Universidade Federal de Goiás
Instituto de Informática

Leonel Diógenes Carvalhaes Alvarenga

Uso de Seleção de Características da Wikipédia na Classificação Automática de Textos

Goiânia
2012

Leonel Diógenes Carvalhaes Alvarenga

Uso de Seleção de Características da Wikipédia na Classificação Automática de Textos

Dissertação apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Computação.

Área de concentração: Recuperação de Informação.

Orientador: Prof. Dr. Thierson Couto Rosa

Goiânia
2012

Leonel Diógenes Carvalhaes Alvarenga

Graduou-se em Ciência da Computação na Universidade de Rio Verde (FESURV). Atua como Professor no campus Rio Verde do Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Dedico este trabalho aos meus Pais Walter Alvarenga dos Santos e Lúcia Carvalhaes Alvarenga que sempre me incentivaram na conquista de meus ideais, sempre pautados pelo trabalho e pela ética. Dedico também à minha esposa Lídia Nunes de Ávila Carvalhaes, que sempre está ao meu lado, tanto nos momentos alegres quanto nas horas difíceis.

Agradecimentos

Primeiramente agradeço a Deus por sempre me dar forças para seguir perseverante e vencendo os obstáculos com determinação.

Agradeço a minha esposa Lídia Nunes de Ávila Carvalhaes por sempre me motivar e compreender os momentos de ausência.

Agradeço a todos os colegas do mestrado que me auxiliaram, direta ou indiretamente.

Agradeço ao meu Orientador, professor Dr. Thierson Couto Rosa pela motivação e pela dedicação a este trabalho estando sempre pronto a me auxiliar em todas as situações.

Agradeço à Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) da qual obtive apoio na forma de bolsa de formação, sem a qual seria muito difícil custear os gastos com o deslocamento de Rio Verde à Goiânia, alimentação e hospedagem durante o decorrer do Mestrado.

Agradeço ao Instituto Federal Goiano por sempre me apoiar e incentivar na busca de qualificação.

Finalmente, agradeço a minha família e amigos pelo auxílio nessa jornada.

"A ciência que não se transforma em conhecimento apoiador do desenvolvimento da sociedade é como uma lâmpanda acesa em uma gaveta fechada."

Leonel Diógenes Carvalhaes Alvarenga,
Em reflexão sobre o papel dos trabalhos científicos.

Resumo

Alvarenga, Leonel Diógenes Carvalhaes. **Uso de Seleção de Características da Wikipédia na Classificação Automática de Textos**. Goiânia, 2012. 114p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

Os métodos tradicionais de classificação de textos normalmente representam documentos apenas como um conjunto de palavras, também conhecido como BOW (do inglês, Bag of Words). Vários estudos têm mostrado bons resultados ao utilizar-se de tesauros e enciclopédias como fontes externas de informações, objetivando expandir a representação BOW a partir da identificação de relacionamentos de sinonímia e hiponímia entre os termos presentes em uma coleção de documentos. Todavia, o processo de expansão pode introduzir termos que conduzam a uma classificação errônea do documento. No presente trabalho, propõe-se a aplicação de medidas de avaliação de termos para a seleção de características extraídas da Wikipédia, com o objetivo de melhorar a eficácia de sua utilização durante o processo de expansão de documentos. O estudo também propõe uma medida de seleção de características denominada Fator de Tendência a uma Categoria (FT1C), de modo que os experimentos realizados demonstraram que esta medida apresenta desempenho competitivo com as medidas *Information Gain*, *Gain Ratio* e *Chi-squared*, neste processo, apresentando os melhores ganhos de $microF_1$ e $macroF_1$, na maioria dos experimentos realizados. O uso integral das características selecionadas neste processo, demonstrou auxiliar a classificação de forma mais estável, ao passo que apresentou menor desempenho ao se restringir sua inserção somente aos documentos das classes em que estas características são bem pontuadas pelas medidas de seleção. Ao ser aplicada nas coleções Reuters-21578, Ohsumed *first*-20000 e 20Newsgroups, a abordagem com seleção de características permitiu a redução da inserção de ruídos inerentes do processo de expansão e potencializou o uso de hipônimos, assim como demonstrou que as relações de sinonímia da Wikipédia também podem ser utilizadas na expansão de documentos, elevando a eficácia da classificação automática de textos.

Palavras–chave

Recuperação de informação, classificação de textos, seleção de características, expansão de documentos, aprendizado de máquina.

Abstract

Alvarenga, Leonel Diógenes Carvalhaes. **Selection of Wikipedia features for automatic text classification**. Goiânia, 2012. 114p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

The traditional methods of text classification typically represent documents only as a set of words, also known as "Bag of Words"(BOW). Several studies have shown good results on making use of thesauri and encyclopedias as external information sources, aiming to expand the BOW representation by the identification of synonymy and hyponymy relationships between present terms in a document collection. However, the expansion process may introduce terms that lead to an erroneous classification. In this paper, we propose the use of feature selection measures in order to select features extracted from Wikipedia in order to improve the effectiveness of the expansion process. The study also proposes a feature selection measure called Tendency Factor to One Category (TF1C), so that the experiments showed that this measure proves to be competitive with the other measures *Information Gain*, *Gain Ratio* and *Chi-squared*, in the process, delivering the best gains in $microF_1$ and $macroF_1$, in most experiments. The full use of features selected in this process showed to be more stable in assisting the classification, while it showed lower performance on restricting its insertion only to documents of the classes in which these features are well punctuated by the selection measures. When applied in the Reuters-21578, Ohsumed *first - 20000* and 20Newsgroups collections, our approach to feature selection allowed the reduction of noise insertion inherent in the expansion process, and improved the results of use hyponyms, and demonstrated that the synonym relationship from Wikipedia can also be used in the document expansion, increasing the effectiveness of the automatic text classification.

Keywords

Information retrieval, text classification, feature selection, document expansion, machine learning.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

1.1 Contextualização

Desde o surgimento das civilizações a organização da informação textual tem sido uma preocupação dos seres humanos [3]. Com o passar do tempo a quantidade de documentos textuais aumentou de forma considerável. A atividade de organizar documentos textuais em categorias ou classes é usualmente denominada *classificação de textos* ou ainda, *categorização de textos*.

Até antes do surgimento dos computadores digitais a classificação de textos era uma tarefa essencialmente humana. Um exemplo desta atividade está na classificação de livros em uma biblioteca. Os bibliotecários geralmente utilizam uma classificação hierárquica, em que livros de uma mesma área do conhecimento humano são colocados em um mesmo conjunto de estantes. Nas estantes, os livros são agrupados por subáreas e dentro de uma subárea, os livros são geralmente agrupados por ordem alfabética por nome de autor.

Com o surgimento do computador digital, vários documentos foram digitalizados e outros criados diretamente em formato digital. De acordo com [52], ainda na década de 60 surgiram os primeiros esforços no sentido de automatizar a classificação de documentos digitais.

Entretanto, até a década de 80 vigorava uma técnica semiautomática que se baseava em *engenharia do conhecimento* para classificação de documentos. Esta técnica consiste em definir manualmente um conjunto de regras que codificam o conhecimento de um especialista sobre como classificar documentos em um conjunto de classes. As regras tinham a seguinte forma:

se (*fórmula FND*) **então** *classes*

A fórmula FND (fórma normal disjuntiva) é um conjunto de operações conjuntivas de cláusulas disjuntivas e serve para combinar características de um documentos que são úteis para determinar se este documento pertence à classe indicada após o termo " então" na regra. Esta abordagem é semiautomática, porque

uma vez que o conjunto de regras tenha sido gerado manualmente, ele pode ser codificado como um programa de computador, utilizando-se uma linguagem de programação. Este programa recebe como entrada documentos e aplica regras que possuam as características presentes em um documento de entrada e utiliza as mesmas para decidir a qual classe o documento pertence. Uma possibilidade é utilizar a classe que aparece na maioria das regras aplicáveis ao documento.

A abordagem semiautomática de classificação de documentos tem uma grande limitação em relação à aquisição de conhecimento para a construção do classificador. Esta limitação está principalmente na necessidade de haver dois especialistas humanos necessários no processo: um especialista em classificar documentos no conjunto de classes pré-definido e um *engenheiro do conhecimento*, capaz de codificar em uma linguagem de programação o conhecimento do especialista representado como um conjunto de regras na forma FND. Claramente, esta abordagem não é flexível, pois se houver alterações nas classes ou se surgir um novo conjunto de classes, os dois profissionais devem ser chamados para que outro programa classificador seja gerado.

No início da década de 90 uma abordagem completamente automática começou a ser utilizada para classificar documentos, pelo menos, em nível de pesquisa acadêmica. A classificação automática de texto (CAT) utiliza técnicas de *aprendizagem por máquina* (do inglês, *machine learning*) para gerar um classificador automático. A aprendizagem por máquina dispensa o especialista e o engenheiro de conhecimento na geração do classificador, mas requer um conjunto de documentos classificados manualmente, denominado *conjunto de treino*. Técnicas que geralmente se baseiam em estatísticas e probabilidades são utilizadas para gerar automaticamente um classificador que é capaz de relacionar um conjunto de características presentes nos vários documentos do conjunto de treino com as classes desses documentos.

As técnicas de CAT por si mesmas despertavam o interesse de pesquisadores, que tinham objetivo de desenvolver heurísticas mais eficazes para a CAT. Contudo, ainda na década de 90 um evento impulsionou ainda mais o interesse em pesquisas em CAT. Este evento corresponde ao surgimento da *World Wide Web* (Web). A Web não somente corresponde a uma gigantesca e dinâmica coleção não classificada de textos (hipertextos) e de objetos de outras mídias, como também influenciou o surgimento de aplicações e problemas que por si mesmos requerem o uso da CAT.

Mais recentemente, a ampliação das possibilidades de comunicação via Web (por meio de sítios de redes sociais como Facebook, tweeter, etc) e a popularização do correio eletrônico aumentaram ainda mais o número de aplicações de CAT na Web. Entre estas aplicações pode-se citar: filtragem de mensagens *spam* [51, 71], detecção de conteúdo impróprio para menores de idade [8, 13, 21], classificação automática de

documentos em bibliotecas digitais [1, 11] e aprendizagem automática de ordenação de documentos em máquinas de busca [34].

De acordo com Sebastiani [52], dois aspectos são observados, visando a avaliar o desempenho de métodos de CAT : a eficácia e a eficiência da classificação automática. A eficácia corresponde à habilidade de um classificador automático decidir corretamente a classe de determinado documento. A eficiência, por sua vez, corresponde ao tempo gasto na classificação automática e pode ser avaliada de dois modos distintos. Um deles corresponde ao tempo gasto pelo método de aprendizagem para gerar um classificador com base nas características dos documentos de treino (eficiência do treinamento). O outro modo corresponde ao tempo gasto pelo classificador gerado para determinar a classe de um documento não pertencente ao conjunto de treino (eficiência do classificador).

Para se avaliar os aspectos de eficiência e eficácia, é necessário que se tenha um conjunto de documentos não pertencentes ao conjunto de treino. Esse conjunto é denominado *conjunto de teste*. A classe dos documentos do conjunto de teste é conhecida pela pessoa que avalia a classificação, mas não pode ser utilizada como entrada ao processo de aprendizagem por máquina que gera do classificador e também não pode ser utilizada como informação de entrada ao classificador, para se garantir uma avaliação correta da classificação

Quanto às técnicas de aprendizado por máquina para geração de classificadores, um vasto número de algoritmos tem sido propostos ao longo das duas últimas décadas. Dentre eles pode-se citar o *naive bayes* [36], k-vizinhos mais próximos (do inglês *k-nearest neighbor*) [60], máquinas de vetor de suporte (do inglês *support vector machines*) [25], algoritmo de aceleração (do inglês, *boosting*) [48] e algoritmos de aprendizado de regras (do inglês *rule learning algorithms*) [55].

No entanto, apesar de todos os esforços no aperfeiçoamento de algoritmos para que estes consigam gerar classificadores eficazes, observa-se que a eficácia dos classificadores também é fortemente dependente da forma como os documentos são representados [54] [58] [19], ou seja, da qualidade dos componentes textuais utilizados como informações no processo de treino para a classificação. Estes elementos são denominados características dos documentos. As características mais comuns da CAT são os termos dos documentos. Dessa forma, quanto mais se utiliza características relevantes para classificação na representação do documento, maiores serão as chances de se ter um aumento na eficácia do método.

1.2 Representação de Documentos

Os métodos tradicionais de CAT normalmente são baseados na representação de documentos utilizando a abordagem de conjunto de palavras (BOW, do inglês *Bag of Words*) [22] [35] de forma que a classificação é baseada na presença ou na ausência de termos-chave na matriz documento-termos que representa cada documento, como exposto por Sebastiani [52]. O motivo disso é a simplicidade, eficiência e relativa eficácia do paradigma BOW.

No entanto, no método BOW, importantes relações semânticas entre os termos chave são descartadas [22]. Outro aspecto refere-se ao fato de que ao se visualizar um documento representado pelo método descrito, a ordem exata dos termos é ignorada [18]. Como resultado do que foi exposto, se dois documentos utilizam diferentes conjuntos de palavras chave para descrever sobre um mesmo tópico os dois podem ser classificados como sendo de categorias diferentes, entretanto as palavras chave utilizadas por ambos são provavelmente sinônimas ou semanticamente associadas de alguma outra forma [23]. Consequentemente, tais observações impulsionaram as pesquisas em CAT de modo a buscar melhores representações para documentos de texto que conseguissem captar tais relacionamentos e que contribuíssem com a eficácia do processo de classificação.

Entre representações alternativas podemos citar aquelas que utilizam características do próprio texto, distintas dos termos, por exemplo: coocorrência sequencial de n termos (n-gramas) e coocorrências não sequenciais de n termos (conjuntos de termos – termsets) . Outros trabalhos têm explorado recursos externos ao texto. Como exemplo, pode-se citar o crescente interesse em técnicas que trabalham com a geração de características (FG, do inglês *Feature Generation* ou *Feature Construction*) também conhecida por expansão de documentos (do inglês *Document Expansion*) ou enriquecimento de documentos (do inglês *Document Enrichment*), por meio do qual adiciona-se novos termos aos documentos, melhorando a representação BOW através da inserção de características mais informativas na matriz documento-termos deste [18].

Diversos métodos que utilizam FG têm conseguido bons resultados em CAT por meio da extração de relações semânticas de sinonímia, polissemia, hiponímia [1] e relações associativas entre conceitos, presentes em Enciclopédias, Tesauros [2], Páginas Web, dentre outros [18, 19, 22, 63, 64].

[1] Relação semântica em que uma palavra está num plano hierárquico inferior a outra, uma vez que a outra corresponde a uma categoria ou espécie que a inclui ao nível do significado. Ex: sardinha, salmão, carapau são hiponímias de peixe.

[2] Dicionário que registra uma lista de palavras que são associadas semanticamente a outras, apresentando geralmente sinônimos e, algumas vezes, antônimos.

1.3 Problemas de Pesquisa e Objetivos

A Wikipédia é uma enciclopédia em formato digital composta por conceitos e que faz uso extensivo de metadados para representar relacionamentos entre tais conceitos. Como exemplos de metadados da Wikipédia pode-se citar: ligações de redirecionamento (ou sinonímia) entre conceitos, ligações de categoria entre conceitos (hiponímia). Em trabalhos recentes, o uso de metadados da Wikipédia como fonte adicional de características, tem gerado uma melhor qualidade na representação BOW de documentos de coleções, conseguindo-se melhorar tarefas de agrupamento (do inglês, *clustering*) e de classificação de documentos [7, 50, 67].

Entretanto, [64] reporta que a expansão de documentos utilizando relações de sinonímia de conceitos da Wikipédia gera grande quantidade de ruídos[3], o que degrada a qualidade da classificação quando comparada com a não expansão de características. Por outro lado, o uso de relações de hiponímia provindas das categorias dos conceitos da Wikipédia tem se mostrado útil para a CAT [19, 33, 40, 62] , inclusive na expansão da representação de documentos [64]. Nos dois casos apresentados os autores não utilizam nenhum método de seleção de característica (do inglês, *feature selection*), a fim de selecionar somente sinônimos e categorias que sejam mais relevantes ao processo de classificação automática de textos.

O objetivo geral desse trabalho é o de melhorar o processo de utilização de sinônimos e categorias extraídos da Wikipédia para o uso na expansão de características de documentos a serem classificados pelo processo de CAT. Com o intuito de alcançar o objetivo exposto, abordamos 3 problemas de pesquisa, descritos a seguir.

Problema de Pesquisa 1 *A aplicação de um método de seleção de características consegue melhorar a eficácia da utilização das relações de sinonímia e de categorias provindas da Wikipédia durante o processo de expansão de documentos, reduzindo a inserção de ruídos e potencializando a adição de características boas discriminadoras de classes?*

Em relação à questão acima, as hipóteses são que a utilização de métodos de seleção das características pode reduzir a inserção de ruídos provenientes do processo de expansão de documentos com características provindas de conceitos sinônimos extraídos da Wikipédia, tornando-os úteis ao processo de classificação textual. Também acredita-se que a seleção de características consiga melhorar os resultados obtidos com a expansão de documentos por categorias diretas dos

[3]Neste contexto, os ruídos são características que atrapalham a classificação correta de um documento.

conceitos provindos da mesma enciclopédia. Para confirmar ou refutar as hipóteses, foi definido o seguinte objetivo decorrente do Problema de Pesquisa 1:

- Melhorar a eficácia da CAT utilizando expansão de documentos em conjunto com medidas de seleção de características, visando a enriquecer a representação BOW de documentos por meio dos conceitos sinônimos e das categorias dos conceitos sinônimos, ambos provenientes da Wikipédia.

O objetivo decorrente do Problema de Pesquisa 1 possui os seguintes objetivos específicos:

- Avaliar a eficácia da utilização de seleção de características durante a expansão de documentos utilizando apenas conceitos sinônimos da Wikipédia e que coocorrem nos documentos a serem classificados.
- Avaliar a eficácia da utilização de seleção de características durante a expansão de documentos utilizando apenas categorias dos conceitos sinônimos da Wikipédia e que coocorrem nos documentos a serem classificados.
- Avaliar a eficácia da utilização de seleção de características durante a expansão de documentos utilizando conceitos sinônimos juntamente com as categorias dos conceitos sinônimos da Wikipédia e que coocorrem nos documentos a serem classificados.

