UNIVERSIDADE DO VALE DO RIO DOS SINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM
COMPUTAÇÃO APLICADA – PIPCA

Uso de Informações Lingüísticas na etapa de pré-processamento em Mineração de Textos

por

CASSIANA FAGUNDES DA SILVA

Dissertação de mestrado submetida a avaliação
como requisito parcial para obtenção do grau de
Mestre em Computação Aplicada

Profª. Drª. Renata Vieira
Orientadora

Profº. Drº. Fernando Santos Osório
Co-orientador

São Leopoldo, fevereiro de 2004.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Silva, Cassiana Fagundes

Uso de Informações Lingüísticas na etapa de pré processamento em Mineração de Textos/ por Cassiana Fagundes da Silva. – São Leopoldo: Centro de Ciências Exatas e Tecnológicas da UNISINOS, 2004.

109 f.: il.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos. Ciências Exatas e Tecnológicas Programa Interdisciplinar de Pós-Graduação em Computação Aplicada – PIPCA, São Leopoldo, BR – RS, 2004. Orientador: Vieira, Renata.

I. Vieira, Renata. II. Título.

Mãe, Pai e Pepe,
dedico este trabalho a vocês.

Agradecimentos

Em especial a profª. Renata Vieira, minha adorada orientadora, pela amizade, dedicação, atenção, compreensão, carinho e palavras de consolo nos momentos difíceis. A você meu agradecimento especial.

Ao meu co-orientador profº. Fernando Santos Osório. Obrigada pela atenção, paciência, orientação e amizade.

Ao profº. Adelmo Cechin, pela amizade, sinceridade e atenção. Meu muitíssimo obrigado.

À minha mãe, Sirlei, por ter confiado e por ter apostado na realização de mais um objetivo em minha vida. Devo parte dessa conquista a ti. Obrigada por todo apoio, amor, compreensão, preocupação (ás vezes exagerada, ☺), e dedicação. Ao meu pai, Nelson, por sua preocupação, dedicação e apoio financeiro durante esse tempo fora de casa.

Ao Pedro meu eterno namorado, pelo amor, carinho, dedicação, incentivo, paciência, amizade, e apoio. Sabes que tenho um carinho muito especial por ti e que parte deste mestrado eu devo a você.

Aos meus queridos colegas do laboratório de Engenharia da Linguagem, Cássia dos Santos, Cláudia Pérez, Sandra Collovini, César Coelho, Douglas M. da Silva, Rodrigo Goulart, Fábio Okuyama, Sandro Rigo. Pela amizade, pelos momentos de descontração, pela paciência, pelas trocas de conhecimento, pelos lanches no meio da tarde, e convívio diário. Muito obrigado a todos vocês, com certeza ficarão guardados no meu coração.

Ao corpo docente do PIPCA, pela interação e disposição.

Aos administradores de rede do PIPCA, Fábio Luciano, Farlei, Souto pela disposição, ajuda e amizade.

À UNISINOS que tornou possível a realização dessa pesquisa, através do auxílio financeiro.

À CAPES pela viagem a Lisboa, a qual possibilitou o aprofundamento do meu objeto de estudo.

A Deus, que me guiou e permitiu que eu pudesse atingir meus objetivos.

Resumo

Este trabalho apresenta estudos, com realização de experimentos e análise de resultados, da aplicação de informações lingüísticas na etapa de pré-processamento no processo de Mineração de Textos para as tarefas de Categorização e Agrupamento de Documentos.

Usualmente, o pré-processamento utilizado no processo de Mineração de Textos para estas tarefas consiste na remoção de termos irrelevantes (tais como, preposição, artigos, pronomes, entre outros), normalização morfológica e seleção dos termos (ao que denominamos baseado em métodos usuais). Propõe-se, ao longo deste trabalho, um pré-processamento que faz o uso de informações lingüísticas, ou seja, um pré-processamento baseado em combinações gramaticais, visando avaliar a repercussão do uso dessas informações nos resultados de tarefas de Mineração de Textos.

Foram realizados diversos experimentos para a validação da abordagem proposta junto à língua portuguesa. O corpus utilizado nos experimentos consiste de um extrato do corpus NILC (Núcleo Interinstitucional de Lingüística Computacional), formado por textos jornalísticos do ano de 1994 das seções: Esporte, Imóveis, Informática, Política e Turismo, escritos em português do Brasil.

Experimentos com as combinações gramaticais: substantivos; substantivos-adjetivos; substantivos-nomespróprios; substantivos-nomespróprios-adjetivos; e finalmente nomespróprios-adjetivos são descritos. A análise dos resultados é detalhada, apresentando comparações entre os resultados obtidos a partir do pré-processamento usual e os resultados obtidos a partir da seleção por combinações gramaticais aqui proposta.

Com o resultado dos experimentos, pode-se verificar que o uso de informações lingüísticas na etapa de pré-processamento apresentou melhorias em ambas tarefas de categorização e agrupamento de textos. Para os experimentos de categorização a menor taxa de erro (18,01%) foi obtida através da seleção de substantivos-nomespróprios para o aprendizado simbólico. Para os experimentos de agrupamento o uso de informações lingüísticas possibilitou a identificação de um maior numero de grupos.

Abstract

This work presents studies through experiments and analysis of their results of the application of linguistic information to the pre-processing phase in Text Mining tasks such as categorization and clustering.

Usually, the pre-processing phase is based on the removal of irrelevant terms (such as prepositions, articles, pronouns), stemming and terms selection. We propose the use of linguistic information as an alternative for this phase, that is pre-processing based on grammatical combinations aiming at the evaluation of the impact of this information in the results of text mining tasks.

Many experiments were undertaken to verify the adequacy of the approach concerning the Portuguese language. The corpus is part of the NILC Corpus, and is composed by newspaper articles from the year of 1994. The sections are Sports, Informatics, Politics, Property and Tourism, written in Brazilian Portuguese.

Experiments are described in which the term selection is based on nouns; nouns-adjectives; nouns-proper names; nouns-proper names-adjectives; and finally, proper names-adjectives. The results analysis is detailed: we present the comparison of the results obtained with the usual pre-processing based on words and the results obtained when using our proposal of linguistic information.

The results show that the use of linguistic information in the pre-processing phase has brought some improvements both in categorization and in the clustering tasks In the categorization experiments we obtained the lowest error rate for PD2 when the pre-processing phase was based on the selection of nouns and adjectives (18,01%) for the symbolic learning. For the clustering experiments the use of linguistic information allowed the identification of a larger number of groups.

Sumário

Lista de Tabelas

Lista de Figuras

Abreviaturas

AD	Árvore de Decisão
AM	Aprendizado de Máquina
EI	Extração de Informações
FT	Freqüência do Termo
FD	Freqüência do Documento
IA	Inteligência Artificial
KDD	Knowledge Discovery from Databases
KDT	Knowledge Discovery from Text
MD	Mineração de Dados
MLP	Multi-Layer Perceptron
MT	Mineração de Textos
PLN	Processamento de Linguagem Natural
RI	Recuperação de Informações
RNA	Redes Neurais Artificiais
SRI	Sistemas de Recuperação de Informação
SVM	Support Vector Machines
TF-IDF	Freqüência do Termo - Freqüência Inversa de Documentos
VSM	Modelo de Espaço Vetorial

1 Introdução

Atualmente, o aumento considerável de bases de dados textuais tem motivado o desenvolvimento de técnicas e ferramentas para Recuperação de Informação (RI), as quais visam buscar meios de permitir um acesso rápido e fácil às informações relevantes aos usuários [Sparck et al., 1997].

Contudo, mesmo com a eficiência na recuperação das informações (documentos), ainda é necessário examinar os documentos relevantes para encontrar a informação desejada. A grande dificuldade vem do fato de que documentos são geralmente insatisfatórios como respostas, por serem grandes e difusos em geral [Wilkinson, 1994]. Além disso, ferramentas de RI costumam produzir como resposta uma quantidade muito grande de documentos, ocasionando a chamada "sobrecarga de informações" (*information overload)*, que ocorre quando o usuário tem muita informação ao seu alcance, mas não tem condições de tratá-la ou de encontrar o que realmente deseja ou lhe interessa [Chen,1994].

Como conseqüência da evolução da área de Recuperação de Informação surgiu a área de Mineração de Textos (MT) ou *Text Mining* [Tan, 1999]. Termos como Descoberta de Conhecimento de Textos (*Knowledge Discovery from Text-KDT*), Mineração de Dados Textuais [Hearst, 1999] também são utilizados para o mesmo fim. Neste trabalho será adotado o termo Mineração de Textos.

A Mineração de Textos (MT) vem sendo amplamente adotada no campo de Inteligência Artificial, envolvendo áreas como Recuperação de Informação (RI), Extração de Informação (EI), Agrupamento, Categorização, Aprendizado de Máquina (AD) e Mineração de Dados (MD), entre outras. Segundo Tan (1999), esta nova área é definida como o processo de extrair padrões ou conhecimentos, interessantes e não-triviais, a partir de um conjunto de documentos textuais.

Nesse contexto, esta dissertação de mestrado apresenta um estudo na área de Mineração de Textos e propõe a aplicação de diferentes técnicas para as etapas de pré-processamento no processo de MT. O estudo utiliza uma coleção de documentos da língua portuguesa do Brasil, nas tarefas de Categorização e Agrupamento de Documentos. Usualmente, o pré-processamento utilizado no processo de MT é baseado

em métodos usuais (por meio de técnicas como remoção de termos irrelevantes, normalização morfológica e seleção dos termos). Alternativamente, propomos um pré-processamento que faz o uso de informações lingüísticas, visando avaliar a repercussão do uso dessas informações nos resultados de tarefas de Mineração de Textos.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral desta dissertação é verificar a adequação de uma abordagem baseada no uso de informações lingüísticas na etapa de pré-processamento para descobrir conhecimento em textos e testar a hipótese de que esse tipo de abordagem tem vantagens sobre abordagens baseadas em métodos usuais.

1.1.2 Objetivos Específicos

- Revisar as etapas de pré-processamento baseado em métodos usuais nas tarefas de categorização e agrupamento de textos no processo de MT aplicadas neste trabalho;
- Avaliar o uso de informações lingüísticas como alternativa no pré-processamento para as tarefas de categorização e agrupamento;
- Definir um processo de identificação de informações lingüísticas para seleção de termos relevantes;
- Aplicar a MT em uma coleção de textos da língua portuguesa do Brasil;
- Avaliar o grau de acerto no processo de MT baseado no pré-processamento baseado em métodos usuais em comparação ao uso de informações lingüísticas;
- Analisar resultados obtidos utilizando as técnicas de aprendizado simbólico (Árvores de Decisão) e conexionistas (Redes Neurais Artificiais) para o processo de categorização de textos e o algoritmo *K-means* para o agrupamento de textos, considerando as abordagens propostas.

1.2 Organização do Texto

Esta dissertação está organizada em 6 capítulos, sendo o primeiro a presente introdução.

O capítulo 2 descreve o processo de Mineração de Textos (MT). São apresentadas diversas características relacionadas a esse processo, principalmente na etapa de pré-processamento dos textos, bem como as tarefas de MT adotadas nessa dissertação.

O capítulo 3 apresenta uma nova abordagem baseada no uso de informações lingüísticas na etapa de pré-processamento no processo de MT. Abordando de forma detalhada a metodologia aplicada neste pré-processamento, bem como os recursos utilizados para a extração de categorias gramaticais.

No capítulo 4, é descrito o *corpus* (coleção de documentos) e os métodos utilizados para prepará-los para a realização dos experimentos com cada uma das tarefas de MT: categorização e agrupamento, bem como a metodologia utilizada na concepção, treinamento e avaliação dos classificadores e grupos para a coleção.

No capítulo 5, são relatados e discutidos os resultados dos experimentos realizados, objetivando de comparar o desempenho das tarefas de categorização e agrupamento utilizando a abordagem baseada em informações lingüísticas em comparação a abordagem baseada em palavras.

As conclusões deste trabalho, incluindo menção às etapas que lhe poderão dar continuidade, são discutidas no capítulo 6.

Finalmente, são listadas as referências bibliográficas.

2 Mineração de Textos

A Mineração de Textos (MT), ou Mineração de Dados Textuais [Hearst, 1999], é definida como o processo de extrair padrões ou conhecimentos, interessantes e não-triviais, a partir de um conjunto de documentos textuais [Tan, 1999].

O processo de extrair padrões em MT se preocupa em buscar informações dos textos e tratá-las de forma a apresentar ao usuário algum tipo de conhecimento, eficientemente, útil e novo. Para que se possa descobrir esse conhecimento útil, e que esse apresente um bom desempenho no processo de MT, o pré-processamento dos textos torna-se uma etapa fundamental.

Neste capítulo é apresentada uma introdução ao processo de MT, bem como as principais etapas e tarefas relacionadas ao processo. É dada ênfase à etapa de pré-processamento e as tarefas de categorização e agrupamento de bases textuais, uma vez que esse é o objetivo de estudo desta dissertação.

2.1 Etapas do Processo de Mineração de Textos (MT)

A MT, é relativamente recente em comparação a outras áreas, porém, atualmente, vem sendo amplamente adotada no campo de Inteligência Artificial (IA), envolvendo áreas como Recuperação de Informação (RI), Análise de Textos, Extração de Informação (EI), Agrupamento, Categorização, Sumarização, Aprendizado de Máquina (AD) e Mineração de Dados (MD).

A partir da análise de publicações correlacionadas à MT é possível reconhecer uma metodologia para esse processo. Tal procedimento técnico assemelha-se ao processo de MD, que atua em dados estruturados, enquanto que a MT trabalha com dados não estruturados, normalmente na forma de textos ou documentos, havendo, portanto um tratamento diferenciado.

Deste modo, independente dos vários fatores que acarretam na complexidade das tarefas de MT (como, por exemplo, linguagem, estilo ou conteúdo do documento escrito), algumas etapas são indispensáveis à maioria das tarefas de MT (dentre elas,

Categorização, Agrupamento, Sumarização, entre outras). Estas etapas são ilustradas na Figura 2-1.

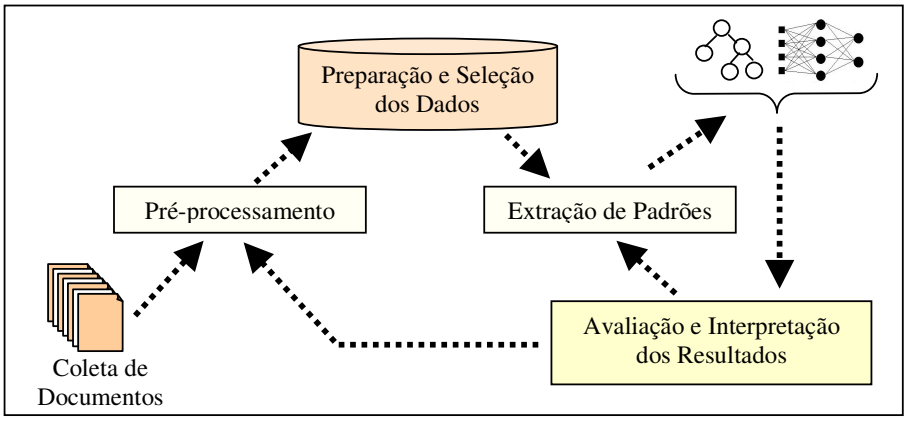

Figura 2-1 Etapas do Processo de Mineração de Textos

A primeira etapa do processo é coleta de documentos. Esta consiste na busca de documentos relevantes ao domínio de aplicação do conhecimento a ser extraído. Tais documentos estão disponíveis virtualmente na Internet ou fisicamente nos livros. Com a finalidade de auxiliar a investigação na Internet, são desenvolvidos diversos sites de busca, entre eles os mais populares estão Altavista[1], Google[2], Yahoo[3].

No entanto, o sucesso dessa atividade depende, em parte, da cooperação de um especialista. Este não só fornece conhecimento sobre o domínio, como também apóia a tarefa de encontrar os objetivos almejados [Fayyad, 1996].

Após a coleta de documentos é necessário formatar os documentos selecionados, pois eles serão submetidos aos algoritmos de extração automática de conhecimento. Essa segunda etapa denomina-se pré-processamento. Ela é responsável por exibir uma representação estruturada dos documentos, freqüentemente no formato de uma tabela atributo-valor.

A tabela atributo-valor caracteriza-se pela alta dimensionalidade, uma vez que cada termo do documento poder representar um possível elemento do conjunto de atributos da tabela. Logo, é imprescindível a seleção dos dados, a fim de reduzir a extensão da tabela atributo-valor, de modo que se alcance dados efetivamente expressivos. Mas isso será tratado de forma mais minuciosa na terceira etapa: preparação e seleção dos dados.

[1] http://www.altavista.com.br/
[2] http://www.google.com.br/
[3] http://www.yahoo.com/

Na etapa de extração de padrões podem ser utilizados sistemas de aprendizado de máquina com a finalidade de encontrar conhecimentos, tendências, similaridades úteis e desconhecidas nos documentos.

A etapa de avaliação e interpretação dos resultados é a última etapa do processo de MT. Geralmente é realizada com o auxílio do usuário e/ou especialista do domínio. Nessa etapa, bem como no processo de MD, os padrões encontrados são analisados e interpretados, com a finalidade de constatar se o objetivo da aplicação foi alcançado. Em caso negativo, é iniciada uma investigação de quais etapas do processo ocasionam no resultado não satisfatório. Através desta investigação é possível a adequação de algumas etapas, provocando uma melhora significativa dos resultados.

A seguir é apresentado em maiores detalhes o pré-processamento de textos, preparação e seleção dos dados, bem como as tarefas de MT (categorização e agrupamento), focos de estudo desta dissertação.

2.1.1 Pré-processamento

Considerando que a etapa de coleta de documentos tenha sido cumprida, e conseqüentemente, os documentos encontram-se disponíveis, é necessário aplicar o pré-processamento. A etapa de pré-processamento demanda a maior parte do tempo do processo de extração de conhecimento. Além disso, ela exige planejamento e processamento. Sem esses requisitos não se obtém uma performance na MT.

A etapa de pré-processamento consiste nas fases: análise léxica, eliminação de termos considerados irrelevantes, ou *stopwords* [Korfhage, 1997; Kowalski, 1997; Salton, 1983], bem como a normalização morfológica dos termos (remoção de prefixos e sufixos).

Na análise léxica é feita uma adaptação do documento. Essencialmente, eliminam-se os dígitos e os sinais de pontuações e isolam-se os termos e efetua-se a conversão das letras de maiúsculas para minúscula.

Além disso, na MT, quando se examinam documentos textuais, os conectores[4] textuais e lexicais – conjunções e preposição – são considerados entidades com menos relevância, por isso são removidas. Pelo mesmo motivo – a relevância – também se

[4] Nos textos existem duas categorias que são responsáveis pela articulação entre palavras, frase e parágrafos: as conjunções e as preposições.

extrai os artigos. Tanto estes, quanto aqueles, são denominados termos irrelevantes. Segundo Salton (1983), uma das grandes vantagens da eliminação dos irrelevantes é a redução de 40 a 50% da extensão documentos.

Alguns estudos disponibilizam listas de termos irrelevantes (denominadas *stoplist* ou *dicionários negativos*) que podem ser utilizadas na elaboração de ferramentas que aplicam este processo. Existem atualmente várias listas de termos irrelevantes. Para a língua inglesa a mais utilizada é a recomendada pelo Laboratório de Recuperação de Informações da Universidade de Massachusetts em Amherst, relatada no trabalho de Lewis (1991). Segundo o autor, ela é composta por 266 termos. Para a língua portuguesa pode ser adotada a lista disponibilizada no trabalho de Neves (2001).

Outra etapa realizada no pré-processamento é a eliminação da variação morfológica de um termo, também denominada de *lematização* ou algoritmo de *stemming* [Frakes, 1992; Kraaij, 1996]. Entretanto, esta variação depende da tarefa de MT adotada, na categorização de documentos, por exemplo, a variação morfológica pode ser extremamente importante, pois aumenta a discriminação entre documentos [Riloff, 1995], devido à redução de variantes de um mesmo radical para um mesmo conceito.

A aplicação de um algoritmo de eliminação da variação morfológica *(stemming)*, visa remover as vogais temáticas, as desinências, os prefixos e os sufixos de um termo. Por exemplo, quando o algoritmo encontra os termos: **conectar**, **conectado**, **conectando;** ele os reduz a um único termo: o radical **conect.** Tal procedimento não só diminui o número de entradas, como também aumenta a ocorrência do radical nos documentos.

Os algoritmos de eliminação da variação morfológica mais conhecidos são os algoritmos de Porter (1980), Lovins (1968) e o de Paice (1994) que removem sufixos de termos da língua inglesa. O algoritmo de Porter tem sido usado e referenciado nos últimos anos, várias implementações do algoritmo estão disponíveis na Web, assim como a página oficial escrita e mantida pelo autor[5]. O algoritmo de Lovins consiste em um único passo, desenvolvido a partir de um conjunto exemplo de termos, removendo no máximo um sufixo por termo, retirando o sufixo mais longo conectado ao termo. Em Paice (1994) há uma comparação entre os algoritmos destes dois autores com o de

[5] Disponível em http://www.tartarus.org/Martin/PorterStemmer

Porter. Para a língua portuguesa, dispomos dos algoritmos de Porter[6] e Viviane Orengo [Orengo, 2001]. Apesar destes algoritmos serem aplicados com o intuito de melhorar o desempenho de sistemas de RI e de MT, alguns estudos mostraram que a redução de afixos em uma coleção de documentos não apresenta uma melhora significativa no desempenho da tarefa de mineração aplicada [Harman, 1991; Lennon et. al, 1981] e sim uma redução no espaço de representação dos dados.

Para esta dissertação o algoritmo de eliminação da variação morfológica de Porter para a língua portuguesa apresenta alguns problemas, como: a não retirada da acentuação do termo, impossibilitando que termos de mesmo conceito sejam reduzidos a um único radical. Como exemplo do problema considere os termos, **representação** e **representações**, onde o algoritmo reduz somente as vogais ***o*** e ***es*** dos termos, resultando no radical **representaçã** e **representaçõ** respectivamente. Como o objetivo deste estudo não é avaliar o desempenho de um algoritmo de variação morfológica, esse problema foi ignorado.

Diferentemente das etapas apresentadas nessa seção, para o pré-processamento proposto nesta dissertação, foram utilizadas informações lingüísticas mais elaboradas. A relevância de tal uso de informação lingüística será analisada no capitulo 3.

Na seção que segue, é apresentada em maiores detalhes a etapa de preparação e seleção dos dados.

2.1.2 Preparação e Seleção dos Dados

A etapa de preparação e seleção dos dados consiste em identificar e filtrar nos dados pré-processados os termos mais representativos para as tarefas de MT. Considerando que em uma coleção de documentos extensos, ou com um grande número de documentos, o tamanho do vocabulário poderá ser intratável pelas técnicas de MT. Dessa forma, faz-se necessária à busca de termos relevantes que possam melhor discriminar os documentos.

Várias etapas são realizadas a fim de reduzir a quantidade de termos, visando uma melhor representatividade dos documentos no desempenho do sistema. As etapas necessárias para a preparação dos dados são ilustradas na Figura 2-2 e são mais detalhadas nas subseções a seguir.

[6] Disponível para várias línguas http://snowball.sourceforge.net/

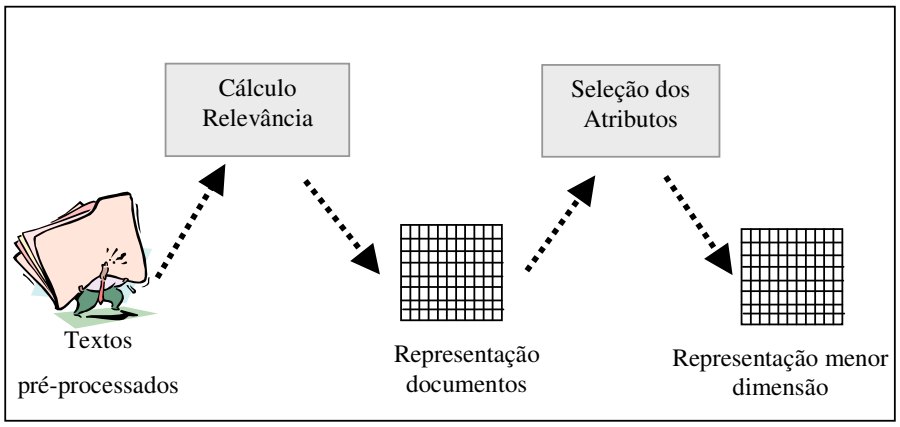

Figura 2-2 Etapas da Preparação e Seleção dos Dados

2.1.2.1 Cálculo de Relevância

Conforme descrito anteriormente, muitos termos não são suficientemente significativos para descrever o assunto de um texto [Meadow et al., 2000]. Logo, utilizar uma medida para calcular a relevância dos termos nos documentos é de grande valia.

Os métodos mais usuais de cálculo de relevância são baseados na freqüência dos termos [Rijsbergen, 1979; Salton, 1987], tais como: freqüência absoluta, freqüência relativa, freqüência do termo - freqüência inversa de documentos.

