Publishing Co. Citado na página 66.

Lamping, J., Rao, R., e Pirolli, P. (1995). A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. Em *Proceedings of the XIII Conference on Human Factors in Computing Systems - CHI*, páginas 401–408, New York - EUA. ACM Press/Addison-Wesley Publishing Co. Citado na página 18.

Landauer, T. K., Foltz, P. W., e Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284. Citado na página 14.

Larocca, J., Santos, A. D., Kaestner, C. A. A., e Freitas, A. A. (2000). Document clustering and text summarization. Em *Proceedings of the* 4th *International Conference on Practical Applications of Knowledge Discovery and Data Mining - PADD*, páginas 41–55, London - UK. The Practical Application Company. Citado na página 21.

Larsen, B. e Aone, C. (1999). Fast and effective text mining using linear-time document clustering. Em *Proceedings of the Knowledge Discovery and Data Mining - KDD*, páginas 16–22. Citado na página 65.

Lawrie, D. e Croft, W. B. (2000). Discovering and comparing topic hierarchies. Em *Proceedings of the* 6th *Computer-Assisted Information Retrieval - RIAO*, páginas 314–330. College de France. Citado na página 22.

Lawrie, D., Croft, W. B., e Rosenberg, A. (2001). Finding topic words for hierarchical summarization. Em *Proceedings of the XXIV Annual International Conference on Research and Development in Information Retrieval - SIGIR*, páginas 349–357, New York - EUA. ACM. Citado nas páginas 22 e 85.

Librelotto, G. R., Ramalho, J. C., e Henriques, P. R. (2004). Tm-builder : um construtor de ontologias baseado em topic maps. *Centro Latinoamericano de Estudios en Informática [Electronic Journal]*, 2. Citado na página 20.

Lima, V. L. S., Nunes, M. G., e Vieira, R. (2007). Desafios do processamento de línguas naturais. Em *Proceedings of the* 34th *Seminário Integrado de Software e Hardware - SEMISH*, páginas 2202–2216. Sociedade Brasileira de Computação - SBC. Citado na página 39.

Liu, L., Kang, J., Yu, J., e Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering - NLP-KE*, páginas 597–601. Citado na página 31.

Liu, T., Liu, S., Chen, Z., e Ma, W.-Y. (2003). An evaluation on feature selection for text clustering. Em *Proceedings of the* 20th *International Conference on Machine Learning - ICML*, páginas 488–495. AAAI Press. Citado nas páginas 31 e 58.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31. Citado na página 29.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *Journal of Research and Development - IBM*, 2(2):159–165. Citado nas páginas 15, 31, e 44.

Maedche, A. (2002). Ontology learning for the semantic web. volume 665 of *The Springer International Series in Engineering and Computer Science*. Springer. Citado nas páginas 17 e 20.

Manning, C. D., Raghavan, P., e Schütze, H. (2008). Language models for information retrieval. Em *An Introduction to Information Retrieval*, capítulo 12. Cambridge University Press. Citado na página 29.

Manning, C. D. e Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge - EUA. Citado nas páginas 14, 16, 30, 45, 46, 54, e 85.

Marcacini, R. M. (2008). Um ambiente interativo para análise visual de agrupamentos hierárquicos. Monografia conclusão de curso de graduação, Instituto de Ciências Matemáticas e de Computação – ICMC – USP. Citado nas páginas 34, 39, e 107.

Marcacini, R. M., Moura, M. F., e Rezende, S. O. (2007). Biblioteca digital do ifm: uma aplicação para a organização da informação através de agrupamentos hierárquicos. Em *Proceedings of Workshop de Bibliotecas Digitais - WDL, XIII Brazilian Symposium on Multimedia and the Web - WebMidia*, volume CD-ROM, páginas 1–16. UFRGS. Citado nas páginas 35 e 108.

Marcacini, R. M., Moura, M. F., e Rezende, S. O. (2009). Uma abordagem para seleção de grupos significativos em agrupamento hierárquico de documentos. Em *Proceedings of Encontro Nacional de Inteligência Artificial - ENIA*, volume CD-ROM, páginas 1–16. UFRGS. Citado nas páginas 32, 38, 39, e 107.

Marcacini, R. M., Rossi, R. G., Moura, M. F., e Rezende, S. O. (2008). Identificação de competências por meio de agrupamentos hierárquicos. Em *Anais do Congresso da Academia Trinacional de Ciências - C3N*, volume CD-ROM. Unioeste. Citado nas páginas 35 e 108.

Mardia, K. V., Kent, J. T., e Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press. Citado nas páginas 14, 16, e 33.

Martins, C. A. (2003). *Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos - SP. Citado na página 28.

Martins, C. A., Matsubara, E. T., e Monard, M. C. (2003). Um estudo de caso utilizando uma ferramenta computacional que auxilia na redução da dimensão da representação de documentos em tarefas preditivas de mineração de textos. *Proceedings of Fourth Workshop on Advances & Trends in Artificial Intelligence for Problem Solving - IASTED*, páginas 21–27. Citado na página 30.

Matsubara, E. T., Martins, C. A., e Monard, M. C. (2003). Pre-TexT: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Relatório Técnico 209, Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos - SP. Citado nas páginas 29, 35, 44, 84, 92, e 96.

Matsubara, E. T., Monard, M. C., e Batista, G. E. A. P. A. (2005). Utilizando algoritmos de aprendizado semi-supervisionados multi-visão como rotuladores de textos. Em

Proceedings of XXV Congresso da Sociedade Brasileira de Computação - SBC, volume CD-ROM, páginas 2108–2117. UNISINOS. Citado na página 16.

Mazzari, C. A. (2007). Gestão de pessoas e identificação de competências estratégicas em unidades descentralizadas da Embrapa - o caso Embrapa pecuária sudeste. Projeto de Pesquisa e Desenvolvimento, Situação: Concluído; Natureza: Pesquisa. Embrapa Pecuária Sudeste. Citado nas páginas 35 e 108.

Miao, Y., Keselj, V., e Milios, E. E. (2005). Document clustering using character n-grams: a comparative evaluation with term-based and word-based clustering. Em *Proceedings of the International Conference on Information and Knowledge Management - CIKM*, páginas 357–358. ACM. Citado na página 42.

Mihalcea, R. (2005). Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. Em *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT*, páginas 411–418, Morristown - EUA. Association for Computational Linguistics. Citado na página 66.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Higher Education. Citado na página 15.

Müller, A., Dörre, J., Gerstl, P., e Seiffert, R. (1999). The TaxGen framework: Automating the generation of a taxonomy for a large document collection. Em *Proceedings of* 32nd *Hawaii International Conference on System Sciences - HICSS*, páginas 2034–2042, New York - EUA. IEEE Press. Citado nas páginas 21 e 69.

Monard, M. C. e Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. Em Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, capítulo 4, páginas 89–114. Manole, 1 edição. Citado na página 15.

Moura, M. F. (2006). Uma abordagem para a construção e atualização de taxonomias de tópicos a partir de coleções de textos dinâmicas. Monografia de Qualificação de Doutorado, Instituto de Ciências Matemáticas e de Computação – ICMC – USP. Citado na página 19.

Moura, M. F., Marcacini, R. M., Nogueira, B. M., Conrado, M. S., e Rezende, S. O. (2008a). A proposal for building domain topic taxonomies. Em *Proceedings of I Workshop on Web and Text Intelligence - WTI, XIX Simpósio Brasileiro de Inteligência Artificial - SBIA*, páginas 83–84, São Carlos - SP. Instituto de Ciências Matemáticas e de Computação – ICMC – USP. Citado nas páginas 33, 64, e 105.

Moura, M. F., Marcacini, R. M., Nogueira, B. M., Conrado, M. S., e Rezende, S. O. (2008b). Uma abordagem completa para a construção de taxonomias de tópicos em um domínio. Relatório técnico 329, Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos - SP. Citado na página 105.

Moura, M. F., Marcacini, R. M., e Rezende, S. O. (2008c). Easily labelling hierarchical document clusters. Em *Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados - WAAMD, XXIII Simpósio Brasileiro de Banco de Dados - SBBD*, páginas 37–45, Porto Alegre - RS. Porto Alegre : SBC. Citado nas páginas 33, 35, 37, 67, 70, 92, 103, 104, 106, 107, e 128.