Problema de Pesquisa 2 *A utilização de uma medida de avaliação de termos que pontue positivamente a abundância de uma característica na classe a qual pertence o documento de treino a ser expandido e utilize como penalização a abundância relativa desta mesma característica nas outras classes da coleção, pode se mostrar como opção competitiva na seleção de características provindas da Wikipédia na forma de conceitos sinônimos e categorias?*

Neste trabalho é proposta uma medida de avaliação de termos para seleção de características provindas da Wikipédia denominada Fator de Tendência a uma Categoria (FT1C). Por meio desta medida, quanto maior for a abundância de uma característica t_i em uma categoria c_j pertencente ao conjunto de categorias C, e menor o valor da abundância de t_i nas demais categorias de C, maior será o fator FT1C de t_i em c_j. Por outro lado, quanto menor a abundância de t_i em c_j, e quanto maior a abundância de t_i nas demais categorias de C, menor será o fator de tendência a uma classe FT1C de t_i em c_j. A medida FT1C será vista com detalhes na Seção 3.2

Em relação à questão apresentada pelo Problema de Pesquisa 2, a hipótese é que a utilização da medida de avaliação FT1C possibilite avaliar bem as características boas discriminadoras de classes, assim como imprimir um menor valor para

características pouco relevantes para o processo de classificação e gerando boa estabilidade ao processo de enriquecimento com características provindas da Wikipédia. Espera-se que o método se adapte bem às grandes variações na distribuição de documentos pelas diversas classes da coleção. Para confirmar ou refutar a hipótese foi definido o seguinte objetivo decorrente do Problema de Pesquisa 2:

- Comparar a eficácia da medida de avaliação de termos FT1C durante o processo de seleção de características provindas da Wikipédia, confrontando-a com outras já consolidadas na literatura.

O objetivo decorrente do Problema de Pesquisa 2 possui os seguintes objetivos específicos:

- Avaliar os melhores ganhos obtidos por meio da expansão de documentos com características provindas da Wikipédia as quais foram selecionadas por meio da medida FT1C, comparando-a com os ganhos obtidos com a utilização das medidas de seleção de características *Information Gain*, *Gain Ration* e *Chi-squared*.
- Comparar a estabilidade da medida de seleção FT1C, com as medidas *Information Gain*, *Gain Ration* e *Chi-squared*, quando utilizadas no processo de seleção de características provindas da Wikipédia.

Além de analisar a aplicação de métodos de seleção de características como exposto pelo Problema de Pesquisa 1, e comparar a medida proposta para avaliação de característica com outras medidas já consagradas na literatura como explanado pelo Problema de Pesquisa 2, avalia-se também o método de utilização das características eleitas pelo processo de seleção de características.

Ao aplicar um método de seleção de característica, pode-se utilizar uma filtragem adicional, para a qual uma característica só será utilizada na expansão de um documento d_k se tal característica obtiver um valor mínimo na medida de seleção de característica para classe c_j e d_k esteja entre o conjunto de documentos pertencentes a c_j, como em [14]. O presente trabalho utiliza a referida abordagem de restrição de classe (CRC) e levanta o Problema de Pesquisa 3:

Problema de Pesquisa 3 *A utilização de um método o qual permita a expansão de documentos somente com características bem avaliadas na classe do documento de treino a ser expandido, poderia aumentar a eficácia da classificação de documentos enriquecidos com características provindas da Wikipédia?*

Com relação a questão acima, a hipótese é que assim como no trabalho de [14], a restrição apresentada pelo Problema de Pesquisa 3 consiga melhorar

a qualidade das características provindas da Wikipédia e utilizadas na expansão de documentos. Com o uso desta restrição (CRC), espera-se que características provindas da referida enciclopédia e que são boas discriminadoras de uma classe c_k não sejam utilizadas para enriquecer um documento pertencente a uma classe c_j, melhorando a eficácia da CAT. Com o intuito de confirmar ou refutar a supracitada hipótese, foi definido o seguinte objetivo decorrente do Problema de Pesquisa 3:

- Avaliar o desempenho da utilização da metodologia de expansão com restrição de classe CRC após a aplicação de um método de seleção de características, confrontando seus resultados com os obtidos sem o uso desta restrição de classe, referenciada como abordagem SRC.

Especificamente, objetiva-se avaliar a eficácia da CAT ao se utilizar as medidas de avaliação de termos FT1C, *Information Gain*, *Gain Ratio*, *Chi-squared*, em conjunto com as metodologias CRC e SRC, podendo de forma empírica constatar o desempenho de cada abordagem nos diversos ambientes experimentais.

Com o intuito de avaliar a qualidade das abordagens propostas, o algoritmo de classificação SVM (do inglês, *Support Vector Machine*) [61] foi aplicado nas coleções Reuters, Ohsumed e 20Newsgroups. Dessa forma, foram coletados os resultados antes e após a geração de características nestas coleções de dados. Escolheu-se o SVM por ser um das técnicas de aprendizado de máquina que gera classificadores mais eficazes para CAT [16, 26, 64, 14].

1.4 Principais contribuições do Trabalho

O presente trabalho tem as seguintes contribuições para a CAT:

- Comprovação da eficácia da utilização de métodos de seleção de características para melhorar a eficácia do uso das relações de sinonímia e de categorias provindas da Wikipédia durante o processo de expansão de documentos para melhoria da CAT.
- Demonstração experimental de que os conceitos sinônimos extraídos da Wikipédia são de grande importância para a melhoria da CAT, desde que seja aplicado um método que selecione as características mais significativas, contrariando a afirmativa de [64] sobre a má qualidade destes elementos para melhorar o processo de classificação automática de documentos.
- Demonstração experimental de que o uso das categorias extraídas da Wikipédia geram melhor eficácia à CAT quando aplicado um método que selecione as características mais significativas.

- Proposição e análise de uma medida de avaliação de características denominada FT1C para a seleção de características da Wikipédia. A FT1C se posicionou como opção competitiva para o processo de seleção de características provindas da Wikipédia particularmente quando comparada com às medidas *Information Gain*, *Gain Ratio*, e *Chi-squared*. Comprovou-se experimentalmente a melhor estabilidade da medida de avaliação de características FT1C frente às demais medidas avaliadas, quando aplicada às versões uni-rótulo das coleções Reuters-21578, Ohsumed *first*-20000 e 20Newsgroups - *All 20000 documents*.
- Realização de um estudo comparativo e conclusivo sobre a eficácia e estabilidade da expansão de documentos utilizando a abordagem sem restrição de classe (SRC), quando comparado à abordagem com restrição de classe (CRC). A utilização de nossas abordagens, proporcionou bons resultados na melhoria da CAT, principalmente na coleção Ohsumed *first*-20000, para a qual obteve-se um ganho de $7,67\%$ na medida $microF_1$ e $15,08\%$ na medida $macroF_1$. Dado que esta coleção é reconhecida por sua dificuldade de classificação, tais ganhos se mostram ainda mais significativos. Todos os experimentos foram realizados sobre o classificador SVM, o qual na maioria dos casos eleva a linha-base utilizada, provinda da representação BOW. Dessa forma, mesmo os menores ganhos obtidos podem ser considerados como importantes.

1.5 Organização do Trabalho

O Capítulo 2 apresenta uma revisão bibliográfica dos conceitos e técnicas diversas utilizados no decorrer deste trabalho, assim como são apresentados os trabalhos que possuem correlação direta com o tema abordado. No Capítulo 3 são apresentados a abordagem proposta pelo corrente trabalho para a extração de características da enciclopédia Wikipédia e seu uso na CAT. É apresentada ainda a medida proposta para seleção de características (FT1C), e a metodologia de expansão de características sem restrição de classe (SRC) e com restrição de classe (CRC). Ademais, também são discutidos alguns detalhes pertinentes à implementação dessas abordagens. O Capítulo 4 apresenta os resultados alcançados com a utilização da medida de seleção de característica FT1C, confrontando seus resultados com outras três medidas consolidadas na literatura. Também compara-se eficácia das metodologias de expansão de características CRC e SRC ao serem aplicadas após a etapa de seleção de características. Por fim, o Capítulo 5 apresenta as conclusões acerca dos resultados obtidos com o trabalho, assim como propõe possíveis trabalhos futuros.

CAPÍTULO 2

Revisão Bibliográfica

Este capítulo aborda conceitos e técnicas diversas utilizados no decorrer deste trabalho. A Seção 2.1 apresenta o modelo VSM *do inglês, Vector Space Model* para a representação textual. Na Seção 2.2 são introduzidos os conceitos referentes ao modelo baseado em grafos para representação de documentos. A Seção 2.3 apresenta as principais características da Enciclopédia Wikipédia, incluindo detalhes da linguagem de marcação *Wiki Markup* que forem pertinentes ao presente trabalho. A Seção 2.4 apresenta as teorias relacionadas à área de classificação automática de documentos utilizando aprendizado de máquinas, classificação uni-classe e multi-classe, apresenta a teoria introdutória sobre o algoritmo de classificação SVM, além de introduzir os conceitos relacionados aos métodos de avaliação no que tange à classificação de documentos. Por fim, na Seção 2.5 são discutidos os trabalhos relacionados ao tema proposto pelo presente estudo.

2.1 O Modelo Espaço Vetorial

Durante o processo de classificação textual, o primeiro e vital passo é a representação textual a qual converte o conteúdo de um documento de texto em um formato compacto de modo que o mesmo possa ser identificado e classificado por um computador ou classificador. Neste contexto, o Modelo Espaço Vetorial (VSM, do inglês *Vector Space Model*) consiste em um modelo algébrico para a representação de textos, por meio do qual documentos e consultas são representados como vetores compostos por termos [28] [47] [58]. Este tipo de representação é largamente utilizado em Recuperação de Informação, tanto em tarefas de recuperação textual e ordenação de documentos por relevância (do inglês, *ranking*), quanto em tarefas de classificação de documentos. A utilização da representação por vetor torna possível o uso de qualquer operação algébrica aplicável a este tipo de estrutura, possibilitando comparar consultas com documentos, ou comparar a semelhança entre dois documentos, como exposto por Salton [46] [35], razões pelas quais neste trabalho optou-se por utilizar este tipo de representação.

2.1.1 Representação de características de documentos

Termos, também chamados de características (do inglês *features*) [52], são unidades indexáveis usadas para identificar o conteúdo do documento de texto, podendo ser descritos em vários níveis de granularidade, como sílabas, uma palavra (uni-gramas), várias palavras (bigramas, trigramas ou n-gramas, de forma mais geral), frases, ou qualquer outra unidade semântica e/ou sintática mais elaborada [28].

A representação de características mais utilizada na CAT, também conhecida como representação do conjunto de termos, ou também como BOW (do inglês, *bag of words*), considera apenas palavras como termos. No entanto, esta abordagem desconsidera importantes relações semânticas entre os termos [22] [64]. De fato, elementos como "Casa Branca" ou "*Bill Gates*" são representadas no BOW como palavras desassociadas. Ao se analisar a representação BOW de um dado documento no qual ocorrem as palavras *Bill* e *Gates*, poderia-se sugerir que tal documento trata de assuntos como contabilidade devido à palavra *Bill* (que significa conta, em inglês) ou sobre construções devido à palavra *Gates* (que significa portões, em inglês), de forma que seria muito difícil associá-lo a programas de computadores. Entretanto, se a representação do mesmo documento contiver o conjunto de palavras "*Bill Gates*" como sendo um termo, dificilmente o leitor confundiria o tema tratado pelo documento [4].

Com a intenção de encontrar uma representação que conseguisse expressar de forma mais correta os conceitos tratados em um texto, Lewis [29], utilizou um parser para criar frases sintáticas como termos indexáveis. Tais frases correspondem a pares de palavras que mantêm algum dos muitos específicos relacionamentos sintáticos no documento original. Como exemplo pode-se citar o verbo e o substantivo chave de algum assunto, substantivo e adjetivo, etc. De maneira intuitiva, a utilização de frases contribui para a redução de incertezas no significado de palavras. Como exemplo pode-se perceber que em "java Script" o significado da palavra java só pode ser "um tipo de linguagem de programação" evitando que a palavra seja confundida com a ilha da Indonésia também de nome java. Entretanto, existe um número muito grande de frases distintas nas coleções, porém a frequência de cada frase é pequena, limitando a contribuição deste tipo de representação. Outro ponto relevante é o fato de que a representação por frases sintáticas é altamente redundante, ou seja, há um grande número de frases que possuem essencialmente o mesmo significado. Além disso, esse tipo de representação tem se mostrado muito ruidosa.

Vários trabalhos demostraram que o uso de frases como características possui eficácia inferior quando se comparado à representação BOW apenas com termos [2] [30] [31]. As razões dos resultados ruins para esta abordagem foram apresentadas

por [31] o qual constatou que a pequena ocorrência 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.

Muitos esforços têm sido realizados visando a contornar os possíveis problemas apresentados no uso de frases e alguns trabalhos têm gerado resultados que mostram que a adição de n-gramas (sequência de n palavras) tem melhorado o processo de CAT, quando comparado com a representação BOW, conforme exposto por Tan [59].

A representação BOW pode ser enriquecida criando novos termos formados por n-gramas cujas palavras formadoras já ocorrem na representação de conjunto de termos, como exposto por Mladenic [41] e Tan [59]. Neste método, adicionam-se à representação BOW já existente, tanto os n-gramas como as palavras que os formam, de modo a enriquecer a representação BOW. Outra forma de utilização dos n-gramas é a sua adição à representação BOW, porém as palavras que ajudam a compor os n-gramas não são incluídas. Tan [59] analisou ambas as abordagens demonstrando que a não utilização dos termos formadores dos n-gramas, na maioria dos casos degrada o desempenho da BOW, enquanto que a utilização destes elementos pode, potencialmente, melhorar os resultados da CAT.

Neste trabalho, utiliza-se a representação BOW enriquecida com unigramas, bigramas e trigramas oriundos da Wikipédia. Dessa forma, tanto os n-gramas quanto as palavras que formam os mesmos, estão presentes na representação BOW. Mais detalhes sobre o modo como os n-gramas são identificados, serão mostrados nas seções seguintes.

Independente da granularidade utilizada na representação das características, ou seja, tanto no uso de uma palavra ou n-gramas, cada termo distinto se posiciona como um item do conjunto de termos T (também denominado vocabulário) da coleção de documentos. Nesse sentido, o conjunto de documentos D de uma coleção são representados no VSM como pontos em um espaço euclidiano multidimensional, no qual cada dimensão corresponde a um termo distinto dessa coleção. Por este motivo diz-se que o vocabulário define um espaço vetorial $|T|$-dimensional e cada documento é representado com um vetor neste espaço.[52]. Na Seção 2.1.4 será mostrado como quantificar a importância de cada termo para cada documento.

2.1.2 Termos não discriminativos

Durante o processo de representação de documentos, um dos procedimentos que visam a melhoria do conjunto de características utilizadas é a remoção de termos que não colaboram com o processo de classificação. Palavras que são extremamente comuns não trazem nenhuma melhoria para a CAT [35]. Como exemplo pode-se citar

artigos, advérbios, conjunções e qualquer outro elemento os quais não caracterizam nenhum tópico específico, pois seu uso é apenas funcional para a adequação às regras sintáticas da língua.

O conjunto de palavras que possuem estas características é denominado de lista de exclusão (do inglês, *stop list*), assim como as palavras que fazem parte deste conjunto são denominadas de palavras de exclusão (do inglês, *stop words*). Tais elementos podem ser removidos do vocabulário da coleção sem que haja prejuízo para a CAT. Cada idioma possui sua própria *stop list*, cada qual podendo conter adjetivos, pronomes, advérbios, verbos comuns, substantivos comuns, etc. Palavras que apresentam uma grande incidência em uma determinada coleção também podem ser incluídas na *stop list* utilizada durante a classificação da mesma. Entretanto, a exclusão de *stop words* nem sempre é utilizada. Há casos em que utiliza-se a pesquisa por frases, em que todos os termos devem ser indexados de forma que se saiba inclusive a posição em que os mesmos ocorrem [35].

2.1.3 Expansão de características de documentos

Os métodos que adicionam novas características com o objetivo de melhorar a representação BOW de documentos têm mostrado grande potencial na melhoria da classificação automática de textos. Tais métodos se apoiam na geração de características (FG, do inglês *Feature Generation* ou *Feature Construction*) por meio do qual é possível a expansão de documentos (do inglês *Document Expansion*) ou enriquecimento de documentos (do inglês *Document Enrichment*), adicionando-se novos termos aos documentos, de forma a melhorar a representação BOW através da inserção de características mais informativas na matriz documento-termos [18].

A Expansão de características de documentos pode ser realizada de diversas formas, utilizando tanto elementos disponíveis dentro do próprio documento e coleção, como abordado por [14] [17] [41] ou por meio da utilização de fontes externas de informações (do inglês, *external corpus*) como realizado por [18] [23] [33] [64].

2.1.4 Medidas de importância dos termos

No modelo espaço vetorial, uma coleção D de documentos é dada por $D = \{d_1, d_2, ..., d_{|D|}\}$ onde $|D|$ representa o total de documentos distintos em toda a coleção. Por sua vez, o vocabulário T de todos os termos distintos que aparecem nos documentos de D é o conjunto dado por $T = \{t_1, t_2, ...t_{|T|}\}$ onde $|T|$ é a quantidade total de termos distintos que ocorrem no vocabulário da coleção. Diferentes termos possuem diferentes graus de importância para o texto do documento, de modo que, para todo par (t_i, d_j), em que $t_i \in T$ e $d_j \in D$, associa-se um peso w_{ij}. O Peso

w_{ij} visa a expressar o quanto o termo t_i contribui com o significado semântico do documento d_j. Dessa forma, um documento d_j é representado como um vetor de pesos de termos, sendo $d_j = \left(w_{1j}, w_{2j}, ...w_{|T|j}\right)$ [28] [35] [45] [46] [47].

Na abordagem VSM, uma coleção de documentos (também chamada de *corpus*) é mapeada como uma matriz documento-termo M, na qual cada linha representa um documento d_j do conjunto de documentos D e cada coluna representa um termo t_i do conjunto de todos os termos distintos T presentes na coleção. Cada posição (j, i) representada nesta matriz equivale ao peso para cada termo em cada documento [57], conforme ilustrado na equação 2-1:

$$M(D,T) = \begin{pmatrix} w(t_1,d_1) & \dots & w(t_{|T|},d_1) \\ \vdots & \ddots & \vdots \\ w(t_1,d_{|D|}) & \dots & w(t_{|T|},d_{|D|}) \end{pmatrix} \tag{2-1}$$

A partir da Matriz documento-termo M, é possível construir a matriz termo-documento M' a qual mapeia, para cada termo, qual o peso dado a este em cada documento. A matriz termo-documento é particularmente útil na criação de um índice de consultas por termos, o qual também é conhecido como índice invertido.

O método de atribuição de pesos a termos é um procedimento de extrema importância no intuito de melhorar a eficácia da CAT. Para tanto, é necessário que seja atribuído a cada termo um peso que de fato corresponda à sua importância no processo de classificação.

A métrica binária é a maneira mais simples de se atribuir peso a um termo t_i que ocorre no documento d_j. Como exposto por [35], nesta metodologia utiliza-se os valores 1 e 0 como possíveis pesos, de modo que se um termo t_i aparece no documento d_j, terá peso 1 ou, se não aparece no mesmo, terá peso 0, como visto na equação 2-2.

$$w_{bin}(t_i,d_j) = \begin{cases} 1, & \text{se o termo } t_i \text{ aparece no documento } d_j \\ 0, & \text{caso contrário} \end{cases} \tag{2-2}$$

Como pode ser percebido, todos os termos presentes no vocabulário T e que aparecem no documento d_j, possuem o mesmo peso e portanto o mesmo grau de importância, mesmo que isto não represente a realidade semântica dos termos. A abordagem binária também ignora o número de ocorrências dos termos presentes no documento, como exposto por Lan et al. [28].