Cálculos como, ganho de informação [Yang and Pederson, 1997], coeficiente de correlação [Ng et al., 1997], a técnica do qui-quadrado [Ng et al., 1997], são baseados na teoria da informação, e, nesse caso, quanto maior a probabilidade de um termo ser representativo ao documento, maior é o seu peso. Existem ainda métodos que envolvem análise de correlação entre documentos ou termos e podem envolver técnicas de agrupamento de documentos ou termos [Yang and Pederson, 1997]. São exemplos desses métodos, o *term strength* [Wilkinson, 1992] e *discrimination value* [Salton, 1983].

As fórmulas mais simples e comuns, que servem para praticamente todo o tipo de aplicação, são apresentadas a seguir.

A *freqüência absoluta*, também conhecida por *Term Frequency* (TF), é quantidade de vezes que um termo ocorre em um documento. Essa é a medida de peso mais simples, porém não é aconselhada por ser incapaz de fazer distinção entre os termos que ocorrem em poucos documentos e os termos que com alta freqüência em

muitos documentos. Em alguns casos esse tipo de análise poderia ser extremamente importante, pois os termos que ocorrem em muitos documentos não são capazes de diferenciar um documento de outro. Além disso, a freqüência absoluta não leva em conta a quantidade de termos existentes no documento. Por exemplo, num documento pequeno, determinado termo que ocorre muitas vezes, pode ser menos representativo do que outros que ocorrem poucas vezes em documentos extensos e vice-versa.

Nesse contexto, a *freqüência relativa* busca solucionar esse problema, levando em conta o tamanho do documento (quantidade de termos que ele possui) e normalizando os pesos de acordo com essa informação. Sem essa normalização, os documentos grandes e pequenos acabam sendo representados por valores em escalas diferentes. Com isso os documentos maiores possuem melhores chances de serem recuperados, já que receberão valores maiores no cálculo de similaridades [Salton, 1987], e vice-versa.

A freqüência relativa de um termo *x* em um documento qualquer, pode ser calculada dividindo-se a freqüência do termo *(Tf)* pelo número total de termos no mesmo documento *(N)*:

$$Frelx = \frac{Tf(x)}{N} \qquad (2.1)$$

Outra medida utilizada é a *freqüência do documento* (*Document Frequency),* que busca solucionar outro problema encontrado na freqüência absoluta, onde a quantidade de documentos em que um termo aparece não é considerada. Com esse cálculo é possível indicar a quantidade de documentos que um termo ocorre, baseando-se no cálculo da freqüência do termo e na remoção do espaço de características dos termos cuja freqüência de documentos é inferior a um determinado limiar. Por exemplo, dados os termos mais freqüentes, são selecionados os documentos cujo termo aparece no mínimo *x* vezes. Essa técnica é considerada menos complexa para redução de termos, além disso é facilmente escalável para conjuntos bem maiores de textos com uma complexidade computacional aproximadamente linear em relação ao número de documentos.

Com base nos cálculos de freqüência do termo e freqüência dos documentos é possível obter a *freqüência do termo - freqüência inversa de documentos (Term Frequency - Inverse Document Frequency (Tf-Idf))*, capaz de aumentar a importância

de termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem em muitos documentos, justamente pelo fato dos termos de baixa freqüência de documentos serem, em geral, mais discriminantes [Salton, 1983]. Existem várias maneiras de se identificar o peso do termo através deste método: uma das mais utilizadas é obtida pela aplicação da seguinte fórmula [Salton, 1983]:

$$Tf - Idf = \frac{Freq_{td}}{DocFreq_t}. \qquad (2.2)$$

Onde, $Freq_{td}$ corresponde ao número de vezes que o termo *t* aparece no documento *d* e $DocFreq_t$ corresponde ao número de documentos que o termo *t* aparece.

Uma consideração sobre os pesos identificados em uma coleção de documentos é a de que eles são válidos por determinado período de tempo [Kowalski, 1997]. Isso porque a coleção pode variar devido à adição de novos documentos ou devido a mudanças no conteúdo dos documentos (que podem ser modificados).

Na seção que segue, são apresentados em maiores detalhes os modelos de representação de documentos.

2.1.2.2 Representação dos Documentos

Com base na relevância dos termos nos textos, é possível realizar a codificação dos dados e escolher um modelo de representação de documentos utilizados pelos métodos estatísticos. A codificação dos documentos é escolhida, ou baseada nas indicações de um especialista, ou ainda, pela maximização de algum critério escolhido que reflita propriedades interessantes dos dados em relação à finalidade da modelagem.

A codificação dos termos e a representação dos documentos consistem na escolha de um modelo que possa ser compreendido pelas técnicas de MT utilizadas. Em várias pesquisas de mineração de coleções de documentos, os documentos são freqüentemente codificados pela técnica *bag of words* [Bekkerman et. al, 2003], na qual cada documento é representado por um vetor composto de termos presentes no texto. A codificação *bag of words* é considerada uma simplificação de toda a abundância de informações expressa por um documento, não fornecendo portanto uma descrição fiel do conteúdo.

Modelos de representação de documentos foram desenvolvidos na área de Recuperação de Informação (RI), e são conhecidos como Modelo Booleano e Modelo de Espaço Vetorial [Baeza-Yates and Ribeiro-Neto, 1999].

O modelo booleano é uma das representações de documentos mais clássicas utilizadas em RI. Essa abordagem avalia a presença ou ausência do termo no documento, sendo binários, isto é, {0,1} os pesos atribuídos a esses termos. A maior vantagem deste modelo é sua simplicidade e necessidade de pouco espaço de armazenamento.

O modelo de espaço de vetorial (VSM) visa contornar as limitações do modelo booleano utilizando pesos e, conseqüentemente, permitindo uma similaridade parcial entre os documentos e os termos. Nesse modelo cada documento é representado por um vetor, cada dimensão diz respeito a um conceito ou termo encontrado no texto dos documentos. O peso de cada termo é usado como coordenada para aquela dimensão. O VSM e suas variantes são freqüentemente a forma mais comum de representar os documentos textuais na mineração de coleções de documentos. Isto ocorre devido às operações de vetores serem executadas muito rapidamente e existir algoritmos padronizados eficientes para realizar a seleção do modelo, a redução de dimensionalidade e visualização de espaços de vetores [Baeza-Yates and Ribeiro-Neto, 1999]. A Figura 2-3 mostra geometricamente a representação dos documentos *doc1*, *doc2*, *doc3*, *doc4* e *doc5* em relação aos termos *linguagem*, *natural* e *processamento.*

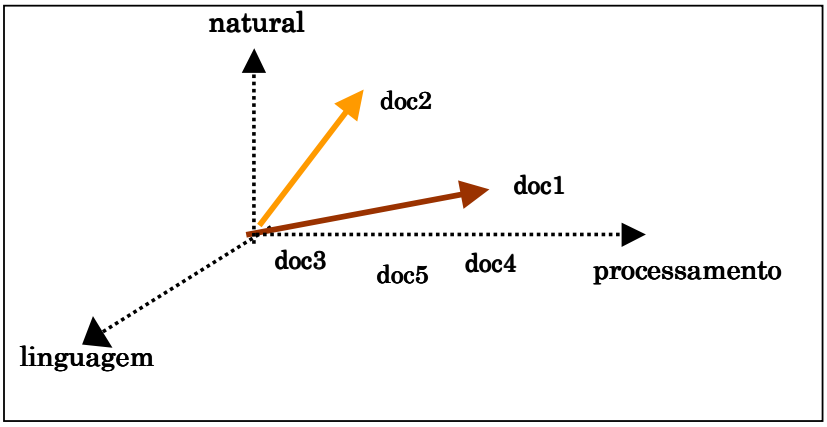

Figura 2-3 Modelo de Espaço Vetorial

Observa-se na figura que o termo *processamento* pertence ao vetor de documento *doc1*, ao contrário do termo *natural* que esta presente no documento *doc2*.

2.1.2.3 Seleção de Atributos

Considerar todos os termos de um documento para a geração de sua representação pode prejudicar o desempenho da tarefa de MT e, também, apresentar-se computacionalmente inviável.

Desta forma, a seleção de atributos consiste em eliminar termos que não são representativos, ou então combinar mais de um termo em um único atributo. A seleção também serve para diminuir o número de elementos que compõem os vetores dos documentos. Porém, esta etapa deve ser realizada com cautela. Algumas aplicações são influenciadas pelos termos de menor importância (agrupamento, classificação e sumarização, por exemplo) [Yang and Pederson,1997]. Cabe portanto ao especialista da aplicação ou ao usuário decidir se esses termos são relevantes ou não para o seu experimento. Sob esta perspectiva, vários métodos e fórmulas foram desenvolvidos com o propósito de selecionar os termos mais relevantes ao contexto da aplicação. Os métodos de seleção mais comumente usados são: filtragem baseada no peso do termo, seleção baseada no peso do termo, seleção por indexação semântica latente, seleção por linguagem natural, entre outras.

A *filtragem baseada no peso do termo,* consiste em filtrar os termos inferiores a um limiar (*threshold*) estabelecido pelo usuário ou pela aplicação [Yang and Pederson, 1997]. Além disso, mesmo depois de filtrados, o número de termos escolhidos ainda pode ser alto. Para resolver a alta dimensionalidade dos termos pode ser aplicada a seleção dos *n* termos mais relevantes. Essa técnica é denominada *seleção do peso do termo* ou *truncagem* e estabelece um número máximo de características a serem utilizadas para caracterizar um documento e todas as outras são eliminadas.

Também, é necessário que as características estejam ordenadas de acordo com seu grau de relevância. Conseqüentemente, somente as primeiras *x* características são utilizadas. Essa seleção pode ser aplicada em técnicas de descoberta de conhecimento a fim de aumentar a performance dos algoritmos, já que o número de características é proporcional ao tempo de processamento, ou seja, quanto maior o número de características a ser comparado, maior o tempo empregar no processo. No entanto, um dos problemas encontrados nessa seleção é determinar a quantidade mínima de termos necessária para a melhorar a atuação das descrições dos documentos, sem que suas características mais relevantes sejam perdidas no processo.

Outro método de seleção é a *indexação semântica latente*, desenvolvida com o intuito de reduzir o número de dimensões utilizadas pelo modelo espaço vetorial, fazendo uma análise da estrutura co-relacional de termos na coleção de documentos. Essa redução é feita identificando-se as dimensões mais similares (próximas). Uma vez identificadas, elas são aproximadas por um processo matemático de rotação, fazendo com que termos mais similares acabem na mesma dimensão. Em alguns casos, sinônimos e outros termos de forte correlação acabam sendo colocados na mesma dimensão, minimizando os problemas relacionados à diferença de vocabulário.

Além disso, é possível aplicar algumas técnicas de análise de linguagem natural [Salton, 1983] para identificar os termos mais relevantes de um documento. Essas técnicas são denominadas de *seleção por linguagem natural* e incluem a análise sintática e a análise semântica dos documentos.

Com uma gramática bem definida para um domínio específico (um *léxico* [Guthrie, 1996]), é possível realizar uma análise sintática em orações não muito complexas. Os objetos que compõem uma oração (uma frase) costumam ter posições sintáticas definidas. É possível influenciar o peso dos termos encontrados nessas posições a fim de torná-los mais ou menos relevantes. Pode-se, também, simplesmente selecionar os termos mais importantes, de acordo com sua categoria sintática (sujeitos e complementos, por exemplo), e ignorar os outros. Porém, esse tipo de técnica exige uma base de conhecimento contendo todas as combinações sintáticas possíveis (uma gramática). Além disso, a aplicação de uma técnica desse tipo envolve mais uma etapa de processamento, o que pode não ser prático em algumas aplicações (pelo fato de tornar o processo de indexação ainda mais demorado).

No contexto dessa dissertação, propõe-se aplicar na etapa de pré-processamento a seleção dos termos baseado em sua categoria sintática com o intuito de criar diversas combinações gramaticais a fim de verificar o desempenho dessas no processo de MT. O cálculo de relevância e a representação dos documentos aplicados a seleção dos termos são a freqüência relativa, Tf-Idf e espaço vetorial respectivamente.

2.1.3 Extração de Padrões

A extração de padrões consiste na escolha e aplicação de um método de mineração. Sendo que para cada método escolhido existe um tipo diferente de extração de conhecimento de textos. Esses métodos precisam ser definidos para que as tarefas

cabíveis possam ser executadas a fim de identificar padrões e relacionamento entre os documentos. Dentre as principais tarefas relacionadas ao processo de MT, destacam-se:

Categorização: esta técnica visa identificar os tópicos principais em um documento e associar este documento a uma ou mais categorias pré-definidas [Yang and Pedersen, 1997]. Muitas das técnicas de extração de padrões utilizadas em categorização de documentos são similares às utilizadas em MD [Sebastiani, 2002; Joachims, 2002].

Agrupamento: esta técnica busca agrupar, em um ou mais grupos, um conjunto de exemplos de acordo com a similaridade ou dissimilaridade de seu conteúdo. A função de similaridade entre os exemplos é definida através dos termos que aparecem nos documentos.

Sumarização: é uma técnica que identifica os termos e frases mais importantes de um documento ou conjunto de documentos gerando a partir destes um resumo ou sumário [Fayyad, 1996]. Segundo Habn and Mani (2000), esta tarefa pode ser considerada como o processo de redução da quantidade de texto em um documento, porém mantendo seus significados-chave.

Na seção que segue os métodos de avaliação e interpretação dos resultados para as tarefas de MT serão discutidos.

2.1.4 Avaliação e Interpretação dos Resultados

A etapa de avaliação e interpretação dos resultados consiste na fase de validação das descobertas obtidas durante o processo de MT. Nesta fase, os resultados podem ser analisados com a finalidade de constatar se o objetivo foi almejado, semelhante ao processo de MD. Para analisar os resultados dispõe-se de alguns recursos como medidas de desempenho, ferramentas de visualização e conhecimento de especialistas para auxiliar na validação das descobertas.

As medidas de avaliação do desempenho de um sistema são adotadas da área de RI e baseadas na idéia de relevância: se um documento atender a necessidade de informação do usuário, ele é considerado relevante à solicitação do mesmo. Basicamente, as medidas mais importantes para a avaliação do resultado e do desempenho são: *abrangência*, *precisão* e *F-measure* [Korfhage, 1997; Kowalski, 1997; Salton, 1983] e podem ser obtidas através das seguintes relações [Rijsbergen, 1979]:

Abrangência – consiste em avaliar a habilidade do sistema em recuperar os documentos mais relevantes para o usuário [Lancaster, 1968], medindo a quantidade de itens recuperados, dentre os relevantes na base de dados.

$$abrangência = \frac{n\_recuperados\_relevantes}{n\_relevantes} \tag{2.3}$$

Precisão – avalia a habilidade do sistema manter os documentos irrelevantes fora do resultado de uma consulta [Lancaster, 1968].

$$precisão = \frac{n\_recuperados\_relevantes}{n\_total\_recuperados} \tag{2.4}$$

O exame das medidas de Precisão e Abrangência separadamente pode levar a uma má avaliação do sistema, pois em geral, ao se aumentar a Precisão de um sistema, diminui-se sua Abrangência. Portanto, há a necessidade de se investigar outras formas de avaliar o sistema de modo a obter a configuração mais adequada.

As medidas do ponto de equilíbrio (*breakeven point*) [Yang and Liu,1999] e a medida do *F-measure* [Rijsbergen, 1979] combinam os valores de Precisão e Abrangência de modo a se obter o desempenho geral do sistema. O ponto de equilíbrio já foi bastante utilizado em sistemas de categorização: através do traçado dos vários pares de Precisão e Abrangência obtidos, pode-se obter por interpolação o ponto de equilíbrio, isto é, o ponto em que a Precisão e a Abrangência se igualam. A medida do *F-measure* foi definida por Rijsbergen, (1979) e permite um balanceamento entre os valores de Precisão e Abrangência através da expressão:

$$F = \frac{(\beta^2 + 1) * P * C}{\beta^2 * (P + C)} \tag{2.5}$$

Onde β é o parâmetro que permite a atribuição de diferentes pesos para as medidas de Precisão (P) e Abrangência (C), sendo 1 o valor geralmente adotado. O valor de F é maximizado quando a Precisão e a Abrangência são iguais ou muito próximas, de modo que nesta situação, por definição, o valor do *F-measure* é o próprio valor da Precisão ou da Abrangência, que por sua vez, é o ponto de equilíbrio do sistema.

No entanto, as técnicas e ferramentas para visualização de dados visam melhorar a compreensão dos resultados obtidos e a comunicação entre os usuários [Rezende et al., 1998], tornando-se instrumentos indispensáveis ao processo de MT. Segundo Fayyad et al. (2002), poderosas ferramentas de visualização que consigam gerar diversas formas de visualização (árvores, regras, gráficos 2D, 3D) combinadas com técnicas de mineração de textos podem melhorar muito o processo de MT.

Entretanto, o conhecimento de um especialista é importante ao longo do processo de extração de conhecimento em mineração de textos, e tem como objetivo auxiliar a resolver situações de conflito, indicando os melhores caminhos e complementando informações.

As diferentes medidas de avaliação de resultados podem fazer uso de variadas técnicas de MT. A seção que segue apresenta mais detalhadamente as tarefas de MT adotadas nesta dissertação: categorização e agrupamento de bases textuais.

2.2 Tarefas de Mineração de Textos

Cada tipo de tarefa de MT extrai um tipo diferente de informação dos textos e independente da tarefa escolhida o objetivo é descobrir conhecimento útil e inovador dos textos de forma a auxiliar os usuários na obtenção de informações relevantes.

Nesta dissertação será adotado as tarefas de categorização e agrupamento de bases textuais, na seção que segue estes serão melhor detalhados.

2.2.1 Categorização de Textos

A crescente quantidade de informação disponibilizada eletronicamente, principalmente através da Internet, e a dificuldade de recuperação das mesmas têm motivado o desenvolvimento de pesquisas na área de RI. O objetivo é auxiliar os usuários na localização de informações relevantes aos seus interesses.

Inúmeros sistemas de busca na Internet adotam as técnicas de recuperação baseadas em palavras-chave. Nestes sistemas, não é realizada a análise mais detalhada dos conteúdos dos textos e muita das informações retornadas aos usuários são de pouca ou nenhuma relevância. Sob essa perspectiva, uma das alternativas consiste na categorização de textos. Neste processo, é feita uma análise detalhada da informação que compõe os conteúdos [Lewis, 1991; Yang and Liu, 1999], (por exemplo, esporte,

economia, lazer, turismo, entre outras), sendo os mesmos, então, organizados em categorias pré-definidas [Rijsbergen, 1979]. Em sistemas, tais como o *Yahoo*[7], a categorização é adotada, mas feita de forma manual. Assim, categorizar grandes volumes de conteúdos manualmente tem se tornado inviável.

Atualmente, tem sido proposta a aplicação de técnicas de IA na automatização do processo de categorização. Abordagens utilizando técnicas de aprendizado de máquina, tais como Árvores de Decisão (ADs), Redes Neurais Artificiais (RNAs), classificadores Bayesianos, o algoritmo *Nearest Neighbor,* e S*upport Vector Machines (SVM)* vêm sendo exploradas. Nestas abordagens, as ferramentas são treinadas para oferecer suporte a problemas de decisão (tais como classificação ou agrupamento), baseadas em dados de treinamento. Esse treinamento consiste na apresentação dos exemplos e os resultados esperados ao algoritmo de aprendizado, de modo que este aprenda a reproduzir os resultados com o mínimo de erro. A Figura 2-4, ilustra o processo de categorização com base em um conjunto de dados amostrais.

De um modo geral, o processo de categorização automática de textos vem sendo amplamente adotado em conteúdos textuais [Sebastiani, 2002]. Todavia, dentre as aplicações que podem ser exploradas pode-se citar a organização de mensagens de e-mail, o filtro de mensagens (evitando as do tipo *spam*, por exemplo) e notícias, a organização hierárquica e a recomendação de documentos.

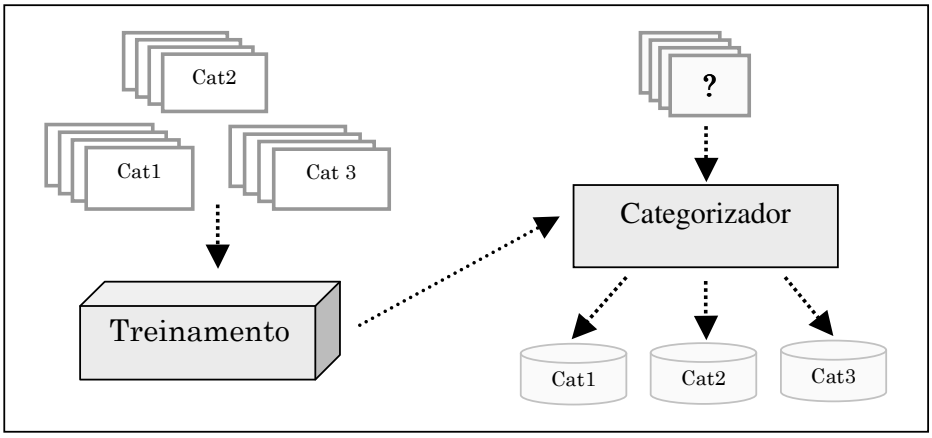

Figura 2-4 Processo de Categorização

[7] http://www.yahoo.com

No contexto desta dissertação é abordado o processo automático de categorização de textos, com o intuito de avaliar o desempenho dos classificadores no processo de MT. Os classificadores são baseados nas técnicas de aprendizado de máquina Árvores de Decisão, utilizando o algoritmo C4.5 e Redes Neurais Artificiais, com o algoritmo *Backpropagation*.

Árvores de decisão vem sendo amplamente utilizada na área de IA desde o desenvolvimento do algoritmo ID3 criado por Quinlan (1986). Sua construção é realizada a partir de exemplos utilizando um aprendizado não incremental. O conjunto de exemplos de treinamento é apresentado ao sistema que induz a árvore, construída de cima para baixo (da raiz para as folhas) com base na informação contida nos exemplos. Os nós nas árvores correspondem aos atributos, utilizados na representação dos objetos, enquanto que os ramos representam valores alternativos pré-determinados para estes atributos.

As árvores de decisão são inspiradas no uso de um algoritmo de particionamento recursivo, tais como ID3 [Quinlan, 1986], C4.5 [Quinlan, 1993] e CART [Steinberg and Colla, 1995]. Nesses algoritmos, a indução da árvore baseia-se na divisão recursiva do conjunto de exemplos de treinamento em subconjuntos mais representativos, utilizando a métrica de ganho de informação. Após a construção da árvore, esta poderá ser utilizada para a classificação de novos exemplos, descritos em termos dos mesmos atributos usados na sua representação. Está classificação é feita percorrendo-se a árvore, até se chegar à folha, que determina a classe a que o exemplo pertence ou sua probabilidade de pertencer àquela classe.

Em tarefas de categorização de textos, a indução de árvores de decisão tem sido um grande sucesso. Alguns exemplos são encontrados em [Lewis and Ringuette 1994; Moulinier et al., 1996; Joachims, 1998; Yang and Pederson, 1999].

Joachims (1998) utilizou o algoritmo C4.5 na construção da árvore de decisão, e o desempenho obtido foi comparado em relação aos resultados obtidos por outros classificadores. Em Yang (1999), compara-se o desempenho da aplicação de árvores de decisão em atividades de categorização de documentos nas várias versões da coleção *Reuters.*

Oliveira and Castro (2000) propõem a aplicação de ADs para a categorização múltipla de textos. O processo de aprendizado é efetuado sobre todas as categorias dos

textos, gerando um único classificador capaz de associar cada novo documento a uma ou mais categorias pré-definidas. São usadas para os experimentos as coleções *Reuters – 21578* (um conjunto de 21.578 artigos da agência de notícias *Reuters*, devidamente classificados) e *Medline* (formada por artigos da área médica).

Melcop et al (2002) adotam as ADs na classificação de páginas *web*. É apresentado um sistema para recuperação e classificação de páginas de domínios específicos, a partir da utilização de uma ontologia do domínio escolhido. No sistema de classificação, as árvores geradas são capazes de definir se um determinado documento pertence ou não a uma dada categoria (categorização binária). Experimentos para o domínio da culinária são apresentados.

Outra técnica de aprendizado de máquina adotada neste trabalho são as RNAs.

Para Haykin (2001), as redes neurais são sistemas paralelos distribuídos compostos por unidades de processamento simples (nodos) que calculam determinadas funções matemáticas. Tais unidades são dispostas em uma ou mais camadas e interligadas por um grande número de conexões, geralmente unidirecionais, sendo que na maioria das vezes estas conexões estão associadas a pesos. Estes pesos guardam o conhecimento de uma rede neural e são usados para definir a influência de cada entrada recebida por um neurônio na sua respectiva saída. Ao ajustar os seus pesos a rede neural assimila padrões e esta é capaz de fazer generalizações, isto é, produzir saídas consistentes para entradas não apresentadas anteriormente. Possuem a capacidade de aprender a partir dos exemplos de treinamento, e agrupar ou classificar novos exemplos, com base na generalização do que aprenderam.

Uma rede neural básica geralmente possui os seguintes elementos em sua estrutura:

- Uma camada de entrada, composta de várias unidades, de acordo com o número de características utilizadas na representação dos exemplos considerados;

- Uma ou mais camadas intermediárias compostas por alguns neurônios responsáveis pela modelagem de relações não lineares entre as unidades de entrada e saída;

- Uma camada de saída que fornece a resposta do sistema;

- Ligações entre as várias camadas (pesos), responsáveis pela propagação dos sinais entre essas, que são aprendidos durante a fase de treinamento do sistema.

