O efeito do uso de diferentes formas de extração de termos na compreensibilidade e representatividade dos termos em coleções textuais na língua portuguesa

Merley da Silva Conrado

O efeito do uso de diferentes formas de extração de termos na compreensibilidade e representatividade dos termos em coleções textuais na língua portuguesa

Merley da Silva Conrado

Orientadora: *Profª Drª Solange Oliveira Rezende*

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

USP - São Carlos
Julho/2009

Dedicatória

Aos meus amados pais,
Lindinalva e Conrado.

Agradecimentos

A Deus, por tudo e sempre.

Aos meus amados pais, Lindinalva (Lili) e Gecé Delmiro Conrado, por serem meu esteio em todas as horas da vida. Por "segurarem a barra", pelo apoio, direção, incentivo, amor e outras mil palavras mais. Esta conquista é nossa! À minha querida irmã Midiam (Mi) e cunhado Carlos (Caio), pelo amor e por acreditarem que eu seria capaz. À minha linda sobrinha Tássyla (Nanica) por alegrar sempre meus dias. Agradeço também aos meus queridos irmãos, Gecé Júnior, Márcio e Marcelo, e ao mais novo sobrinho Lucas! Mesmo a distância nossos corações continuam sempre juntos!

À minha orientadora Solange Oliveira Rezende, que mesmo estando muito ocupada, atendeu-me em todas as horas. Obrigada pelas inúmeras conversas, seja de assunto profissional ou pessoal. Agradeço também pelo conhecimento transferido e por sua amizade conquistada. Realmente valeu a pena!

Ao Víctor Laguna (Peru), pelo amor, compreensão e conversas. Com seu jeito engraçado nos momentos descontraídos e seu jeito adulto nos momentos sérios, agradeço o seu importante apoio e presença nas fases difíceis nesse tempo que estamos juntos.

Ao grupo de pesquisa "mais modesto" de Mineração de Textos, pela colaboração imensa no desenvolvimento desse trabalho, pelas reuniões e descontrações, a saber: Fabiano Santos, Rafael Rossi, Solange Rezende e Tatiane Nogueira. Especialmente ao Bruno Nogueira, Maria Fernanda Moura e Ricardo Marcacini.

Ao professor Thiago Pardo, pela simpatia sempre, atenção e colaboração para este trabalho. Ao professor Eduardo Hruschka pela colaboração na qualificação deste trabalho. Ao meu ex-orientador Edmundo Spoto e ao professor Auri Vincenzi.

À equipe do Projeto Agência Embrapa, em especial a Adriana Delfino, Maria Fernanda Moura e Leandro Oliveira; pelos dados, dicas de interpretação e conversão da Agência para o nosso formato. À Binagri, pela elaboração e manutenção do Thesagro; em especial a Lucia Santos e Neuza Rangel, pelo fornecimento da versão XML mais recente do Thesagro. Aos especialistas, que devido a sua disposição possibilitaram efetivamente a avaliação subjetiva deste trabalho.

Aos antigos e de sempre amigos, Aline Fávero (Linão), Andréia Mesquita, Cláudia Casagrande (Claudinha), Danilo Dini (Dani), Gabriela Reis (Gabi), Jeniffer Nayara, Laura

Maria e Nádia Castilho (Nádião). Ao Anderson Marciano pelo carinho e aprendizado compartilhado. Aos amigos que cultivei durante esta jornada, Bruno Nogueira, Gustavo Cervini (Guga), Marcos Cintra, Maria Fernanda Moura, Marília, Sandro Bianchini (KLB) - também pelo logo da ferramenta e Vanessa Borges (Van).

Aos amigos do "predinho", Celso, Eloísa Regueiro, Priscila Aleixo, Rafael, especialmente à Juliana Salvador (Jú) e ao Luis Antônio (Potó).

Aos companheiros do LABIC e outros laboratórios. Para citar alguns, Alfredo Roa (Shakira), André Abe Vicente (Magrelo), André Domingues (André Maluquinho), André Maletzke, Cristina Oliveira (Kika), Débora Medeiros, Edson Takashi, Ellen Barbosa, Everton Cherman, Ígor Braga, Leonardo Almeida, Matheus Soares (Caneca), Paulo Gabriel (Gambi), Leandro Paganotti (Nanico), Lucía Castro, Marcella Letícia, Márcio Basgalupp, Marco Aurélio (Marcão), Marllos Prado, Paula Donegan, Rafael Giusti e Robson Motta.

Aos professores e funcionários do ICMC-USP.

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), pelo apoio financeiro.

Finalmente, agradeço a todos que me ajudaram direta ou indiretamente.

Resumo

A extração de termos em coleções textuais, que é uma atividade da etapa de Pré-Processamento da Mineração de Textos, pode ser empregada para diversos fins nos processos de extração de conhecimento. Esses termos devem ser cuidadosamente extraídos, uma vez que os resultados de todo o processo dependerão, em grande parte, da "qualidade" dos termos obtidos. A "qualidade" dos termos, neste trabalho, abrange tanto a representatividade dos termos no domínio em questão como sua compreensibilidade. Tendo em vista sua importância, neste trabalho, avaliou-se o efeito do uso de diferentes técnicas de simplificação de termos na compreensibilidade e representatividade dos termos em coleções textuais na Língua Portuguesa. Os termos foram extraídos seguindo os passos da metodologia apresentada neste trabalho e as técnicas utilizadas durante essa atividade de extração foram a radicalização, lematização e substantivação. Para apoiar tal metodologia, foi desenvolvida uma ferramenta, a ExtraT (Ferramenta para Extração de Termos). Visando garantir a "qualidade" dos termos extraídos, os mesmos são avaliados objetiva e subjetivamente. As avaliações subjetivas, ou seja, com o auxílio de especialistas do domínio em questão, abrangem a representatividade dos termos em seus respectivos documentos, a compreensibilidade dos termos obtidos ao utilizar cada técnica e a preferência geral subjetiva dos especialistas em cada técnica. As avaliações objetivas, que são auxiliadas por uma ferramenta desenvolvida (a TaxEM - Taxonomia em XML da Embrapa), levam em consideração a quantidade de termos extraídos por cada técnica, além de abranger também a representatividade dos termos extraídos a partir de cada técnica em relação aos seus respectivos documentos. Essa avaliação objetiva da representatividade dos termos utiliza como suporte a medida CTW (*Context Term Weight*). Oito coleções de textos reais do domínio de agronegócio foram utilizadas na avaliação experimental. Como resultado foram indicadas algumas das características positivas e negativas da utilização das técnicas de simplificação de termos, mostrando que a escolha pelo uso de alguma dessas técnicas para o domínio em questão depende do objetivo principal pré-estabelecido, que pode ser desde a necessidade de se ter termos compreensíveis para o usuário até a necessidade de se trabalhar com uma menor quantidade de termos.

Abstract

The task of term extraction in textual domains, which is a subtask of the text pre-processing in Text Mining, can be used for many purposes in knowledge extraction processes. These terms must be carefully extracted since their quality will have a high impact in the results. In this work, the quality of these terms involves both representativity in the specific domain and comprehensibility. Considering this high importance, in this work the effects produced in the comprehensibility and representativity of terms were evaluated when different term simplification techniques are utilized in text collections in Portuguese. The term extraction process follows the methodology presented in this work and the techniques used were radicalization, lematization and substantivation. To support this metodology, a term extraction tool was developed and is presented as ExtraT. In order to guarantee the quality of the extracted terms, they were evaluated in an objective and subjective way. The subjective evaluations, assisted by domain specialists, analyze the representativity of the terms in related documents, the comprehensibility of the terms with each technique, and the specialist's opinion. The objective evaluations, which are assisted by TaxEM and by Thesagro (National Agricultural Thesaurus), consider the number of extracted terms by each technique and their representativity in the related documents. This objective evaluation of the representativity uses the CTW measure (Context Term Weight) as support. Eight real collections of the agronomy domain were used in the experimental evaluation. As a result, some positive and negative characteristics of each techniques were pointed out, showing that the best technique selection for this domain depends on the main pre-established goal, which can involve obtaining better comprehensibility terms for the user or reducing the quantity of extracted terms.

Sumário

Sumário i
Lista de Figuras iii
Lista de Tabelas v
Lista de Algoritmos vii

1 Introdução **1**

2 Processo de Mineração de Textos **5**
2.1 Considerações Iniciais 5
2.2 Etapas do Processo de Mineração de Textos 6
2.2.1 Identificação do Problema 6
2.2.2 Pré-Processamento 6
2.2.3 Extração de Padrões 9
2.2.4 Pós-Processamento e Utilização do Conhecimento 10
2.3 Algumas aplicações da Mineração de Textos 10
2.4 Exemplo de instanciação da Mineração de Textos: Metodologia TopTax . . 11
2.5 Considerações Finais 13

3 Extração de Termos para a Mineração de Textos **15**
3.1 Considerações Iniciais 15
3.2 Técnicas de Simplificação de Termos 16
3.2.1 Radicalização 16
3.2.2 Lematização 19
3.2.3 Substantivação 20
3.3 Extração de Termos Simples e Compostos 21
3.4 Vocabulário Controlado 23
3.5 Trabalhos Relacionados com a Extração de Termos 24
3.6 Considerações Finais 26

4 Metodologia para a Utilização de Diferentes Formas de Extração de Termos a partir de Coleções Textuais **27**
4.1 Considerações Iniciais 27

4.2 Descrição da Metodologia para a Utilização de Diferentes Formas de Extração de Termos . . . 27
4.2.1 Fase 1: Preparação dos Textos . . . 28
4.2.2 Fase 2: Extração dos Termos . . . 30
4.3 Ferramenta ExtraT . . . 36
4.4 Considerações Finais . . . 39

5 Avaliação dos Termos Extraídos 41
5.1 Considerações Iniciais . . . 41
5.2 Abordagem para Avaliação dos Termos Extraídos . . . 41
5.2.1 Avaliações Subjetivas dos Termos Extraídos . . . 42
5.2.2 Avaliações Objetivas dos Termos Extraídos . . . 45
5.3 Bases de Textos Utilizadas . . . 50
5.4 Avaliação Experimental . . . 51
5.4.1 Avaliação 1 - O uso da Metodologia de Extração de Termos Proposta 51
5.4.2 Avaliação 2 - Quantidade de Termos Obtidos Utilizando as Técnicas de Simplificação de Termos . . . 53
5.4.3 Avaliação 3 - Representatividade Objetiva dos Termos Extraídos . . 53
5.4.4 Avaliação 4 - Representatividade Subjetiva dos Termos Extraídos . 55
5.4.5 Avaliação 5 - Compreensibilidade dos Termos Extraídos . . . 59
5.4.6 Avaliação 6 - Preferência dos Especialistas . . . 60
5.4.7 Avaliação 7 - Complexidades dos Algoritmos das Técnicas de Simplificação de Termos . . . 61
5.5 Considerações Finais . . . 63

6 Conclusões e Trabalhos Futuros 65

Lista de Figuras

2.1 Etapas do processo de Mineração de Textos (**?**) 6
2.2 Etapas do Pré-Processamento da coleção de textos - Adaptada de **?** 7
2.3 TopTax - Metodologia de extração de taxonomias de tópicos (**?**) 12

4.1 Preparação dos textos para a metodologia de extração de termos 28
4.2 Extração de termos 30
4.3 Exemplo da obtenção das palavras substantivadas 32
4.4 Ferramenta para extração de termos - ExtraT 38

5.1 Exemplo de taxonomias com a mesma estrutura hierárquica utilizando as diferentes técnicas de simplificação de termos 43
5.2 Avaliação do ramo selecionado 45
5.3 Processo para a preparação do conteúdo das Agências 48
5.4 Exemplos de ***RT*** e ***USE*** 48
5.5 Vocabulário expandido 49
5.6 Exemplo da visualização da hierarquia 50
5.7 Redução do número de termos utilizando a técnica de radicalização 52
5.8 Redução do número de termos utilizando a técnica de lematização 52
5.9 Redução do número de termos utilizando a técnica de substantivação . . . 53
5.10 Ramos selecionados para avaliação da técnica de radicalização 56
5.11 Ramos selecionados para avaliação da técnica de lematização 57
5.12 Ramos selecionados para avaliação da técnica de substantivação 57
5.13 Avaliação subjetiva quanto a compreensibilidade dos termos obtidos utilizando as técnicas 60
5.14 Avaliação subjetiva quanto a técnica de preferência dos especialistas 60

Lista de Tabelas

2.1 Padrão de matriz atributo-valor 8

3.1 Algoritmos para radicalização - Adaptada de **?** 18
3.2 Exemplos de lematização 19

4.1 Parágrafos retirados de um documento original da base de textos 29
4.2 Parágrafos, referentes aos parágrafos mostrados na Tabela 4.1, preparados para serem utilizados na metodologia de extração de termos 29
4.3 Palavras radicalizadas 31
4.4 Palavras lematizadas 31
4.5 Palavras substantivadas 31
4.6 Trigramas obtidos com a técnica de radicalização a partir dos parágrafos de um documento exemplo mostrados na Tabela 4.1 33
4.7 Trigramas obtidos com a técnica de lematização a partir dos parágrafos de um documento exemplo mostrados na Tabela 4.1 34
4.8 Trigramas obtidos com a técnica de substantivação a partir dos parágrafos de um documento exemplo mostrados na Tabela 4.1 34
4.9 Exemplos dos valores obtidos pela aplicação do teste da razão de máxima verossimilhança para alguns bigramas e trigramas 35

5.1 Exemplo do uso da medida CTW neste trabalho 49
5.2 Descrição das bases de textos 51
5.3 Quantidade de unigramas, bigramas e trigramas extraídos 54
5.4 Pontuação CTW para cada técnica 55
5.5 Agrupamento das notas nos ramos 58

6.1 O uso das técnicas de simplificação de termos 67

List of Algorithms

4.1 Substantivação das palavras da coleção de textos 32
5.1 Radicalização 61
5.2 Lematização 62
5.3 Substantivação (**?**) 63

Capítulo

1

Introdução

A quantidade de informação adicionada anualmente no universo digital durante o período de 2006 a 2010 tem sido aproximadamente de 988 hexabytes (**?**). Esta rápida expansão na quantidade de informação é notada no cotidiano das pessoas.

Em se tratando de informações textuais, estudos indicam que 80% da informação das corporações no mundo é representada por documentos textuais[1], considerando que essa é a forma mais natural de armazenar informações. Muitos desses documentos são armazenados em meios eletrônicos e uma boa parte é lançada diariamente na Web, formando grandes coleções de dados, como relatórios, especificações de produtos, resumos, notas, correspondência eletrônica e toda sorte de publicações eletrônicas textuais - bibliotecas virtuais e acervos documentais variados (**?**).

Nesses documentos estão contidas informações relevantes que podem ser utilizadas para os mais diversos objetivos, como para servir de vantagem competitiva ou como suporte à tomada de decisão dentro das empresas. Nesse sentido, torna-se necessário o investimento em técnicas capazes de lidar com essa quantidade de documentos. Tais técnicas devem ser capazes de transformar essas informações de forma automática ou semi-automática em conhecimento útil organizado. Para isso, pode-se fazer uso do processo de Mineração de Textos.

A Mineração de Textos visa, com a utilização de termos que representam os documentos[2], identificar padrões, tendências e regularidades em textos escritos em língua natural. Sendo que seu processo é dividido em cinco grandes etapas: Identificação do Problema, Pré-Processamento, Extração de Padrões, Pós-Processamento e Utilização do Conhecimento. Essas etapas podem ser instanciadas de acordo com os objetivos do processo (**?**).

A etapa de Pré-Processamento é uma das etapas mais importantes da Mineração

[1]Delphi Group - http://www.delphigroup.com/

[2]Os termos *texto* e *documento* são utilizados indistintamente neste trabalho.

de Textos, uma vez que as atividades realizadas nessa etapa influenciam diretamente a "qualidade" dos resultados finais de todo o processo de Mineração de Textos. No Pré-Processamento ocorrem transformações necessárias para a adequação do formato dos documentos disponíveis para a extração de conhecimento. Essas transformações abrangem o tratamento, a limpeza e a redução dos dados dos documentos, bem como a extração de termos que representem a coleção textual. Esses termos devem ser cuidadosamente extraídos, uma vez que os resultados de todo o processo dependerão, em grande parte, da "qualidade" dos termos obtidos. A "qualidade" dos termos, neste trabalho, abrange tanto a representatividade dos termos no domínio em questão como sua compreensibilidade.

Nem sempre todos os termos do documento são importantes para descrever seu conteúdo. Dessa forma, quando se tem sistemas manuais, tais termos ou expressões são determinados manualmente por especialistas do domínio da coleção de textos. Entretanto, quando se tem sistemas automáticos ou semi-automáticos, são utilizadas técnicas específicas para este fim.

As técnicas que podem ser aplicadas para auxiliar a extração de termos são as técnicas de simplificação de termos, que visam reduzir as formas como as palavras dos documentos aparecem. A radicalização ("stemmização" ou *stemming*), por exemplo, visa reduzir as palavras às suas formas inflexionáveis e as vezes reduzir às suas derivações. A técnica de lematização (ou redução à forma canônica) tem como objetivo agrupar as variantes de um termo em um único lema. E, por fim, tem-se a técnica de substantivação (ou "nominalização"), na qual são derivados substantivos de palavras de outras categorias morfológicas.

O impacto dessas técnicas utilizadas para extração de termos é bem perceptível em tarefas de organização de informação em que a compreensibilidade, representatividade e o número de termos extraídos têm impacto direto na interpretabilidade dos modelos extraídos.

Assim, este trabalho tem como objetivo avaliar o efeito do uso dessas três diferentes técnicas para a extração de termos de um domínio específico considerando a compreensibilidade e representatividade dos termos em coleções textuais reais e não-rotuladas na Língua Portuguesa.

Para facilitar essa comparação é apresentada uma metodologia proposta para apoiar a extração de termos utilizando diferentes técnicas de simplificação dos termos. Como auxílio a aplicação desta metodologia foi desenvolvida a ferramenta ExtraT (Ferramenta para Extração de Termos). Para essa extração, primeiramente, deve-se visar pela garantia da "qualidade" desta coleção de textos a ser trabalhada por meio da preparação dos documentos, pois estes podem prover de diferentes repositórios (bases de textos). Em seguida, é realizada a extração de termos importantes do domínio dessa coleção, sendo que durante essa extração são criados novos conjuntos de termos que possuam significado importante para a coleção textual de um determinado domínio, utilizando para isso as técnicas de simplificação de termos separadamente.

Adicionalmente, esses conjuntos de termos são avaliados, visando ressaltar algumas das características positivas e negativas da utilização das técnicas de simplificação de termos. Essa avaliação abrange avaliações subjetivas e objetivas e ambas fazem uso de taxonomias *gold*. Neste trabalho, as taxonomias *gold* são consideradas como taxonomias validadas e que podem ser utilizadas como modelo para experimentos.

As avaliações subjetivas foram auxiliadas por especialistas do domínio e abrangem a representatividade dos termos em seus respectivos documentos; a compreensibilidade dos termos obtidos ao utilizar cada técnica de simplificação; e qual técnica os especialistas indicam para ser utilizada para determinado domínio. As avaliações objetivas levam em consideração a quantidade de termos extraídos por cada técnica, além de abranger também a representatividade dos termos extraídos a partir de cada técnica em relação aos seus respectivos documentos. A avaliação da representatividade objetiva utiliza como suporte a medida CTW (*Context Term Weight*) (**?**). Para auxiliar nas avaliações objetivas foi desenvolvida a ferramenta TaxEM (Taxonomia em XML da Embrapa), detalhada no Capítulo 5.

Finalmente, os termos extraídos e avaliados podem ser utilizados para os mais diversos fins. Uma aplicação direta dos termos extraídos neste trabalho diz respeito ao Projeto TopTax[3] (metodologia de extração de taxonomias de tópicos), o qual tem por objetivo auxiliar especialistas de um domínio específico a organizar e manter a informação do mesmo, por meio da criação e manipulação de uma taxonomia de tópicos para domínios específicos.

A principal hipótese deste trabalho é que alguma das formas de extração de termos será de auxílio efetivo à geração automática ou semi-automática de uma taxonomia de tópicos de um domínio específico, apoiando os especialistas dos domínios de conhecimento na organização e manutenção da informação.

Como hipótese secundária, considera-se também que, para cada uma das formas de extração de termos será obtido um resultado (taxonomia de tópicos final), conseqüentemente, com a comparação desses resultados serão gerados diferentes impactos nos especialistas, o que influenciará na escolha dos especialistas por alguma técnica de simplificação de termos. Espera-se também que as técnicas de substantivação e lematização serão eleitas pelos especialistas por, possivelmente, gerarem termos mais compreensíveis do que a técnica de radicalização devido a agressividade da simplificação desta última técnica.

Outra hipótese deste trabalho é que a complexidade dos algoritmos de cada técnica será um fator que influenciará os especialistas na escolha de alguma delas.

Adicionalmente, com o apoio da metodologia de extração de termos proposta será possível diminuir consideravelmente a quantidade de termos. Considera-se que a utilização das técnicas de lematização e substantivação obterão quantidades maiores de termos do que quando utilizada a técnica de radicalização. O uso dessa metodologia obterá também uma quantidade satisfatória de termos representativos (importantes) para o domínio em

[3]TopTax - http://labic.icmc.usp.br/projects/researchproject.2008-06-04.9415524093

questão além de termos representativos segundo os especialistas do domínio.

Para apresentar o trabalho desenvolvido, esta dissertação está organizada como segue. Neste capítulo foram apresentados o contexto, os objetivos e as hipóteses deste trabalho. No Capítulo 2, o processo de Mineração de Textos é definido, destacando-se a etapa de Pré-Processamento de documentos, que é o foco deste trabalho. No Capítulo 3, a extração de termos para a Mineração de Textos é discutida e, para dar suporte a esta extração, são apresentadas três técnicas de simplificação de termos, que são a radicalização, a lematização e a substantivação. A metodologia proposta neste trabalho para a extração de termos, utilizando diferentes formas de simplificação de termos, a partir de coleções textuais é descrita no Capítulo 4. E a abordagem proposta para avaliação dos termos extraídos, bem como a avaliação experimental juntamente com os seus resultados e análises são descritas no Capítulo 5. No Capítulo 6, são descritos as conclusões e os trabalhos futuros. Por fim, são apresentadas as referências bibliográficas utilizadas.

Capítulo

2

Processo de Mineração de Textos

2.1 Considerações Iniciais

O processo de Mineração de Textos (MT) é utilizado para descobrir padrões e conhecimento útil em um conjunto de dados textuais. As bases textuais utilizadas no processo de Mineração de Textos são coleções de documentos em linguagem natural, sem formato predefinido para seus conteúdos (**?**). Os textos que não obedecem a um padrão de formatação são chamados **não-estruturados**; os que seguem algum padrão, como textos científicos que são divididos em tópicos pré-definidos e livros divididos em capítulos, são denominados textos **semi-estruturados**. Neste trabalho, tratamos os textos representados em uma linguagem de marcação como **textos estruturados**.

A Mineração de Textos foi inspirada na Mineração de Dados e por vezes considerada uma subárea da mesma. A Mineração de Dados pode ser definida como a extração de conhecimento implícito, previamente desconhecido e potencialmente útil, ou a busca por relações e padrões globais existentes em bases de dados (**?**). Mineração de Textos é uma área interdisciplinar, relativamente nova, que engloba: processamento de língua natural, mais particularmente a lingüística computacional; aprendizado de máquina; recuperação de informação; mineração de dados; estatística e visualização de informação. A MT oferece a possibilidade de se trabalhar com dados textuais não estruturados, enquanto a Mineração de Dados trabalha com dados estruturados.

O foco deste trabalho é a extração de termos (atributos, características), atividade desenvolvida na etapa de Pré-Processamento do processo de Mineração de Textos, na qual são executadas as transformações necessárias para a adequação do formato dos textos disponíveis para a extração de conhecimento.

A seguir, são descritas as etapas do processo de Mineração de Textos.

2.2 Etapas do Processo de Mineração de Textos

A Mineração de Textos é dividida em cinco grandes etapas: Identificação do Problema, Pré-Processamento, Extração de Padrões, Pós-Processamento e Utilização do Conhecimento, conforme ilustrado na Figura 2.1. Este processo pode ser instanciado de acordo com o objetivo pré-estabelecido. Cada etapa é discutida a seguir, bem como as definições e conceitos necessários à sua compreensão.