Com o propósito de estabelecer uma relação mais realista a cerca da importância de um termo dentro do documento no qual o mesmo ocorre, e consequente-

mente, em busca de obter uma melhor eficácia no processo de CAT, foram propostas várias abordagens na determinação dos pesos w_{ij} dos termos, dentre as quais está a frequência do termo (*TF* do inglês, *Term Frequency*) dado por $tf(t_i, d_j)$ o qual representa o número de vezes que o termo t_i ocorre no documento d_j. A partir desta medida, quanto maior a quantidade de ocorrências do termo t_i no documento d_j, maior a importância de t_i termo neste documento.

Outra importante métrica é o cálculo da frequência inversa em documentos (*IDF*, do inglês *Inverse Document Frequency*), o qual é dado pela equação 2-3:

$$idf(t_i) = log\left(\frac{|D|}{df(t_i)}\right) \tag{2-3}$$

onde $|D|$ é o total de elementos distintos de uma coleção e $df(t_i)$ é a quantidade de documentos da coleção D onde o termo t_i ocorre. A métrica *IDF*, parte do princípio de que se um termo ocorre em um número muito grande de documentos de uma coleção, este termo tende a não ser um bom discriminador dos documentos nos quais o mesmo ocorre. Por outro lado, se um termo ocorre em poucos documentos, o mesmo tende a ser um bom discriminador do temas sobre o qual tais documentos tratam.

A utilização em conjunto das métricas *TF* e *IDF* tem sido um dos métodos mais utilizados na determinação da importância de termos no processo de classificação [47] [52].

O cálculo do *TF-IDF* é representado pela equação 2-4:

$$tfidf(t_i, d_j) = tf(t_i, d_j).idf(t_i) \tag{2-4}$$

Em muitos trabalhos, observa-se o uso do *TF-IDF* sobre o qual é aplicada a normalização por cosseno, o que visa a manter a atribuição de pesos dentro da faixa entre 0 e 1. Este procedimento tem por objetivo minimizar o efeito causado pelo tamanho dos documentos, visto que documentos maiores tendem a possuir um número maior de repetição, o que naturalmente aumentaria de forma exagerada os pesos dos termos presentes no mesmo, conforme exposto por [52] [47]. No presente trabalho utiliza-se o *TF-IDF* normalizado por cosseno para o devido cálculo do peso w.

O cálculo do *TF-IDF* normalizado por cosseno, pode ser visto na equação 2-5:

$$tfidf\ normalizado\, w(t_i, d_j) = \frac{tfidf(t_i, d_j)}{\sqrt{\sum_{s=1}^{|T|} \left(tfidf(t_s, d_j)\right)^2}} \tag{2-5}$$

onde o denominador da equação representa a distância euclidiana aplicada a todos os termos do conjunto de termos T com seus respectivos valores de *TF-IDF*.

2.1.5 Medidas de similaridade entre documentos

O modelo VSM possibilita a verificação da similaridade entre dois documentos, a qual pode ser conseguida por meio do cálculo da distância Euclidiana, ou do cosseno, dentre outros.

O cálculo da distância Euclidiana é mostrado na equação 2-6:

$$euc(d_j, d_k) = \sqrt{\sum_{i=1}^{|T|} (w_{ij} - w_{ik}))^2} \tag{2-6}$$

onde w representa o peso de um termo em um dado documento, como visto na Seção 2.1.4.

Uma medida muito utilizada para calcular a similaridade entre dois documentos é o cálculo do cosseno do ângulo θ formado pelos vetores correspondentes a esses documento no espaço $|T|$-dimensional. A medida dos cossenos é expressa pela Equação 2-7:

$$cos(d_j, d_k) = \frac{\sum_{i=1}^{|T|} (w_{ij} \times w_{ik})}{\sqrt{\sum_{i=1}^{|T|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|T|} w_{ik}^2}} \tag{2-7}$$

que consegue extrair a similaridade entre dois vetores por meio do cálculo do cosseno do ângulo existente entre eles. O ângulo 0 (zero) entre dois vetores que representam documentos, informa que ambos possuem uma representação VSM igual. Para o ângulo 0, o cálculo do cosseno terá o valor 1.

Diante do exposto, quanto mais similares são determinados dois documentos, mais próximo de 1 será o valor do cosseno. Por outro lado, um resultado de cosseno igual a 0 indica que os vetores correspondentes são ortogonais entre si e que os documentos não possuem termos em comum [53].

2.1.6 Métodos de Seleção de Características

Conforme já visto na Seção 2.1.4, no Modelo Espaço Vetorial a dimensionalidade do espaço de representação das características possui o tamanho do vocabulário da coleção, podendo chegar a dezenas de milhares de termos. De acordo com Yang [69], a alta dimensionalidade pode se tornar proibitivamente alta para muitos algoritmos de aprendizado durante o processo de classificação de documentos. Portanto,

é altamente desejável reduzir o número de termos sem que este procedimento resulte em diminuição da eficácia do processo de classificação. Neste contexto, os métodos utilizados para a redução do vocabulário visam a manter termos bons discriminadores de classes presentes nos documentos, ao passo que termos pouco discriminativos são removidos. Este processo é conhecido como *seleção de características* (do inglês, *feature selection*) ou seleção de termos (do inglês, *term selection*).

Uma seleção de características eficaz deve ser capaz de definir valores de importância dos termos no processo de classificação. Dessa forma, a partir de um conjunto de termos T são mantidos apenas os k termos que possuem maior poder discriminativo de categoria. Tais elementos formam o subconjunto T'. Neste processo são descartados os termos que não contribuem com a classificação ou até impactam negativamente no processo de CAT [69] [15].

De acordo com Debole e Sebastiani [12], o fator de redução do conjunto de termos T é dado pela equação 2-8:

$$\xi = \frac{|T| - |T'|}{|T|} \tag{2-8}$$

Ainda segundo [12], usualmente as técnicas de seleção de termos consistem em atribuir uma nota a um termo t de forma a valorar o poder discriminativo do mesmo na coleção. Para esta tarefa utiliza-se uma função de avaliação de termos f. O próximo passo é a escolha dos k termos mais bem avaliados pela função f, os quais serão formadores do conjunto T'.

As subseções seguintes apresentam três funções de avaliação de termos bastante utilizadas na literatura, as quais são utilizadas no presente trabalho.

Information Gain

Information Gain(IG) mede o ganho de informação trazido pela presença ou ausência de um dado termo $t_i \in T$ e uma classe $c_j \in C$, conforme equação 2-9:

$$IG(t_i, c_j) = \sum_{c \in \{c_j, \overline{c}_j\}} \sum_{t \in \{t_i, \overline{t}_i\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)} \tag{2-9}$$

onde $\overline{c_j}$ representa alguma classe diferente de c_j e $\overline{t_i}$ representa um termo distinto de t_i. O valor $P(t, c)$ corresponde à probabilidade conjunta do termos t e da classe c. Se o valor de IG de um termo para um termo t_i para uma classe c_j for alto, tal termo é considerado importante e informativo para a classe c_j. Por outro lado, se o valor obtido for baixo, este termo não traz informações a respeito do tópico relacionado à classe c_j e consequentemente pode ser removido. Utilizando a medida *information gain* pode-se eliminar um grande número de termos reduzindo $|T|$ a um

número de 100 a 300 termos, ao mesmo tempo que consegue-se melhorar o processo de classificação.

Gain Ratio

Gain Ratio (GR) é uma variante normalizada do *Information Gain*. O *Gain Ratio* tenta resolver uma deficiência apresentada pelo *Information Gain*, o qual apresenta valores elevados não somente com a elevação da dependência entre t e c, mas também com o aumento da entropia de c. Esta deficiência deixa de existir visto que GR normaliza o IG com o valor da entropia da classe, como pode ser visto na equação 2-10:

$$GR(t_i, c_j) = \frac{IG(t_i, c_j)}{-\sum_{c \in \{c_j, \bar{c}_j\}} P(c) \log_2 P(c)} \tag{2-10}$$

Chi-Squared(X^2)

A medida *Chi-squared* caracteriza-se por ser um teste estatístico que mede a divergência existente entre uma distribuição esperada ao assumir-se que a ocorrência de um termo t_i é independente de determinada classe c_j. Quanto maior o valor de *Chi-squared*(X^2), maior é a dependência entre termo e classe. Um valor igual a 0 indica independência entre tais elementos [15]. Por ser este um teste estatístico, podem ocorrer comportamentos errôneos quando esta medida é aplicada a termos raros na coleção, ou quando o número de exemplos positivos no treino é muito escasso para um determinado conceito. A medida *Chi-squared* é representada pela equação 2-11:

$$X^2(t_i, c_j) = \frac{[P(t_i, c_j)P(\bar{t}_i, \bar{c}_j) - P(t_i, \bar{c}_j)P(\bar{t}_i, c_j)]^2}{P(t_i)P(\bar{t}_i)P(c_j)P(\bar{c}_j)} \tag{2-11}$$

Todas as funções acima tentam capturar a intuição de que os termos mais valiosos para a caracterização de uma categoria c_j são aqueles que estão distribuídos o mais diferentemente possível entre os exemplos negativos e positivos de documentos de c_j

É importante salientar que as métricas acima relacionadas indicam a importância de um termo t_i para uma classe c_j. Dessa forma, um mesmo termo pode possuir $|C|$ valores diferentes de *Information Gain, Gain Ratio* ou *Chi-squared*. Neste contexto, de acordo com Sebastiani [52] é necessário a aplicação de um método que objetiva extrair um valor global $f_{global}(t_i)$ a partir dos valores obtidos localmente para cada classe por meio da função local de avaliação $f(t_i, c_j)$. As técnicas de globalização mais comuns são as seguintes:

$$f_{sum}(t_i) = \sum_{j=1}^{|C|} f(t_i, c_j) \tag{2-12}$$

$$f_{wsum}(t_i) = \sum_{j=1}^{|C|} P(c_j) f(t_i, c_j) \tag{2-13}$$

$$f_{max}(t_i) = \max_{j=1}^{|C|} f(t_i, c_j) \tag{2-14}$$

as quais representam, respectivamente, a soma dos valores obtidos pelas funções locais (2-12); a soma ponderada pela probabilidade de ocorrência da classe específica (2-13); e por fim o valor de importância global que é dado pelo maior valor de função local encontrado para um determinado termo t_i ao se comparar os valores obtidos para todas as categorias do conjunto C (2-14). No presente trabalho utiliza-se a medida de valoração global para importância do termo, de modo que para esta tarefa optou-se pelo método de valor máximo local como exposto na equação 2-14.

2.2 O Modelo Baseado em Grafos

Coleções de documentos que possuem elementos que interconectam documentos podem ser modeladas como um grafo direcionado $G = (D, E)$, onde o conjunto de vértices D representa o conjunto de documentos e o conjunto de arestas direcionadas E representa o conjunto de ligações (do inglês, *links*) entre os documentos. O uso de *links* como descrito acima pode ser encontrado em várias coleções como: páginas Web, bibliotecas digitais, enciclopédias, etc.

Por se tratar de um grafo direcionado, para cada documento d podem existir arestas de entrada, representadas pelos *links* com origem em outros documentos e que apontam para d (do inglês, *in-links*), e podem existir arestas de saída, as quais têm origem em d e apontam para outros documentos o (do inglês, *out-links*) [24]. A estrutura descrita é ilustrada por meio da Figura 2.1.

Como exposto por [10, 44, 56], cada *out-link* presente no documento denota uma citação que tal documento faz a um outro documento para o qual o *out-link* aponta. Da mesma forma, um *in-link* direcionado a um documento, denota que o mesmo foi citado por outro documento. Neste contexto, os textos âncoras (do inglês, *anchor texts*) também presentes na Figura 2.1, funcionam como apelidos utilizados pelo documento que realizou a citação. Com isso, é possível nomear o documento citado, de forma arbitrária. Os *links* denotam uma relação semântica de afinidade entre o conteúdo do documento que realiza a citação e do documento citado. Os

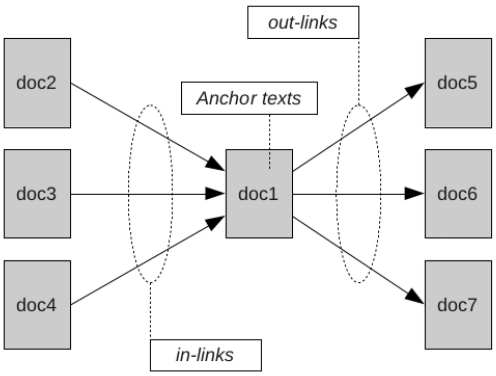

Figura 2.1: *Ligações entre documentos utilizando links*

textos âncoras por sua vez representam uma relação semântica de sinonímia entre a expressão formadora do texto âncora e o tema do documento para o qual o *link* aponta [11, 37].

2.3 Enciclopédia Wikipédia

A Wikipédia[1] é uma Enciclopédia Web colaborativa voluntária, que utiliza o formato Wiki, disponível mundialmente com versões em várias línguas, chegando a um total de 280 idiomas [65]. Lançada em 15 de Janeiro de 2001 por Jimmy Wales e Larry Sanger, sua facilidade de acesso via Internet e a liberdade para edição do conteúdo possibilitam que seus artigos sejam atualizados constantemente sendo que novos artigos são adicionados diariamente, alcançando um total de 20 milhões de artigos, espalhados pelas diversas versões da enciclopédia, que correspondem a idiomas distintos. A maior versão é a inglesa, a qual, em novembro de 2011, possuía mais de 3,7 milhões de artigos diferentes.

O conteúdo da Wikipédia é criado a partir de uma linguagem de marcação própria para conteúdo colaborativo do tipo Wiki chamada de linguagem Wikitext, também conhecida como *Wiki Markup*. A partir do código feito nesta linguagem um sistema o converte em uma representação HTML para a devida interpretação em navegadores Web. Por meio desta linguagem é possível a criação de *links* internos entre os artigos da Wikipédia. Para tanto esta enciclopédia utiliza o que é denominado links livres (do inglês, *free links*). Na linguagem *Wiki Markup* cria-se um *free link* utilizando-se colchetes duplos [[...]], sendo que o conteúdo dentro dos colchetes é efetivamente o nome oficial de outro artigo que também está presente na

[1] http://www.wikipedia.org

Wikipédia para o qual se deseja criar um link. Na Figura 2.2 é possível visualizar um exemplo deste tipo de estrutura.

Código *Wiki Markup*	Como o código será Visualizado
`London has [[public transport]].`	London has public transport.

Figura 2.2: *Criação de free links utilizando Wikitext.*

Como visto na Figura 2.2, o texto "*public transport*" atua como um *hyperlink* para o artigo de mesmo nome.

Sabendo que cada artigo da Wikipédia corresponde a um conceito bem definido, a possibilidade de interconectar os artigos por meio de *hyperlinks* representa a possibilidade de estabelecer relacionamentos entre conceitos [43], assim como exposto na Seção 2.2, também possuindo *in-links* e *out-links*. De acordo com Milne [39], os *hyperlinks* entre os artigos da Wikipédia conseguem capturar grande parte das relações semânticas definidas pela ISO 2788 que tratam de padrões internacionais para Tesauros[2].

Os textos âncoras presentes em páginas HTML também possuem seu correspondente na linguagem *Wikitext*. Este elemento é criado a partir do uso do símbolo *pipe* (|) dentro dos colchetes duplos que compõe um *free link*. O nome correto do artigo referenciado pelo *free link* aparece à esquerda do *pipe* e à direita encontra-se o texto âncora que será utilizado como apelido para o *free link* verdadeiro, como mostrado na Figura 2.3.

Código *Wiki Markup* com anchor	Como o código será Visualizado
`Mais um gol do [[Pelé| Rei do Futebol]].`	Mais um gol do Rei do Futebol.

Figura 2.3: *Criação de textos âncoras utilizando free links.*

A nomeação arbitrária gerada pelos textos âncoras possibilita a utilização de palavras ou expressões totalmente diferentes dos nomes reais dos artigos, como no caso expressado pela Figura 2.3, tornando possível o estabelecimento de relações de sinonímia entre o texto âncora e o título do artigo.

Cada conceito da Wikipédia deve ser representado por um único artigo, de modo que artigos que representam o mesmo conceito apenas redirecionam o leitor para o artigo principal. Estes tipos de artigos são criados a partir dos *links* de redirecionamento, que são inseridos nos artigos dos quais se deseja redirecionar. Esta situação é ilustrada pela Figura 2.4:

[2] Um tesauro consiste em um tipo de dicionário que define um conjunto de conceitos além de especificar as relações entre tais conceitos.

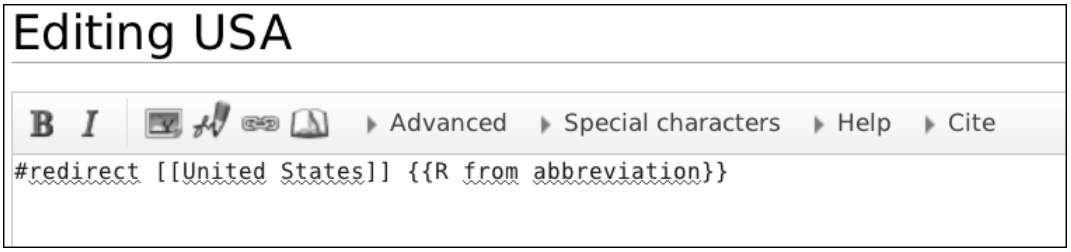

Figura 2.4: *Links de redirecionamento de páginas na Wikipédia.*

na qual pode ser observado que a presença da palavra-chave **#redirect** torna o *hyperlink* como do tipo de redirecionamento. Neste exemplo, portanto, o artigo "USA" serve apenas para levar o leitor para o artigo "United States".

Vários trabalhos destacam a relação de sinonímia existente entre o texto do título do artigo de um determinado conceito da Wikipédia e os textos do título dos seus artigos de redirecionamento [22, 64, 67]. Como exemplo da importância dos redirecionamentos, pose-se destacar o caso do artigo "United States", para o qual foram observados 78 artigos de redirecionamento.

Cada artigo da Wikipédia pertence a uma ou mais categorias, e esta relação é expressa por meio dos *links* de categorias. Tais *links* são escritos em *Wikitext* também utilizando o formato *free links*, porém com a adição da palavra chave **Category:** antecedendo o nome da categoria a qual o artigo foi associado. (Ex.: [[Category:Cities]]). As próprias categorias podem pertencer a outras categorias [65]. Esta estrutura gera uma relação de afiliação entre um conceito (representado pelo artigo) e uma categoria, o que permite derivar relações de hiponímia entre ambos. A partir desta mesma relação entre as categorias, é possível extrair uma estrutura hierárquica entre estes elementos [39].