A Figura 2-5 ilustra a estrutura de uma rede neural com quatro entradas ([x_1, x_2, x_3, x_4]), uma saída ([y]) e quatro neurônios na camada intermediária, as conexões que ligam as entradas às camadas intermediárias são denominadas pesos (w).

O neurônio é o elemento processador da rede neural. Cada neurônio gera uma saída com base na combinação de sinais de entrada recebidos de outros neurônios, com os quais está conectado, ou a partir de sinais externos. Na maior parte dos modelos à saída de um neurônio é, o resultado de uma função de ativação aplicada a soma ponderada de suas entradas.

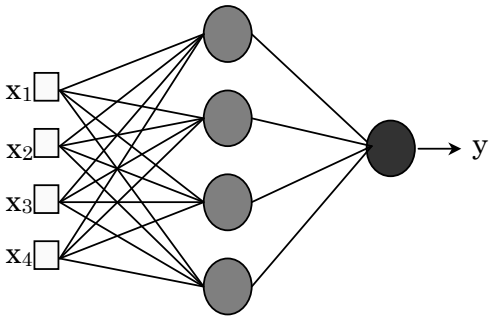

Figura 2-5 Visão esquemática de uma rede neural

A topologia de uma rede é definida pelo número de camadas da rede, o número de neurônios em cada camada, e o tipo de conexão entre nodos. A fase de treinamento de redes neurais corresponde a determinação dos pesos das conexões entre os neurônios. Elas possuem a capacidade de aprender por exemplos e fazer interpolações e extrapolações do que aprenderam. No aprendizado conexionista, não se procura obter regras como a abordagem simbólica da IA, mas sim determinar a intensidade de conexões entre neurônios.

Para o treinamento de uma rede neural podem ser utilizados três diferentes tipos de algoritmos de aprendizagem:

- Aprendizagem supervisionada;
- Aprendizagem não-supervisonada;
- Aprendizagem por reforço.

No conjunto de treinamento de uma aprendizagem supervisionada estão contidos os valores das unidades de entrada e das unidades de saída para cada exemplo. Os pesos da rede devem ser ajustados até que os valores das unidades de saída sejam semelhantes aos valores apresentados na entrada da rede. Em um conjunto de treinamento de uma aprendizagem não-supervisionada, estão contidos somente os valores das unidades de entrada e a rede sozinha executa um agrupamento, a fim de aprender as classes presentes no conjunto de treinamento. Por fim, em uma aprendizagem por reforço, mesmo que no conjunto de treinamento estejam contidos apenas os valores das unidades de entrada, a cada ciclo do processamento da rede, algumas dicas sobre a saída desejada são fornecidas, ao invés de sua resposta direta.

Trabalhos utilizando Redes Neurais Artificiais nas tarefas de categorização de textos podem ser encontrados em Rizzi et al. (2000). Este trabalho propõe a categorização de textos na organização e filtragem de informações na área empresarial, com o objetivo de validar a aplicação do processo, especialmente nas tarefas de análise e distribuição das informações. Um conjunto de experimentos, utilizando a coleção *Reuters – 21578* e a RNA *Multi-layer Perceptron* (MLP), com o algoritmo de aprendizado *Backpropagation* (BP), é apresentado.

Corrêa and Ludemir (2002) apresentam e comparam o uso das RNAs, dos tipos MLP (com algoritmo de aprendizado BP*)* e *Self-organizing Maps* (SOM) e das técnicas tradicionalmente aplicadas no processo de categorização, tais como ADs e classificadores *Naive Bayes*. Foram utilizadas nos experimentos as coleções de dados *K1* (uma coleção de 2.340 páginas *web* sobre notícias diversas), *PubsFinder* (um conjunto de 1.300 páginas *web* classificadas como sendo ou não páginas de publicação de artigos técnicos e científicos) e a sub-coleção *Metais* da *Reuters – 21578*.

Duarte et al. (2002) propõem o uso de um agente neural para a coleta e classificação de informações disponíveis na Internet. O agente coleta as informações, submetendo-as a um classificador neural que utiliza uma RNA do tipo MLP. Foram utilizados os algoritmos *Levemberg Marquadt* e MOBJ para o treinamento da rede. Experimentos para o domínio da economia são apresentados.

Na seção que segue será detalhada a tarefa de agrupamento de textos.

2.2.2 Agrupamento

O agrupamento tem sido amplamente utilizado em padrões de análises exploratórias de dados, tomada de decisão e situações de aprendizado de máquina, incluindo mineração de dados, recuperação de informação, segmentação de imagens e classificação de padrões.

O agrupamento é uma técnica de aprendizado não supervisionado, portanto não existem rótulos pré-determinados para os padrões de treinamento. Técnicas de agrupamento buscam agrupar em um ou mais grupos, um conjunto de exemplos similares entre si. Ao final do processo, cada grupo irá conter um representante que permite identificar aquele grupo. O resultado do agrupamento é dependente da medida de similaridade (critério) adotada.

Theodoridas (1998) define agrupamento como sendo um dos processos mais primitivos do ser humano, e desenvolvido para auxiliar a processar grandes quantidades de informações e agrupá-las de acordo com seus atributos comuns.

Conforme descrito em Jain and Dubes (1999), para se executar uma tarefa de agrupamento normalmente os seguintes passos devem ser seguidos:

- Selecionar as características;
- Estabelecer uma medida de similaridade;
- Determinar um critério de agrupamento;
- Escolher um algoritmo de agrupamento;
- Validar os resultados do agrupamento;
- Interpretar os resultados.

Faz-se necessário selecionar características devido aos métodos de agrupamento estarem de alguma forma baseados em alguma medida de similaridade entre os exemplos. Portanto, para que os exemplos possam ser agrupados, é preciso identificar as características dos exemplos e agrupá-los de acordo com a quantidade de características similares entre eles. A seleção das características é baseada nas medidas de relevância descritas na seção 2.1.2.1.

A segunda etapa do processo de agrupamento consiste em analisar todos os exemplos com o objetivo de selecionar semelhanças entre os exemplos. O grau de semelhança entre os exemplos é dado, em geral, por uma fórmula de similaridade. Sendo que essa fórmula analisa todas as características semelhantes que os exemplos possuem e retorna um valor, indicando um grau de similaridade entre os exemplos.

Conforme descrito em Cole (1998), para agrupar objetos de acordo com sua similaridade, deve-se definir uma medida de quão próximos dois objetos estão, ou quão bem seus valores se comparam. Uma pequena distância entre os objetos deve indicar uma alta similaridade.

Como a distância é uma função que envolve somente atributos relativos a dois exemplos, o processo de agrupamento baseado em similaridade pode ser feito de forma simples e sem necessitar de nenhum tipo de conhecimento sobre o assunto tratado. Entretanto, devido a variedade de tipos e escalas dos valores dos atributos, a medida de similaridade deve ser escolhida cuidadosamente. As características utilizadas para definir similaridade são especificadas e combinadas em uma medida a ser calculada para todos os exemplos

Na literatura existem várias medidas de similaridade, tais como *Euclidiana*, *Manhattan* e *Minkowski*. Dentre essas medidas [Cole 1998] destacam-se a distância *Euclidiana,* que é a mais utilizada.

Depois de concluída a etapa de cálculo de similaridades (ou distância), tem-se uma matriz que indica os valores de similaridade entre os exemplos. Com base nesta matriz é possível identificar os grupos de exemplos, especificando algum tipo de regra de relacionamento entre os mesmos e então escolher um dos tipos de agrupamentos (tais como, K-means, Cobweb, etc.). Por fim, os resultados do agrupamento são validados e interpretados com o intuito verificar o desempenho do processo.

Fundamentalmente, existem dois tipos de agrupamento quando nos referimos à MT: um baseado em termos e outro em documentos. No primeiro tipo, os grupos de termos similares são identificados, com o intuito de construir um dicionário de termos que definam assuntos similares. Em contraste, o agrupamento por documentos visa identificar os documentos de assuntos similares e alocá-los em um grupo. Esse método é extremamente útil quando não se tem uma idéia dos assuntos (das classes) tratados em cada documento e deseja-se separá-los por assunto.

Logo, o agrupamento tem como finalidade identificar os exemplos (tais como, documentos ou termos) que apresentem características em comum, agrupando-os em subconjuntos de exemplos similares, sendo que estes possam ser os mais variados possíveis antes de agrupados.

Além disso, o agrupamento pode ser classificado em relação à forma como os grupos são construídos. De acordo com Cutting (1992), quanto à forma, há dois tipos de agrupamento: o agrupamento por partição e o agrupamento hierárquico, ambos relatam à forma na qual os grupos são constituídos.

Na primeira abordagem, os documentos são distribuídos em classes distintas, não havendo relação direta entre as classes, sendo esta denominada de agrupamento por partição total *(flat partition).* Essa técnica consiste em agrupar os documentos em um número pré-determinado de grupos distintos. Os documentos são agrupados de forma que todos os elementos de um mesmo grupo possuem no mínimo um grau de semelhança, que é indicado pelo número de características em comum dos mesmos. Além de constituir grupos distintos, permite a possibilidade de colocar ou não determinado documento em mais de um grupo. Quando os documentos são atribuídos a um único grupo diz-se que o processo é *disjunto*. Caso um documento seja atribuído a mais de um grupo, por possuir forte relação com mais de um grupo, diz-se que o processo não é disjunto. A Figura 2-6 ilustra como exemplo o resultado de um agrupamento por partição total disjunta.

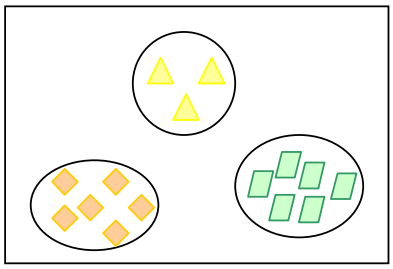

Figura 2-6 Exemplo de um agrupamento por partição

Uma desvantagem encontrada neste método é devido a não existência de estruturas que indiquem o co-relacionamento entre os grupos, impossibilitando o usuário de identificar os assuntos mais específicos e os mais abrangentes.

No agrupamento hierárquico *(hierarchic partition)*, o processo de identificação de grupos é geralmente realimentado recursivamente, utilizando tanto documentos quanto grupos já identificados previamente como entrada para o processamento. Neste

contexto, os grupos identificados são recursivamente analisados, fazendo com que as relações entre os grupos também sejam identificadas. Para que a hierarquia seja identificada algoritmos específicos podem ser utilizados, assim como algoritmos de partição total, aplicados de forma recursiva. Este método consiste em oferecer estruturas de navegação hierárquica entre os grupos, facilitando a localização de informações. Porém, essa vantagem exige um tempo de processamento maior, já que o algoritmo de agrupamento deve passar analisar os grupos identificados várias vezes.

Portanto, o agrupamento hierárquico ou não hierárquico possui vários algoritmos e aplicações a fim de facilitar a organização e a recuperação de informações. Um método de agrupamento bastante utilizado na literatura e o utilizado neste trabalho para realizar o agrupamento é o algoritmo *K-means*, descrito detalhadamente na seção que segue.

2.2.2.1 Algoritmo *K-means*

Em Theodoridas (1998), o método de agrupamento *k-means* é baseado em centróide para particionar um conjunto de n objetos em k grupos tal que a similaridade intracluster resultante é alta, mas a similaridade intercluster é baixa. A similaridade de grupos é medida em respeito ao valor médio dos objetos em um grupo, que pode ser visto como o centro de gravidade do grupo.

O algoritmo *k-means* trabalha da seguinte forma: primeiro, ele, aleatoriamente seleciona *k* objetos, cada um dos quais, inicialmente, representa a média do grupo. Para cada um dos objetos remanescentes, é feita a atribuição ao grupo ao qual o objeto é mais similar, baseado na distância entre o objeto e a média do grupo. Então, novas médias para cada grupo são computadas. Esse processo se repete até que a função critério venha a convergir. Normalmente, o critério do erro quadrado é usado. A Figura 2-7 ilustra os passos do algoritmo *k-means*.

1. Determinar a quantidade k desejada de grupos;
2. Inicializar as médias m_1, m_2 .... m_k;
3. Repetir até não haver mais variações no k-médias (m_1,m_2,..,m_k);
   a. Utilizar as médias estimadas para classificar os grupos;
   b. Faça i = 1 até k
      i. Atribuir m_i à média dos exemplos

Figura 2-7 Passos do Algoritmo K-means

O método *k-means*, entretanto, pode ser aplicado somente quando a média de um grupo é definida. Isto pode não ser o caso em algumas aplicações, tais como quando dados com atributos categóricos (nominais) estão envolvidos.

2.3 Considerações Finais

MT é uma área recentemente nova e seus domínios de aplicações são numerosos, principalmente devido à proliferação de documentos digitais, tais como bibliotecas digitais, Intranets e a própria Internet.

O processo de MT é utilizado para descobrir padrões interessantes e úteis em um conjunto de dados textuais. Apesar de similar ao processo de MD, que trabalha com dados estruturados (numéricos), o processo de MT difere, principalmente, por trabalhar com dados não estruturados em formato textual. Assim, para que esses dados textuais possam ser submetidos a algoritmos de aprendizado de máquina empregados em tarefas de MD, é necessário um tratamento diferenciado na etapa de pré-processamento de dados textuais.

Normalmente esta etapa de pré-processamento é bastante custosa, devido aos textos não apresentarem um formato único (por exemplo, formato HTML, txt, doc, ppt, entre outros), ocasionando uma padronização dos mesmos para então serem pré-processados. As fases que constituem a etapa de pré-processamento em dados textuais geralmente são compostas por análise léxica, remoção de irrelevantes e normalização morfológica.

As tarefas de MT escolhidas para a avaliar o desempenho do pré-processamento dos textos nesta dissertação são a categorização e agrupamento. Além disso, as técnicas de aprendizado de máquina adotadas para ambas tarefas são: árvores de decisão, redes neurais artificiais e algoritmo *k-means*, respectivamente.

Cabe ressaltar que, o bom desempenho de extração de conhecimento em tarefas de MT está diretamente relacionado a qualidade dos dados que descrevem uma situação do mundo real, bem como a aplicação correta das etapas usuais de pré-processamento de textos.

Neste contexto, o próximo capitulo apresenta uma nova abordagem baseada no uso de informações lingüísticas na etapa de pré-processamento no processo de MT.

Serão detalhados a metodologia adotada no pré-processamento e os recursos utilizados para a extração de categorias gramaticais.

3 Uso de Informações Lingüísticas em MT

Para executar o processo de MT, é necessário transformar os documentos em uma boa representação estruturada e para a aplicação de algumas técnicas de aprendizado é preciso reduzir a alta dimensionalidade do espaço de atributos. Para essa transformação e redução, a etapa de pré-processamento é fundamental e bastante custosa.

Com base em estudos no pré-processamento de MT, a proposta desta dissertação é verificar a adequação do uso de informações lingüísticas nesta etapa, analisando as vantagens e desvantagens dessa abordagem em relação aos métodos usuais.

Os métodos usuais de pré-processamento normalmente consistem nas fases de eliminação de termos irrelevantes dos textos, assim como na aplicação de algoritmos de variação morfológica.

Para tanto, para aplicar a nova abordagem baseada em extração de informações lingüísticas na etapa de pré-processamento dos textos, foram estabelecidos os seguintes procedimentos: uma primeira fase para a análise sintática dos textos, e posteriormente uma segunda fase para a extração das informações gramaticais.

Este capítulo apresenta uma introdução ao uso de informações lingüísticas em MT e está organizado como segue. Na seção 3.1 são comentados os métodos tradicionais de pré-processamento. A seção 3.2 apresenta a análise sintática dos textos, com a tarefa de anotar os textos sintaticamente a fim de extrair informações. Na seção 3.3, são apresentados a metodologia aplicada a seleção das estruturas lingüísticas dos textos, por fim, na seção 3.4 as considerações finais obtidas em relação ao capítulo.

3.1 Pré-processamento baseado em Termos

Conforme comentado, o pré-processamento baseado em termos consiste basicamente das etapas ilustrada na Figura 3-1. Em Riloff (1995), dependendo da aplicação de domínio, essas etapas podem variar sua ordem ou simplesmente não ocorrer.

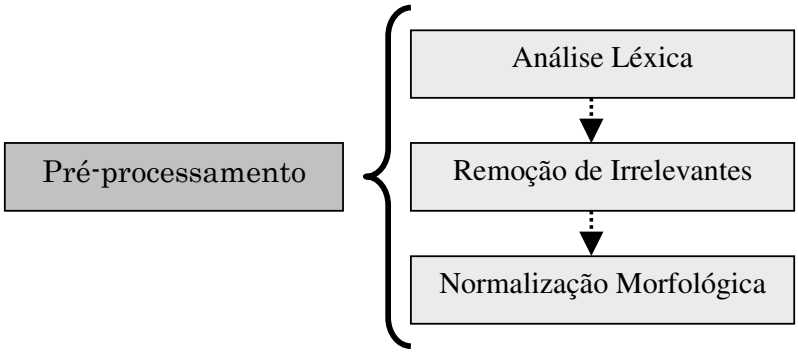

Figura 3-1 Etapas do Pré-processamento

Para facilitar o entendimento destas etapas, considere uma frase de um documento referente ao assunto Dinheiro.

"Janeiro começa com grandes liquidações."

Conforme descrito na seção 2.1.1, a análise léxica consiste na limpeza dos textos. A frase resultante da aplicação desta etapa (ilustrada abaixo) encontra-se sem as aspas duplas e o ponto final, devido a estes não serem relevantes para o documento.

Janeiro começa com grandes liquidações

Geralmente a análise léxica é aplicada quando os textos não estão em um formato único, por exemplo, textos em HTML, ou convertidos para um formato txt, apresentam muitos caracteres insignificantes, prejudicando o entendimento do texto.

A remoção de termos irrelevantes visa a retirada dos termos de pouca importância para a representatividade dos documentos em um processo de MT. Com base na frase exemplificada acima, o resultado da remoção dos termos irrelevantes ficaria:

Janeiro começa grandes liquidações

Nesta etapa e exemplo, observa-se que a preposição *com* foi removida. Geralmente os termos que compõem uma lista de termos irrelevantes são constituídos de: preposições, conjunções, verbos (tais como: ser e estar), preposições, entre outros.

A etapa referente à normalização morfológica em vez de eliminar os termos irrelevantes, reduz os radicais dos termos restantes da etapa anterior. Por exemplo, para a frase dada como exemplo, a normalização morfológica dos termos seria:

```
Janeir começ grande liquidaç
```

Analisando a frase resultante, observa-se que os termos apresentam-se reduzidos ao seu radical, possibilitando que termos com o mesmo radical sejam considerados unicamente.

Após as etapas de pré-processamento, os termos resultantes são submetidos a preparação e seleção dos dados, para então serem aplicados as tarefas de MT. Estas etapas foram detalhadas nas seções 2.1.1 e 2.1.2, respectivamente.

Neste contexto, esta dissertação busca comparar esses métodos tradicionais em relação ao uso de informações lingüísticas no pré-processamento. Porém, para que informações lingüísticas possam ser extraídas faz-se necessária a análise sintática dos textos.

A seção que segue detalha a metodologia aplicada neste processo.

3.2 Análise sintática

A sintaxe é o componente[8] do sistema lingüístico que determina as relações formais que interligam os constituintes da sentença, atribuindo-lhe uma estrutura.

Nessa estrutura encontra-se o sintagma, ou seja, unidade da análise sintática composta de um núcleo (por exemplo, um verbo, um nome, um adjetivo etc.) e de outros termos que a ele se unem, formando uma locução que entrará na formação da oração. O nome do sintagma depende da classe da palavra que forma seu núcleo, havendo assim sintagma nominal (núcleo substantivo), sintagma verbal (núcleo verbo), sintagma adjetival (núcleo adjetivo), sintagma preposicional (núcleo preposição) e sintagma adverbial (núcleo advérbio). Na teoria gerativa existem sintagmas formados por núcleos mais abstratos, como tempo, concordância etc.

Nas estruturas frasais observa-se uma hierarquia, sendo que esta pode ser representada por meio de um diagrama arbóreo: no topo encontra-se a unidade maior (a sentença), nos níveis intermediários os elementos sintáticos que formam a sentença (constituintes) e na base da estrutura, os itens lexicais correspondentes (as palavras). A Figura 3-2 exemplifica um diagrama arbóreo de uma frase simples em língua portuguesa.

[8] Segundo a gramática generativa, as partes constituintes da gramática de uma língua são o componente sintático, o componente fonológico e o componente semântico.

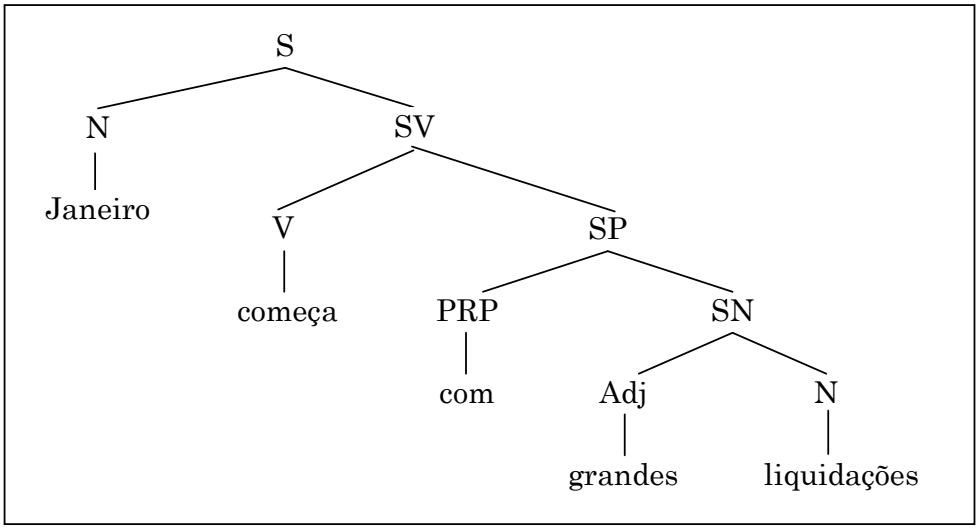

Figura 3-2 Estrutura arbórea da sentença *Janeiro começa com grandes liquidações*

Nesse exemplo a derivação da estrutura da sentença se constitui dos elementos indicados por N, SV, V, SP, PRP e SN que, por sua vez, realizam os itens lexicais da base. Os elementos dos níveis intermediários são a expressão de um conjunto de informações necessário para a composição da sentença.

Para construir representação em árvore de uma sentença, é preciso conhecer quais as estruturas legais da língua, isto é, a gramática da língua (o conjunto de regras). Com o diagrama arbóreo é possível, observar que, sob a aparente diversidade, todas as estruturas possuem uma organização interna que obedece a princípios gerais bem definidos.

Um conjunto de regras gramaticais descrevem quais as são estruturas permitidas. Essas regras dizem que um certo símbolo pode ser expandido em árvore, pela seqüência de outros símbolos. Cada símbolo é um constituinte da sentença que pode ser composto de uma ou mais palavras.

A nível computacional para construir o diagrama arbóreo de uma sentença faz-se necessário um processamento sintático, ou analisador sintático, que é alcançado por intermédio da análise morfossintática[9].

Enquanto, o analisador léxico-morfológico trabalha com a estrutura dos termos e com a classificação dos mesmos em diferentes categorias[10], o analisador sintático

[9] Nível da estrutura que engloba a morfologia (estudo das formas) e a sintaxe (regras de combinação que regem a formação de frases).

trabalha no nível do sintagma, tentando validar o agrupamento de termos que compõem as frases.

O processamento sintático é responsável por construir (ou recuperar) uma estrutura sintática válida para a sentença de entrada. Para tanto, é orientado por uma representação da gramática da língua em questão. Em se tratando de uma língua natural, freqüentemente se adota uma gramática "parcial", que, embora não abranja todas as construções da língua, contempla aquelas construções válidas de interesse para a aplicação. Assim, evita-se o grande volume de informações gramaticais que pode aumentar demasiadamente a complexidade de sua representação, bem como a complexidade do próprio processo de análise.

Para auxiliar na construção de um processamento sintático, utiliza-se um analisador sintático. Esse, por sua vez é um procedimento que pesquisa os vários modos de como combinar regras gramaticais, com a finalidade de gerar uma estrutura de árvore que represente a estrutura sintática da sentença analisada. Caso a sentença for ambígua, o analisador sintático poderá obter todas as possíveis estruturas sintáticas que a representam. No entanto, os analisadores sintáticos podem apresentar variações de acordo com (a) a relação que estabelecem com o usuário; (b) os recursos disponíveis; e (c) as estratégias de análise.

O analisador sintático utilizado neste trabalho para extração de informações lingüísticas dos documentos é o desenvolvido por Eckhard Bick [Bick, 2000] para a língua portuguesa e dito PALAVRAS. Ele realiza tarefas como tokenização, processamento léxico-morfológico, análise sintática e faz parte de um grupo de analisadores sintáticos do projeto VISL (*Visual Interactive Syntax Learning*)[11], do *Institute of Language and Communication* da *University of Southern Denmark*.