Moura, M. F., Nogueira, B. M., Conrado, M. S., dos Santos, F. F., e Rezende, S. O. (2008d). Making good choices of non-redundant n-gram words. Em *Proceedings of I International Workshop on Data Mining and Artificial Intelligence - DMAI, XI International Conference on Computer and Information Technology - ICCIT*, páginas 64–71. IEEE Xplore Digital Library. Citado nas páginas xiii, 32, 51, 95, e 106.

Moura, M. F. e Rezende, S. O. (2007). Choosing a hierarchical cluster labelling method for a specific domain document collection. Em Neves, J., Santos, M. F., e Machado, J. M., editores, *New Trends in Artificial Intelligence*, Encontro Português de Inteligência Artificial - EPIA, capítulo 11, páginas 812–823. Associação Portuguesa para Inteligência Artificial - APIA, 1 edição. Citado nas páginas 33, 42, 87, 104, 106, e 128.

Moura, M. F. e Rezende, S. O. (2009a). A new and simple method for labeling hierarchical document clusters. Em *Proceedings of International Conference on Data Mining series - ICDM*, Miami - EUA. Trabalho submetido, em revisão. Citado na página 106.

Moura, M. F. e Rezende, S. O. (2009b). Novos algoritmos e métodos de avaliação objetiva para modelos de rotulação em agrupamentos hierárquicos de documentos. Relatório Técnico 346, Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos - SP. Citado na página 106.

Moura, M. F. e Rezende, S. O. (2009c). RLUM: a robust labeling up method for hierarchical document clusters. *Data Mining and Knowledge Discovery*. Trabalho submetido, em revisão. Citado na página 107.

Nogueira, B. M. (2009). Avaliação de métodos não-supervisionados de seleção de atributos para mineração de textos. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos. Citado nas páginas 31, 39, 44, 52, e 107.

Nogueira, B. M., Moura, M. F., Conrado, M. S., e Rezende, S. O. (2008a). Avaliação de métodos não-supervisionados de seleção de atributos para mineração de textos. Em *Proceedings of I Workshop on Web and Text Intelligence - WTI, XIX Simpósio Brasileiro de Inteligência Artificial - SBIA*, páginas 59–66, São Carlos - SP. Instituto de Ciências Matemáticas e de Computação – ICMC – USP. Citado nas páginas 31, 58, e 107.

Nogueira, B. M., Moura, M. F., Conrado, M. S., Rossi, R. G., Marcacini, R. M., e Rezende, S. O. (2008b). Winning some of the document preprocessing challenges in a text mining process. Em *Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados - WAAMD, XXIII Simpósio Brasileiro de Banco de Dados- SBBD*, páginas 10–18, Porto Alegre - RS. SBC. Citado nas páginas 44, 53, 58, e 107.

Nunes, M. G. V. (1996). The design of a lexicon for brazilian portuguese: Lessons learned and perspectives. *Proceedings of the Second Workshop on Computational Processing of Written and Spoken Portuguese - PROPOR*, páginas 61–70. Citado na página 30.

Orengo, V. M. e Huyck, C. (2001). A stemming algorithm for the portuguese language. *Proceedings of* 8th *International Symposium on String Processing and Information Retrieval - SPIRE*, páginas 183–193. Citado na página 30.

Ou, S., Khoo, C. S. G., e Goh, D. H. (2006). Automatic multi-document summarization for digital libraries. Em *Proceedings Asia-Pacific Conference on Library & Information Education & Practice - A-LIEP*, páginas 72–82, Singapore. Citado na página 22.

Peltonen, J., Sinkhonen, J., e Kaski, S. (2002). Discriminative clustering of text documents. Em *Proceedings of IX International Conference on Neural Information Processing - ICONIP*, volume 4, páginas 1956–1960. IEEE. Citado na página 17.

Pidcok, W. (2003). What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology and a meta-model? Disponível em http://www.metamodel.com/article.php?story=200030115211223271, [Agosto, 2009]. Citado nas páginas 9 e 10.

Pierre, J. M. (2002). Mining knowledge from text collections using automatically generated metadata. Em *Proceedings of the* 4th *International Conference on Practical Aspects of Knowledge Managment - PAKM*. Disponível em http://www.sukidog.com/jpierre/pakm2002.pdf, [Julho, 2004]. Citado na página 15.

Popescul, A. e Ungar, L. (2000). Automatic labeling of document clusters. Disponível em http://citeseer.nj.nec.com/popescul100automatic.html, [Agosto, 2009]. Citado nas páginas 66, 70, e 75.

Porter, M. F. (1997). An algorithm for suffix stripping. *Readings in Information Retrieval*, páginas 313–316. Citado na página 29.

Punera, K., Rajan, S., e Ghosh, J. (2005). Automatically learning document taxonomies for hierarchical classification. Em *Proceedings of the* 14th *International Conference on World Wide Web - WWW*, páginas 1010–1011. ACM. Citado na página 23.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc, San Francisco - EUA. Citado na página 15.

Ramsey, P. H. (1993). Multiple comparisons of independent means. Em Edwards, L. K., editor, *Applied analysis of variance in behavioral science*. Marcel Dekker, New York - EUA. Citado na página 84.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *Proceedings of the Empirical Methods in Natural Language Processing Conference - EMNLP*, páginas 491–497. Citado na página 30.

Rezende, S. O., Moura, M. F., e Marcacini, R. M. (2008). Uso de mineração de textos para gerenciamento do conhecimento produzido. Em *Proceedings of II Assembléia Geral - Instituto Fábrica do Milênio - IFM, Monitoramento das necessidades das indústrias - WP01*, capítulo 5, páginas 39–40. IFM, 1 edição. Citado nas páginas 35 e 108.

Rezende, S. O., Oliveira, R., Félixa, L. C. M., e Rocha, C. (1998). Visualization for knowledge discovery in database. Em *Data Mining*, páginas 81–95. WIT Press Computational Mechanics Publications, 1 edição. Citado na página 34.

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., e Paula, M. F. (2003). Mineração de dados. Em Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, capítulo 12, páginas 307–335. Manole, 1 edição. Citado nas páginas 2, 8, e 9.

Rijsbergen, C. J. v. (1979). *Information Retrieval*. Butterworth. Citado na página 13.

Rino, L. H. M. e Pardo, T. A. S. (2003). A sumarização automática de textos: Principais características e metodologias. Em Vieira, R., editor, *Proceedings of Jornada de Atualização em Inteligência Artificial - JAIA*, páginas 203–245. SBC, viii edição. Citado na página 17.

Rossi, R. G., Rezende, S. O., e de Andrade Lopes, A. (2009). IESystem - uma ferramenta para extração de elementos de artigos científicos. Relatório técnico, (em preparação), Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos - SP. Citado na página 108.

Rychlý, P. e Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). Em *Proceedings of the* 45th *Annual Meeting of the Association for Computational Linguistics - ACL*. The Association for Computer Linguistics. Citado na página 9.

Salton, G. e Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Relatório técnico, Ithaca - EUA. Disponível em http://www.ncstrl.org:8900/ncstrl/servlet/search?formname=detail&id=oai%3Ancstrlh%3Acornellc%3ACORNELLCS%3ATR87-881, [Janeiro, 2008]. Citado nas páginas 12 e 13.

Salton, G. e McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc, New York - EUA. Citado na página 31.

Salton, G., Yang, C. S., e Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Association Science*, 1(26):33–44. Citado nas páginas 13, 15, 31, 44, 87, e 96.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing - NEMLAP*, páginas 44–49. Citado na página 30.

Searle, S. R. (1971). *Linear models*. J. Wiley, New York - NY. Citado nas páginas 33, 80, 83, 84, e 129.

Segal, E., Koller, D., e Ormoneit, D. (2001). Probabilistic abstraction hierarchies. Em *Advances in Neural Information Processing Systems*, páginas 913–920. MIT Press - Canada. Citado na página 65.

Silberschatz, A. e Tuzhilin, A. (1995). On subjective measures of interestingness in knowledge discovery. Em *Proceedings of I International Conference on Knowledge Discovery and Data Mining - KDD*, páginas 275–281. Citado nas páginas 18 e 34.

Silla Jr, C. N. e Kaestner, C. A. A. (2002). Estudo de métodos automáticos para sumarização de textos. *Anais do Simpósio de Tecnologias de Documentos - STD*, 1:45–48. Citado na página 29.

Slonim, N. e Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. Em *Proceedings of the XXIII Annual International Conference on Research and Development in Information Retrieval - SIGIR*, páginas 208–215, New York - EUA. ACM. Citado na página 14.