Devido às várias características mostradas acima, como relações associativas entre *links*, possibilidade de se derivar relações de sinonímia entre conceitos, além da estrutura hierárquica estabelecida pelas categorias, a Wikipédia tem se tornado objeto de inúmeras pesquisas acerca do uso destas e de outras informações para a construção de tesauros [24, 39, 42, 43]. Neste contexto, esta enciclopédia tem se mostrado como um excelente repositório de informações relevantes para a melhoria do processo de CAT [33], de forma que palavras-chave associadas de alguma forma podem ser consideradas durante este processo, fazendo com que documentos que compartilhem tais palavras tendam a ser classificados em uma mesma classe.

2.4 Classificação de Documentos

A tarefa de classificação se faz necessária em uma ampla gama de atividades humanas, e em seu sentido mais geral, o termo poderia cobrir qualquer contexto em que alguma decisão ou previsão é feita com base em informações atualmente disponíveis. Dessa maneira, um procedimento de classificação é, portanto, algum método formal para repetidamente fazer julgamentos sempre que novas situações são apresentadas [38].

A classificação de documentos tem como objetivo categorizar documentos de acordo com um conjunto de categorias predefinidas, a partir da análise do conteúdo de tais documentos. Categorias também são referenciadas como classes ou rótulos [31, 35]. Dessa forma, dada uma coleção de documentos $D = \{d_1, d_2, \ldots d_{|D|}\}$ e um conjunto fixo de categorias $C = \{c_1, c_2, \ldots c_{|C|}\}$, a classificação de textos é a tarefa de atribuir um valor booleano (1 ou 0) para cada par $(d_j, c_i) \in D \times C$. Quando $(d_j, c_i) = 1$ (ou T, do inglês *True*) o documento d_j está rotulado como pertencente à categoria c_i e quando $(d_j, c_i) = 0$ (ou F, do inglês *False*) o documento d_j não está rotulado como pertencente à categoria c_i. Este processo corresponde à função de classificação $\Phi : D \times C \rightarrow \{T, F\}$.

2.4.1 Classificação Automática de Documentos utilizando Aprendizado de Máquinas

A partir dos anos 90 a abordagem baseada em engenharia do conhecimento para a classificação de documentos, começou a perder espaço para outro paradigma, o aprendizado de máquinas (do inglês, *Machine Learning*). O aprendizado de máquinas (AM) utiliza-se de um processo indutivo geral para construir automaticamente um classificador para uma categoria $c_i \in C$. Este processo é realizado por meio da análise das características de um conjunto de documentos já classificados manualmente por um especialista de domínio (conjunto de treino) como sendo pertencentes à c_i ou não pertencentes ($\overline{c_i}$). Após o processo de aprendizagem das características proporcionado pela indução, um novo documento cuja categoria não se sabe, pode ser automaticamente classificado como pertencente ou não à c_i. Por ser supervisionado pelo conhecimento prévio das categorias dos documentos do conjunto de treino, esse processo de construção do classificador é denominado *supervisionado* por vários autores [2, 52, 68].

A abordagem de aprendizado de máquinas para a classificação de documentos tem se tornado atrativa, principalmente devido ao vasto número de aplicações na Web que demandam a classificação textual [52, 35, 38]. Dentre elas pode-se citar:

- Os catálogos de recursos Web;
- A detecção de *spans* (mensagens indesejadas e por diversas vezes mal intencionadas) em e-mails;
- A organização temática de canais de notícias com o intuito de satisfazer as preferências dos usuários;
- A personalização de publicidades por áreas de interesse, além do auxílio no diagnóstico de doenças de acordo com determinados quadros clínicos;
- A indexação de documentos com base em um vocabulário controlado;
- A filtragem de documentos;
- A geração automática de metadados;
- A desambiguação de sentidos de palavras.

A classificação automática de documentos por meio da abordagem de aprendizado de máquinas requer, inicialmente, a disponibilidade de um *corpus* inicial $\Omega = \{d_1, d_2, ..., d_{|\Omega|}\} \subset D$ de documentos pré-classificados manualmente por um especialista, em um conjunto de categorias $C = \{c_1, c_2, \dots c_{|C|}\}$. Dessa forma, o resultado da função de classificação objetivo $\Psi : D \times C \rightarrow \{T, F\}$ tem seus valores conhecidos para todo par $(d_j, c_i) \in \Omega \times C$.

A partir do processo de aprendizagem, deriva-se então função de classificação $\Phi : D \times C \rightarrow \{T, F\}$ que mapeia documentos em classes, também conhecido como classificador. Portanto, formalmente, a tarefa de classificação busca aproximar o máximo possível a função de classificação Φ com o valor desconhecido da função objetivo $\Psi : D \times C \rightarrow \{T, F\}$ de forma que o resultado de Φ e Ψ coincidam o máximo possível. Dessa forma, após a obtenção do classificador Φ por meio do processo de aprendizagem, é necessário avaliar a eficácia comparando os resultados obtidos com os resultados esperados da função Ψ [52].

Para se realizar o treino e avaliação do classificador são necessários dois subconjuntos distintos de Ω, T_r e T_e, tais que $T_r \cap T_e = \emptyset$):

- **Conjunto de Treino** T_r - utilizado para se obter o classificador Φ. O classificador é treinado aprendendo as características dos documentos do conjunto de treino o qual já foi classificado manualmente.
- **Conjunto de Teste** T_e - utilizado para avaliar a eficácia do classificador obtido Φ. Para cada documento d_j pertencente ao conjunto de teste é conhecida a classe (ou classes) a que pertence, porém esta informação não é repassada ao classificador Φ criado na etapa de treino. Cada documento d_j de T_e é submetido ao classificador Φ que atribui uma ou mais classes de C a d_j, comparando as características presentes em d_j com as características aprendidas durante a etapa de treino.

A próxima etapa é a avaliação de cada decisão realizada pelo classificador Φ para cada par (d_j, c_i), a qual é comparada com a decisão esperada $\Psi(d_j, c_i)$, de forma que quanto maior o número de decisões de Φ que forem iguais às decisões de Ψ, mais eficaz é o classificador criado.

Um classificador Φ deve possuir uma boa capacidade de generalização, de forma a eliminar erros causados por sobre-ajuste (do inglês, *overfitting*). Este tipo de problema ocorre quando um classificador se adapta à características pontuais dos documentos de treino, o que pode diminuir a taxa de acerto na classificação de novos documentos.

O processo de avaliação da classificação será abordado com maiores detalhes na Seção 2.4.4.

2.4.2 Classificação Uni-classe e Multi-Classe

A classificação de textos pode estar sujeita a diferentes restrições, dependendo de sua aplicação. Dentre tais restrições está a limitação do número de classes pertencentes ao conjunto de categorias C as quais um documento pode ser associado pelo classificador. Para o caso em que somente uma categoria deve ser associada a cada $d_j \in \Omega$, dá-se o nome de classificação uni-classe. Um caso especial de classificação uni-classe é a classificação binária, por meio da qual cada documento $d_j \in \Omega$ deve ser classificado como pertencente à categoria c_i ou ao seu complemento $\overline{c_i}$. Como exemplo da classificação binária, pode-se citar filtros de *spams*, por meio do qual mensagens recebidas devem ser classificadas como sendo do tipo *spam* ou não-*spam*. Para o caso em que qualquer número de categorias, de 0 a $|C|$, podem ser atribuídas a um documento $d_j \in \Omega$ dá-se o nome de classificação multi-classe.

Teoricamente, a abordagem binária é dita como sendo uma abordagem mais geral que a abordagem multi-classe, visto que a classificação binária pode ser empregada em problemas de classificação multi-classe. Para tanto, basta transformar um problema de classificação multi-classe, com documentos podendo pertencer à $(c_1, \ldots, c_{|C|})$, em $|C|$ problemas independentes de classificação binária sobre c_i ou $\overline{c_i}$, com $i = 1, \ldots, |C|$. Neste caso, $\overline{c_i}$ é formado pelos documentos que pertencem a todas as outras categorias e é chamado como metodologia *um contra todos os outros* (do inglês, *one against others*). Esta interpretação só é possível quando as categorias envolvidas são estocasticamente independentes, ou seja, a classificação de um documento em uma categoria não exige que este mesmo documento também seja categorizado em outra.

De acordo com exposto por Shen [54], sendo a CAT um tipo específico de problema de classificação de padrões, o algoritmo de classificação, juntamente

com a metodologia de representação dos documentos, são aspectos essenciais que contribuem com a eficácia deste processo.

Ao longo de algumas décadas um vasto número de algoritmos tem sido proposto para CAT utilizando aprendizado de máquinas. Dentre eles pode-se citar o *naive bayes* [36], k-vizinho mais próximo (do inglês *k-nearest neighbor - KNN*) [60], máquinas de vetor de suporte (do inglês *support vector machines - SVM*) [25], *boosting* [48] e algoritmos de aprendizado de regras (do inglês, *rule learning algorithms*) [55], os quais têm sido amplamente utilizados.

Vários trabalhos que abordam a classificação de documentos reportaram comparações de desempenho entre os diversos algoritmos disponíveis. Figueiredo [14] utilizou em seu trabalho os algoritmos *KNN*, *Naive Bayes* e *SVM*, sendo que o último apresentou melhor desempenho na maioria dos casos analisados de classificação de textos, até mesmo ao se comparar a linha base deste com os melhores resultados dos outros algoritmos após a aplicação do método proposto.

Nos resultados apresentados por Zaiane [70] e Lewis [32] também é possível observar um melhor desempenho geral do algoritmo SVM.

Em seus trabalhos, Wang [63, 64], Gantner e Schmidt-Thieme [20], Gabrilovich [18, 19] e Bekkerman [5] optaram por utilizar apenas o classificador *SVM*, por ser mostrar um algoritmo do estado da arte em classificação de documentos. [63, 68].

No presente trabalho, também optou-se por utilizar apenas o classificador SVM, haja vista seu alto desempenho em classificação textual, como exposto nos parágrafos anteriores. O algoritmo já possui uma linha base alta, sendo o mesmo excelente para demonstrar a validade do método proposto.

2.4.3 Algoritmo de Classificação SVM

O algoritmo de classificação *Support Vector Machines* (SVM) é um método relativamente recente, introduzido por Vapnik [61] e utilizado na classificação de documentos primeiramente por Joachims [25], de modo que sua utilização em classificação de documentos apoia-se em características apontadas por [6, 35], como:

- Boa capacidade de generalização;
- Robustez em situações de alta dimensionalidade;
- Capacidade de lidar bem com dados ruidosos;
- Uma base matemática solidamente fundamentada.

Dado o conjunto de treino $T_r = \{d_j, c_i\}_{j=1}^{|T_r|} \subset D$, e o conjunto de teste $T_e = \{d_j, c_i\}_{j=1}^{|T_e|} \subset D$, tal que $d_j \in \mathbb{R}^{|T|}$, onde D é uma coleção de documentos, T é o conjunto de termos distintos da coleção D e $c_i \in \{1, -1\}$ sendo que o rótulo

1 indica um exemplo positivo e -1 indica um exemplo negativo. Cada documento da coleção D é representado por um ponto d_j no espaço euclidiano $\mathbb{R}^{|T|}$ e gerado de forma independente e identicamente distribuída em relação a uma probabilidade desconhecida $P_r(d_j, c_i)$ [49][27]. Todos os documentos são mapeados neste espaço euclidiano $|T|$-dimensional de acordo com sua representação no modelo espaço vetorial de forma que, a etapa de aprendizado do algoritmo busca encontrar um hiperplano que separe as duas classes, o qual possua a maior margem possível [35]. Novos documentos são mapeados neste mesmo espaço euclidiano de forma que são classificados em uma das categorias, baseando-se em qual dos lados do hiperplano (também conhecido como hiperplano de decisão) o novo documento foi mapeado, como pode ser visualizado na Figura 2.5, a qual ilustra a separação linear entre as classes.

Intuitivamente, a maximização da margem intenciona minimizar erros de classificação, visto que quanto mais próximo do hiperplano maior o grau de incerteza em relação à qual classe o documento pertence. Dessa forma, o classificador possui uma margem de segurança a qual garante que pequenos erros na classificação ou pequenas variações em características de documentos não irão causar classificações errôneas [35].

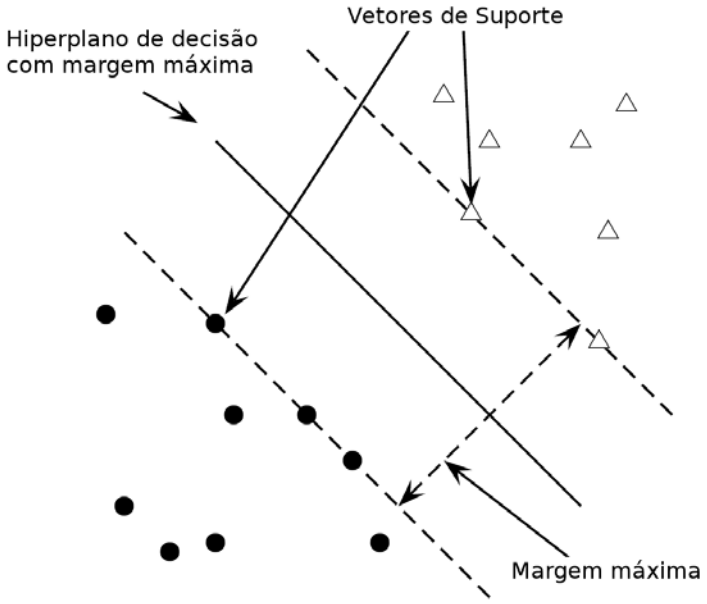

Figura 2.5: *Documentos de duas classes representados em um espaço euclidiano dividido pelo hiperplano de decisão com margem máxima.*

O hiperplano que separa as duas classes é descrito por:

$$(\overrightarrow{w} \cdot \overrightarrow{x}) + b = 0, \tag{2-15}$$

para a qual $\overrightarrow{w} \cdot \overrightarrow{x}$ é o produto escalar entre $\overrightarrow{w}$ e $\overrightarrow{x}$, sendo que $\overrightarrow{x}$ é um ponto arbitrário na Figura 2.5 o qual representa um documento a ser classificado, e o vetor $\overrightarrow{w}$ denominado vetor de peso (do inglês, *weight vector*) representa o vetor normal perpendicular ao hiperplano de decisão e ao ponto $\overrightarrow{x}$e o termo b possibilita deslocar o hiperplano paralelamente a este ponto.

Como mostrado na Figura 2.6, o parâmetro $\frac{b}{\|\overrightarrow{w}\|}$ determina o deslocamento (do inglês, *offset*) do hiperplano da origem ao longo do vetor normal $\overrightarrow{w}$. Deseja-se escolher $\overrightarrow{w}$ e b de modo a maximizar a margem o quanto for possível mantendo a separação das classes.

Objetivando determinar a qual categoria $c_i \in \{+1, -1\}$ pertence um determinado documento representado pelo vetor $\overrightarrow{x}$, é necessário verificar a sua posição relativa ao hiperplano através das restrições abaixo:

$$+1, \quad \text{se } (\overrightarrow{w} \cdot \overrightarrow{x}) + b \geq 0 \tag{2-16}$$

$$-1, \quad \text{se } (\overrightarrow{w} \cdot \overrightarrow{x}) + b < 0 \tag{2-17}$$

desta forma, a classificação de um vetor é alcançada aplicando-se a função de decisão expressa em 2-18:

$$f(\overrightarrow{x}) = sign((\overrightarrow{w} \cdot \overrightarrow{x}) + b) \tag{2-18}$$

O conjunto de documentos de treino que incidem nos hiperplanos marginais das classes são denominados como vetores de suporte (do inglês, *support vectors*). O hiperplano marginal $\overrightarrow{w} \cdot \overrightarrow{x} + b = 1$ é incindido pelos vetores de suporte pertencentes à classe especificada pela equação 2-16, sendo que a distância entre estes vetores de suporte até a origem do hiperplano de decisão é dada por $\frac{|1-b|}{\|\overrightarrow{w}\|}$.

De forma similar, o hiperplano marginal $\overrightarrow{w} \cdot \overrightarrow{x} + b = -1$ é incindido pelos vetores de suporte pertencentes à classe especificada pela equação 2-17, onde a distância entre estes vetores de suporte até a origem do hiperplano de decisão é dada por $\frac{|-1-b|}{\|\overrightarrow{w}\|}$. como mostrado na Figura 2.6.

Dessa forma, a distância entre os vetores de suporte e a origem do hiperplano de decisão é estabelecida por $\frac{1}{\|\overrightarrow{w}\|}$, e portanto a distância entre os vetores de suporte das duas classes é dado por $\frac{2}{\|\overrightarrow{w}\|}$

A maximização da margem é alcançada resolvendo-se um problema de otimização quadrática, em termos dos vetores de suporte na forma de:

$$\overrightarrow{w} = \sum_i v_i \overrightarrow{x_i} \tag{2-19}$$

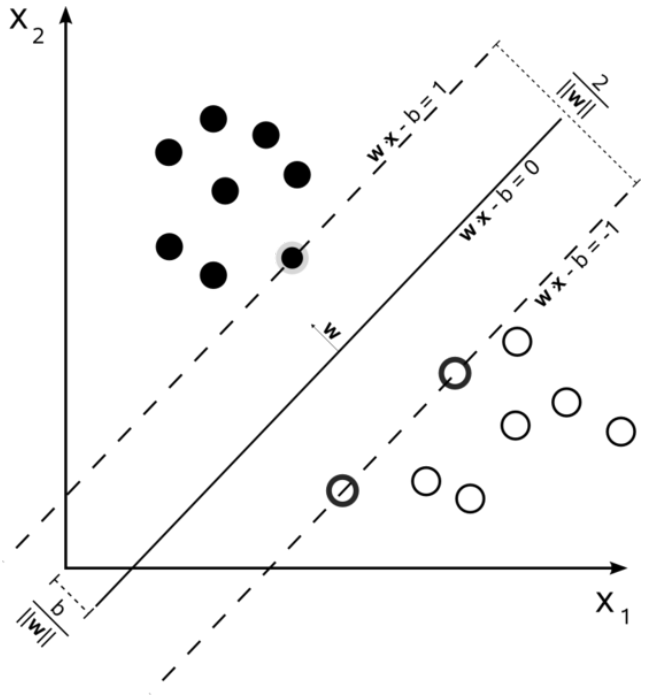

Figura 2.6: *Distância entre os dois hiperplanos marginais de classe.*

na qual cada v_i representa um parâmetro aprendido e cada x_i é um vetor de suporte. A função de decisão pode, então, ser escrita como:

$$f(\overrightarrow{x}) = sign(\sum_i v_i(\overrightarrow{x_i} \cdot \overrightarrow{x}) + b) \tag{2-20}$$

No espaço de dados original, também conhecido como espaço de entrada, pode ocorrer de as classes não serem separáveis por meio de um hiperplano. Entretanto, os vetores de dados originais (os quais representam os documentos no modelo espaço vetorial) podem ser mapeados em um espaço de alta dimensionalidade, chamado de espaço de características, ao invés do espaço de entrada. A função de decisão final é dada por:

$$f(\overrightarrow{x}) = sign(\sum_i v_i k(\overrightarrow{x_i} \cdot \overrightarrow{x}) + b) \tag{2-21}$$

onde k é a função núcleo.