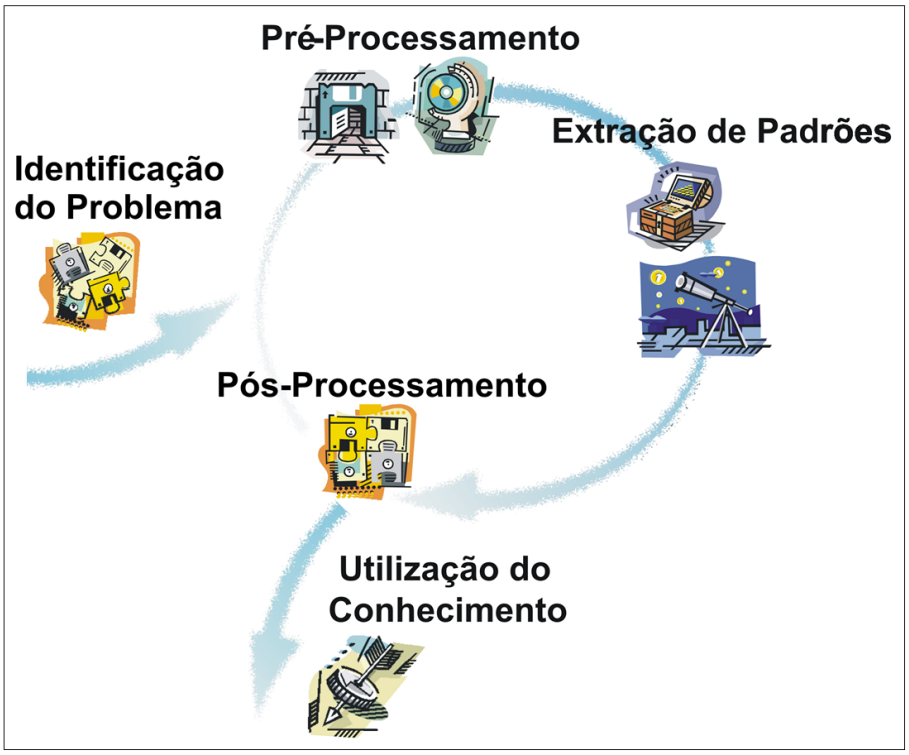


Figura 2.1: Etapas do processo de Mineração de Textos (**?**)

2.2.1 Identificação do Problema

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 subdomínio do problema, a coleção de textos a ser analisada ou sua fonte de busca, a existência de algum conhecimento prévio de 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 Textos, pois a mesma fornece subsídios a todo o processo, permitindo identificar requisitos e possíveis ferramentas para cada passo.

2.2.2 Pré-Processamento

Na etapa de Pré-Processamento encontra-se a principal diferença entre o processo de Mineração de Dados e o de Mineração de Textos. Nos dois casos, o problema resume-

se ao fato dos dados não estarem sempre em um formato adequado para a Extração de Padrões; é necessário adequá-los para um formato manipulável por algoritmos de extração de conhecimento, além de aplicar-lhes um processo de tratamento, limpeza e, em geral, redução do volume de textos, sempre lhes preservando as características necessárias para que os objetivos do processo de mineração sejam cumpridos (**?**).

Nessa etapa, a partir da coleção de textos, ocorre a preparação dos textos para poder extrair os termos importantes da coleção. Estes termos devem discriminar a coleção de textos que, por sua vez, é representada na forma matriz atributo-valor, conforme ilustrado na Figura 2.2 e detalhado a seguir.

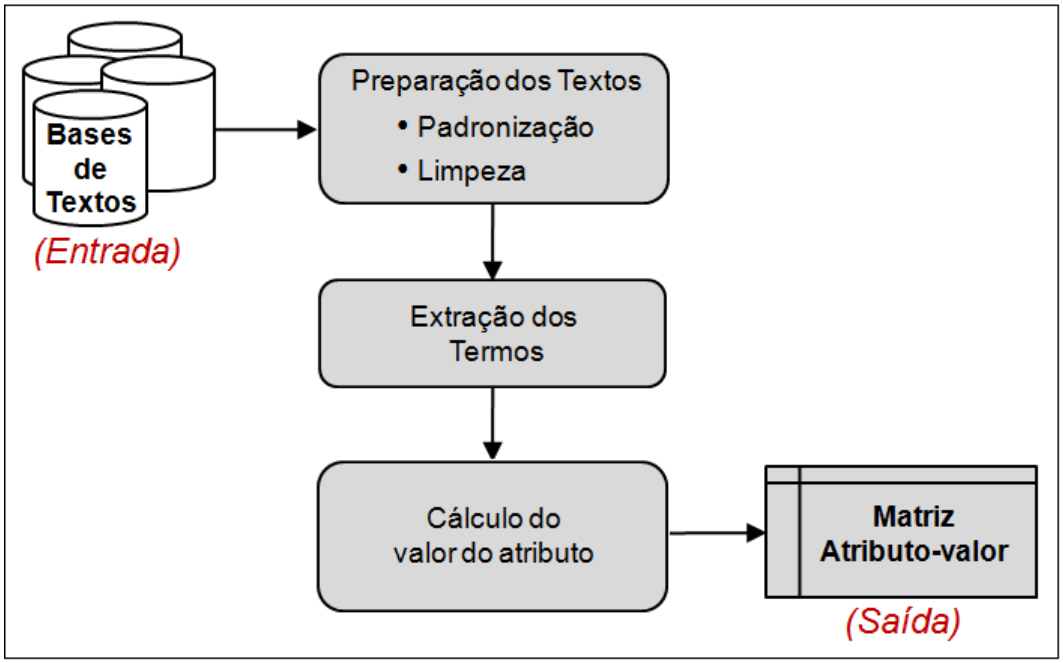


Figura 2.2: Etapas do Pré-Processamento da coleção de textos - Adaptada de **?**

Os dados de entrada correspondem à coleção de documentos de interesse (**coleção de textos**). Esta coleção pode estar em diferentes formatos, como documentos hipertextos ou formatos *.pdf*. Dependendo de como os documentos foram armazenados ou gerados, há a necessidade de padronizar as formas em que se encontram. Na **padronização dos textos**, geralmente, os documentos são convertidos para o forma de texto plano sem formatação.

Em seguida, na **limpeza dos textos**, elimina-se dos mesmos as *stopwords*, que são aquelas palavras que nada acrescentam à representatividade dos termos ou que sozinhas nada significam, como artigos, pronomes e advérbios. Sua definição depende dos objetivos estabelecidos e do domínio do conhecimento, e seu uso pode ser considerado como uma "forma de seleção inicial de termos, pois ajuda a reduzir a dimensão final do vocabulário eliminando termos não significativos para a análise" (**?**). O conjunto de *stopwords* é chamado de *stoplist*. Segundo **?**, tal eliminação reduz significativamente a quantidade de palavras que são armazenadas, podendo-se deduzir que a quantidade de palavras que são processadas e o tempo gasto para isto também são reduzidos.

O número de termos que compõem os textos da coleção após a remoção das *stopwords* pode ainda ser muito grande e, em geral, alguns termos se referem ao mesmo conceito. Assim, na **extração dos termos**, faz-se necessária a busca por padrões que simplifiquem as diversas formas de apresentação de termos com o mesmo significado essencial. Entre as técnicas mais utilizadas para este fim, encontram-se a radicalização, a lematização e a substantivação. Além disso, a partir dos termos simplificados, existe a possibilidade de extrair termos compostos e um vocabulário controlado, como explicado na Seção 3.2 do Capítulo 3.

Uma vez extraídos os termos que representam a coleção, passa-se à estruturação da coleção textual em uma representação que seja manipulável por algoritmos utilizados na etapa de Extração de Padrões. A representação mais comumente aplicada à coleção de textos, em tarefas de Mineração de Textos com ênfase em métodos estatísticos, leva à forma matricial dos documentos (**?**). Esta forma matricial dos documentos engloba os termos utilizados e algum **valor do atributo**, sendo que este valor pode ser, basicamente, binário (presença ou ausência do termo no texto) ou freqüência de ocorrências do termo na coleção de textos, entre outras (**?**). A escolha e cálculo desse valor do atributo depende do objetivo, e possibilita calcular a representatividade de cada termo em seu respectivo documento e/ou na coleção de textos.

Normalmente, os termos extraídos de uma coleção textual são tratados como um conjunto desordenado de palavras independentes - conhecido como *bag of words*. A principal limitação dessa representação é considerar apenas a co-ocorrência dos termos. Por exemplo, polissemias (ocorrências de termos cuja grafia é idêntica ou muito próxima), mas cujo significado depende do contexto, são tratadas como termos únicos e sinônimos como termos independentes (**?**), não obtendo, portanto, uma descrição fiel do conteúdo do documento. O efeito negativo das ocorrências de polissemias na coleção de textos pode ser bastante reduzido quando a coleção vem de um *corpus de especialidade*, ou seja, de uma coleção de textos proveniente de um domínio de conhecimento específico.

A representação *bag of words* pode ser estruturada por meio da **matriz 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 (termo), a_{ij} é o valor do j-ésimo atributo no i-ésimo documento. O y_i, que representa a classe do i-ésimo documento, não está presente na matriz quando a coleção de textos é não-rotulada.

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

Tabela 2.1: Padrão de matriz atributo-valor

A matriz atributo-valor é caracteristicamente esparsa e possui alta dimensionalidade, dado que cada palavra presente nos documentos é candidata a atributo da tabela. Entre-

tanto, várias dessas palavras não estão presentes em todos os documentos, fazendo com que seus valores internos sejam nulos - ou inexistentes.

Essa representação (matriz atributo-valor) permite o emprego de um grande leque de ferramentas de aprendizado de máquina, aplicadas ao contexto de Mineração de Textos.

Após a representação matricial dos documentos, pode-se, ainda, diminuir a quantidade de termos a ser trabalhada utilizando abordagens de seleção de termos, como o método de Luhn, Salton, *Term Variance*. Estes métodos são detalhados no trabalho de **?**.

Deve-se ressaltar, que esta etapa de Pré-Processamento pode ser redefinida e então repetida após as próximas etapas, uma vez que a descoberta de alguns padrões pode levar a estabelecer melhorias a serem empregadas sobre o valor do atributo utilizado na matriz atributo-valor. Por exemplo, o uso de algum fator de ponderação - estabelecido a partir da presença ou ausência de um conjunto de atributos no texto; o uso de algum tipo de taxonomia após a descoberta da mesma; o descarte ou inclusão de algum vocábulo ou combinação deles.

2.2.3 Extração de Padrões

Na etapa de Extração de Padrões são definidas as tarefas a serem realizadas de acordo com o objetivo do processo. Essas tarefas podem ser preditivas ou descritivas.

As **tarefas preditivas** consistem na generalização de exemplos ou experiências passadas com respostas conhecidas. Essas tarefas utilizam os chamados modelos de aprendizado de máquina supervisionado, uma vez que as categorias são sempre pré-conhecidas e disponíveis junto aos dados. Estes modelos podem ser divididos em tarefas de classificação, referente ao processo em que o atributo classe tem valor categórico, e tarefas de regressão, na qual busca-se predizer valores de variáveis que possuem valores contínuos.

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. 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. Já como exemplo de uma tarefa de regressão, pode-se considerar a predição do ganho ou da perda de produção em uma cultura, com base no estudo de diferentes adubações no qual a medida de produtividade é um atributo contínuo.

Quando apenas parte dos exemplos são rotulados, podem ser aplicados modelos de aprendizado de máquina semi-supervisionados, em que a informação dos rotulados ajuda a gerar modelos para rotular os dados 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 (**??**).

As **tarefas descritivas**, por sua vez, consistem na identificação de comportamentos intrínsecos da coleção de textos, sendo que esses dados são exemplos não rotulados ou

tratados como não rotulados. Nestas tarefas, são utilizados modelos de aprendizado de máquina não-supervisionado, e as principais tarefas são regras de associação, agrupamento (*clustering*), sumarização e visualização.

Nesta etapa de Extração de Padrões, os resultados genéricos podem ser validados por métricas objetivas ou, ainda, por julgamento subjetivo de especialistas do domínio em questão, o que pode auxiliar na condução da análise desses resultados. As validações sob esses resultados genéricos podem indicar a necessidade de repetir passos do Pré-Processamento ou mesmo refazê-lo.

2.2.4 Pós-Processamento e Utilização do Conhecimento

O Pós-Processamento é a etapa de validação das descobertas efetuadas. É a etapa de avaliação do conhecimento obtido e a apresentação do mesmo, seja por ferramentas de visualização ou simplesmente por tabelas de resultados. A análise minuciosa dos resultados obtidos permite que se valide a sua utilidade e até mesmo o próprio processo, determinando a necessidade de retomar passos anteriores e reestruturando-os se necessário. Nesta etapa o especialista do domínio e o de Mineração de Textos devem trabalhar juntos, procurando responder a questões como: representatividade do conhecimento obtido; o que há de novo nos resultados encontrados; de que maneira o conhecimento do especialista difere do obtido; validação dos resultados obtidos; identificação da adequação de procedimentos nas etapas anteriores para tentar melhorar os resultados; e de que maneira os resultados obtidos devem ser utilizados.

Na etapa de Utilização do Conhecimento os resultados estão validados e aptos a serem utilizados. Dessa forma, o conhecimento extraído pode ser aplicado para apoiar algum processo de tomada de decisão, conforme objetivo pré-estabelecido na etapa de Identificação do Problema.

2.3 Algumas aplicações da Mineração de Textos

Mineração de Textos tem como principal função identificar, nos mesmos, informações singulares, e não necessariamente informações principais, como é o caso da recuperação e preservação da idéia central. Todavia, ela pode ser aplicada para diversos fins, como em algumas das aplicações listadas a seguir.

- **Sumarização para documentos não estruturados** corresponde a geração de sumário, um resumo, com o objetivo de transmitir ou comunicar somente o que é importante de uma fonte textual de informação (**?**);

- **Classificação de textos** consiste em classificar novos textos (documentos) pertencentes a um conjunto de documentos pré-definido em uma ou mais categorias ou classes (**?**);

- **Recuperação de informação** tem como objetivo encontrar documentos que contenham uma determinada informação procurada a partir de uma coleção de textos (**?**);

- **Extração de informação** consiste em extrair a partir de documentos alguma informação relevante (**?**);

- **Indexação** visa identificar termos convenientes para a recuperação de informação. No trabalho de **?** foi feita uma comparação entre indexação manual e indexação utilizando uma ferramenta de Mineração de Textos, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação.

- **Agrupamento de documentos** almeja descobrir agrupamentos naturais e, então, apresentar as possíveis classes em uma coleção de textos (**?**).

2.4 Exemplo de instanciação da Mineração de Textos: Metodologia TopTax

Como exemplo de um processo instanciado de Mineração de Textos, pode-se citar a metodologia para extração de taxonomia de tópicos TopTax (*Topic Taxonomy Environment*) (**??**). Esta metodologia tem como objetivo auxiliar especialistas de um domínio específico a organizar e manter a informação do mesmo. Tal organização é possível devido à criação de uma taxonomia de tópicos. A taxonomia considerada é uma classificação hierárquica de tópicos extraídos de uma coleção de textos, na qual os tópicos superiores são *pais* dos inferiores, ou seja, os inferiores são especializações dos tópicos superiores. Além disso, a cada nível da taxonomia pode-se associar recursos da base textual referentes ao seu domínio, facilitando, assim, a organização da informação sob essa taxonomia.

Para isto, é necessário analisar o conhecimento do domínio e construir a hierarquia dos tópicos específicos do domínio, isto é, uma taxonomia de tópicos sobre o conhecimento do domínio representado pela coleção de textos, ou os metadados sobre outros tipos de recursos, bem como fornecer critérios que possibilitam aos especialistas decidirem sobre o crescimento da coleção de textos (o limite do aumento da quantidade de textos na taxonomia existente).

Pode-se haver algumas confusões em relação a definição de taxonomia e ontologia. Para a comunidade de Inteligência Artificial, ontologia é "uma especificação de uma conceituação" (**?**), isto é, é uma especificação explícita, com um vocabulário formal e regida por axiomas, dos conceitos em um domínio e as relações entre eles. Considera-se como axiomas as regras pertinentes ao domínio em questão. Neste sentido, uma taxonomia pode se tornar uma ontologia com a inclusão de axiomas.

A TopTax, para atingir seus objetivos, conforme mostrado na Figura 2.3, segue as etapas do processo de Mineração de Textos.

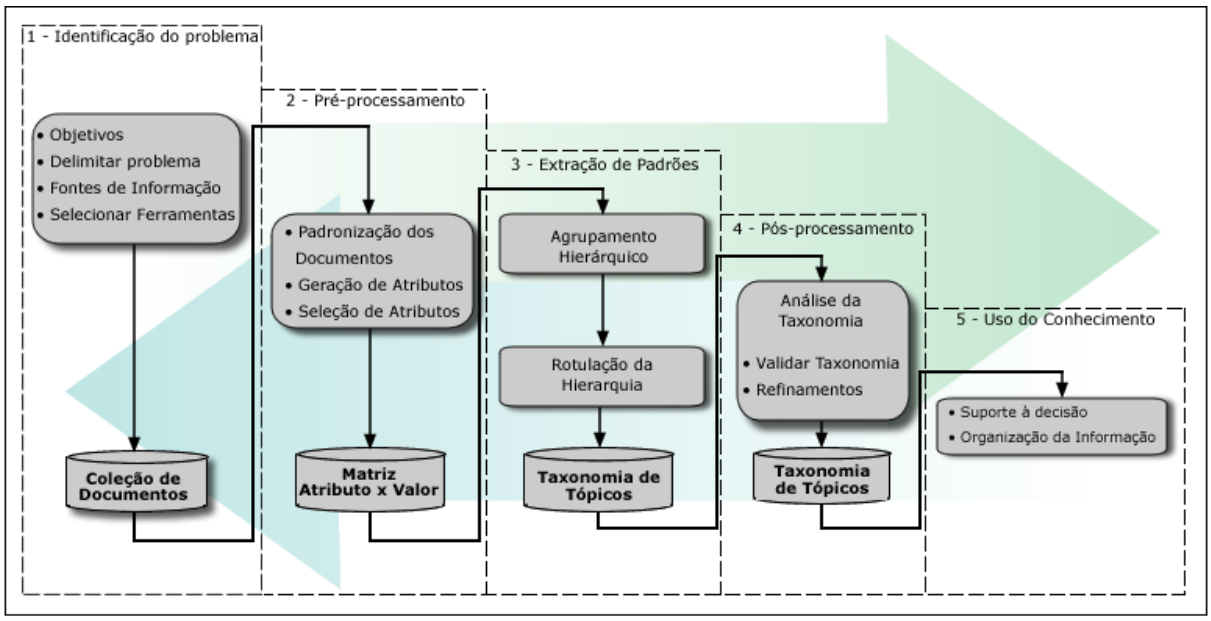


Figura 2.3: TopTax - Metodologia de extração de taxonomias de tópicos (**?**)

A etapa de Identificação do Problema visa delimitar o problema a ser abordado, que no caso é fornecer auxílio aos especialistas de um domínio específico a organizar e manter a informação do mesmo. A partir deste problema, pretende-se encontrar uma solução para o mesmo, traçando o objetivo a ser alcançado para que este problema possa ser resolvido, que no caso é a construção de uma taxonomia de tópicos do domínio em questão. Após a definição do problema e do objetivo, pode-se, então, com a ajuda de especialistas do domínio, recuperar os documentos do mesmo que formarão a base de textos a ser trabalhada, considerando que esses documentos podem vir de bases e formatos distintos. Além disso, nesta etapa são selecionadas as ferramentas que auxiliarão no processo de Mineração de Textos.

Já na etapa do Pré-Processamento, os documentos da base de textos obtida são preparados para servirem de entrada para as ferramentas que serão utilizadas. Esta preparação abrange as atividade de limpeza, padronização, extração de termos e cálculo do valor do atributo que será utilizado na matriz atributo-valor. Durante a padronização, os documentos que não estiverem em formato textual serão descartados, mas os que estiverem em fontes de informação em formato de figuras e filmes, serão considerados seus metadados, caso existam. Na atividade de extração de termos que descrevem a base de textos, os termos são representados em uma matriz atributo-valor. Mas como a quantidade de termos é muito elevada, utilizam-se técnicas que simplifiquem tais termos, o que diminuirá a dimensionalidade dessa matriz, facilitando o seu processamento e, como conseqüência, na compreensão dos termos. Mesmo assim, a dimensionalidade da matriz que representa tais termos pode continuar elevada. Neste caso, a TopTax sugere utilizar métodos de seleção de termos, com o objetivo de diminuir a quantidade de termos a ser trabalhada. Como exemplos desses métodos, pode-se citar os métodos de Luhn, Salton, e *Term Variance*,

que são detalhados no trabalho de **?**.

Em seguida, na etapa de Extração de Padrões, visando a construção de uma taxonomia de tópicos, efetua-se o processo de agrupamento hierárquico de documentos. Para isso, aplicam-se métodos de agrupamento hierárquico, como o *complete linkage*, *average likage* ou *single linkage* (**?**).

Com a hierarquia, os agrupamentos obtidos retêm tópicos ou sub tópicos aos quais os documentos se referem. Em seguida, conforme descrito em **?**, gera-se os rótulos para cada grupo encontrado pela obtenção dos termos mais significativos, sendo possível adicionar a cada nó recursos de informação de tópicos, como documentos, vídeos e imagens associadas. Por fim, na etapa do Pós-Processamento, esta hierarquia obtida é visualizada e validada para que a mesma possa ser utilizada para dar suporte às decisões e organização da informação ali contida. Tal visualização e validação é apoiada pela ferramenta TaxTools (**?**).

2.5 Considerações Finais

Neste capítulo foram abordados alguns conceitos das etapas do processo de Mineração de Textos, com ênfase na etapa de Pré-Processamento, que é o foco deste trabalho. Além disso, destacou-se a elevada importância da representação dos documentos neste processo.

Foram apresentadas algumas das aplicações da Mineração de Textos, bem como um resumo da metodologia TopTax, que é um exemplo de processo instanciado de Mineração de Textos. Deve-se ressaltar que os passos da metodologia de extração de termos apresentada neste trabalho (detalhada no Capítulo 4), além de outros fins, são utilizados na etapa de Pré-Processamento da TopTax. A metodologia proposta neste trabalho é semi-automatizada já que em determinados momentos ocorre interação com especialistas.

No decorrer do desenvolvimento deste trabalho, poder-se-á notar que a escolha de alguns procedimentos serve, também, para amenizar o problema do espaço de dimensão muito alta no processo de Mineração de Textos. Como exemplo desses procedimentos, pode-se citar a limpeza da base de textos, que acarreta na diminuição do espaço de armazenamento dos dados trabalhados mantendo a qualidade, visando melhorar o tempo de processamento e o espaço de armazenamento necessário para os dados.

No capítulo a seguir é detalhada uma das tarefas do Pré-Processamento da Mineração de Textos, que é a Extração de Termos.

Capítulo

3

Extração de Termos para a Mineração de Textos

3.1 Considerações Iniciais

O resultado da etapa do **Pré-Processamento**, no processo de Mineração de Textos, é uma representação da coleção de textos a ser analisada em um formato adequado à etapa de Extração de Padrões. A qualidade dos resultados depende principalmente do trabalho realizado nesta etapa, mais especificamente nas tarefas de seleção e extração dos termos.

O **termo**, também chamado de **característica** ou **atributo**, pode ser uma palavra simples ou composta. Quando o termo é composto por apenas uma palavra, este é denominado de *unigrama* ou termo simples, e quando é composto por mais de uma palavra, é chamado de *n-grama* (termo composto ou combinação). Como exemplos de termos simples, pode-se citar: *inteligencia*, *artificial*, *processo*; já como exemplos de termos compostos, têm-se: *inteligencia_artificial* e *processo_mineracao_textos*. Deve-se ressaltar que o significado semântico desses termos pode diferenciar quando são compostos de uma única palavra, como *inteligencia* e quando são compostos de mais de uma palavra, como *inteligencia artificial*.

Na etapa de Pré-Processamento, deve-se extrair termos (simples e/ou compostos) que conceitualmente melhor representam tais coleções. Isso ajudará a reduzir o número de termos utilizados, restringindo-os a um conjunto mais representativo da coleção, a fim de que o processamento desses dados seja uma tarefa computacionalmente mais simples e semanticamente adequada ao domínio de conhecimento.