Snedecor, G. W. e Cochran, W. G. (1967). *Statistical methods, 6th ed.* Iowa State University Press, Iowa - EUA. Citado nas páginas 81, 84, 89, e 95.

Soucy, P. e Mineau, G. W. (2005). Beyond tfidf weighting for text categorization in the vector space model. Em Kaelbling, L. P. e Saffiotti, A., editores, *Proceedings of the XIX International Joint Conferences on Artificial Inteligence - IJCAI*, páginas 1130–1135. Professional Book Center. Citado na página 13.

Souza, M. I. F., Moura, M. F., e Santos, A. D. (2005). Similaridade e complementaridade dos metadados nos sistemas agência de informação Embrapa e acervo documental do ainfo. Em *Proceedings of Congresso Brasileiro de Biblioteconomia - CBB*, página 12. FEBAB. Citado na página 28.

Souza, M. I. F., Santos, A. D., Moura, M. F., e Alves, M. D. R. (2006). Agência de informação Embrapa: uma aplicação para a organização da informação e gestão do conhecimento. Em *Proceedings of II Workshop on Digital Libraries - WDL*, páginas 51–56, Florianópolis - SC. SBC. Citado nas páginas 20 e 37.

Sphinx Brasil (2008). Sphinx. Disponível em http://www.sphinxbrasil.com.br/po/, [Agosto, 2009]. Citado na página 30.

Tang, L., Liu, H., Zhang, J., Agarwal, N., e Salerno, J. J. (2008). Topic taxonomy adaptation for group profiling. *Transactions on Knowledge Discovery from Data*, 1(4):1–28. Citado na página 24.

Tang, L., Zhang, J., e Liu, H. (2006). Acclimatizing taxonomic semantics for hierarchical content classification from semantics to data-driven taxonomy. Em Eliassi-Rad, T., Ungar, L. H., Craven, M., e Gunopulos, D., editores, *Proceedings of the* 12th *International Conference on Knowledge Discovery and Data Mining - SIGKDD*, páginas 384–393. ACM. Citado na página 24.

Treeratpituk, P. e Callan, J. (2006). Automatically labeling hierarchical clusters. Em *Proceedings of* 7th *Annual International Conference on Digital Government Research*, páginas 167–176, San Diego - CA. ACM. Citado na página 66.

Weiss, S. M., Indurkhya, N., Zhang, T., e Damerau, F. J. (2005). *Text Mining - Predictive Methods for Analizing Unstructured Information*. Springer Science+Business Media, Inc. Citado nas páginas 7, 17, 42, e 64.

Witten, I. H. e Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edição. Citado na página 2.

Wives, L. K. (2004). *Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos*. Tese de Doutorado, Universidade Federal do Rio Grande do Sul - Instituto de Informática. Citado na página 17.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., e Steinberg, D. (2008). Top 10 algorithms in data mining. *Journal Knowledge and Information Systems*, 14(1):1–37. Citado na página 15.

Yao, Y. Y. (2003). Information-theoretic measures for knowledge discovery and data mining. Em *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, páginas 115–136. Springer, Berlim - Alemanha. Citado na página 13.

Yu, L. e Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224. Citado na página 13.

Zhang, C., Wang, H., Liu, Y., e Xu, H. (2009). Document clustering description extraction and its application. Em *Proceedings of* 22nd *International Computer Processing of Oriental Languages - ICCPOL*, volume 5459 of *Lecture Notes in Computer Science*, páginas 370–377. Springer. Citado na página 65.

Zhang, X. e Zhu, X. (2007). A new type of feature - loose n-gram feature in text categorization. Em *Proceedings of* 3rd *Iberian Conference on Pattern Recognition and Image Analysis - IbPRIA*, páginas 378–385. Citado nas páginas 42, 60, 61, e 110.

Zhao, Y. e Karypis, G. (2005). Topic-driven clustering for document datasets. Em *Proceddings of the* 5th *International Conference on Data Mining - SIAM*, volume 119. Citado na página 24.

Apêndice

A

Apêndice - Tabelas de comparação das medidas de interesse de resultados de EMIM e busca

Neste apêndice encontram-se todos os dados de medidas de interesse obtidas a partir de recuperação de informações e de esperança da informação mútua, para todas as coleções de textos utilizadas nos experimentos do Capítulo 5. Cada seção contém uma pequena contextualização a fim de evitar a consulta constante ao Capítulo 5. Para cada tabela é apresentada a análise dos resultados. Ainda, as coleções de textos utilizadas são as resumidas na Tabela 5.11, Capítulo 5.

A.1 Comparações múltiplas de médias de EMIM

Para estudar o comportamento da EMIM em cada nó para cada método ajusta-se o seguinte modelo linear generalizado:

$$\hat{emim}_{n_i} = \hat{\mu} + \hat{n}_i + \hat{m} + \hat{\epsilon}$$

onde,

- $\hat{emim}_{n_i}$: emim inferida para cada nó em cada método e cada coleção de textos;
- $\hat{\mu}$: a média geral de dispersão estimada para inferir a $emim_{n_i}$;
- $\hat{n}_i$: estimativa do efeito de cada nó n_i em cada coleção de textos na inferência da $emim_{n_i}$. O nó funciona como efeito de categoria para cada EMIM, fazendo com que as comparações de médias dos efeitos dos métodos sejam justas por estarem sendo realizadas nó a nó;

- $\hat{m}$: estimativa do efeito do método de rotulação aplicado, isto é, quanto a média geral de dispersão desvia-se da EMIM final em relação a esse efeito;

- $\hat{\epsilon}$: erro aleatório de ajuste do modelo linear.

Com esse modelo, a seguinte hipótese é testada para cada coleção de textos:

$$H_0 : EMIM(RLUM) \geq EMIM(P) \approx EMIM(M) \geq EMIM(D)$$

Realizando-se a comparação múltipla de médias das estimativas de EMIM em relação aos efeitos de cada método, para cada coleção de textos é apresentada uma tabela. Cada tabela contem o nome da coleção, o número de graus de liberdade gl utilizado na comparação múltipla de médias, a variância média vm utilizada pelo teste SNK, o número de pontos n utilizado para estimar a EMIM em cada coleção, a estimativa da média de EMIM ($\bar{emim}$) para cada método em cada coleção e uma letra para *grupo*. Letras iguais representam grupos estatisticamente iguais e, a ordem dos grupos é a alfabética.

Col. de Textos AI $gl = 147, vm = 160140.0445$				*Col. de Textos HCI* $gl = 108, vm = 286839.4677$			
método	n	$\bar{emim}$	*grupo*	método	n	$\bar{emim}$	*grupo*
RLUM	50	1089.899965	*a*	*RLUM*	37	885.627493	*a*
PopeUngar	50	893.817625	*b*	*PopeUngar*	37	765.692200	*a*
MostFreq	50	893.742870	*b*	*MostFreq*	37	762.961843	*a*
RLDM	50	557.042183	*c*	*RLDM*	37	373.175243	*b*
Col. de Textos Hard $gl = 36, vm = 185914.7144$				*Col. de Textos Sec* $gl = 159, vm = 487370.7896$			
método	n	$\bar{emim}$	*grupo*	método	n	$\bar{emim}$	*grupo*
RLUM	13	1104.943006	*a*	*RLUM*	54	1613.500940	*a*
PopeUngar	13	909.675011	*a*	*PopeUngar*	54	1337.548719	*a*
MostFreq	13	906.329243	*a*	*MostFreq*	54	1337.383964	*a*
RLDM	13	77.376233	*b*	*RLDM*	54	636.411828	*b*

Tabela A.1: EMIM - resultados para o domínio de Ciência da Computação

Col. de Textos AnaC $gl = 150,\ vm = 626822.0993$				*Col. de Textos InoC* $gl = 153,\ vm = 1e+06$			
method	n	$\bar{emim}$	*group*	*method*	n	$\bar{emim}$	*group*
RLUM	51	1944.724002	*a*	*RLUM*	52	2235.525562	*a*
PopeUngar	51	1447.356021	*b*	*PopeUngar*	52	1664.561465	*b*
MostFreq	51	1446.846145	*b*	*MostFreq*	52	1661.026177	*b*
RLDM	51	747.228784	*c*	*RLDM*	52	1308.392552	*b*
Col. de Textos OrgC $gl = 105,\ vm = 1e+06$				*Col. de Textos Poly* $gl = 141,\ vm = 274807.2374$			
method	n	$\bar{emim}$	*group*	*method*	n	$\bar{emim}$	*group*
RLUM	36	2638.297039	*a*	*RLUM*	48	1083.452986	*a*
MostFreq	36	2129.767519	*a*	*PopeUngar*	48	885.448869	*a b*
PopeUngar	36	2113.688907	*a*	*MostFreq*	48	884.329346	*a b*
RLDM	36	303.138253	*b*	*RLDM*	48	656.565679	*b*