O método SVM somente realiza classificação binária. Assim, para a classificação multi-classe é necessário utilizar a metodologia 1 Contra Todos [61], descrita na Seção 2.4.2. Neste trabalho, utilizamos o classificador SVM disponibilizado pelo pacote SVM^{perf} descrito por Joachims em [26] por meio da metodologia supracitada.

2.4.4 Avaliação de Classificação

De acordo com Sebastiani [52], visando a avaliar o desempenho de métodos de CAT, são observados dois aspectos, sendo estes a eficácia e a eficiência do classificador automático. A eficácia é uma medida que avalia a habilidade de um classificador automático decidir corretamente a categoria (ou classe) de determinado documento. A eficiência, por sua vez, geralmente é uma medida que avalia o tempo gasto por um classificador automático para decidir a categoria de determinado documento.

Com a intenção de verificar a eficácia de um classificador, normalmente são utilizadas medidas que podem ser compreendidas por meio de uma *tabela de contingência*. Através da Tabela de contingência 2.1 adaptada de [52], dada uma categoria c_i qualquer, é possível visualizar as possibilidades de resposta de um classificador (Φ) ao decidir sobre os documentos da coleção de teste, comparando tais decisões com o que seria esperado como resposta correta de acordo com o julgamento (Ψ) previamente atribuído por especialistas do domínio da classe c_i.

Categoria c_i		Julgamentos Corretos	
		$+1$	-1
Julgamentos do Classificador	$+1$	TP_i	FP_i
	-1	FN_i	TN_i

Tabela 2.1: *Tabela de Contingência para a classificação dos documentos de teste para a classe c_i.*

Nesta abordagem:

- TP_i dito como Verdadeiros Positivos (do inglês, *True Positives*) da classe c_i, representa o número de documentos **corretamente classificados** na categoria c_i;
- TN_i dito como Verdadeiros Negativos (do inglês, *True Negatives*) da classe c_i, representa o número de documentos **corretamente não classificados** na categoria c_i;
- FP_i dito como Falsos Positivos da classe c_i, representa o número de documentos **incorretamente classificados** na categoria c_i;
- FN_i dito como Falsos Negativos da classe c_i, representa o número de documentos **incorretamente não classificados** na categoria c_i;

Utilizam-se os dados da Tabela 2.1 para calcular a eficácia do classificador utilizando as métricas de precisão p (do inglês, *precision*) e cobertura ou revocação r (do inglês, *recall*).

Medidas de Precisão e Cobertura

A partir da medida de precisão é possível estabelecer a proporção entre os documentos que foram classificados corretamente como sendo da classe c_i (TP_i) com relação a todos os documentos classificados como sendo da classe c_i ($TP_i + FP_i$), como mostrado na equação 2-22:

$$p_i = \frac{TP_i}{TP_i + FP_i} \tag{2-22}$$

A medida de cobertura ou revocação, por sua vez, estabelece a proporção entre os documentos que foram classificados corretamente como sendo da classe c_i (TP_i) com relação a todos os documentos que deveriam ter sido classificados como sendo da classe c_i ($TP_i + FN_i$), como mostrado na equação 2-23:

$$r_i = \frac{TP_i}{TP_i + FN_i} \tag{2-23}$$

A Figura 2.7 mostra a representação gráfica destas duas medidas.

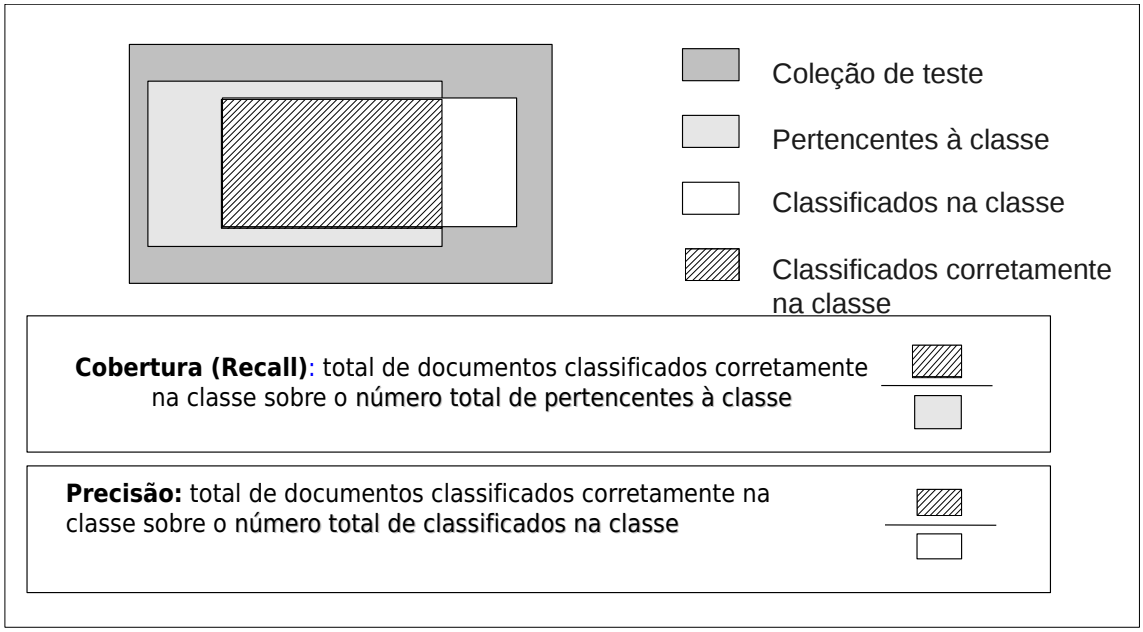

Figura 2.7: *Representação gráfica das medidas de precisão e cobertura.*

Métrica-F

A métrica-F (do inglês, *F-measure*), de acordo com Yang e Liu[68], combina os valores de precisão e cobertura criando uma medida única, como definido na

Equação 2-24:

$$F_\alpha(c) = \frac{(\alpha^2 + 1)pr}{\alpha^2 p + r} \tag{2-24}$$

na qual α define a importância relativa da precisão p e cobertura r. Quando $\alpha = 0$, somente a precisão é considerada. Quando $\alpha = \infty$, somente a cobertura é considerada. Quando $\alpha = 0.5$ a cobertura possui a metade da importância da precisão, e assim por diante.

A métrica-F utilizada no presente trabalho é a métrica-F_1 a qual é obtida atribuindo pesos iguais para precisão e cobertura. Para isso, α é definido com valor 1. A equação 2-25 mostra o cálculo da métrica-F_1

$$F_1(c) = \frac{2pr}{p + r} \tag{2-25}$$

A F_1 considera o desempenho de um classificador em relação a apenas uma categoria. Para considerar todas as categorias, um único valor para F_1 pode ser derivado. Também é comum derivar a métrica F_1 de modo a avaliar o desempenho geral do classificador por meio do cálculo da média das F_1 calculadas para cada classe. Duas médias são normalmente utilizadas com esse propósito: média $microF_1$ e média $macroF_1$ [68].

O cálculo da média $microF_1$ é realizado levando em conta o valor global de precisão e cobertura, por meio do somatório das variáveis TP, TN, FP, FN de todas as classes. Dessa forma, é possível obter o valor global de precisão e cobertura. A precisão global p_g é calculada por meio da equação 2-26:

$$p_g = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FP_i)} \tag{2-26}$$

Por sua vez a cobertura global r_g é calcula como:

$$r_g = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FN_i)} \tag{2-27}$$

Nesse sentido, o cálculo de $microF_1$ é dado por:

$$microF_1 = \frac{2p_g r_g}{p_g + r_g} \tag{2-28}$$

Dessa forma, a média $microF_1$ pondera as medidas F_1 de cada classe com base na representatividade da classe na coleção de acordo com o número de documentos em cada classe.

O cálculo da média $macroF_1$ é realizado a partir dos resultados de F_{1i}, o qual representa o valor de F_1 para cada categoria c_i da coleção. O cálculo desta métrica é dado por:

$$macroF_1 = \frac{\sum_{i=1}^{|C|} F_{1i}}{|C|} \tag{2-29}$$

É importante notar que a média $macroF_1$ parte do princípio de que cada classe possui importância igual na coleção. Por este motivo, atribui-lhes pesos iguais, independente da quantidade de documentos contidos em cada uma. Já a média $microF_1$ estabelece que cada documento é igualmente importante, porém em coleções onde se observa uma distribuição de documentos em classes de forma muito irregular, observa-se que a métrica privilegia classes maiores.

A partir das diferentes abordagens estabelecidas pelas duas métricas, pode-se perceber que se a maioria das classes em uma coleção contiver proporcionalmente poucos documentos em relação ao todo, então a média $macroF_1$ possui uma maior relevância, pois são raros os casos em que é adequado subestimar a importância de uma vasta densidade de classes. Caso contrário, a média $microF_1$ é uma métrica tipicamente mais significativa.

Método de Validação Cruzada

A obtenção de bons classificadores depende em grande parte da escolha do conjunto de treino e teste a ser utilizada. Para tanto, algumas coleções de dados utilizadas para se avaliar classificadores textuais apresentam divisões padrão entre treino e teste visando a tornar comparáreis os experimentos realizados nestas coleções. Entretanto, para coleções que não possuem uma divisão padrão e até mesmo para as que a possuem, pode-se utilizar o método de validação cruzada para avaliar o desempenho de um classificador ao ser aplicado a uma coleção de dados.

A validação cruzada tem se tornado um método padrão para a avaliação de classificação de documentos [52] [38].

- **Validação cruzada com k partições** (do inglês, *k-fold cross validation*): Este método consiste em construir k diferentes classificadores: $\Phi_1, \Phi_2, \ldots, \Phi_k$ a partir da divisão do corpus inicial Ω apresentado na Seção 2.4.1, em k conjuntos disjuntos: $Te_1, Te_2, \ldots, Te_k$ com aproximadamente $\frac{|\Omega|}{k}$ documentos em cada conjunto. Cada classificador Φ_i é treinado usando $\Omega - Te_i$ e avaliado utilizando o conjunto de teste Te_i. Cada classificador é avaliado usualmente

utilizando as medidas de precisão, cobertura e F_1, e finalmente a avaliação geral é dada pela média das k avaliações realizadas. O valor mais utilizado de k tem sido 10, o qual é denominado como validação cruzada com 10 partições *10-fold cross validation*.

- **Validação cruzada estratificada com k partições** (do inglês, *Stratified k-Fold Cross Validation*): Este método é similar ao anterior, sendo que ao dividir a coleção de documentos Ω em k conjuntos, a proporção de documentos em cada uma das categorias é considerada na constituição dos conjuntos. Neste contexto, verifica-se o número de documentos de cada categoria com relação ao total de documentos da coleção. Cada partição k deve ser composta respeitando esta mesma proporção de distribuição de categorias entre os documentos que compõem a partição.

A partir dos métodos de validação cruzada é possível verificar o comportamento do classificador para cada partição utilizada. O presente trabalho utiliza partições fixas para as coleções Reuters e Ohsumed, e utiliza a validação cruzada para a coleção 20Newsgroups, como será abordado na Seção 4.2.

2.5 Trabalhos Relacionados

Nesta seção serão apresentados alguns trabalhos relacionados ao enriquecimento da representação de documentos buscando aumentar o desempenho do processo de classificação automática de documentos.

Vários trabalhos na área de classificação de documentos propuseram melhorias frente ao modelo de conjunto de palavras (BOW) tradicional, com a finalidade de obter maior eficácia na construção de classificadores. Com este intuito, a expansão de características se mostrou, em diversos trabalhos, muito propensa a ajudar neste processo. Nas próximas linhas serão apresentados os trabalhos que enriquecem a representação BOW, seja com informações extraídas de dentro da própria coleção ou provindas de fontes externas.

Mladenic e Grobelnik [41], utilizaram o aprendizado de máquinas no processo de enriquecimento do BOW com n-gramas de comprimento até 3 (também chamados de 3-gramas ou trigramas) identificados dentro do próprio documento. Os autores constataram que o uso de n-gramas pode melhorar a eficácia da classificação automática de documentos. Ganhos mais acentuados provindos do uso destes elementos foram obtidos a partir de 2-gramas (também chamados de bigramas). No referido trabalho, os autores reportam que n-gramas maiores que 3 não se mostraram úteis na melhoria da classificação, devido principalmente à quantidade de n-gramas pouco relevantes ao processo de classificação.

Em seu trabalho, Fürnkranz e Grobelnik[17] utilizaram o enriquecimento do BOW adicionando n-gramas a esta representação a partir da análise de palavras consecutivas que compõem a matriz documento-termo dos documentos, conseguindo melhores resultados quando comparado aos obtidos por meio da representação de conjunto de palavras. Entretanto, sequências de comprimento maior que 3 (trigramas) não possibilitaram melhorias nos resultados, de modo que em alguns casos esta metodologia gerou inclusive a degradação do processo de classificação. Naquele trabalho, a frequência da ocorrência dos n-gramas dentro dos documentos também foi considerada, sob a argumentação de que essa informação tenderia a melhorar os resultados.

Apesar dos bons resultados reportados pelos dois trabalhos acima relacionados, os autores não utilizam nenhum tipo de medida de identificação de n-gramas mais discriminativos, bem como nenhuma técnica de seleção de característica ou filtragem de n-gramas ruidosos ou pouco discriminativos.

Gabrilovich et al. [19] propuseram um método de utilização de informações provindas de enciclopédias na melhoria dos sistemas de classificação de textos. Neste trabalho os autores utilizaram especificamente a enciclopédia Wikipédia e o ODP (do inglês, *Open Directory Project*), um serviço de diretório aberto de categorização de conteúdo Web. Primeiramente, construiu-se um classificador de texto de cunho auxiliar por meio do qual relacionam-se os documentos a serem classificados com o conjunto de artigos mais relevantes da Wikipédia a fim de encontrar similaridade textual entre os dois elementos. Após isto, enriqueceu-se a representação convencional BOW com novas características, as quais correspondem a conceitos, em sua maioria títulos dos artigos. Os resultados empíricos mostraram que esta abordagem conseguiu melhorar a eficácia do processo de classificação de documentos em diversas coleções de dados como Reuters-21578, RVC1, 20NG e *Movie Reviews*. Entretanto os autores não fizeram uso de todos os ricos relacionamentos existentes na Wikipédia, tais como as relações de hiponímia e sinonímia.

Wang et al.[63] construíram um tesauro informativo com dados extraídos da Wikipédia, por meio do qual explicitamente derivaram relações de sinonímia, polissemia, hiponímia e relações associativas entre conceitos desta enciclopédia. Este tesauro foi utilizado para introduzir informações semânticas nos documentos, mostrando-se com um poder de cobertura muito mais amplo do que qualquer tesauro construído manualmente, como no caso do WordNet[3]. As relações de sinonímia, polissemia e hiponímia foram extraídas da forma mostrada na Seção 2.3, a partir de

[3]WordNet corresponde a um grande banco de dados léxico para a língua inglesa. Substantivos, verbos, adjetivos e advérbios são agrupados em conjuntos de sinônimos cognitivos denominados synsets, cada um expressando um conceito distinto, fazendo o papel de dicionário e tesauro.

uma varredura nos documentos da coleção a ser classificada de modo a encontrar conceitos da Wikipédia presentes nestes. Entretanto, no referido trabalho não se conseguiu utilizar de forma satisfatória o enriquecimento de documentos por meio de sinônimos derivados da Wikipédia. Por outro lado o enriquecimento de documentos com relações de hiponímia gerou bons resultados. Como justificativa para o baixo desempenho da utilização dos sinônimos, os autores reportam o excesso de ruídos provindos deste tipo de característica assim como a impossibilidade de filtrar os conceitos sinônimos de baixa qualidade. Os autores não reportam a utilização de nenhuma medida de seleção de características a qual poderia selecionar um conjunto de sinônimos mais discriminativos e menos ruidosos para o processo de classificação de documentos. Ademais, Wang et al. [63] apenas relatam a necessidade de melhoria no método de adição de sinônimos à representação BOW de modo que seja minimizada a inserção de ruídos.

Em seu trabalho, Figueiredo et al.[14] utilizam o critério de *Predominância* como medida de seleção de características visando a estimar a pertinência de um documento ser expandido por um novo termo. A partir desta medida, os autores quantificam a probabilidade global de um termo candidato estar exclusivamente associado a uma classe. Visando a garantir um bom poder de generalização do método, o referido trabalho utiliza ainda um valor fixo mínimo de ocorrência de termos em documentos do conjunto de treino da coleção, referenciado por [9] como suporte mínimo. Entretanto, o suporte mínimo utilizado não garante um número de ocorrências mínimas em documentos dentro de uma classe. Da mesma forma, o suporte mínimo não considera a quantidade de documentos de treino na classe, utilizando o mesmo valor para classes muito pequenas, com 4 documentos, por exemplo, e classes muito grandes com mais 2000 documentos, por exemplo. Por meio da medida de Predominância, Figueiredo et al.[14] selecionam duplas de termos não necessariamente adjacentes (referenciado por eles como c-termos), que possuem valor acima de um limiar mínimo nesta medida. Ademais, o referido trabalho também impõe a seguinte restrição adicional para utilização dos c-termos: Um determinado c-termo só será utilizado na expansão de um documento d_k se tal c-termo obtiver valor de Predominância mínima para classe c_j e d_k esteja entre o conjunto de documentos pertencentes a c_j. A partir da restrição de classe imposta por [14], e dos bons resultados apresentados pelos referidos autores, o presente trabalho utiliza a referida abordagem de restrição de classe (CRC), comparando os resultados obtidos sem a utilização desta restrição(SRC).

O Capítulo 3 apresenta a metodologia proposta pelo presente trabalho para a expansão de características por meio da utilização da coocorrência de n-gramas nos documentos a serem classificados e na Wikipédia na forma de sinônimos e/ou

categorias, propondo também uma medida de importância para estes de modo a minimizar a inserção de ruídos durante o processo de expansão.

CAPÍTULO 3

Uso da Wikipédia para Expansão de Características

Neste capítulo apresentam-se as abordagens propostas pelo corrente trabalho para a extração de características da enciclopédia Wikipédia e seu uso na CAT. Ademais, também são discutidos alguns detalhes pertinentes à implementação dessas abordagens. A Seção 3.1 apresenta a abordagem proposta para extração de características da Wikipédia que também ocorrem nos documentos a serem classificados. São explicitados os pré-processamentos necessários, a indexação da enciclopédia e os algoritmos de identificação de conceitos presentes em documentos. A Seção 3.2 apresenta a medida de seleção de característica FT1C, proposta por esta pesquisa. A Seção 3.3 apresenta a metodologia de utilização da medida FT1C, assim como explana sobre as metodologias de expansão de característica SRC e CRC. A Seção 3.4 apresenta a metodologia de utilização das categorias da Wikipédia em CAT.