Ressalta-se que alguns pesquisadores da área consideram a extração de termos como sinônimo para geração de termos, porém, neste trabalho, estas palavras são consideradas como distintas. A geração de termos é o processo de obter um conjunto de palavras com

significado importante para a coleção de textos de um determinado domínio, sendo que estas palavras não são modificadas e, sim somente obtidas da coleção. Já a extração de termos, que é o caso deste trabalho, está relacionado à criação de um novo conjunto de termos que possua significado importante para a coleção de textos de um determinado domínio, como termos radicalizados, lematizados ou substantivados.

A etapa de Pré-Processamento, na qual a atividade de extração de termos ocorre, exige um cuidadoso planejamento e acompanhamento. Esse processo é interativo e bastante trabalhoso, e pode ter alto custo computacional.

3.2 Técnicas de Simplificação de Termos

A extração de termos, que visa reconhecer os candidatos a termos em uma coleção de textos, pode ser auxiliada por meio de buscas por padrões que simplifiquem as diversas formas de apresentação de termos com o mesmo significado essencial ou termos que utilizados em conjunto modifiquem o significado dos mesmos isoladamente. Entre os padrões ou simplificações mais utilizados encontram-se as técnicas de radicalização, lematização e substantivação, além da possibilidade de utilizar termos compostos e vocabulário controlado, como explicado a seguir.

3.2.1 Radicalização

A radicalização, também conhecida como "Stemmização" ou *Stemming*, é uma técnica antiga muito utilizada. O primeiro trabalho encontrado na literatura sobre esta técnica é o de **?**. A radicalização tem como objetivo reduzir as palavras às suas formas inflexionáveis e às vezes reduzir às suas derivações (**?**). Para isto, a radicalização reduz cada palavra do texto ao seu provável radical, ou seja, palavra raiz (*stem*), em que cada palavra é analisada isoladamente. Segundo **?**, a radicalização pode ser vista como *radicalização inflexional*, em que se considera apenas as remoções de flexões verbais, ou *radicalização para a raiz*, na qual se realiza a remoção de todas as formas de prefixos e sufixos dos termos, sendo esta última a forma mais agressiva de radicalização. A seguir é mostrado um exemplo de radicalização para a raiz.

Frase exemplo: Brasileiros pesquisam perfil do estudante.

Considerando a remoção de *stopwords*, como resultado da radicalização para este exemplo tem-se:

brasil pesquis perfil estudant

O processo de radicalização pode depender da linguagem, por normalmente necessitar de conhecimento lingüístico (**?**). 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 resultando possivelmente em alguns erros.

Os algoritmos de radicalização realizam a eliminação de prefixos e sufixos das palavras ou a transformação de um verbo para sua forma infinitiva. Porém, durante este processo, podem ocorrer dois tipos de erros: *overstemming* e *understemming*. O ***overstemming*** acontece quando a parte removida da palavra não é um sufixo, e sim parte do seu radical. Este erro pode acarretar na possibilidade da combinação de palavras não relacionadas. Já o erro de ***understemming*** acontece quando não se remove completamente um sufixo da palavra. Ao contrário do *overstemming*, quando ocorre *understemming* pode-se fazer com que não haja a combinação de palavras relacionadas. Por exemplo, o *stem* correto da palavra *inteligencia* é *intelig*, mas quando ocorre o erro de *overstemming*, o resultado da aplicação da técnica de radicalização pode ser *intel*; e quando ocorre o erro de *understemming* o resultado pode ser *inteligenc*.

Como mostrado na Tabela 3.1, existem vários algoritmos de radicalização destinados a diferentes línguas. Dentre os mais conhecidos na literatura, podem-se citar o Método de Lovins (**?**), o Método de Porter (*Porter Stemming Algorithm*) (**?**) e o Método *Stemmer* S (**?**). Sendo estes métodos desenvolvidos para a Língua Inglesa.

O **método de Lovins** é executado em um único passo, removendo no máximo um sufixo por palavra (o sufixo mais longo). Este método é considerado mais agressivo do que os métodos de Porter e *Stemmer* S.

O **Método de Porter** foi originalmente proposto para a formação de radicais para a Língua Inglesa, isto é, geração dos radicais a partir da remoção dos sufixos das palavras. É considerado um algoritmo simples e muito eficiente para a radicalização de termos. Enquanto o Método de Lovins é executado em um único passo, este método é executado em cinco passos, sendo que cada passo realiza uma transformação sobre o termo alvo. Cada passo é formado por um conjunto de regras do tipo: *se um termo* ***t*** *possui mais do que* ***s*** *sílabas e termina com o sufixo* ***SUFIX****, o sufixo* ***SUFIX*** *é substituído por* ***SUF***. Ao final dessas substituições, espera-se obter o radical do termo.

***Stemmer* S** é considerado um método simples, conservador e raramente surpreende o usuário, pois somente remove alguns finais de palavras, como *ies*, *es* e *s*.

Já para a Língua Portuguesa, pode-se citar os algoritmos: Porter - Português, *PortugueseStemmer*, *Pegastemming* e STEMBR.

Porter - Português foi desenvolvido na linguagem de programação *Snowball*[1] em 2005, pelo mesmo autor do algoritmo de Porter para a Língua Inglesa, sendo baseado em regras de remoção de sufixos.

PortugueseStemmer, desenvolvido por Viviane Orengo e Christian Huyck (**?**), mesmo não sendo baseado no algoritmo de Porter, utiliza regras para a remoção de sufixos. Além disso, o *PortugueseStemmer* trata palavras exceções por meio do uso de um dicionário de 32 (trinta e dois) mil termos.

O ***Pegastemming***[2], desenvolvido por Gonzalez, realiza a remoção simples de sufixos comuns, sem se preocupar com artigos, preposições e conjunções.

[1] *Snowball* - http://snowball.tartarus.org/index.php
[2] *Pegastemming* - http://www.inf.pucrs.br/~gonzalez/ri/pesqdiss/analise.htm

Língua	***Algoritmo***	***Autoria***
Inglês	Dawson Stemmer S Lovins KStem Paice/Husk Porter Porter 2	Dawson Harman Lovins Krovetz Paice e Husk Porter Porter
Português	STEMBR Pegastemming PortugueseStemmer Porter - Português	Alvares Gonzalez Orengo Porter
Alemão	Porter - Alemão Porter - Alemão - Variação	Porter Porter
Amárico (etíope)	Alemayehu-Willett	Alemayehu e Willett
Búlgaro	BulStem	Nakov
Dinamarquês	Porter - Dinamarquês	Porter
Esloveno	Popovic-Willett	Popovic e Willett
Espanhol	Honrado et al. Porter - Espanhol	Honrado et al. Porter
Finlandês	Porter - Finlandês	Porter
Francês	Porter - Francês	Porter
Galego	Galician stemmer	Brisaboa
Holandês	Kraaij-Pohlmann Porter - Holandês	Kraaij e Pohlmann Porter
Italiano	Porter - Italiano	Porter
Latim	Schinke et al.	Schinke et al.
Norueguês	Carlberger et al. Porter - Norueguês	Carlberger et al. Porter
Russo	Porter - Russo	Porter
Sueco	Porter - Sueco	Porter
Turco	Ekmekçioglu et al.	Ekmekçioglu et al.

Tabela 3.1: Algoritmos para radicalização - Adaptada de **?**

O **STEMBR** (**?**), mesmo não sendo baseado no método de Porter, também trabalha com conjunto de regras para a extração do *stem*. O STEMBR remove os prefixos e sufixos das palavras por meio do tratamento baseado em estudo estatístico das freqüências das palavras contidas em páginas Web até o ano de 2005.

Como exemplos de aplicações dos algoritmos descritos, pode-se citar o Stemmer (**?**), PreTexT (**?**) e Lucene (**?**).

A ferramenta **Stemmer** (**?**) foi desenvolvida no LABIC[3] (Laboratório de Inteligência Computacional do ICMC/USP) baseada no algoritmo de Porter e extrai *stems* de palavras do português do Brasil, para isso a ferramenta remove os sufixos e terminações destas palavras.

[3]LABIC - http://labic.icmc.usp.br/

A ferramenta **PreTexT**, desenvolvida no LABIC inicialmente por **?** e posteriormente atualizada por **?** (PreTexT II), tem como objetivo auxiliar na etapa de Pré-Processamento de uma coleção de documentos, apresentando facilidades para reduzir a dimensionalidade do conjunto de termos. Para isso, possui uma implementação do algoritmo do Porter utilizando o paradigma de orientação a objetos em Perl. Tal implementação possibilita extrair *stems* de palavras nas Línguas Portuguesa, Espanhola e Inglesa. O algoritmo da PreTexT verifica se os sufixos da palavra possuem comprimento mínimo estabelecido, considerando algumas regras pré-estabelecidas. Caso possuem, estes sufixos são eliminados da palavra. Porém, devido às línguas provenientes do latim terem formas verbais conjugadas em sete tempos, cada uma com seis terminações diferentes, foi necessário um tratamento para estas terminações. Então, para as Línguas Portuguesa e Espanhola, caso não seja possível eliminar, de acordo com essas regras, nenhum desses sufixos, as terminações verbais da palavra são analisadas. A ferramenta disponibiliza também uma lista de *stopwords* que pode ser incrementada manualmente pelo usuário. Quanto ao uso de termos, a PreTexT possibilita gerar os termos simples (*unigrama*) ou compostos (mais de *unigrama*) e, tem como saída vários arquivos com informações úteis para o usuário, como freqüência dos *stems*, o quanto cada documento é esparso, freqüência das palavras que originam os *stems* e outros. Além disso, permite, também, o uso de métodos de seleção de termos, como os cortes de Luhn (**?**). Para aplicá-los, a PreTexT oferece uma opção de utilizar somente os *stems* que estão em um determinado intervalo de freqüência ou usar os pontos de corte superior e inferior que são encontrados empiricamente pelo usuário (**?**).

O **Lucene** (**?**) é uma API que contém classes desenvolvidas utilizando a linguagem de programação Java que executam atividades de Mineração de Textos. Dentre estas classes há duas específicas para realizar a radicalização em textos na Língua Portuguesa, a *BrazilianStemFilter* e a *BrazilianStemmer*, que são baseadas no algoritmo de Porter.

3.2.2 Lematização

A técnica de lematização, também conhecida como Redução à Forma Canônica, tem como objetivo agrupar as variantes de um termo em um único lema, ou seja, transformar verbos para sua forma no infinitivo, e substantivos e adjetivos para o masculino singular. Pode-se observar um exemplo da redução das palavras ao seu lema na Tabela 3.2, no qual são mostrados os lemas e exemplos de flexões das mesmas.

Lema	*SingularFem.*	*PluralFem.*	*PluralMasc.*
brasileiro	brasileira	brasileiras	brasileiros
pesquisa	pesquisa	pesquisas	pesquisas
perfil	perfil	perfis	perfis
estudante	estudante	estudantes	estudantes

Tabela 3.2: Exemplos de lematização

Para a Língua Portuguesa, foram encontrados alguns etiquetadores morfossintáticos que podem auxiliar no processo de lematização. No processo de etiquetagem cada termo

de um texto é associado à uma etiqueta (*tag*), que corresponde a sua classe gramatical, como verbo, substantivo e adjetivo. Segundo **?**, o processo de etiquetagem, normalmente, tem custo de tempo alto e está sujeito à erros. Os etiquetadores encontrados são: o etiquetador de BRILL (**?**) e o MXPOST (**?**).

O **etiquetador de BRILL** é um marcador morfossintático de palavras de um texto baseado em aprendizado computacional, ou seja, o aprendizado de uma série de regras contextuais que são utilizadas na etiquetagem.

O **MXPOST** (*Maximum entropy pos tagger*) (**?**) é um etiquetador morfossintático disponível na Web para uso não comercial e foi implementado, usando a linguagem de programação Java (JDK 1.1), por um grupo de pesquisadores da Universidade da Pensilvânia. Seu objetivo é fazer uma análise sintática, colocando em arquivos textos as marcações *tag* que identificam a classificação gramatical da palavra dentro da frase.

Após a identificação das classes gramaticais dos termos a partir do processo de etiquetagem, é possível, então, reduzir tais palavras ao seu lema. Existem ferramentas de lematização encontradas na literatura, que são descritas a seguir, como a TreeTagger (**?**), o Lematizador de Nunes (**?**), o FLANOM (**?**), a FORMA (**?**) e o **Sphinx**[4].

O **TreeTagger**[5] (**?**) foi desenvolvido por Helmut Schmid em 1994 para o Projeto TC do Instituto para Computação Lingüística da Universidade de Stuttgart. É uma ferramenta para etiquetagem morfossintática e um lematizador, podendo ser utilizado para as Línguas Alemã, Búlgara, Chinesa, Espanhola, Francesa, Grega, Holandesa, Inglesa, Italiana, Portuguesa e Russa.

O **Lematizador de Nunes** é uma ferramenta disponível gratuitamente desenvolvida por Nunes e seus colaboradores (**?**) direcionado à Língua Portuguesa.

O **FLANOM** (*Flexionador y lematizador automático de formas nominales*) é um lematizador de palavras na Língua Espanhola desenvolvido por **?**.

A ferramenta **FORMA**[6], desenvolvida por **?**, também é direcionada à Língua Portuguesa. Essa ferramenta primeiramente *toqueniza* as palavras do texto, em seguida, as etiqueta morfologicamente para, então, lematizá-las.

O software proprietário **Sphinx**, versão 4, possibilita a aplicação da técnica de lematização nos textos das Línguas Francesa e Inglesa.

3.2.3 Substantivação

É um processo, também conhecido por "Nominalização", na qual as palavras passam a exibir um comportamento sintático/semântico semelhante àquele próprio de um nome[7]. Deve-se ressaltar que a maioria das palavras do português podem ser nominalizadas com o uso de artigos. A seguir, é mostrado um exemplo de substantivação.

[4]Sphinx - http://www.sphinxbrasil.com.br/

[5]TreeTagger - http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6]FORMA - http://www.inf.pucrs.br/ gonzalez/tr+/forma/

[7]Texto sobre Gramática Tradicional e Categorização Lexical - http://www.dacex.ct.utfpr.edu.br/paulo3.htm

Frase exemplo: Técnicas relacionadas à Inteligência Artificial.

Considerando a remoção de *stopwords* e limpeza do texto, tem-se como resultado da substantivação para este exemplo:

tecnica relacionar inteligencia artificial

Para a Língua Portuguesa, pode-se citar a combinação das ferramentas CHAMA e FORMA, desenvolvidas por **?**. A ferramenta FORMA tem como objetivo *toquenizar* e etiquetar morfologicamente as palavras dos textos, resultando em palavras lematizadas. Este resultado serve como entrada para a ferramenta CHAMA que é responsável pela nominalização de adjetivos, advérbios e verbos nos textos, ou seja, a transformação destas palavras em substantivos.

3.3 Extração de Termos Simples e Compostos

A extração de termos consiste em, a partir da extração de palavras de documentos de um domínio, obter um novo conjunto de termos que representam tal domínio a ser trabalhado. A extração de termos, de acordo com **?**, possui três abordagens principais, que são: estatística, lingüística e híbrida.

A abordagem **Estatística** utiliza somente métodos baseados em conhecimento estatístico e é utilizada sobre a forma de representação de termos *bag of words*, nos quais os termos são tratados como um conjunto desordenado de palavras independentes. Assim, termos são considerados independentes entre si e todas as inferências são realizadas sobre algum valor dado a esses termos, como por exemplo, suas respectivas freqüências na coleção de textos.

A abordagem **Lingüística** utiliza métodos baseados em conhecimento lingüístico. Estes métodos podem fazer uso de recursos que contêm diferentes informações lingüísticas para a extração dos termos, como informações lexicográficas (dicionários de termos e *stoplist*), informações morfológicas (padrões de estrutura interna das palavras), informações morfossintáticas (categorias morfossintáticas e funções sintáticas), informações semânticas (classificações semânticas) e informações pragmáticas (representações tipográficas e informações de disposição do termo no texto). E, por fim, a abordagem **Híbrida** faz uso das duas abordagens (estatística e lingüística).

Independente da abordagem escolhida para ser utilizada, a atividade de extração de termos é completada quando se tem somente os termos que representam a coleção de textos ou a maioria destes. Estes termos podem ser termos simples ou combinações de termos, ou seja, seqüências de duas ou mais palavras que possuem características sintáticas e semânticas de uma unidade.

O significado exato e desambíguo ou conotação destas palavras não pode ser diretamente derivado dos significados ou conotações de seus componentes. Deve-se, portanto,

considerar o comportamento e sentido especial destas palavras consecutivas (**?**). Para melhor entendimento, pode-se observar os diferentes significados das combinações a seguir: *inteligencia*, *inteligencia emocional*, *inteligencia artificial*, *inteligencia policial*, *inteligencia musical* e *inteligencias multiplas*. Um método simples para encontrar essas combinações em um texto é a contagem de ocorrência das mesmas. Este método considera que se duas palavras ocorrem juntas diversas vezes, então é evidente que elas tenham um significado especial juntas, que não é o mesmo quando separadas. Deve-se notar que esse processo é puramente estatístico, pois as combinações das palavras são encontradas em um processo estocástico de co-ocorrência.

Entretanto, deve-se assumir que somente selecionando os *n-gramas* (seqüência de 'n' *tokens*) mais freqüentes, não levará a um resultado completamente satisfatório, pois pode haver, por exemplo, uma alta freqüência das palavras "*e o*", que dependendo do objetivo, não significará nada. Por isso, no processo de limpeza dos textos, anteriormente citado, é interessante elaborar uma lista de *stopwords*, para eliminar palavras com menos significado. Outra solução interessante é combinar um pouco de conhecimento lingüístico, que pode auxiliar a correta identificação das funções sintáticas de cada termo (**?**).

Ressalta-se também que, ao escolher a quantidade máxima de *gramas* que comporá o termo, deve-se levar em consideração que tal quantidade é proporcional ao número de possibilidades de combinações entre termos simples (*unigramas*) e, conseqüentemente, proporcional ao número de termos extraídos.

Para a obtenção de *n-gramas*, pode-se utilizar a ferramenta PreTexT, anteriormente citada, ou o pacote NSP. O pacote NSP (*Ngram Statistics Package*) (**?**) foi implementado na linguagem de programação Perl e tem sido apoiado pela *National Science Foundation Faculty Early Career Development Program* (CAREER). Em versões anteriores (v0.1, v0.3, v0.4), este pacote era conhecido como ***Bigram*** *Statistics Package* (BSP), mas com o aumento da capacidade de trabalhar com *n-gramas* e não mais somente *Bigramas*, seu nome foi alterado para *N-gramas*. Este pacote é composto por um conjunto de programas que auxilia na análise de *n-gramas* em arquivos textos, ou seja, permite a identificação de *n-gramas* nesses arquivos.

Esses *n-gramas* correspondem aos candidatos a termos da coleção de textos. Como a quantidade de candidatos a termos é muito elevada, faz-se necessário utilizar algum método para escolher os que devem ser removidos. Há alguns testes estatísticos que podem ser utilizados para este fim. A idéia fundamental é testar a dependência entre as palavras consideradas como partes do *n-grama*. Para isso considera-se a freqüência de ocorrência de cada grama e de todas as suas combinações. Para testar a hipótese de dependência o estimador mais utilizado é o qui-quadrado, no entanto, para dados bastante esparsos, o estimador mais adequado é o logaritmo da razão de verossimilhança (**?**).

3.4 Vocabulário Controlado

Uma possível técnica de redução do número de termos em uma coleção de textos é o reconhecimento de sinônimos ou termos hierarquicamente superiores ou inferiores aos termos analisados, bem como a substituição dos termos analisados por esses, ou melhor, o uso de taxonomias ou *thesaurus*. Considera-se que dois termos são sinônimos se existir algum contexto em que ambos puderem ser substituíveis sem provocar alteração substancial do significado (**?**).

Uma taxonomia é uma coleção de vocabulários controlados e organizados hierarquicamente (**?** apud **?**), enquanto um *thesaurus* pode ser definido como um vocabulário controlado que representa sinônimos, hierarquias e relacionamentos associativos entre termos.

O *thesaurus* utiliza listas pré-compiladas de termos importantes para um determinado contexto, em que cada termo da lista é representado pela ligação (seja por relações de equivalência, hierarquia e/ou associação) com vários outros. Dessa forma, o seu uso facilita aos usuários encontrar as informações que necessitam, mesmo que façam várias consultas com termos distintos (que estejam ligados), obtendo os mesmos resultados devido às relações dos termos. Além disso, o *thesaurus* diminui a quantidade de termos-índices quando utilizado no processo de normalização (como ocorre na técnica de radicalização).

Pode-se fazer uma comparação do *thesaurus* com a técnica de simplificação de termos observando o resultado final de ambos. O *thesaurus* resume várias palavras em apenas um termo, substituindo essas palavras por termos ligados entre si (como sinônimos). E técnicas de simplificação de termos, como por exemplo a radicalização, também resumem várias palavras em apenas um termo no momento em que substitui estas palavras por seus radicais.

Como exemplos de *thesaurus*, pode-se citar o Thesagro, o *Thesaurus* da Língua Portuguesa do Brasil e o Tep.

O **Thesagro** (*Thesaurus* Nacional Agrícola), publicado pela BINAGRI[8] (Biblioteca Nacional de Agricultura, órgão da Secretaria de Executiva do Ministério da Agricultura, Pecuária e Abastecimento), é o único *thesaurus* brasileiro especializado em literatura agrícola, além disso, contém informações sobre as relações hierárquicas dos seus termos, que são explicadas a seguir. Para exemplificar essas relações, são utilizados alguns termos extraídos do Thesagro[9]. A **relação de associação** (*related term* (*RT*) - termo relacionado) é empregado para estabelecer associação entre um termo cujo significado se relaciona semanticamente com outro termo, mas sem nenhuma ligação hierárquica entre si. Como exemplo, pode-se citar o termo *enologia* cujo termo relacionado é *vinho*. A **relação de equivalência** (*USE*) é utilizada para indicar o termo correto a ser utilizado. Por exemplo: para o termo *abacaxizeiro* deve-se usar (USE) o termo *abacaxi*. A **relação hierárquica**

[8]BINAGRI - http://www.agricultura.gov.br/

[9]Thesagro - http://www.agricultura.gov.br/portal/page?_pageid=33,959135&_dad=portal&_schema=PORTAL

genérica (*broader term* (*BT*) - termo genérico) é empregado para indicar um termo mais amplo, mais abrangente. Para o termo *inseticida*, por exemplo, o termo genérico correspondente é *defensivo*. A **relação hierárquica específica** (*narrower term* (*NT*) - termo específico) é utilizado para indicar termos mais definidos. Para o termo *intoxicação*, por exemplo, tem-se como termos específicos a *intoxicação animal* e a *intoxicação vegetal*.

O ***Thesaurus* da Língua Portuguesa do Brasil**[10] foi construído manualmente no ano de 2000 por Valdir Jorge e disponibilizado livremente na Web.

O **TeP** (*Thesaurus* Eletrônico para o Português do Brasil) foi desenvolvido por **?** e posteriormente detalhado no trabalho de **?**. É considerado um *thesaurus* eletrônico para o português do Brasil, sendo considerado um dicionário eletrônico de sinônimos e antônimos, composto por substantivos, adjetivos, verbos e advérbios. A base de dados lexicais do TeP (**?**) foi desenvolvida para servir como ponto de partida da rede WordNet, sendo, portanto, feita segundo o modelo da rede WordNet para o português do Brasil (WordNet.Br).

Na obtenção de sinônimos e antônimos, quando um determinado verbete é buscado na base de dados lexicais do TeP, caso este esteja presente, são retornados os conjuntos de sinônimos e antônimos do mesmo. Por exemplo: para a entrada (verbete) *recordar*, é retornado seu conjunto correspondente, que é {*lembrar, recordar*}, sendo *lembrar* considerado como seu sinônimo. Caso o verbete ainda não exista na base do TeP, o mesmo é inserido como entrada na base, possibilitando a geração automática de novos verbetes, incluindo os seus conjuntos de sinônimos e de antônimos, se houver.

A base de dados lexicais do TeP possui mais de 19 mil conjuntos, que indexam 44 mil entradas distribuídas em 17 mil substantivos, 15 mil adjetivos, 11 mil verbos e 1 mil advérbios. Essa base é utilizada na **WordNet.Br** (WordNet para o Português do Brasil) (**?**) o que possibilita substituir palavras sinônimas em diversas frases do texto. Por exemplo, quando deseja-se evocar em um texto o sentido de "*examinar cuidadosamente*", a WordNet.Br procura no próprio texto frases que contenham este sentido, como, por exemplo, os verbos: *fiscalizar*, *patrulhar*, *policiar* e r*ondar*.

