FÁBIO SOARES FIGUEIREDO

CONSTRUÇÃO DE EVIDÊNCIAS PARA CLASSIFICAÇÃO AUTOMÁTICA DE TEXTOS

Belo Horizonte
11 de abril de 2008

FÁBIO SOARES FIGUEIREDO
Orientador: Wagner Meira Jr.

CONSTRUÇÃO DE EVIDÊNCIAS PARA CLASSIFICAÇÃO AUTOMÁTICA DE TEXTOS

Proposta de dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Belo Horizonte
11 de abril de 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Construção de Evidências para Classificação Automática de Textos

FÁBIO SOARES FIGUEIREDO

Proposta de dissertação defendida e aprovada pela banca examinadora constituída por:

Ph. D. Wagner Meira Jr. – Orientador
Universidade Federal de Minas Gerais

Ph. D. Marcos André Gonçalves – Co-orientador
Universidade Federal de Minas Gerais

Ph. D. Nívio Ziviani
Universidade Federal de Minas Gerais

Ph. D. André Carlos Ponce de Leon Ferreira de Carvalho
Universidade de São Paulo, São Paulo

Belo Horizonte, 11 de abril de 2008

De tudo ficaram três coisas:

A certeza de que estava sempre começando,

A certeza de que era preciso continuar,

A certeza de que seria interrompido antes de terminar.

Portanto, devemos:

Fazer da interrupção um caminho novo.

Da queda, um passo da dança,

Do medo, uma escada,

Do sonho, uma ponte,

Da procura, um encontro.

Fernando Sabino

Agradecimentos

Embora uma dissertação seja um trabalho majoritariamente individual, há contribuições importantes de origens diversas que precisam ser realçadas. Por esse motivo, desejo expressar meus sinceros agradecimentos.

Agradeço primeiramente aos meus pais, que, antes de qualquer um, sempre acreditaram em mim. Agradeço também aos meus orientadores Wagner Meira Jr. e Marcos André Gonçalves, assim como a Leonardo C. D. Rocha, com os quais tive o prazer de trabalhar e discutir em prol do desenvolvimento do trabalho. Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa concedida durante os dois anos do curso.

Enfim, agradeço, de modo geral, a todos os meus amigos por me acompanharem nesta encantadora caminhada.

Resumo

Desde a popularização de documentos digitais, a classificação automática de textos é considerada um importante tópico de pesquisa. Apesar dos esforços na área, ainda há espaço para aperfeiçoar o desempenho de classificadores. A maior parte da pesquisa em classificação automática de texto foca em desenvolver algoritmos de classificação. Porém, não há muitos esforços concentrados em aperfeiçoar a representação das bases de dados usadas para treinar classificadores automáticos de texto. Este tipo de esforço, por sua vez, é o foco deste trabalho.

Nós propomos uma estratégia de tratamento de dados, baseada em extração de características, que precede a tarefa de classificação, a fim de introduzir em documentos características discriminativas de cada classe capazes de melhorar a eficácia da classificação.

Nossa estratégia é baseada em co-ocorrência de termos visando à geração de termos compostos discriminativos, chamados de c-termos, que podem ser incorporados aos documentos para facilitar a tarefa de classificação. A idéia é que, quando usados em conjuntos com os termos isolados, a ambigüidade e ruído inerente aos termos que compõem os c-termos é reduzida, portanto tornando-os mais úteis para separar classes em partições mais homogêneas.

Contudo, o custo computacional da extração de características pode tornar o método inviável. Neste trabalho, elaboramos um conjunto de mecanismos que torna a estratégia computacionalmente viável ao mesmo tempo em que aperfeiçoamos a eficácia dos classificadores.

Nós testamos essa abordagem com diversos algoritmos de classificação e coleções de texto que são referência na literatura. Resultados experimentais demonstram ganhos em quase todos os cenários testados, desde os algoritmos mais simples, como *k-Nearest Neighbors* (kNN) (46% de ganho em micro-média F_1 sobre a coleção 20 Newsgroups 18828) até o algoritmo mais complexo, estado da arte, *Support Vector Machine* (SVM) (10,7% de ganho em macro-média F_1 na coleção OHSUMED).

Abstract

Since the popularization of digital documents, automatic text classification is considered an important research topic. Despite the research efforts, there is still a demand for improving the performance of classifiers. Most of the research in automatic text classification focus on the algorithmic side, but there are few efforts focused on enhancing the datasets used for training the automatic text classifiers, which is the focus of this paper.

We propose a data treatment strategy, based on feature extraction, that precedes the classification task, in order to enhance documents with discriminative features of each class capable of increasing the classification effectiveness.

Our strategy is based on term co-occurrences to generate new discriminative features, called compound-features (or c-features), that can be incorporated to documents to help the classification task. The idea is that, when used in conjunction with single-features, the ambiguity and noise inherent to c-features components are reduced, therefore making them more helpful to separate classes into more homogeneous partitions.

However, the computational cost of feature extaction may make the method unfeasible. In this paper, we devise a set of mechanisms that make the strategy computationally feasible while improving the classifier effectiveness.

We test this approach with several classification algorithms and standard text collections. Experimental results demonstrated gains in almost all evaluated scenarios, from the simplest algorithms such as k-Nearest Neighbors (kNN) (46% gain in micro

average F_1 in the 20 Newsgroups 18828 collection) to the most complex one, the state of the art Support Vector Machine (SVM) (10,7% gain in macro average F_1 in the collection OHSUMED).

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Classificação Automática de Documentos

Nas últimas duas décadas, tarefas relacionadas à gestão do conteúdo de documentos ganharam destaque proeminente na Computação. A ampliação do uso da Internet para distribuição de todo tipo de informação resultou em um crescente volume de dados a serem armazenados e acessados por meio da WWW e por outras ferramentas. Estes dados são freqüentemente organizados como documentos textuais e têm sido o principal alvo de máquinas de busca e outras ferramentas de recuperação, que executam tarefas como busca por documentos potencialmente relevantes e filtragem de textos com base em conteúdos específicos.

Uma técnica popular e tradicional Sebastiani (2002) associada à recuperação de documentos é a classificação automática de textos (CAT), que consiste em identificar a quais categorias pré-definidas um documento de texto discorre, a fim de agrupar documentos semanticamente relacionados. Mais formalmente, dado um documento d_j e um conjunto de classes $C = \{c_1, c_2, \ldots, c_K\}$, a CAT consiste em identificar um subconjunto $C_{d_j} \subseteq C$, tal que C_{d_j} representa as classes sobre as quais o documento d_j discorre. Assim, a classificação de texto permite, por exemplo, ajudar usuários e ferramentas a localizar os documentos classificados por meio de tópicos.

A CAT remonta ao início dos anos 60, mas somente nos anos 90 tornou-se um campo

importante da computação, graças à sua crescente aplicabilidade em diversos problemas e à disponibilidade de hardwares mais poderosos. Atualmente, a CAT é a base de diversas aplicações importantes, como filtragem de *spam* Sculley e Wachman (2007); Zhang et al. (2004), detecção de conteúdo adulto Rongbo Du (2003); Mohamed Hammami (2004); Chandrinos et al. (2000), organização de documentos em tópicos de bibliotecas digitais Couto et al. (2006); Amati et al. (1997), e tipicamente qualquer aplicação que requer organização ou seleção de documentos.

Até o fim dos anos 80, a abordagem mais popular usada para CAT era baseada em sistemas especialistas Sebastiani (2002). Essa estratégia consistia em desenvolver sistemas capazes de aplicar uma série de regras, introduzidas manualmente por especialistas de um domínio específico de problema, com o objetivo de serem usadas para classificar automaticamente documentos em categorias pré-definidas Hayes et al. (1990). Porém, uma grande desvantagem de sistemas baseados nessa estratégia é que eles sofrem do gargalo da aquisição de conhecimento (*knowledge acquisition bottleneck*) Cullen e Bryman (1988), i.e., dificuldade de representar fielmente massas de dados, à medida em que crescem, por meio de regras de classificação manualmente introduzidas. Essa dificuldade torna-se mais evidente em função de:

- um profissional recorrentemente precisar intervir à medida em que as massas de dados são atualizadas, em decorrência da necessidade de se criar, remover e adaptar regras de classificação então existentes;
- um especialista de outro domínio precisar intervir caso os dados sejam portados para outro domínio (ex.: mudança de taxonomia), o que exigiria que o trabalho fosse completamente refeito;
- o especialista tornar-se o gargalo em termos de custo e tempo à medida em que as massas de dados crescem.

Como resultado, sistemas especialistas foram gradativamente perdendo espaço para solucionar problemas de CAT no início dos anos 90, principalmente na comunidade aca-

dêmica, em favor do paradigma baseado em aprendizado de máquina Sebastiani (2002). Por meio desse novo paradigma, a CAT passou a ser então implementada por meio de técnicas supervisionadas de aprendizado, i.e., existe um conjunto de treino composto por exemplos previamente classificados, de forma que esses exemplos sejam usados para automaticamente criar um classificador por meio de aprendizado indutivo. Conseqüentemente, o paradigma baseado em aprendizado de máquina automatizou o trabalho até então feito por especialistas. Dessa maneira, o classificador induzido automaticamente é usado para determinar a(s) melhor(es) classe(s) para um novo documento não classificado Mitchell (1997b). Logo, as vantagens do aprendizado de máquina sobre sistemas especialistas são diversas, como: (i) economia de tempo e custos; (ii) velocidade e praticidade para gerir um classificador; (iii) ausência de subjetividade humana; e (iv) alta eficácia de classificação em diversos domínios, inclusive superior a alguns sistemas especialistas Wielinga et al. (1990).

Em decorrência disso, atualmente o procedimento de criação de um classificador é visto tipicamente como o processo de determinar uma função que pondera automaticamente um conjunto de características que particionem os documentos de treino em grupos que sejam tão homogêneos quanto possível no que diz respeito às suas categorias. Essas características podem envolver diversas evidências encontradas comumente em documentos textuais, tais como:

- ocorrência de termos/palavras,
- co-ocorrência seqüencial de termos (n-gramas),
- co-ocorrência não seqüencial de termos (*itemsets*),
- ocorrência de frases completas, analisadas sintática ou semanticamente,
- referências bibliográficas,
- influência do tempo sobre a relevância de características.

Porém, um desafio comum para se induzir bons classificadores de documentos é a possível insuficiência de características discriminativas, o que resulta em modelos com informações ambíguas ou escassas. Há também limitações do classificador em função da complexidade dos padrões que ele é capaz de empregar. Assim, a seção a seguir introduz qualitativamente diversos aspectos que dizem respeito à utilização de características em CAT.

1.2 Representação da Informação

A CAT é uma sub-área importante da Recuperação de Informação e muito conhecimento nesse domínio tem se acumulado nas últimas décadas. Até o início desta década, por larga margem de diferença, a representação da informação textual para fins de CAT empregava apenas termos, tratados independentemente entre si, como características para indução de classificadores. Essa modelagem é também conhecida como *bag of words*, em que os termos são tratados independentemente uns dos outros. Desde os primórdios da CAT Salton e McGill (1983), diversas pesquisas focaram no desenvolvimento da teoria e prática de algoritmos de classificação, acarretando desde então o surgimento de diversos bons classificadores, como *Support Vector Machine* (SVM) Vapnik (1998); Joachims (1998, 1999), Adaboost Freund e Schapire (1995), Naïve-Bayes McCallum e Nigam (1998); Lewis (1998) e *k-Nearest Neighbors* (kNN) Yang e Pedersen (1997).

Como resultado, apesar do grande avanço decorrente da criação de diversos classificadores, a maioria das pesquisas simplificou a representação dos documentos como um conjunto de termos independentes (*bag of words*), de forma que a classificação fosse baseada na presença ou na ausência de termos chaves. O motivo disso é a simplicidade, eficiência e relativa eficácia do paradigma *bag of words*. Por exemplo, até 2003, o melhor resultado de classificação automática de texto para a muito estudada coleção Reuters-21578 era baseada na representação *bag of words* Bekkerman e Allan (2004).

Porém, a maior desvantagem do paradigma *bag of words* é a não utilização de

relações semânticas e sintáticas entre os termos, bem como a desconsideração de outras estruturas complexas, como redes de citações entre artigos ou a diferença do real significado de um termo quando redigido em épocas diferentes. Conseqüentemente, essas observações impulsionaram as pesquisas em CAT de modo a buscar melhores representações para dados contendo linguagem natural.

Neste trabalho, nós propomos e avaliamos uma estratégia para extração de características, que visa a identificar e introduzir evidências tanto nos documentos de treino como de teste a fim de facilitar a tarefa de classificação. Intuitivamente, nossa proposta incrementa documentos ao adicionar termos compostos (chamados de "c-termos" a partir de agora), que representam a co-ocorrência de dois ou mais termos que fornecem melhor habilidade de discriminação. A fim de ilustrar nosso raciocínio, vamos assumir que tenhamos uma biblioteca digital contendo documentos sobre diversos tópicos. O termo *título* provavelmente deve ocorrer em diversos contextos distintos, como Esportes, Redação e Mercado Financeiro. Nessas circunstâncias, um segundo termo (e.g., {título, taça}, {título, margem} e {título, ações}, respectivamente) ajudaria a determinar a categoria do documento. Observe que, em todos os três casos, qualquer termo sozinho não diz muito em função de possuir diversos significados, ao passo que a consideração combinada especifica a informação muito mais focadamente. Esse exemplo ilustra um dos principais ganhos fornecidos por um método de extração de características.

Nós ilustramos os pontos discutidos por meio de três técnicas mais tradicionais de CAT, descritas a seguir: kNN, Naïve-Bayes e SVM. Essas técnicas empregam uma função de particionamento que considera todas as características, gerando um *ranking* das possíveis categorias às quais atribuir um novo documento. Conseqüentemente, características ambíguas afetam a acurácia do modelo porque elas diminuem o poder discriminativo geral do modelo. Como resultado, a insuficiência de características discriminativas não deve acarretar boa acurácia. Dessa forma, torna-se uma grande questão determinar como minimizar o impacto de características não discriminativas

na construção de modelos.