O analisador sintático recebe como entrada o conjunto de sentenças de um corpus, e gera a análise sintática das sentenças. A Figura 3-3 mostra a marcação sintática do analisador sintático PALAVRAS, para a sentença em língua portuguesa “Janeiro começa com grandes liquidações”.

---

[10] Por exemplo, as categorias, ou classes gramáticas, em língua portuguesa são dez: verbo, substantivo, adjetivo, advérbio, artigo, preposição, interjeição, pronome, numeral e conjunção.

[11] Disponível em: http://visl.sdu.dk/visl/pt/parsing/automatic/

```
STA:fcl
=SUBJ:n('janeiro' M S)     Janeiro
=P:v-fin('começar' PR 3S IND)    começa
=ADVL:pp
==H:prp('com')  com
==P<:np
===>N:adj('grande' F P)    grandes
===H:n('liquidação' F P)   liquidações
=.
```

Figura 3-3 Marcação do analisador sintático PALAVRAS

No exemplo, podemos verificar as seguintes etiquetas morfossintáticas: 'SUBJ' = o sujeito, 'P' = predicado verbal, o 'ADVL' = adjunto adverbial, após o ':' existe a forma sintática para grupo de palavras ('np' = sintagma nominal, 'pp' = sintagma preposicional) e etiquetas para palavras únicas ('v-fin' = verbo flexionado, 'prp' = preposição, 'n'= substantivo, 'art' = artigo, etc); entre parênteses a forma canônica da palavra e as etiquetas de flexão (gênero, número, entre outras); após os parênteses existe a palavra como aparece no corpus. Os sinais de '=' no começo de cada linha representam o nível da expressão[12].

Com base nas marcações do analisador sintático, um conjunto de programas foi desenvolvido em cooperação com a Universidade de Évora[13]: a ferramenta Xtractor – apresentada em [Gasperin, 2003]. A ferramenta engloba a análise do corpus, por meio do PALAVRAS, e o tratamento da saída do parser, com a geração de três arquivos XML.

O primeiro arquivo XML, chamado de *Words,* contém as palavras do corpus com elementos <word> e atributos id que representam um identificador único para cada termo (Figura 3-4) de um texto.

```
<words>
<word id="word_1">Janeiro</word>
<word id="word_2">começa</word>
<word id="word_3">com</word>
<word id="word_4">grandes</word>
<word id="word_5">liquidações</word>
<word id="word_6">.</word>
</words>
```

Figura 3-4 Arquivo Words (Termos do Corpus)

[12] O conjunto de etiquetas está disponível em http://visl.hum.sdu.dk/visl/pt/info/symbolset-manual.html

[13] Programas desenvolvidos pelo Prof. Paulo Quaresma, através de cooperação CAPES/GRICES.

O segundo arquivo apresenta as informações morfo-sintáticas das palavras do texto, denominado de POS (P*part-of-Speech)* (Figura 3-5), onde o elemento <n> indica um substantivo, e o elemento <v> um verbo.

```
<words>
<word id="word_1">
<n canon="janeiro" gender="M" number="S"/>
</word>
<word id="word_2">
<v canon="começar">
<fin tense="PR" person="3S" mode="IND"/>
</v>
</word>
<word id="word_3">
<prp canon="com"/>
</word>
<word id="word_4">
<adj canon="grande" gender="F" number="P"/>
</word>
<word id="word_5">
<n canon="liquidação" gender="F" number="P"/>
</word>
</words>
```

Figura 3-5 Arquivo de POS (Informações morfo-sintáticas)

Por fim, o terceiro arquivo consiste nas estruturas e subestruturas sintáticas das sentenças, representadas por *chunks.* Um *chunk* representa a estrutura interna da sentença e pode conter sub-chunks, como ilustrado na Figura 3-6.

```
<text>
<paragraph id="paragraph_1">
<sentence id="sentence_1" span="word_1..word_6">
<chunk id="chunk_1" ext="sta" form="fcl" span="word_1..word_5">
<chunk id="chunk_2" ext="subj" form="n" span="word_1">
</chunk>
<chunk id="chunk_3" ext="p" form="v_fin" span="word_2">
</chunk>
<chunk id="chunk_4" ext="advl" form="pp" span="word_3..word_5">
<chunk id="chunk_5" ext="h" form="prp" span="word_3">
</chunk>
<chunk id="chunk_6" ext="p" form="np" span="word_4..word_5">
<chunk id="chunk_7" ext="n" form="adj" span="word_4">
</chunk>
<chunk id="chunk_8" ext="h" form="n" span="word_5">
</chunk>
</chunk>
</chunk>
</chunk>
</sentence>
</paragraph>
</text>
```

Figura 3-6 Estrutura de *Chunks*

Com base nos arquivos gerados, aplica-se a extração de estruturas lingüísticas através de folhas de estilos, descrita em maiores detalhes na seção que segue.

3.3 Extração de Estruturas

Gerados os arquivos no formato XML, a extração de estruturas é realizada aplicando folhas de estilos XSL[14] (*eXtensible Stylesheet Language*) nos arquivos de POS (informações morfo-sintáticas).

XSL é um conjunto de instruções destinadas à visualização de documentos XML. Desta forma, dado um documento XSL é possível transformar um documento XML em diversos formatos, incluindo RTF, TeX, PostScript, HTML e TXT. A linguagem XSL auxilia a identificação dos elementos (nodos) de um documento XML, permitindo a simplificação do processamento de transformação desses elementos em outros formatos de apresentação. Contudo, é possível criar múltiplas representações da mesma informação a partir de vários documentos XSL aplicados a um único documento XML.

Uma folha de estilos é composta por um conjunto de regras, chamado *templates,* ativados no processamento de um documento XML. A Figura 3-7 exemplifica o conteúdo de uma folha de estilos XSL, que extrai a forma canônica de todos os substantivos de um arquivo de POS. Para cada word em POS ela seleciona aquelas com elemento filho `<n>` que indica ser um substantivo, enquanto que o atributo `@canon` de cada elemento contém a forma canônica[15] da palavra.

Para realização deste trabalho, foram implementadas folhas de estilo para extrair automaticamente dos textos algumas categorias sintáticas para construção da tabela atributo-valor. As combinações de categorias extraídas dos documentos para a realização dos experimentos foram: substantivos (*noun*), substantivos e adjetivos, substantivos e nomes próprios, substantivos, nomes próprios e adjetivos e por fim nomes próprios e adjetivos.

---

[14] Linguagem desenvolvida pelo W3C (*World Wide Web Consortium*) disponiível em: http://www.w3.org/Style/XSL/

[15] A forma canônica de uma palavra significa a forma base de tal palavra, sem flexões de gênero, número ou grau.

```
<xsl:style sheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:param name='pattern'>parametro</xsl:param>
<xsl:output method="text" encoding="ISO-8859-1" indent="no"/>

<xsl:template match="/words">
    <xsl:apply-templates select="word"/>
</xsl:template>
<xsl:template match="word">
    <xsl:apply-templates select="n"/>
</xsl:template>
<xsl:template match="n">
<xsl:value-of select="@canon"/><xsl:value-of select="'&#010;'"/>
</xsl:template>

</xsl:stylesheet>
```

Figura 3-7 Folha de Estilo XSL

A folha de estilo aplicada sobre os arquivos dos POS, para a extração dos substantivos, tem como resultado os termos ilustrados na Figura 3-8. Os substantivos resultam da aplicação das etapas de análise sintática e seleção das estruturas, na seguinte frase do texto:

```
Janeiro começa com grandes liquidações.
```

```
Janeiro liquidação
```

Figura 3-8 Substantivos extraídos do texto

Com base nas estruturas extraídas, os termos são apresentados à etapa de preparação e seleção dos atributos no processo de MT com o intuito de aplicar as tarefas de categorização e agrupamento de textos, abordadas neste estudo. As avaliações dos resultados obtidos nas tarefas de mineração adotadas são baseadas nas medidas de desempenho descritas na seção 2.1.4.

3.4 Considerações Finais

Em um processo de mineração de textos, normalmente a seleção de características é baseada em métodos tradicionais de pré-processamento como, remoção de termos irrelevantes e redução da variação morfológica. No entanto, é possível aplicar a etapa de pré-processamento métodos utilizados em Processamento de Linguagem Natural (PLN), tais como, análise sintática ou analise semântica na seleção dos termos relevantes.

Nesse contexto, um dos objetivos deste trabalho consiste na proposta e aplicação de uma metodologia baseada no uso de informações lingüísticas nas etapas de pré-processamento de textos, visando descobrir conhecimento útil em um conjunto de dados não rotulados, usando as tarefas de categorização e agrupamento de textos no processo de MT.

Para que informações lingüísticas possam ser extraídas e comparadas nas tarefas de MT foi aplicada a análise sintática nos textos, bem como a seleção de estruturas utilizando folhas de estilo XSL.

No próximo capítulo, será descrito a coleção de documentos utilizada e os métodos aplicados para prepará-los para a realização dos experimentos com cada uma das tarefas de MT: categorização e agrupamento, bem como a metodologia utilizada na concepção, treinamento e avaliação dos classificadores e grupos para a coleção.

4 Métodos aplicados aos Experimentos

Neste capítulo são descritos o *corpus*[16] e os métodos utilizados no pré-processamento dos textos, de modo a prepará-los para os experimentos com cada uma das tarefas de MT: categorização e agrupamento. A metodologia utilizada na concepção, treinamento e avaliação dos classificadores e grupos para a coleção é apresentada.

Na tarefa de categorização, foram utilizadas as técnicas de aprendizado supervisionado do tipo simbólico e conexionista, tais como: Árvores de Decisão, com o algoritmo J48, re-implementação do algoritmo C4.5 [Quinlan, 1993], e as Redes Neurais MLP [Beale and Jackson, 1997] respectivamente. Para realizar o treinamento com as redes MLP foi escolhido o algoritmo *Backpropagation com momentum*, devido ao relativo sucesso e popularidade em vários domínios [Rumelhart and McClelland, 1986]. Para a realização dos experimentos com o agrupamento, foi adotado o algoritmo de particionamento *K-means* [Theodoridas, 1998], descrito anteriormente na seção 2.2.2.1.

Para a realização dos experimentos foi adotada a ferramenta *Weka (Waikato Environment for Knowledge Analysis)* [Witten, 2000], formada por uma coleção de algoritmos de Aprendizado de Máquina para resolução de problemas reais de Mineração de Dados (MD).

4.1 O Corpus

O corpus utilizado para realização dos experimentos foi fornecido pelo Núcleo Interinstitutional de Lingüística Computacional - NILC[17]. Esse conjunto de textos contém mais de 4000 documentos escritos em português do Brasil, divididos nos tópicos: didáticos, jornalísticos, jurídicos, literários e técnicos/científicos. Desses documentos, foram selecionados 5093 documentos do tópico jornalístico, totalizando

[16] Por corpus entende-se uma coleção de textos, de tamanhos variados. Os textos nessa coleção estão organizados de acordo com a (s) aplicação (ões) a que se destina o corpus.

[17] Disponível em http://www.nilc.icmp.usp.br/nilc

1.323.700 palavras. Os assuntos relacionados a este tópico foram extraídos da Folha Jornal de São Paulo do ano de 1994, dos quais 855 textos foram escolhidos e classificados em cinco seções: Informática, Imóveis, Esporte, Política e Turismo. Cada documento é um arquivo texto (extensão txt) com tamanho entre 4Kbytes e 250 Kbytes, com um mínimo de 80 e um máximo de 1200 termos (incluindo números e termos irrelevantes, por exemplo, eleitoral, multimídia, aluguel, com, de, entre outros). A Tabela 4-1 mostra a distribuição original dos documentos e o número de termos por seção (existência de termos repetidos).

Tabela 4-1 Corpus NILC: Distribuição de documentos e número de termos por seção

Seções	Informática	Imóveis	Esporte	Política	Turismo	Total
Documentos	171	171	171	171	171	**855**
Nº Termos	42.640	45.912	41.278	45.403	50.399	**225.632**

Na seção que segue é detalhado o pré-processamento aplicado à tarefa de categorização de textos.

4.2 Pré-processamento em Categorização de Textos

Resgatando os conceitos descritos na seção 2.1.1, o pré-processamento de textos é uma tarefa árdua, pois além de apresentar um alto custo computacional exige um cuidadoso planejamento e processamento para obtenção de um bom desempenho no processo de MT.

No processo de categorização de textos o pré-processamento é normalmente formado por um conjunto de etapas fundamentais, ilustradas na Figura 3-1 pág.31.

Para que o resultado do processo de categorização não seja tendencioso, primeiramente o corpus original foi distribuído em três ordens diferentes, resultando em três conjuntos de dados com todos os 855 documentos, visando verificar a variação dos resultados em diferentes distribuições. Para cada versão (V1, V2 e V3) o corpus foi dividido em dois conjuntos de documentos específicos: um para fase de treino e outro para a fase de teste. Cada seção (categoria) do corpus foi particionada em aproximadamente 2/3 dos documentos para o conjunto de treino e 1/3 dos documentos para o conjunto de teste. O objetivo deste particionamento é tornar os testes mais

realistas, já que o algoritmo é testado em uma parte diferente do conjunto de dados do qual foi treinado. A Tabela 4-2 mostra a distribuição dos documentos do conjunto de treino e teste e o número de documentos existentes em cada distribuição do corpus, enquanto que a Tabela 4-3 mostra a quantidade de termos para cada uma das categorias e variações do corpus.

Tabela 4-2 Número de documentos no conjunto de treino e teste por categoria

Categoria	Informática	Imóveis	Esporte	Política	Turismo	Total
Base Treino	114	114	114	114	114	**570**
Base Teste	57	57	57	57	57	**285**
V1,V2,V3	171	171	171	171	171	**855**

Tabela 4-3 Distribuição dos termos para cada categoria e variação do corpus

Corpus	Categoria	Informática	Imóveis	Esporte	Política	Turismo
V1	**Treino**	29.109	31.033	26.875	32.406	32.422
	Teste	13.185	14.537	14.061	12.655	17.635
V2	**Treino**	27.435	30.497	26.512	27.768	33.249
	Teste	14.589	15.073	14.424	17.293	16.808
V3	**Treino**	28.047	29.610	28.485	29.948	34.443
	Teste	14.247	15.960	14.451	15.113	15.614

Após a definição e divisão do corpus em um conjunto de treino e teste, os exemplos selecionados foram pré-processados por um conjunto de programas, que contemplam as sub-etapas correspondentes (Figura 3-1, pág.31) gerando os arquivos de entrada (com os documentos devidamente codificados) que são submetidos a ferramenta de aprendizado. Um programa, implementado na linguagem de programação C++, contempla o processo de análise léxica e remoção dos termos irrelevantes. A lista de *stopwords* foi disponibilizada pela Universidade de Évora[18] e adaptada para o português do Brasil, contendo 476 termos (tais como: artigos, preposições, verbos ser e estar, pronomes, entre outros). A Tabela 4-4 mostra a

[18] Abrigo do projeto de cooperação Desenvolvimento e Integração de recursos para o Processamento do Português CAPES/FCT - DIRPI

quantidade de termos nos conjuntos de treino e teste após a remoção de termos irrelevantes para cada variação do corpus.

Tabela 4-4 Distribuição do número de termos após remoção de irrelevantes

Corpus	Categoria	Informática	Imóveis	Esporte	Política	Turismo
V1	**Treino**	12.963	13.541	11.666	14.582	13.442
	Teste	5.816	6.231	6.191	5.796	7.691
V2	**Treino**	12.225	13.222	11.634	12.675	14.420
	Teste	6.554	6.550	6.223	7.703	6.713
V3	**Treino**	12.370	12.781	12.414	13.499	14.404
	Teste	6.409	6.991	5.443	6.879	6.729

Analisando a distribuição dos termos apresentados na Tabela 4-2, observa-se que a remoção de irrelevantes permite reduzir em aproximadamente 40% o número de termos em uma coleção de dados. Restringindo os documentos a apenas os termos mais representativos.

Com base nos termos resultantes, foi aplicada a extração de afixos dos termos. Para esta extração foi adotado o algoritmo proposto por Martin Porter[19], que remove as letras finais dos termos da língua portuguesa que possuem a mesma variação morfológica e de flexão.

Dessa forma, estes procedimentos descritos acima constituem os programas aplicados aos métodos usuais de pré-processamento. A abordagem referente ao uso de informações lingüísticas na etapa de pré-processamento será descrita a seguir.

Para o pré-processamento usando informações lingüísticas foi utilizada a metodologia proposta na seção 3.2, baseada na análise sintática dos textos e seleção de estruturas através de folhas de estilos XSL.

Inicialmente, todos os documentos que compõem os conjuntos de treino e teste representados na Tabela 4-2 foram submetidos ao analisador sintático PALAVRAS, a fim de gerar a análise sintática das sentenças presentes nos documentos. Com base na marcação do corpus e auxílio da ferramenta *Xtractor*, foram gerados três arquivos XML

[19] Disponível em várias línguas em http://snowball.sourceforge.net

(Words, POS e Chunks) para cada conjunto de treino e teste das categorias Informática, Imóveis, Esporte, Política e Turismo.

Na abordagem proposta por esta dissertação, o pré-processamento consiste na seleção de estruturas gramaticais dos textos marcados com análise sintática das sentenças. Assim, para que estas estruturas sejam extraídas, folhas de estilo XSL foram implementadas e aplicadas aos arquivos de POS gerados.

A primeira folha de estilo XSL implementada (Figura 3-7) e aplicada aos arquivos de POS, buscava extrair automaticamente dos textos a categoria gramatical substantivo. Essa extração foi baseada em estudos realizados por Kuramoto (2002), que propos o uso de sintagmas nominais[20] como meio de acesso a informações de Sistemas de Recuperação de Informações (SRI). A Tabela 4-5 mostra a quantidade de termos nos conjuntos de treino e teste após a seleção da estrutura gramatical substantivo para cada variação do corpus.

Tabela 4-5 Distribuição de termos após seleção da estrutura: substantivo

Corpus	Categoria	Informática	Imóveis	Esporte	Política	Turismo
V1	**Base Treino**	5.994	6.526	4.539	6.243	6.795
	Base Teste	3.191	3.511	2.503	2.617	3.154
V2	**Base Treino**	6.277	6.741	4.500	5.717	6.654
	Base Teste	2.908	3.296	2.542	3.143	3.295
V3	**Base Treino**	6.099	6.807	5.045	5.760	6.449
	Base Teste	3.086	3.230	1.997	3.100	3.500

Analisando o resultado obtido na Tabela 4-5 e traçando um comparativo com a Tabela 4-4, observa-se que através da seleção de estruturas gramaticais dos textos, o número de termos do corpus reduz significativamente, possibilitando uma melhor representatividade nos documentos.

Visando verificar quais estruturas gramaticais são mais representativas para o corpus em questão, novas combinações gramaticais e folhas de estilos XSL foram criadas. A Tabela 4-6 apresenta a distribuição dos termos resultantes após a seleção das

[20] O núcleo de um sintagma nominal em uma sentença normalmente é um substantivo.

combinações gramaticais: substantivos - adjetivos, extraídas do trecho da folha de estilo XSL mostrada na Figura 4-1.

```
...
</xsl:template>
<xsl:template match="n">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
<xsl:template match="adj">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
...
```

Figura 4-1 Trecho da Folha de Estilo XSL para: substantivos-adjetivos

Nesta folha cada word em POS seleciona aquelas com elemento filho `<n>` e `<adj>` que indicam ser um substantivo e adjetivo, respectivamente, enquanto que o atributo `@canon` de cada elemento contém a forma canônica da palavra.

Tabela 4-6 Distribuição de termos para a seleção: substantivos-adjetivos

Corpus	Categoria	Informática	Imóveis	Esporte	Política	Turismo
V1	**Treino**	7.111	8.084	5.657	7.626	8.329
	Teste	3.799	4.286	3.114	3.111	3.981
V2	**Treino**	7.485	8.318	5.621	6.820	8.284
	Teste	3.425	4.052	3.150	3.917	4.026
V3	**Treino**	7.224	8.333	6.264	7.208	8.007
	Teste	3.686	4.032	2.507	3.709	4.303

As estruturas gramaticais nomes próprios-adjetivos adicionadas aos substantivos representam uma nova combinação gramatical. Os termos restantes após a aplicação desta combinação são mostrados na Tabela 4-7 e o trecho da folha de estilo XSL para esta extração é mostrado na Figura 4-2.

```
...
</xsl:template>
<xsl:template match="n">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
<xsl:template match="prop">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
<xsl:template match="adj">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
...
```

Figura 4-2 Trecho da Folha de Estilo para: substantivos-nomes próprios-adjetivos

À medida que novas estruturas são inseridas junto aos substantivos, o número de termos restantes após a extração aumenta.

Tabela 4-7 Distribuição de termos para: substantivos-nomes próprios-adjetivos

Corpus	Categoria	Informática	Imóveis	Esporte	Política	Turismo
V1	**Treino**	9.114	9.585	8.240	10.022	10.700
	Teste	4.844	4.838	4.340	4.237	5.032
V2	**Treino**	9.486	9.662	8.143	9.137	10.492
	Teste	4.472	4.761	4.437	5.122	5.240
V3	**Treino**	9.316	9.599	8.777	9.359	10.272
	Teste	4.642	4.824	3.803	4.900	5.460

Os trechos das folhas de estilos correspondentes a extração dos termos sintáticos: substantivos - nomes próprios e nomes próprios-adjetivos, são mostradas nas Figuras 4-3 e 4-4.

```
...
</xsl:template>
<xsl:template match="n">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
<xsl:template match="prop">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
...
```

Figura 4-3 Trecho da Folha de Estilo para: substantivos-nomes próprios

```
...
</xsl:template>
<xsl:template match="prop">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
<xsl:template match="adj">
    <xsl:value-of select="@canon"/><xsl:value-of
select="'&#010;'"/>
</xsl:template>
...
```

Figura 4-4 Trecho da Folha de Estilo para seleção: nomes próprios-adjetivos

Os resultados obtidos após a extração das combinações gramaticais são mostrados nas Tabelas 4-8 e 4-9. Com base nesses resultados, pode-se verificar que o número de termos relevantes é inferior aos obtidos no pré-processamento baseado em métodos usuais (Tabela 4-4).

Tabela 4-8 Distribuição de termos para a seleção: substantivo-nomes próprios

Corpus	Categoria	Informática	Imóveis	Esporte	Política	Turismo
V1	**Treino**	7.997	8.027	7.122	8.639	9.166
	Teste	4.236	4.063	3.729	3.743	4.205
V2	**Treino**	8.278	8.085	7.022	8.034	8.862
	Teste	3.955	4.005	3.829	4.348	4.509
V3	**Treino**	8.191	8.068	7.558	8.091	8.714
	Teste	4.042	4.022	3.293	4.291	4.657

Tabela 4-9 Distribuição de termos para: nomes próprios-adjetivos

Corpus	Categoria	Informática	Imóveis	Esporte	Política	Turismo
V1	**Treino**	3.120	3.059	3.702	3.779	3.905
	Teste	1.653	1.327	1.837	1.620	1.878
V2	**Treino**	3.209	2.921	3.644	3.420	3.838
	Teste	1.564	1.465	1.895	1.979	1.945
V3	**Treino**	3.217	2.792	3.732	3.599	3.923
	Teste	1.556	1.594	1.807	1.800	1.960

A Tabela 4-10 apresenta uma análise comparativa das distribuições dos termos extraídos nas etapas de pré-processamento baseado em métodos usuais e no uso de informações lingüísticas. Esta Tabela apresenta a soma do número total de termos extraídos nos conjuntos de treino e teste nas variações do corpus para ambas abordagens. Cabe ressaltar que, o total de termos extraídos é o mesmo para cada versão do corpus. As categorias gramaticais que estão sendo comparadas são: substantivos; substantivos-adjetivos (subst-adj); substantivos-nomes próprios-adjetivos (subst-nprop-adj); substantivos-nomes próprios (subst-nprop); nomes próprios-adjetivos (nprop-adj) e métodos usuais de pré-processamento.

Tabela 4-10 Comparação entre os termos extraídos das etapas de pré-processamento

Corpus	Informática	Imóveis	Esporte	Política	Turismo	**Total**
substantivos	9.185	10.037	7.042	8.860	9.949	**45.073**
subst-adj	10.910	12.370	8.771	10.737	12.310	**55.098**
subst-nprop-adj	13.958	14.423	12.580	14.259	15.732	**70.952**
subst-nprop	12.233	12.090	10.851	12.382	13.371	**60.927**
nprop-adj	4.773	4.386	5.539	5.399	5.783	**25.880**
usuais (palavras)	18.779	19.772	17.857	20.378	21.133	**97.919**

Considerando os comparativos traçados na Tabela 4-10, pode-se verificar que comparado a Tabelas 4-3 houve uma redução em torno de 40% no número de termos do corpus.

Na seção que segue é descrita a metodologia utilizada na preparação dos dados para ambas as abordagens de pré-processamento para o processo de categorização de textos.

4.2.1 Preparação e Seleção dos Dados

Para ambos pré-processamentos (métodos usuais e informações lingüísticas), os termos mais relevantes foram identificados nos conjuntos de treino com o auxílio de scripts desenvolvidos na linguagem shell (Linux) e Perl, implementando o cálculo de relevância: freqüência absoluta, freqüência relativa e Tf-Idf. Adotou-se como seleção de termos a técnica de *truncagem*, já que, além de ser implementada mais facilmente,

não influi negativamente nos resultados, conforme testes realizados por Schütze (1997), além de oferecer um ganho de performance no algoritmo.