3.5 Trabalhos Relacionados com a Extração de Termos

Devido a importância de se extrair termos nas mais diversas línguas, várias ferramentas e algoritmos de extração de termos têm sido desenvolvidos. Como exemplo, pode-se citar o trabalho de **?** no qual foi desenvolvida uma ferramenta lexográfica (Xtract) para extrair colocações da Língua Inglesa e, posteriormente, estendida com o nome CXtract, para a Língua Chinesa no trabalho de **?**. Já **?** desenvolveram uma ferramenta que tem como objetivo auxiliar os terminologistas na identificação e tradução de termos técnicos.

No trabalho de **?** foi desenvolvido um ambiente para extrair terminologia de forma híbrida a partir de laudos médicos, denominado *Term Pattern Discover* (TP-Discover). Tal ambiente, resumidamente, seleciona palavras e frases que aparecem com uma determinada

[10] *Thesaurus* da Língua Portuguesa do Brasil - http://alcor.concordia.ca/ vjorge/Thesaurus/

freqüência e, para isso, a técnica de lematização foi aplicada utilizando o lematizador TreeTagger (**?**). Então, selecionam-se os termos com uma determinada propriedade sintática.

Já como algoritmos de extração de termos que combinam a abordagem híbrida (técnicas estatísticas e conhecimento lingüístico), pode-se citar o algoritmo proposto por **?**. O trabalho de **?** também utilizou a abordagem híbrida, atribuindo peso aos candidatos a termo de acordo com sua classe gramatical; já no ano de 2000, para extração de termos, esses autores consideraram o termo candidado e o termo de contexto (termo que aparece dentro de uma janela de tamanho fixo), utilizando três tipos de informação de contexto: sintática (atribui pesos para as diferentes classes gramaticais a que o termo candidato pertence), terminológica (atribui um peso ao termo candidato baseando-se nos termos de contexto dele) e semântica (mede a similaridade entre o termo candidato e os termos de contexto) (**?**).

No trabalho de **?**, as diferenças teóricas existentes entre as técnicas de lematização e radicalização são ressaltadas. Os autores afirmam que a lematização existe puramente no contexto lexicográfico, pois esta representa os adjetivos e substantivos por seu masculino singular e os verbos por seus infinitivos. Já a radicalização não existe puramente no contexto lexicográfico, pois esta remove os sufixos do radical, segundo o algoritmo de Porter. Mesmo assim, eventualmente, estas duas técnicas podem gerar resultados graficamente semelhantes.

? fez uma análise de precisão de dois algoritmos de radicalização de palavras pertencentes à Língua Portuguesa. Para tal análise, foram considerados os algoritmos *Pegastemming* e *PortugueseStemmer*, e 500 *stems* de palavras diversas foram obtidas manualmente para, em seguida, aplicar o processo automático de radicalização nestas mesmas palavras utilizando, separadamente, cada um desses algoritmos. Considerando que cada algoritmo foi desenvolvido para aplicações específicas conforme a necessidade de seu respectivo autor, foram apresentados diversos resultados positivos e negativos de cada algoritmo. Como por exemplo, o algoritmo *Pegastemming* apresentou uma melhor precisão quando processadas palavras da categoria substantivo, porém o *PortugueseStemmer* obteve a melhor precisão quando processados advérbios.

No trabalho de **?**, 5.000 artigos publicados em jornais na Língua Finlandesa foram agrupados por quatro métodos de agrupamento hierárquico aglomerativo e, durante este processo, as técnicas de radicalização e lematização foram aplicadas. Foram obtidos melhores resultados quando utilizada a técnica de lematização, por outro lado, com o uso da radicalização a similaridade entre os documentos aumentou devido a junção maior de diferentes palavras quando utilizada esta técnica, já que o número de palavras discriminantes aumentou.

? desenvolveu um lematizador somente para verbos a partir do Banco de Conjugações de Verbos da Língua Portuguesa em sua versão 1.1, que faz parte do software livre Conjugue[11]. Este lematizador foi aplicado aos verbos da base de textos a fim de unifor-

[11] Conjugue - http://www.ime.usp.br/~ueda/br.ispell/

mizar as regras aprendidas para as tarefas de classificação. Como resultado, a aplicação do lematizador nos verbos obteve uma redução do tempo de treinamento e aumentou a abrangência do conjunto de regras aprendidas, mas acarretou em uma contribuição pouco significativa em termos de eficácia no resultado da aplicação das regras aprendidas.

No trabalho de **?**, foram apresentados a técnica de substantivação e um novo lematizador, ambos voltados para a Língua Portuguesa e implementados pelos autores nas ferramentas FORMA e CHAMA. Estas técnicas foram comparadas à radicalização com foco na recuperação de informação, sendo que para a radicalização utilizou-se o algoritmo *PortugueseStemmer* (**?**). Nesta comparação, o uso da técnica de substantivação obteve diferença significativa positiva em relação às técnicas de radicalização e lematização.

Pode-se citar também a OntoLP (**?**) que é um *plug-in* desenvolvido para auxiliar de forma semi-automática os engenheiros de ontologias de Língua Portuguesa, mostrando sugestões de termos, conceitos e de organização de hierarquias da ontologia, com base no conhecimento contido em base textual ou corpus de um domínio específico. Este *plug-in* serve para ser utilizado no editor de ontologias Protégé (**?**), que oferece suporte à construção de ontologias, seguindo as tecnologias da Web Semântica, como a construção de ontologias OWL *Web Ontology Language*.

3.6 Considerações Finais

Durante a atividade de extração de termos a partir de coleções textuais, que foi abordada neste capítulo, pode-se utilizar combinações das técnicas de simplificação dos termos, a fim de melhorar a representação dos mesmos na coleção de textos, conseqüentemente melhorar o resultado final do objetivo proposto pelo usuário.

Deve-se ressaltar que a opção pelo uso de formas de simplificação depende das metas pré-estabelecidas e tem como benefício melhor representar a coleção em questão, bem como auxiliar na redução da dimensionalidade da forma de representação dos termos extraídos, que no caso é a matriz atributo-valor. Dessa forma, pode-se minimizar um dos maiores problemas que é o de trabalhar com uma enorme quantidade de termos. Além de melhorar relativamente a busca de uma palavra pelo usuário em uma coleção de textos, pois possibilita retornar como resultado, não mais somente uma forma desta palavra, devido suas variações (plurais, formas de gerúndio, sufixos), e sim uma combinação entre uma palavra da consulta e uma palavra do documento, aumentando, portanto, a gama de busca deste usuário.

Com os conceitos apresentados neste capítulo, nota-se a necessidade de se escolher adequadamente técnicas de simplificações dos termos para o domínio e/ou o uso de algum *thesaurus* para serem utilizadas na atividade de extração de termos.

Capítulo

4

Metodologia para a Utilização de Diferentes Formas de Extração de Termos a partir de Coleções Textuais

4.1 Considerações Iniciais

Neste capítulo é descrita a metodologia para a utilização de diferentes formas de extração de termos a partir de coleções textuais de domínio específico. Essas diferentes formas englobam a utilização de três técnicas que simplificam os termos extraídos: a radicalização, a substantivação e a lematização. Tal metodologia, além de poder ser utilizada para outros fins, visa apoiar a metodologia proposta por **?** e **?**, denominada TopTax, descrita na Seção 2.4 do Capítulo 2.

4.2 Descrição da Metodologia para a Utilização de Diferentes Formas de Extração de Termos

O impacto das técnicas de extração de termos é bem perceptível em tarefas de organização de informação em que a compreensibilidade, representatividade e o número de termos extraídos têm impacto direto na interpretabilidade dos modelos gerados. Assim, neste trabalho avalia-se a extração de termos, para que estes sejam utilizados, principalmente, no contexto de extração de taxonomias de tópicos, como no Projeto TopTax[1], o qual tem por objetivo auxiliar especialistas de um domínio específico a organizar e manter a informação do mesmo, por meio da criação e atualização de uma taxonomia de tópicos

[1]TopTax - http://labic.icmc.usp.br/projects/researchproject.2008-06-04.9415524093

para domínios específicos.

A metodologia para aplicação de diferentes formas de extração de termos divide-se em duas principais fases, a saber: preparação dos textos e extração de termos. Primeiramente, é necessário delimitar os textos nos quais irão trabalhar, sendo que estes podem prover de diferentes repositórios (bases de textos). Após a obtenção dos textos, deve-se condensá-los em uma base de textos e garantir a qualidade desta base por meio da **preparação dos textos** para posteriormente realizar a **extração de termos** importantes do domínio desta base textual.

4.2.1 *Fase 1: Preparação dos Textos*

A preparação dos textos para a extração de termos, mostrada na Figura 4.1, afeta o resultado final da metodologia de extração de termos caso suas atividades não sejam cuidadosamente executadas.

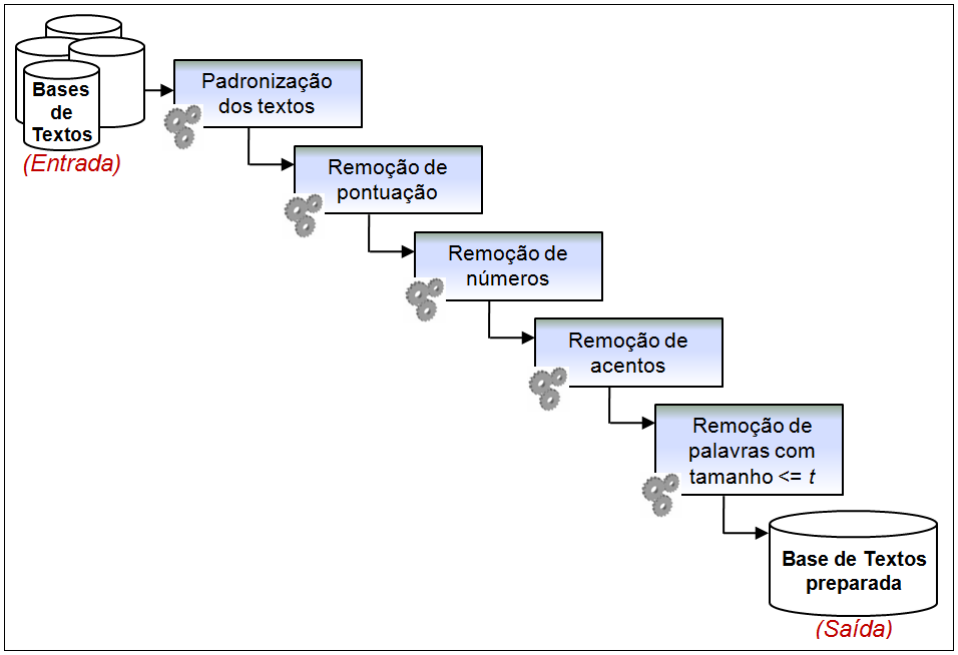


Figura 4.1: Preparação dos textos para a metodologia de extração de termos

Considerando que os documentos podem vir de bases distintas, pode haver formatos diferentes de arquivos, sendo necessário a **padronização dos textos**, colocando a base padronizada a ser trabalhada em um único repositório.

A padronização dos textos consiste em transformar todo o conteúdo destes documentos para sua forma minúscula e padronizar os documentos para o formato de texto plano. Alguns documentos podem não permitir tal padronização, por exemplo, o documento pode estar protegido contra cópias ou estar em formato de figuras e imagens. Sendo assim, deve-se analisar subjetivamente o número de documentos não danificados e, caso este número seja considerado insuficiente, deve-se buscar por mais documentos. Todo o processo pode ser repetido até se obter uma coleção textual satisfatória para atingir os objetivos pré-estabelecidos.

Em seguida, é realizada a **remoção de pontuação** dos textos, dado que a pontuação somente aumentaria a quantidade de termos extraídos, já que os algoritmos utilizados na extração de termos não diferenciam pontuação de palavras, extraindo, assim, termos compostos por pontuações.

Pode-se também realizar a **remoção de números** dependendo do objetivo, ou seja, quando não é necessário trabalhar com números nos textos. Já para a melhor compreensão e junção das palavras iguais por técnicas próprias (como a radicalização, a lematização e a substantivação), é feita a **remoção de acentos** das palavras dos textos.

Remoção de palavras com tamanho <= *t*, sendo t o número de caracteres de uma palavra. Tal remoção é útil quando se tem caracteres sem nenhum significado, como por exemplo os caracteres *z*, *u* e *rc*.

Na Tabela 4.1, para melhor exemplificar as atividades da preparação dos textos, são mostrados dois parágrafos de um documento original da base de textos utilizada, como exemplo, neste trabalho. Na Tabela 4.2 são mostrados esses parágrafos preparados para serem utilizados, ou seja, o resultado dos mesmos após a preparação dos textos seguindo os passos da fase 1 e após a remoção de *stopwords*. Mesmo para apenas dois parágrafos de um documento, pode-se notar a redução do tamanho do mesmo, o que afeta a dimensionalidade da matriz atributo-valor.

"Em virtude dos bons resultados com animais em crescimento, os fazendeiros passaram alimentar com mistura de cana e uréia as vacas em lactação durante o período seco do ano."
"Nestes sistemas de pastejo extensivo de produção de leite, em que as vacas são alimentadas com cana-de-açúcar e uréia, espera-se uma produção de leite elevada, não considerando o leite mamado pelo bezerro, além de ao final do período seco as vacas apresentarem boa condição corporal e fertilidade adequada."

Tabela 4.1: Parágrafos retirados de um documento original da base de textos

virtude bons resultados animais crescimento fazendeiros passaram alimentar mistura cana ureia vacas lactacao periodo seco ano
sistemas pastejo extensivo producao leite vacas alimentadas cana acucar ureia esperase producao leite elevada considerando leite mamado bezerro final periodo seco vacas apresentarem boa condicao corporal fertilidade adequada

Tabela 4.2: Parágrafos, referentes aos parágrafos mostrados na Tabela 4.1, preparados para serem utilizados na metodologia de extração de termos

Após este passo, a base de textos a ser trabalhada é considerada preparada, permitindo, então, efetuar a extração de termos de um domínio, que é explicada na Fase 2.

4.2.2 Fase 2: Extração dos Termos

Dado o objetivo deste trabalho, que é avaliar o efeito do uso de diferentes formas de extração de termos, o processo proposto aqui faz uso de três técnicas de simplificação de termos, descritas conceitualmente no Capítulo 3: a radicalização por meio da ferramenta PreTexT II; a lematização, usando a base de lemas do Lematizador de Nunes; e a substantivação, utilizando as ferramentas FORMA e CHAMA. Este processo pode ser estendido para outras diferentes técnicas.

O processo de extração de termos proposto, conforme ilustrado na Figura 4.2, inicia-se com a base de textos considerada preparada para ser utilizada.

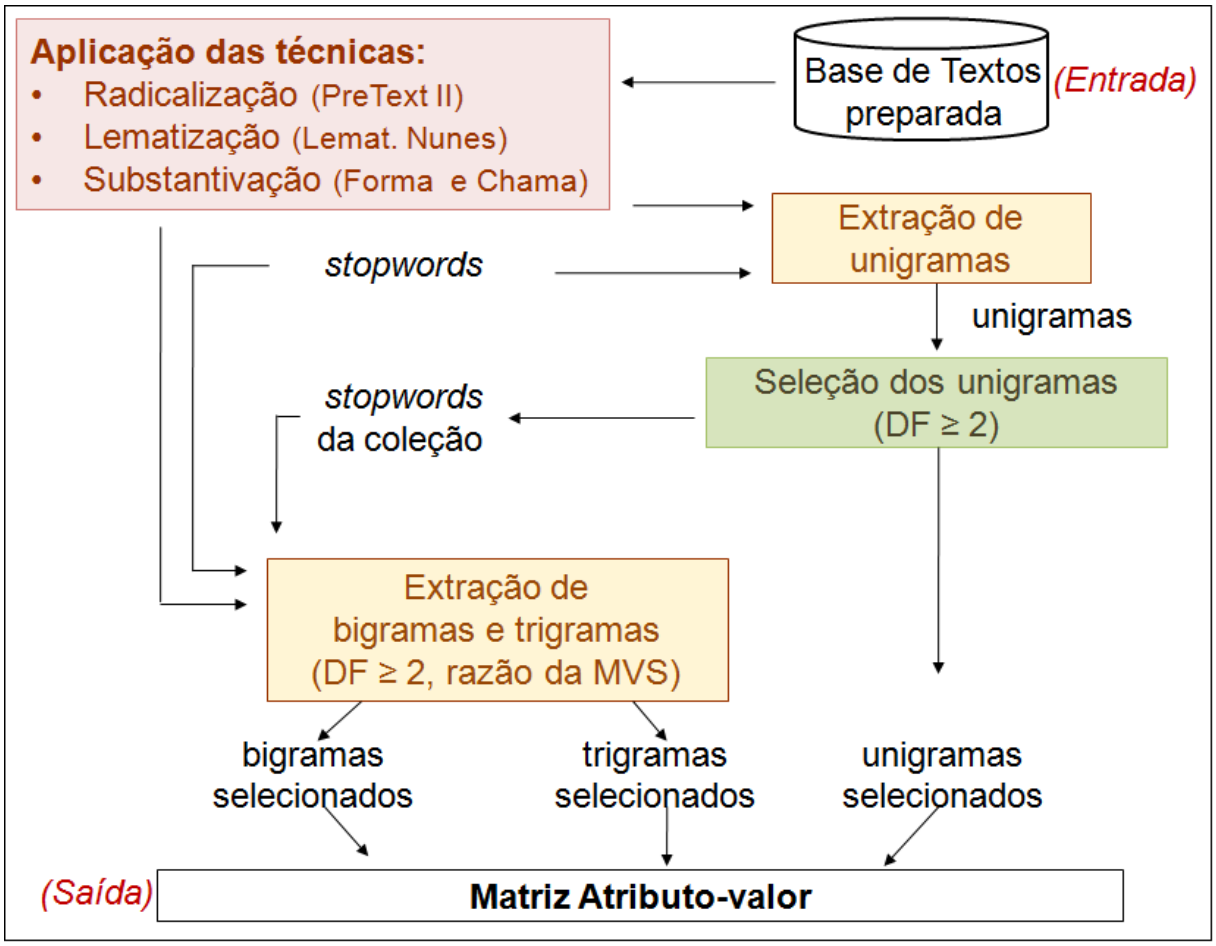


Figura 4.2: Extração de termos

O usuário pode escolher se deseja aplicar alguma técnica de simplificação de termos durante esta fase. Caso positivo, deve escolher qual técnica mais contribui para seu objetivo final. A seguir, são mostrados exemplos de palavras após a aplicação de cada uma destas técnicas, considerando como base os parágrafos de um documento mostrados na Tabela 4.2.

Quando a técnica de radicalização é aplicada, todas as palavras da coleção de textos são transformadas em unigramas radicalizados com o uso da ferramenta PreTexT II. Essa transformação pode ser vista nas palavras que se encontram radicalizadas, que é mostrada

na Tabela 4.3.

virtud bom result anim cresciment fazende pass *aliment mistur can ure vac lactaca period sec ano* *sistem pastej extens produca leit vac aliment can acuc* *ure esperas produca leit elev consider leit mam bezerr* *final period sec vac apresent boa condica corporal fertil adequ*

Tabela 4.3: Palavras radicalizadas

Para a aplicação da técnica de lematização, utiliza-se a base de lemas do Lematizador de Nunes, que é composta por palavras na Língua Portuguesa e seus respectivos lemas. Como resultado, obtém-se todas as palavras da coleção de textos transformadas em palavras lematizadas, conforme mostrado na Tabela 4.4.

virtude resultado animal crescimento fazendeiro passar *alimentar misturar cana ureia vaca lactacao periodo seco ano* *sistema pastejo extensivo producao leite vaca alimentar cana acucar* *ureia esperase producao leite elevar considerar leite mamar bezerro* *alar final periodo seco vaca apresentar condicao corporal fertilidade adequar*

Tabela 4.4: Palavras lematizadas

Já para a aplicação da técnica de substantivação são utilizadas as ferramentas FORMA e CHAMA. Como resultado, tem-se todas as palavras transformadas para seus substantivos correspondentes, conforme mostrado na Tabela 4.5.

virtude resultados animais crescimento fazendeiros passagem *alimentar mistura cana ureia vacas lactacao periodo secura ano* *sistemas pastejo extensividade producao leite vacas alimentacao cana acucar* *ureia esperase producao leite elevacao consideracao leite mamacao bezerro* *final periodo secura vacas apresentacao bondade condicao corporal fertilidade adequacao*

Tabela 4.5: Palavras substantivadas

Para a obtenção de um documento substantivado, conforme mostrado na Figura 4.3, na qual utiliza como exemplo de entrada as palavras *vaca*, *alimentadas* e *cana*, primeiramente efetua-se a (a) aplicação das ferramentas FORMA e CHAMA em cada palavra do texto. Esta aplicação resulta em: <*palavra_original*> <*lema*> <*substantivo_abstrato*> <*substantivo_concreto*> <*classe_gramatical*>. Cada palavra original do texto (indicada por <*palavra_original*>) é transformada para seu lema (<*lema*>), bem como é gerado, a partir dessa palavra original, o substantivo abstrato correspondente a essa palavra (indicado por <*substantivo_abstrato*>). Gera-se também o correspondente substantivo concreto (<*substantivo_concreto*>) e indica-se a classe gramatical (<classe_gramatical>) a qual a palavra original pertence.

Além disso, é colocada uma *tag* em cada palavra original, sendo que neste caso as *tags* colocadas foram: _SUB (indica substantivo) e _AP (indica particípio passado). O procedimento para a obtenção das palavras substantivadas desenvolvido neste trabalho e detalhado no Algoritmo 4.1, (b) considera como palavra substantivada o substantivo abstrato da palavra original, caso não possua, é considerado o substantivo concreto. Se as ferramentas não encontrarem nenhum destes substantivos, então, é considerado o lema, já que a substantivação de algumas das palavras originais corresponde ao seu próprio lema, como o caso das palavras *vaca* e *cana*.

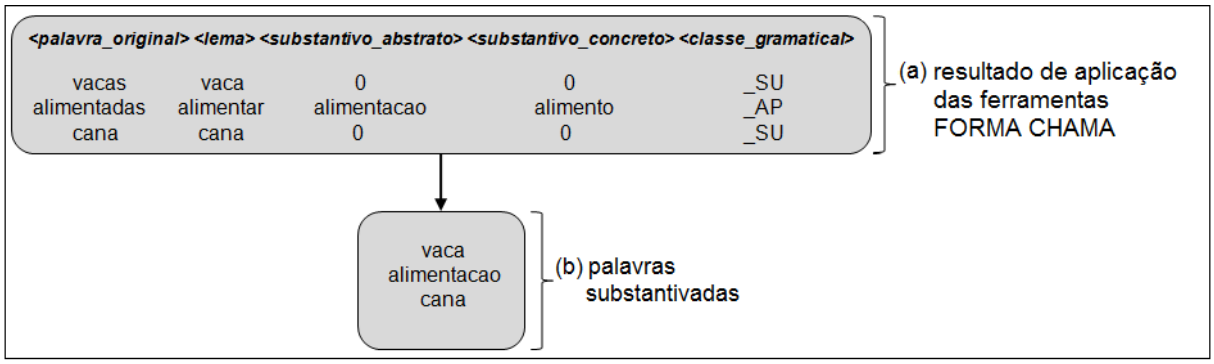


Figura 4.3: Exemplo da obtenção das palavras substantivadas

Algoritmo 4.1 Substantivação das palavras da coleção de textos

Require: coleção de documentos original D $\{d_1, d_2, \ldots, d_n\}$
Ensure: coleção de documentos com palavras substantivadas Ds
1: $D \leftarrow \{d_1, d_2, \ldots, d_n\}$
2: $Ds \leftarrow \{\}$
3: **for all** $d_i \in D$ **do**
4: $\quad d'_i \leftarrow$ aplica ferramentas FORMA e CHAMA (d_i)
5: $\quad ds_i \leftarrow$ pega palavras substantivadas (d'_i)
6: **end for**

Após a aplicação das técnicas supra citadas, são geradas três coleções de textos contendo palavras simplificadas, uma coleção para cada técnica. Para a obtenção de uma lista de unigramas a partir dessas palavras simplificadas, uma lista para cada coleção textual, aplica-se, separadamente, a ferramenta PreTexT II, com a opção de não radicalizar as palavras contidas nos documentos dessas coleções.