A seção a seguir traz um breve resumo de como as questões referentes à extração de características por meio de relacionamentos entre termos são tratadas neste trabalho.

1.3 Co-ocorrência entre Termos

É interessante compreender como a extração de características auxilia o modelo de classificação. Ao adicionar características de bom poder discriminativo, nós aumentamos a acurácia do modelo ao fornecer melhores evidências para os classificadores. Assim, uma questão chave para aperfeiçoar o classificador é como obter um conjunto de características que sejam discriminativas a ponto de facilitar a classificação de texto. Porém, um aspecto importante associado ao relacionamento entre termos é o custo computacional inerente dessa tarefa, uma vez que o número máximo de combinações entre termos em uma coleção contendo T termos é 2^T.

É importante comparar como a extração de características ajuda a melhorar o classificador em comparação à seleção de características (*feature selection*). De fato, as estratégias são complementares. Enquanto a segunda seleciona conjuntos de características de acordo com seu poder discriminativo a fim de descartar padrões irrelevantes e ruidosos Yang e Pedersen (1997), a primeira adiciona novas evidências potencialmente relevantes para promover a geração de melhores classificadores. Nossa estratégia explora co-ocorrência de termos nos documentos de uma dada classe, ou seja, se um c-termo ocorre em documentos de uma dada classe e apenas nessa classe, então o c-termo possui bom poder discriminativo, mesmo quando os termos que o compõem não são bons discriminantes. Assim, um efeito imediato da construção de características é que ela pode reduzir o impacto de termos ambíguos, já que podemos estreitar os possíveis sentidos de um dos termos quando consideramos a interação entre eles.

A exploração de co-ocorrência de termos em sistemas de recuperação de informação tem sido pesquisada há algum tempo Pôssas et al. (2005); Rak et al. (2005); Zaïane e Antonie (2002); Feng et al. (2005). Porém, nossa estratégia é diferente por-

que fornece um mecanismo computacionalmente viável para CAT que incrementa o texto com c-termos flexíveis, tendo em vista que os relacionamentos entre os termos correspondentes podem aparecer em qualquer parte do documento, independentemente da distância ou ordem entre os termos Chomsky (1957). Além disso, nossa estratégia pode ser vista como um processo de tratamento mais rico da representação do texto. Como resultado, a estratégia se torna independente de algoritmo de classificação, pois é aplicada em uma fase anterior ao processo de indução de classificadores. Dessa forma, ao incrementar a representação da coleção de textos e induzir a construção de um classificador, estamos transparente e indiretamente quebrando o paradigma *bag of words*. Como conseqüência, qualquer classificador induzido pode usufruir das evidências construídas a fim de potencializar a qualidade da classificação, o que efetivamente fazem conforme será demonstrado em nossos experimentos e caracterização.

Em resumo, podemos antecipar quatro principais contribuições associadas a este trabalho:

1. Uma estratégia eficaz para incrementar a representação de documentos textuais ao empregar co-ocorrências, facilitando a construção de classificadores e criando modelos mais precisos, mesmo para classes contendo pequena quantidade de documentos.

2. A proposta de tornar a estratégia computacionalmente viável, evitando a explosão combinatória comumente associada a estratégias que empregam co-ocorrências.

3. A quantificação dos ganhos ao aplicar tanto a estratégia(1) como a proposta(2) sobre quatro coleções de referência, algumas largamente conhecidas como desafiadoras em termos de CAT.

4. A análise detalhada do impacto da estratégia sobre vários algoritmos (SVM, Naïve-Bayes e kNN) aplicados a bases de dados reais.

1.4 Estrutura da Dissertação

Este trabalho é organizado como se segue. O Capítulo 2 fornece uma discussão sobre trabalhos relacionados. O Capítulo 3 apresenta nossa estratégia para classificação automática de texto usando técnicas de mineração de dados. O Capítulo 4 descreve nosso estudo de caso para quatro coleções de texto de referência: Reuters-21578 8C, 20 Newsgroups 18828, OHSUMED 18302 e ACM 11C, mostrando os principais resultados associados à aplicação de nossa estratégia. O Capítulo 5 apresenta uma caracterização sobre o impacto da estratégia sobre diferentes algoritmos de classificação. Finalmente, o Capítulo 6 apresenta as conclusões e delineia trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

A pesquisa em extração automática e semi-automática de características em textos com o objetivo de aperfeiçoar a qualidade da CAT remonta ao início da década de 90. Com o desenvolvimento de hardwares mais poderosos e o advento da Internet, representações mais complexas que a *bag of words* foram introduzidas e estudadas. Dessa maneira, direta ou indiretamente, a maior parte dessas novas representações objetivava reduzir a ambigüidade inerente aos termos individuais (a partir de agora denominados "s-termos") a fim de induzir modelos de classificação superiores.

As seções a seguir discutem diversos trabalhos relacionados sobre extração de características para fins de aperfeiçoar a qualidade da CAT. As seções 2.1, 2.2 e 2.3 apresentam diversas abordagens distintas que objetivam melhorar a classificação de documentos por meio de um ponto em comum: identificar relacionamentos semânticos ou sintáticos entre termos de modo a descobrir novas características úteis além dos termos originais tratados independentemente. Para tanto, decidimos apresentar essas três abordagens sob uma perspectiva evolutiva. Iniciamos por discutir, na Seção 2.1, os primeiros trabalhos de impacto que objetivaram aperfeiçoar a representação *bag of words*, tendo então utilizado frases como características. Em seguida, na Seção 2.2, discutimos estratégias que buscam extrair co-ocorrências entre termos adjacentes (n-gramas) que sejam úteis para fins de CAT. Finalizamos a discussão ao apresentar, na Seção 2.3, estratégias sobre extração de co-ocorrência entre termos não adjacentes (*itemsets*)

para aperfeiçoar a qualidade da CAT. Esta última estratégia é a base deste trabalho. Por fim, na Seção 2.4, a título de complementaridade, apresentamos outros trabalhos sobre extração de características, que utilizam redes de referências bibliográficas de modo a beneficiar a qualidade da CAT em domínios que fornecem essa informação, como artigos científicos.

2.1 Frases

No início da década de 90, alguns estudos analisaram a utilização de frases como características a fim de testar a classificação de documentos. Em um número razoável de experimentos (Lewis (1992c,b,a); Apté et al. (1994)), uma análise sintática do texto para extrair frases foi aplicada. Porém, em todos esses trabalhos foi demonstrado que o emprego de frases tipicamente proporcionava degradação nos resultados em relação ao paradigma *bag of words*. Lewis (1992c) apresenta detalhes mais aprofundados sobre as razões desse comportamento. Em seu trabalho, foram apontadas algumas propriedades desejáveis para uma classificação de qualidade, como:

1. pouca redundância de termos para o mesmo significado;
2. pouca irregularidade na distribuição de termos por entre classes;
3. pouca ambigüidade;
4. evitar a indexação de termos ruidosos (pouco discriminativos).

Com base nessas propriedades, foi relatado que frases tendem a obedecer bem o critério 4, porém não obedecem bem os 3 primeiros conforme a metodologia empregada por Lewis, pois: (i) a ocorrência de frases se mostrou muito irregular entre classes; (ii) a ambigüidade geral do modelo foi prejudicada em função da utilização de sinônimos para se estabelecer um critério de equivalência entre frases. Neste último caso, observe que se cada s-termo de uma frase que possua T s-termos contiver, em média,

S sinônimos, então S^T possíveis frases poderiam ter o mesmo significado, o que aumenta drasticamente o universo de frases distintas para definir o mesmo sentido. Como resultado, Lewis (1992b) concluiu que a baixa freqüência de frases semanticamente distintas aliada à alta dimensionalidade do espaço e da alta taxa de sinônimos superaram as vantagens que elas tenderiam a introduzir para representar textos.

Em linha com os estudos anteriores, Dumais et al. (1998) demonstraram que nenhum ganho foi obtido ao se extrair frases como características a partir da estrutura sintática do texto. Não obstante esses resultados desencorajadores, Scott e Matwin (1999) apresentaram um estudo semelhante aos de Lewis, porém alterando alguns detalhes de implementação a fim de selecionar mais rigorosamente apenas as melhores frases da coleções. A hipótese dos autores é que algoritmos baseados em regras de inferência (*rule-based algorithms*) poderiam induzir classificadores mais ricos em regras de classificação úteis caso apenas frases discriminativas sejam extraídas e introduzidas. Contudo, apesar de apresentar ganhos marginais na coleção Reuters-21578, seus resultados ainda foram consistentemente piores que outros algoritmos de classificação aplicados sobre o paradigma *bag of words*. Além disso, na outra coleção utilizada nesse mesmo trabalho (DigiTrad), a linha de base[1] foi tão baixa (36% de acurácia) que mesmo os ganhos apresentados (em torno de 6 pontos percentuais) não são suficientes para confirmar sua metodologia como eficaz para CAT.

Apesar ainda desses contratempos, inúmeras outras pesquisas sobre extração de seqüências de s-termos como características surgiram na expectativa de solucionar as lacunas deixadas em aberto. Dessa forma, a grande maioria dos trabalhos passaram a focar em extração de seqüências de s-termos de tamanhos variados, porém desconsiderando a análise semântica em que frases morfologicamente distintas poderiam ser igualadas por meio de sinônimos ou hiperônimos[2], conforme anteriormente feito por

[1] a partir de agora, a expressão "linha de base" será utilizada no sentido de caracterizar a qualidade da classificação ao empregarmos o paradigma *bag of words*.

[2] termo que apresenta um significado mais abrangente do que o do seu hipônimo (ex.: "doença" é hiperônimo de "gripe").

Lewis (1992c,b,a); Scott e Matwin (1999). A seção a seguir discute alguns trabalhos relacionados nessa linha.

2.2 Co-ocorrências Adjacentes (n-gramas)

Mladenić e Grobelnik (1998) investigaram o uso de seqüências de s-termos (n-gramas) de tamanhos variados como características. Em seu trabalho, foram usados apenas n-gramas freqüentes na coleção compostos por, no máximo, 5 s-termos. Em seus experimentos, foram construídos classificadores probabilísticos que empregavam n-gramas obtidos. Eles concluíram que n-gramas com três ou menos s-termos eram úteis para aperfeiçoar a eficácia da classificação e que n-gramas maiores não resultavam em ganhos. Contudo, sua metodologia foi testada apenas sobre uma única coleção de texto, pouco referenciada na literatura e de baixíssima linha de base, o que facilita o desafio de obter algum ganho. Conseqüentemente, essas circunstâncias levantam dúvidas se seus resultados generalizariam para outras coleções. Um trabalho parecido foi feito por Fürnkranz (1998), porém a freqüência da ocorrência dos n-gramas dentro dos documentos também foi considerada, sob a argumentação de que essa informação tenderia a melhorar os resultados. Todavia, foram observadas diversas degradações nos resultados e poucos ganhos, todos inexpressivos, em duas coleções de alta linha de base comumente empregadas na literatura: Reuters-21578 e 20 Newsgroups. Mais tarde, Caropreso et al. (2001) apresenta um estudo sobre extração de n-gramas também utilizando a coleção Reuters-21578. Nesse trabalho, é aplicado *stemming*[3] sobre os n-gramas na expectativa de identificar n-gramas lexicalmente distintos, mas semanticamente equivalentes. Além disso, também não foi permitido, ao contrário dos estudos anteriores, que a introdução de um n-grama em um documento substituísse seus s-termos correspondentes. Essa estratégia foi definida com base na intuição de que as relações entre termos fornecidas por n-gramas apenas auxiliam a desambiguação dos

[3]Em recuperação de informação: reduzir as variantes morfológicas de um s-termo para sua forma raiz (ex.: "*loving*" para "*lov*").

respectivos s-termos, mas não substituem o caráter discriminativo individual de cada s-termo. No entanto, dos 48 resultados apresentados, em 28 foram observadas depreciações nos resultados. Nos 20 restantes, ganhos estatisticamente insignificantes foram atingidos, o que aponta para falta de consistência dessa abordagem.

Tan et al. (2002) apresentam uma abordagem diferente para extração e utilização de n-gramas. Em analogia a Caropreso et al. (2001), n-gramas não substituem os respectivos s-termos que os compõem, mas são usados em adição a eles. Um aspecto interessante desse trabalho é que os critérios empregados para seleção de n-gramas são muito rigorosos, objetivando incrementar a representação de texto apenas com n-gramas excepcionalmente discriminativos. Como conseqüência, em relação à quantidade de características previamente existente por meio da representação *bag of words*, essa estratégia introduziu apenas 2% de características extras. Nesse trabalho, ganhos razoáveis e estatisticamente significativos foram apresentados. Apesar disso, esse estudo não empregou algoritmos de classificação eficazes em seus experimentos. Como conseqüência, seus resultados ainda ficaram bastante abaixo de algoritmos de classificação mais eficazes aplicados sobre representações *bag of words*. Em todo caso, os resultados apresentados serviram para sugerir que ganhos poderiam ser obtidos se métodos de classificação mais sofisticados fossem utilizados. Em analogia à Tan et al. (2002), Crawford et al. (2004) aplicou uma metodologia de n-gramas para classificar e-mails. Porém, não obteve ganhos generalizados. Além disso, em ambos os trabalhos, as linhas de base eram relativamente baixas, o que pode ter facilitado a tarefa de conquistar melhorias.

Em razão do discutido no paráfrago anterior e com o objetivo de reduzir a margem de dúvida deixada por Tan et al. (2002), Bekkerman e Allan (2004) utilizaram um algoritmo de classificação considerado estado da arte (SVM) e aplicaram uma metodologia de extração de n-gramas sobre a coleção 20 Newsgroups, cuja linha de base é muito alta ao se utilizar a representação *bag of words*. Contudo, conforme a própria descrição do autor: “as melhorias são estatisticamente insignificantes. Todavia, este

resultado é o melhor (de acordo com nosso conhecimento) alcançado para a coleção 20 Newsgroups". Dessa maneira, Bekkerman e Allan (2004) suscitou mais dúvidas que esclarecimentos ao demonstrar que seus ganhos não são estatisticamente significantes, embora tenha apontado que talvez fosse possível melhorar consistentemente a qualidade da classificação mesmo em circunstâncias adversas (alta linha de base) e usando um algoritmo já eficaz (SVM).