A representação dos documentos foi elaborada com base no espaço vetorial. Vetores locais foram construídos com base nos termos restantes do pré-processamento (métodos usuais ou utilizando informações lingüísticas), conforme mostra a Figura 4-5. Cabe destacar que os vetores locais foram elaborados a partir dos *n* termos mais freqüentes do conjunto de treinamento correspondentes aos documentos de cada categoria. Dessa forma, os vetores globais foram construídos através da união dos vetores e serviram como índices para os vetores de cada exemplo e as posições correspondentes representaram a importância da mesma no documento.

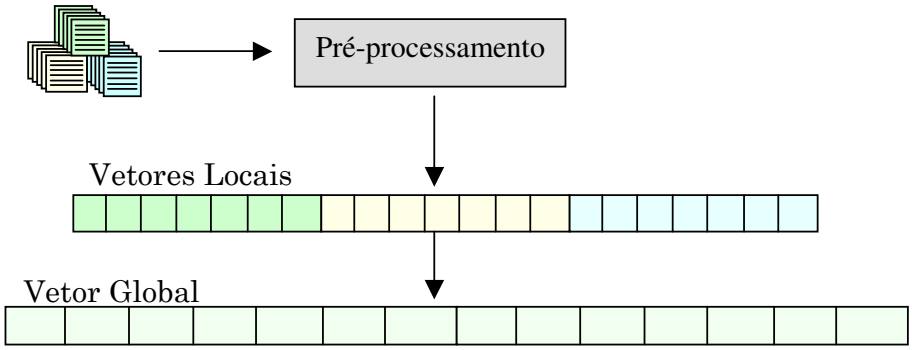

Figura 4-5 Construção dos vetores locais e globais

Nesta dissertação foi adotada a categorização múltipla, dentre várias categorias, um exemplo é associado à categoria mais representativa, cujos conceitos associados mais representam o exemplo em questão. Neste tipo de categorização, é adotado o vetor global para a codificação de cada exemplo, baseado na freqüência relativa e Tf-Idf. Após a codificação dos exemplos, foram gerados os scripts para a ferramenta de suporte *Weka* (descrita na seção 4.5). As técnicas de aprendizado adotadas neste trabalho para o processo de categorização de textos são: Árvores de Decisão e Redes Neurais Artificiais.

A ferramenta *Weka* é responsável pela geração de árvores de decisão através do algoritmo J48 que re-implementa o funcionamento do algoritmo C4.5 desenvolvido por Quinlan (1993), e o treinamento neural que comporta o algoritmo de aprendizado *Backpropagation* (BP). Os parâmetros utilizados nos algoritmos, bem como os experimentos realizados com o uso da ferramenta são comentados no próximo capítulo

. Na seção que segue é descrito o pré-processamento adotado no agrupamento de textos.

4.2 Pré-processamento em Agrupamento de Textos

O pré-processamento em um agrupamento de textos é constituído basicamente das mesmas etapas ilustradas na Figura 3-1 pág.31, porém após a preparação e seleção dos dados, os vetores de documentos são submetidos a uma medida de similaridade. Esta medida busca identificar nos vetores de documentos as características em comum para dois exemplos. Esta etapa é necessária devido ao agrupamento não possuir categorias pré-definidas como na categorização de textos.

Assim como no processo de categorização o corpus dividido em três ordens diferentes de distribuições, para cada uma destas distribuições o corpus foi particionado em dois conjuntos de treino e teste, respeitando a mesma quantidade de documentos descrita anteriormente (2/3 e 1/3). A Tabela 4-11 mostra a distribuição do número de documentos do conjunto de treino e teste.

A metodologia e programas utilizados nas etapas de pré-processamento tradicionais e baseados em informações lingüísticas são os mesmos adotados na seção anterior 4.1.

Tabela 4-11 Distribuição dos documentos do conjunto de treino e teste

Conjunto	Total
Treino	**570**
Teste	**285**

A Tabela 4-12 mostra a o número de termos para cada conjunto de treino e teste para as versões do V1, V2 e V3 do corpus antes e após a remoção de irrelevantes. Enquanto que a Tabela 4-13 mostra a quantidade de termos resultantes após extração das estruturas gramaticais: substantivos; substantivos-adjetivos; substantivos–nomes próprios-adjetivos; substantivos–nomespróprios e nomespróprios-adjetivos.

Comparando-se os resultados obtidos nas extrações dos termos, observa-se que na Tabela 4-12, houve um decréscimo representativo no número de termos. Sendo que, os termos sintáticos extraídos das combinações gramaticais nomes próprios - adjetivos são os que apresentam um número menor de termos.

Tabela 4-12 Quantidade de termos para cada conjunto nas variação do corpus

Variações	Conjunto	Nº Termos	
		Corpus	Remoção de Irrelevantes
V1	**Treino**	151.842	66.171
	Teste	72.076	31.725
V2	**Treino**	115.144	64.234
	Teste	78.457	33.743
V3	**Treino**	150.533	65.468
	Teste	73.565	32.509

Tabela 4-13 Número de termos após seleção de estruturas gramaticais

Variações	Conjuntos	Nº Termos após Seleção das Estruturas				
		Substantivo	Substantivo Adjetivo	Substantivo N.Próprio Adjetivo	Substantivos Nomes Próprios	Adjetivos Nomes Próprios
V1	**Treino**	30.097	36.807	47.661	40.951	17.565
	Teste	14.976	18.291	23.271	19.976	8.315
V2	**Treino**	29.889	36.528	46.920	40.281	17.032
	Teste	15.184	18.570	24.032	20.646	8.848
V3	**Treino**	30.160	36.861	47.383	40.622	17.263
	Teste	14.913	18.237	23.629	24.347	8.717

4.2.1 Preparação e Seleção dos Dados

Para ambos pré-processamento (métodos tradicionais e informações lingüísticas), após a identificação dos termos mais relevantes, seleção baseada na *truncagem* (por exemplo, os *n* termos mais freqüentes do conjunto de treino), e representação dos documentos, parte-se para a análise de similaridade entre estes documentos.

Todo o processo de agrupamento está baseado em algum tipo de similaridade entre os documentos, pois os agrupa (ou separa-os) em grupos de documentos que

possuam alguma semelhança (similaridade) entre si. Esta é a etapa mais crucial do processo, e sua eficiência depende muito das características identificadas como relevantes. Caso as etapas anteriores não tenham identificado características realmente relevantes todo o processo pode ser comprometido, já que o resultado pode não ser o ideal para a coleção de documentos sendo processada.

A medida de similaridade adotada neste trabalho, é baseada na função *coseno (cosine).* Pela função *coseno*, a distância (similaridade) entre dois documentos é dada pela fórmula abaixo:

$$similaridade(Q,D) = \frac{\sum_{k=1}^{n} w_{qk} * w_{dk}}{\sqrt{\sum_{k=1}^{n} (w_{qk})^2 * \sum_{k=1}^{n} (w_{dk})^2}} \qquad (2.6)$$

Onde, Q corresponde ao vetor de termos do documento X, D vetor de termos do documento Y, w_{qk} são os pesos dos termos do documento X e w_{dk} são os pesos dos termos do documento Y.

Depois de realizado o cálculo de similaridade entre os vetores de documentos e gerados os vetores de entrada para a ferramenta *Weka*, estes são submetidos ao algoritmo *K-means* adotado neste trabalho com o intuito de agrupar os documentos similares. Os parâmetros utilizados no algoritmo, são os sugeridos pela ferramenta, com semente aleatória igual a 10 e o número de grupos igual a 5.

Na seção que segue a ferramenta adotada para a realização dos experimentos será detalhada.

4.3 Ferramenta utilizada nos Experimentos

A ferramenta utilizada na realização dos experimentos é denominada *Weka*[21] (*Waikato Environment for Knowledge Analysis*). *Weka* é um ambiente para extração de padrões desenvolvido pela Universidade de Waikato na Nova Zelândia que permite ao usuário criar, rodar, modificar, e analisar experimentos com diferentes algoritmos de aprendizado, de uma forma mais conveniente do que quando processamos os experimentos individualmente.

[21] Disponível em http://www.cs.waikato.ac.nz/ml/weka/

A ferramenta Weka é formada por uma coleção de algoritmos de Aprendizado de Máquina para resolução de problemas reais de Mineração de Dados (MD). Estes algoritmos são implementados na linguagem Java, apresentando como características a portabilidade, podendo rodar nas mais variadas plataformas e aproveitar os benefícios de uma linguagem orientada a objetos (tais como: modularidade, polimorfismo, encapsulamento, reutilização de código entre outros) e ser um software domínio público.

Os algoritmos implementados na ferramenta suportam métodos de aprendizado supervisionado (tais como, Árvores de Decisão, Regras de Associação, Redes Neurais Artificiais, *Naive Bayes*, *Support Vector Machine*), métodos de aprendizado não-supervisionado (como *K-means*, EM, e Cobweb), filtros (transformação dos dados, seleção de atributos), ferramentas de visualização de dados que permitem plotar em forma gráfica os resultados obtidos nos experimentos. Os métodos de aprendizado de máquina adotados nessa dissertação são do tipo: supervisionado (Árvores de Decisão e Redes Neurais Artificiais) e não supervisionado (*K-means).*

As execuções dos algoritmos podem ser realizadas por linhas de comando ou módulo gráfico. Para que os dados possam ser aplicados a um dos algoritmos é necessário que estes sejam convertidos ao formato de entrada do Weka. O Weka apresenta um formato de entrada de dados próprio, denominado ARFF. Esse formato consiste basicamente de duas partes: uma lista de atributos e uma lista dos exemplos. A primeira parte contém uma lista de todos os atributos, onde são definidos os tipos de atributos ou os valores que podem ser representados, quando utilizamos valores estes devem estar entre "{ }" separados por vírgulas. A segunda parte consiste nos exemplos, ou seja, os registros a serem minerados com o valor dos atributos para cada exemplo separado por vírgula, sendo que a ausência de um item em um registro deve ser representada pelo símbolo "?".

Um exemplo de um arquivo no formato ARFF, é mostrado na Figura 4-6, onde a primeira linha do arquivo representa o nome do conjunto de dados atribuído pelo comando @relation script, em seguida é apresentado à relação dos atributos com o nome do atributo e o tipo ou seus possíveis valores, @attribute nome_do_atributo tipo ou {valores} e por fim os dados @data, onde cada linha representa um exemplo.

```
@relation script

@attribute banc real
@attribute cdrom real
@attribute corr real
@attribute cr$ real
@attribute dia real
@attribute equip real
@attribute jog real
@attribute multimíd real
@attribute program real
@attribute category {informatica,dinheiro,esporte}

@data
0,0,0,0,0,0,0,1,0,informatica
3,1,0,0,0,0,0,8,1,informatica
0,1,0,0,0,0,0,4,2,informatica
0,5,0,0,0,2,3,6,3,informatica
1,0,0,0,0,1,0,1,2,informatica

.....
```

Figura 4-6 Arquivo de Entrada no Formato ARFF

Para a construção desse arquivo de entrada os atributos foram selecionados baseados na *truncagem* (neste caso, os 3 termos mais relevantes de cada categoria), e as categorias são identificadas no arquivo através do atributo *category*. Na linha onde os dados (@data) são apresentados os vetores de documentos foram construídos baseados no modelo de espaço vetorial e codificados de acordo com a freqüência do termo nos documentos de cada categoria: informática, esporte e dinheiro.

Com base nesse arquivo de entrada (Figura 4-6), pode-se analisar o resultado da geração de Árvores de Decisão com o algoritmo J48, que re-implementa o funcionamento do C4.5. Analisando os dados de saída do treinamento (Anexo A.1), observa-se os dados utilizados para o treinamento, como: nome do algoritmo executado, os parâmetros do treinamento (weka.classifiers.j48.J48 -C 0.25 -M 2), o número de atributos (Attributes: 10), a quantidade de exemplos (Instances: 33), o tamanho da árvore, o número de folhas, bem como as regras geradas para a criação da árvore. Este exemplo possui um número de 33 documentos, referente as categorias dinheiro, esporte e informática sendo que cada categoria apresenta 11 documentos.

As classes corretamente e incorretamente classificadas são mostradas em porcentagem, as medidas de precisão, abrangência e F-measure são descritas nos detalhes da acurácia por categoria, assim como a matriz de confusão mostra a quantidade de documentos por categoria.

O algoritmo J48 não necessita de um ponto inicial de busca, dessa forma pode ser executado apenas uma vez. Os valores dos parâmetros de treinamento do algoritmo foram os sugeridos pela ferramenta (limiar de confidência para poda igual a 0,25 e número mínimo de instâncias por folha igual a 2). Os mesmos parâmetros são adotados para os experimentos nesta dissertação.

Em contraste, o resultado obtido com o treinamento de uma rede neural para o algoritmo *Backpropagation*, é ilustrado no Anexo A.2. Neste exemplo, podem ser observados os parâmetros utilizados para o treinamento da rede, os pesos atribuídos aos padrões de entrada, bem como as categorias classificadas corretamente e incorretamente. Para o treinamento de um classificador utilizando Redes Neurais inicialmente é necessária uma busca para a melhor arquitetura (parâmetros de treinamento) para cada topologia e posteriormente à escolha da melhor topologia. A melhor topologia de uma rede neural é escolhida baseada no erro obtido nas simulações realizadas para cada topologia gerada.

Em ambos resultados destes 33 exemplos o treinamento foi aplicado baseado na validação cruzada, ou *cross-validation* sobre os dados de entrada. Nesta validação, os exemplos são aleatoriamente divididos em k partições mutuamente exclusivas de tamanho aproximadamente igual a $\frac{N}{k}$ exemplos, sendo N o número total de exemplos, utilizando-se então (k-1) partições para treinamento a hipótese induzida é testada na partição restante. Esse processo é repetido k vezes, cada vez considerando uma partição diferente para teste.

No contexto desta dissertação, o treinamento dos dados é baseado no conjunto de treinamento (2/3 dos documentos) e teste (1/3 dos documentos). Primeiramente, os dados que representam o conjunto de treinamento são submetidos ao algoritmo de aprendizado e, com base nos resultados obtidos, o conjunto de teste mede a generalização do que foi aprendido.

4.4 Avaliação dos Resultados

A avaliação dos resultados para o processo de categorização e agrupamento de textos são baseados nas medidas de desempenho utilizadas em RI, descrito na seção 2.1.4.

No processo de categorização para os classificadores treinados por Árvores de Decisão e Redes Neurais gerados pelas etapas de pré-processamento tradicionais e com informações lingüísticas foram observados os seguintes aspectos:

Erro de Classificação - Percentagem de padrões no conjunto de teste incorretamente classificados pela árvore gerados e pela rede treinada.

F-measure Teste por categoria – Percentagem de erro no conjunto de teste obtido pela árvore gerada ;

O Erro de Classificação no Teste avalia o desempenho na classificação dos padrões no conjunto de teste, e estes são os parâmetros usados para comparar o desempenho dos modelos gerados pelos algoritmos de aprendizagem simbólica e conexionista utilizados neste trabalho.

A medida F-Measure no Teste mede a capacidade de generalização do modelo gerado, permitindo verificar se durante o treinamento este assimilou do conjunto de treinamento características significativas que permitem um bom desempenho em outros conjuntos de dados, ou se concentrou em peculiaridades. Neste trabalho esta medida é utilizada por permitir um balanceamento entre os valores de Precisão e Abrangência nas categorias. O valor da F-Measure é maximizado quando a Precisão e a Abrangência são iguais ou muito próximas.

Para avaliar os resultados obtidos pelos classificadores do corpus em relação as abordagens de pré-processamento propostas, foi adotada a matriz de confusão gerada nos resultados do conjunto de teste (conforme Figuras A.1 e A.2). A organização desta matriz parte do princípio de que um sistema realiza n decisões binárias, sendo que cada uma delas tem ou não exatamente uma resposta correta. Medidas de eficácia para sistemas de categorização de textos como: Abrangência, Precisão, Falha, Acurácia, entre outras podem ser extraídas a partir desta matriz.

A avaliação dos resultados obtidos pelos agrupamentos dos documentos, também será baseada na matriz de confusão. A ferramenta *Weka* mostra após a execução do algoritmo de agrupamento os exemplos pertencentes a cada grupo. Com base nesses exemplos será calculada a Abrangência e Precisão dos Grupos.

A comparação de desempenho entre as técnicas de aprendizado supervisionado e não-supervisonado será realizada por meio da comparação do desempenho do melhor

classificador obtido por cada técnica para o corpus em termos do número de atributos contidos no vetor global.

Para se determinar qual a técnica de aprendizado que melhor se adequa aos termos resultantes do pré-processamento baseado em métodos tradicionais e em informações lingüísticas, serão comparados os menores erros de classificação para a seleção dos termos em comum.

4.5 Considerações Finais

Os experimentos realizados nesta dissertação avaliam as etapas de pré-processamento das tarefas de categorização e agrupamento de textos no processo de MT. O desempenho dessas tarefas são comparados com base em diferentes técnicas de pré-processamento.

O corpus utilizado nos experimentos foi fornecido pelo Núcleo Interinstitutional de Lingüística Computacional – NILC, contendo 855 documentos referentes às seções jornalísticas: Informática, Imóveis, Esporte, Política e Turismo do ano de 1994.

O corpus inicialmente foi dividido em três partes distintas, visando verificar a variação dos resultados em diferentes distribuições. Para cada distribuição o corpus foi dividido em dois conjuntos de documentos específicos: um para fase de treino e outro para a fase de teste, com o objetivo de tornar os testes mais realistas, já que o algoritmo é testado em uma parte diferente do conjunto de dados do qual foi treinado.

Para a realização dos experimentos os documentos do corpus foram pré-processados utilizando duas abordagens: métodos tradicionais (análise léxica, remoção de irrelevantes, normalização morfológica) e seleção com base em informações lingüísticas (substantivos, adjetivos, nomes próprios, etc) do corpus. Realizado o pré-processamento os termos extraídos de ambas abordagens de pré-processamento foram submetidos à preparação e seleção dos dados, gerando os arquivos de entrada para a ferramenta de aprendizado.

Os algoritmos de aprendizado utilizados neste trabalho para verificar a performance do processo de MT são: Árvores de Decisão e Redes Neurais para a tarefa de Categorização de Textos e *K-means* para a tarefa de agrupamento. Os parâmetros utilizados para a geração de AD são os oferecidos pela ferramenta, para as Redes Neurais foi variada a topologia da rede e o número de neurônios na camada

intermediaria e por fim, no algoritmo *k-means* foi alterado o número de grupos. Estes algoritmos são implementados na ferramenta *Weka*, adotada para realizar o aprendizado dos experimentos.

Para avaliar o desempenho das técnicas de aprendizado adotadas neste trabalho, optou-se por comparar o melhor classificador gerado por cada uma delas levando em conta o erro de classificação no teste.

No próximo capitulo serão apresentados os resultados obtidos com base na metodologia proposta nesta seção.

5 Resultados e Análise dos Experimentos

Neste capítulo serão descritos e analisados os resultados dos experimentos utilizando as abordagens baseadas em palavras e informações lingüísticas na etapa de pré-processamento nas tarefas de categorização e agrupamento no processo de MT.

A metodologia aplicada na concepção, treinamento e avaliação dos resultados para o corpus pode ser vista no capítulo 4.

Nas seções que seguem são mostrados os resultados obtidos utilizando os algoritmos: J48, *Backpropagation* (Categorização) e *K-means* (Agrupamento), comparando as duas abordagens de pré-processamento propostas nesta dissertação.

5.1 Análise dos Resultados para a Categorização de Textos

Nesta seção são apresentados os resultados obtidos no processo de resolução do problema de categorização dos documentos do corpus NILC para cada uma das abordagens de pré-processamento: baseada em palavras e em informações lingüísticas.

5.1.1 Pré-processamento baseado em Métodos Usuais

Nos experimentos realizados com os termos extraídos do pré-processamento usual foram consideradas variações nos seguintes parâmetros: tipo de codificação, número de termos relevantes que compõem o vetor global e tipo de algoritmo de aprendizado.

Em todos os experimentos foi utilizada a categorização múltipla dos exemplos. Para cada exemplo é adotado um vetor global representado pelo modelo de espaço vetorial, utilizando as codificações por freqüência relativa e Tf-Idf, e a seleção de 6, 12, 18, 24 e 30 termos mais freqüentes por categoria, correspondentes ao 30, 60, 90, 120 e 150 termos que representam o vetor global.

Os primeiros experimentos foram realizados com as Árvores de Decisão (ADs) utilizando-se a ferramenta *Weka*. Essa ferramenta re-implementa o algoritmo C4.5, no denominado J48 nesse ambiente.

A média dos resultados percentuais obtidos na geração das árvores, em relação ao erro de classificação, no conjunto de teste, e o número médio de nodos de cada árvore para as variações do corpus (V1, V2 e V3) são mostrados na Tabela 5-1.

Tabela 5-1 Média do Erro de Classificação para as Variações do Corpus

	Número de Termos									
	30		60		90		120		150	
Codificação	Erro	Nodos	Erro	Nodos	Erro	Nodos	Erro	Nodos	Erro	Nodos
F.Relativa	21,64	70	21,99	73	20,47	71	20,35	74	**19,77**	73
TF-IDF	21,75	67	21,17	74	19,88	71	**19,18**	77	20,12	76

Conforme pode ser observado na Tabela 5-1 a variação do número de termos relevantes implicou em alguns casos no aumento e em outros na diminuição do erro de classificação. Pode-se verificar que os melhores resultados são obtidos com 120 e 150 termos, considerando as duas codificações, e que o número médio de nodos de cada experimento é próximo de 70.

Pode-se observar que no processo de categorização de textos utilizando os métodos usuais de pré-processamento, um maior número de termos selecionados resultou um melhor desempenho do categorizador. A Figura 5-1 mostra a F-measure média das categorias: Esporte, Imóveis, Informática, Política e Turismo correspondente ao número de termos por categoria que obtiveram o melhor e pior erro de generalização sobre o conjunto de teste com base na freqüência relativa.

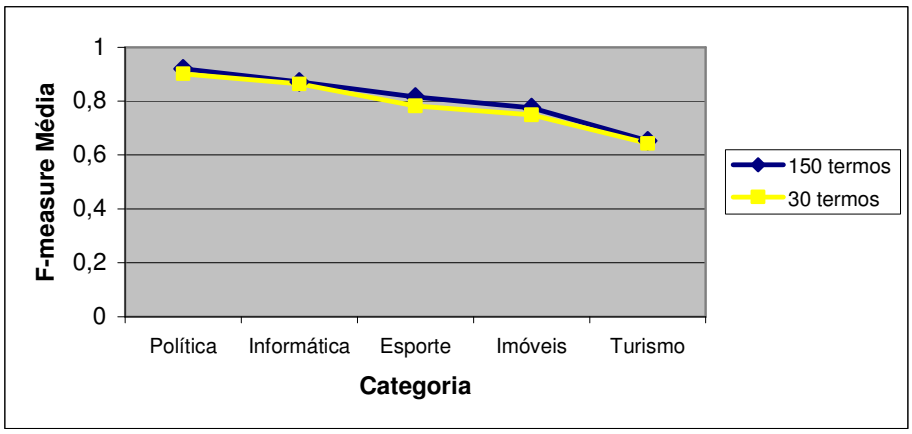

Figura 5-1 F-measure Média utilizando Codificação por Freqüência Relativa

A F-measure média para a codificação Tf-Idf do melhor e pior desempenho do categorizador são mostradas na Figura 5-2.

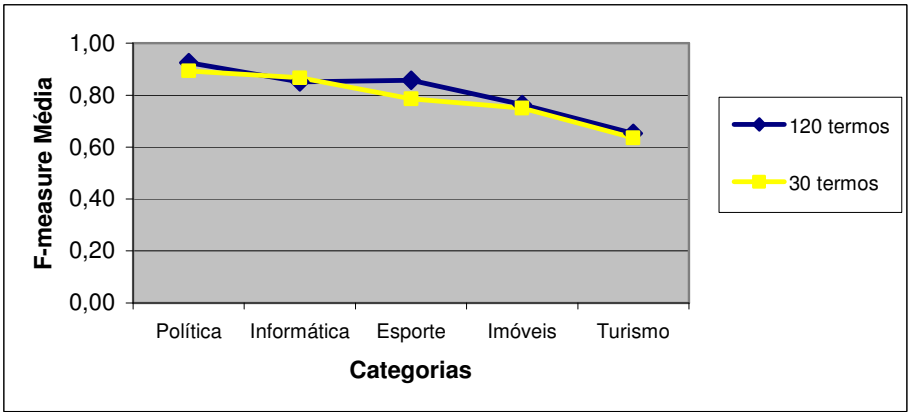

Figura 5-2 F-measure Média utilizando Codificação Tf-Idf

Comparando as melhores taxas de erros obtidas através das codificações por freqüência relativa e Tf-Idf, pode-se verificar na Figura 5-3 que, embora a codificação Tf-Idf apresente uma taxa de erro mais significativa em relação à codificação por freqüência, a precisão dos dados manteve-se semelhante em todos os experimentos realizados com o número de termos selecionados por categoria.