Mesmo com a lista de unigramas contendo somente termos simplificados de acordo com cada técnica (uma lista para cada técnica separadamente) tem-se um elevado número de unigramas, porém nem todos representam positivamente os documentos. Para a **extração dos unigramas** mais representativos e os correspondentes *n-gramas* que deles são derivados, pode-se considerar apenas os termos que aparecem, no mínimo, em d documentos na coleção textual (***Document Frequency*** - DF), sendo que d é fornecido pelo usuário e corresponde à quantidade mínima de documentos no qual o termo deve aparecer. Além disso, remove-se uma lista de ***stopwords*** padrão para português (com artigos,

interjeições, etc). Após a remoção dos unigramas a partir da utilização da *Document Frequency* e da remoção das *stopwords*, obtém-se os **unigramas selecionados**.

Em seguida, efetua-se a **extração dos *n-gramas*** (bigramas e trigramas). Para isso, os unigramas que foram removidos a partir da *Document Frequency* (ou qualquer outra medida que o usuário aplicar para remover os unigramas não representativos da coleção) formam uma nova lista de termos, denominada ***stoplist* da coleção** - lista de *stopwords* da coleção.

A *stoplist* da coleção visa remover as palavras dos documentos que não são indicadas para formarem bons bigramas e trigramas, já que estas palavras foram removidas na formação dos unigramas, o que indica que tal palavra não contribui para a representação da coleção.

Assim, cada uma das técnicas de simplificação terá uma lista de **bigramas** e uma lista de **trigramas selecionados**. Para exemplificar a diferença do uso de cada técnica na obtenção dessas combinações (*n-gramas*), a seguir são mostrados os trigramas extraídos dos parágrafos de um documento exemplo apresentados na Tabela 4.2. Os trigramas obtidos utilizando a técnica de radicalização a partir desses parágrafos são mostrados na Tabela 4.6.

virtud_bom_result	period_sec_ano	elev_consider_leit
bom_result_anim	sistem_pastej_extens	consider_leit_mam
result_anim_cresciment	pastej_extens_produca	leit_mam_bezerr
anim_cresciment_fazende	extens_produca_leit	mam_bezerr_final
cresciment_fazende_pass	produca_leit_vac	bezerr_final_period
fazende_pass_aliment	leit_vac_aliment	final_period_sec
pass_aliment_mistur	vac_aliment_can	period_sec_vac
aliment_mistur_can	aliment_can_acuc	sec_vac_apresent
mistur_can_ure	can_acuc_ure	vac_apresent_boa
can_ure_vac	acuc_ure_esperas	apresent_boa_condica
ure_vac_lactaca	ure_esperas_produca	boa_condica_corporal
vac_lactaca_period	esperas_produca_leit	condica_corporal_fertil
lactaca_period_sec	produca_leit_elev	corporal_fertil_adequ
	leit_elev_consider	

Tabela 4.6: Trigramas obtidos com a técnica de radicalização a partir dos parágrafos de um documento exemplo mostrados na Tabela 4.1

Já na Tabela 4.7, são apresentados os trigramas obtidos com a aplicação da técnica de lematização. Nota-se que as palavras ***bons*** e ***boa***, após a lematização, correspondem a palavra ***bom***, sendo esta eliminada pela lista de *stopwords* padrão. Com isso os trigramas *virtud_bom_result*, *bom_result_anim*, *vac_apresent_boa*, *apresent_boa_condica* e *boa_condica_corporal* não são construídos para a lematização e sim, formam-se diferentes trigramas com a remoção da palavra ***bom***, que são: *virtude_resultado_animal* e *apresentar_condicao_corporal*.

Na Tabela 4.8, é mostrado um exemplo dos trigramas obtidos utilizando a técnica de substantivação.

virtude_resultado_animal	sistema_pastejo_extensivo	leite_elevar_considerar
resultado_animal_crescimento	pastejo_extensivo_producao	elevar_considerar_leite
animal_crescimento_fazendeiro	extensivo_producao_leite	considerar_leite_mamar
crescimento_fazendeiro_passar	producao_leite_vaca	leite_mamar_bezerro
fazendeiro_passar_alimentar	leite_vaca_alimentar	mamar_bezerro_final
passar_alimentar_misturar	vaca_alimentar_cana	bezerro_final_periodo
alimentar_misturar_cana	alimentar_cana_acucar	final_periodo_seco
misturar_cana_ureia	cana_acucar_ureia	periodo_seco_vaca
cana_ureia_vaca	acucar_ureia_esperase	seco_vaca_apresentar
ureia_vaca_lactacao	ureia_esperase_producao	vaca_apresentar_condicao
vaca_lactacao_periodo	esperase_producao_leite	apresentar_condicao_corporal
lactacao_periodo_seco	producao_leite_elevar	condicao_corporal_fertilidade
periodo_seco_ano		corporal_fertilidade_adequar

Tabela 4.7: Trigramas obtidos com a técnica de lematização a partir dos parágrafos de um documento exemplo mostrados na Tabela 4.1

virtude_bons_resultados	periodo_secura_ano	elevacao_consideracao_leite
bons_resultados_animais	sistemas_pastejo_extensividade	consideracao_leite_mamacao
resultados_animais_crescimento	pastejo_extensividade_producao	leite_mamacao_bezerro
animais_crescimento_fazendeiros	extensividade_producao_leite	mamacao_bezerro_final
crescimento_fazendeiros_passagem	producao_leite_vacas	bezerro_final_periodo
fazendeiros_passagem_alimentar	leite_vacas_alimentacao	final_periodo_secura
passagem_alimentar_mistura	vacas_alimentacao_cana	periodo_secura_vacas
alimentar_mistura_cana	alimentacao_cana_acucar	secura_vacas_apresentacao
mistura_cana_ureia	cana_acucar_ureia	vacas_apresentacao_bondade
cana_ureia_vacas	acucar_ureia_esperase	apresentacao_bondade_condicao
ureia_vacas_lactacao	ureia_esperase_producao	bondade_condicao_corporal
vacas_lactacao_periodo	esperase_producao_leite	condicao_corporal_fertilidade
lactacao_periodo_secura	producao_leite_elevacao	corporal_fertilidade_adequacao
	leite_elevacao_consideracao	

Tabela 4.8: Trigramas obtidos com a técnica de substantivação a partir dos parágrafos de um documento exemplo mostrados na Tabela 4.1

Devido ao elevado número de combinações obtidas (bigramas e trigramas) com a aplicação de cada uma das técnicas de simplificação de termos, na tentativa de eliminar as combinações que não representam muito a coleção, são extraídas somente as combinações de palavras que contenham na lista de unigramas finais. Para isso, é utilizada a *stoplist* da coleção, que é a lista de palavras eliminadas no processo de construção dos unigramas finais da coleção.

Mesmo assim, quando se trata de coleções de textos de tamanhos elevados, o número dessas combinações (bigramas e trigramas) é ainda alto e grande parte delas não tem significado semântico. Neste sentido, é necessário aplicar-lhes **métodos estatísticos**, como a medida (***Document Frequency* - DF**) e o **teste da razão de máxima verossimilhança - MVS**.

O teste da razão de máxima verossimilhança, disponível no pacote NSP, visa detectar se as combinações são mais do que simples co-ocorrências casuais nos documentos, fornecendo, para isso, uma lista de todos os candidatos à combinações. Segundo **?** e **?**,

para a elaboração desta lista (para o caso de bigramas w^1w^2, por exemplo), faz-se necessário a formulação das duas hipóteses mostradas a seguir. Sendo que H = hipótese, P = probabilidade e w = palavra, que pertence à combinação (grama).

$$H1 : P(w^1|w^2) = P(w^1|\neg w^2)$$
$$H2 : P(w^1|w^2) \neq P(w^1|\neg w^2)$$

A hipótese 1 ($H1$) é a formalização da independência, isto é, a ocorrência de w^2 é independente da ocorrência de w^1. A hipótese 2 ($H2$) é a formalização da dependência, e quando ela é satisfeita significa que pode ter sido encontrada uma combinação interessante.

Por exemplo, assumindo que o bigrama *cana_acucar* é uma combinação, tem-se:

$$H2 : P(cana|acucar) \neq P(cana|\neg acucar)$$
$$H1 : P(cana|acucar) = P(cana|\neg acucar)$$

Espera-se que a hipótese de independência $H1$ seja falsa, indicando que provavelmente foi encontrada uma combinação interessante.

O resultado da aplicação deste teste é um *rank* dos *n-gramas* em ordem decrescente de importância de acordo com o valor obtido pelo teste para cada *n-grama*. Considerando ainda o documento utilizado como exemplo anteriormente, na Tabela 4.9, são listados alguns dos bigramas e trigramas obtidos com o uso da técnica de radicalização a partir de tal documento, juntamente com as respectivas freqüências no documento e os valores do teste da razão de máxima verossimilhança. Segundo os valores desse teste, o termo *sistem_pastej_extens* é considerado como mais importante do que o termo *leit_vac*.

N-gramas	Freqüências	Valores do teste
sistem_pastej_extens	4	18.9027
sistem_pastej	3	9.4990
pastej_extens	3	9.4990
can_acuc	4	6.7264
vac_aliment	7	3.0061
leit_vac	8	2.0609

Tabela 4.9: Exemplos dos valores obtidos pela aplicação do teste da razão de máxima verossimilhança para alguns bigramas e trigramas

Por fim, a base de textos, agora representada pelos termos extraídos da mesma, é estruturada em uma **matriz atributo-valor**, na qual o documento a que o termo pertence é colocado como linha da matriz, os termos extraídos como colunas, e a matriz é preenchida, neste caso, com a freqüência absoluta deste termo no respectivo documento. Deve-se ressaltar que o sucesso de tarefas de Extração de Padrões é diretamente afetado pela qualidade dos termos que compõem esta matriz. Ressalta-se, portanto, a importância de se obter termos representativos do domínio, sendo, então, necessário escolher uma técnica adequada para a extração desses termos.

4.3 Ferramenta ExtraT

Para a aplicação da metodologia de extração de termos proposta aqui foi desenvolvida uma ferramenta denominada ExtraT (Ferramenta para Extração de Termos) utilizando a linguagem de programação Java juntamente com o ambiente integrado para desenvolvimento de software Eclipse[2] Versão 3.3.2. Para a aplicação da técnica de lematização utiliza-se a base de palavras e seus respectivos lemas contidos no Lematizador de Nunes. Na aplicação da técnica de substantivação são utilizadas conjuntamente as ferramentas FORMA e CHAMA. Já para a aplicação da radicalização e a obtenção dos *n-gramas*, independente de qual técnica o usuário escolher, é utilizada a ferramenta PreTexT II. Por fim, a aplicação do método estatístico do teste da razão de máximo verossimilhança faz uso do pacote NSP.

A opção por utilizar a PreTexT é devido ao fato que esta, ao contrário das outras ferramentas de radicalização, possibilita também a combinação dos *stems* (*n-gramas*). Já a escolha pelo uso do Lematizador de Nunes é que este apresentou bons resultados em trabalhos anteriores, como no trabalho de **?**. Além disso, o FLANOM foi desenvolvido somente para a Língua Espanhola e o software Sphinx é proprietário. Quanto ao uso dos etiquetadores morfossintáticos citados, não é interessante utilizá-los para auxiliar na aplicação da lematização devido a necessidade de se aplicar outra ferramenta para lematizar o resultado destes etiquetadores.

Deve-se ressaltar que o trabalho aqui proposto tem um maior foco na abordagem estatística já que, segundo **?**, o uso de métodos lingüísticos têm como conseqüência sua aplicação somente a uma língua e, às vezes, até mesmo a uma única variante. Porém, a técnica de radicalização, por exemplo, pode ser aplicada em outras diversas línguas além do Português, já a substantivação, encontrada na literatura, só é aplicável à Língua Portuguesa. Tanto as *stopwords*, que podem ser obtidas ou construídas para outras línguas, quanto o *thesaurus*, ambos utilizados neste trabalho, estão inseridos na abordagem lingüística. O *thesaurus* é utilizado na abordagem proposta de avaliação de termos extraídos, explicada no Capítulo 5.

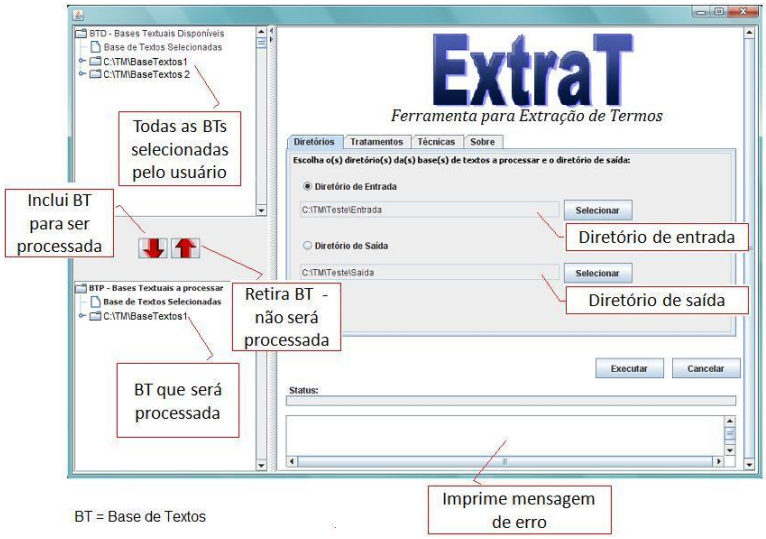
Dentro da abordagem estatística, este trabalho faz uso do teste da razão de máxima verossimilhança que visa indicar se um termo candidato é uma boa colocação para ser utilizada na coleção de textos em questão. Este teste foi escolhido para ser utilizado pois, segundo **?**, é o mais adequado para ser utilizado quando se trabalha com dados bastante esparsos, que é o caso da Mineração de Textos. Além disso, o estimador χ^2 é menos recomendado para trabalhar com coleções textuais pequenas e é menos interpretável, se comparado ao teste razão de máxima verossimilhança, pois este último fornece um *ranking*, não sendo necessário utilizar uma tabela, como no χ^2 (**?**).

Esta ferramenta permite que o usuário execute a extração de termos utilizando diferentes técnicas de simplificação dos termos para bases de textos da sua escolha. Os passos

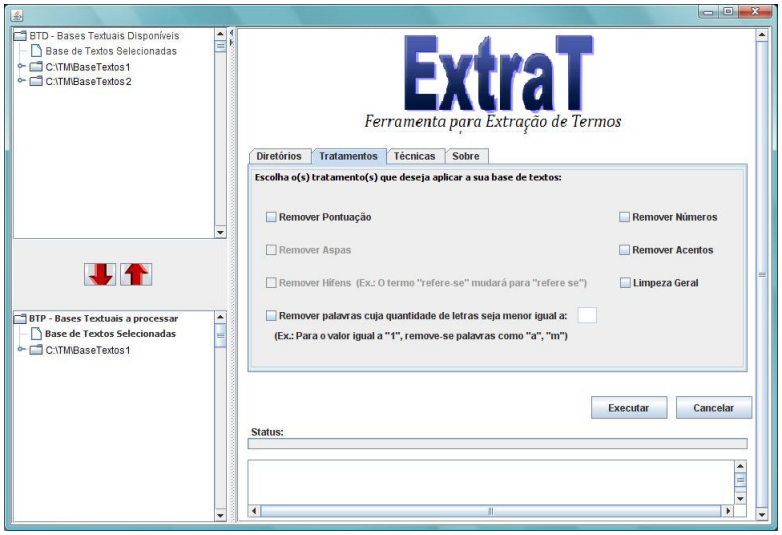
[2]Eclipse - http://www.eclipse.org/platform

para essa execução seguem as fases desta metodologia conforme descritas neste capítulo. Na Figura 4.4(a) é ilustrada a especificação dos diretórios de entrada e saída escolhidos pelo usuário, ou seja, como diretório de entrada tem-se as bases de textos que o usuário deseja processar e o diretório de saída corresponde ao diretório que o usuário deseja que sejam gerados os resultados do processamento.

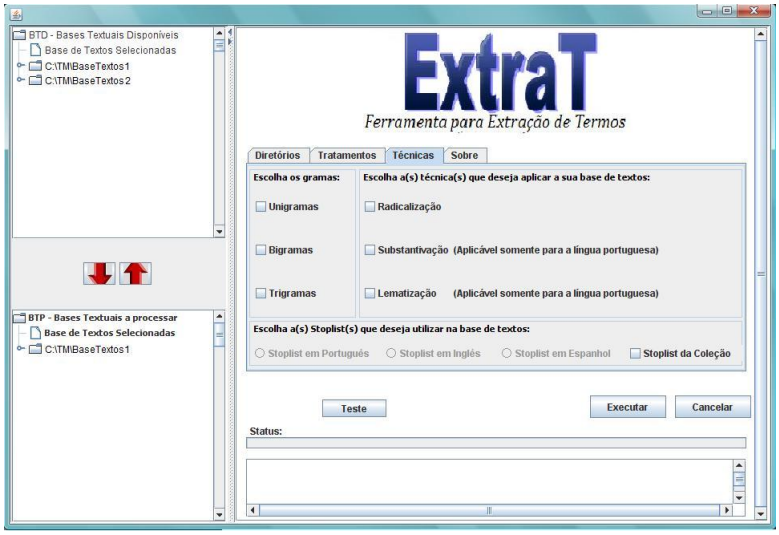
A primeira fase da metodologia, que é a preparação dos textos, pode ser executada na ferramenta ExtraT conforme mostrada na Figura 4.4(b), e a segunda fase da metodologia, que é a extração dos termos é ilustrada na Figura 4.4(c).



(a) Escolha de diretórios na ferramenta ExtraT



(b) Preparação dos textos na ferramenta ExtraT



(c) Extração dos termos na ferramenta ExtraT

Figura 4.4: Ferramenta para extração de termos - ExtraT

4.4 Considerações Finais

Neste capítulo foi descrita a metodologia para a aplicação de diferentes formas de extração de termos a partir de coleções textuais de domínio específico, bem como a ferramenta ExtraT, que foi desenvolvida para apoiar a utilização de tal metodologia. Essas diferentes formas englobam a utilização de três técnicas que simplificam os termos extraídos: a radicalização, a substantivação e a lematização.

Esta metodologia engloba duas fases, que é a preparação dos documentos da base de textos da qual é trabalhada e a extração dos termos utilizando as técnicas anteriormente citadas. Após a obtenção dos termos extraídos, estes devem ser avaliados garantindo que seu uso contribua positivamente para atingir os objetivos pré-estabelecidos pelo usuário. Sendo assim, neste trabalho também é proposta uma abordagem de avaliação, descrita no Capítulo 5, que faz uso de taxonomias *gold* e a medida de avaliação de termos CTW (*Context Term Weight*). Além disso, a metodologia de extração de termos pode ser estendida para outras técnicas de simplificação de termos além das três supra citadas que foram encontradas na literatura.

Capítulo

5

Avaliação dos Termos Extraídos

5.1 Considerações Iniciais

Na avaliação dos termos extraídos é verificado se os resultados obtidos estão em conformidade com os resultados esperados por meio de avaliações subjetivas e objetivas. Tanto para as avaliações subjetivas como para uma avaliação objetiva, faz-se uso de taxonomias *gold*. Neste trabalho, as taxonomias *gold* são taxonomias de referências, ou seja, são consideradas como taxonomias validadas e que podem ser utilizadas como modelo para experimentos. Como exemplo de trabalho que utilizou taxonomias *gold* para avaliação, pode-se citar o trabalho de **?**.

Neste capítulo também é descrita a avaliação experimental realizada utilizando os passos da metodologia de extração de termos detalhada no Capítulo 4. Para esse experimento realizado foram feitas sete avaliações, englobando avaliações subjetivas, conforme descrito na Seção 5.2.1; e avaliações objetivas, de acordo com a descrição da Seção 5.2.2. São apresentados os resultados obtidos nessa avaliação experimental e as análises dos mesmos. O experimento tem como objetivo avaliar as formas de extração de termos em um domínio específico. Além de apontar, por meio de avaliações subjetivas e objetivas, algumas das características positivas e negativas da utilização das técnicas de simplificação de termos (radicalização, lematização e substantivação), quando a compreensibilidade, a representatividade e a quantidade de termos têm impacto direto na interpretação e uso dos resultados.

5.2 Abordagem para Avaliação dos Termos Extraídos

Os termos extraídos a partir de coleções de textos podem ser utilizados para diversos fins, além de seu uso em taxonomias do domínio. Como o foco deste trabalho é avaliar

diferentes formas de extração de termos de um domínio específico, sendo que estas diferentes formas correspondem ao uso principalmente de três técnicas de simplificação de termos (radicalização, lematização e substantivação), deve-se avaliar a compreensibilidade dos mesmos, já que cada técnica de simplificação trabalha com métodos diferentes, extraindo, portanto, termos distintos. Também é importante avaliar se os termos extraídos correspondem ao domínio em questão.

Neste trabalho, a "qualidade" dos termos extraídos abrange tanto a representatividade dos termos no domínio como sua compreensibilidade. Neste sentido, para avaliar os termos extraídos são feitas algumas avaliações subjetivas e objetivas em relação ao domínio em questão, considerando cada uma das técnicas de simplificação de termos separadamente. As avaliações subjetivas, ou seja, com o auxílio de especialistas, abrangem: (i) a representatividade dos termos em seus respectivos documentos; (ii) a compreensibilidade dos termos obtidos ao utilizar cada técnica; e (iii) a preferência geral subjetiva dos especialistas em cada técnica. Como exemplo de avaliação subjetiva pode-se citar o trabalho de **?**, no qual foram feitas avaliações semelhantes para o domínio de Inteligência Artificial. No trabalho de **?** também foram feitas avaliações subjetivas, bem como objetivas, ambas voltadas para termos do domínio de agronegócio.

As avaliações objetivas levam em consideração (iv) a quantidade de termos extraídos por cada técnica, além de abranger também (v) a representatividade dos termos extraídos a partir de cada técnica em relação aos seus respectivos documentos.

Após feitas todas estas avaliações, considera-se ser possível ressaltar características positivas e negativas da utilização das técnicas de simplificação de termos.

5.2.1 Avaliações Subjetivas dos Termos Extraídos

As avaliações subjetivas dos termos extraídos tem como objetivo avaliar (i) o quão bem os termos extraídos representam seus respectivos textos, considerando cada uma das técnicas; (ii) a compreensibilidade dos termos extraídos ao utilizar cada técnica de simplificação de termos; e (iii) a preferência geral subjetiva dos especialistas em relação às técnicas.

Para que tais avaliações sejam possíveis, são efetuados três passos: (1) a **geração de taxonomias de tópicos utilizando as diferentes técnicas de simplificação de termos**, em seguida, é feita uma (2) **análise subjetiva das taxonomias por especialistas do domínio**, com uma forma clara de apresentação dos termos extraídos, permitindo avaliar qual forma é mais adequada para apoiar os especialistas do domínio em seus objetivos e possibilitar a implementação de melhorias na abordagem como um todo. Após esta análise, é efetuado um (3) **cálculo de precisão da representatividade dos termos** extraídos em relação ao domínio em questão. Estes três passos são explicados detalhadamente a seguir.

Passo 1: Geração de taxonomias de tópicos utilizando as diferentes técnicas de simplificação de termos

A metodologia para construção de taxonomias de tópicos proposta em **?** é aplicada às coleções de textos. Cada coleção é pré-processada e os termos representantes da mesma são extraídos utilizando separadamente as técnicas de simplificação de termos (radicalização, lematização e substantivação), conforme detalhado no Capítulo 3. Dessa forma, trabalha-se com três conjuntos de termos representados por suas respectivas matriz atributo-valor, sendo que o primeiro conjunto contém termos radicalizados; o segundo é composto por termos lematizados; e o terceiro contém termos substantivados.

Para a geração das taxonomias, durante a etapa de Extração de Padrões, para cada conjunto de termos, utiliza-se o algoritmo de agrupamento hierárquico *average-link* sobre cada matriz atributo-valor dos conjuntos de termos. Uma vez obtido um agrupamento hierárquico para cada conjunto de termos, as taxonomias são geradas a partir da rotulação da hierarquia, na qual são considerados os termos mais freqüentes. Os termos são selecionados e ordenados de acordo com sua freqüência em cada grupo de documentos.

Desta forma, para cada coleção textual são obtidas três versões de taxonomias com a mesma estrutura hierárquica, porém rotuladas com os termos obtidos em cada técnica de simplificação, conforme ilustrado na Figura 5.1.