Assim, embora alguns pesquisadores tenham reportado ganhos, muitos deles são apenas marginais ou condicionados a circunstâncias muito específicas. Com isso, podemos concluir que a nuvem de dúvida sobre a eficácia de n-gramas ainda persiste. Em coleções bem conhecidas na literatura, como Reuters-21578 e 20 Newsgroups, por exemplo, ganhos estatisticamente significativos nunca foram consistentemente reportados pela comunidade acadêmica Bekkerman e Allan (2004). Para justificar essa falha, Bekkerman e Allan (2004) conjecturam que essas duas coleções já são naturalmente simples de classificar com a representação *bag of words*, o que acarreta altas linhas de base que impossibilitam melhorar caso representações mais sofisticadas fossem empregadas. Contudo, demonstramos empiricamente neste trabalho que essa proposição é falsa, pois apresentamos uma metodologia capaz de proporcionar ganhos estatisticamente significativos nessas duas coleções, mesmo com altíssimas linhas de base da representação *bag of words*. Além disso, colocamos nossa metodologia a prova em outras duas coleções e com três algoritmos distintos de classificação, tendo obtido ganhos estatisticamente significativos em todos os cenários avaliados.

Tendo em vista tudo que foi mencionado, abordagens baseadas em n-gramas não apresentaram sucesso consistente desde sua concepção. Dessa forma, em nosso trabalho, relaxamos as premissas dos paradigmas baseados em n-gramas e frases a fim de testar estratégias inovadoras. Em nossos estudos, não exploramos a ordem da ocorrência dos s-termos ou a distância entre eles para extrairmos relacionamento entre termos. Isso porque termos podem ser correlatos independentemente da distância e ordem entre eles Chomsky (1957). Como resultado dessa flexibilização, eliminamos a restrição

de extrair apenas conjuntos de termos adjacentes. Esse tipo de co-ocorrências entre s-termos é comumente referenciada na literatura como *termset* ou *itemset* Pôssas et al. (2005). Porém, para facilitar a distinção entre termos simples e termos compostos, utilizamos a expressão s-termos para palavras simples (individuais) e c-termos para co-ocorrências não seqüenciais entre s-termos. Além disso, em nosso trabalho, consideramos apenas c-termos contendo dois s-termos, pois isso é considerado bastante para uma desambiguação eficaz, uma vez que essa estratégia é suficiente para o propósito de deliminar adequadamente o real significado de um dos termos na maioria dos casos Kaplan (1955); Choueka e Lusignan (1985).

A seguir, discutimos trabalhos relacionados sobre construção de características utilizando-se c-termos.

2.3 Co-ocorrências não Adjacentes (*itemsets*)

Feng et al. (2005) consideraram o cruzamento entre sentenças a fim de identificar relacionamento entre s-termos. Esse trabalho se baseia na intuição de que escritores tendem a enfatizar idéias relevantes ao repetir alguns termos em sentenças distintas de um mesmo texto, independente da distância entre elas. Contudo, em coleções em que documentos são curtos, como a classificação de artigos com base no resumo, por exemplo, o pequeno número de sentenças encontrados em cada documento pode não ser suficiente para detectar bons relacionamentos. Por essa razão, em nossa abordagem, usamos o texto completo para identificar relacionamentos entre s-termos.

Zaïane e Antonie (2002) usaram regras de associação para construir classificadores de texto que empregavam conjuntos freqüentes de termos (c-termos), na expectativa de melhor caracterizar as classes. Eles também filtraram alguns dos c-termos que eram freqüentes em muitas categorias em seu método. Todavia, nas coleções testadas (Reuters-21578 and OHSUMED multi-rotuladas), seu método não apresentou ganhos consistentes quando comparado a classificadores mais tradicionais que usavam apenas termos simples (s-termos) como características. Em seguida a esse trabalho, Rak et al.

(2005) publicam outro estudo análogo, mas que considera o ponderamento das características geradas, argumentando que, para alguns problemas, a freqüência de um termo é uma evidência mais forte que sua presença. Cheng et al. (2007) mostrou, ao usar bases de dados não textuais, que c-termos medianamente freqüentes são mais prováveis de serem bons discriminantes e conseqüentemente melhor ajudar a eficácia da classificação em comparação aos pouco e muito freqüentes. Em nosso trabalho, questionamos essa proposição para classificação de texto e mostramos que c-termos infreqüentes são importantes como características discriminantes em algumas circunstâncias. Além disso, em todos esses três últimos trabalhos, combinações de todos os tamanhos entre termos foram usadas. Como conseqüência, seus métodos geram uma quantidade exponencial de c-termos (pares, triplas, etc), assim tornando sua estratégia inviável para grandes coleções de texto.

Em nosso trabalho, consideramos a interação entre apenas dois s-termos como sendo suficiente para desambiguação de termos, uma vez que essa estratégia é, na maioria dos casos, suficiente para reduzir os possíveis significados de um dos termos Kaplan (1955); Choueka e Lusignan (1985). Dessa maneira, nós empregamos estratégias de extração e seleção de *itemsets* muito mais restritivas, focadas nos *itemsets* que trazem maior poder de desambiguação (de tamanho 2) na expectativa de aperfeiçoar os classificadores que presumem independência entre termos. Além disso, essa abordagem representa um bom compromisso entre viabilidade computacional e qualidade de classificação, conforme demonstrado por nossos resultados experimentais. Outro aspecto positivo de nossa estratégia é sua capacidade de aperfeiçoar a qualidade da classificação de algoritmos de CAT fundamentados em premissas bastante distintas, testado sobre coleções de texto reais e com propriedades diversas.

2.4 Citações

Em bibliotecas digitais, principalmente as de domínio científico, é muito comum em documentos a existência de referências bibliográficas para citar trabalhos relevantes.

Ao mesmo tempo, diversas bibliotecas digitais precisam organizar seus documentos em tópicos diversos. Porém, simultaneamente, algumas sofrem de problemas de escassez de características, necessárias para uma classificação automática de qualidade. Em casos assim, é bastante desafiadora a tarefa de classificar conteúdo por meio de técnicas mais tradicionais de classificação que empregam o paradigma *bag of words*. Como conseqüência, objetivando contornar esse problema, a construção de características a partir de uma rede de citações bibliográficas pode servir de informação útil para potencializar a classificação. Couto et al. (2006); Zhang et al. (2005) investigam a valia da informação provida por citações científicas para prover a categoria de documentos. Em ambos os trabalhos, ganhos consistentes foram apresentados ao se empregar características extraídas a partir da rede de citações bibliográficas.

Porém, uma inconveniência no que diz respeito a informações de referências bibliográficas é que existem diversos domínios em que essa informação simplesmente inexiste. Como conseqüência, é importante desenvolver outras técnicas gerais que extraiam características capazes de potencializar a qualidade da CAT.

Capítulo 3

Estratégia de Extração de Características

Neste capítulo, descrevemos como executamos a extração de características e também discutimos alguns detalhes de implementação chave para aperfeiçoar o desempenho da estratégia.

Na seção 3.1 discutimos, em linhas gerais, nossa metodologia para extração e utilização de c-termos como evidências úteis a serem empregadas para maximizar a qualidade da CAT. Na seção 3.2, apresentamos alguns pontos críticos associados à eficiência dessa metodologia, bem como fundamentos para sua utilização.

3.1 Extração de c-termos

Nesta seção descrevemos a metodologia geral para extração de características para coleções unirótulo[1] (*unilabel*). A intuição é incrementar documentos, representados originalmente apenas por s-termos, com c-termos que agreguem informação de co-ocorrência, melhorando a eficácia de classificadores que não consideram tal informação. Esse raciocínio é sumarizado por meio das Figuras 3.1 e 3.2. A Figura 3.1 ilustra o processo mais tradicional de algoritmos de aprendizado de máquina aplicados a CAT.

[1] coleções cujos documentos pertencem a uma única classe

Por sua vez, a Figura 3.2 apresenta as diferenças de nosso processo em relação ao mais tradicional.

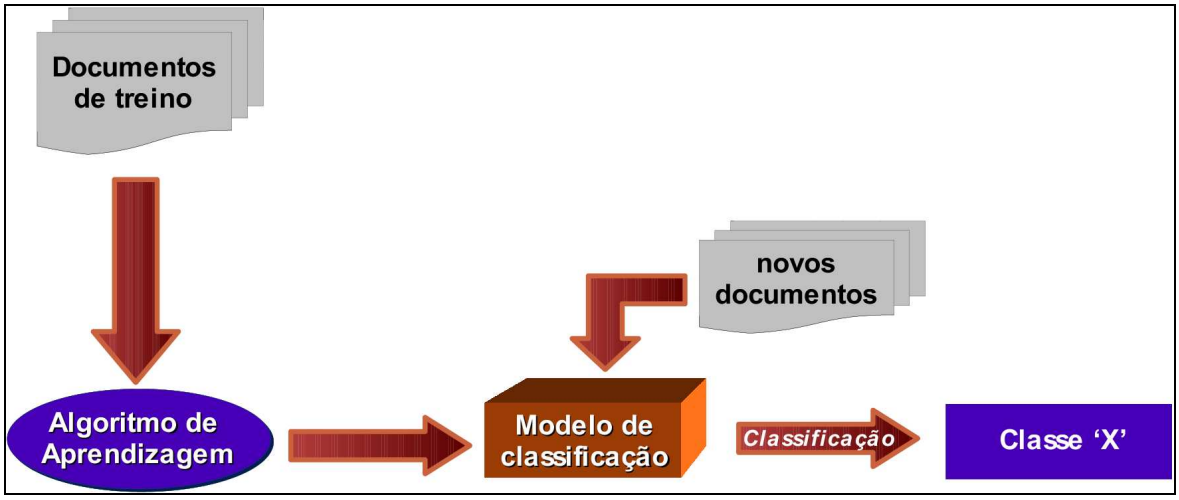

Figura 3.1: Visão geral do processo de classificação baseado em aprendizado de máquina

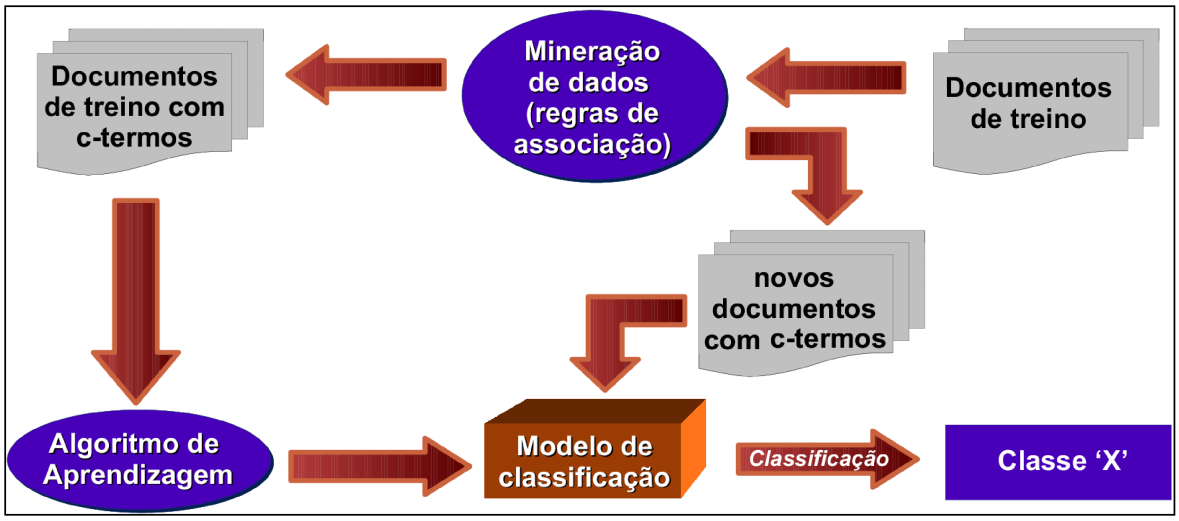

Figura 3.2: Visão geral do processo de extração de c-termos para subseqüente utilização em sistemas baseados em aprendizado de máquina

Nossa metodologia executa tarefas tanto no estágio de treino quanto no de teste. Começaremos descrevendo as tarefas executadas durante o treino do classificador, divididas em três passos:

O primeiro passo é a enumeração, quando determinamos como os s-termos contidos nos documentos devam ser combinados a fim de se gerar c-termos. É possível realizar a enumeração utilizando diversas estratégias. Nós começamos por meio de uma abor-

dagem *bottom-up*, criando pares de s-termos e então poderíamos combinar os pares em triplas e assim sucessivamente. Para essa tarefa, empregamos dois critérios a fim de determinar quais combinações devem ser geradas: a determinação do conjunto de s-termos utilizados e o suporte. O primeiro critério determina quais s-termos são os melhores para serem combinados a fim de se gerar c-termos relevantes. O segundo critério é uma medida de significância para c-termos e define o número mínimo de documentos nos quais um c-termo deve ocorrer a fim de ser considerado como relevante.

O segundo passo é o *ranking*, i.e., determinar os c-termos que possuam poder discriminativo significante, no sentido de que sejam bons indicadores de que o c-termo esteja fortemente associado a uma classe. A fim de determinar tais c-termos, geramos e avaliamos as regras de classificação associadas aos c-termos, i.e., regras que possuam c-termos como antecedentes e apenas a classe do documento como conseqüente. O processo de *ranking* é então baseado no critério de *Predominância*, que estima a pertinência de um documento ser incrementado com um c-termo Zaïane e Antonie (2002). Quanto menor for o número de classes em que um c-termo ocorre, maior será a *Predominância*. Dessa maneira, podemos obter um *ranking* das regras com base em sua *Predominância*, que quantifica o grau com que um c-termo está exclusivamente associado a uma dada classe. Formalmente, seja $T = \{t_1, t_2, t_3, \ldots, t_M\}$ o conjunto dos c-termos associados a uma coleção; $C = \{c_1, c_2, \ldots, c_K\}$ o conjunto das classes existentes na coleção; $df(t_i, c_j)$ o número de documentos de treino associados à classe c_j que contém t_i. A *Predominância* é então definida como se segue:

$$Predominância(t_i, c_j) = \frac{df(t_i, c_j)}{\sum_{j=1}^{K} df(t_i, c_j)}$$

Predominância é particularmente interessante porque ela pode ser usada para filtrar c-termos distribuídos irregularmente entre diversas classes, garantindo assim que incrementemos os documentos apenas com características discriminativas.