Tabela A.2: EMIM - resultados para o domínio de Química

Col. de Textos BioP $gl = 102,\ vm = 4e+06$				*Col. de Textos GeoP* $gl = 150,\ vm = 309004.6045$			
method	*n*	$\bar{emim}$	*group*	*method*	*n*	$\bar{emim}$	*group*
RLUM	35	3305.903151	*a*	*RLUM*	51	1644.596988	*a*
RLDM	35	2353.240417	*b*	*PopeUngar*	51	1209.551453	*b*
PopeUngar	35	1842.456371	*b*	*MostFreq*	51	1209.089450	*b*
MostFreq	35	1842.456371	*b*	*RLDM*	51	1069.912500	*b*
Col. de Textos Mech $gl = 138,\ vm = 457452.0106$				*Col. de Textos Quan* $gl = 117,\ vm = 284639.2019$			
method	*n*	$\bar{emim}$	*group*	*method*	*n*	$\bar{emim}$	*group*
RLUM	47	1335.448216	*a*	*RLUM*	40	1276.068038	*a*
PopeUngar	47	1030.787221	*a*	*PopeUngar*	40	988.725002	*b*
MostFreq	47	1030.787221	*a*	*MostFreq*	40	985.847125	*b*
RLDM	47	414.650337	*b*	*RLDM*	40	654.964093	*c*

Tabela A.3: EMIM - resultados para o domínio de Física

Col. de Textos WP01 $gl = 90,\ vm = 2e+07$			
method	*n*	$\bar{emim}$	*group*
RLUM	31	5190.859414	*a*
PopeUngar	31	3429.286832	*a b*
MostFreq	31	3429.286832	*a b*
RLDM	31	1897.535148	*c*
Col. de Textos WP02 $gl = 102,\ vm = 4e+06$			
method	*n*	$\bar{emim}$	*group*
RLUM	35	3305.903151	*a*
RLDM	35	2353.240417	*b*
PopeUngar	35	1842.456371	*b*
MostFreq	35	1842.456371	*b*
Col. de Textos WP04 $gl = 78,\ vm = 3e+06$			
method	*n*	$\bar{emim}$	*group*
RLUM	27	3790.885026	*a*
PopeUngar	27	2475.731493	*b*
MostFreq	27	2475.731493	*b*
RLDM	27	2094.232192	*b*

Tabela A.4: EMIM - resultados para as coleções de textos do IFM

Os valores da Tabela A.1 resumem as comparações múltiplas de médias para as coleções de textos de Ciência da Computação. Deve-se notar que a EMIM de RLUM é estatisticamente maior que as demais apenas para a coleção de textos de IA. Nas demais coleções de textos o resultado é estatisticamente igual ao apresentado pelos métodos de Popescul & Ungar e dos Mais Freqüentes. Já o RLDM apresenta-se sempre na última posição, exceto para a primeira coleção de textos (IA).

Na Tabela A.2 encontram-se resumidas as comparações múltiplas de médias para as coleções de textos de Química. Deve-se notar que a EMIM de RLUM é indiscutivelmente maior que os demais apenas para a coleção de textos de química analítica (AnaC), nas demais coleções de textos o resultado é estatisticamente igual ao apresentado pelos métodos de Popescul & Ungar e dos Mais Freqüentes. Já o RLDM apresenta-se isoladamente na última posição ou, na melhor das classificações empatado com o Popescul & Ungar e

Mais Freqüentes. Ainda, vale observar que, para coleção de Ciência dos Polímeros (Poly), a hipótese inicial é bastante evidente. Nessa coleção, o RLUM empata com o Popescul & Ungar e Mais Freqüentes ou pode ser considerado melhor. Isso é um grande indício de que, realmente, o RLUM provê um vocabulário representativo; dado que em avaliações subjetivas sempre foi observado que o Mais Freqüentes provê uma maior gama de vocábulos, o que auxilia a interpretação dos tópicos Moura e Rezende (2007).

Os valores da Tabela A.3 referem-se às comparações múltiplas de médias para as coleções de textos do domínio de Física. Nesse domínio, a EMIM de RLUM foi melhor que os demais em três das coleções de textos, empatando apenas em uma delas com os métodos de Popescul & Ungar e Mais Freqüentes. Já o RLDM em duas coleções empatou com os métodos de Popescul & Ungar e Mais Freqüentes, porém nas demais se apresenta na última posição.

Finalmente, na Tabela A.4 estão resumidas as comparações múltiplas de médias para as coleções de textos do IFM. Nessa coleção não foi possível obter o valor de EMIM para a coleção WP03; após dois dias rodando o programa ainda não havia terminado. Para implementar a EMIM necessita-se de um algoritmo de complexidade exponencial; logo, em uma coleção de textos com 198 textos, 9302 atributos e até 2300 elementos no vocabulário; tem-se uma tarefa muito custosa. Como esse dado não seria responsável por uma mudança tão grande na classificação dos métodos, foi ignorado. E, novamente, o RLUM mostrou o mesmo comportamento que nas outras coleções.

Deve-se observar que, sempre o Popescul & Ungar e Mais Freqüentes apresentam comportamento similar, dado que os vocabulários por eles selecionados são sempre muito próximos. Assumiu-se o Mais Freqüentes como uma boa estimativa de variabilidade do vocabulário devido às análises subjetivas anteriormente realizadas (Moura e Rezende (2007) e Moura et al. (2008c)). E, finalmente, que a hipótese inicial ($H_0 : EMIM(RLUM) \geq EMIM(P) \approx EMIM(M) \geq EMIM(D)$) foi confirmada em todos os experimentos.

A.2 Conjuntos de rótulos como expressões de busca

Para cada método, hierarquia e coleção de textos varia-se a expressão de busca, podendo-se utilizar:

- rótulo genérico: todos os conjuntos de rótulos são tratados como expressões de busca, nos quais o operador "*and*" é usado entre os termos de um mesmo conjunto. Por exemplo, o conjunto de rótulos $L_s(n_{21}) = \{husbandry, product\} \bigcup L_g(n_{21}) = \{pantanal, embrapa, catle, beef\}$ é usado como uma expressão de busca da seguinte forma: "$husbandry \wedge product \wedge pantanal \wedge embrapa \wedge catle \wedge beef$".

- rótulo específico: todos os conjuntos de rótulos são tratados como expressões de busca, nos quais o operador "*and*" é usado entre os termos de um mesmo conjunto. Por exemplo, o conjunto de rótulos $L_s(n_{21}) = \{husbandry, product\}$ é usado como uma expressão de busca da seguinte forma: "$husbandry \wedge product$".

Para verificar experimentalmente uma hipótese sobre o comportamento das medidas de interesse, sobre a Tabela 5.5, para cada um dos métodos de rotulação, calculam-se

as medidas para cada um dos nós mantidos nas quatro taxonomias. Tabulam-se essas medidas junto ao número do nó e ao método que as gerou, e, então se ajusta um modelo linear generalizado Searle (1971) para analisar a variância e obter estimativas para cada medida. O modelo considera os efeitos do nó e do método de rotulação, além da média geral:

$$\hat{m_e} = \hat{\mu} + \hat{n_{ij}} + \hat{l_m} + \hat{\epsilon}$$

- $\hat{m_e}$: estimativa da medida de avaliação após ajuste do modelo para a coleção;
- $\hat{\mu}$: média geral do modelo, sem qualquer outro efeito, para a coleção;
- $\hat{n_{ij}}$: estimativa do efeito do nó n_{ij}, que é o quanto a medida $\hat{m_e}$ desvia-se da média geral devido a esse efeito, para a coleção;
- $\hat{l_m}$: estimativa do efeito do método de rotulação,que é o quanto a medida $\hat{m_e}$ desvia-se da média geral devido à aplicação de um dado método, para a coleção;
- $\hat{\epsilon}$: efeito aleatório associado a cada estimativa na coleção.