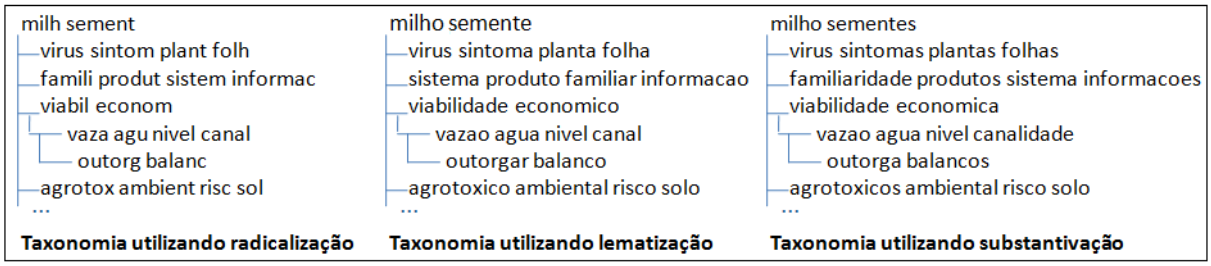


Figura 5.1: Exemplo de taxonomias com a mesma estrutura hierárquica utilizando as diferentes técnicas de simplificação de termos

Devido à existência da mesma estrutura hierárquica das taxonomias, para cada conjunto de termos, é possível selecionar ramos que representam um mesmo grupo de documentos para avaliação. Isto permite que os usuários comparem a compreensibilidade dos termos utilizados nos rótulos de cada grupo sobre um mesmo cenário, uma vez que o mesmo conceito é apresentado em cada grupo. Além disso, espera-se que bons grupos sejam compactos de modo que seus documentos apresentem alta similaridade, enquanto que a similaridade com os elementos de outros grupos seja a menor possível. Neste caso, a similaridade foi calculada utilizando a distância de cosseno.

Como as taxonomias geradas podem possuir um tamanho elevado para serem avaliadas subjetivamente pelos especialistas, são selecionados os 10 (dez) grupos da hierarquia que apresentam a menor variabilidade interna da similaridade entre os documentos (os grupos mais coesos) para facilitar a análise subjetiva do conceito representado naquele grupo por

parte dos especialistas e permitir a avaliação da qualidade dos termos ali selecionados.

Passo 2: Análise subjetiva das taxonomias por especialistas do domínio

Como o objetivo deste trabalho não é a geração de taxonomias e, sim de termos, não é considerada a avaliação de taxonomias em si, pois para isso seria necessário uma pesquisa abrangente sobre as formas e métodos disponíveis para geração das mesmas. Por isso, neste trabalho as taxonomias são utilizadas para auxiliar a avaliação dos termos extraídos.

Sendo assim, com o auxílio da ferramenta TaxTools (**?**), os especialistas do domínio em questão devem avaliar os ramos escolhidos das taxonomias obtidas e examinar os documentos e descritores (termos) obtidos, sendo que a partir de cada observação, os especialistas devem eleger uma nota ao ramo, entre um (1) e quatro (4) correspondente a:

1. Nota 1 - Muito ruim, termos nada discriminativos
2. Nota 2 - Termos pouco discriminativos
3. Nota 3 - Termos bem discriminativos
4. Nota 4 - Termos muito bons

Esta nota diz respeito a representatividade subjetiva dos termos em relação ao domínio em questão, isto é, o quão os termos extraídos representam seus respectivos documentos.

Em seguida, passa-se para a segunda parte da avaliação, conforme ilustrado na Figura 5.2. A questão (a) avalia a compreensibilidade dos termos, ou seja, qual (ou quais) técnica obteve termos mais compreensíveis - fáceis de serem entendidos pelos especialistas (usuários finais dos termos). A questão (b) avalia qual das técnicas os especialistas consideram como a mais adequada para ser utilizada na coleção textual no domínio.

Assim, considera-se ser possível avaliar os termos extraídos de acordo com a compreensibilidade dos mesmos a partir da análise dos especialistas do domínio e ainda por sua preferência por utilizar alguma das técnicas.

Passo 3: Cálculo da representatividade e compreensibilidade dos termos

As notas atribuídas pelos especialistas são tabuladas de acordo com as notas dadas nos ramos em cada taxonomia de cada coleção textual. Dessa forma, é possível realizar testes comparativos, com base em análise de variância das notas em função de cada um dos efeitos envolvidos na avaliação, que são: em relação ao ramo (nó) avaliado; em relação às diferentes técnicas de simplificação de termos utilizadas; e em relação ao avaliador.

Estes testes comparativos são realizados por meio de comparações múltiplas de médias, indicando se há ou não diferença estatisticamente significativa entre as médias de cada técnica utilizada. Com isso, é possível indicar se alguma técnica obteve resultados melhores nestes testes em relação às outras.

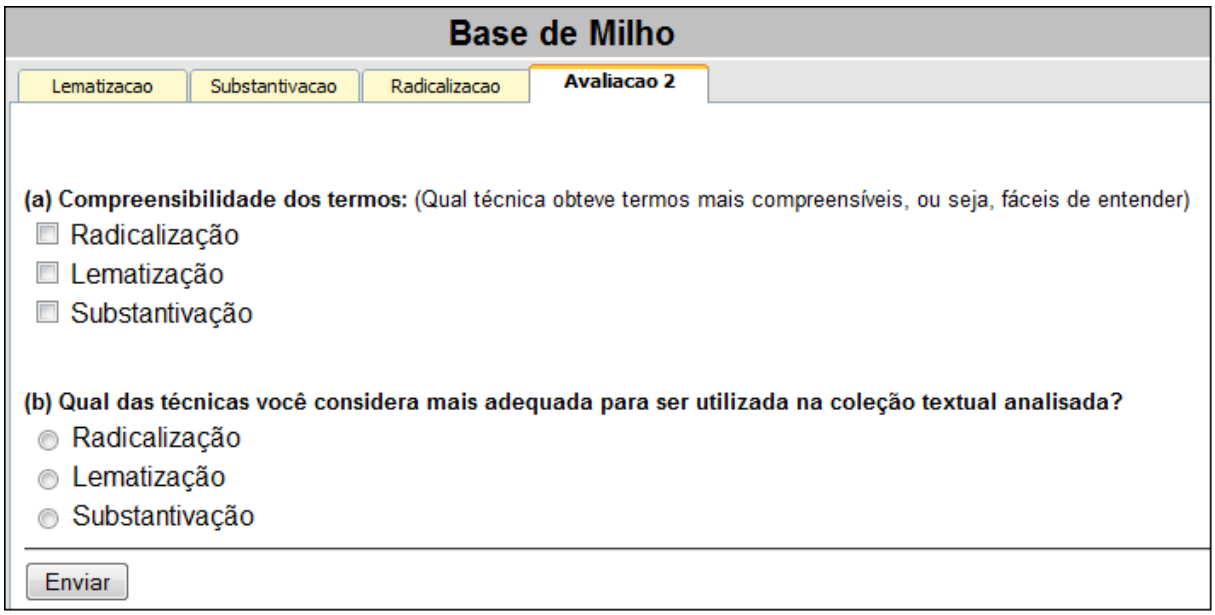


Figura 5.2: Avaliação do ramo selecionado

5.2.2 Avaliações Objetivas dos Termos Extraídos

A quantidade de termos extraídos com o uso de diferentes técnicas de simplificação de termos (radicalização, lematização e substantivação), utilizando a mesma base de textos, é diferente para cada técnica. Neste sentido, é necessário considerar na avaliação dos termos, além da compreensibilidade e representatividade dos termos, (iv) a quantidade de termos extraídos por cada técnica.

Para avaliar objetivamente (v) a representatividade dos termos extraídos a partir de cada técnica separadamente em relação a um determinado domínio, é utilizado um vocabulário expandido do mesmo domínio.

Foram escolhidas, para caso de uso, algumas árvores publicadas pela Agência de Informação Embrapa[1] como taxonomias *gold*. Devido a algumas diferenças entre formas de armazenamento dos dados da Agência e dos dados no ambiente TopTax, foi feita uma preparação dos dados das Agências de Informação Embrapa para o padrão necessário à comparação com o vocabulário automaticamente obtido. Ressalta-se que, em trabalhos futuros, esta preparação poderá ser utilizada para auxiliar na avaliação automática de taxonomias de tópicos automaticamente geradas contra taxonomias de tópicos *gold*. Essa avaliação poderá ser efetuada comparando-se, por exemplo, cada um dos tópicos de uma taxonomia gerada automaticamente com cada um dos tópicos de uma taxonomia *gold*, assim, poder-se-á avaliar a rotulação feita para a geração da taxonomia, bem como os termos contidos em cada tópico da mesma.

Para melhor compreensão dos passos necessários para a avaliação objetiva dos termos e, possivelmente, de taxonomias de domínios específicos, é detalhada a taxonomia utilizada do domínio de agronegócio, que é a taxonomia de tópicos da Agência de Informação

[1] Agência de Informação Embrapa - http://www.agencia.cnptia.embrapa.br/

Embrapa. Bem como, são descritas a adaptação desta taxonomia para ser utilizada como base na avaliação (taxonomia *gold*) com a utilização do Thesagro, e as duas possíveis formas de avaliação disponíveis com esse processo.

A escolha pela utilização do Thesagro é devida ao fato que este atende às necessidades da metodologia de avaliação apresentada aqui. Tais necessidades são: (i) a obtenção de um vocabulário expandido; e (ii) o uso de bases de textos do mesmo domínio do vocabulário expandido. Nesse sentido, como tem-se disponíveis as bases de textos do domínio de agronegócio e o Thesagro, que é do domínio agrícola validado e consolidado, torna-se viável e interessante o seu uso.

A Taxonomia de Tópicos da Agência de Informação Embrapa

Uma das formas de divulgação de informação na Embrapa é a Agência de Informação, que "é um sistema Web que possibilita a organização, o tratamento, o armazenamento, a divulgação e o acesso à informação tecnológica e ao conhecimento gerados pela Embrapa e outras instituições de pesquisa"[2] (**?**).

Os dados e informações contidos na Agência são organizados em hierarquias, denominadas árvores do conhecimento. Nos primeiros níveis das hierarquias estão os conhecimentos mais genéricos e, nos níveis mais profundos, estão os conhecimentos específicos (**?**). Cada nó da hierarquia corresponde a um tema (tópico) descrito por um texto, que resulta da compilação do conhecimento produzido por pesquisadores, técnicos extensionistas e agricultores; e, ainda faz referências a outras obras que complementam essa informação. Tais hierarquias são visualizadas por meio de árvores hiperbólicas ou navegação por hipertexto das páginas Web. Adicionalmente, são fornecidas ferramentas de busca alimentadas por palavras-chave.

Utilizar árvores da Agência de Informação Embrapa como taxonomias *gold* é interessante, uma vez que a informação é pública, consolidada, validada e completa. Mesmo assim, algumas adaptações foram necessárias, devido às características das taxonomias envolvidas no processo.

A primeira característica é que para cada nó, em uma taxonomia de tópicos automaticamente obtida, o tópico possui um ou mais termos como palavras-chave a serem validadas, e cada um desses termos possui termos similares, que poderiam ter sido também automaticamente encontrados. Na árvore de conhecimento da Agência, cada nó é representado apenas pelos termos mais específicos, que foram subjetivamente selecionados a partir de um *thesaurus*. Por exemplo, o termo *Abacaxi*, neste domínio, poderia também ser representado pelo termo *Ananas Comosus*, assim, o nó deveria conter os termos *Abacaxi* e *Ananas Comosus*. Dessa forma, para comparar a taxonomia automaticamente gerada com uma *gold*, sem perda de informação, foi necessário expandir os termos em cada nó das árvores da Agência, buscando termos sinônimos e/ou relacionados no *thesaurus* utilizado, no caso, o Thesagro (*Thesaurus* Nacional Agrícola).

[2]Agência de Informação Embrapa - http://www.agencia.cnptia.embrapa.br/

A segunda característica é que, taxonomias automaticamente obtidas agrupam documentos em tópicos, e, a seguir procuram palavras-chave para descrever os tópicos nesses documentos. Na Agência, em cada nó da hierarquia há uma página Web contendo uma síntese assunto do nó em questão e, possivelmente mais documentos relacionados ao nó são referenciados, bem como metadados desses documentos. Novamente, uma adaptação faz-se necessária, a fim de abstrair esses textos como documentos agrupados sob o nó.

Adaptação da Árvores da Agência a uma Taxonomia Gold

O processo de obtenção automática de taxonomias de tópicos necessita da extração de termos capazes de representar com qualidade cada nó. Além disso, esses termos podem e devem ser utilizados como palavras-chave dos documentos, podendo alimentar expressões de busca na coleção de textos ou em coleções similares. Deve-se ressaltar que, não necessariamente, esses termos representam adequadamente o domínio de conhecimento completo, pois são obtidos a partir de uma coleção restrita de documentos.

Adicionalmente, espera-se que o volume de termos extraídos automaticamente da coleção seja maior do que um conjunto de termos selecionados por humanos em um *thesaurus*. Assim, para avaliar a "qualidade" dos termos extraídos das mesmas coleções de documentos que geraram algumas árvores da Agência, implementaram-se os processos de adaptação de algumas árvores da Agência a uma taxonomia *gold*. Para efetuar as adaptações foi desenvolvida a ferramenta TaXEm (Taxonomia em XML da Embrapa), que utiliza como base de vocabulário controlado o contido no Thesagro. Deve-se ressaltar que, apenas Agências de produto foram utilizadas, a fim de garantir que o *thesaurus* utilizado fosse o Thesagro; dado que essas árvores são baseadas em cadeia produtiva de produtos agrícolas.

Conforme apresentado na Figura 5.3, o processo de adaptação proposto neste trabalho consiste em: a partir dos arquivos que armazenam a estrutura hierárquica das Agências, reestruturá-la de modo que seja possível facilmente vincular os documentos relacionados a cada nó; e que, cada nó seja rotulado com os termos que o representam, viabilizando uma busca mais ampla.

Assim, para cada entrada, a hierarquia dos nós para o produto é preparada em formato *.xml*, e em cada nó, os documentos relativos ao mesmo são referenciados. Estes documentos são gerados em formato de texto plano em três diferentes categorias: *conteúdo*, *mais detalhes* e *recursos*. Os documentos pertencentes à categoria *conteúdo* descrevem o tópico, nos moldes da Agência; os documentos da categoria *mais detalhes* são os metadados sobre os recursos; e os pertencentes a *recursos* são os próprios recursos disponíveis do nó, isto é, as informações complementares citadas no texto que descreve o nó.

Os termos originais da hierarquia são expandidos com a utilização do Thesagro, isto é, são aumentadas as possibilidades de cada palavra-chave acrescentando-lhes sinônimos ou termos relacionados. As relações existentes no Thesagro foram explicadas mais detalhadamente na Seção 3.4 do Capítulo 3. Utilizando a notação do Thesagro, na expansão de cada termo T (descritor), consideraram-se os termos relacionados (RT) ao termo buscado

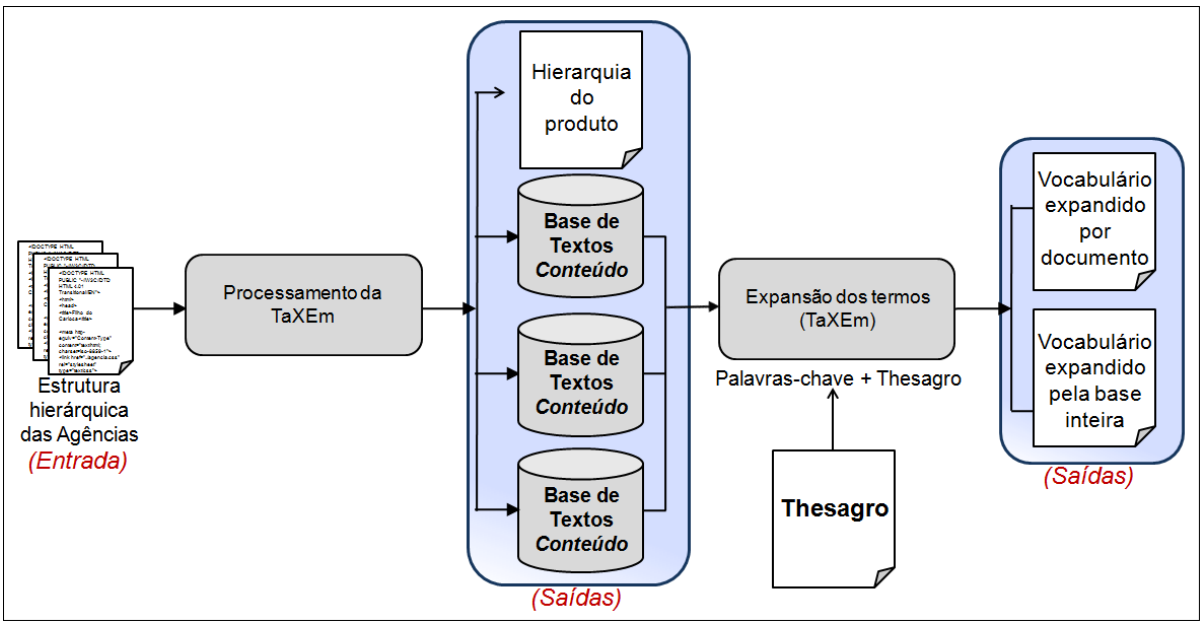


Figura 5.3: Processo para a preparação do conteúdo das Agências

T; sendo que, os RT possuem significados que se relacionam semanticamente com o termo T, mas sem nenhuma ligação hierárquica entre si. Também são considerados os termos que possuem alguma relação de equivalência (denominados USE), que são classificados, neste trabalho, como os termos que possuem alguma relação de equivalência ao termo T. Na Figura 5.4 são mostrados exemplos de (a) RT e (b) USE. O termo (descritor) *enologia* possui um significado semântico relacionado ao termo (RT) *vinho*, e o termo (descritor) *abacaxizeiro* possui uma relação de equivalência com o termo *abacaxi*, ou seja, o termo *abacaxizeiro* pode ser usado (USE) como *abacaxi*.

```
<TERMO>
        <ID>0022</ID>
        <DESCRITOR>ENOLOGIA</DESCRITOR>
        <RT>VINHO</RT>
</TERMO>
```

(a) Exemplo de *RT*

```
<TERMO>
        <ID>2966</ID>
        <DESCRITOR>ABACAXIZEIRO</DESCRITOR>
        <USE>ABACAXI</USE>
</TERMO>
```

(b) Exemplo de *USE*

Figura 5.4: Exemplos de ***RT*** e ***USE***

O vocabulário expandido, obtido com a utilização do Thesagro, é apresentado de duas formas: (a) separado por documento e (b) referente à base inteira, conforme ilustrado na Figura 5.5.

Formas de Avaliação Objetiva dos Termos Extraídos

A primeira forma de avaliação objetiva visa analisar objetivamente a representatividade dos termos extraídos em relação ao domínio da coleção de textos que os mesmos foram extraídos. Para isso, utilizou-se como suporte a medida CTW (*Context Term Weight*)

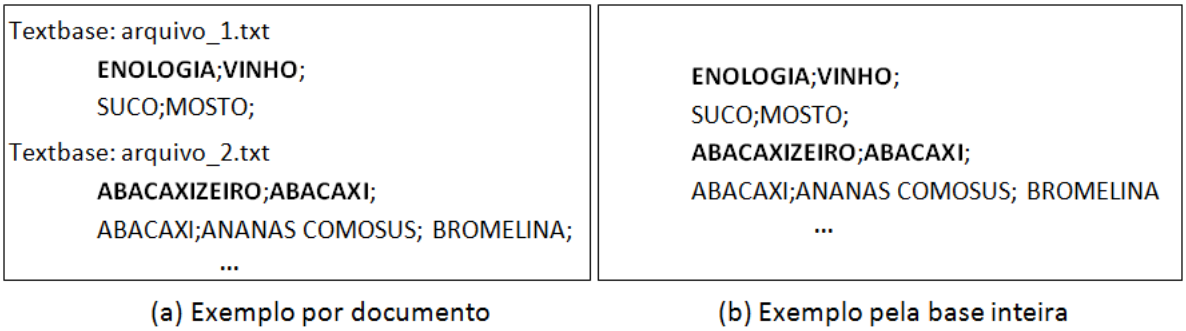


Figura 5.5: Vocabulário expandido

(**?**), que avalia a quantidade de vezes (a freqüência do termo na coleção) em que um termo extraído aparece em um determinado contexto. Neste caso, o contexto é representado por um vocabulário expandido obtido pela TaXEm, ou seja, os termos representativos e consagrados do domínio juntamente com seus sinônimos, que são obtidos do Thesagro. Sendo assim, para a avaliação dos termos, os mesmos são recuperados no vocabulário expandido. A descrição formal da medida CTW (*Context Term Weight*) adaptada para este trabalho é:

$$CT(a) = \sum_{d \epsilon T_a} f_a(d) \tag{5.1}$$

no qual a é o termo extraído seguindo os passos da metodologia de extração de termos apresentada neste trabalho; T_a é o conjunto das palavras do vocabulário expandido que coincidem com os termos extraídos; d é a palavra de T_a; e $f_a(d)$ é a freqüência de d como um termo a, que é obtida durante a extração de termos seguindo os passos da metodologia apresentada neste trabalho.

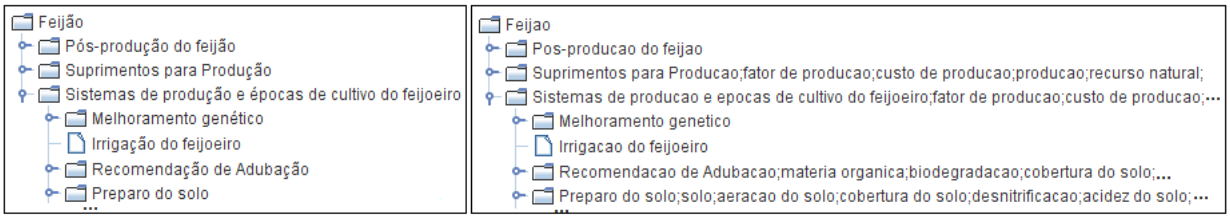
Na Tabela 5.1 é mostrado um exemplo do uso da medida CTW (*Context Term Weight*) neste trabalho. Neste exemplo, a pontuação CTW é igual a 30, que é obtida somente a partir das freqüências do termos *abacaxi* (20) e *ananas* (10), pois o termo *ruim* não está presente no vocabulário expandido, não sendo, portanto, considerado como um termo importante para o domínio do vocabulário.

Termos extraídos		*Termos do vocabulário expandido*	*Pontuação da CTW*
Termos (a)	Freqüências (f_a)		
abacaxi	20	abacaxi, ananas, abacaxizeiro.	
ananas	10	maçã, macieira.	30
ruim	5		

Tabela 5.1: Exemplo do uso da medida CTW neste trabalho

Além disso, com o auxílio da TaXEm é possível haver uma segunda forma de avaliação objetiva, na qual considera como base a taxonomia *gold*, com o objetivo de verificar se os termos extraídos representam os tópicos e sub-tópicos do domínio. Dessa forma, as comparações podem ser feitas verificando os termos de cada nó em cada hierarquia ou verificando a posição dos nós das hierarquias. Isso é possível devido ao padrão da

hierarquia preparada na TaXEm, o que permite sua fácil visualização, utilizando, por exemplo, a ferramenta TaxTools (**?**), como ilustrada na Figura 5.6 que usa como exemplo o produto Feijão, conforme publicado na Agência Embrapa. Nesta figura é mostrada (a) a hierarquia da árvore do Feijão reestruturada pela TaXEm contendo somente os termos originais e (b) a mesma hierarquia contendo os termos expandidos com o vocabulário.



(a) Hierarquia com termos originais (b) Hierarquia com termos expandidos

Figura 5.6: Exemplo da visualização da hierarquia

Como nesta segunda forma de avaliação, cada nó da taxonomia *gold* é comparado com o nó da taxonomia obtida. Nesta avaliação, considera-se os termos extraídos, a forma como foi gerada a taxonomia com tais termos, como a rotulação utilizada em cada uma, já que este pode afetar a posição de cada nó e, conseqüentemente, se o resultado obtido desta avaliação foi afetado por isso. Devido ao foco deste trabalho, não será considerada esta segunda forma de avaliação.