A aplicação das codificações Tf-Idf e freqüência relativa não resultou muita diferença, isso pode ser devido ao fato do corpus não apresentar uma grande variação no tamanho dos documentos.

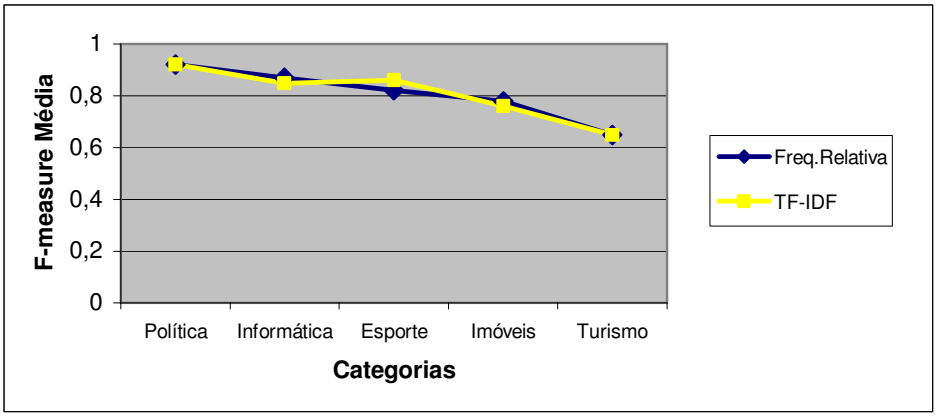

Figura 5-3 F-measure média das Codificações para as melhores Taxas de Erro

Os experimentos seguintes foram conduzidos com as RNAs. Nestes experimentos, assim como nos iniciais, foram variados o número de termos relevantes selecionadas (6, 12, 18, 24 e 30) e utilizada somente a codificação por freqüência relativa. A RNA utilizada foi a MLP, sendo aplicado o algoritmo de aprendizado *Backpropagation* (BP).

As topologias das redes são formadas por 3 ou mais camadas. O número de neurônios na camada de entrada corresponde ao número de termos no vetor global (dependente do número de termos relevantes selecionadas); o número de neurônios da camada intermediária varia conforme o algoritmo de aprendizado utilizado; e o número de neurônios na camada de saída (5) corresponde às categorias: Informática, Imóveis, Esporte, Política e Turismo. A condição de parada do aprendizado foi o número máximo de 3000 épocas, baseado em experimentos anteriores.

Nos experimentos com o algoritmo BP, foi utilizado um valor de 0.9 para o *momentum,* 0.1 para a taxa de aprendizado e variado o número de neurônios na camada intermediária (2, 4, 8 e 16). Para cada topologia, foram realizadas 10 simulações, variando-se a semente aleatória e mantendo-se os demais parâmetros de configuração da rede.

A Tabela 5-2 apresenta os valores médios (para as 10 simulações) dos menores erros de generalização obtidos no processo de aprendizagem para as três variações do corpus.

Tabela 5-2 Média do menor erro de generalização para as variação do corpus

	Número de Termos				
	30	60	90	120	150
Nº de Neurônios	Erro	Erro	Erro	Erro	Erro
2	32,85%	40,22%	40,35%	49,68%	**64,32%**
4	20,81%	16,36%	21,42%	35,43%	49,82%
8	20,26%	16,56%	**14,64%**	32,32%	55,37%
16	20,68%	16,74%	16,19%	24,28%	54,23%

Conforme Tabela 5-2, a menor taxa de erro (14,64%) foi obtida utilizando-se 8 neurônios na camada intermediária, a codificação freqüência relativa e um número de 18 palavras relevantes. Quanto a pior taxa (64,32%), esta foi obtida utilizando-se 30 termos relevantes e 2 neurônios na camada intermediária. Pode-se afirmar que 8

neurônios na camada intermediária são suficientes para a categorização utilizando o pré-processamento baseado métodos usuais, oferecendo taxas de erro, em geral, inferiores as obtidas com o uso de 2, 4 e 16 neurônios. Também pode-se observar que existe uma grande variação no erro de generalização à medida que o número de exemplos de entrada da rede aumenta. Essa variação no erro é decorrente do aumento no número de exemplos de entrada, pois muitas informações semelhantes são apresentadas a rede e esta não consegue distinguir o que pertence a uma classe distinta, classificando muitos padrões de forma errada.

Com base nos resultados obtidos, foi feita uma comparação entre as técnicas. A Tabela 5-3 apresenta os parâmetros utilizados na obtenção das menores taxas de erros. Os valores considerados para o algoritmo BP foram os obtidos com o uso de 8 neurônios na camada intermediária.

Tabela 5-3 Comparação da Média do menor Erro de Teste

Técnica	Taxa	Codificação	Nº Termos
Árvores de Decisão	19,18%	Tf-Idf	24
RNA *Backpropagation*	14,64%	Freqüência Relativa	18

Conforme comparação entre as técnicas de aprendizado mostrado na Tabela 5-1 e 5-2, as melhores taxas de classificação para o pré-processamento baseado em métodos usuais, foram obtidas com o algoritmo RNA MLP-BP. As taxas de erro das RNAs foram obtidas a partir do ajuste fino de seus parâmetros (com base em experimentos preliminares). Assim, observou-se uma maior dependência das RNAs aos seus parâmetros do que a apresentada pelas Árvores de Decisão. Ainda, o número de experimentos aplicado as RNAs foi superior ao necessário para as Árvores de Decisão. Por fim, a codificação freqüência relativa implicou nas melhores taxas de classificação e precisão no desempenho do categorizador, e foram adotadas para realização dos experimentos utilizando o pré-processamento baseado em informações lingüísticas.

Na seção que segue será apresentado o resultado dos classificadores para os experimentos realizados com informações lingüísticas na etapa de pré-processamento.

5.1.2 Pré-processamento baseado em Informações Lingüísticas

Os experimentos realizados com a extração de informações lingüísticas na etapa de pré-processamento seguem as mesmas variações do experimento anterior. Para os

parâmetros: número de termos relevantes selecionados por categoria e tipo de algoritmo de aprendizado.

A classificação múltipla foi adotada em todos os experimentos, assim como as codificação por freqüência relativa para a seleção dos 30, 60, 90, 120 e 150 termos que constituem o vetor global

Os primeiros experimentos foram feitos com as árvores de decisão utilizando-se a estrutura gramatical substantivo. A média dos resultados percentuais obtidos na geração das árvores, em relação ao erro médio de classificação, no conjunto de teste (erro na generalização), e os nodos médios para as variações do corpus são mostrados na Tabela 5-4.

Tabela 5-4 Média da Taxa de Erro para a estrutura gramatical: substantivos

	Número de Termos				
	30	60	90	120	150
Erro	24,91%	**21,75%**	23,98%	23,51%	22,69%
Nodos	64,33	77,67	71,67	75	79,67

Conforme pode ser observado na Tabela 5-4, os melhores resultados foram obtidos com 60 termos mais relevantes, obtendo uma taxa de erro de 21,75%, enquanto que o pior é para os 30 termos mais relevantes, com uma taxa de erro de 24,91%. A Figura 5-4 mostra a capacidade de generalização da árvore gerada para as categorias: Política, Informática, Esporte, Imóveis e Turismo.

Analisando a generalização do conjunto de teste, pode-se verificar que durante o treinamento os melhores resultados foram obtidas para as categorias Política e Informática, permitindo um bom desempenho no conjunto de dados. Os piores resultados estão relacionados às categorias Imóveis e Turismo. A conseqüência da baixa precisão nestas categorias, esta relacionada à similaridade dos termos, por exemplo, os substantivos *apartamento*, *cidade*, *preço*, *casa* apresentam uma alta ocorrência em ambas categorias.

Além disso, comparando os resultados obtidos com os 60 termos mais relevantes para os termos extraídos utilizando ambos pré-processamentos: baseados em palavras (21,17%) e em informações lingüísticas (21,75%), pode-se verificar que não há um aumento ou diminuição significativa no erro de generalização. Assim, com o intuito de identificar uma combinação gramatical que apresente um melhor

desempenho em relação aos métodos usuais foram adicionados aos substantivos os nomes próprios encontrados no corpus.

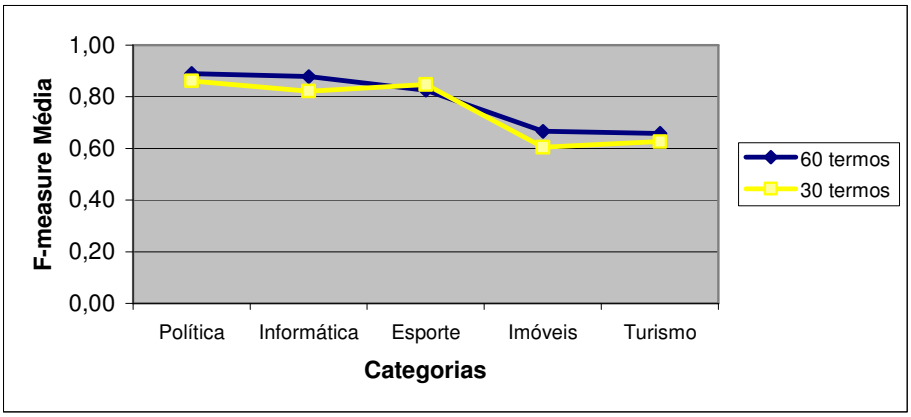

Figura 5-4 F-measure Média das Categorias

Os resultados percentuais dos erros médios de classificação obtidos na geração das árvores, e a média dos nodos (para as versões V1, V2 e V3 do corpus) baseados nas combinações gramaticais: substantivos- nomes próprios são mostrados na Tabela 5-5.

Tabela 5-5 Média da Taxa de Erro para estrutura: substantivos-nomes próprios

	Número de Termos				
	30	60	90	120	150
Erro	24,09%	24,56%	22,80%	**22,45%**	22,80%
Nodos	57	81	67	70,33	71,06

Nesse treinamento os resultados obtidos através das combinações gramaticais: substantivos – nomes próprios apresentaram um desempenho melhor no erro de generalização (22,45%) em relação à estrutura gramatical substantivos (23,50%). Porém, comparando a taxa de erro aos termos extraídos do pré-processamento baseado em métodos usuais (20,35%), para os 120 termos mais relevantes, houve um aumento significante no erro de generalização.

No entanto, aplicando uma nova combinação gramatical baseada em substantivos-adjetivos (Tabela 5-6), pode-se verificar melhores resultados com os 60 termos mais relevantes, e piores resultados com 150 termos, com taxas de erro de generalização de 18,01% e 23,15%, respectivamente para as variações do corpus.

Tabela 5-6 Média da Taxa de Erro no Teste para estrutura: substantivos - adjetivos

	Número de Termos				
	30	60	90	120	150
Erro	23,15%	20,35%	**18,01%**	19,18%	18,71%
Nodos	59	61,66	61	71,66	65,66

A comparação dos resultados obtidos para a melhor taxa de erro através dos erros de generalização no pré-processamento baseado em combinações gramaticais de acordo com a seleção do número de termos por categoria, são mostrados na Figura 5-5.

Portanto, conforme análise dos resultados na Figura 5-5, a extração da combinação gramatical substantivos-adjetivos na etapa de pré-processamento contempla melhor desempenho no categorizador em relação às outras combinações. As menores taxas de erro de generalização no conjunto de teste para geração de árvores de decisão são obtidas na seleção dos 60, 120 e 150 termos mais relevantes. A Abrangência e Precisão dos termos que melhor caracterizam as categorias na combinação gramatical substantivos-adjetivos são mostrados na Figura 5-6. Novamente, as categorias Imóveis e Turismo, são as que apresentam piores resultados.

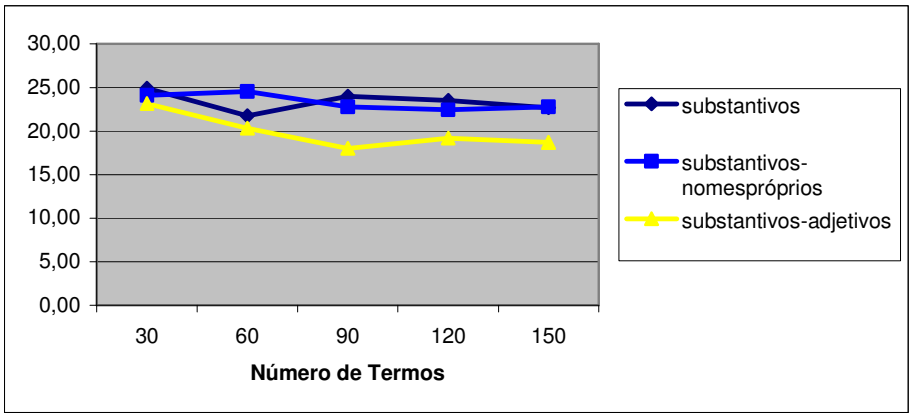

Figura 5-5 Comparação entre as Taxas de Erro das Estruturas Gramaticais

Outras combinações gramaticais utilizadas na realização dos experimentos são baseadas em: substantivos-nomes próprios-adjetivos e nomes próprios-adjetivos. Os resultados obtidos para essas combinações são mostrados nas Tabelas 5-7 e 5-8.

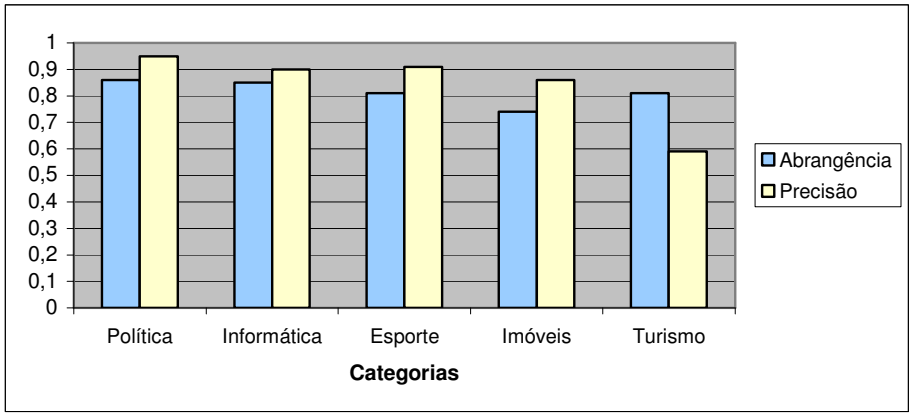

Figura 5-6 Abrangência e Precisão na estrutura: substantivos-adjetivos

Tabela 5-7 Média da Taxa de Erro para: substantivos-nomes próprios-adjetivos

	Número de Termos				
	30'	60	90	120	150
Erro	**20,82%**	22,92%	20,94%	21,05%	21,17%
Nodos	57	71,67	67,67	69	69,67

Tabela 5-8 Média da Taxa de Erro para: nomes próprios-adjetivos

	Número de Termos				
	30	60	90	120	150
Erro	**47,01%**	46,34%	32,51%	33,21%	32,86%
Nodos	97,66	106,33	125	136,33	126,33

Comparando a melhor taxa de erro de generalização da Tabela 5-7 em relação às outras estruturas gramaticais extraídas no pré-processamento baseado em informações lingüísticas e baseadas em palavras para a seleção dos 30 termos mais relevantes, pode-se verificar que esta categoria gramatical apresenta um desempenho superior no categorizador. No entanto, a extração de termos baseados nas combinações nomes próprios - adjetivos (Tabela 5-8), apresentaram as piores taxas de erro (47,01%) em comparação aos demais experimentos. A alta taxa de erro é decorrente da distribuição dos nomes próprios no corpus. Em um conjunto de documentos os nomes próprios que fazem parte do conjunto de treinamento necessariamente não são os mesmos presentes em um conjunto de testes. Ocasionando uma grande diferença nos dados de aprendizado.

Por fim, cabe ressaltar que as informações lingüísticas mais relevantes para categorização de documentos utilizando ADs são as baseadas nas combinações: substantivos - adjetivos para a seleção dos 60, 120 e 150 termos mais relevantes e substantivos – nomes próprios - adjetivos para a seleção dos 30 termos mais relevantes. Enquanto que as combinações gramaticais: substantivos e substantivos – nomes próprios apresentam os resultados semelhantes ou piores aos métodos usuais de pré-processamento, conforme mostra a Figura 5-7.

As mesmas combinações gramaticais adotadas na geração de Árvores de Decisão foram utilizados nos experimentos realizados com RNAs, variando o número de termos relevantes selecionadas (30, 60, 90, 120 e 150) e o tipo de codificação (freqüência relativa). A RNA utilizada foi a MLP, sendo aplicado o algoritmo de aprendizado *Backpropagation* (BP).

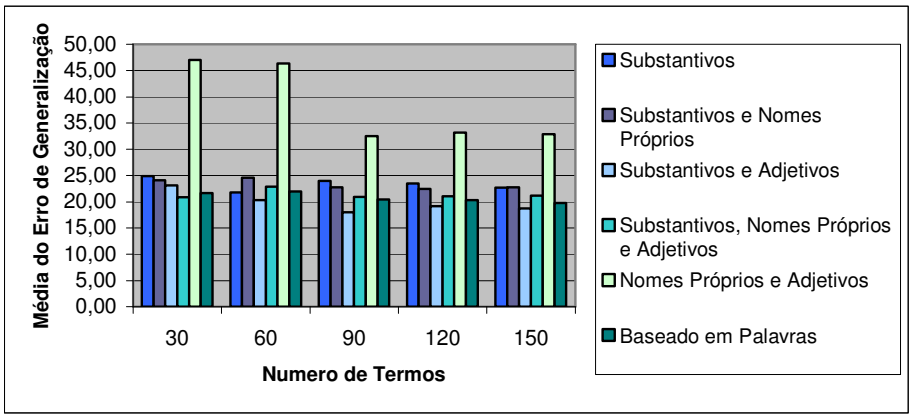

Figura 5-7 Comparação do Erro de Generalização para os Pré-processamentos

As topologias das redes são formadas por 3 ou mais camadas. O número de neurônios na camada de entrada corresponde ao número de termos no vetor global, o número de neurônios da camada intermediária varia conforme o algoritmo de aprendizado utilizado; e o número de neurônios na camada de saída (5) corresponde as categorias: Informática, Imóveis, Esporte, Política e Turismo. A condição de parada do aprendizado foi o número máximo de 3000 épocas.

Os parâmetros utilizados no algoritmo BP foram: 0.9 para o *momentum,* 0.1 para a taxa de aprendizado e variado o número de neurônios na camada intermediária (2, 4, 8 e 16). Para cada topologia, foram realizadas 10 simulações, variando-se a semente aleatória e mantendo-se os demais parâmetros de configuração da rede.

O primeiro experimento realizado no processo de aprendizagem para as três variações do corpus foram baseados na combinação gramatical substantivos. A Tabela 5-8 apresenta os valores médios (para as 10 simulações) dos menores erros de generalização correspondente ao número de neurônios na camada escondida.

Conforme Tabela 5-9, a menor taxa de erro (19,22%) foi obtida utilizando-se 16 neurônios na camada escondida, com a codificação por freqüência relativa e um número de 60 termos mais relevantes. Quanto a pior taxa (69%), esta foi obtida utilizando-se 150 termos relevantes e 2 neurônios na camada intermediaria. Pode-se afirmar que 16 neurônios na camada intermediária foram melhores para os substantivos extraídos, oferecendo taxas de erro, em geral inferiores às obtidas com o uso de 2, 4 e 8. Porém, comparando-se a taxa de erro (19,22%) obtida no treinamento de termos baseados em substantivos com os resultados obtidos do treinamento utilizando os métodos tradicionais de pré-processamento (16,74%), não obteve-se uma redução significativa no erro para o mesmo número de termos selecionados (60).

Tabela 5-9 Média do menor erro de generalização para: substantivos

	Número de Termos				
	30	60	90	120	150
Nº de Neurônios	Erro	Erro	Erro	Erro	Erro
2	32,06%	43,27%	49,84%	63,84%	68,40%
4	21,26%	21,53%	30,96%	57,30%	69%
8	22,85%	19,49%	21,54%	48,01%	**71,13%**
16	23,09%	**19,22%**	20,27%	54,38%	67,54%

Assim, cabe ressaltar que no treinamento de um corpus de pequeno porte para um número médio de 60 atributos, 16 neurônios na camada intermediária foram suficientes para a categorização, utilizando os métodos tradicionais de pré-processamento.

Com base nesses resultados, uma nova estrutura gramatical substantivos-nomes próprios foi apresentada ao processo de aprendizagem. Os resultados percentuais médios obtidos (10 simulações) são mostrados na Tabela 5-10. Analisando estes resultados, pode-se verificar que a menor taxa de erro (19,01%) foi obtida utilizando-se 8 neurônios na camada intermediária e um número de 90 termos relevantes.Quanto a

pior taxa (70,06%) foi obtida utilizando 150 termos mais relevantes para 2 neurônios na camada intermediária.

Tabela 5-10 Média da Taxa de Erro para: substantivos–nomes próprios

	Número de Termos				
	30	60	90	120	150
Nº de Neurônios	Erro	Erro	Erro	Erro	Erro
2	35,06%	42,36%	46,32%	60,49%	**70,06%**
4	23,15%	19,91%	28,02%	47,51%	67,42%
8	22,27%	19,18%	21,62%	49,15%	65,78%
16	22,94%	19,77%	**19,01%**	38,82%	66,58%

Comparando as taxas de erros no processo de aprendizagem para as estruturas gramaticais mostradas nas Tabelas 5-9 e 5-10, pode-se verificar que adicionando os nomes próprios ao treinamento obteve-se uma taxa de erro semelhante. No entanto, comparando-se aos métodos tradicionais de pré-processamento a estrutura gramatical substantivos - nomes próprios apresenta uma taxa de erro superior, tornando-se o método tradicional mais apropriado.

Utilizando as combinações gramaticais, substantivos-adjetivos no treinamento de uma rede neural (Tabela 5-11), pode-se observar que os resultados obtidos foram os mesmos do experimento anterior. A melhor taxa de erro (19,05%) foi obtida para os 90 termos mais relevantes e 16 neurônios na camada intermediária, enquanto que a pior taxa (66,43%) permaneceu entre os 150 termos mais relevantes e 2 neurônios na camada intermediária. Porém, se comparado à estrutura substantivos, para o mesmo número de termos relevantes, observa-se uma redução na taxa de erro. Assim, pode-se concluir que para o treinamento de uma rede neural baseada em termos extraídos de informações lingüísticas a inclusão dos *adjetivos* apresenta um melhor resultado. Mas ainda não se igualam aos resultados obtidos pelos termos extraídos do pré-processamento original.

As combinações gramaticais substantivo-nomespróprios-adjetivos apresentaram uma taxa de erro (16,96%) para os 60 termos mais relevantes e 16 neurônios na camada intermediária. Enquanto que a pior taxa de erro (67,17%) encontra-se entre os 150 termos mais freqüentes e 2 neurônios na camada intermediaria, conforme mostra a Tabela 5-12.

Tabela 5-11 Média da Taxa de Erro para: substantivos-adjetivos

	Número de Termos				
	30	60	90	120	150
Nº de Neurônios	Erro	Erro	Erro	Erro	Erro
2	34,33%	44,49%	49,93%	54,31%	**66,43%**
4	22,01%	20,18%	28,19%	43,55%	59,03%
8	21,86%	19,66%	21,59%	34,82%	56,38%
16	22,56%	19,18%	**19,05%**	32,35%	56,52%

Tabela 5-12 Média da Taxa de Erro para: substantivos-nomespróprios–adjetivos

	Número de Termos				
	30	60	90	120	150
Nº de Neurônios	Erro	Erro	Erro	Erro	Erro
2	33,22%	45,06%	45,42%	53,03%	**67,17%**
4	21,91%	19,51%	28,60%	40,63%	59,53%
8	21,39%	18%	17,55%	33,29%	56,46%
16	22,44%	**16,96%**	18,27%	37,98%	54,21%

Analisando os resultados percentuais médios obtidos para a estrutura gramatical mostrada na Tabela 5-12, observou-se que esta apresenta uma taxa de erro (16,96%) muito próxima à encontrada no processo de extração de termos tradicional (16,74%). E em comparação com as taxas de erro obtidas nas demais combinações, a inclusão dos nomes próprios - adjetivos aos substantivos, mostrou-se com o melhor desempenho no processo de aprendizagem para o uso de informações lingüísticas nas etapas de pré-processamento.

Por fim, as combinações gramaticais: nomes próprios - adjetivos foram as taxas de erro que apresentaram os piores resultados se comparados as demais estruturas. A menor taxa de (35,57%) foi obtida utilizando os 90 termos mais relevantes e 8 neurônios na camada intermediária, e a maior taxa de erro (74,04%) para os 150 termos mais relevantes utilizando no processo de aprendizagem 2 neurônios na camada intermediária.

Com base nos resultados obtidos na geração de ADs e no processo de aprendizagem de uma RNA com o algoritmo BP através do uso de informações lingüísticas, foi feito um comparativo entre as técnicas. A Tabela 5-13 apresenta a

comparação entre as técnicas de aprendizado utilizadas no treinamento dos classificadores utilizando informações lingüísticas. Os valores considerados para o algoritmo BP foram os obtidos com o uso de 16 neurônios na camada intermediária correspondente a estrutura gramatical substantivos-nomes próprios-adjetivos.