Com o uso dessas estimativas para cada medida, $\hat{m_e}$, para cada método de rotulação $\hat{l_m}$ espera-se obter comparações mais estatisticamente confiáveis. Aplicando-se os testes de comparação múltiplas de médias a cada modelo, tem-se 128 tabelas, duas para cada coleção e 4 para cada medida em cada caso. Os valores são resumidos nas Tabelas de A.5 a A.12. Cada pequena tabela interna às maiores contem a comparação múltipla de médias das estimativas de uma das medidas em relação aos efeitos de cada método, para cada coleção de textos. Essas tabelinhas contem o nome da coleção, o modelo generalizado para a medida, a expressão de busca utilizada, o número de graus de liberdade gl utilizado na comparação múltipla de médias, a variância média $sqrt(e)$ utilizada pelo teste SNK, o número de pontos n utilizado para estimar a medida em cada coleção, a estimativa da medida em cada coleção e uma letra para *grupo*. Novamente, letras iguais representam grupos estatisticamente iguais e, a ordem dos grupos é a alfabética. Ainda, quando não foi possível calcular a estimativa com o modelo, é fornecida sua média e desvio padrão dp, ou na falta dessa estimativa o valor "." (*missing value*).

A expressão de busca com o rótulo genérico é tabulada na parte esquerda das Tabelas de A.5 a A.12. Nesse caso, observa-se a ausência de pontos calculados para a medida F para as coleções WP01, WP03, IA, HCI, Sec, AnaC, Poly, BioP, Mech e Quan. Em geral, isso ocorre quando a expressão de busca retorna conjuntos de resultados vazios, pois tanto *recall* quanto *precision* ficam zeradas e a F tem seu quociente zerado. E, em alguns casos o desbalanceamento entre o número de pontos é bem grande. Em qualquer desses casos, o RLUM sempre apresenta valores que estão nas melhores posições. Quando o valor de F para o RLUM não é o melhor, o número de pontos definidos para RLUM é maior que para os outros métodos, o que compensa os grandes valores dos demais. Logo, a medida F para o RLUM é sempre a melhor para o conjunto de rótulos genéricos.

Ainda com o rótulo genérico como expressão de busca, examinando os valores de *recall* observa-se que para o RLUM ele é sempre o melhor e muito próximo a um (valor

máximo). Enquanto o resultado de *recall* para os demais métodos é ruim. Observa-se o mesmo comportamento para *precision*; e para o *falso alarme* um comportamento bastante aceitável em todos os métodos, estatisticamente equivalentes. Logo, o RLUM tem um melhor desempenho ao se considerar o rótulo genérico do nó como expressão de busca.

Os resultados para os rótulos específicos como expressão de busca estão tabulados do lado direito das Tabelas de A.5 a A.12. Nota-se claramente a melhoria de desempenho do RLDM, enquanto os demais métodos mantêm seus desempenhos anteriores. Logo, o RLDM se aproxima do desempenho do RLUM. Assim, não há outras mudanças significativas de comportamento dos métodos e o RLUM mantém seus resultados, provando experimentalmente ter bom desempenho também quando a expressão de busca utilizada é apenas o rótulo específico do nó.

WP01 – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 1$				$L_s(n_{ij})$ $\sqrt{e} = 0.0282, gl = 60$			
método	n	F	*grupo*	método	n	F	*grupo*
$MostFreq$	1	0.697248	a	$RLDM$	30	0.807975	a
$PopeUngar$	1	0.697248	a	$RLUM$	30	0.701492	b
$RLDM$	1	0.367816	b	$PopeUngar$	17	0.401681	c
$\overline{F}_{RLUM}$	30	1.000000	*singular*	$MostFreq$	17	0.401681	c

WP01 – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0179, gl = 64$				$L_s(n_{ij})$ $\sqrt{e} = 0.0371, gl = 89$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
$RLUM$	30	0.590959	a	$RLDM$	31	0.801459	a
$RLDM$	6	0.166667	b	$RLUM$	30	0.579107	b
$PopeUngar$	31	0.032258	c	$PopeUngar$	31	0.197366	c
$MostFreq$	31	0.032258	c	$MostFreq$	31	0.197366	c

WP01 – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0007, gl = 89$				$L_s(n_{ij})$ $\sqrt{e} = 0.0541, gl = 89$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
$RLUM$	30	1.000000	a	$RLUM$	30	1.000000	a
$PopeUngar$	31	0.017265	b	$RLDM$	31	0.848667	b
$MostFreq$	31	0.017265	b	$PopeUngar$	31	0.326405	c
$RLDM$	31	0.007269	b	$MostFreq$	31	0.326405	c

WP01 – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0042, gl = 89$				$L_s(n_{ij})$ $\sqrt{e} = 0.0041, gl = 89$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
$MostFreq$	31	0.064665	a	$MostFreq$	31	0.065600	a
$PopeUngar$	31	0.064665	a	$PopeUngar$	31	0.065600	a
$RLUM$	30	0.056270	a	$RLUM$	30	0.059169	a
$RLDM$	31	0.003314	b	$RLDM$	31	0.009922	b

WP02 – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0204, gl = 82$			
método	n	F	*grupo*	método	n	F	*grupo*
$MostFreq$	1	0.645833	a	$RLUM$	34	0.940728	a
$PopeUngar$	1	0.645833	a	$RLDM$	34	0.856283	b
$RLDM$	1	0.030303	b	$PopeUngar$	26	0.781311	b
$\overline{F}_{RLUM}$	30	1.000000	*singular*	$MostFreq$	26	0.781311	b

WP02 – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0564, gl = 7$				$L_s(n_{ij})$ $\sqrt{e} = 0.0205, gl = 82$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
$RLDM$	1	1.000000	a	$RLDM$	34	1.000000	a
$RLUM$	34	0.925414	a	$RLUM$	34	0.925414	a
$PopeUngar$	5	0.200000	b	$PopeUngar$	26	0.907051	a
$MostFreq$	5	0.200000	b	$MostFreq$	26	0.907051	a

WP02 – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0014, gl = 101$				$L_s(n_{ij})$ $\sqrt{e} = 0.0582, gl = 101$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
$RLUM$	34	1.000000	a	$RLUM$	34	1.000000	a
$PopeUngar$	35	0.013626	b	$RLDM$	35	0.788195	b
$MostFreq$	35	0.013626	b	$PopeUngar$	35	0.559885	c
$RLDM$	35	0.000440	b	$MostFreq$	35	0.559885	c

WP02 – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0099, gl = 101$				$L_s(n_{ij})$ $\sqrt{e} = 0.0100, gl = 101$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
$RLUM$	34	0.051110	a	$RLUM$	34	0.051110	a
$PopeUngar$	35	0.001891	a	$PopeUngar$	35	0.004620	a
$MostFreq$	35	0.001891	a	$MostFreq$	35	0.004620	a
$RLDM$	35	0.000000	a	$RLDM$	35	0.000000	a

Tabela A.5: Resultados das medidas de interesse para **WP01 e WP02**

WP03 – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 1$

método	n	F	*grupo*
RLUM	109	0.913249	*a*
PopeUngar	2	0.686275	*b*
MostFreq	2	0.686275	*b*

$L_s(n_{ij})$, $\sqrt{e} = 0.0211, gl = 260$

método	n	F	*grupo*
RLUM	109	0.914064	*a*
RLDM	107	0.852992	*b*
PopeUngar	78	0.718027	*c*
MostFreq	79	0.718027	*c*

WP03 – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0000, gl = 1$

método	n	*prec*	*grupo*
PopeUngar	2	1.00000	*a*
MostFreq	2	1.00000	*a*
RLUM	109	0.895565	*b.*

$L_s(n_{ij})$, $\sqrt{e} = 0.0221, gl = 260$

método	n	*prec*	*grupo*
RLDM	107	1.000000	*a*
MostFreq	79	0.938608	*b*
PopeUngar	79	0.937821	*b*
RLUM	109	0.893730	*b*

WP03 – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0017, gl = 326$

método	n	*rec*	*grupo*
RLUM	109	0.994648	*a*
PopeUngar	110	0.009504	*b*
MostFreq	110	0.009504	*b*
RLDM	110	0.000000	*b*

$L_s(n_{ij})$, $\sqrt{e} = 0.0533, gl = 326$

método	n	*rec*	*grupo*
RLUM	109	0.997706	*a*
RLDM	110	0.780400	*b*
MostFreq	110	0.464428	*c*
PopeUngar	110	0.461398	*c*

WP03 – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.00, gl = 326$

método	n	F_a	*grupo*
RLUM	109	0.031854	*a*
PopeUngar	110	0.000000	*b*
RLDM	110	0.000000	*b*
MostFreq	110	0.000000	*b*