5.3 Bases de Textos Utilizadas

As bases de textos a ser utilizadas podem vir de qualquer domínio, pois tal metodologia de extração termos apresentada neste trabalho é independente do domínio. Para experimento, foram utilizadas bases de textos reais disponíveis na Embrapa, devido à necessidade de possuir os textos das bases e seus respectivos termos originais para poder avaliar se os termos aqui extraídos obtiveram resultados satisfatórios.

Para a avaliação objetiva dos termos extraídos, foi necessário utilizar um *thesaurus* do mesmo domínio das bases de textos, sendo assim, houve um reforço positivo pela escolha destas bases, pois seu domínio é o mesmo do domínio do *thesaurus* disponível (o Thesagro). Dessa forma, o experimento foi conduzido utilizando oito bases de textos na Língua Portuguesa referentes a oito produtos diferentes cultivados na Embrapa. A quantidade de documentos destas bases é mostrada na Tabela 5.2. Estas bases foram pré-processadas, conforme descrição do Capítulo 2, e, como conseqüência, alguns documentos foram removidos por não estarem adequados para a extração de termos. Portanto, a quantidade de documentos de algumas bases foi diminuída (documentos finais).

Base de Textos (Produto)	*# Docs Iniciais*	*# Docs Finais*	*Média de palavras por texto*
Milho	511	510	705.04
Cana	431	391	1272.81
Feijão	352	348	801.63
Leite	332	332	278.40
Maçã	231	230	895.97
Caupi	204	198	1933.70
Eucalipto	100	100	277.20
Caju	40	40	531.50

Tabela 5.2: Descrição das bases de textos

5.4 Avaliação Experimental

Os termos foram extraídos a partir de coleções textuais do domínio de agronegócio de três diferentes formas. Essa extração seguiu os passos da metodologia descrita no Capítulo 4. Ressaltando que cortes, como Luhn e Salton, não foram utilizados no experimento deste trabalho, mesmo que a metodologia apresentada aqui possibilita o uso dos mesmos. Tal escolha tem o objetivo de diminuir a subjetividade da metodologia, já que estes exigem que sejam fornecidos os pontos de corte.

A primeira forma de extração de termos utilizou para a simplificação dos termos a técnica de radicalização; a segunda forma fez uso da técnica de lematização; e a terceira utilizou a técnica de substantivação. As avaliações são descritos a seguir.

5.4.1 Avaliação 1 - O uso da Metodologia de Extração de Termos Proposta

Objetivo: verificar se a extração de termos seguindo os passos descritos na metodologia de extração de termos contribui consideravelmente para a diminuição da quantidade de termos, já que esta é um problema existente no processo de Mineração de Textos.

Hipótese: a extração de termos seguindo os passos descritos na metodologia de extração de termos contribui consideravelmente para a diminuição da quantidade de termos.

Descrição: primeiramente, para cada base de textos e técnica utilizada para a extração de termos, foram extraídos (a) os *termos iniciais* (unigramas, bigramas e trigramas), removendo-se somente a lista de *stopwords* padrão para português disponível na PreTexT, as conjugações do verbo SER e as palavras compostas por apenas um caracter. Logo em seguida, foram extraídos (b) os termos finais, que foram obtidos seguindo os passos da metodologia descrita no Capítulo 4. Para a extração dos termos finais, foram considerados apenas os termos que aparecem, no mínimo, em dois documentos na coleção textual (*Document Frequency* - $DF >= 2$); e para o teste da razão de máxima verossimilhança adotou-se $p\_value = 0.05$, visando manter somente os bigramas e trigramas que contenham algum significado semântico.

Com a extração de *termos iniciais* e *finais* separadamente é possível observar a redução da quantidade de termos obtida quando a extração de termos é feita seguindo os passos da metodologia proposta neste trabalho, ou seja, quando se trabalha com os termos finais. Tal redução, utilizando as três técnicas de simplificação de termos separadamente, é mostrada

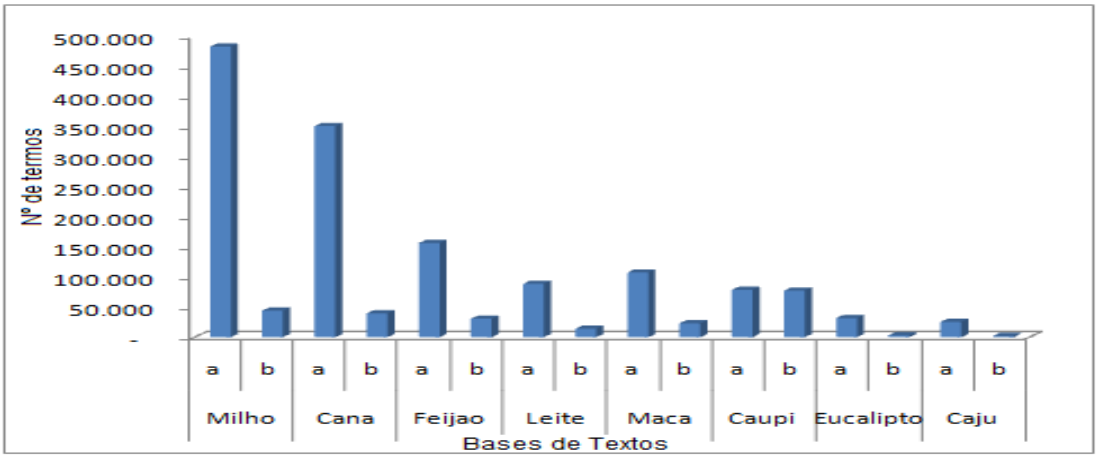


Figura 5.7: Redução do número de termos utilizando a técnica de radicalização

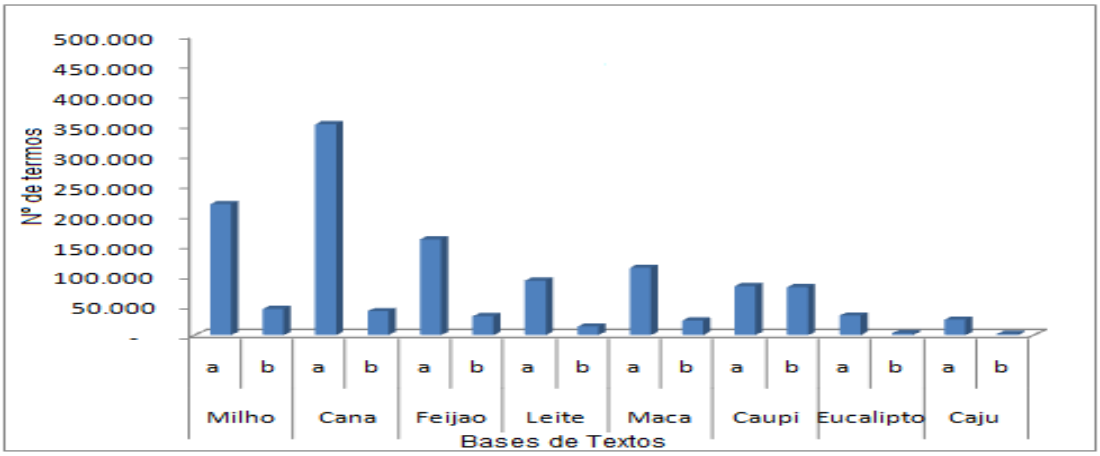


Figura 5.8: Redução do número de termos utilizando a técnica de lematização

nas Figuras 5.7, 5.8 e 5.9; e as quantidades *iniciais* e *finais* de unigramas, bigramas e trigramas extraídos (número de termos e porcentagem de redução) são mostradas na Tabela 5.3. A porcentagem de redução mostrada nessa Tabela, diz respeito ao percentual de redução da quantidade de termos quando comparado o total de *termos iniciais (a)* com o total de *termos finais (b)* referentes à cada técnica.

Análise e Resultados: a extração de termos seguindo os passos descritos na metodologia de extração de termos contribui consideravelmente para a diminuição da quantidade de termos quando comparados os *termos inciais* e *finais*. Esta característica deve-se ao fato que a metodologia visa manter os termos que realmente interessam para o domínio e, para isso, faz uso de métodos estatísticos descritos no Capítulo 4. Ressalta-se que, fixando-se os métodos estatísticos utilizados, seria interessante avaliar o efeito dessa diminuição da quantidade de termos no final do processo, por exemplo, avaliar esse efeito em uma taxonomia final obtida que utiliza tais termos.

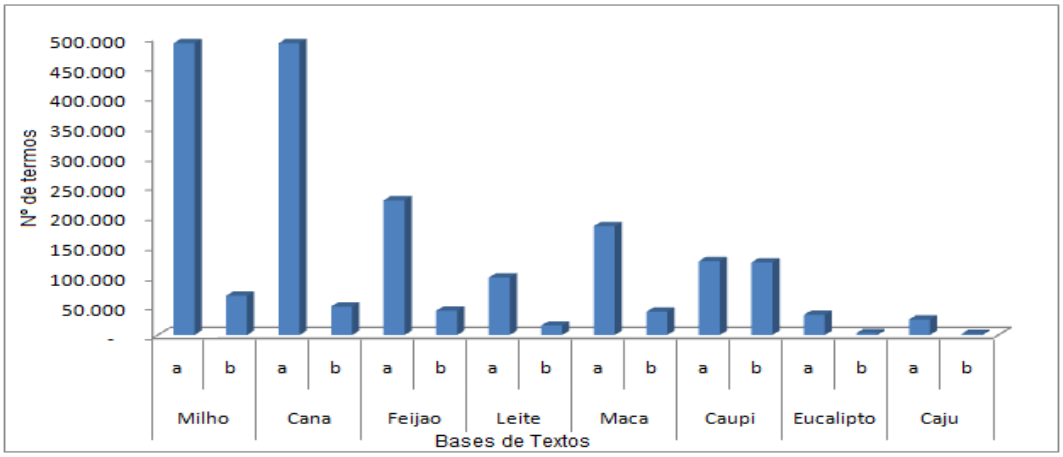


Figura 5.9: Redução do número de termos utilizando a técnica de substantivação

5.4.2 *Avaliação 2 - Quantidade de Termos Obtidos Utilizando as Técnicas de Simplificação de Termos*

Objetivo: avaliar a quantidade de termos obtidos por cada técnica de simplificação de termos.

Hipótese: o uso das técnicas de lematização e substantivação obtêm quantidades de termos maiores do que quando utilizada a técnica de radicalização.

Descrição: conforme descrito no avaliação 1, foram extraídos termos seguindo os passos da metodologia apresentada neste trabalho. Para cada base de textos, foram utilizadas as três técnicas de simplificação de termos, visando observar as diferentes quantidades de termos extraídos com cada técnica. Para melhor comparação das quantidades de termos obtidos por cada técnica, na Figura 5.7 é mostrada a quantidade de termos obtidos quando utilizada a técnica de radicalização. A quantidade de termos obtidos com o uso da lematização é apresentada na Figura 5.8. Por fim, na Figura 5.9 é a mostrada a quantidade quando utilizada a técnica de substantivação. Já na Tabela 5.3, essas quantidades são apresentadas como mais detalhes, ou seja, são mostrados os números de unigramas, bigramas e trigramas obtidos separadamente para cada técnica. Ainda nessa tabela, para cada base de textos, estão destacadas (em cor azul) as quantidades de termos menores obtidas utilizando as técnicas de simplificação de termos.

Análise e Resultados: o uso da técnica de radicalização geralmente obtém uma quantidade de termos inferior do que quando utilizada a técnica de lematização. Por último, o uso da técnica de substantivação gera mais termos do que as duas outras técnicas. Isso pode ser explicado pelo fato da técnica de radicalização ser mais agressiva para simplificar os termos em relação as técnicas de lematização e substantivação.

5.4.3 *Avaliação 3 - Representatividade Objetiva dos Termos Extraídos*

Objetivo: verificar objetivamente se o processo de extração de termos utilizando cada uma das três técnicas (radicalização, substantivação e lematização) obteve termos

		Radicalização		*Lematização*		*Substantivação*	
Bases	***Gramas***	(a)	(b)	(a)	(b)	(a)	(b)
Milho	Unigramas	19.883	6.526	15.791	6.934	23.556	7.744
	Bigramas	167.412	15.838	93.603	21.548	197.077	28.884
	Trigramas	295.105	21.808	109.478	15.107	270.678	29.543
	Total	482.400	44.172	218.872	43.589	491.311	66.171
	% Redução∗	**91%**		**80%**		**87%**	
Cana	Unigramas	20.249	8.316	21.960	8.697	23.556	9.501
	Bigramas	177.858	21.019	153.088	21.067	197.077	24.785
	Trigramas	152.414	10.157	177.460	10.197	270.678	13.738
	Total	350.521	39.492	352.508	39.961	491.311	48.024
	% Redução∗	**89%**		**89%**		**90%**	
Feijão	Unigramas	10.542	5.401	11.631	5.798	12.599	6.243
	Bigramas	78.915	14.321	67.835	14.572	88.311	17.993
	Trigramas	66.881	10.918	80.585	11.233	125.817	16.652
	Total	156.338	30.640	160.051	31.603	226.727	40.888
	% Redução∗	**80%**		**80%**		**82%**	
Leite	Unigramas	5.419	3.051	6.403	3.454	6.568	3.524
	Bigramas	36.948	6.663	37.936	6.681	40.228	7.257
	Trigramas	46.210	4.130	46.789	4.132	49.807	5.127
	Total	88.577	13.844	91.128	14.267	96.603	15.908
	% Redução∗	**84%**		**84%**		**84%**	
Maçã	Unigramas	7.988	4.537	9.412	5.174	9.595	5.225
	Bigramas	51.661	12.075	53.863	12.426	74.347	17.556
	Trigramas	47.723	6.545	49.115	6.654	99.375	16.284
	Total	107.372	23.157	112.390	24.254	183.317	39.065
	% Redução∗	**78%**		**78%**		**79%**	
Caupi	Unigramas	10.826	10.826	12.064	11.989	12.196	12.180
	Bigramas	33.180	31.991	34.267	32.994	49.424	48.044
	Trigramas	34.534	34.534	35.607	35.169	62.811	62.029
	Total	78.540	77.351	81.938	80.152	124.431	122.253
	% Redução∗	**2%**		**2%**		**2%**	
Eucalipto	Unigramas	3.185	1.456	3.659	1.564	3.692	1.630
	Bigramas	13.243	1.263	13.546	1.210	14.149	1.231
	Trigramas	15.294	413	15.431	411	16.102	434
	Total	31.722	3.132	32.636	3.185	33.943	3.295
	% Redução∗	**90%**		**90%**		**90%**	
Caju	Unigramas	2.703	1.142	2.868	1.152	3.053	1.228
	Bigramas	10.513	662	10.658	670	10.862	589
	Trigramas	12.160	148	12.261	147	12.281	135
	Total	25.376	1.952	25.787	1.969	26.196	1.952
	% Redução∗	**92%**		**92%**		**93%**	

∗Porcentagem de redução da quantidade de termos quando comparados os totais de *(a)* e *(b)* referentes à cada técnica.

Tabela 5.3: Quantidade de unigramas, bigramas e trigramas extraídos

representativos / importantes para o domínio em questão.

Hipótese: o processo de extração de termos obtém uma quantidade satisfatória de termos representativos/importantes para o domínio em questão.

Descrição: para os termos extraídos com o uso de cada técnica separadamente utilizando cada base de textos é verificado objetivamente se os mesmos representam o domínio. Tal verificação é feita conforme detalhada na Seção 5.2.2, com a utilização do vocabulário expandido do mesmo domínio.

Análise e Resultados: os resultados desta avaliação são mostrados na Tabela 5.4, na qual a pontuação CTW para cada técnica aplicada em cada base de textos e seus respectivos vocabulários expandidos são apresentados.

Bases	*# Docs*	*# Termos do Vocabulário Expandido*	*Radicalização*	*Lematização*	*Substantivação*
Milho	510	1.028	358	306	349
Cana	391	11.161	1.561	1.139	1.526
Feijão	348	8.128	809	499	588
Leite	332	634	1487	413	993
Maçã	230	405	909	379	252
Caupi	198	5469	43864	14220	19304
Eucalipto	100	237	482	248	348
Caju	40	65	11	3	7

Tabela 5.4: Pontuação CTW para cada técnica

As pontuações CTW (*Context Term Weight*) obtidas com o uso da técnica de radicalização foi superior para todas bases de textos do que as pontuações obtidas quando utilizadas as técnicas de lematização e substantivação. Ou seja, houve uma quantidade de recuperação de termos no vocabulário expandido maior quando utilizada a técnica de radicalização do que as quantidades das outras duas técnicas. Assim, pode-se perceber que a técnica de radicalização geralmente é mais eficaz na recuperação de termos do vocabulário do domínio, indicando, portanto, que esta gerou uma maior quantidade de termos importantes para este domínio em relação as outras técnicas. Este resultado pode ser explicado, assim como no avaliação 1, pelo fato que a mesma é mais agressiva na simplificação dos termos em relação às outras duas técnicas. Ressalta-se que a escala das pontuações é proporcional ao tamanho do vocabulário expandido.

5.4.4 Avaliação 4 - Representatividade Subjetiva dos Termos Extraídos

Objetivo: verificar junto à especialistas do domínio se o processo de extração de termos utilizando cada uma das três técnicas de simplificação de termos (radicalização, substantivação e lematização) obteve termos representativos / importantes para o domínio em questão.

Hipótese: o processo de extração de termos obtém termos representativos/importantes para o domínio em questão segundo os especialistas do domínio.

Descrição: para cada uma das oito bases de textos aplicou-se separadamente as três técnicas de simplificação de termos, gerando, para cada base, três conjuntos de termos. No total, foram obtidos 24 (vinte e quatro) conjuntos de termos, conforme mostrado no avaliação 1. Por exemplo: para base de textos do Milho foram gerados três conjuntos de termos, sendo o primeiro gerado a partir do uso da técnica de radicalização, o segundo gerado com a lematização e o terceiro com a substantivação.

Para cada conjunto de termos obtido, foi gerada um taxonomia para possibilitar que os especialistas do mesmo domínio da base avaliassem a representatividade destes termos em

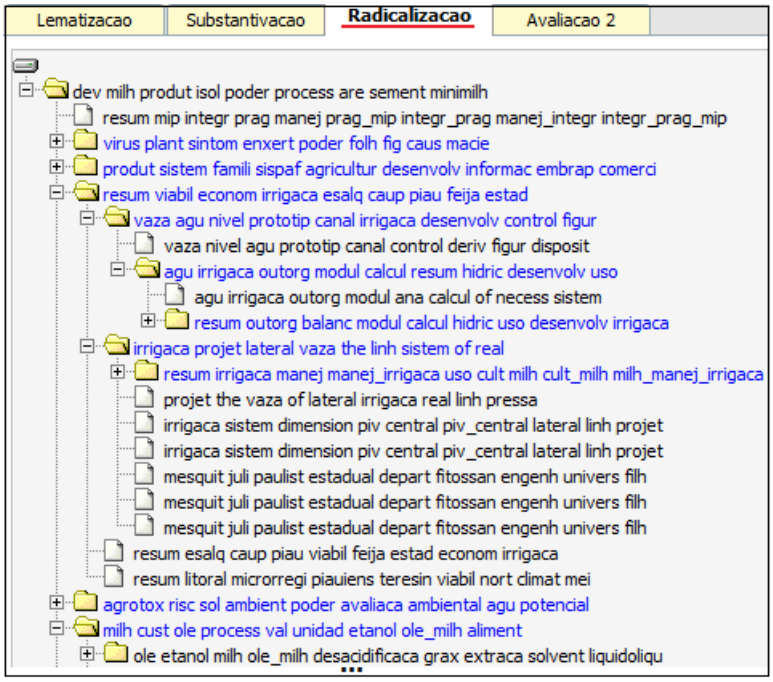


Figura 5.10: Ramos selecionados para avaliação da técnica de radicalização

relação aos documentos da base. Conforme a descrição da Seção 5.2.1, foram selecionados dez ramos de cada taxonomia. Para melhor entendimento, para a base do Milho, nas Figuras 5.10, 5.11 e 5.12 são mostradas as taxonomias geradas e os ramos escolhidos para avaliação das três técnicas (nas figuras, estes ramos são representados pela cor azul).

Onze especialistas do domínio atribuíram à cada ramo das taxonomias uma nota - de um (1) a quatro (4) - em relação a representatividade dos termos. Esta avaliação subjetiva foi auxiliada por meio da ferramenta TaxTool, que possibilidade aos especialistas abrir e ler os documentos que os termos devem representar, permitindo, assim, uma melhor avaliação.

Para analisar as notas dadas pelos especialistas, foi utilizado um modelo linear generalizado (**?**) da seguinte forma:

$$\widehat{nota} = \widehat{\mu} + \widehat{noh} + \widehat{tecnica} + \widehat{avaliador} + \widehat{\epsilon}$$

no qual, $\widehat{\mu}$ é a média geral das notas atribuídas pelos especialistas. Os efeitos do modelo, que correspondem aos desvios em relação à média geral, são representados por $\widehat{noh}$, que corresponde ao nó (ramo) da taxonomia que foi avaliado; $\widehat{tecnica}$, que corresponde a técnica avaliada; e $\widehat{avaliador}$, que corresponde ao avaliador. Por fim, tem-se o erro aleatório ($\widehat{\epsilon}$), sendo que este não é explicado pelos efeitos citados.

As comparações múltiplas de médias foram realizadas com o uso do teste SNK (**?**), considerando 95% de certeza, no qual os grupos de notas foram representados por letras, ou seja, na tabela que segue, letras iguais correspondem a grupos estatisticamente iguais. Na Tabela 5.5 encontram-se os resultados das comparações das notas das três técnicas utilizadas para cada base de textos, bem como os valores do quadrado médio do

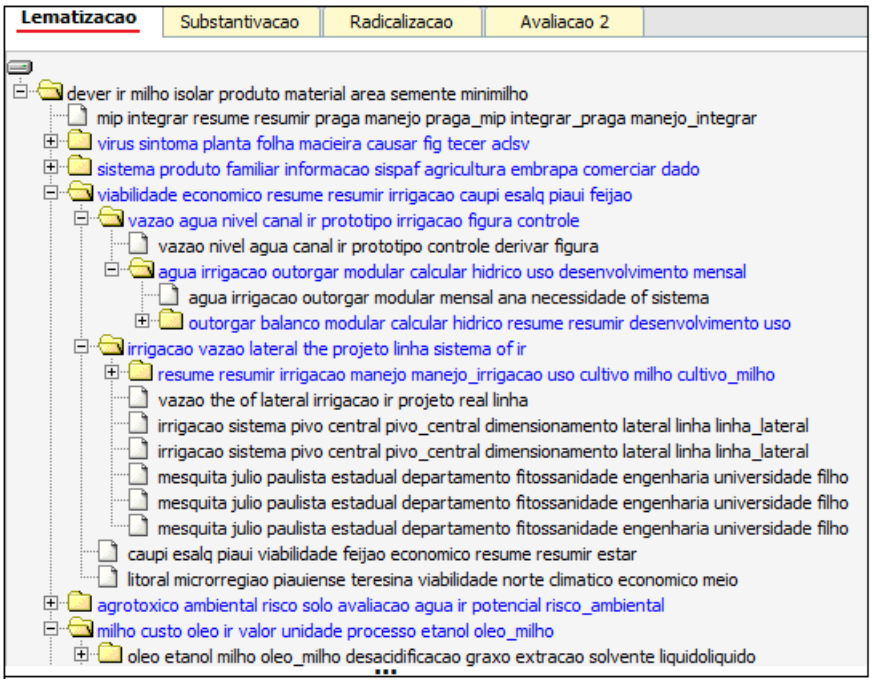


Figura 5.11: Ramos selecionados para avaliação da técnica de lematização

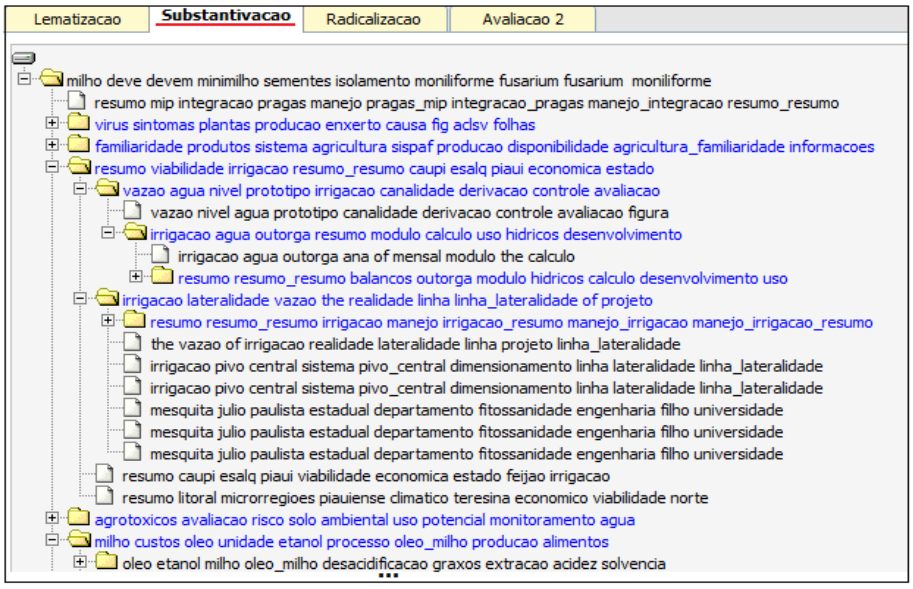


Figura 5.12: Ramos selecionados para avaliação da técnica de substantivação

Bases	*g.l*	*qme*	*Técnicas*	*# notas dadas*	*Ramos*	*Grupos*
Milho	307	0.3875	Radicalização	110	2.754545	milho_b
			Substantivação	110	3.009091	milho_a
			Lematização	109	2.926606	milho_a
Cana	290	0.3634	Radicalização	104	2.355769	cana_b
			Substantivação	104	2.634615	cana_a
			Lematização	104	2.596154	cana_a
Feijão	276	0.3837	Radicalização	98	2.551020	feijao_b
			Substantivação	96	2.979167	feijao_a
			Lematização	104	2.634615	feijao_b
Leite	307	0.3875	Radicalização	110	2.754545	leite_b
			Substantivação	110	3.009091	leite_a
			Lematização	109	2.926606	leite_a
Maçã	291	0.3720	Radicalização	105	2.409524	maca_b
			Substantivação	105	3.085714	maca_a
			Lematização	106	3.009434	maca_a
Caupi	267	0.3296	Radicalização	97	2.381443	caupi_b
			Substantivação	97	2.793814	caupi_a
			Lematização	97	2.752577	caupi_a
Eucalipto	185	0.4381	Radicalização	57	2.543860	eucalipto_b
			Substantivação	57	2.929825	eucalipto_a
			Lematização	99	3.141414	eucalipto_a
Caju	333	0.3837	Radicalização	118	2.669492	caju_c
			Substantivação	118	3.211864	caju_a
			Lematização	120	3.025000	caju_b

Tabela 5.5: Agrupamento das notas nos ramos

erro aleatório (qme) e os graus de liberdade (g.l) de cada base. Para este teste foi utilizado o software livre MODLIN pertencente à Rede de Software Livre para Agropecuária[3] (Software Científico - SOC) da Embrapa.