3.1 Extração de termos-chaves da Wikipédia

Nesta seção é descrita a abordagem utilizada neste trabalho para a extração de características da Wikipédia visando a melhorar o enriquecimento de coleções Uni-rótulo [1] (do inglês, *uni-label*). As características extraídas da Wikipédia são utilizadas para expandir a representação BOW de documentos das coleções a serem classificadas, com o objetivo de se obter uma melhoria na eficácia da classificação de documentos dessas coleções.

Para o entendimento do método proposto é interessante estabelecer uma comparação com a abordagem comumente adotada para classificação de textos, utilizando aprendizado de máquina, como pode ser visto na Figura 3.1.

No modelo ilustrado na Figura 3.1, os documentos de treino e de teste são representados através do modelo VSM, como descrito na Seção 2.1. Neste modelo,

[1] Coleções em que cada documento pertence a somente uma única categoria.

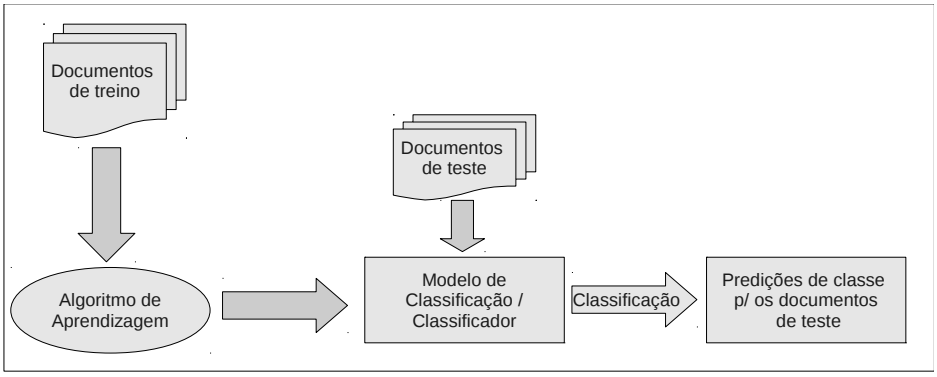

Figura 3.1: *Modelo tradicional de classificação de textos baseado em aprendizado de máquina.*

os termos são representados apenas por meio do conjunto de palavras que ocorrem diretamente nos documentos, conhecido como BOW. Como visto na Subseção 2.1.1, a representação BOW não consegue estabelecer relações semânticas entre termos, o que limita o desempenho dos classificadores.

Em nosso trabalho, como ilustrado pela Figura 3.2, ampliamos a matriz BOW, que no método tradicional contém apenas palavras da própria coleção a ser classificada, adicionando conceitos e/ou categorias extraídas da Wikipédia, tanto nos documentos de treino quanto nos documentos de teste. Os documentos são expandidos com as características da Wikipédia relacionadas a conceitos desta enciclopédia que ocorrem nesses documentos. Como pode ser visualizado na Figura 3.2, várias tarefas devem ser realizadas durante a metodologia proposta, as quais serão abordadas adiante.

3.1.1 Pré-processamento da Wikipédia

O primeiro passo da metodologia é o pré-processando dos dados da Wikipédia visando à extração de características úteis ao processo de classificação. A partir da versão em XML da Wikipédia são extraídos todos os títulos dos conceitos (artigos) que compõem esta enciclopédia. Dentre estes nem todos realmente representam conceitos que podem ser úteis na CAT, e por esses motivos títulos de artigos que apresentam alguns padrões são descartados, tais como “list of”, “th century”, “(decade)”, nomes de meses do ano, datas, anos, títulos compostos apenas por *stop words*, sobre letras do alfabeto, somente números, dentre outros. No restante do texto, os títulos que foram mantidos após a filtragem são denominados como *w-conceitos*. São extraídos também os títulos dos conceitos de redirecionamento, os quais denotam sinonímia para com os w-conceitos, como informado na Seção 2.3. Os títulos de redirecionamento que apontam para conceitos descartados no procedimento de filtragem também são descartados.

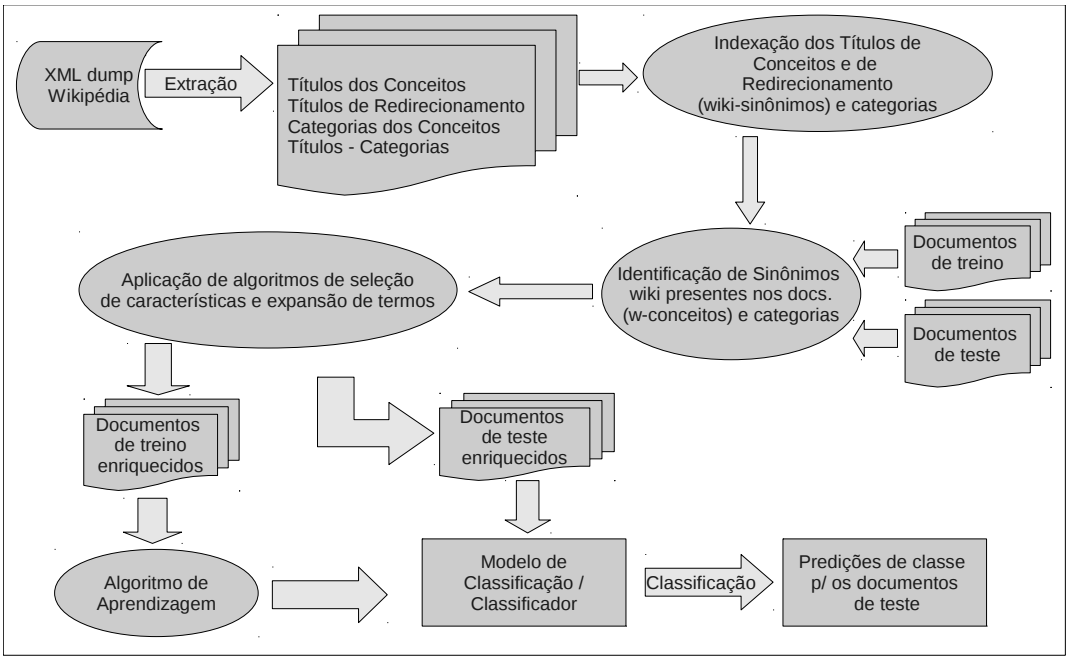

Figura 3.2: *Modelo de abordagem proposto para a classificação de textos baseado em aprendizado de máquina.*

Por meio deste processo também são extraídos todos os títulos das categorias presentes nos conceitos da Wikipédia. Este procedimento é realizado fazendo-se uso dos *links* de categorias presentes nos artigos relacionados a estes w-conceitos. Neste trabalho, utilizou-se somente as categorias de nível mais baixo na hierarquia de categorias, isto é, utilizou-se somente as categorias que aparecem explicitamente ligadas a um documento pela *tag* **Category:** que aparece em cada documento. Como cada categoria também é um conceito da Wikipédia, uma categoria pode pertencer a outra categoria, formando níveis hierárquicos de categorias. A investigação de expansão com características com outros níveis mais altos na hierarquia é estabelecida como trabalho futuro.

3.1.2 Grupos de conceitos sinônimos da Wikipédia

Após o pré-processamento é feito um agrupamento por sinonímia de todos os w-conceitos e dos títulos de documentos de redirecionamento que foram extraídos da Wikipédia. Cada grupo é formado por um w-conceito e por todos os conceitos de redirecionamento que possuem ligação para o mesmo por meio de um *link* de redirecionamento. Um identificador único w_j é associado ao w-conceito representante de cada grupo g_j. Esta metodologia possibilita captar a noção de sinonímia empregada pelos redirecionamentos. Cada componente de um grupo é denominado *wiki-sinônimo*, inclusive o w-conceito que representa o grupo.

Um índice posicional φ é criado para permitir consultas por frase no agrupamento. O índice permite verificar se uma dada frase corresponde a um wiki-sinônimo e, em caso afirmativo, retorna o identificador w_j correspondente ao grupo de sinônimos ao qual a frase pertence. Para tanto, uma função f de consulta ao índice é utilizada, a qual recebe como parâmetros o índice φ e a frase q a ser consultada. Frases formadas apenas por *stop-words* são desconsideradas. A função f é definida como:

$$f(\varphi, q_i) = \begin{cases} w_j, & \text{se } q_i \text{ corresponde a um wiki-sinônimo em } \varphi \\ 0, & \text{caso contrário} \end{cases} \quad (3\text{-}1)$$

A função f e o índice posicional φ são utilizados para verificar se um wiki-sinônimo ocorre em um documento de uma coleção de textos a ser classificada, conforme explica a Seção 3.1.3.

3.1.3 Identificação dos w-conceitos em textos da coleção a ser classificada

A identificação da presença de w-conceitos é realizada tanto nos documentos de treino quanto nos documentos de teste da coleção a ser classificada. Para realizar este processo, utiliza-se uma adaptação do método utilizado por [64], por meio de um algoritmo de janelas de termos que é aplicado aos textos dos documentos da coleção. Os w-conceitos identificados neste processo são chamados de w-conceitos candidatos.

O primeiro passo da identificação de w-conceitos é a divisão de um documento em vários trechos. Cada trecho S é composto por um conjunto de palavras que serão objeto de constituição de n-gramas visando a encontrar wiki-sinônimos. Estes trechos são delimitados por meio das ocorrências de caracteres de pontuações e símbolos, com exceção de apóstrofos, palavras com hífen, e abreviações que utilizam ponto, como U.S., por exemplo. Este aspecto visa a impossibilitar a formação de n-gramas a partir de palavras pertencentes a trechos semânticos diferentes, aumentando a eficácia e a performance na busca por wiki-sinônimos. Os trechos derivados deste processo e que são compostos somente por números são descartados.

A Tabela 3.1 mostra um exemplo em que parte do documento 0003908 da Coleção Ohsumed é dividido em trechos, conforme a explicação acima.

Uma vez obtidos os trechos, o algoritmo de janelas de termos, descrito a seguir, é aplicado a cada trecho, o qual é representado como um vetor de termos. O algoritmo de janelas funciona do seguinte modo: seja t um dado trecho de um

The presence of chlamydial deoxyribonucleic acid (dna) was evaluated by dna hybridization in endocervical cells of infertile and normal fertile women. chlamydial dna was detected in 49 of 186 (26.3%) infertile patients, which is significantly more common than in fertile control individuals (12.5%, or 8 of 64 individuals).	
trecho-1	The presence of chlamydial deoxyribonucleic acid
trecho-2	dna
trecho-3	was evaluated by dna hybridization in endocervical cells of infertile and normal fertile women
trecho-4	Chlamydial dna was detected in 49 of 186
trecho-5	infertile patients
trecho-6	which is significantly more common than in fertile control individuals
trecho-7	or 8 of 64 individuals

Tabela 3.1: *Exemplo de divisão do texto em trechos.*

documento d, uma janela de tamanho $n = N_{max}$ é posicionada mais à esquerda sobre o trecho t, se t tiver tamanho igual ou superior a n. Caso o tamanho de t seja menor do que n, o algoritmo utiliza um tamanho de janela n igual ao tamanho de t. A sequência q de n termos que aparece na janela é consultada no índice posicional φ, utilizando-se a função $f(\varphi, q)$ definida na Seção 3.1.2. Se $f(\varphi, q) = w_j$, para algum $w_j \neq 0$, o valor de w_j é inserido no conjunto de w-conceitos candidatos à expansão de d. Neste caso, a janela é deslocada à direita n posições em t, e o algoritmo busca a próxima janela com tamanho $n = N_{max}$, com o objetivo de procurar um novo w-conceito. Se $f(\varphi, q) = 0$ e $n > N_{min}$, então nesse caso, um w-conceito não foi encontrado e o tamanho da janela é diminuído de um, excluindo-se da mesma o termo mais a sua direita. Se, entretanto, $f(\varphi, q) = 0$ e $n = N_{min}$, a janela é deslocada à direita n posições e o tamanho n e o algoritmo busca a próxima janela com tamanho $n = N_{max}$. Esse processo continua até que não seja mais possível deslocar a janela à direita em t.

O algoritmo é aplicado a todos os trechos de um documento d, gerando um conjunto de w-conceitos candidatos a expandir d. Neste trabalho, utilizou-se $N_{min} = 1$ e $N_{max} = 4$, conseguindo cobrir $87,8\%$ dos wiki-sinônimos da Wikipédia indexados neste trabalho. Wiki-sinônimos maiores que 4 são muito raros, podendo gerar degradação dos resultados [41].

A Tabela 3.2 mostra o resultado da aplicação do algoritmo de janelas aos trechos obtidos na Tabela 3.1. Por uma questão de melhor visualização, são mostrados os títulos dos w-conceitos encontrados e não os identificadores dos

mesmos, retornados pelo algoritmo de janelas.

Trechos	wiki-sinônimos	w-conceitos
trecho-1	deoxyribonucleic acid	dna
trecho-2	dna	dna
trecho-3	dna hybridization	nucleic acid thermodynamics
	endocervical	canal of the cervix
	cells	cell
	infertile	infertility
	fertile	fertility
	women	woman
trecho-4	chlamydial	chlamydia
	dna	dna
trecho-5	infertile	infertility
	patients	patient
trecho-6	fertile	fertility
	individuals	individual
trecho-7	individuals	individual

Tabela 3.2: *Relação entre wiki-sinônimos e w-conceitos extraídos dos trechos da Tabela 3.1.*

Por meio da Tabela 3.2, é possível visualizar algumas características importantes na busca de wiki-sinônimos em documentos. Ambas as expressões *deoxyribonucleic acid* no trecho-1 e *dna* no trecho-2 são wiki-sinônimos pertencentes ao mesmo grupo do w-conceito *dna*. Dessa forma, mesmo que a primeira expressão ocorra em um documento e a segunda ocorra em outro, ambas serão tratadas como sendo o w-conceito *dna*, associando tais documentos. Também pode ser visualizado nos demais trechos que o método auxilia na associação entre termos que estão no plural com seus equivalentes no singular, assim como variações léxicas como no caso de *fertile* e *fertility*.

O método de uso de w-conceitos possui grande possibilidade de auxiliar a CAT, porém, ainda é necessário uma filtragem de w-conceitos candidatos que não são bons discriminadores das classes na coleção a ser expandida. Na Seção 3.2, será visto a descrição desta etapa do processo.

3.2 Filtragem de w-conceitos não discriminativos

Após a identificação dos w-conceitos candidatos, uma parcela destes não apresenta-se como bons discriminadores dos documentos nos quais aparecem. A utilização destes w-conceitos em adição com a representação BOW do documento, pode gerar classificadores com baixa eficácia. Dessa forma, faz-se necessário um método que impeça a inserção destes w-conceitos no documento que será expandido.

Em seu trabalho, Wang et. al.[64] também utiliza w-conceitos (referenciado apenas como sinônimos). No entanto, tais elementos não contribuíram com o processo de classificação ao serem utilizados na expansão de documentos. O referido autor reporta em seus resultados que a utilização de w-conceitos prejudica a classificação de documentos devido à inserção de termos ruidosos durante este processo. Todavia, o autor não executa nenhum tipo de seleção de característica (do inglês, *feature selection*) a fim de inserir apenas termos com bom potencial para serem discriminadores de categorias.

Neste contexto, o presente trabalho busca utilizar um método de *feature selection* utilizando uma função de avaliação de termos (como descrito na secção 2.1.6 do Capítulo 2) capaz de selecionar apenas w-conceitos bons discriminadores de classes ao mesmo tempo que descarta w-conceitos que não são bons discriminadores. Tal método é aplicado apenas aos documentos de treino da coleção de forma que os w-conceitos aprovados nesta fase são definidos como w-conceitos eleitos.

Como exposto na Seção 2.1.6, as medidas de avaliação de termos são fundamentais no processo de filtragem de características, visto que representam critérios a serem seguidos durante o processo de seleção de termos.

3.2.1 Fator de Tendência a uma categoria - FT1C

O presente trabalho propõe uma função de avaliação de termos para o problema de seleção de w-conceitos candidatos, a qual se adeque bem às variações quantitativas das classes, gerando maior capacidade de adaptação às diversas coleções. Por meio desta função de avaliação, tenta-se garantir que os documentos sejam enriquecidos com w-conceitos que, além de tenderem a apenas uma classe, possuam também uma abundância relativa suficiente dentro da mesma classe.

Seja D uma coleção de documentos particionada em dois conjuntos: D_{tr} o conjunto de treinamento, para o qual se conhece a classe de cada documento $d \in D_{tr}$, e D_{te}, o conjunto de documentos de teste, em que seus documentos não são classificados. Pretende-se estabelecer uma medida que represente a predominância de um w-conceito dentro da categoria do documento no qual ocorre. Esta medida é provida pelo conceito de *Predominância-Local*, por meio do qual w-conceitos muito raros dentro da classe podem ser filtrados. Seja $T = \{t_1, t_2, \ldots, t_M\}$ o conjunto de tamanho M formado por w-conceitos candidatos a enriquecer os documentos da coleção D. O conjunto T é obtido pelo processo de extração de características descrito na Seção 3.1.3. Seja $C = \{c_1, c_2, \ldots, c_K\}$ o conjunto de K classes da coleção D. Cada documento d de D_{tr} pertence a uma única classe $c_i \in C$. A *Predominância-*

Local é formalizada na Equação 3-2:

$$P_{local}(t_i, c_j) = \frac{df(t_i, c_j)}{td(c_j)} \tag{3-2}$$

sendo $df(t_i, c_j)$ o número de documentos que o w-conceito t_i é candidato a enriquecer dentro da classe c_j, e $td(c_j)$ o total de documentos contidos na classe c_j. Por conseguinte, dado um w-conceito t_i e uma categoria c_j, a *Predominância Local* mede qual a probabilidade de um documento x de c_j conter t_i. O método se adapta bem tanto em classes pequenas quanto em grandes classes, visto que para um mesmo valor de *Predominância-Local*, uma classe composta por uma maior quantidade de elementos deve possuir mais documentos contendo o w-conceito t_i do que uma classe menor.

Entretanto, somente o valor da *Predominância-Local* não é capaz de determinar se um w-conceito é realmente bom discriminador de uma classe, apenas que o mesmo é abundante na referida classe. Como visto na Seção 2.1.4, um determinado w-conceito que é abundante em várias categorias, e portanto ocorre em vários documentos, tem seu poder de discriminação diminuído, o que de fato a medida de *Predominância-Local* não consegue representar.