$L_s(n_{ij})$, $\sqrt{e} = 0.000, gl = 326$

método	n	F_a	*grupo*
RLUM	109	0.031854	*a*
PopeUngar	110	0.000000	*b*
RLDM	110	0.000000	*b*
MostFreq	110	0.000000	*b*

WP04 – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 1$

método	n	F	*grupo*
RLUM	26	0.986111	*a*
PopeUngar	2	0.476191	*b*
MostFreq	2	0.476191	*b*
RLDM	1	0.040816	*c*

$L_s(n_{ij})$, $\sqrt{e} = 0.00, gl = 59$

método	n	F	*grupo*
RLUM	26	0.960012	*a*
RLDM	27	0.890224	*a*
PopeUngar	18	0.690344	*b*
MostFreq	18	0.690344	*b*

WP04 – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0137, gl = 15$

método	n	*prec*	*grupo*
RLDM	1	1.000000	*a*
RLUM	26	0.976923	*a*
PopeUngar	9	0.166667	*b*
MostFreq	9	0.166667	*b*

$L_s(n_{ij})$, $\sqrt{e} = 0.0, gl = 63$

método	n	*prec*	*grupo*
RLDM	27	1.000000	*a*
RLUM	26	0.938462	*a*
PopeUngar	20	0.760000	*b*
MostFreq	20	0.760000	*b*

WP04 – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0024, gl = 77$

método	n	*rec*	*grupo*
RLUM	26	1.000000	*a*
PopeUngar	27	0.025926	*b*
MostFreq	27	0.025926	*b*
RLDM	27	0.000772	*b*

$L_s(n_{ij})$, $\sqrt{e} = 0.0621, gl = 77$

método	n	*rec*	*grupo*
RLUM	26	1.000000	*a*
RLDM	27	0.852006	*b*
MostFreq	27	0.455556	*c*
PopeUngar	27	0.455556	*c*

WP04 – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0004, gl = 77$

método	n	F_a	*grupo*
MostFreq	27	0.018075	*a*
PopeUngar	27	0.018075	*a*
RLUM	26	0.002584	*b*
RLDM	27	0.000000	*b*

$L_s(n_{ij})$, $\sqrt{e} = 0.0005, gl = 77$

método	n	F_a	*grupo*
MostFreq	27	0.018880	*a*
PopeUngar	27	0.018880	*a b*
RLUM	26	0.006801	*a c*
RLDM	27	0.000000	*c*

Tabela A.6: Resultados das medidas de interesse para **WP03 e WP04**

Hard – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0181, gl = 40$			
método	n	F	*grupo*	método	n	F	*grupo*
$RLUM$	25	0.872534	a	$RLUM$	26	0.897844	a
$PopeUngar$	1	0.630769	b	$PopeUngar$	9	0.884900	a
$MostFreq$	1	0.630769	b	$MostFreq$	9	0.884900	a
$RLDM$	2	0.521978	c	$RLDM$	25	0.872534	a

Hard – modelo:$\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0331, gl = 40$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
$PopeUngar$	1	1.00000	a	$RLDM$	26	1.000000	a
$MostFreq$	1	1.00000	a	$MostFreq$	9	1.000000	a
$RLDM$	2	1.00000	a	$PopeUngar$	9	1.000000	a
$RLUM$	25	0.858595	$b.$	$RLUM$	25	0.858595	a

Hard – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.00117, gl = 74$				$L_s(n_{ij})$ $\sqrt{e} = 0.0750, gl = 74$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
$RLUM$	25	1.000000	a	$RLUM$	25	1.000000	a
$RLDM$	26	0.039326	b	$RLDM$	26	0.873941	a
$MostFreq$	26	0.017718	b	$MostFreq$	26	0.286949	b
$PopeUngar$	26	0.017718	b	$PopeUngar$	26	0.286949	b

Hard – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0111, gl = 74$				$L_s(n_{ij})$ $\sqrt{e} = 0.0111, gl = 74$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
$RLUM$	25	0.074451	a	$RLUM$	25	0.074451	a
$PopeUngar$	26	0.000000	a	$PopeUngar$	26	0.000000	a
$RLDM$	26	0.000000	a	$RLDM$	26	0.000000	a
$MostFreq$	26	0.000000	a	$MostFreq$	26	0.000000	a

IA – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0079, gl = 100$			
método	n	F	*grupo*	método	n	F	*grupo*
$MostFreq$	1	0.553846	a	$MostFreq$	27	0.953846	a
$PopeUngar$	1	0.553846	a	$PopeUngar$	27	0.953846	a
$RLDM$	1	0.021053	b	$RLDM$	50	0.931746	a
$RLUM$	48	0.9178	$dp = 0.2246$	$RLUM$	49	0.923607	a

IA – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0113, gl = 100$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
$MostFreq$	1	1.000000	a	$MostFreq$	27	1.000000	a
$PopeUngar$	1	1.000000	a	$PopeUngar$	27	1.000000	a
$RLDM$	1	1.000000	a	$RLDM$	50	0.982759	a
$RLUM$	48	0.9230	$dp = 0.2398$	$RLUM$	49	0.911579	b

IA – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0087, gl = 146$				$L_s(n_{ij})$ $\sqrt{e} = 0.0752, gl = 146$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
$RLUM$	49	0.945578	a	$RLUM$	49	0.986395	a
$PopeUngar$	50	0.007660	b	$RLDM$	50	0.928546	a
$MostFreq$	50	0.007660	b	$PopeUngar$	50	0.500993	b
$RLDM$	50	0.000213	b	$MostFreq$	50	0.500993	b

IA – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0049, gl = 146$				$L_s(n_{ij})$ $\sqrt{e} = 0.0046, gl = 146$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
$RLUM$	49	0.034920	a	$RLUM$	49	0.036507	a
$PopeUngar$	50	0.000000	a	$RLDM$	50	0.005556	b
$RLDM$	50	0.000000	a	$PopeUngar$	50	0.000000	b
$MostFreq$	50	0.000000	a	$MostFreq$	50	0.000000	b

Tabela A.7: Resultados das medidas de interesse para **Hard e IA**

HCI – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0106, gl = 81$			
método	n	F	*grupo*	método	n	F	*grupo*
MostFreq	1	0.515152	*a*	*RLUM*	26	0.897844	*a*
PopeUngar	1	0.492308	*b*	*PopeUngar*	9	0.884900	*a*
RLDM	25	0.115385	*c*	*MostFreq*	9	0.884900	*a*
RLUM	36	0.9183	$dp = 0.2429$	*RLDM*	25	0.872534	*a*

HCI – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0157, gl = 81$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
PopeUngar	1	1.00000	*a*	*MostFreq*	24	1.000000	*a*
MostFreq	1	1.00000	*a*	*PopeUngar*	24	1.000000	*a*
RLDM	1	1.00000	*a*	*RLDM*	37	1.000000	*a*
RLUM	36	0.9076	$dp = 0.2694$	*RLUM*	36	0.907559	*b*

HCI – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0005, gl = 107$				$L_s(n_{ij})$ $\sqrt{e} = 0.0727, gl = 107$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
RLUM	36	1.000000	*a*	*RLUM*	36	1.000000	*a*
MostFreq	37	0.009377	*b*	*RLDM*	37	0.921925	*a*
PopeUngar	37	0.008825	*b*	*MostFreq*	37	0.621989	*b*
RLDM	37	0.001655	*b*	*PopeUngar*	37	0.621438	*b*

HCI – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0049, gl = 146$				$L_s(n_{ij})$ $\sqrt{e} = 0.0071, gl = 107$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
RLUM	49	0.034920	*a*	*RLUM*	36	0.046655	*a*
PopeUngar	50	0.000000	*a*	*PopeUngar*	37	0.000000	*a*
RLDM	50	0.000000	*a*	*RLDM*	37	0.000000	*a*
MostFreq	50	0.000000	*a*	*MostFreq*	37	0.000000	*a*

SEC – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0164, gl = 92$			
método	n	F	*grupo*	método	n	F	*grupo*
MostFreq	1	0.540230	*a*	*MostFreq*	21	0.952709	*a*
PopeUngar	1	0.540230	*a*	*PopeUngar*	21	0.952709	*a*
RLDM	1	0.015625	*b*	*RLDM*	54	0.895924	*a b*
RLUM	53	0.8319	$dp = 0.3157$	*RLUM*	53	0.838208	*b*