O terceiro passo é o incremento de documentos de treino, o qual visa a introduzir c-termos que auxiliam o classificador a realizar seu trabalho. Um ponto chave aqui é

decidir, para cada c-termo, se ele deve ser introduzido a um documento ou não. Nós então definimos a *Predominância* como um critério de *ranking* que indica quais c-termos possuem melhor poder discriminativo. Então, nós apenas introduzimos c-termos de alta *Predominância* a um documento que necessariamente possua os s-termos que compõe o c-termo.

Uma vez concluído o passo de incremento de documentos, podemos então empregar qualquer método de classificação, como SVM, Naïve-Bayes e kNN. O classificador resultante é então usado para classificar novos documentos.

As Figuras 3.3 e 3.4 ilustram como a extração de c-termos por meio do processo descrito pode ser de grande valia para enriquecer a representação de documentos textuais. Conforme podemos observar pela Figura 3.3, um documento-frase contendo a sentença apresentada na figura muito remotamente seria classificado em sua classe correta por meio da representação *bag of words* em razão da ambigüidade inerente a seus termos e das fortes associações entre os termos independentes a assuntos diversos. No exemplo em questão, o termo "Lula" prevalentemente está associado à categoria "Política" por ser o nome próprio de um político recorrentemente citado, ao passo que o termo "baía" tende a estar mais correlato a assuntos de "Geografia" por ser tema de ensino e pesquisa da área. Porém a semântica do texto aponta claramente para o tema "Ecologia". Por meio de extração de c-termos, podemos identificar relacionamentos semânticos entre s-termos, conforme demonstra a Figura 3.4, que identificou o relacionamento "Lula gigante". Nesse caso, ao possuirmos um acervo de ecologia contendo alguma descrição sobre esse animal, então poderíamos induzir classificadores a automaticamente criar modelos mais precisos de modo a classificar o documento à sua classe verdadeira.

Dessa forma, a fim de usar o modelo de classificação, também é necessário incrementar documentos de teste com c-termos. Nesse caso, porém, tendo em vista que não conhecemos a categoria a que pertence o documento, precisamos nos precaver de não extrair c-termos que representem mais um ruído do que um padrão propriamente útil. Sob essa restrição, nós adicionamos todos os c-termos induzidos pelos s-termos contidos

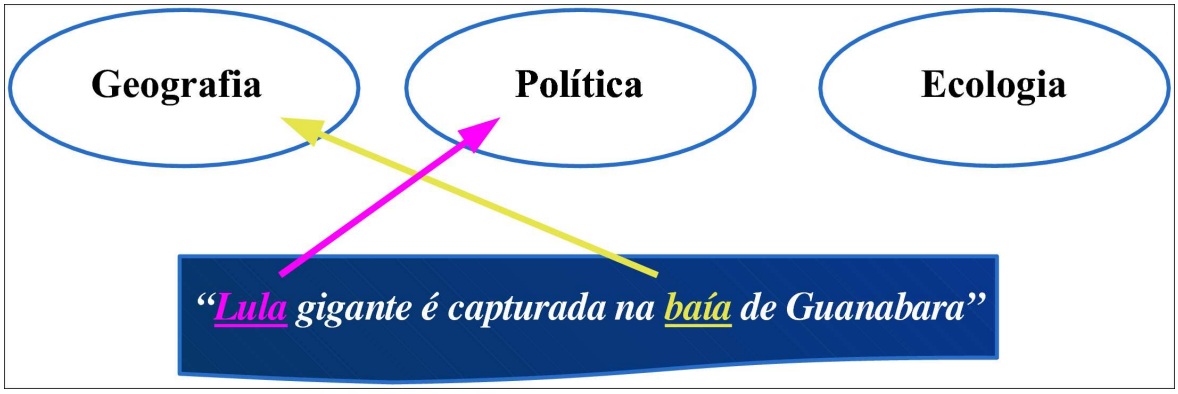

Figura 3.3: Importância da extração de c-termos: ambigüidade inerente aos s-termos.

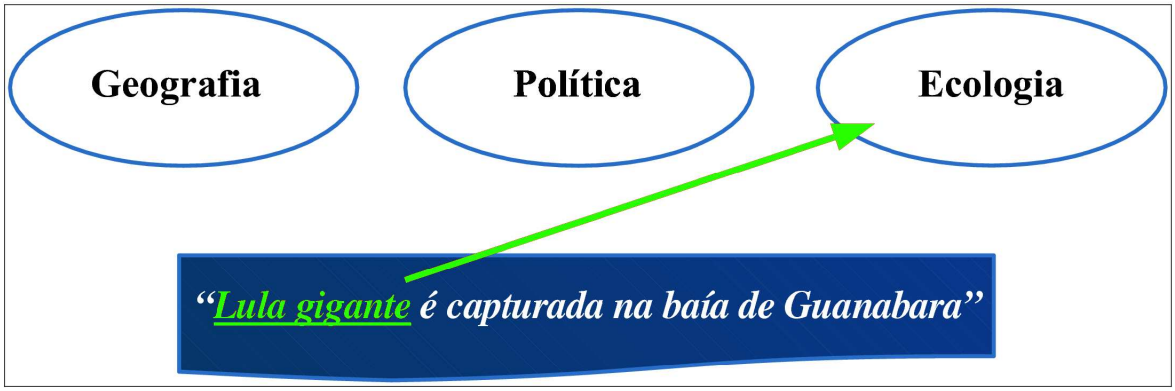

Figura 3.4: Importância da extração de c-termos: delimitação semântica por meio de c-termos.

no documento, desde que possuam alta *Predominância*. Utilizamos a *Predominância* como critério de seleção de c-termos porque as coleções de refências utilizadas possuem documentos que pertencem exclusivamente a uma classe e a *Predominância* quantifica diretamente a pertinência de um c-termo pertencer a apenas uma classe. As Figuras 3.5 e 3.6 exemplificam dois aspectos complementares importantes da *Predominância* na seleção de características. Por meio da Figura 3.5, pode ser notado que existem c-termos extraídos que são naturalmente pouco discriminantes, pois podem estar associados a assuntos distintos e pouco ajudariam ou até depreciariam a qualidade do modelo de classificação induzido. Por outro lado, a Figura 3.6 exemplifica como um filtro baseado em *Predominância* auxilia a selecionar apenas padrões úteis à classificação ao identificar c-termos específicos no que diz respeito à homogeneidade de classes.

Com base nessas observações, discutimos, a seguir, alguns aspectos pertinentes

Figura 3.5: Importância da *Predominância* na seleção de c-termos: descarte de c-termos ruidosos.

Figura 3.6: Importância da *Predominância* na seleção de c-termos: seleção de c-termos úteis.

sobre (i) detalhes de implementação, (ii) eficiência da metodologia de extração de c-termos, bem como (iii) a estratégia adotada para inspecionar a influência dos fatores experimentais sobre os c-termos extraídos e sobre a qualidade da classificação.

3.2 Eficiência e Detalhes de Implementação

Nesta seção, discutimos alguns detalhes de implementação de nossa estratégia, assim como aspectos práticos que surgem durante a o processo de extração de características.

Uma questão fundamental é a explosão combinatorial associada à enumeração de c-termos, como conseqüência da alta dimensionalidade dos dados, uma vez que cada s-termo é uma dimensão. Conforme discutido a seguir, empregamos uma heurística de três fases para reduzir o custo computacional do primeiro passo da metodologia.

A primeira fase é ordenar os s-termos a título de enumeração com base em seu respectivo poder discriminativo. Usamos ganho de informação (infogain)Mitchell (1997a) como um critério, mas outros se aplicam. Essa abordagem se baseia na observação de que combinar os s-termos mais informativos constroem c-termos mais informativos em comparação à combinação de s-termos pouco informativos, conforme discutido no capítulo de resultados experimentais. Como resultado, esse fato torna possível definir um subconjunto menor de vocabulário com o objetivo de construir uma quantidade significativa de bons c-termos.

A segunda fase determina a porção do *ranking* de s-termos que deve ser usada para extração de c-termos. Para tal, definimos um parâmetro N, que representa o número de s-termos consideradas em ordem decrescente de ganho de informação. Empiricamente, demonstramos que aumentar N tende a gerar mais c-termos úteis e conseqüentemente acarretar melhores classificações. Contudo, incluir s-termos pouco discriminativos no subconjunto delimitado por N não incrementa significativamente a qualidade da classificação, muito embora aumente o custo computacional e adiciona c-termos ruidosos ao modelo.

A terceira fase é a enumeração das características que extraímos. Para tanto, neste trabalho, enumeramos apenas pares de s-termos, uma vez que a interação entre apenas dois termos é suficiente na maioria dos casos para uma boa desambiguação de um dos termos Kaplan (1955); Choueka e Lusignan (1985), ajudando, assim, a especificar melhor o verdadeiro significado dos termos. Então, dado todo o vocabulário de treino contendo V s-termos, tal que V é tipicamente maior que N, nós reduzimos o custo (limite superior de possíveis c-termos introduzidos na coleção) de $(2^V - V)$ para $\frac{N*(N-1)}{2}$ (combinações de tamanho 2), mas mesmo assim ainda aptos a aperfeiçoar a eficácia do classificador, conforme demonstrado experimentalmente.

Uma idéia análoga foi usada para verificar o impacto do fator *Predominância* em nosso processo de extração de características. A fim de selecionar apenas c-termos caracterizados por possuir alto poder discriminativo, nunca usamos valores menores

que 70%, tendo em vista que nosso objetivo é enfatizar relacionamentos intra-classe ao passo que minimizamos similaridades inter-classe.

Capítulo 4

Resultados Experimentais

Neste capítulo, apresentamos os resultados obtidos por meio de nossa metodologia para extração de c-termos. Para tanto, primeiramente introduzimos, na seção 4.1, as coleções e métodos de classificação usados. Logo em seguida, na seção 4.2, debatemos a metodologia de avaliação e discutimos formas para interpretação dos resultados. Por fim, na seção 4.3, avaliamos os resultados obtidos.

4.1 Coleções e Métodos

Para experimentalmente avaliar nossa estratégia, empregamos quatro coleções de texto de referência comumente discutidas na literatura: 20 Newsgroups 18828[1], OHSUMED 18302[2], Reuters-21578 8C[3] e ACM 11C[4]. Em todas as coleções, *stopwords* foram removidas, assim como documentos com múltiplas categorias, exceto na coleção 20 Newsgroups 18828, a qual já é unirotulada.

A versão da coleção 20 Newsgroups empregada possui 18.828 documentos, distribuídos quase uniformemente entre 20 categorias de fórum de discussões. O número de documentos por classe varia de 628 a 999, conforme ilustra a Figura 4.1. Documen-

[1]at: http://people.csail.mit.edu/jrennie/20Newsgroups
[2]at: http://ai-nlp.info.uniroma2.it/moschitti/corpora.htm
[3]at: http://kdd.ics.uci.edu/databases/reuters21578
[4]at: http://www.acm.org

tos correspondem a mensagens eletrônicas enviadas a fóruns de discussão sobre temas como ciências, religião e política.

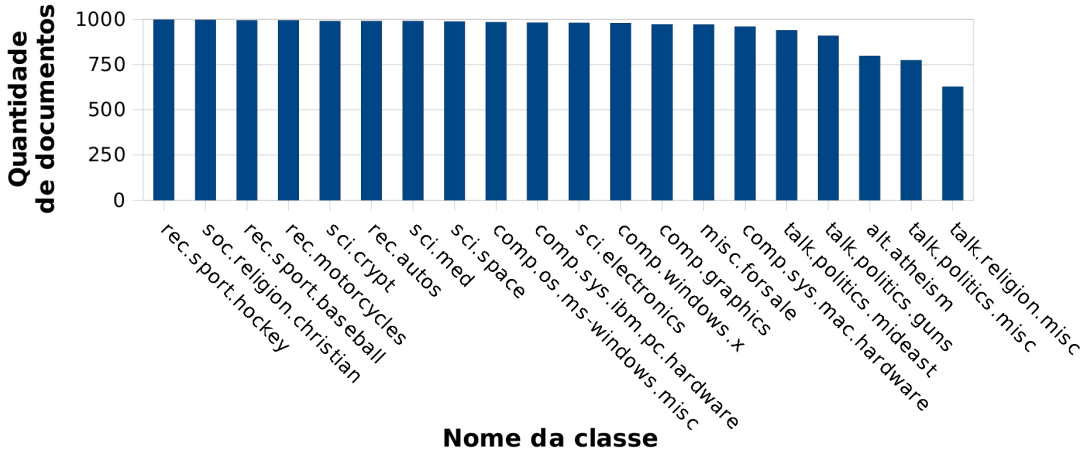

Figura 4.1: Distribuição de documentos por classes da coleção 20 Newsgroups 18828.

A coleção OHSUMED 18302 contém documentos médicos coletados em 1991 relativos a 23 classes sobre doenças diversas. A versão usada possui 18.302 documentos, distribuídos muito irregularmente entre categorias que variavam entre 56 a 2.876 documentos por categoria, conforme ilustra a Figura 4.2.