Diante deste contexto, propõe-se nesta seção reduzir o valor da *Predominância-Local* $P_{local}(t_i, c_j)$ de um w-conceito t_i em uma classe c_j, ao se deduzir deste valor a soma dos valores de *Predominância-Local* alcançados por t_i nas classes restantes da coleção D. Esta abordagem é definida como *Fator de tendência a uma categoria* ($FT1C$) a qual é dada pela equação 3-3:

$$FT1C(t_i, c_j) = \frac{df(t_i, c_j)}{td(c_j)} - \sum_{m=1}^{|C|} \frac{df(t_i, c_m)}{td(c_m)} \; \forall \; c_m \in C \text{ e } c_m \neq c_j \tag{3-3}$$

Por meio desta equação, é possível perceber que cada valor de *Predominância Local* obtido da ocorrência do termo t_i em categorias diferentes de c_j é utilizado como um peso de depreciação de $FT1C(t_i, c_j)$. Quanto mais categorias diferentes t_i ocorrer e quanto maiores os valores de *Predominância-Local* que t_i obtiver nas demais categorias de C, menor será o Fator de tendência a uma classe - $FT1C$ de t_i em c_j. Por outro lado, quanto maior for a *Predominância Local* em c_j e menos categorias diferentes de c_j que t_i ocorre e quanto menor o valor da *Predominância Local* de t_i nestas ocorrências, maior será o fator $FT1C$ na categoria c_j.

3.3 Enriquecimento da coleção a partir de w-conceitos eleitos

As medidas *Information Gain*, *Gain Ratio*, *Chi-squared* e $FT1C$ são utilizadas alternadamente como função f de avaliação no processo de seleção de características (w-conceitos e/ou categorias). O próximo passo é a escolha das k características mais bem avaliadas pela função f as quais serão formadoras do conjunto T' de w-conceitos eleitos. Todas as medidas podem gerar até $|C|$ valores diferentes para cada w-conceito t_i, visto que todas as medidas relacionadas acima indicam a importância de um termo t_i para uma classe c_j. Diante do exposto, utiliza-se a equação 2-14 para extrair o valor global $f_{global}(t_i)$ a partir do maior valor obtido pela função local de avaliação $f(t_i, c_j)$. Os k w-conceitos que possuem o maior valor $f_{global}(t_i)$, integram o conjunto T' de w-conceitos eleitos, independentemente da categoria por meio da qual o valor máximo foi obtido.

Um w-conceito t_n eleito para compor T', mesmo sendo muito importante para caracterizar uma classe c_j pode ocorrer também em uma classe c_m para a qual o valor de $f(t_n, c_m)$ seja menor que o valor de qualquer um dos k elementos de T'. A partir do exposto, seja D_m o conjunto de documentos de treino associados à categoria c_m, se um w-conceito eleito t_n ocorre em um documento de treino $d \in D_m$, criam-se duas abordagens de inserção de w-conceitos eleitos no conjunto de treino:

- **SRC - Sem Restrição de Classe:** O w-conceito eleito $t_n \in T'$ será utilizado no enriquecimento de qualquer documento em que o mesmo ocorra, inclusive nos documentos das classes para as quais $f(t_n, c_m)$ é menor que o valor de $f_{global}(t_n)$.
- **CRC - Com Restrição de Classe:** O w-conceito eleito $t_n \in T'$ só será utilizado no enriquecimento de um documento d em c_m se $f(t_n, c_m)$ obtiver valor maior ou igual ao menor valor f_{global} dentre os elementos de T'.

Neste estudo foram comparados os desempenhos das medidas *FT1C*, *Information Gain*, *Gain Ratio* e *Chi-squared*, sendo que para cada uma destas, comparou-se o desempenho das duas abordagens SRC e CRC de inserção de w-conceitos eleitos no conjunto de treino.

No conjunto de teste, como não se sabe a qual categoria um documento pertence, um dado w-conceito candidato é eleito para enriquecer um documento deste conjunto se este w-conceito tiver enriquecido algum documento na etapa de treino, como pode ser visto na Figura 3.3.

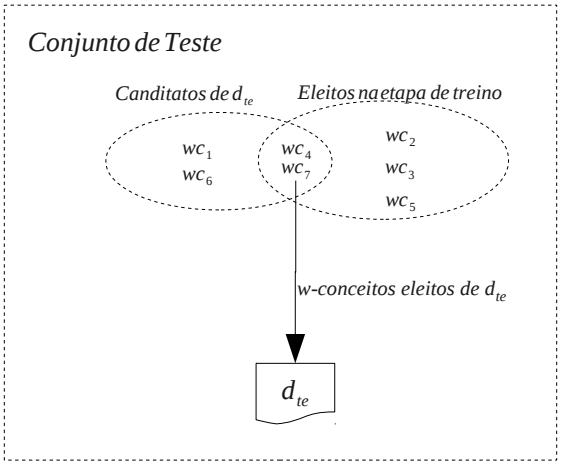

Figura 3.3: *Processo de enriquecimento dos documentos do conjunto de teste.*

3.4 Utilização das Categorias da Wikipédia no Enriquecimento de documentos

Como visto na Seção 2.3, cada conceito da Wikipédia está relacionado a uma ou várias categorias, gerando uma relação de hiponímia entre tais elementos. Nesse sentido, conceitos que compartilham uma mesma categoria tendem a estar semanticamente relacionados.

No presente trabalho, compara-se a eficácia da utilização de w-conceitos com o uso das categorias diretas destes w-conceitos no processo de enriquecimento de documentos. Para tanto, duas abordagens são analisadas:

Primeiramente, para cada w-conceito candidato a enriquecer um documento de determinada categoria, extrai-se da Wikipédia o conjunto de categorias as quais tal w-conceito candidato está diretamente ligado. Dessa maneira, cada w-conceito candidato é substituído pelas categorias da Wikipédia às quais o mesmo está diretamente relacionado. A partir deste processo, gera-se um conjunto de categorias candidatas a enriquecer os documentos das coleções.

A partir desta etapa, aplica-se uma das medidas de avaliação de termos como descrito na Seção 3.2.1 de modo que utiliza-se apenas as categorias que foram aprovadas nesta etapa. As categorias eleitas são, então, utilizadas no enriquecimento dos conjuntos de treino e teste conforme exposto na Seção 3.3 analogamente ao procedimento utilizado para os w-conceitos.

A segunda forma de abordagem de utilização das categorias realiza a união entre o conjunto de w-conceitos candidatos com suas respectivas categorias. A partir deste ponto todas as etapas já descritas para os w-conceitos são realizadas para este novo conjunto de características candidatas.

Ao realizar as duas formas de abordagens, objetiva-se comparar o desempenho das funções de avaliação de termos *FT1C*, *Information Gain*, *Gain Ratio* e *Chi-Squared* em diferentes coleções textuais para o processo de expansão de documentos.

O Capítulo 4 trata do ambiente experimental utilizado, bem como dos resultados obtidos com a utilização das abordagens propostas.

CAPÍTULO 4

Resultados Experimentais

Neste Capítulo apresentam-se os resultados de experimentos com classificação de textos expandidos com características oriundas da Wikipédia. Na Seção 4.1 são apresentadas as informações acerca da versão da Wikipédia utilizada no presente trabalho. Na Seção 4.2 são apresentadas as coleções de dados Reuters, Ohsumed e 20Newsgroups, utilizadas nos experimentos. Na Seção 4.3 apresenta-se a implementação utilizada do SVM, denominada SVM^{perf}. A Seção 4.4 aborda a metodologia utilizada nos experimentos de validação do método proposto. Finalmente, a Seção 4.5 apresenta os resultados alcançados pelos experimentos realizados, discutindo sobre seus ganhos e relacionando-as aos problemas de pesquisa abordados pelo presente trabalho.

4.1 Características Experimentais da Wikipédia

A versão utilizada na pesquisa é a de língua inglesa, por ser a de maior volume de conceitos e porque os textos das coleções utilizadas nos experimentos são escritos em inglês. A data da criação desta versão é de 17 de agosto de 2010, a qual possui um volume de 26.7GB de dados [1].

Foram indexados 6.540.651 wiki-sinônimos diferentes, sejam estes conceitos principais ou de redirecionamento, distribuídos dentre os vários grupos, juntamente com os w-conceitos representantes de cada grupo, conforme definição na Seção 3.1.2 do Capítulo 3. Quanto ao comprimento dos wiki-sinônimos presentes na versão da Wikipédia indexada, os mesmos são distribuídos da seguinte forma:

- Quantidade de wiki-sinônimos de tamanho um (unigramas): 925.808.
- Quantidade de wiki-sinônimos de tamanho dois (bigramas): 2.548.162.
- Quantidade de wiki-sinônimos de tamanho três (trigramas): 1.524.605.
- Quantidade de wiki-sinônimos de tamanho quatro (4-gramas): 749.392.

[1] Disponível em http://dumps.wikimedia.org/enwiki/20100817/.

O restante dos wiki-sinônimos são distribuídos em n-gramas de tamanho superior a quatro. Desta forma, n-gramas com os comprimentos de um a quatro compõe 87.8% dos wiki-sinônimos indexados.

Como visto no Capítulo 2, n-gramas de comprimentos maiores não trazem ganhos concretos ao processo de CAT, de modo que em alguns casos foram observadas depreciações nos resultados. Dessa forma, a presente pesquisa se concentrou nesta faixa de comprimento de conceitos. Após a filtragem descrita na Seção 3.1.1 restaram ainda 3.418.739 wiki-sinônimos. Os que foram eliminados representam conceitos ruidosos, de uso interno pela Wikipédia, ou com funções de ajuda ao usuário desta Enciclopédia.

4.2 Coleções Utilizadas na Validação da Abordagem

Para experimentalmente avaliar nossa estratégia, empregamos três coleções de texto de referência comumente discutidas na literatura:

- Reuters-21578[2] com divisão Aptè de 90 categorias;
- Ohsumed *first*-20000[3];
- 20Newsgroups - *All 20000 documents*[4];

Em todas as coleções, as *stop-words* só foram removidas após o processo de busca por wiki-sinônimos dentro dos documentos, visto que tais termos participam da composição de n-gramas que formam um conceito da Wikipédia. Ademais, como pré-processamento, documentos pertencentes à múltiplas categorias também foram removidos. Assim, todas as coleções resultantes são formadas apenas por documentos uni-rotulados. Foram removidas todas as categorias que não possuem pelo menos um documento no conjunto de treino e um no conjunto de teste.

A coleção Reuters utilizada possui originalmente 12,902 documentos distribuídos em 90 classes utilizando a divisão ModApte [2]. Ao se aplicar as restrições de pré-processamento, esta coleção passou a possuir 9.129 documentos sendo 6.559 no conjunto de treino e 2.570 no conjunto de teste, distribuídos em 52 categorias. Os documentos desta coleção representam notícias apresentando título, corpo do texto, localização geográfica, e data de publicação, dentre outros atributos. A distribuição de documentos pelas categorias pode ser analisada na Figura 4.1, por meio da qual

[2]Disponível em: http://disi.unitn.it/moschitti/corpora/Reuters21578-Apte-90Cat.tar.gz
[3]Disponível em: http://disi.unitn.it/moschitti/corpora/ohsumed-first-20000-docs.tar.gz
[4]Disponível em: http://disi.unitn.it/moschitti/corpora/20_newsgroups.tar.gz

é possível verificar o alto grau de desbalanceamento desta coleção. O número de documentos dentro de uma categoria varia de 1 para a classe *platinum* até 2.840 para a classe *earn*. É possível verificar também que as classes *earn* e *acq* concentram 67.63% de todo o conjunto de treino.

A coleção Ohsumed contém documentos médicos coletados em 1991 relativos a 23 classes. A versão utilizada contém os primeiros 20.000 documentos divididos em 10.000 documentos para o conjunto de treino e 10.000 para o conjunto de teste. Após a etapa de pré-processamento, removemos os documentos multi-classes, de forma que o número total de documentos resultante foi de 7.400 documentos, sendo 3.357 no conjunto de treino e 4.043 no conjunto de teste, distribuídos de forma irregular entre as 23 categorias. A distribuição de documentos no conjunto treino pode ser vista na Figura 4.2.

A coleção 20Newsgroups possui 19.997 artigos contidos em 20 categorias. O conteúdo dos documentos desta coleção é constituído de um conjunto de textos de grupos de discussão provindos da rede Usenet. Esta coleção apresenta uma gama de temas bem diversificados, incluindo categorias pouco relacionadas, assim como categorias fortemente relacionadas entre si, como o caso de Sistemas de Hardware de computadores PC (*comp.sys.ibm.pc*) e Sistemas de Hardware de computadores Macintosh (*comp.sys.mac.hardware*). A 20Newsgroups apresenta um grande vocabulário e palavras que possuem mais de um significado. Ao mesmo tempo, o estilo de escrita de seus documentos corresponde a diálogos por e-mail, o que a coloca bem distante de outras coleções de textos mais técnicos. Após o pré-processamento desta coleção o número total de documentos passou a ser de 19.582. Nesta coleção não há uma divisão de treino e teste padronizada, de forma que nesta coleção utiliza-se o método de validação cruzada, exposta na Seção 2.4.4. Por conseguinte, várias divisões são realizadas para cada uma das k-partições. Neste estudo utiliza-se a validação cruzada de 5-partições, comumente utilizada na literatura. A distribuição geral dos documentos nesta coleção pode ser observada na Figura 4.3.

Após o enriquecimento das coleções, e estas haverem passado pelo processo de eliminação de *stop-words*, é criada a matriz documento-termos conforme exemplificado pela equação 2-1. Nesta abordagem, se um w-conceito ou categoria aparece mais de uma vez no documento, o mesmo será introduzido neste, tantas quanto forem as ocorrências deste no referido documento.

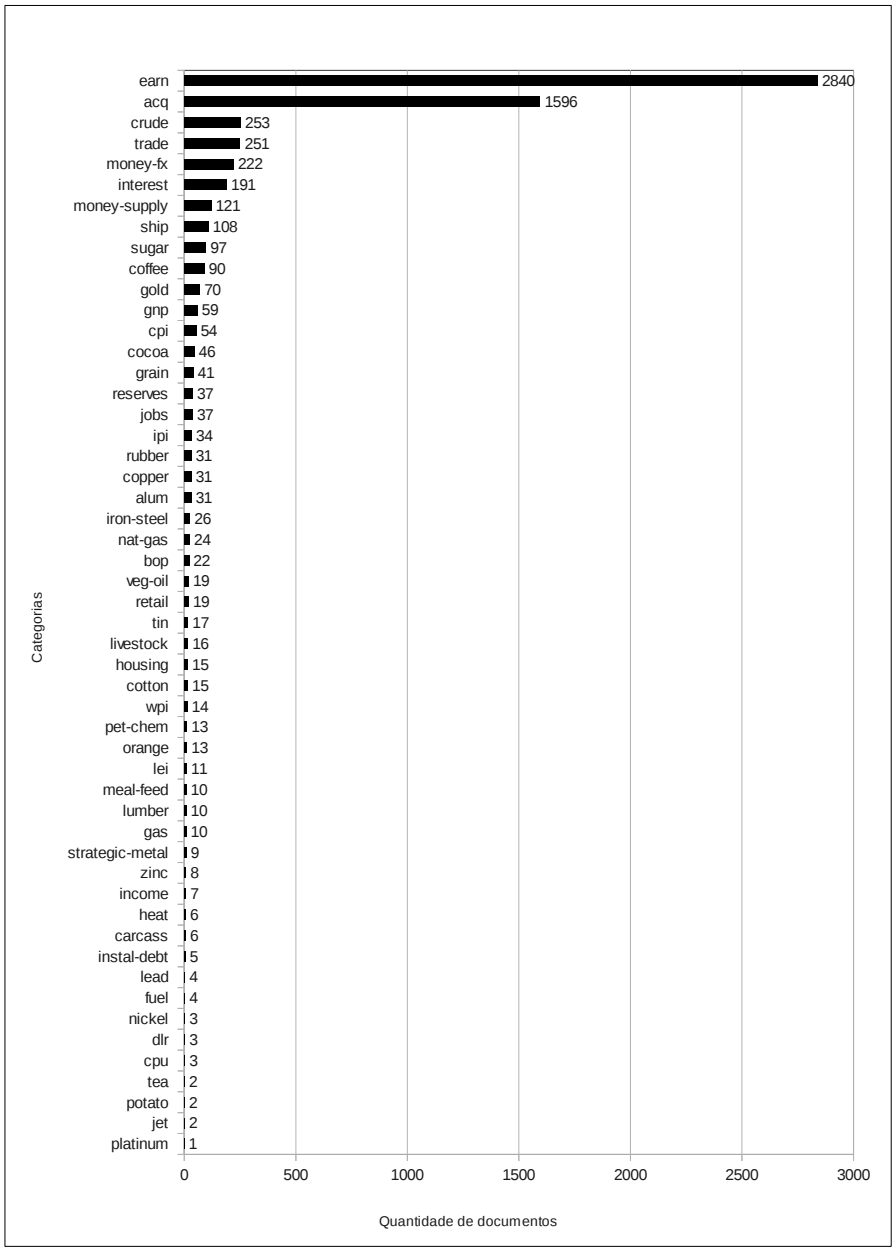

Figura 4.1: *Distribuição dos documentos no conjunto de treino da coleção Reuters-21578 após o pré-processamento.*

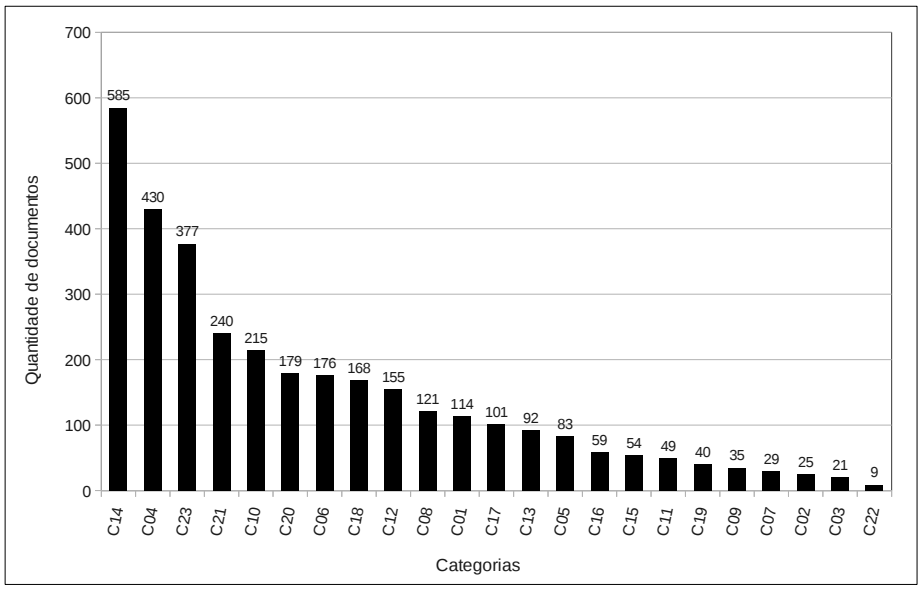

Figura 4.2: *Distribuição dos documentos no conjunto de treino da coleção Ohsumed após o pré-processamento.*

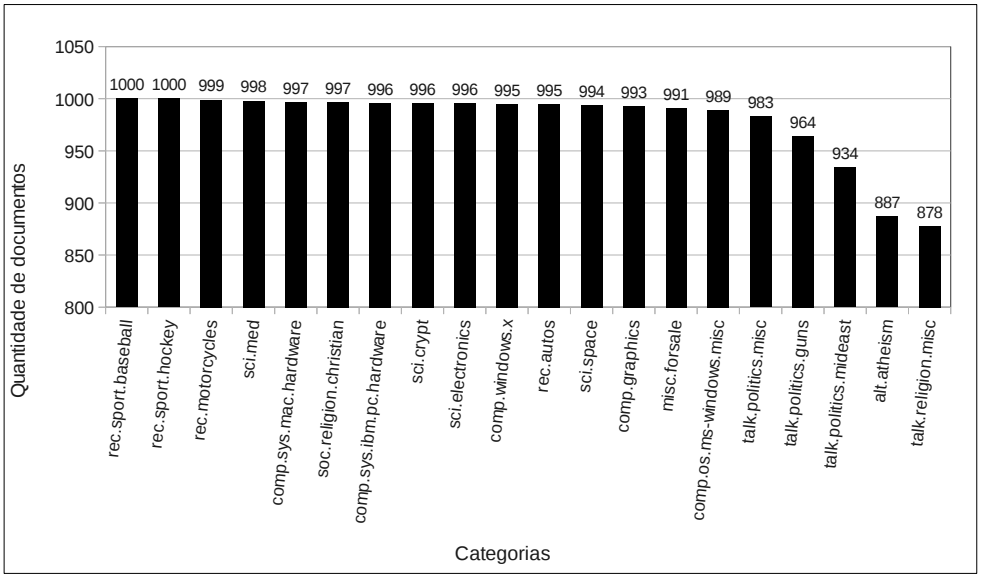

Figura 4.3: *Distribuição dos documentos da coleção 20Newsgroups após o pré-processamento.*

4.3 Ambiente experimental de classificação com SVM

Como explanado na Seção 2.4.3, utilizamos o algoritmo de classificação SVM. Para tanto, fizemos uso do pacote (SVM^{Perf})[5] disponibilizado por Joachims [27]. Este pacote implementa uma versão eficiente do classificador Support Vector Machine (SVM), o qual pode ser treinado em um tempo que aumenta linearmente com o tamanho do conjunto de treino. O SVM^{Perf} é disponibilizado gratuitamente para uso científico.