SEC – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0365, gl = 92$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
MostFreq	1	1.000000	*a*	*MostFreq*	21	1.000000	*a*
PopeUngar	1	1.000000	*a*	*PopeUngar*	21	1.000000	*a*
RLDM	1	1.000000	*a*	*RLDM*	54	0.990741	*a*
RLUM	53	0.9843	$dp = 0.0818$	*RLUM*	53	0.814195	*b*

SEC – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0022, gl = 158$				$L_s(n_{ij})$ $\sqrt{e} = 0.0750, gl = 158$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
RLUM	53	0.984277	*a*	*RLUM*	53	0.993711	*a*
PopeUngar	54	0.006853	*b*	*RLDM*	54	0.865269	*b*
MostFreq	54	0.006853	*b*	*PopeUngar*	54	0.361792	*c*
RLDM	54	0.000146	*b*	*MostFreq*	54	0.361792	*c*

SEC – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0147, gl = 158$				$L_s(n_{ij})$ $\sqrt{e} = 0.0147, gl = 158$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
RLUM	53	0.089954	*a*	*RLUM*	53	0.089954	*a*
PopeUngar	54	0.000000	*b*	*RLDM*	54	0.000448	*b*
RLDM	54	0.000000	*b*	*PopeUngar*	54	0.000000	*b*
MostFreq	54	0.000000	*b*	*MostFreq*	54	0.000000	*b*

Tabela A.8: Resultados das medidas de interesse para **HCI e SEC**

AnaC – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0206, gl = 86$			
método	n	F	*grupo*	método	n	F	*grupo*
RLUM	50	0.922434	*a*	*RLUM*	50	0.929100	*a*
MostFreq	1	0.611111	*b*	*RLDM*	50	0.914546	*a*
PopeUngar	1	0.611111	*b*	*PopeUngar*	20	0.912222	*a*
RLDM	2	0.519608	*c*	*MostFreq*	20	0.912222	*a*
AnaC – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0188, gl = 86$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
PopeUngar	1	1.00000	*a*	*MostFreq*	20	1.000000	*a*
MostFreq	1	1.00000	*a*	*PopeUngar*	20	1.000000	*a*
RLDM	2	1.00000	*a*	*RLDM*	50	0.988333	*a*
RLUM	50	0.921118	*b*	*RLUM*	50	0.921118	*a*
AnaC – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0075, gl = 149$				$L_s(n_{ij})$ $\sqrt{e} = 0.0734, gl = 149$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
RLUM	50	0.983333	*a*	*RLUM*	50	0.993333	*a*
RLDM	51	0.020000	*b*	*RLDM*	51	0.873268	*b*
MostFreq	51	0.008627	*b*	*MostFreq*	51	0.341961	*c*
PopeUngar	51	0.008627	*b*	*PopeUngar*	51	0.341961	*c*
AnaC – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0072, gl = 149$				$L_s(n_{ij})$ $\sqrt{e} = 0.0072, gl = 149$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
RLUM	50	0.041660	*a*	*RLUM*	50	0.041660	*a*
PopeUngar	51	0.000000	*a*	*PopeUngar*	51	0.000000	*a*
RLDM	51	0.000000	*a*	*RLDM*	51	0.000000	*a*
MostFreq	51	0.000000	*a*	*MostFreq*	51	0.000000	*a*
InoC – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0101, gl = 115$			
método	n	F	*grupo*	método	n	F	*grupo*
RLUM	51	0.948548	*a*	*RLUM*	51	0.948548	*a*
MostFreq	1	0.689189	*b*	*RLDM*	51	0.921489	*a*
PopeUngar	1	0.689189	*b*	*MostFreq*	34	0.907245	*a*
RLDM	2	0.520202	*c*	*PopeUngar*	34	0.907245	*a*
InoC – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0103, gl = 115$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
MostFreq	1	1.000000	*a*	*MostFreq*	34	1.000000	*a*
PopeUngar	1	1.000000	*a*	*PopeUngar*	34	1.000000	*a*
RLDM	2	1.000000	*a*	*RLDM*	51	0.990196	*a*
RLUM	51	0.934309	*b*	*RLUM*	51	0.934309	*b*
InoC – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0060, gl = 152$				$L_s(n_{ij})$ $\sqrt{e} = 0.0694, gl = 152$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
RLUM	51	1.000000	*a*	*RLUM*	51	1.000000	*a*
PopeUngar	52	0.019627	*b*	*RLDM*	52	0.889179	*b*
MostFreq	52	0.010111	*b*	*PopeUngar*	52	0.559791	*c*
RLDM	52	0.010111	*b*	*MostFreq*	52	0.559791	*c*
InoC – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$							
$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0008, gl = 152$				$L_s(n_{ij})$ $\sqrt{e} = 0.0009, gl = 152$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
RLUM	51	0.013860	*a*	*RLUM*	51	0.013860	*a*
PopeUngar	52	0.000000	*a*	*RLDM*	52	0.000614	*a*
RLDM	52	0.000000	*a*	*PopeUngar*	52	0.000000	*a*
MostFreq	52	0.000000	*a*	*MostFreq*	52	0.000000	*a*

Tabela A.9: Resultados das medidas de interesse para **AnaC e InoC**

OrgC – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0433, gl = 34$			
método	n	F	*grupo*	método	n	F	*grupo*
$MostFreq$	1	0.544118	a	$RLDM$	36	0.866026	a
$RLDM$	1	0.476923	b	$RLUM$	35	0.814839	a
$PopeUngar$	1	0.322034	c	$PopeUngar$	1	0.544118	a
$RLUM$	35	0.8148	$dp = 0.3431$	$MostFreq$	1	0.322034	a

OrgC – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0695, gl = 34$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
$PopeUngar$	1	1.00000	a	$MostFreq$	1	1.000000	a
$MostFreq$	1	1.00000	a	$PopeUngar$	1	1.000000	a
$RLDM$	1	1.00000	a	$RLDM$	36	1.000000	a
$RLUM$	35	0.7918	$dp = 0.3729$	$RLUM$	35	0.791819	a

OrgC – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0002, gl = 104$				$L_s(n_{ij})$ $\sqrt{e} = 0.0209, gl = 104$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
$RLUM$	35	1.000000	a	$RLUM$	35	1.000000	a
$MostFreq$	36	0.010382	b	$RLDM$	36	0.809624	b
$RLDM$	36	0.008698	b	$MostFreq$	36	0.010382	c
$PopeUngar$	36	0.005331	b	$PopeUngar$	36	0.005331	c

OrgC – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0128, gl = 104$				$L_s(n_{ij})$ $\sqrt{e} = 0.0128, gl = 104$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
$RLUM$	35	0.102751	a	$RLUM$	35	0.102751	a
$PopeUngar$	36	0.000000	b	$PopeUngar$	36	0.000000	b
$RLDM$	36	0.000000	b	$RLDM$	36	0.000000	b
$MostFreq$	36	0.000000	b	$MostFreq$	36	0.000000	b

Poly – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0024, gl = 104$			
método	n	F	*grupo*	método	n	F	*grupo*
$MostFreq$	1	0.689189	b	$RLDM$	46	0.978261	a
$PopeUngar$	1	0.689189	b	$RLUM$	47	0.966277	a
$RLDM$	.	.	.	$MostFreq$	31	0.966157	a
$RLUM$	47	0.9663	$dp = 0.1515$	$PopeUngar$	31	0.966157	a

Poly – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0047, gl = 104$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
$MostFreq$	1	1.000000	a	$MostFreq$	31	1.000000	a
$PopeUngar$	1	1.000000	a	$PopeUngar$	31	1.000000	a
$RLDM$	.	.	.	$RLDM$	46	0.987578	a
$RLUM$	47	0.9594	$dp = 0.1732$	$RLUM$	47	0.959433	a

Poly – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0013, gl = 140$				$L_s(n_{ij})$ $\sqrt{e} = 0.0785, gl = 140$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
$RLUM$	47	1.000000	a	$RLUM$	47	1.000000	a
$PopeUngar$	48	0.010833	b	$RLDM$	48	0.942708	a
$MostFreq$	48	0.010833	b	$PopeUngar$	48	0.611528	b
$RLDM$	48	0.000000	b	$MostFreq$	48	0.611528	b

Poly – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0019, gl = 140$				$L_s(n_{ij})$ $\sqrt{e} = 0.0019, gl = 140$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
$RLUM$	47	0.015764	a	$RLUM$	47	0.015764	a
$PopeUngar$	48	0.000000	a	$RLDM$	48	0.000614	a
$RLDM$	48	0.000000	a	$PopeUngar$	48	0.000000	a
$MostFreq$	48	0.000000	a	$MostFreq$	48	0.000000	a