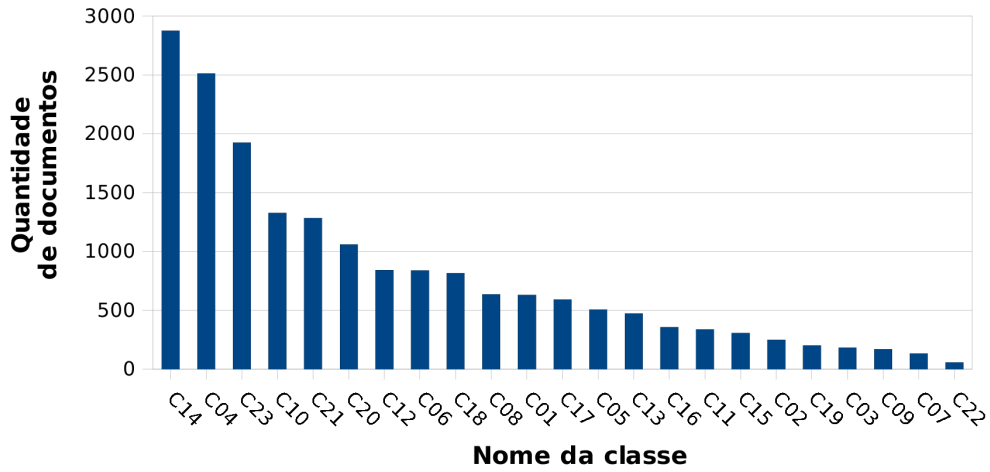

Figura 4.2: Distribuição de documentos por classes da coleção OHSUMED.

Para a coleção Reuters-21578 8C, foram usados 8.184 documentos que representam notícias, compreendendo título, corpo do texto, localização geográfica dentre outros

atributos. Como mencionado anteriormente, documentos multirótulo foram removidos. Além disso, classes extremamente específicas contendo apenas 1 documento foram descartadas. Isso resultou em uma coleção com 8 classes. A distribuição de documentos nessa coleção é muito irregular; o número de documentos por classe varia de 113 a 3.930. A Figura 4.3 ilustra essa distribuição.

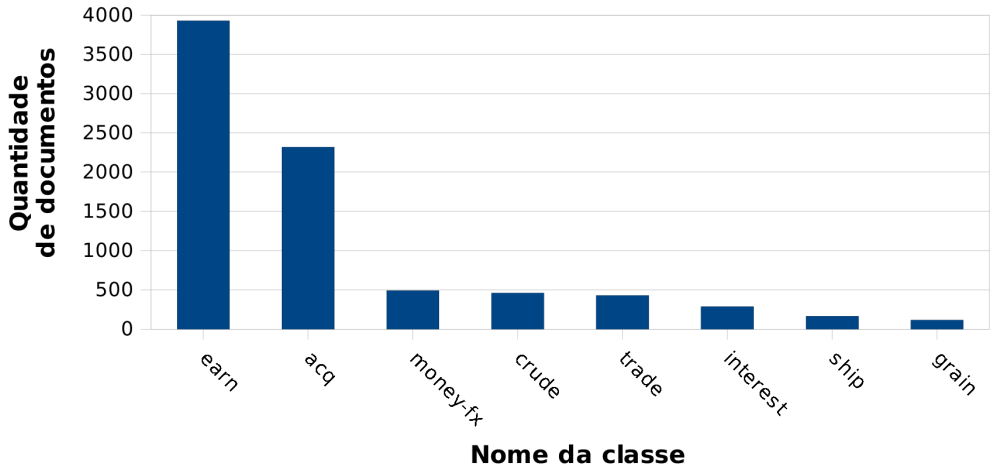

Figura 4.3: Distribuição de documentos por classes da coleção Reuters-21578 8C.

A coleção ACM 11C é uma sub-coleção de artigos científicos da biblioteca digital da *Association for Computing Machinery* (ACM). Todo o texto contido no título ou resumo do artigo, quando disponíveis, foram usados como índices dos documentos. A coleção resultante é um conjunto com 29.570 documentos, sem *stopwords*. Os documentos são classificados nas 11 categorias existentes da biblioteca digital da ACM. Dentre todas coleções testadas, esta é a mais irregular: a quantidade de documentos por classe varia de 93 a 7.585, conforme pode ser analisado pela Figura 4.4.

Como métodos de classificação de nossos experimentos, foram utilizados os algoritmos kNN e Naïve-Bayes implementados pelo arcabouço de classificação *libbow* McCallum (1996). Além desses métodos, também foi empregado o SVM-Perf Joachims (2006), um pacote que implementa uma versão eficiente do Support Vector Machine (SVM) e que pode ser treinado em tempo linear. Além disso, foi utilizada a abordagem um-contra-todos Vapnik (1998) a fim de adaptar o classificador binário do SVM para

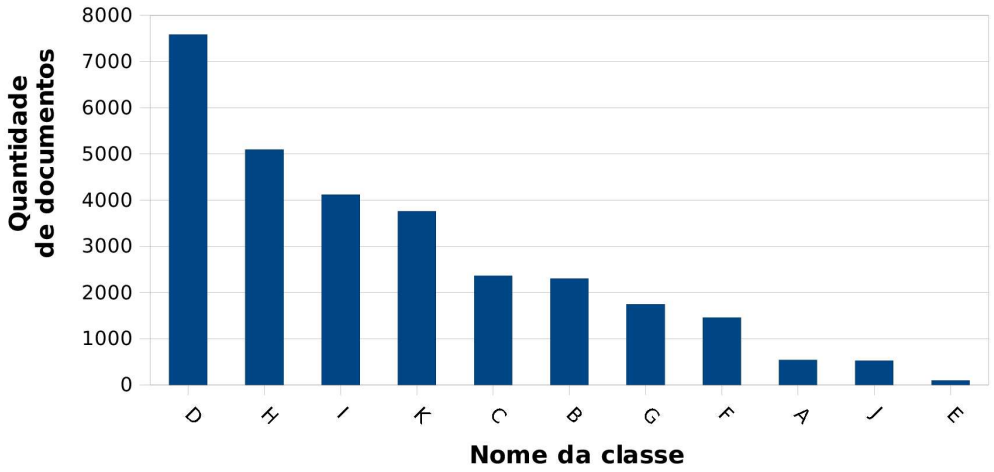

Figura 4.4: Distribuição de documentos por classes da coleção ACM 11C.

classificação multi-classe, uma vez que nossas coleções possuem mais de duas classes.

O objetivo de empregar esses vários classificadores em nossos experimentos foi analisar a generalidade de nossa metodologia, tendo em vista que cada um desses algoritmos é inspirado em estratégias completamente distintas, variando de abordagens probabilísticas a baseada em espaço vetorial.

Outro aspecto importante é comparar a eficácia relativa entre diferentes classificadores aplicados sobre as coleções de texto incrementadas fornecidas pela nossa metodologia de extração de características.

4.2 Metodologia de Avaliação

Nesta seção discutimos as métricas de avaliação usadas nos experimentos, bem como os motivos de empregá-las para avaliar os parâmetros investigados neste trabalho.

4.2.1 Métricas de Avaliação

A eficácia de nosso método foi avaliada utilizando métricas padrão da área de recuperação de informação: precisão, revocação e F_1 Lewis (1995).

Revocação, r, é definida como a fração dos documentos de uma classe corretamente

classificados. Precisão, p, por sua vez, é definida como a fração de documentos corretamente classificados dentre todos os documentos atribuídos pelo classificador a uma classe.

Portanto, uma revocação perfeita para uma dada classe é alcançada caso todos os documentos da classe em questão sejam nela classificados, independentemente se outros documentos de outras classes sejam também atribuídos à ela. Por outro lado, uma boa precisão é atingida ao evitar que documentos oriundos de diferentes classes sejam atribuídos a uma só.

Em decorrência dessa multiplicidade de aspectos de avaliação, uma abordagem mais usual para avaliar a eficácia da classificação é F_1, uma combinação entre precisão e revocação dada pela média harmônica dessas duas métricas. Como resultado, dada uma classe, c_j, seu respectivo escore F_1 é matematicamente definido como:

$$F_1(c_j) = \frac{2p_j r_j}{p_j + r_j}$$

Nós então sumarizamos nossos resultados gerais por meio de duas métricas tradicionalmente usadas para esse fim: macro-média F_1 e micro-média F_1. A macro-média F_1 mensura a eficácia da classificação com base em uma média aritmética dos F_1 de cada classe. Por sua vez, a micro-média F_1 pondera os F_1 de cada classe com base na representatividade da classe na coleção de acordo com o número de documentos em cada classe.

Em termos qualitativos, a macro-média F_1 parte do princípio que cada classe é igualmente importante, e, por isso, atribui-lhes pesos iguais, independente da quantidade de documentos contidos em cada uma. Por outro lado, a micro-média F_1 parte da premissa que cada documento é igualmente importante. Como resultado, se a maioria das classes em uma coleção contiver proporcionalmente poucos documentos em relação ao todo, então a macro-média F_1 é uma métrica tipicamente mais relevante, pois são raros os casos em que é adequado subestimar a importância de uma vasta diversidade de classes. Caso contrário, a micro-média F_1 é uma métrica tipicamente

mais significativa.

Outro ponto importante a se ressaltar sobre a avaliação é que não consideramos a freqüência intra-documento tanto dos s-termos como dos c-termos a fim de evitar que um tipo de característica fosse favorecida em detrimento de outra. Nós procedemos dessa forma porque definir a melhor estratégia para pesagem de c-termos é um novo problema por si só, o qual não é tratado neste estudo e deixado como trabalho futuro.

4.2.2 Analisando o Tamanho do Vocabulário para Geração de c-termos

Em nossos experimentos, para todas as coleções, foi empregada uma divisão na proporção de 70%-30% na amostragem de documentos de treino e teste, respectivamente, a fim de facilitar a comparação dos resultados experimentais entre as coleções. Os c-termos são gerados em função apenas dos s-termos de maior ganho de informação encontrados em documentos de treino, em que cada documento é definido como uma transação para efeito de geração de co-ocorrências. Essa estratégia é baseada na observação de que a combinação entre s-termos mais discriminativos resultam em melhores c-termos.

Nós constatamos essa observação ao comparar o poder discriminativo entre os c-termos gerados a partir da melhor metade de s-termos, ordenados conforme ganho de informação, e os c-termos gerados a partir da metade de baixo desse mesma lista ordenada de s-termos. Os resultados, constatados em todas as coleções de referência, demonstraram que nenhum c-termo gerado por s-termos de baixo ganho de informação se posicionou entre a metade de s-termos de maior ganho de informação. Em tempos práticos, isso significa que os c-termos construídos com s-termos de baixo ganho de informação não são melhores que metade dos s-termos já existentes. Por outro lado, mais de 90% da metade superior da lista ordenada de s-termos foi substituída por algum c-termo após combinar s-termos gerados a partir dessa metade da lista.

Esse achado sugere que não é necessário utilizar todo o vocabulário de s-termos

Coleção	Vocabulário de treino completo (V)	N		
20 Newsgroups 18828	$V_1 = 9,1 \times 10^4$	V_1	$\frac{1}{2}V_1$	$\frac{1}{4}V_1$
ACM 11C	$V_2 = 4,8 \times 10^4$	V_2	$\frac{1}{2}V_2$	$\frac{1}{4}V_2$
OHSUMED 18302	$V_3 = 3,9 \times 10^4$	V_3	$\frac{1}{2}V_3$	$\frac{1}{4}V_3$
Reuters-21578 8C	$V_4 = 2,1 \times 10^4$	V_4	$\frac{1}{2}V_4$	$\frac{1}{4}V_4$

Tabela 4.1: Valores de N para cada coleção usada em nossos experimentos.

para efeito de geração de novas características relevantes. Como resultado, uma questão que naturalmente surge em função disso resguarda sobre o melhor tamanho desse vocabulário. Em função disso, nós então analisamos o impacto do tamanho do vocabulário (N), ordenado em ordem decrescente de ganho de informação, para geração de c-termos. Assim, inspecionamos diferentes valores para N em cada coleção, conforme mostrado na Tabela 4.1, com o objetivo de avaliar o custo-benefício de se variar esse fator.

É interessante observar que, para o maior $N = 9.1 \times 10^4$, se associarmos s-termos par a par (i.e., todas as possíveis combinações dois a dois), então teoricamente mais de 4 bilhões de c-termos potencialmente podem ser gerados. Contudo, dada a grande esparsidade da co-ocorrência de s-termos nos documentos, na prática encontramos quantidades construídas de c-termos menores que o limite teórico superior em duas ordens de magnitude. Isso sem contar que esse valor ainda deve diminuir em decorrência de nosso processo de filtragem de c-termos ruidosos, o qual é apresentado a seguir.

4.2.3 Avaliação do Poder Discriminativo dos c-termos

Medir a efetividade de um c-termo está diretamente relacionado à importância relativa desse c-termo perante à informação pré-existente. Em termos gerais, precisamos analisar se esse c-termo introduz mais um ruído à coleção do que uma informação útil. Assim, para permitir uma caracterização mais cuidadosa sobre essa questão, executamos experimentos utilizando diferentes níveis de *Predominância*. Conforme discutido anteriormente, *Predominância* é uma métrica que promove c-termos que enfatizam

dissimilaridades inter-classes. Dessa forma, testamos quatro limiares mínimos de *Predominância*: 70%, 80%, 90% e 100%. Note que, para o nível 100%, apenas c-termos que ocorrem em documentos de apenas uma classe foram gerados. Para os outros níveis, apenas c-termos cuja *Predominância* fosse maior ou igual ao respectivo limiar foram construídas.

Outra questão que deve significativamente influenciar a tarefa de classificação é a freqüência da característica. De acordo com Cheng et al. (2007), o uso de padrões infreqüentes em classificação previne o modelo de classificação de se generalizar bem, uma vez que esse tipo de padrão é baseado em observações pouco significativas. Na literatura, esse fenômeno é referido como *overfitting*. Ainda de acordo com o mesmo trabalho, para classificação, o poder discriminativo de um padrão está intimamente relacionado a seu suporte, i.e., sua ocorrência por entre diferentes registros. Todavia, um ponto contrastante a se ressaltar é que muitos algoritmos de classificação, senão a grande maioria, emprega a métrica *idf* como parte de uma estratégia para pesar a relevância das características, baseando-se na intuição de que quão mais documentos um termo ocorre, mais genérico ele é Sebastiani (2002), e, portanto, menos útil para especificar assuntos e classificar documentos.