Conforme comparação entre as técnicas de aprendizado mostrado para as etapas de pré-processamento baseadas na extração de informações lingüísticas, as melhores taxas de classificação foram obtidas com o algoritmo RNA MLP-BP. As taxas de erro das RNAs foram obtidas a partir de um ajuste fino de seus parâmetros, observando uma dependência maior aos seus parâmetros do que a apresentada pelas Árvores de Decisão.

Tabela 5-13 Comparação Média do Menor Erro de Teste

Técnica	Taxa	Codificação	Nº Termos
Árvores de Decisão	18,01%	Freqüência Relativa	90
RNA *Backpropagation*	**16,96%**	Freqüência Relativa	90

A Figura 5-8 apresenta a média dos menores erros de generalização, o número de termos selecionados por categoria e o número de neurônios na camada intermediária para o treinamento das redes MLP – BP utilizando as estruturas gramaticais adotadas nos experimentos.

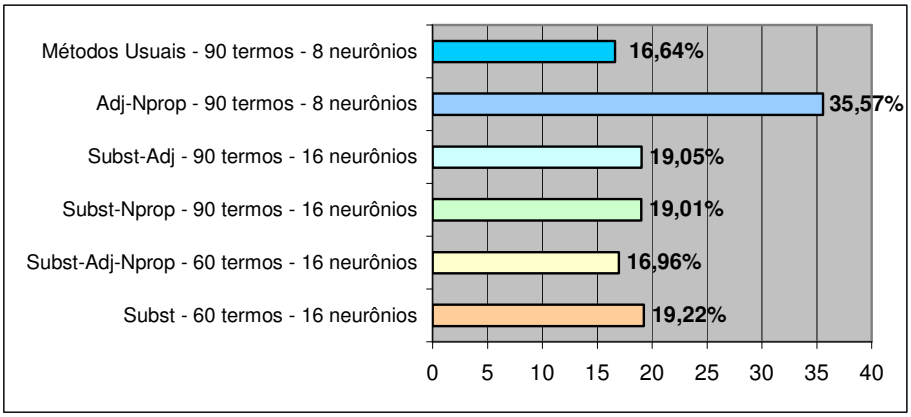

Figura 5-8 Média das Menores Taxas de Erro para RNAs

Quanto ao tempo de treinamento dos classificadores para as técnicas de aprendizado adotadas, o treinamento da Rede MLP resultou em aproximadamente 1068 segundos para a menor taxa de erro da seleção dos termos substantivos–nomes próprios -adjetivos, e para a geração das árvores de decisão como o algoritmo J48 obteve-se uma média de 92 segundos para a menor taxa de erro.

Na seção que segue será apresentado os resultados obtidos nos experimentos realizados com o agrupamento dos textos.

5.2 Análise dos Resultados para o Agrupamento de Textos

Nesta seção são apresentados os resultados obtidos no agrupamento de textos do corpus NILC para cada uma das abordagens de pré-processamento: baseada em métodos usuais e em informações lingüísticas, utilizando o algoritmo de agrupamento *K-means*.

5.2.1 Pré-processamento baseado em Métodos Usuais

Nos experimentos realizados com os termos extraídos do pré-processamento usual foram consideradas variações nos seguintes parâmetros: número de termos relevantes, variação do corpus, o tipo de codificação e o algoritmo de agrupamento *K-means*.

Diversos experimentos foram feitos adotando-se um vetor global representado pelo modelo de espaço vetorial, utilizando a codificação por freqüência relativa, e a seleção dos 30, 90 e 150 termos relevantes para o conjunto de treinamento. Essa seleção foi baseada na seleção dos 6, 18 e 30 termos mais relevantes por categoria utilizados nos experimentos de categorização. Para a obtenção dos termos relevantes foram utilizados três conjuntos diferenciados de treino (2/3 dos documentos) e teste (1/3 de documentos). Essas variações resultaram em 9 experimentos de agrupamento, utilizando os parâmetros: semente randômica igual a 10 e o número grupos igual a 5. Desses 9 experimentos em apenas um dos casos foi possível identificar dois grupos com mais clareza. A Tabela 5-14 apresenta a matriz de confusão para esse resultado obtido com os 150 termos relevantes e versão V2 do corpus.

Tabela 5-14 Matriz de Confusão para a versão V2 do corpus e 150 termos

	Cluster 0	Cluster 1	Cluster2	Cluster 3	Cluster 4
Esporte	1	**31**	2	0	23
Imóveis	2	0	4	0	51
Informática	0	0	1	0	55
Política	0	0	2	**39**	16
Turismo	5	0	17	0	33

Com base nos resultados obtidos para o agrupamento utilizando os métodos tradicionais de pré-processamento, pode-se verificar que adotando o algoritmo *k-means*

com 5 grupos apenas 2 foram identificados. Analisando os grupos formados, observou-se que os documentos pertencem às seções Esporte e Política do corpus jornalístico.

Devido aos resultados dos experimentos de categorização, onde as categorias Turismo e Imóveis apresentaram sempre as piores medidas para as taxas de precisão dos termos que as caracterizavam, optou-se por eliminar as seções Imóveis e Turismo do conjunto de treino e teste (restando 342 e 171 documentos em cada conjunto, respectivamente), refazendo os experimentos com os mesmos parâmetros utilizados no grupo identificado anteriormente: semente randômica igual a 10, número de grupos igual a 5 e selecionando os 150 termos relevantes.

Inicialmente, a semente randômica foi salva e aplicada na inicialização dos demais experimentos variando o corpus no treinamento. Os resultados obtidos para o agrupamento das versões V1, V2 e V3 do corpus no conjunto de teste são mostrados na Tabela 5-15.

Tabela 5-15 Matriz de Confusão para as versões V1, V2 e V3 do Corpus

Corpus	Seções	Grupo0	Grupo 1	Grupo 2	Grupo 3	Grupo 4
V1	Esporte	0	1	0	**55**	1
	Informática	0	**55**	0	2	0
	Política	0	5	0	3	**49**
V2	Esporte	0	0	0	**57**	0
	Informática	0	0	**54**	3	0
	Política	**50**	1	4	2	0
V3	Esporte	**57**	0	0	0	0
	Informática	7	0	**50**	0	0
	Política	**43**	14	0	0	0

Conforme pode ser analisado na Tabela 5-15, três grupos distintos foram formados, para as versões V1 e V2 do corpus. Como são variados os documentos que pertencem às versões do corpus, pode-se verificar que os termos selecionados para o agrupamento na versão V3 não foram significativos a ponto de conseguir identificar os grupos Esporte e Política.

As Figuras 5-9, 5-10 e 5-11 mostram as taxas de Abrangência e Precisão para os grupos formados com base nas versões do corpus (V1, V2 e V3).

Nestas figuras, as medidas de precisão são maiores para os assuntos referentes à Política nas versões V1 e V2 do corpus, mostrando que os documentos que pertencem a esse grupo podem ser melhor identificados. Por outro lado, na versão V3 do corpus, a

precisão maior dos grupos foi obtida para os grupos de Informática e Política que permaneceram juntos.

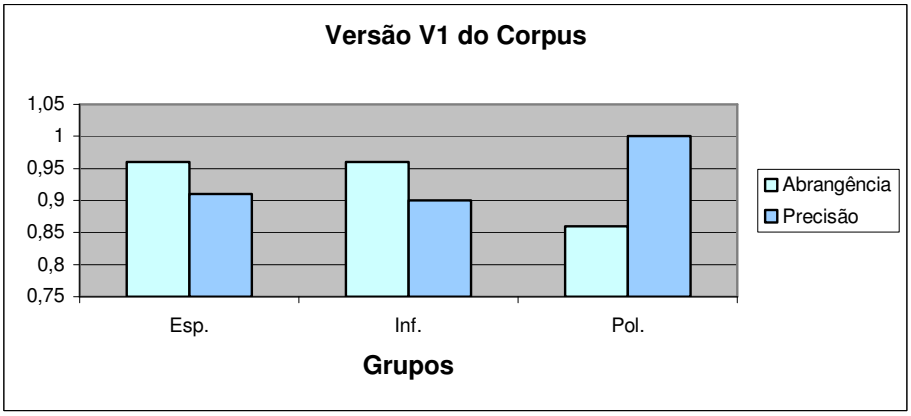

Figura 5-9 Abrangência e Precisão dos Grupos para versão V1

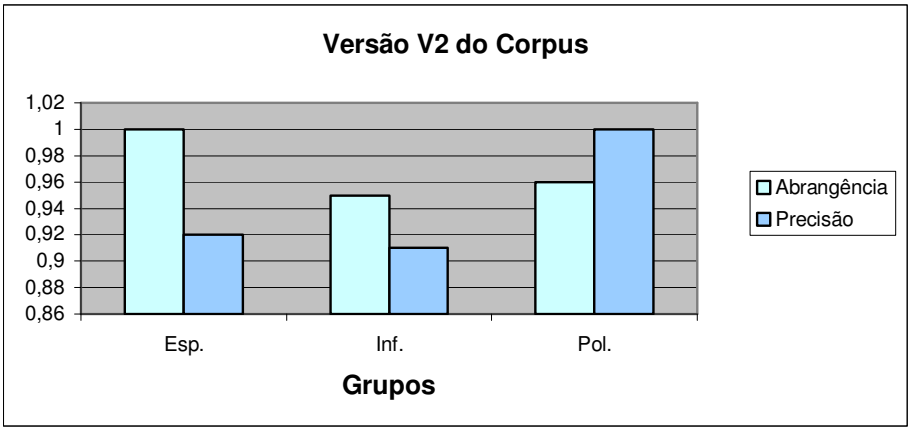

Figura 5-10 Abrangência e Precisão dos Grupos para versão V2

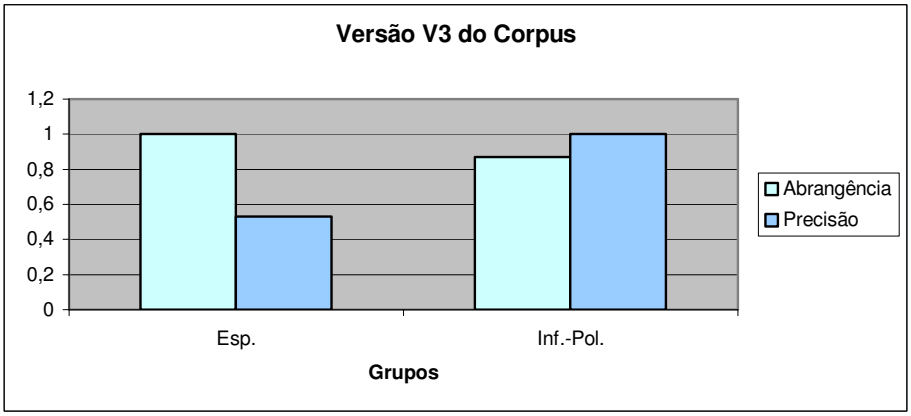

Figura 5-11 Abrangência e Precisão dos Grupos para versão V3

Analisando os agrupamentos obtidos utilizando as etapas tradicionais de pré-processamento, pode-se verificar que não foi possível obter os 5 grupos correspondentes às seções escolhidas no corpus jornalístico. Isso pode ser decorrente da similaridade dos termos que representam o conteúdo dos assuntos referentes a imóveis e turismo. No entanto, com a eliminação destas seções, foi possível identificar 3 grupos distintos utilizando os mesmos parâmetros adotados nos primeiros experimentos.

Na seção que segue novos experimentos baseados em combinações gramaticais usando informações lingüísticas são apresentados.

5.2.2 Pré-processamento baseado em Informações Lingüísticas

Para os experimentos realizados com extrações de informações lingüísticas no processo de agrupamento de textos, foram adotados os parâmetros: número de grupos igual a 5 e semente randômica igual a 10.

A realização dos experimentos contempla a construção de um vetor global representado pelo modelo de espaço vetorial e codificação por freqüência relativa com base nos 30, 90 e 150 termos mais relevantes no conjunto de treino. Três conjuntos diferenciados de treino e teste foram aplicados aos experimentos, permitindo verificar a variação dos grupos em diferentes distribuições.

Os primeiros experimentos realizados para o agrupamento foram baseados na seleção dos substantivos. Novamente 9 experimentos foram realizados, objetivando encontrar cinco grupos de documentos correspondentes aos assuntos de Esporte, Imóveis, Informática, Política e Turismo. No entanto, apenas dois grupos referentes aos assuntos de Política e Turismo foram identificados nas versões V1 e V2 do corpus para o número de 150 termos selecionados.

Com base nesses resultados, uma nova combinação gramatical foi testada. Esta nova combinação equivale à seleção dos termos sintáticos: substantivos–nomes próprios. O resultado obtido para este agrupamento, é mostrado nas Tabela 5-16 e 5-17.

Analisando os grupos formados nas Tabelas 5-16 e 5-17, pode-se verificar que os termos sintáticos substantivos–nomes próprios permitem identificar 3 grupos distintos para duas seleções de termos (90 e 150) na versão do corpus V2. Assim, pode-se concluir que esta estrutura apresentou melhores grupos se comparado aos substantivos. No entanto, podemos observar que os termos correspondentes aos

assuntos de Imóveis e Turismo apresentam uma similaridade alta, ocasionando a não identificação das classes pelos grupos.

Tabela 5-16 Matriz de Confusão da versão V2 do Corpus para os 90 termos

	Grupo 0	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Esporte	26	**31**	0	0	0
Imóveis	43	0	13	0	1
Informática	25	0	0	0	**32**
Política	21	0	0	**36**	0
Turismo	57	0	0		0

Tabela 5-17 Matriz de Confusão da versão V2 do Corpus para os 150 termos

	Grupo 0	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Esporte	0	**38**	19	0	0
Imóveis	11	0	44	1	1
Informática	0	0	19	0	**38**
Política	0	1	20	**36**	0
Turismo	0	0	57	0	0

Para a combinação gramatical: substantivos-adjetivos, 3 grupos distintos foram formados na versão V1 do corpus para os 150 termos mais relevantes do conjunto de treino. No entanto, a adição dos adjetivos aos substantivos, mostrou na Tabela 5-18 que os documentos pertencentes aos assuntos de Imóveis e Esporte foram identificados como um grupo, ao contrário de Informática, Política e Turismo que permaneceram juntos. Com base na matriz de similaridade gerada após a construção do vetor global, é possível verificar que muitos dos termos selecionados estão presentes em mais de um assunto, resultando assim na má identificação dos grupos.

Tabela 5-18 Matriz de Confusão da versão V1 do Corpus para os 150 termos

	Grupo 0	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Esporte	**41**	16	0	0	0
Imóveis	0	**56**	0	1	0
Informática	6	0	40	2	9
Política	4	0	20	1	**25**
Turismo	2	0	55	0	0

A Tabela 5-19 mostra o agrupamento resultante da combinação gramatical substantivos-nomes próprios-adjetivos para os 3 grupos formados na versão V1 do corpus e 150 termos relevantes selecionados.

Tabela 5-19 Matriz de Confusão para os 150 termos da versãoV1

	Grupo 0	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Esporte	16	0	0	**41**	0
Imóveis	54	0	2	0	1
Informática	22	0	**34**	1	0
Política	16	0	0	1	**40**
Turismo	57	0	0	0	0

Considerando os resultados obtidos com as demais combinações gramaticais, é possível verificar que para os 5 grupos selecionados no treinamento do algoritmo *K-means*, no máximo 3 grupos são formados, sendo que estes correspondem aos conteúdos de Esporte, Política e Informática. Isto é decorrente da similaridade dos conteúdos existentes nos documentos referentes a Imóveis e Turismo.

A última combinação gramatical testada para o agrupamento é baseada na seleção dos termos sintáticos nomes próprios-adjetivos. Para este treinamento, foi possível identificar apenas um grupo distinto, sendo este referente aos assuntos de Informática. Cabe ressaltar que, a má formação do agrupamento utilizando essa estrutura é decorrente da distribuição de nomes próprios nos conjuntos de treino e teste. Os resultados obtidos na categorização de textos com esta estrutura apresentaram altos erros, devido a não similaridade entre os nomes próprios encontrados no texto.

Outros experimentos variando o número de grupos para 6, 7 e 8 foram realizados, com base nas mesmas estruturas apresentadas. No entanto, mesmo aumentando a variação do número de grupos, os grupos formados pertenceram aos assuntos de Esporte, Informática e Política.

Assim, como o objetivo deste trabalho não é analisar a performance do algoritmo de agrupamento e sim verificar quais etapas de pré-processamento formam o número de grupos desejados, optou-se por reduzir o número de documentos, eliminando as seções Imóveis e Turismo a fim de trabalhar melhor com as variações e experimentos.

Os novos experimentos são realizados sem as seções Imóveis e Turismo do conjunto de treino e teste do corpus, com base na seleção dos 150 termos mais relevantes, devido a este apresentar na maioria dos casos o maior número de grupos identificados.

Para a realização dos experimentos, inicialmente a semente randômica foi salva e aplicada na inicialização dos demais experimentos variando o corpus no treinamento. Os resultados obtidos para o agrupamento da versões V1, V2 e V3 do corpus no conjunto de teste para a seleção dos termos sintáticos: substantivos–nomes próprios-adjetivos são mostrados na Tabela 5-20. De acordo com os resultados obtidos na Tabela 5-20, pode-se verificar que a versão V1 do corpus foi a que melhor agrupou os documentos. Assim, pode-se concluir que os termos mais representativos para esses experimentos utilizando esta estrutura gramatical estão presentes na primeira variação do corpus.

As medidas de desempenho Abrangência e Precisão são mostradas nas Figuras 5-12, 5-13 e 5-14 para os grupos formados com base nas versões do corpus (V1, V2 e V3).

Tabela 5-20 Matriz de Confusão para as versões V1, V2 e V3 do Corpus

Corpus	Seções	Grupo 0	Grupo 1	Grupo 2	Grupo 3	Grupo 4
V1	Esporte	1	3	**53**	0	0
	Informática	**54**	0	2	1	0
	Política	3	0	4	0	**50**
V2	Esporte	0	1	1	**55**	0
	Informática	0	0	**54**	3	0
	Política	0	0	**54**	3	0
V3	Esporte	0	**48**	9	0	0
	Informática	0	1	**56**	0	0
	Política	9	0	27	0	**21**

.

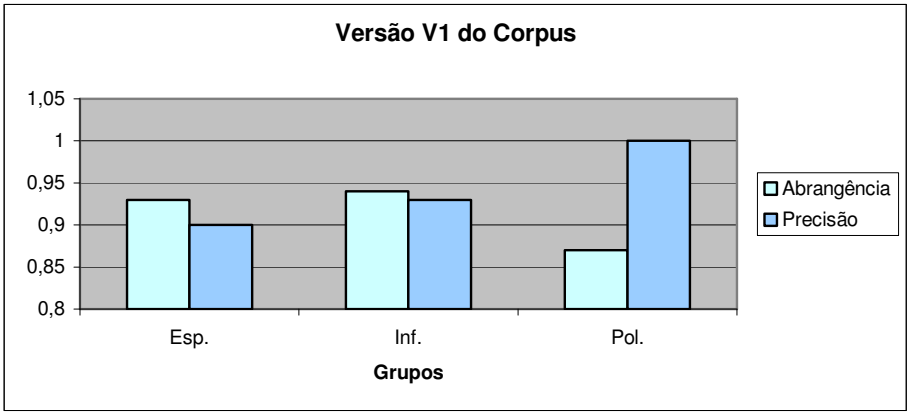

Figura 5-12 Abrangência e Precisão dos Grupos para versão V1

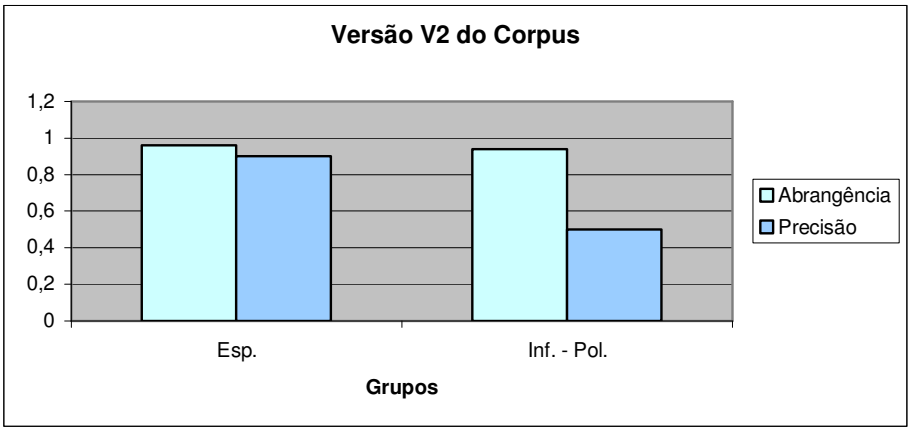

Figura 5-13 Abrangência e Precisão dos Grupos para versão V2

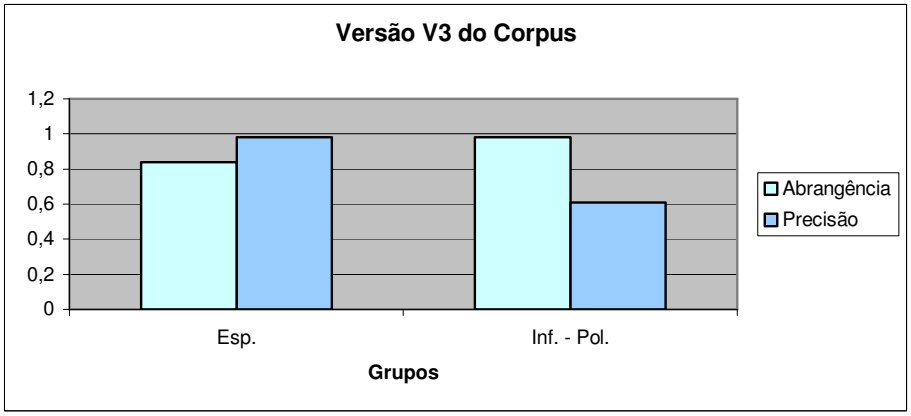

Figura 5-14 Abrangência e Precisão dos Grupos para versão V3

Analisando os resultados obtidos através das medidas de precisão, pode-se salientar que para a seleção dos termos baseados em substantivos-nomespróprios-adjetivos, foram os termos mais representativos para a formação dos grupos nas versões V1 e V3 do corpus são referentes aos assuntos Política e Esporte, respectivamente.

Por fim, pode-se verificar que o uso do algoritmo *K-means* com base na seleção de informações lingüísticas para a formação dos 5 grupos utilizando todas as seções do corpus mostrou-se mais significativa. Por outro lado, os resultados obtidos no agrupamento das seções Esporte, Informática e Política utilizando informações lingüísticas em comparação aos métodos usuais de pré-processamento para a seleção dos termos, mostrou-se semelhante. Pode-se concluir que, a utilização de somente nomes próprios e substantivos não são suficientes para um bom desempenho no agrupamento em comparação as combinações de aos termos sintáticos substantivos-nomes próprios-adjetivos.

Cabe ressaltar, que duas sementes aleatórias foram aplicadas aos experimentos, uma para a formação dos grupos baseados na seleção de termos através de informações lingüísticas e outra para os experimentos baseados nos métodos tradicionais de agrupamento.

Na seção que segue, as considerações finais a respeito dos experimentos realizados no processo de categorização e agrupamento de textos são apresentadas.

5.3 Considerações Finais

Os experimentos realizados neste capítulo utilizaram técnicas de aprendizado simbólica e conexionista para as tarefas de categorização, e o algoritmo *K-means* para o agrupamento de textos, comparando os termos extraídos das etapas de pré-processamento baseado em métodos usuais e em informações lingüísticas.

Os resultados obtidos para o pré-processamento baseado em métodos usuais no processo de categorização de textos do corpus NILC para a geração de ADs, mostrou que para um corpus de pequeno porte, quanto maior o número de atributos menor o erro de generalização no conjunto de teste (19,18%), independentemente da codificação adotada: freqüência relativa e Tf-Idf. Porém, comparado ao resultado obtido com o algoritmo RNA MLP-BP, este apresentou a menor taxa de erro (14,64%) obtida utilizando-se 8 neurônios na camada intermediária, a codificação freqüência relativa e um número de 18 palavras relevantes. Por outro lado, os experimentos realizados com o pré-processamento baseado em informações lingüísticas para geração das ADs as estruturas gramaticais que obtiveram como resultados percentuais as melhores taxas de erro em comparação aos métodos tradicionais de pré-processamento foram: substantivos-adjetivos e substantivos-nomes próprios-adjetivos.

Quanto ao uso de RNAs no treinamento dos termos selecionados do pré-processamento de informações lingüísticas, estes obtiveram novamente uma taxa de erro de generalização menor do que ADs. Mas, se comparadas aos métodos tradicionais baseados em palavras, o treinamento das combinações gramaticais substantivos-nomes próprios-adjetivos são similares para os 60 termos mais relevantes, com um número de 16 neurônios na camada intermediária. Além disso, esta combinação gramatical apresenta os menores erros de generalização do conjunto de testes se comparado as demais estruturas gramaticais utilizadas nos experimentos com 16 neurônios na camada intermediária e contemplando um vetor global de 90 termos. Assim, pode-se concluir

que no processo de categorização de textos, o uso de informações lingüísticas nas etapas de pré-processamento apresenta um ganho na taxa do erro de generalização comparado aos métodos tradicionais para a geração de ADs. No entanto, para o processo de aprendizado utilizando RNAs, os métodos tradicionais obtém os melhores resultados com 16 neurônios na camada intermediária.