Tabela A.10: Resultados das medidas de interesse para **OrgC e Poly**

BioP – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0092, gl = 105$			
método	n	F	*grupo*	método	n	F	*grupo*
MostFreq	1	0.560606	*a*	*MostFreq*	28	0.953355	*a*
PopeUngar	1	0.429752	*b*	*PopeUngar*	28	0.948682	*a*
RLDM	1	0.137255	*c*	*RLUM*	53	0.914427	*a*
RLUM	53	0.9144	$dp = 0.2281$	*RLDM*	53	0.906316	*a*

BioP – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0136, gl = 105$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
PopeUngar	1	1.00000	*a*	*MostFreq*	28	1.000000	*a*
MostFreq	1	1.00000	*a*	*PopeUngar*	28	1.000000	*a*
RLDM	1	1.00000	*a*	*RLDM*	53	0.982556	*a*
RLUM	53	1.00000	$dp = 0.000$	*RLUM*	53	0.896655	*b*

BioP – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0003, gl = 158$				$L_s(n_{ij})$ $\sqrt{e} = 0.0798, gl = 158$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
RLUM	53	1.000000	*a*	*RLUM*	53	1.000000	*a*
MostFreq	54	0.007212	*b*	*RLDM*	54	0.883155	*b*
RLDM	54	0.005068	*b*	*MostFreq*	54	0.482521	*c*
PopeUngar	54	0.001365	*b*	*PopeUngar*	54	0.480377	*c*

BioP – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0039, gl = 158$				$L_s(n_{ij})$ $\sqrt{e} = 0.0044, gl = 158$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
RLUM	53	0.036176	*a*	*RLUM*	53	0.036176	*a*
PopeUngar	54	0.000000	*b*	*RLDM*	54	0.009972	*b*
RLDM	54	0.000000	*b*	*PopeUngar*	54	0.000000	*b*
MostFreq	54	0.000000	*b*	*MostFreq*	54	0.000000	*b*

GeoP – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0007, gl = 1$				$L_s(n_{ij})$ $\sqrt{e} = 0.0172, gl = 107$			
método	n	F	*grupo*	método	n	F	*grupo*
RLUM	50	0.916928	*a*	*RLUM*	50	0.916928	*a*
MostFreq	2	0.802159	*a*	*MostFreq*	30	0.904588	*a*
PopeUngar	2	0.776120	*a*	*PopeUngar*	30	0.902852	*a*
RLDM	1	0.040404	*b*	*RLDM*	51	0.888623	*a*

GeoP – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0009, gl = 3$				$L_s(n_{ij})$ $\sqrt{e} = 0.0223, gl = 109$			
método	n	*prec*	*grupo*	método	n	*prec*	*grupo*
RLDM	1	1.000000	*a*	*RLDM*	51	1.000000	*a*
RLUM	50	0.907434	*b*	*PopeUngar*	31	1.956989	*a b*
MostFreq	3	0.666667	*c*	*MostFreq*	31	1.956989	*a b*
PopeUngar	3	0.666667	*c*	*RLUM*	50	0.907434	*a b*

GeoP – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0073, gl = 149$				$L_s(n_{ij})$ $\sqrt{e} = 0.0730, gl = 149$			
método	n	*rec*	*grupo*	método	n	*rec*	*grupo*
RLUM	50	0.996000	*a*	*RLUM*	50	0.996000	*a*
MostFreq	51	0.028098	*b*	*RLDM*	51	0.860535	*b*
PopeUngar	51	0.027087	*b*	*MostFreq*	51	0.510124	*c*
RLDM	51	0.000404	*b*	*PopeUngar*	51	0.509113	*c*

GeoP – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0076, gl = 149$				$L_s(n_{ij})$ $\sqrt{e} = 0.0076, gl = 149$			
método	n	F_a	*grupo*	método	n	F_a	*grupo*
RLUM	50	0.051354	*a*	*RLUM*	50	0.051354	*a*
PopeUngar	51	0.000213	*b*	*PopeUngar*	51	0.000420	*b*
MostFreq	51	0.000213	*b*	*MostFreq*	51	0.000420	*b*
RLDM	51	0.000000	*b*	*RLDM*	51	0.000000	*b*

Tabela A.11: Resultados das medidas de interesse para **BioP e GeoP**

Mech – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0116, gl = 81$			
método	n	F	$grupo$	método	n	F	$grupo$
$MostFreq$	1	0.555556	a	$MostFreq$	19	0.922222	a
$PopeUngar$	1	0.555556	a	$PopeUngar$	19	0.922222	a
$RLDM$	1	0.016949	b	$RLUM$	46	0.905314	a
$RLUM$	46	0.905300	$dp = 0.2562$	$RLDM$	47	0.893066	a

Mech – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0220, gl = 81$			
método	n	$prec$	$grupo$	método	n	$prec$	$grupo$
$PopeUngar$	1	1.00000	a	$MostFreq$	19	1.000000	a
$MostFreq$	1	1.00000	a	$PopeUngar$	19	1.000000	a
$RLDM$	1	1.00000	a	$RLDM$	47	1.000000	a
$RLUM$	46	0.8910	$dp = 0.2813$	$RLUM$	46	0.891024	b

Mech – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0007, gl = 137$				$L_s(n_{ij})$ $\sqrt{e} = 0.0773, gl = 137$			
método	n	rec	$grupo$	método	n	rec	$grupo$
$RLUM$	46	1.000000	a	$RLUM$	46	1.000000	a
$PopeUngar$	47	0.008183	b	$RLDM$	47	0.861529	b
$MostFreq$	47	0.008183	b	$MostFreq$	47	0.359247	c
$RLDM$	47	0.000182	b	$PopeUngar$	47	0.359247	c

Mech – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0095, gl = 137$				$L_s(n_{ij})$ $\sqrt{e} = 0.0095, gl = 137$			
método	n	F_a	$grupo$	método	n	F_a	$grupo$
$RLUM$	46	0.060359	a	$RLUM$	46	0.060359	a
$PopeUngar$	47	0.000000	b	$PopeUngar$	47	0.000000	b
$RLDM$	47	0.000000	b	$RLDM$	47	0.000000	b
$MostFreq$	47	0.000000	b	$MostFreq$	47	0.000000	b

Quan – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0115, gl = 72$			
método	n	F	$grupo$	método	n	F	$grupo$
$MostFreq$	1	0.573913	a	$RLUM$	39	0.967562	a
$PopeUngar$	1	0.573913	a	$MostFreq$	18	0.965217	a
$RLDM$	1	0.047619	b	$PopeUngar$	18	0.965217	a
$RLUM$	39	0.967600	$dp = 0.1414$	$RLDM$	40	0.909484	a

Quan – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0078, gl = 72$			
método	n	$prec$	$grupo$	método	n	$prec$	$grupo$
$MostFreq$	1	1.000000	a	$PopeUngar$	18	1.000000	a
$PopeUngar$	1	1.000000	a	$MostFreq$	18	1.000000	a
$RLDM$	1	1.000000	a	$RLDM$	40	1.000000	a
$RLUM$	50	0.960600	$dp = 0.1718$	$RLUM$	39	0.960570	a

Quan – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0011, gl = 116$				$L_s(n_{ij})$ $\sqrt{e} = 0.0846, gl = 116$			
método	n	rec	$grupo$	método	n	rec	$grupo$
$RLUM$	39	0.994872	a	$RLUM$	39	0.994872	a
$MostFreq$	40	0.010061	b	$RLDM$	40	0.889360	a
$PopeUngar$	40	0.010061	b	$PopeUngar$	40	0.426728	b
$RLDM$	40	0.000610	b	$MostFreq$	40	0.426728	b

Quan – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0006, gl = 116$				$L_s(n_{ij})$ $\sqrt{e} = 0.0006, gl = 116$			
método	n	F_a	$grupo$	método	n	F_a	$grupo$
$RLUM$	39	0.011072	a	$RLUM$	39	0.011072	a
$PopeUngar$	40	0.000000	a	$PopeUngar$	40	0.000000	a
$RLDM$	40	0.000000	a	$RLDM$	40	0.000000	a
$MostFreq$	40	0.000000	a	$MostFreq$	40	0.000000	a

Tabela A.12: Resultados das medidas de interesse para **Mech e Quan**