Além disso, em coleções de texto, é comum existirem diversas pequenas classes contendo prevalentemente características de baixo suporte em virtude da relativa pequena quantidade de documentos contidos nessas classes. Conseqüentemente, sobrevalorizar o suporte como métrica de relevância das características extraídas tende a favorecer as classes maiores (com mais documentos). Isso ocorre pois a expectativa de classes pequenas serem capazes de gerar características tão freqüentes como as das classes grandes é muito mais baixa. Isso então sugere que a estratégia apresentada por Cheng et al. (2007) para medir a relevância das características extraídas por meio do suporte possui potencial teórico para ocasionar *overfitting*.

Isso nos levou a testar a hipótese de que características de baixo suporte possam ser importantes para aperfeiçoar a qualidade de classificação principalmente para pequenas

classes. Nessa linha de raciocínio, nós então definimos um fator para quantificar o suporte mínimo que um c-termo deve possuir a fim de ser selecionado para o modelo de classificação. Nós nomeamos esse fator de Min_Supp e experimentamos quatro valores para ele: 2, 4, 6 and 8.

4.2.4 Configuração Experimental

Em resumo, nossos experimentos testaram os seguintes fatores anteriormente discutidos: N, *Predominância* e Min_Supp, empregando uma configuração experimental do tipo fatorial Jain (1991), que considera todas as possíveis combinações entre valores de níveis de cada fator.

Nós testamos essa configuração 30 vezes com divisões aleatórias entre documentos de treino e teste para cada uma das quatro coleções. Então, os classificadores são aplicados a cada um dos 30 conjuntos incrementados com c-termos, de forma que pudéssemos comparar a média desses resultados contra a classificação original sem o uso de c-termos. A fim de se obter resultados mais precisos, os experimentos em que novas características não foram construídas (*bag of words*) também foram repetidos 30 vezes. Assim, um teste-t de dupla cauda com 99% de confiança foi aplicado para comparar os resultados usando ou não extração de características.

4.3 Análise dos Resultados

Apresentamos nesta seção os melhores resultados obtidos em nossos experimentos. As tabelas 4.2 e 4.3 apresentam os resultados da classificação utilizando a macro-média F_1 e micro-média F_1, respectivamente. Em ambas as tabelas, a coluna com o cabeçalho "l.b." (linha de base) representa a qualidade da classificação ao utilizar apenas s-termos, ao passo que a coluna com o cabeçalho "e.c." (extração de características) apresenta valores para a classificação que também emprega c-termos. A coluna com o título "v.r." (variação relativa), por sua vez, quantifica, para um dado classificador, se houve

ganho positivo ou negativo ao empregar c-termos na classificação. Por fim, a coluna com o título "tt" (teste-t) atesta se houve ganhos estatisticamente significativos, dada uma confiança de 99% em um teste-t de dupla cauda, entre a classificação envolvendo a extração de características e a classificação mais tradicional, que envolve apenas s-termos. Os desvios-padrão relativos aos resultados apresentados são tabelados no Apêndice B.

Devemos ressaltar que estamos primordialmente interessados em analisar quão abrangente o processo de extração de características proposto melhora a qualidade de um classificador, e não em comparar a eficácia entre classificadores distintos. Portanto, ao analisar as tabelas 4.2 e 4.3, devemos focar nas variações da qualidade de classificação entre um dado algoritmo aplicado sobre a coleção original (*bag of words*) contra a qualidade do mesmo algoritmo sobre a coleção incrementada com os c-termos extraídos.

Macro Média F_1	*k*NN				Naïve-Bayes				SVM			
	$macF_1$ (%)				$macF_1$ (%)				$macF_1$ (%)			
COLEÇÕES	l.b.	e.c.	v.r.	tt	l.b.	e.c.	v.r.	tt	l.b.	e.c.	v.r.	tt
20 newsgroups	59,5	81,6	+37	▲	88,0	90,0	+2,3	▲	90,0	91,9	+2,0	▲
ACM 11C	34,0	47,7	+40	▲	53,1	55,0	+3,4	▲	56,6	58,5	+3,3	▲
OHSUMED	38,5	57,3	+49	▲	32,1	49,8	+55	▲	61,6	68,2	+10,7	▲
Reuters-21578	79,2	86,2	+8,8	▲	80,4	88,9	+10,5	▲	91,4	92,7	+1,4	▲

Tabela 4.2: Impacto da extração de c-termos sobre a macro-média F_1.

Micro Média F_1	*k*NN				Naïve-Bayes				SVM			
	$micF_1$ (%)				$micF_1$ (%)				$micF_1$ (%)			
COLEÇÕES	l.b.	e.c.	v.r.	tt	l.b.	e.c.	v.r.	tt	l.b.	e.c.	v.r.	tt
20 newsgroups	55,3	80,7	+46	▲	88,7	90,4	+1,9	▲	90,2	91,9	+2,0	▲
ACM 11C	45,9	62,9	+37	▲	69,9	72,5	+3,8	▲	70,7	72,2	+2,1	▲
OHSUMED	50,8	64,4	+27	▲	54,4	65,4	+20	▲	68,3	73,5	+7,7	▲
Reuters-21578	92,7	94,3	+1,8	▲	92,9	95,1	+2,3	▲	96,0	96,4	+0,4	●

Tabela 4.3: Impacto da extração de c-termos sobre a micro-média F_1.

O primeiro aspecto importante a ser observado nas Tabelas 4.2 e 4.3 é que nossa estratégia de extração de características resultou em melhorias nos classificadores em

quase todos os cenários testados quando comparados aos classificadores que não empregaram extração de características. Para macro-média F_1, houve ganhos estatisticamente significativos em todos os 12 resultados.

Em termos de micro-média F_1, houve ganhos estatisticamente significativos ao empregar extração de características em 11 dos 12 resultados. Na exceção, houve equivalência estatística apontada pelo teste-t.

Um ponto interessante a se observar nas Tabelas 4.2 e 4.3 é que os maiores ganhos foram obtidos em coleções mais difíceis de se classificar, em que as linhas de base usando-se apenas s-termos não foram muito eficazes. Por exemplo, para as coleções ACM 11C e OHSUMED, as quais são mais difíceis de se classificar ao usar apenas s-termos, os ganhos do SVM após usar o processo de extração de características foi mais significativo do que para as outras duas coleções.

Além disso, mesmo nos casos em que os resultados das linhas de base eram altos e conseqüentemente melhorias tendem a ser mais limitadas, foram alcançados ganhos significativos. Vide o caso do melhor classificador observado, SVM, em que alcançamos ganhos estatisticamente significativos em 7 de seus 8 resultados, sendo o melhor ganho de 10,7% na macro-média F_1 sobre a coleção OHSUMED.

4.3.1 Analizando os Melhores Resultados

Uma questão fundamental a ser levantada é que, conforme podemos observar ao analisar ambas as Tabelas 4.2 e 4.3, a metodologia proposta neste trabalho tende a melhorar mais significativamente a macro-média F_1 do que a micro. Isso implica que nossa estratégia tipicamente aperfeiçoa a qualidade de classificação mais intensamente sobre classes pequenas do que sobre classes grandes, conforme a quantidade de documentos na classe. Novas evidências sobre esse fato são mostradas na Figura 4.5. Nesse gráfico, para cada coleção, nós agrupamos suas respectivas classes em três quantis (classes pequenas, médias e grandes), com base no número de documentos em cada classe. Então, para cada quantil, tiramos a média do F_1 de suas respectivas classes e comparamos esse

valor ao do quantil respectivo da linha de base. Os resultados da Figura 4.5 claramente confirmam nossa hipótese, verificada em todas as coleções.

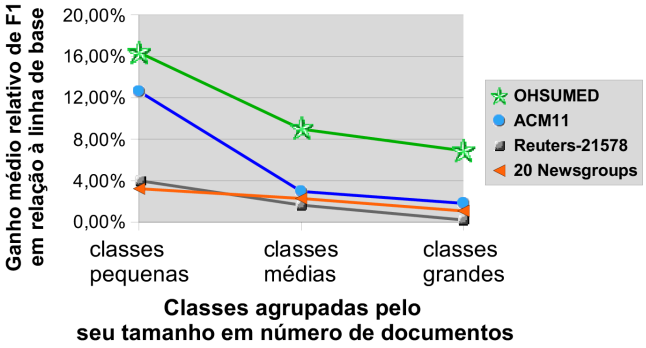

Figura 4.5: Ganhos do SVM fragmentados pelo tamanho das classes.

Esses resultados são, de certa forma, contra-intuitivos, uma vez que é mais provável construir mais características de classes grandes do que a partir de classes pequenas. Contudo, o que ocorreu é que novos c-termos que foram incorporados por nossa metodologia ajudaram a reduzir o efeito de desbalanceamento entre o tamanho das classes. De fato, quão mais desbalanceada for uma coleção, mais as classes pequenas se beneficiam da extração de características. Na Tabela, 4.4, a coluna "desbalanceamento da coleção" foi medida dividindo-se o número de documentos da maior classe sobre o número da menor. Os ganhos relativos "$\Delta ganho$" foram computados ao subtrair o ganho médio em F_1 das classes grandes do ganho médio das classes pequenas. Com base nisso, a Tabela 4.4 demonstra que houve uma correspondência direta entre desbalanceamento e ganhos. Um motivo importante para esses resultados é devido ao uso do fator Suporte Mínimo (Min_Supp), o qual é analisado a seguir.

4.3.2 Analisando Suporte

O fator Suporte Mínimo, (Min_Supp) é um critério para delimitar a quantidade mínima de documentos em que um c-termo deve ocorrer a fim de ser selecionado para o modelo de classificação. Como conseqüência, valores altos para Min_Supp filtram c-termos de baixo suporte. No entanto, padrões baseados em observações infreqüentes

Coleção	Desbalanceamento da Coleção	$\Delta ganho$
ACM 11C	81,6	10,8%
OHSUMED 18302	51,4	9,5%
Reuters-21578 8C	34,8	3,8%
20 newsgroups 18828	1,6	2,1%

Tabela 4.4: Relação entre ganhos de classes pequenas e grandes de acordo com o desbalanceamento da coleção.

são mais decisivos para pequenas classes, as quais, em relação às grandes, são incapazes de gerar c-termos de alto suporte. Corroborando com essa hipótese, a Figura 4.6, fragmentada pelo tamanho das classes de forma análoga à Figura 4.5, revela que c-termos de baixo suporte são notavelmente mais críticos para corretamente classificar documentos de classes pequenas. Nós usamos a coleção OHSUMED para sumarizar esse resultado, mas resultados análogos foram encontrados em todas as 4 coleções de referência. As figuras das outras 3 coleções de referência encontram-se no Apêndice A.

Como resultado, isso está de acordo com nossas hipóteses de que c-termos de baixo suporte ajudam a evitar overfitting, dada sua propriedade de reduzir desbalanceamento.

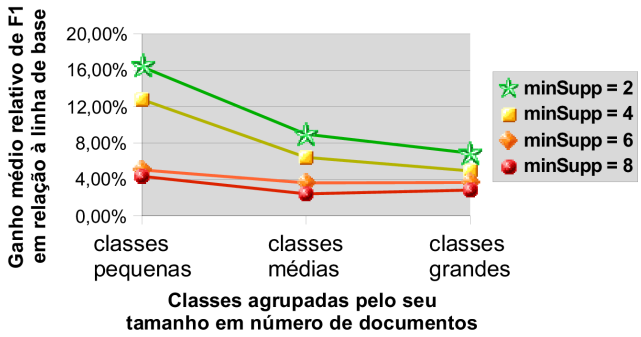

Figura 4.6: SVM: ganhos agrupados conforme o tamanho da classe para cada limiar mínimo de suporte ($MinSupp$). Coleção: OHSUMED

4.3.3 Analisando o Tamanho do Vocabulário - N

Nesta subseção, mostramos que os melhores s-termos no do que diz respeito a ganho de informação são bons o suficiente para gerar c-termos de alta qualidade. No intuito

de sustentar esse argumento, conduzimos um experimento em que variamos a quantidade de s-termos (N) considerados para extração de c-termos e calculamos os ganhos obtidos em relação à linha de base (classificação usando-se apenas s-termos). Então repetimos esse experimento para cada coleção e variamos N de $V/4$ até V, em que V é a quantidade de s-termos encontrados no conjunto de treino da coleção. Os resultados desse experimento estão sumarizados na Figura 4.7, a qual demonstra que, aumentar N tipicamente resulta em apenas ganhos marginais na eficácia da classificação. Como conseqüência, podemos usar um N relativamente pequeno comparado ao V e, assim, reduzir o número de c-termos gerados sem sacrificar a qualidade da classificação.

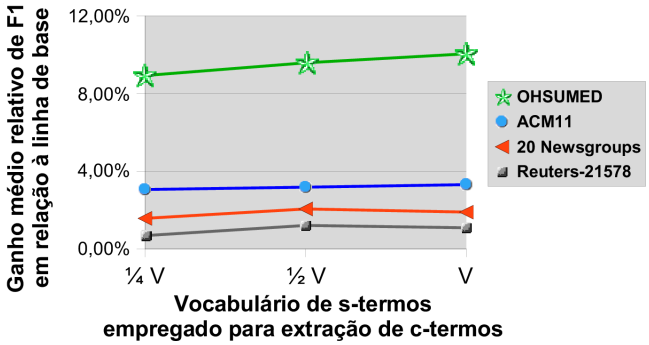

Figura 4.7: SVM: ganhos em termos de macro-média F_1 em função do tamanho do vocabulário de s-termos usados para extração de c-termos.

4.3.4 Analisando *Predominância*

Nesta subseção, analisamos o efeito da *Predominância*, a métrica usada para selecionar c-termos. A idéia por trás dessa métrica é selecionar apenas os c-termos mais discriminativos a fim de ajudar a especificar o sentido original dos termos e induzir uma melhor classificação.