Análise e Resultados: os resultados das comparações das notas das três técnicas utilizadas para cada base de textos mostraram que, para a base do Caju, houve diferença estatisticamente significativa para as três técnicas, sendo que o uso da técnica de substantivação gerou termos mais representativos para esta base, seguida da técnica de lematização e, por fim, da radicalização. Este fato provavelmente é explicado pelo tamanho da base, pois a base do Caju é a menor de todas, com apenas 40 (quarenta) documentos. Como a quantidade de termos é pequena, tem-se menos chance de serem escolhidos termos que não representam a coleção, dessa forma, os termos extraídos com o uso das três técnicas tendem a ser mais semelhantes (exceto pela forma de simplificação). Isso pode fazer com que os especialistas elejam as técnicas cujos termos estejam mais compreensíveis.

Já para a base do Feijão, o uso da técnica de substantivação obteve termos mais representativos do que as técnicas de lematização e radicalização. Para as demais bases,

[3]Rede de Software Livre para Agropecuária - http://www.agrolivre.gov.br/

houve diferença estatisticamente significativa positiva para as técnicas de substantivação e lematização em relação à radicalização.

Esta avaliação mostrou que, segundo os especialistas, exceto para as bases do Feijão e Caju, obtêm-se termos mais representativos quando utilizada a técnica de lematização ou substantivação.

5.4.5 Avaliação 5 - Compreensibilidade dos Termos Extraídos

Objetivo: avaliar a compreensibilidade dos termos extraídos utilizando as diferentes técnicas de simplificação dos termos, bem como analisar os impactos que as diferentes formas de extração de termos causaram nos especialistas.

Hipóteses:

1. com esta avaliação será possível eleger uma ou mais técnica que melhor se aplica ao domínio em questão, segundo a avaliação subjetiva dos especialistas;

2. a extração de termos utilizando diferentes técnicas, gerará impactos diferentes nos especialistas;

3. a técnica de radicalização, provavelmente, será eleita como a que gera termos menos compreensíveis por ser a mais agressiva em sua aplicação.

Descrição: a compreensibilidade dos termos extraídos foi avaliada subjetivamente pelos especialistas do domínio, sendo feita separadamente para cada base de textos, conforme mostrado anteriormente na Figura 5.2. Para isso, os especialistas elegeram uma ou mais técnicas que consideraram gerar termos mais compreensíveis para a base de textos analisada, isto é, qual (ou quais) técnica obteve termos mais fáceis de serem entendidos pelos especialistas (neste caso, os especialistas são os usuários finais dos termos). Ressalta-se que, os termos foram avaliados não levando em consideração as taxonomias, sendo que estas foram utilizadas somente para facilitar à avaliação.

Análise e Resultados: o resultado desta avaliação é mostrado na Figura 5.13. Pode-se observar que a técnica de substantivação foi eleita como a que gera termos bem mais compreensíveis do domínio de agronegócio se comparada com as outras duas técnicas. Sendo, portanto, a mais indicada para ser utilizada neste domínio quando necessita-se de compreensibilidade nos resultados. Enquanto a técnica de radicalização foi eleita como a que gera termos menos compreensíveis, por ser a mais agressiva em sua aplicação.

Provavelmente o impacto causado nos especialistas em relação à compreensibilidade dos termos, afetou também a avaliação subjetiva dos termos quanto às suas representatividades. Uma possível explicação seria que, como em alguns casos, determinados termos foram obtidos de uma forma mais fácil de serem entendidos (menos simplificados) e outros menos fácil de serem entendidos, mesmo os dois termos estarem se referindo ao mesmo tema, os termos mais compreensíveis podem ser entendidos mais facilmente e, portanto, serem eleitos como mais representativos.

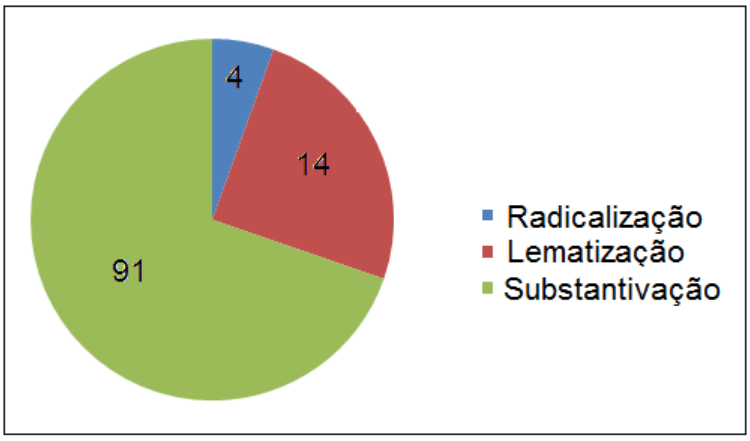


Figura 5.13: Avaliação subjetiva quanto a compreensibilidade dos termos obtidos utilizando as técnicas

5.4.6 Avaliação 6 - Preferência dos Especialistas

Objetivo: apontar, para todas as bases de textos, qual a preferência geral subjetiva dos especialistas em relação às técnicas de simplificação de termos, isto é, qual técnica os especialistas consideraram melhor para ser utilizada neste domínio.

Hipótese: as técnicas de substantivação e lematização serão eleitas pelos especialistas por gerarem, possivelmente, termos mais compreensíveis do que a técnica de radicalização.

Descrição: os especialistas do domínio indicaram suas preferências gerais por alguma das técnicas utilizadas elegendo, para cada base de textos, a técnica que consideraram mais adequada para a base, conforme mostrado anteriormente no item "B" da Figura 5.2.

Análise e Resultados: como todas as bases de textos pertencem a um subdomínio do domínio de agronegócio, estas preferências foram consideradas como preferência geral do domínio, sendo que o resultado pode ser observado na Figura 5.14.

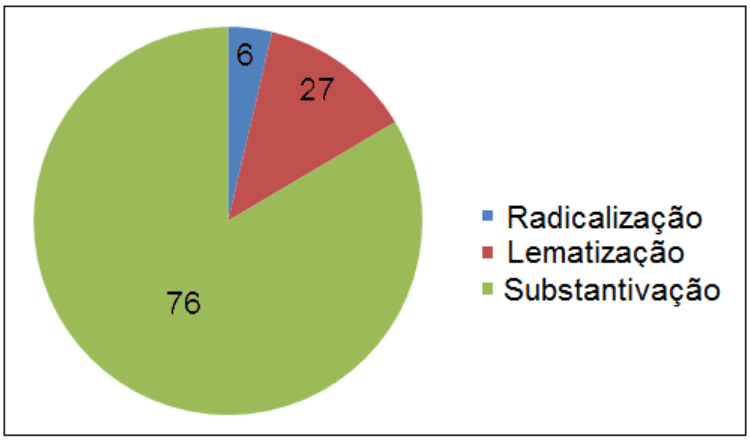


Figura 5.14: Avaliação subjetiva quanto a técnica de preferência dos especialistas

Observa-se que a técnica de substantivação, assim como na avaliação anterior, foi eleita como a técnica que os especialistas preferem para ser utilizada neste domínio, seguida da técnica de lematização e, depois, da radicalização. Tal escolha também é devido a

compreensibilidade dos termos extraídos por esta técnica, pois esta característica atrai o usuário humano.

5.4.7 Avaliação 7 - Complexidades dos Algoritmos das Técnicas de Simplificação de Termos

Objetivo: analisar as complexidades dos algoritmos utilizados para a aplicação das técnicas de simplificação de termos.

Hipótese: com a análise das complexidades de cada técnica é possível escolher qual técnica utilizar para auxiliar a extração de termos.

Descrição: foram analisadas as complexidades, no pior caso, dos algoritmos de radicalização, lematização e substantivação, considerando somente os passos para a simplificação dos termos de cada técnica. A descrição dos algoritmos e de suas complexidades é apresentada a seguir.

Para o algoritmo de radicalização voltado à Língua Portuguesa, **?** criaram uma estrutura de dados para armazenar as palavras do documento e cada palavra desse documento é submetida ao processo de radicalização. Conforme pode ser observado no Algoritmo 5.1, a complexidade do algoritmo de radicalização para cada documento, no pior caso, é $O(|P|)$, sendo $|P|$ a quantidade de palavras do documento.

Algoritmo 5.1 Radicalização

```
Require: documento original
Ensure: documento com palavras radicalizadas
 1: for all palavras do documento do
 2:     if palavra é um verbo irregular then
 3:         retornar o infinitivo do verbo
 4:     else
 5:         if tamanho da palavra é menor ou igual a três then
 6:             retornar a própria palavra
 7:         else
 8:             remover os pronomes oblíquos e o plural da palavra
 9:             remover os sufixos da palavra
10:             if palavra não foi reduzida then
11:                 if palavra é um verbo then
12:                     remover sufixos do verbo
13:                 end if
14:             end if
15:         end if
16:     end if
17: end for
```

O lematizador utilizado neste trabalho faz uso da base de palavras canônicas do Lematizador de Nunes (**?**). Conforme descrito no Algoritmo 5.2, cada palavra presente no documento é substituída por seu respectivo lema buscado na base de palavras. Para isso, foi criada uma estrutura *hash* para armazenar os lemas da base, na qual as buscas são

feitas a partir das palavras originais do documento e são retornados seus respectivos lemas encontrados. Considerando que a complexidade de busca da estrutura *hash* é $O(1)$ devido a sua característica, a complexidade, para cada documento, do algoritmo de lematização desenvolvido neste trabalho é $O(|P|)$, sendo $|P|$ a quantidade de palavras do documento.

Algoritmo 5.2 Lematização

Require: documento original
Ensure: documento com palavras lematizadas

```
1: for all palavras do documento do
2:     if se palavra do documento existe na lista de lemas then
3:         retornar o lema encontrado da palavra
4:     else
5:         retornar a palavra do documento
6:     end if
7: end for
```

Segundo **?**, a complexidade do algoritmo de substantivação para cada documento, descrito no Algoritmo 5.2, é $O(|P|)$, considerando que as árvores utilizadas no algoritmo estão parcialmente balanceadas e que o número de adjetivos, advérbios, verbos ou substantivos é, no máximo, igual a $|P|$, sendo sendo $|P|$ a quantidade de palavras do documento.

Logo, a complexidade dos algoritmos de radicalização, lematização e substantivação, para cada documento, pertence à classe linear $O(|P|)$.

Análise e Resultados: ao contrário da hipótese apresentada nesta avaliação, as complexidades dos algoritmos utilizados para a aplicação de cada técnica não são um fator determinante na escolha de uma das técnicas para auxiliar a extração de termos, já que os três pertencem à classe linear. Porém, deve-se levar em consideração que o algoritmo de substantivação necessita de que as palavras do documento já estejam etiquetadas morfologicamente. Além disso, neste trabalho, após a obtenção dos substantivos por este algoritmo, é necessário escolher qual substantivo será considerado (substantivo abstrato ou concreto), conforme explicado anteriormente na Seção 4.2.2 do Capítulo 4. Estes detalhes podem aumentar o custo computacional da aplicação da técnica de substantivação em relação as outras duas técnicas.

Algoritmo 5.3 Substantivação (?)

Require: documento original
Ensure: documento com palavras substantivadas

```
1: for all p ϵ P = palavras do documento do
2:     verificar etiqueta morfológica de p
3:     if p é advérbio then
4:         transformar em adjetivo correspondente
5:     end if
6:     if p é adjetivo ou verbo then
7:         pesquisar em autômato de exceções implementado em árvore ternária de
           pesquisa com N_E nodos
8:         if pesquisa tem sucesso then
9:             derivar substantivos de p
10:        else
11:            pesquisar p em autômato de adjetivos ou de verbos implementado em
               árvore ternária de pesquisa com N_A (para adjetivos) ou N_V (para verbos)
               nodos
12:            if pesquisa tem sucesso then
13:                derivar substantivos de p
14:            else
15:                não há nominalização (substantivação) para p
16:            end if
17:        end if
18:    else
19:        if p é substantivo then
20:            pesquisar em autômato de sinônimos implementado em árvore ternária
               de pesquisa com N_S nodos
21:            if pesquisa tem sucesso then
22:                derivar sinônimo de p
23:            else
24:                não há sinônimo de p
25:            end if
26:        end if
27:    end if
28: end for
```

5.5 Considerações Finais

Neste capítulo foram descritas as avaliações subjetivas e objetivas adotadas para avaliar os termos extraídos deste trabalho. As avaliações subjetivas tendem a ser custosas, pois demandam mais tempo para sua execução, e além de necessitar da disponibilidade de especialistas do domínio pode ser influenciada por preferências pessoais dos mesmos. Mesmo com estas dificuldades, a presença dos especialistas é vantajosa, já que os mesmos, neste caso, são os usuários finais dos termos. Adicionalmente, estas avaliações subjetivas são necessárias para avaliar a compreensibilidade dos termos obtidos.

Neste capítulo também foi realizada a avaliação experimental seguindo os passos da metodologia de extração de termos. Com isso foi possível observar que, conforme mostrado

na Tabela 5.3, o uso da metodologia contribui consideravelmente para a diminuição da quantidade de termos extraídos, pois além de simplificar os termos com três diferentes técnicas, faz uso de métodos estatísticos para remover as palavras que não são bons candidatos a termos.

Mesmo assim, foram mantidos alguns termos que deveriam ter sido excluídos, como termos que possuem algum verbo que não contribui para a obtenção de termos representativos da coleção (como *manejo_integrar*), ou ainda bigramas e trigramas compostos por palavra repetidas (como *resumo_resumo*). Estes termos podem ser vistos, respectivamente, nas Figuras 5.11 e 5.12 deste capítulo. Uma possível solução para este problema, seria incrementar a metodologia com técnicas lingüísticas capazes de identificar se a presença de determinados verbos é relevante ou não para o termo.

Também foram descritos experimento e avaliações desses experimentos, juntamente com os resultados e análises dos mesmos, sobre a representatividade dos termos extraídos sob a visão subjetiva dos especialistas e sob uma medida objetiva, a CTW (*Context Term Weight*). Bem como, segundo os especialistas, sobre a compreensibilidade dos termos extraídos com o uso de cada técnica e a que os especialistas sugerem para ser utilizada no domínio de agronegócio.

No capítulo a seguir, são apresentadas as conclusões e as principais contribuições alcançadas com o desenvolvimento deste trabalho, bem como sugestões de trabalhos futuros.

Capítulo

6

Conclusões e Trabalhos Futuros

Neste trabalho foi apresentada uma metodologia para apoiar a extração de termos utilizando três diferentes técnicas de simplificação de termos, a radicalização, a lematização e a substantivação. Para apoiar tal extração foi desenvolvida uma ferramenta, a ExtraT. Desses termos depende, em grande parte, a "qualidade" dos resultados do processo de Mineração de Textos. Sendo assim, os mesmos devem ser avaliados e, caso necessário, o processo de extração de termos deve ser refeito. Esses termos foram avaliados objetivamente com o auxílio de uma ferramenta desenvolvida, a TaxEM; e subjetivamente por especialistas do domínios.

A avaliação subjetiva de termos tem como vantagem o auxílio de especialistas do domínio, o que permite uma melhor avaliação dos termos extraídos. Mas, por outro lado, a presença dos especialistas demanda mais tempo para a aplicação da avaliação e um esforço manual dos mesmos.

Neste trabalho, uma avaliação experimental foi realizada utilizando oito coleções textuais do domínio de agronegócio. É importante ressaltar que estas coleções de textos são reais, o que exige um esforço adicional em relação ao seu tratamento. A partir de avaliações em relação ao experimento realizado, pode-se observar, conforme o esperado, que seguindo os passos da metodologia para extrair os termos, pode-se diminuir consideravelmente a quantidade de termos trabalhados quando comparados os termos extraídos seguindo esta metodologia e os extraídos sem seguir os passos da mesma. Neste sentido, a metodologia, aqui apresentada, contribui para melhorar um dos problemas de se trabalhar com grandes quantidades de termos na Mineração de Textos.

Além disso, para estas bases de textos, pode-se observar que o uso da técnica de radicalização geralmente obtém uma quantidade de termos inferior do que quando utilizada a lematização. Por último, o uso da substantivação obtém mais termos do que as duas outras técnicas. Isso pode ser explicado pelo fato da técnica de radicalização ser mais agressiva para simplificar os termos em relação às técnicas de lematização e substantiva-

ção. Utilizar uma técnica que seja capaz de gerar uma menor quantidade de termos ajuda a minimizar o problema da alta dimensionalidade da Mineração de Textos. Já que quando se tem um número menor de termos, o espaço de armazenamento dos dados trabalhados exigido também é menor.

Mas quando se compara as complexidades dos algoritmos utilizados para as três técnicas, percebe-se todas as complexidades lineares, indicando que as complexidades não são um fator determinante para a escolha de qual técnica utilizar.

Os termos extraídos foram avaliados também quanto a representatividade dos mesmos em relação às coleções de textos e, para isso, utilizou-se como suporte a medida CTW (*Context Term Weight*) e um vocabulário expandido do domínio. Como resultado, a técnica de radicalização mostrou ser mais eficaz na recuperação de termos do vocabulário do domínio, indicando, portanto, que esta gerou uma maior quantidade de termos importantes para este domínio em relação às outras técnicas. Este resultado também pode ser explicado pelo fato que a mesma é mais agressiva na simplificação dos termos em relação às outras duas técnicas. Isso significa que a técnica de radicalização reduz mais o número de palavras distintas representadas por um só termo, o que pode ser útil para tarefas de recuperação de informação.

Já quando a representatividade dos termos na coleção foi analisada subjetivamente por especialistas do domínio, para todas as bases, exceto para a base de Caju e Feijão, obtêm-se termos mais representativos quando utilizada a técnica de lematização ou substantivação. A diferença de resultados quanto à representatividade dos termos quando avaliado objetiva e subjetivamente provavelmente é devido ao impacto causado nos especialistas em relação à compreensibilidade dos termos obtidos por cada técnica. Sendo que as técnicas de substantivação e lematização simplificam os termos de uma forma menos agressiva do que a técnica de radicalização. Por isso, acredita-se que os especialistas elegeram a técnica de substantivação como a que obtém termos bem mais compreensíveis do domínio se comparada com as outras duas técnicas, seguida da lematização e, por último, da radicalização.

A melhor compreensibilidade obtida com o uso da substantivação faz com que essa técnica seja a mais indicada para ser utilizada neste domínio quando a compreensibilidade nos resultados é necessária. Ainda provavelmente por este motivo, a substantivação foi eleita como a técnica que os especialistas preferem para ser utilizada neste domínio, seguida da técnica de lematização e, depois, da radicalização. Deve-se ressaltar que ao contrário da lematização e radicalização, a técnica de substantivação aqui aplicada necessita primeiramente etiquetar morfologicamente as palavras do documento, podendo carregar os erros cometidos neste processo.

Já quando a compreensibilidade não é um fator determinante nos resultados, a radicalização é mais indicada para ser utilizada em coleções de textos do domínio de agronegócio. A lematização, por sua vez, segundo os especialistas, obtém termos menos compreensíveis do que a substantivação, porém mais compreensíveis do que a radicalização.

Na Tabela 6.1, tem-se um resumo de quando utilizar cada técnica de simplificação de termos para o domínio de agronegócio, mas deve-se levar em consideração os pontos abordados nas avaliações do uso dessas técnicas. Com este trabalho, pode-se observar que a escolha pelo uso de uma das técnicas depende do objetivo pré-estabelecido.

Técnicas	***Radicalização***	***Lematização***	***Substantivação***
Representatividade objetiva	•		
Representatividade subjetiva		•	•
Compreensibilidade			•
Preferência			•
Número de termos	•		

Tabela 6.1: O uso das técnicas de simplificação de termos

Com este trabalho também foi possível obter contribuições importantes para a área de pesquisa em questão. Além disso, com este trabalho viabiliza-se a escolha de qual técnica de simplificação de termos pode ser utilizada na construção de taxonomias de tópicos para um domínio específico de acordo com o objetivo pré-estabelecido, a TopTax (*Topic Taxonomy Environment*), detalhada na Seção 2.4 do Capítulo 2. Os termos aqui extraídos são utilizados na etapa de Pré-Processamento da TopTax, visando somente fornecer ao usuário final da mesma os termos que realmente representam o domínio em questão e possibilitar a escolha de qual técnica de simplificação é mais aconselhável para o objetivo.

Deve-se ressaltar que esses termos são mostrados como uma matriz atributo-valor e é gerada, por meio da ferramenta PreText, uma lista completa em formato de texto plano de todos os termos de cada técnica de simplificação utilizada. Devido a este formato, é possível que o usuário final visualize uma lista dos termos obtidos, além disso, caso necessário, pode-se aplicar aos termos obtidos algum método de seleção de atributos, como o método de Luhn, Salton e *Term Variance*.

Por fim, como resultado do mestrado em termos de publicação tem-se: (i) artigos publicados em eventos da área de Inteligência Artificial, que são (**?**), (**?**), (**?**), (**?**), (**?**), (**?**) e (**?**); e (ii) relatórios técnicos que descrevem a abordagem completa para a construção de taxonomias de tópicos em um domínio (**?**) e o uso de diferentes formas de extração de termos a partir de coleções textuais (**?**).

Como trabalhos futuros, visando contribuir com a metodologia de extração de termos proposta aqui, os verbos contidos nos documentos podem ser melhor tratados, para que os termos extraídos contenham somente os verbos necessários para a composição dos termos. Além disso, pode-se tratar as palavras com erros gramaticais, pois uma mesma palavra que contenha algum erro gramatical pode se tornar termos distintos.

Adicionalmente, podem ser incorporados à metodologia de extração de termos métodos lingüísticos para refinar mais ainda os termos obtidos, sempre visando manter o balanço entre a "qualidade" dos resultados e o custo de processamento da aplicação. Considerando que os métodos lingüísticos incorporam tanto mais custo de processamento como aumento da "qualidade" nos resultados.