No processo de agrupamentos de textos, diversos experimentos foram feitos utilizando o algoritmo de partição *K-means.* Adotou-se um vetor global representado pelo modelo de espaço vetorial, utilizando a codificação por freqüência relativa, e a seleção dos 30, 90 e 150 atributos relevantes para o conjunto de treinamento. Para a obtenção dos termos relevantes foram utilizados três conjuntos diferenciados de treino (2/3 dos documentos) e teste (1/3 de documentos), resultando em 9 experimentos de agrupamento, utilizando os parâmetros: semente randômica igual a 10 e o número grupos igual a 5. Dentre os 9 experimentos realizados no pré-processamento baseado em métodos usuais, em apenas um dos casos foi possível identificar dois grupos com mais clareza. Com base nesses resultados, optou-se por eliminar as seções Imóveis e Turismo do conjunto de treino e teste, devido a pouca precisão nos termos no processo de categorização.

Novos experimentos foram realizados com os mesmos parâmetros utilizados no grupo identificado anteriormente: semente randômica igual a 10, número de grupos igual a 5 e selecionando os 150 termos relevantes. Para estes experimentos foi possível identificar os 3 grupos distintos referentes aos assuntos Esporte, Informática e Política, nas versões V1 e V2 do corpus. Os experimentos realizados com a seleção dos termos baseados em informações lingüísticas também adotaram um vetor global representado pelo modelo de espaço vetorial, utilizando a codificação por freqüência relativa, e a seleção dos 30, 90 e 150 atributos relevantes para o conjunto de treinamento. Os parâmetros utilizados foram: semente randômica igual a 10 e o número de grupos igual a 5, para cada uma das categorias gramaticais propostas nos experimentos de categorização.

Analisando os resultados obtidos, pode-se verificar que as seções Imóveis e Turismo não são representativas para formar os 5 grupos de documentos. Com base nos parâmetros adotados no treinamento apenas as combinações gramaticais: substantivos – nomes próprios, substantivos – nomes próprios - adjetivos e substantivos - adjetivos, identificaram 3 grupos nas versões V1 e V2 do corpus para os 150 termos selecionados.

Assim, os documentos pertencentes seções Imóveis e Turismos foram eliminados dos conjuntos de treino e teste e os experimentos foram repetidos utilizando os mesmos parâmetros dos experimentos anteriores (grupos igual a 5 e semente igual a 10). As variações do corpus que permitiram a formação dos 3 grupos foram as versões V1 e V3 para os 150 termos mais relevantes e combinação gramatical substantivos-nomes próprios-adjetivos.

Pode-se verificar que o uso do algoritmo *K-means* com base na seleção de informações lingüísticas para a formação dos 5 grupos utilizando todas as seções do corpus mostrou-se mais significativa. Por outro lado, os resultados obtidos no agrupamento das seções Esporte, Informática e Política utilizando informações lingüísticas em comparação aos métodos usuais de pré-processamento para a seleção dos termos, mostrou-se semelhante.

No capítulo que segue, será descrita a conclusão obtida com o estudo realizado, bem como suas contribuições e trabalhos futuros.

6 Conclusão

Este trabalho apresentou um estudo sobre a aplicação de diferentes técnicas e etapas de pré-processamento na Mineração de Textos, nas tarefas de Categorização e Agrupamento de Documentos. O pré-processamento usual consiste na remoção de termos irrelevantes, normalização morfológica e seleção dos termos. O trabalho propôs um pré-processamento baseado em informações lingüísticas, visando a comparação de novas técnicas e etapas com métodos usuais de pré-processamento.

Para a realização do estudo foi adotada uma coleção de documentos do corpus NILC, contendo 855 textos jornalísticos das seções: Informática, Imóveis, Esporte, Política e Turismo do Jornal Folha de São Paulo do ano de 1994. Inicialmente o corpus foi distribuído em três ordens diferentes de documentos e após, para cada ordem de distribuição, dividido em dois conjuntos de documentos: um para a fase de treino e outro para a fase de teste.

Para o estudo do pré-processamento baseado em informações lingüísticas, todos os documentos que compõem os conjuntos de treino e teste foram submetidos ao analisador sintático PALAVRAS, a fim de gerar a análise sintática das sentenças dos documentos. Com base nas marcações do analisador sintático, a ferramenta Xtractor gera três arquivos com elementos XML: o primeiro arquivo contendo as palavras do corpus, o segundo com as informações morfo-sintáticas das palavras do texto, e por fim o arquivo de *chunks.* Gerados estes arquivos, a extração de informações lingüísticas é realizada aplicando folhas de estilos XSL.

Foram implementadas folhas de estilo para extração das seguintes combinações gramaticais: substantivos; substantivos - adjetivos; substantivos, nomes próprios - adjetivos; substantivos - nomes próprios; e nomes próprios - adjetivos. Após essa extração, os termos selecionados foram submetidos à preparação dos dados e então submetidos aos algoritmos de aprendizados. Sendo que, a preparação dos dados foi baseada no cálculo de relevância freqüência relativa e a seleção dos termos pela técnica de *truncagem.* Os algoritmos de aprendizado utilizados para verificar a performance do processo de MT nos experimentos foram: Árvores de Decisão e Redes Neurais

(Categorização) e algoritmo *K-means* (Agrupamento). Para realizar o aprendizado destes algoritmos foi adotada a ferramenta *Weka.*

Os primeiros experimentos foram realizados utilizando a categorização múltipla dos exemplos. Para cada exemplo um vetor global representado pelo modelo de espaço vetorial, utilizando as codificações por freqüência relativa e Tf-Idf, e a seleção dos 6, 12, 18, 24 e 30 termos mais freqüentes por categoria, correspondente ao 30, 60, 90, 120 e 150 termos que constituem o vetor global.

Os resultados obtidos neste processo para o pré-processamento baseado em métodos usuais para a geração de ADs, mostrou que para um corpus de pequeno porte, quanto maior o número de atributos menor o erro de generalização no conjunto de teste, independentemente da codificação adotada: freqüência relativa e Tf-Idf. Porém, comparado ao resultado obtido com o algoritmo RNA MLP-BP, este apresentou a menor taxa de erro utilizando-se 8 neurônios na camada intermediária, a codificação freqüência relativa e um número de 90 termos relevantes. Por outro lado, nos experimentos realizados com o pré-processamento baseado em informações lingüísticas na geração das ADs, as estruturas gramaticais que obtiveram como resultados percentuais as melhores taxas de erro em comparação aos métodos tradicionais de pré-processamento foram: substantivos - adjetivos e substantivos - nomes próprios - adjetivos. Quanto ao uso de RNAs estas obtiveram para as informações lingüísticas uma taxa de erro menor do que as ADs. Mas se considerarmos os 60 termos mais relevantes com um número de 16 neurônios na camada intermediária, o aprendizado com base nas combinações gramaticais substantivos - nomes próprios - adjetivos produz resultados similares aos métodos tradicionais baseados em métodos usuais.

No segundo experimento com o agrupamento de textos adotou-se um vetor global representado pelo modelo de espaço vetorial, utilizando a codificação por freqüência relativa. Foram selecionados os 30, 90 e 150 termos mais relevantes para o conjunto de treinamento, com base na seleção dos 6, 18 e 30 termos selecionados na categorização. Utilizaram-se os seguintes parâmetros: semente randômica igual a 10 e o número de grupos igual a 5. Nos resultados obtidos com o pré-processamento baseado em métodos usuais (ao todo nove experimentos), em apenas um dos casos foi possível identificar a formação de dois grupos. Optou-se por eliminar as seções Imóveis e Turismo do conjunto de treino e teste e realizar novos experimentos com os mesmos parâmetros anteriores. Para estes experimentos foi possível identificar os 3 grupos

distintos referentes aos assuntos Esporte, Informática e Política, nas versões V1 e V2 do corpus.

Para os experimentos utilizando informações lingüísticas foram identificados 3 grupos nas versões V1 e V2 do corpus para os 150 termos selecionados para as combinações gramaticais: substantivos - nomes próprios, substantivos - nomes próprios - adjetivos e substantivos - adjetivos. Ao eliminar as seções Imóveis e Turismos dos conjuntos de treino e teste as variações do corpus que permitiram a formação dos 3 grupos foram as versões V1 e V3 para a combinação gramatical substantivos - nomes próprios - adjetivos.

Dessa forma, pode-se verificar que os resultados obtidos com o uso de informações lingüísticas na fase de pré-processamento apresentaram melhorias em ambas tarefas de categorização e agrupamento de textos. Os algoritmos de aprendizado simbólico mostraram um melhor resultado se comparado aos métodos usuais de pré-processamento, com uma taxa de erro de 18,01% para a seleção de substantivos e adjetivos. Para o aprendizado utilizando RNA MLP-BP, no entanto, os métodos usuais apresentaram um melhor desempenho 14,64% para os 90 termos mais relevantes. O estudo realizado possibilitou verificar a importância da categoria gramatical substantivos na seleção dos termos relevantes, bem como na combinação dessa com os complementos adjetivos e nomes próprios.

Nos experimentos realizados com o agrupamento, as combinações gramaticais permitiram a identificação de um número maior de grupos se comparado aos métodos usuais de pré-processamento utilizando todas as seções do corpus.

Como trabalho futuro pretende-se comparar a metodologia proposta com a extração de sintagmas nominais simples (2 ou 3 palavras) dos textos, bem como a extração de n-grama. Também busca-se comparar os resultados obtidos aqui com outros algoritmos de aprendizado, por exemplo SVM, devido a sua robustez em grandes números de termos. SVM tem sido aplicado para documentos utilizando o português Europeu [Gonçalves and Quaresma, 2003], porém o pré-processamento baseado em combinações gramaticais ainda não foi testado para o processo de categorização. Como as categorias estudadas aqui não eram de domínios muito distintos, outro estudo poderia aplicar a metodologia proposta em corpora de outras línguas ou de outro domínio.

7 Referências Bibliográficas

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley. 1999.

[Bekkerman et. al., 2003] Bekkerman, R., R. El-Yaniv, N. Tishby, and Y. Winter. *Distributional word clusters vs. words for text categorization.* Journal of Machine Learning Research 3, 1183–1208, 2003.

[Beale and Jackson, 1990] Beale, R. Jackson, T. *Neural Computing: An Introduction*. Institute of Physics Publishing. 1990.

[Bick, 2000] Bick, E. *The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework*. Århus University. Århus: Århus University Press,2000.

[Chen, 1994] Chen, Hsinchum. *The vocabulary problem in collaboration*. IEEE Computer, special issue on CSCW, v. 27, n. 5, May 1994.

[Côrrea and Ludemir, 2002] Correa, R.; Ludemir, T. *Categorização Automática de Documentos: Estudo de Caso.* In: XVI Brazilian Symposium on Neural Networks, Porto de Galinhas, 2002.

[Cole, 1998] Cole, R. M. *Clustering with Genetic Algorithms*, M. Sc., Department of Computer Science, University of Western Australia, Australia, 1998.

[Cutting, 1992] Cutting, D. et al. *Constant interaction-time scatter/gather browsing of very large document collections.* In Special Interest Group on Information Retrieval, SIGIR. New York: Association for Computing Machinery, p.126-134, 1993.

[Duarte et al, 2002] Duarte, E.; Braga, A.; Braga, J. *Agente Neural para Coleta e Classificação de Informações Disponíveis na Internet.* In: XVI Brazilian Symposium on Neural Networks(SBRN 2002), 2002, Porto de Galinhas. Proceedings.

[Fayyad, 1996] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.. *From Data Mining to Knowledge Discovery: An Overview*. In: Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996. 611 p. p.11-34.

[Fayyad et. al., 2002] Fayyad, U.; G. G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discivery.* Morgan kaufmann Publishers, 2002.

[Feldman, 1995] Feldman, Ronen; Dagan, Ido. *Knowledge discovery in textual databases (KDT).* In: 1st International Conference on Knowledge Discovery (KDD-95). Montreal, August 1995.

[Frakes, 1992] Frakes, William B. *Stemming Algorithms*. In: Frakes, William B.; Baeza-Yates, Ricardo A. Information Retrieval: Data Structures & Algorithms. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.131-160.

[Gasperin et al., 2003] Gasperin, C.; Vieira, R.; Goulart, R. and Quaresma, P. *Extracting XML Syntactic Chunks from Portuguese Corpo*ra. Proceedings of the Workshop TALN 2003 Natural Language Processing of Minority Languages and Small Languages - Batz-sur-Mer France June 11 – 14, 2003.

[Gonçalves and Qauresma, 2003] Gonçalves, T.; Quaresma, P. *A prelimary approach classification problem of Portuguese juridical documents*. In: In: 11º Portuguese Conference on Artificial Intelligence Lectures Notes In: 11th Portuguese Conference on Artificial Intelligence, 2003, Béja. - Lecture Notes in Artificial Intelligence. Berlim: Springer Verlag

[Guthrie, 1996] Guthrie, Louise et al. *The role of lexicons in natural language processing*. Communications of the ACM, v.39, n.1, p.63-72, 1996.

[Habn and Mani, 2000] Habn, U. and Mani, I. *The challenges of Summarization.* IEEE *Automatic* Computer 33(11), p. 29-36, 2000.

[Harman, 1991] Harman, D. *How effective is suffixing.* Journal of the American Society for Information Science 42 (1), 7–15, 1991.

[Haykin, 2001] Haykin, S. *Redes Neurais : princípios e Prática.* 2ed. Bookman, 2001.

[Jain and Dubes, 1999] Jain, A. K.; Murty, M. N.; and Flynn, P. J. *Data Clustering: A Review*. ACM Computing Surveys 31 (3): 26423, 1999.

[Joachims, 1998] Joachims, T. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. In: Proceedings of European Conference on Machine Learning (ECML), Berlin, p. 137-142, 1998.

[Joachims, 2002] Joachims, T. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers, 2002.

[Korfhage, 1997] Korfhage, Robert R. *Information Retrieval and Storage*. New York: John Wiley & Sons, 1997. 349p.

[Kowalski, 1997] .Kowalski, Gerald. *Information Retrieval Systems: Theory and Implementation.* Boston: Kluwer Academic Publishers, 1997. 282p.

[Kuramoto, 2002] Kuramoto, H.. *Sintagmas Nominais: uma Nova Proposta para a Recuperação de Informação.* In: Revista de Ciência da Informação - v.3 n.1 Fevereiro 2002.

[Kraaij, 1996] Kraaij, Wessel; Pohlmann, Renée. *Viewing stemming as recall enhancement.* In Proceedings of ACM SIGIR, pp. 40-48, 1996. .

[Lancaster, 1968] Lancaster, F. Wilfrid. I*nformation Retrieval Systems: Characteristics*, Testing and Evaluation. New York: John Wiley & Sons, 1968. 222p.

[Lennon et al, 1981] Lennon, M., D. Pierce, B. Tarry, P. Willett .*An evaluation of some conflation algorithms for information retrieval*. Journal of Information Science 3, 177–183, 1981.

[Lewis, 1991] Lewis, David D. *Evaluating Text Categorization*. In Proceedings Speech and Natural Language Workshop, pages 312-318, 1991.

[Lewis, 1992] Lewis, D. D. *Representation and Learning in Information Retrieval.* Thesis of Doctor of Philosophy. Massachusetts: Department of Computer and Information Science, University of Massachusetts, 1992.

[Lewis and Ringuette, 1994] Lewis, D. D.; Ringuette, Marc. *Comparison of Two Learning Algorithms for Text Categorization*. In Proceedings of the ThirdAnnual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994.

[Lovins, 1968] Lovins, J. B. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics 11* (1-2), 22–31, 1968.

[MacCallum and Nigan, 1998] MacCallum, A. and Nigam, K. *A Comparison of events models for naïve bayes text classification.* In: AAAI/ICML – 98 Workshop on Learning for Text Categorization. Technical Report WS-98-05, AAAI Press, p.41-48, 1998.

[Meadow et. al., 2000] Meadow, C., Boyce, B. e Kraft, D. *Text Information Retrieval Systems.* San Diego: Academic Press, 2000.

[Melcop et. al, 2002] Melcop, T.; Costa, I.; Barros, F.; Ramalho, G. *Uma Ferramenta para Recuperação e Categorização de Páginas Web para Domínios Específicos.* In: Workshop de Ontologias para a Construção de Metodologias de Buscas na Web - XIII Simpósio Brasileiro de Informática na Educação (SBIE 2002), São Leopoldo, 2002.

[Mitchell, 1997] Mitchell, T. M. *Machine Learning.* McGraw-Hill. 1997.

[Moulinier et al. 1996] Moulinier, I.; Raskinis, G.; Ganascia, J. G. *Text Categorization: A Symbolic Approach.* In: Proc. SDAIR-96, 5th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, April 1996.

[Neves, 2001] Neves, M. L. *PubsFindes – um agente inteligente para busca e classificação de paginas de publicações.* Marster's thesis, Centro de Informatica – Universidade Federal de Pernambuco, Recife – PE, 2001.

[Ng et al, 1997] Ng, H.; Goh, W.; Low, K. *Feature Selection, Perceptron Learning and Usuability Case Study for Text Categorization.* Proceedings of 20th ACM International Conference on Research Development in Information Retrieval, Philadelphia, PA, USA, p.67-73, 1997.

[Oliveira and Castro, 2000] Oliveira, C.; Castro, P. *Categorização Múltipla com Árvores de Decisão e Regras.* Instituto Militar de Engenharia, Departamento de Engenharia de Sistemas (Relatório Técnico), 2000.

[Orengo, 2001] Orengo, V.; Huyck, C., *A Stemming Algorithm for Portuguese Language.* In: Eigth Symposium on String Processing and Information Retrie(RIJ 79).val (SPIRE 2001), Chile, 2001. pp. 186-193.

[Paice, 1994] Paice, C. *An Evaluation Method for Stemming Algorithms.* Proceedings Conference on Research and Development in Information Retrieval. London: Springer Verlag, p.42-50, 1994.

[Porter, 1980] Porter, M. F. *An Algorithm for Suffix Stripping.* Program, 14(3): 130-137, 1980.

[Quinlan 1986] Quinlan, J. R. *Induction of Decision Trees.* In Readings in Knowledge Acquisition and Learning, Bruce G. Buchnan & David C. Wilkins, Morgan Kaufmann, pp. 349-361, 1986.

[Quinlan, 1993] Quinlan, J. R. *C 4.5 : Programs for Machine Learning.* San Mateo: Morgan Kufmann Publishers, 1993.

[Rezende et al.,1998] Rezende, S. O.; Oliveira, R. B. T.; Feliz,L. C. M.; Rocha, C. A. J. *Visualization for Knowledge Discovery in database.* Em N. F. F. Ebecken (ed.), Data Mining, England, pp 81-95. WT Press – Computational Mechanics Publications, 1998.

[Rijsbergen, 1979] Rijsbergen, C. Van. *Information Retrieval.* 2ed. London: Butterworths, 1979. 147p.

[Riloff, 1995] Riloff, Ellen. *Little words can make big difference for Text Classification*. In Ed- ward A. Fox, Peter Ingwersen, and Raya Fidel, editors, Proceedings of SIGIR-95, 102 18th ACM International Conference on Research and Development in Information Retrieval, pages 130 136, Seattle, US, 1995. ACM Press, New York, US.

[Rizzi et. al, 2000] Rizzi, C.; Wives, L.; Oliveria, J. P., Engel P. *Fazendo uso da Categorização de Textos em Atividades Empresariais.* In: Internationa Symposium on Knowledge Management / Document - ISKDM/DM 2000, 2000, Curitiba. Anais. 2000. p.251-268.

[Rumelhart and McClelland 1986] Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing, volume 1*: Foundations. The MIT Press, 1986.

[Salton, 1983] Salton, G.; MacGill, M. *Introduction to Modern Information Retrieval.* New York: McGRAW-Hill, 1983. 448p.

[Salton, 1987] Salton, Gerard; Buckley, Chris. *Improving Retrieval Performance by Relevance Feedback.* Ithaca, New York: Department of computer science, Cornell University, 1987. (Technical Report).

[Sebastiani, 2002] Sebastiani, F. *Machine learning in automated text categorization.* ACM *Computing Surveys*, vol. 34, no. 1, pp.1-47, 2002.

[Steinberg and Colla, 1995] Steinberg, D. & Colla, P. *CART: Tree-Structured Non-Parametric Data Analysis*. Salford Systems, San Diego, CA. 1995.

[Scarinci, 1997] Scarinci, R. G. *SES: Sistema de Extração Semântica de informações*. Porto Alegre, 1997. 165p. Dissertação de mestrado – Instituto de Informática, UFRGS.

[Schütze et al., 1994] Schütze, H.; Pedersen, J.; Hearst, M. Xerox TREC3 report: *Comining Exact and Fuzzy Predictors.*

[Sparck et al., 1997] Sparck-Jones, Karen; Willet, Peter (eds). *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann, 1997.

[Steinberg and Colla, 1995] Steinberg, D. and Colla, P. *CART: Tree-Strutured Non_Parametric Data Analysis.* Salford Systems, San Diego, CA. 1995.

[Tan, 1999] Tan, A. *Text mining: the state of the art and the challenges.* In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases - PAKDD'99, pages 65-70, Beijing, April 1999.

[Theodoridas and Koutroumbas, 1998] Theodoridas, S. and Koutroumbas, K. *Pattern Recognition*. Academic Press, p. 351-495, 1998.

[Wiener et al, 1995] Wiener, E.; Pedersen, J.; Weigend, A. *A Neural Network Approach to Topic Spotting*. In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, pages 22-34, 1993.

[Wilkinson, 1994] Wilkinson, Ross. *Effective retrieval of structured documents.* In The 17 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 311-317, Dublin, Ireland, 1994. ACM.

[Witten 2000] Witten, I. H. Data mining: *Pratical Machine Learning tools and techniques with Java implementations*. Academic Press, 2000.

[Yang and Pedersen, 1997] Yang, Y; Pederson, J. *A Comparative Study on Feature Selection in Text Categorization*. In Proceedings of 14th International Conference on Machine Learning, pp. 412-420, Morgan Kaufmann Publishers, San Francisco, US.

[Yang and Liu, 1999] Yang, Y. and Liu, X. *An evaluation of statistical approaches to text categorization*. Journal of Information Retrieval, v.1, n.1/2, p.67-88, 1999.

Apêndice A

Saídas geradas pela Ferramenta Weka

Nas seções seguintes, seguem os resultados apresentados como saída pela ferramenta *Wek*: na geração de Árvores de Decisão com o algoritmo J48 e no processo de aprendizado de uma Rede Neural do tipo MLP com o algoritmo *Backpropagation.*

A.1 Exemplo de saída do algoritmo J48

```
=== Run information ===

Scheme:       weka.classifiers.j48.J48 -C 0.25 -M 2
Relation:     script
Instances:    33
Attributes:   10
              banc
              cdrom
              corr
              cr$
              dia
              equip
              jog
              multimíd
              program
              category {informatica,dinheiro,esporte}

Test mode:   10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

cdrom <= 0
|   equip <= 1
|   |   corr <= 0: dinheiro (14.0/3.0)
|   |   corr > 0: esporte (3.0)
|   equip > 1: esporte (7.0)
cdrom > 0: informatica (9.0)

Number of Leaves  : 4
Size of the tree :  7

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          28        84.8485 %
Incorrectly Classified Instances         5        15.1515 %
Kappa statistic                          0.7727
Mean absolute error                      0.1346
Root mean squared error                  0.3072
Relative absolute error                 30.2914 %
Root relative squared error             65.1574 %
Total Number of Instances               33

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.818     0          1         0.818     0.9      informatica
  1         0.182      0.733     1         0.846    dinheiro
  0.727     0.045      0.889     0.727     0.8      esporte

=== Confusion Matrix ===

  a  b  c   <-- classified as
  9  1  1 |  a = informatica
  0 11  0 |  b = dinheiro
  3  8  0 |  c = esporte
```

A.2 Exemplo de saída da RNA MLP com o algoritmo *Backpropagation*

```
=== Run information ===

Scheme:       weka.classifiers.neural.NeuralNetwork -L 0.3
              -M 0.2 -N 5000 -V 0 -S 20 -E 20 -H 2
Relation:     script
Instances:    33
Attributes:   10
              banc
              cdrom
              corr
              cr$
              dia
              equip
              jog
              multimíd
              program
              category {informatica,dinheiro,esporte}

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Sigmoid Node 0
    Inputs    Weights
    Threshold    -1.834118086819874
    Node 3    6.919187179603994
    Node 4    -8.058754732570879
.........
.........
Time taken to build model: 2.2 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          26               78.7879 %
Incorrectly Classified Instances         7               21.2121 %
Kappa statistic                          0.6818
Mean absolute error                      0.1558
Root mean squared error                  0.3437
Relative absolute error                 35.0578 %
Root relative squared error             72.9073 %
Total Number of Instances               33

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.727     0.045      0.889     0.727      0.8      informatica
  0.818     0.182      0.692     0.818      0.75     dinheiro
  0.818     0.091      0.818     0.818      0.818    esporte

=== Confusion Matrix ===

 a b c   <-- classified as
 8 2 1 | a = informatica
 1 9 1 | b = dinheiro
 0 2 9 | c = esporte
```