Um aspecto importante sobre a *Predominância* é descobrir um valor que produza o melhor compromisso entre uma menor quantidade de c-termos extremamente discriminativos e uma maior quantidade de c-termos moderadamente discriminativos. Infelizmente, no entanto, não existe um único valor ótimo global para todas as coleções. Isso acontece em virtude de, embora um valor fixo de *Predominância* selecione c-termos

Coleção	Desorganização
Reuters-21578	$0,7 \times 10^3$
20 newsgroups	$1,1 \times 10^3$
OHSUMED	$1,2 \times 10^3$
ACM 11C	$3,7 \times 10^3$

Tabela 4.5: Desorganização intrínsica das coleções de referência

com poder discriminativo conhecido, o impacto desses c-termos na classificação difere em função das características pré-existentes de cada coleção original. Isso ocorre porque coleções diferentes possuem propriedades distintas: (i) algumas são mais *caóticas*, com seu vocabulário se dispersando por entre diversas classes; (ii) outras coleções são mais *regulares*, pois possuem inúmeros s-termos que apresentam alta *Predominância* e caracterizam claramente uma classe. Dessa forma, de um ponto de vista prático, c-termos moderadamente discriminativos tendem a aperfeiçoar a classificação em coleções mais caóticas, mas provavelmente são inúteis se introduzidos em coleções já muito bem organizadas. Assim, valores diferentes de *Predominância* relacionam-se com essas propriedades intrínsicas das coleções.

Nossas análises apontaram que, para aperfeiçoar uma coleção já muito organizada, apenas c-termos extremamente discriminativos devem ser introduzidos nos documentos. Por outro lado, em relação a uma coleção mais caótica, introduzir uma maior quantidade de c-termos moderadamente discriminativos tende a ser mais eficaz para aperfeiçoar a classificação. Como resultado, medimos a desordem intrínsica de cada coleção [5] a fim de estabelecer um relacionamento entre os melhores valores de *Predominância* que funcionam para cada coleção. Essa informação é mostrada na Tabela 4.5. Essa tabela aponta que OHSUMED e ACM 11C são coleções relativamente menos organizadas, ao passo que Reuters-21578 8C e 20 Newsgroups 18828 são relativamente organizadas.

Conseqüentemente, é mais provável que apenas c-termos muito discriminativos (va-

[5] definida como o inverso da média do ganho de informação de seus s-termos da coleção de treino

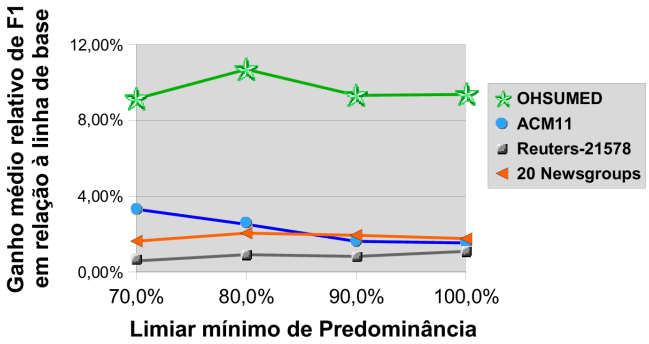

Figura 4.8: SVM: ganhos de macro-média F_1 ao variar limiares de *Predominância*

lores altos para o limiar de *Predominância*) sejam capazes de melhorar a classificação da Reuters-21578 8C e 20 Newsgroups 18828. Porém, para OHSUMED e ACM 11C, a introdução de c-termos moderadamente discriminativos tende a ser melhor para aperfeiçoar sua classificação. A Figura 4.8 valida essa hipótese. Ela mostra que ganhos associados a coleções mais organizadas são mais intensos quando limiares mais altos de *Predominância* são empregados, ao passo que em coleções menos organizadas, ganhos são intensos em limiares menores de *Predominância*. Com base nisso, quando uma nova coleção precisasse ser classificada utilizando c-termos, saberíamos associar, a partir do grau de desordem dessa coleção, um limiar de *Predominância* adequado a ser empregado para a classificação.

Capítulo 5

Discussão

Os resultados anteriores mostraram que a metodologia proposta de extração de características é capaz de proporcionar ganhos na eficácia da classificação. Neste capítulo, investigamos mais profundamente como isso foi possível em nível dos algoritmos de classificação utilizados e mostramos evidências empíricas de que algumas de nossas hipóteses iniciais estavam corretas, i.e., que alguns c-termos realmente possuem maior poder discriminativo que alguns s-termos para fins de CAT.

5.1 SVM

Além de ter atingido os melhores resultados em nossos experimentos, SVM nos fornece um método geral para investigar a qualidade de seus termos. SVM é capaz de aprender uma função linear a partir da qual é possível atribuir pesos para cada s-termo ou c-termo. Assim, dada uma classe, pesos altamente positivos associados a termos indicam que documentos que possuírem tais termos estão provavelmente relacionados à classe em questão, ao passo que pesos negativos indicam o inverso. Em outras palavras, esse método fornece uma forma para o SVM ordenar os termos de modo a nos permitir analisar a relativa importância entre s-termos e c-termos para o SVM.

A Tabela 5.1 demonstra alguns pesos de termos, do maior valor ao menor, para a classe “Pathological Conditions, Signs and Symptoms”, encontrada na coleção OH-

Tipo do termo	Peso	Termo
c-termo	0,167717	{postoperative, pain}
s-termo	0,160238	{deletion}
s-termo	0,150482	{graft}
s-termo	0,142706	{ventricular}
s-termo	0,136462	{bypass}
s-termo	0,135890	{arrhythmias}
...	...	...
c-termo	0,086127	{syndrome, premenstrual}
...	...	...
c-termo	0,077404	{palsy, facial]
...	...	...
s-termo	-0,068059	{tumor}
s-termo	-0,069713	{antibiotics}
s-termo	-0,081312	{cancer}
s-termo	-0,085188	{joint}
s-termo	-0,093884	{defect}
s-termo	-0,094051	{disease}

Tabela 5.1: Termos com os maiores e menores pesos conforme aprendido a partir dos dados de treino da classe "Pathological Conditions, Signs and Symptoms", oriunda da coleção OHSUMED.

SUMED 18302. Conforme pode ser notado pela Tabela 5.1, o termo mais importante para discriminar a classe em questão é o c-termo "{postoperative, pain}". Outros c-termos também aparecem com alto peso no *ranking*. Mais importante, os pesos dos s-termos que compõem o c-termo melhor posicionado no *ranking* são ambos muito baixos (0,009611 para "pain"), ou o s-termo foi completamente ignorado pelo SVM (que foi o caso do s-termo "postoperative", que não ocorreu em nenhum vetor de suporte), uma equivalência de peso nulo. Isso aponta que as a metodologia de extração de características está realmente criando c-termos discriminativos que auxiliam a tarefa de classificação.

5.2 kNN

Com o intuito de avaliar os efeitos da metodologia proposta sobre o kNN, inspecionamos como o processo de classificação foi influencidado por ela.

Com o objetivo de decidir para qual classe c_i um documento de teste d_j deve ser classificado, o kNN analisa se os k documentos de treino mais similares a d_j estão também em c_i. Como resultado, o kNN cria, para cada d_j, um *ranking* de classes. Além disso, tendo em vista que nossas coleções são unirotuladas, o kNN classifica d_j à classe da primeira posição no *ranking*. Conseqüentemente, nós inspecionamos de que forma a metodologia proposta de extração de características influenciou o *ranking* ao analisarmos se ela alterou a posição da classe verdadeira de cada documento de teste. Para tanto, o mesmo conjunto aleatório de documentos de treino e teste foram usados tanto na linha de base como em nossa metodologia a fim de evitar comparações enviesadas. Dessa forma, definimos uma variável aleatória X, a qual representa o número de posições do *ranking* que foram ganhos ou perdidos pela classe verdadeira ao utilizar a metodologia proposta. Para avaliar X, documentos corretamente classificados tanto na linha de base como pelo processo de extração de características não foram considerados. Assim, mais formalmente, temos:

$$X(d_j) = Pos_{e.c.}(d_j) - Pos_{l.b.}(d_j)$$

em que:

$Pos_{l.b.}(d_j)$: posição da classe verdadeira para o documento de teste d_j em uma coleção que emprega apenas s-termos (linha de base).

$Pos_{e.c.}(d_j)$: análogo ao raciocínio anterior, mas com a coleção também usando c-termos.

Dessa maneira, avaliamos a variável aleatória X de todas as coleções de referência testadas. A Figura 5.1 ilustra a distribuição de freqüência de X para a coleção 20 Newsgroups 18828. Como pode ser visto, o valor médio da distribuição é maior que zero, o que significa que nossa estratégia aperfeiçoa a classificação para a maioria dos documentos da coleção. Além disso, também quantificamos a obliqüidade (*skewness*) da distribuição a fim de medir quão assimétrica ela é. Um valor positivo para o valor de obliqüidade significa que a cauda da direita da distribuição é mais pesada que a da

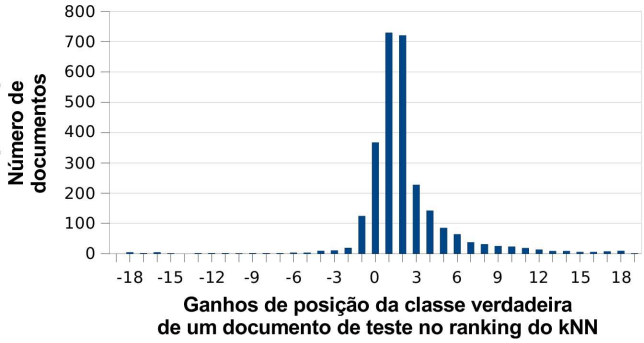

Figura 5.1: Efeito da extração de características sobre o kNN: melhorias no processo de amostragem feito pelo kNN a fim de atribuir documentos de teste à sua classe verdadeira. Coleção: 20 Newsgroups 18828. Média: 2,1; Obliqüidade: 0,3

esquerda. Portanto, para essa tarefa, empregamos o primeiro coeficiente de obliqüidade de Pearson Barry C. Arnold (1995), definido por:

$$Obliqüidade\ de\ Pearson = \frac{3 \times (''média'' - '' moda'')}{\sigma}$$

Assim, tendo em vista que a obliqüidade encontrada para a 20 Newsgroups 18828 foi positiva (+0,3), temos que, esse fato combinado com a média também positiva implica que a extração de características por c-termos é capaz de ajustar a classificação até de documentos cuja classe verdadeira estava longe da primeira posição do *ranking*, e, conseqüentemente, não seriam corretamente classificados se apenas s-termos fossem empregados. Isso pode ser percebido visualmente pela Figura 5.1 quando analisamos a existência de diversos documentos cuja posição da classe verdadeira melhorou expressivamente, mas poucas pioraram.

O efeito discutido anteriormente foi observado em todas as coleções. A Tabela 5.2 demonstra valores encontrados para média e obliqüidade para cada coleção testada.

5.3 Naïve-Bayes

Classificadores probabilísticos são metodologias que computam a probabilidade de um documento d_j pertencer à classe c_i. O algoritmo Naïve-Bayes emprega o Teorema de Bayes Meyer (2000) para medir essa probabilidade, definida como:

Coleções	Ganho médio de posições no *ranking* do kNN	Obliqüidade
20 newsgroups 18828	+2,1	+0,3
ACM 11C	+0,7	+0,4
OHSUMED 18302	+0,6	+0,2
Reuters-21578 8C	+0,5	+0,4

Tabela 5.2: kNN: melhoria média da classe verdadeira no *ranking*.

$$P(c_i|\overrightarrow{d_j}) = P(c_i)\frac{P(\overrightarrow{d_j}|c_i)}{P(\overrightarrow{d_j})}$$

No entanto, a estimativa de $P(\overrightarrow{d_j}|c_i)$ na equação anterior é problemática, uma vez que a quantidade de possibilidades de vetores $\overrightarrow{d_j}$ é exponencial Sebastiani (2002). Como conseqüência, Naïve-Bayes mitiga esse problema ao empregar a premissa de que cada termo é independente. Isso permite computar a probabilidade de cada termo independentemente do restante, tornando a classificação computacionalmente viável.

Como complemento ao Naïve-Bayes, a extração de características identifica relacionamentos discriminativos entre termos, assim quebrando a premissa de independência do algoritmo ao introduzir c-termos discriminativos.

Logo, com o objetivo de mensurar o impacto da metodologia sobre o Naïve-Bayes, medimos como as características inseridas influenciam a probabilidade de um documento de teste para sua classe verdadeira ($P(c_{classe\_verdadeira}|\overrightarrow{d_j})$). Em analogia à análise envolvendo o kNN, os mesmos documentos aleatórios de treino e teste foram empregados para evitar comparações enviesadas. Nós então medimos o número de documentos cuja probabilidade $P(c_{classe\_verdadeira}|\overrightarrow{d_j})$ foi influenciada positiva ou negativamente por, pelo menos, 1 ponto percentual após a inserção de c-termos. Os resultados, apresentados na Tabela 5.3, demonstram a proporção de documentos de teste influenciados pelos c-termos. A referida tabela aponta que houve mais documentos cujas probabilidades para a classe verdadeira foram influenciadas positiva do que

negativamente, para todas as coleções testadas. Portanto, isso resume a maneira em que a extração de características por meio de c-termos afeta o algoritmo Naïve-Bayes: potencializando a probabilidade de um dado documento para sua classe correta.

Coleção	Influência positiva (%)	Influência negativa (%)
20 newsgroups 18828	9,7	5,5
ACM 11C	34,7	15,2
OHSUMED 18302	22,2	6,4
Reuters-21578 8C	7,1	1,8

Tabela 5.3: Naïve-Bayes: influência dos c-termos sobre a probabilidade de um documento para sua classe verdadeira.

Capítulo 6

Conclusão

Podemos distinguir dois fatores principais que tornam a classificação de documentos uma tarefa difícil. O primeiro é o problema de definir um conjunto de características significativas no que diz respeito a melhor distinguir a classe a que cada documento pertence. O segundo é relacionar ao melhor método a ser usado a fim de induzir melhores classificadores de documentos, uma vez que o conjunto de características tenha sido definido.

Neste trabalho, nós não apenas diretamente tratamos o primeiro fator, mas também discutimos como métodos mais tradicionais de classificação podem se beneficiar de nossa estratégia a fim de induzir melhores classificadores.

Especificamente, estivemos preocupados em extrair novas características ao relacionar termos, chamadas de c-termos, que são relevantes para a tarefa de classificação. Os c-termos são conjuntos de termos que co-ocorrem em documentos. No entanto, o tamanho do espaço de busca para c-termos é exponencial. Assim, propusemos uma metodologia viável e eficaz que nos permite extrair c-termos que são pares de s-termos (palavras). A estratégia é composta por três passos. No passo de enumeração, selecionamos os melhores s-termos que serão usados para extrair c-termos. Para essa tarefa, usamos ganho de informação como critério para ordenar os s-termos de forma a empregar apenas os top N s-termos, definido como parâmetro. Na fase de seleção, selecionamos os c-termos extraídos que serão usados para incrementar documentos do

conjunto de treinamento e de teste. Como critério de seleção, usamos c-termos de alta *Predominância* na classe do conjunto de treino. O incremento de documentos de teste é feita ao inserir todos os c-termos de alta *Predominância* cujos s-termos ocorrem no documento. Em ambos os casos, o limiar de *Predominância* é uma função das características da coleção. O compromisso de apenas utilizar c-termos discriminativos na classificação é, além de enxuto, uma necessidade que nos permite melhorar a classificação ao mesmo tempo em que mantemos o processo de extração viável.

Experimentos, conduzidos ao longo de diversas coleções (Reuters-21578 8C, 20 Newsgroups 18828, OSHUMED 18302 e ACM 11C) empregaram diferentes métodos de classificação e demonstraram que a estratégia apresentada neste trabalho consistentemente aperfeiçoa a qualidade da classificação de texto. Os ganhos mais significativos foram obtidos usando-se o algoritmo de classificação kNN, no qual foram atingidos ganhos de até 46% sobre a micro-média F_1 para a coleção 20 Newsgroup 18828. Melhorias expressivas também foram reportadas para o método de classificação considerado estado da arte, SVM, em que ganhos de 10,7% sobre a macro-média F_1 foram reportados na coleção OHSUMED 18302. Os experimentos mostram que nossa metodologia de extração de características não é limitada pelo método de classificação em uso, tampouco pela coleção. Isso sem contar que os maiores ganhos foram verificados nas coleções mais difíceis de se classificar.

Como trabalho futuro, pretendemos investigar outras técnicas de mineração de dados para explorar relacionamentos entre termos. Como exemplo, pretendemos investigar a eficácia de outras técnicas de seleção de características, além da *Predominância* e ganho de informação, para efeito de selecionar os melhores c-termos gerados. Outro tópico de pesquisa interessante refere-se à melhor pesagem a se utilizar aos c-termos extraídos. Embora esse problema tenha sido bastante estudado para os s-termos, em que estratégias como *tf/idf* são comumente utilizadas, o domínio dos c-termos, por serem gerados a partir de co-ocorrências, possuem peculiaridades que problematizam a aplicação direta de estratégias de pesagem usadas para s-termos, como a ausência

de definição do *tf* de um c-termo, por exemplo. Por fim, pretendemos investigar como generalizar mais ainda nossa estratégia de modo que possa ser muito bem aplicada a coleções cujos documentos estão associados a mais de uma classe.

Apêndice A

Gráficos: Efeito do Suporte

O fator Suporte Mínimo, (Min_Supp) é um critério para delimitar a quantidade mínima de documentos em que um c-termo deve ocorrer a fim de ser selecionado para o modelo de classificação. Como conseqüência, valores altos para Min_Supp filtram c-termos de baixo suporte, ao passo que valores baixos permitem que c-termos infreqüentes também sejam considerados no modelo de classificação.

As figuras a seguir apresentam os efeitos do fator Min_Supp sobre o algoritmo de classificação SVM aplicado sobre as seguintes coleções: 20 Newsgroups 18828 (Figura A.1), Reuters-21578 8C (Figura A.2) e ACM11C (Figura A.3).

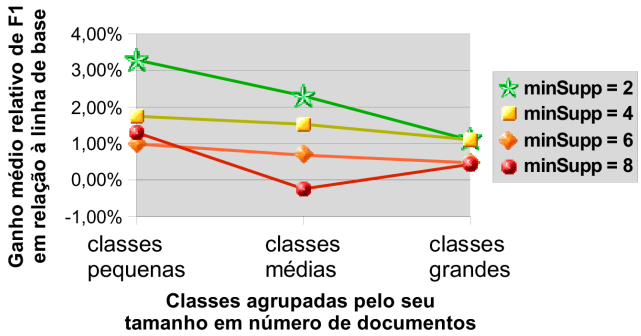

Figura A.1: SVM: ganhos agrupados conforme o tamanho da classe para cada limiar mínimo de suporte ($MinSupp$). Coleção: 20 Newsgroups 18828

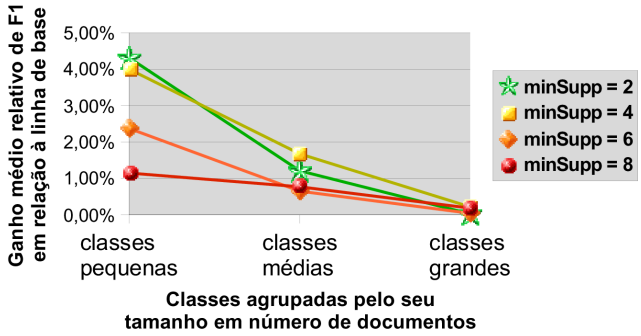

Figura A.2: SVM: ganhos agrupados conforme o tamanho da classe para cada limiar mínimo de suporte (*MinSupp*). Coleção: Reuters-21578 8C

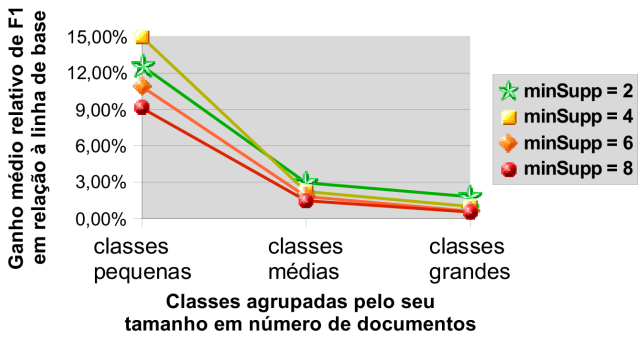

Figura A.3: SVM: ganhos agrupados conforme o tamanho da classe para cada limiar mínimo de suporte (*MinSupp*). Coleção: ACM11C

Apêndice B

Tabelas: Dispersão dos Resultados

Abaixo apresentamos, a título de complementariedade, os desvios-padrão relativos aos resultados experimentais da classificação utilizando-se s-termos e c-termos, apresentados nas Tabelas 4.2 e 4.3, capítulo 4, página 35.

Desvios-padrão da Macro-Média F_1	kNN	Naïve-Bayes	SVM
COLEÇÕES	Desvio P.	Desvio P.	Desvio P.
20 newsgroups	0,7	0,3	0,2
ACM 11C	0,5	0,5	0,4
OHSUMED	0,4	0,3	1,0
Reuters-21578	1,1	0,6	0,3

Tabela B.1: Desvios-padrão da macro-média F_1.

Desvios-padrão da Micro-Média F_1	kNN	Naïve-Bayes	SVM
COLEÇÕES	Desvio P.	Desvio P.	Desvio P.
20 newsgroups	1,1	0,3	0,2
ACM 11C	0,5	0,4	0,3
OHSUMED	0,2	0,5	0,7
Reuters-21578	0,5	0,2	0,2

Tabela B.2: Desvios-padrão da micro-média F_1.

Referências Bibliográficas

Amati, G.; Aloisi, D. D.; Giannini, V. e Ubaldini, F. (1997). A framework for filtering news and managing distributed data. *J.UCS: Journal of Universal Computer Science*, 3(8):1007–1021.

Apté, C.; Damerau, F. e Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst.*, 12(3):233–251.

Barry C. Arnold, R. A. G. (1995). Measuring skewness with respect to the mode. In *The American Statistician*, volume 49, pp. 34–38.

Bekkerman, R. e Allan, J. (2004). Using bigrams in text categorization. Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst.

Caropreso, M. F.; Matwin, S. e Sebastiani, F. (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. pp. 78–102.

Chandrinos, K.; Androutsopoulos, I.; Paliouras, G. e Spyropoulos, C. D. (2000). Automatic web rating: Filtering obscene content on the web. In *ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 403–406, London, UK. Springer-Verlag.

Cheng, H.; Yan, X.; Han, J. e Hsu, C.-W. (2007). Discriminative frequent pattern analysis for effective classification. In *ICDE'07: Proceedings of 2007 International Conference on Data Engineering*, pp. 716–725, Istanbul, Turkey.

Chomsky, N. (1957). *Syntactic structures*. The Houge, Mouton, The Netherlands.

Choueka, Y. e Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities*, 19(3):147–157.

Couto, T.; Cristo, M.; Gonçalves, M. A.; Calado, P.; Ziviani, N.; Moura, E. e Ribeiro-Neto, B. (2006). A comparative study of citations and links in document classification. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pp. 75–84, New York, NY, USA. ACM.

Crawford, E.; Koprinska, I. e Patrick, J. (2004). Phrases and feature selection in e-mail classification. In Bruza, P.; Moffat, A. e Turpin, A., editores, *ADCS*, pp. 59–62. University of Melbourne, Department of Computer Science.

Cullen, J. e Bryman, A. (1988). The knowledge acquisition bottleneck: Time for reassessment? 5:216–225.

Dumais, S.; Platt, J.; Heckerman, D. e Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pp. 148–155, New York, NY, USA. ACM.

Feng, J.; Liu, H. e Zou, J. (2005). Sat-mod: moderate itemset fittest for text classification. In *WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, pp. 1054–1055, New York, NY, USA.

Freund, Y. e Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In Vitanyi, P., editor, *Computational Learning Theory: Second European Conference(EuroCOLT'95)*, pp. 23–37. Springer, Berlin.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. Technical report, OEFAI-TR-9830, Austrian Institute for Artificial Intelligence.

Hayes, P. J.; Andersen, P. M.; Nirenburg, I. B. e Schmandt, L. M. (1990). Tcs: a shell for content-based text categorization. In *Proceedings of the sixth conference on Artificial intelligence applications*, pp. 320–326, Piscataway, NJ, USA. IEEE Press.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis- Thechniques for Experimental Design, Measurement, Simulation and Modeling*. John Wiley & Sons, Inc., New York, NY, USA.

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pp. 137–142, London, UK. Springer-Verlag.

Joachims, T. (1999). Making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*, pp. 169–184. MIT Press, Cambridge, MA, USA.

Joachims, T. (2006). Training linear svms in linear time. In *KDD '06: Proceedings of the 12th ACM SIGKDD: International Conference on Knowledge Discovery and Data Mining*, pp. 217–226, Philadelphia, PA, USA. ACM Press.

Kaplan, A. (1955). An experimental study of ambiguity and context. *Mechanical Translation, 2, 39-46.*, 2(2):39–46.

Lewis, D. D. (1992a). An evaluation of phrasal and clustered representations on a text categorization task. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 37–50, New York, NY, USA. ACM.

Lewis, D. D. (1992b). Feature selection and feature extraction for text categorization. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pp. 212–217, Morristown, NJ, USA. Association for Computational Linguistics.

Lewis, D. D. (1992c). *Representation and learning in information retrieval*. PhD thesis, Amherst, MA, USA.

Lewis, D. D. (1995). Evaluating and optimizing autonomous text classification systems. In *SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pp. 246–254, Seattle, Washington, USA.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pp. 4–15, London, UK. Springer-Verlag.

McCallum, A. e Nigam, K. (1998). A comparison of event models for naive bayes text classification.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. Software available at http://www.cs.cmu.edu/~mccallum/bow/.

Meyer, P. (2000). *Probabilidade - Aplicações a Estatística*. LTC.

Mitchell, T. (1997a). *Machine Learning*. McGraw-Hill, New York, NY, USA.

Mitchell, T. M. (1997b). Does machine learning really work? *AI Magazine*, 18(3):11–20.

Mladenić, D. e Grobelnik, M. (1998). Word sequences as features in text-learning. In *Proceedings of ERK-98, the Seventh Electrotechnical and Computer Science Conference*, pp. 145–148, Ljubljana, SL.

Mohamed Hammami, Dzmitry V. Tsishkou, L. C. (2004). Adult content web filtering and face detection using data-mining based kin-color model. 1:403–406.

Pôssas, B.; Ziviani, N.; Ribeiro-Neto, B. e Meira Jr., W. (2005). The set-based model for information retrieval. *ACM Transactions on Information Systems*, 23(4):397–429.

Rak, R.; Stach, W.; Zaïane, O. R. e Antonie, M. L. (2005). Considering re-occurring features in associative classifiers. In *PAKDD'05: Nineth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 240–248, Hanoi, Vietnam.

Rongbo Du, Reihaneh Safavi-Naini, W. S. (2003). Web filtering using text classification. pp. 325– 330.

Salton, G. e McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Scott, S. e Matwin, S. (1999). Feature engineering for text classification. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 379–388, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Sculley, D. e Wachman, G. M. (2007). Relaxed online svms for spam filtering. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 415–422, New York, NY, USA. ACM.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Tan, C.-M.; Wang, Y.-F. e Lee, C.-D. (2002). The use of bigrams to enhance text categorization. *Inf. Process. Manage.*, 38(4):529–546.

Vapnik, V. (1998). *Statistical Learning Theory*. Jonh Wiley & Sons Inc., New York, NY, USA.

Wielinga, B.; Boose, J.; Gaines, B.; Shereiber, G. e van Someren, M. (1990). *Comparison of Inductive and Naive Bayesian Learning Approaches to Automatic Knowledge Acquisition*. IOS Press, Amsterdam.

Yang, Y. e Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zaïane, O. R. e Antonie, M. L. (2002). Classifying text documents by associating terms with text categories. In *CRPITS '02: Proceedings of the Thirteenth Australasian Conference on Database Technologies*, pp. 215–222, Darlinghurst, Australia. Australian Computer Society, Inc.

Zhang, B.; Chen, Y.; Fan, W.; Fox, E. A.; Gonçalves, M. A.; Cristo, M. e Calado, P. (2005). Intelligent fusion of structural and citation-based evidence for text classification. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 667–668, New York, NY, USA. ACM.

Zhang, L.; Zhu, J. e Yao, T. (2004). An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(4):243–269.