Após o enriquecimento dos documentos de treino e teste de cada uma das coleções utilizadas na validação do método, a matriz documento-termo de cada um deles é convertida no formato de entrada do SVM^{perf}, tanto para a etapa de treino quanto na classificação.

Como trabalhamos com várias classes em uma mesma coleção, o método, de fato, deve escolher dentre todas as classes desta coleção, apenas uma à qual um dado documento será relacionado. Como o classificador SVM trabalha com classificação binária, utilizamos a abordagem um contra todos, vista na Seção 2.4.2.

Como o presente trabalho não objetiva a extrema otimização da classificação e sim verificar a melhoria do processo classificatório por meio do método proposto, utilizamos a configuração padrão do SVM^{perf} para todas as coleções utilizadas.

4.4 Metodologia Experimental

Nesta seção, são apresentadas as metodologias utilizadas nos experimentos, objetivando aplicar a abordagem proposta de seleção de características (FT1C) com o intuito de compará-la com outras medidas comumente utilizadas na literatura, assim como avaliar as duas formas de expansão de características investigadas: sem aplicação de restrição de classe para as características eleitas (SRC) e com a aplicação de restrição de classe às características eleitas (CRC).

Especificamente, foram realizados experimentos aplicando a abordagem proposta com o enriquecimento somente por w-conceitos, somente por categorias dos w-conceitos e a combinação de ambos, nas coleções Reuters, 20Newsgroups e Ohsumed.

Em cada uma das coleções utilizadas foram feitos experimentos de expansão de características de documentos com w-conceitos, categorias, e a combinação de ambos, utilizando as medidas de seleção de características *Information Gain*, *Gain*

[5]Disponível em: http://svmlight.joachims.org/svm_perf.html

Ratio e *Chi-squared*, apresentadas na Seção 2.1.6 do Capítulo 2, além da medida de seleção de característica proposta denominada $FT1C$ descrita na Seção 3.2.1 do Capítulo 3.

Para cada medida de seleção de característica utilizada nos experimentos, comparou-se os dois métodos de expansão das características eleitas denominados SRC e CRC descritos na Seção 3.3 do Capítulo 3.

Após determinar o tipo de característica provinda da Wikipédia a ser utilizada, a coleção a ser enriquecida, o tipo de medida de importância de termos a ser utilizada, e a abordagem de expansão das características selecionadas, é necessário ainda definir a quantidade k de características que devem ser utilizadas na etapa de seleção de características como descrito na Seção 2.1.6. No presente trabalho, com o intuito de facilitar a comparação entre características, o valor de k é definido por meio da limitação da porcentagem total das características de expansão eleitas.

Apesar de terem sido feitos experimentos com a utilização de porcentagens de utilização de 0,5% à 100% de características de expansão eleitas, os resultados de $microF_1$e $macroF_1$ se deterioram rapidamente acima de 19,5% na maioria dos casos. Diante deste quadro e buscando uma melhor visualização da faixa de porcentagens em que os resultados se mostraram melhores, optou-se por reportar apenas os resultados observados na faixa de porcentagens entre 0,5% até 19,5%.

As Tabelas 4.1, 4.2 e 4.3 mostram a quantidade de características correspondentes a cada valor de percentagem utilizada respectivamente para as coleções Reuters-21578, Ohsumed e 20newsgroups.

Ao analisarmos o conteúdo destas tabelas é possível perceber na linha de porcentagem 100% a quantidade de w-conceitos, categorias e união de ambos, candidatos a enriquecer as coleções de dados. Por exemplo, observe a Tabela 4.1 a qual apresenta 18.654 candidatos w-conceitos, 26.404 candidatos categorias e 45.058 candidatos provenientes da união de w-conceitos e categoria. Como os candidatos do tipo categorias são derivados a partir dos candidatos do tipo w-conceitos, é possível constatar que o crescimento do número de conceitos de categorias depende de como os w-conceitos geradores se relacionam na Wikipédia. Dessa forma, se dois ou mais w-conceitos estão ligados à mesma categoria da Wikipédia, apenas um candidato de categoria será criado para substituir os dois w-conceitos utilizados. É notório que se esta situação tende a se repetir com muita frequência, o número de candidatos do tipo categorias tenderá a ser menor que o número de candidatos do tipo w-conceitos, entretanto, haverá mais documentos compartilhando as mesmas características de expansão, de forma que tais documentos estarão relacionados entre si.

Se por outro lado um único w-conceito pertencer a mais de uma categoria, este w-conceito será substituído por todas as categorias às quais o mesmo pertence.

Por meio do mesmo raciocínio aplicado ao caso anterior, se situações com esta se repetirem demasiadamente, a quantidade de candidatos do tipo categorias tenderá a ser maior que os candidatos do tipo w-conceitos.

Ao se observar novamente a linha de 100% da Tabela 4.1 pode-se perceber que para os w-conceitos que coocorrem na coleção Reuters-21578 e na Wikipédia, muitos deles estão ligados a mais de uma categoria da Wikipédia, o que faz com que o número de candidatos do tipo categorias seja maior que o de w-conceitos. Esta mesma situação ocorre para a coleção 20newsgroups, como pode ser visto na Tabela 4.3.

Para a coleção Ohsumed (Tabela 4.2), a relação entre w-conceitos e categorias se mostrou diferente. O número de categorias candidatas nesta coleção é menor que o número de w-conceitos candidatos, o que leva a concluir que vários w-conceitos compartilham a mesma categoria nesta coleção.

Nas linhas seguintes das Tabelas 4.1, 4.2 e 4.3, são mostradas as variações de porcentagens de características e sua respectiva quantidade absoluta, variando de 0,5% a 19,5%.

Coleção Reuters-21578			
Proporção de utilização de Características Candidatas	Tipos de características de expansão		
	w-conceitos	categorias	w-conceitos + categorias
100,0%	18654	26404	45058
0,5%	93	132	225
1,0%	187	264	451
1,5%	280	396	676
2,0%	373	528	901
2,5%	466	660	1126
3,0%	560	792	1352
3,5%	653	924	1577
4,0%	746	1056	1802
4,5%	839	1188	2028
5,0%	933	1320	2253
5,5%	1026	1452	2478
6,0%	1119	1584	2703
6,5%	1213	1716	2929
7,0%	1306	1848	3154
7,5%	1399	1980	3379
8,0%	1492	2112	3605
8,5%	1586	2244	3830
9,0%	1679	2376	4055
9,5%	1772	2508	4281
10,0%	1865	2640	4506
10,5%	1959	2772	4731
11,0%	2052	2904	4956
11,5%	2145	3036	5182
12,0%	2238	3168	5407
12,5%	2332	3301	5632
13,0%	2425	3433	5858
13,5%	2518	3565	6083
14,0%	2612	3697	6308
14,5%	2705	3829	6533
15,0%	2798	3961	6759
15,5%	2891	4093	6984
16,0%	2985	4225	7209
16,5%	3078	4357	7435
17,0%	3171	4489	7660
17,5%	3264	4621	7885
18,0%	3358	4753	8110
18,5%	3451	4885	8336
19,0%	3544	5017	8561
19,5%	3638	5149	8786

Tabela 4.1: *Tabela demostrativa relacionando porcentagem de uso de características de expansão e sua respectiva quantidade absoluta k para a coleção Reuters-21578.*

Coleção Ohsumed			
Proporção de utilização de Características Candidatas	Tipos de características de expansão		
	w-conceitos	categorias	w-conceitos + categorias
100,0%	16074	14929	31003
0,5%	80	75	155
1,0%	161	149	310
1,5%	241	224	465
2,0%	321	299	620
2,5%	402	373	775
3,0%	482	448	930
3,5%	563	523	1085
4,0%	643	597	1240
4,5%	723	672	1395
5,0%	804	746	1550
5,5%	884	821	1705
6,0%	964	896	1860
6,5%	1045	970	2015
7,0%	1125	1045	2170
7,5%	1206	1120	2325
8,0%	1286	1194	2480
8,5%	1366	1269	2635
9,0%	1447	1344	2790
9,5%	1527	1418	2945
10,0%	1607	1493	3100
10,5%	1688	1568	3255
11,0%	1768	1642	3410
11,5%	1849	1717	3565
12,0%	1929	1791	3720
12,5%	2009	1866	3875
13,0%	2090	1941	4030
13,5%	2170	2015	4185
14,0%	2250	2090	4340
14,5%	2331	2165	4495
15,0%	2411	2239	4650
15,5%	2491	2314	4805
16,0%	2572	2389	4960
16,5%	2652	2463	5115
17,0%	2733	2538	5271
17,5%	2813	2613	5426
18,0%	2893	2687	5581
18,5%	2974	2762	5736
19,0%	3054	2837	5891
19,5%	3134	2911	6046

Tabela 4.2: *Tabela demostrativa relacionando porcentagem de uso de características de expansão e sua respectiva quantidade absoluta k para a coleção Ohsumed.*

Coleção 20NG			
Proporção de utilização de Características Candidatas	Tipos de características de expansão		
	w-conceitos	categorias	w-conceitos + categorias
100,0%	55554	61391	116945
0,5%	278	307	585
1,0%	556	614	1169
1,5%	833	921	1754
2,0%	1111	1228	2339
2,5%	1389	1535	2924
3,0%	1667	1842	3508
3,5%	1944	2149	4093
4,0%	2222	2456	4678
4,5%	2500	2763	5263
5,0%	2778	3070	5847
5,5%	3055	3377	6432
6,0%	3333	3683	7017
6,5%	3611	3990	7601
7,0%	3889	4297	8186
7,5%	4167	4604	8771
8,0%	4444	4911	9356
8,5%	4722	5218	9940
9,0%	5000	5525	10525
9,5%	5278	5832	11110
10,0%	5555	6139	11695
10,5%	5833	6446	12279
11,0%	6111	6753	12864
11,5%	6389	7060	13449
12,0%	6666	7367	14033
12,5%	6944	7674	14618
13,0%	7222	7981	15203
13,5%	7500	8288	15788
14,0%	7778	8595	16372
14,5%	8055	8902	16957
15,0%	8333	9209	17542
15,5%	8611	9516	18126
16,0%	8889	9823	18711
16,5%	9166	10130	19296
17,0%	9444	10436	19881
17,5%	9722	10743	20465
18,0%	10000	11050	21050
18,5%	10277	11357	21635
19,0%	10555	11664	22220
19,5%	10833	11971	22804

Tabela 4.3: *Tabela demostrativa relacionando porcentagem de uso de características de expansão e sua respectiva quantidade absoluta k para a coleção 20NG.*

4.5 Análise dos resultados

Nesta seção, são analisados os resultados dos experimentos realizados com o objetivo de responder aos problemas de pesquisa apresentados no Capítulo 1, os quais são transcritos abaixo com o objetivo de facilitar a leitura:

Problema de pesquisa 1: *A aplicação de um método de seleção de características consegue melhorar a eficácia da utilização das relações de sinonímia e de categorias provindas da Wikipédia durante o processo de expansão de documentos, reduzindo a inserção de ruídos e potencializando a adição de características boas discriminadoras de classes?*

Problema de pesquisa 2: *A utilização de uma medida de avaliação de termos que pontue positivamente a abundância de uma característica na classe a qual pertence o documento de treino a ser expandido e utilize como penalização a abundância relativa desta mesma característica nas outras classes da coleção, pode se mostrar como opção competitiva na seleção de características provindas da Wikipédia na forma de conceitos sinônimos e categorias?*

Problema de pesquisa 3: *A utilização de um método o qual permita a expansão de documentos somente com características bem avaliadas na classe do documento de treino a ser expandido, poderia aumentar a eficácia da classificação de documentos enriquecidos com características provindas da Wikipédia?*

Para que seja possível responder ao Problema de Pesquisa 1, deve-se comparar os resultados em termos de $microF_1$ e $macroF_1$ obtidos ao se enriquecer uma coleção com w-conceitos e/ou categorias sem a utilização de nenhuma medida de seleção de características, confrontando-os com os resultados obtidos com a utilização das medidas *Information Gain*, *Gain Ratio*, *Chi-squared*, além da medida *FT1C* proposta neste trabalho. Desta forma, pode-se confirmar ou refutar a hipótese de melhoria na expansão de documentos, apresentada para este problema na Seção 1.

Com o objetivo de responder ao Problema de Pesquisa 2, os experimentos com as diversas medidas de seleção de características são utilizados para avaliar o desempenho da medida FT1C comparando seus resultados em termos de $microF_1$ e $macroF_1$ obtidos ao se enriquecer uma coleção com w-conceitos e/ou categorias, com relação às demais medidas avaliadas, confirmando-se ou refutando-se a hipótese de competitividade da medida FT1C, apresentada para este problema na Seção 2.

Para que se possa responder ao Problema de Pesquisa 3, deve-se aplicar a restrição de classe apresentada pelo problema, denominada CRC, em conjunto com todas as medidas de seleção de características abordadas. Os resultados obtidos em termos de $microF_1$ e $macroF_1$ com a aplicação desta restrição são então confrontados com os resultados obtidos com as mesmas medidas de seleção de características

sem a aplicação da restrição de classe, SRC. A partir destes experimentos é possível confirmar ou refutar a hipótese apresentada para este problema na Seção 3, esperando-se que a restrição CRC poderia melhorar os resultados da CAT, quando comparado com as abordagens sem o uso da restrição.

Foram definidas várias abordagens experimentais com o intuito de conduzir os experimentos de modo a possibilitar as análises do problemas. Para cada uma das coleções Reuters, 20Newsgroup (20NG) e Ohsumed, foram coletados os resultados para as seguintes abordagens:

1. Classificação de cada coleção sem o uso de expansão de documentos, utilizado como linha base;
2. Classificação com a expansão de cada coleção com características da Wikipédia sem o uso de nenhuma medida de seleção de características, ou seja, 100% das características extraídas da Wikipédia são utilizadas.
3. Classificação com a expansão de cada coleção com características da Wikipédia, utilizando cada uma das medidas de seleção de características *Information Gain*, *Gain Ratio* e *Chi-squared*, sem a aplicação da restrição CRC na expansão dos documentos.
4. Classificação com a expansão de cada coleção com características da Wikipédia, utilizando a medida de seleção de características FT1C, proposta neste trabalho, sem a aplicação da restrição CRC na expansão dos documentos.
5. Classificação com a expansão de cada coleção com características da Wikipédia, utilizando cada uma das medidas de seleção de características *Information Gain*, *Gain Ratio* e *Chi-squared*, com a aplicação da restrição CRC na expansão dos documentos.
6. Classificação com a expansão de cada coleção com características da Wikipédia, utilizando a medida de seleção de características FT1C, proposta neste trabalho, sem a aplicação da restrição CRC na expansão dos documentos.

Foram realizados experimentos aplicando as abordagens propostas utilizando características da Wikipédia na forma de w-conceitos, de categorias dos w-conceitos e da combinação de ambos. Com isso, pôde-se analisar o desempenho das abordagens de expansão tanto em diferentes tipos de coleções textuais como também utilizando características da Wikipédia de naturezas diferentes.

A Seção 4.5.1 apresenta a utilização de w-conceitos como características da Wikipédia, analisando os resultados das abordagens experimentais com o intuito de relacionar tais resultados com os Problemas de Pesquisa 1, 2 e 3. O mesmo é realizado para as Seções 4.5.2 e 4.5.3 nas quais utilizam-se as categorias dos w-conceitos e a combinação de ambos, respectivamente.

A Tabela 4.4 apresenta a relação de todos os tipos de abordagens utilizadas neste trabalho. Para cada uma das abordagens relacionadas nesta tabela experimentou-se variar o número k de características eleitas de acordo com a porcentagens do total de candidatas, conforme apresentado nas Tabelas 4.1, 4.2 e 4.3, variando de $0,5\%$ a $19,5\%$.

Abordagens investigadas		
W-conceitos	FT1C	SRC
		CRC
	GAIN RATIO	SRC
		CRC
	INFOGAIN	SRC
		CRC
	CHI-SQUARED	SRC
		CRC
Categorias	FT1C	SRC
		CRC
	GAIN RATIO	SRC
		CRC
	INFOGAIN	SRC
		CRC
	CHI-SQUARED	SRC
		CRC
W-conceitos + Categorias	FT1C	SRC
		CRC
	GAIN RATIO	SRC
		CRC
	INFOGAIN	SRC
		CRC
	CHI-SQUARED	SRC
		CRC

Tabela 4.4: *Relação de abordagens investigas nos experimentos realizados*

Os experimentos foram conduzidos com o intuito de calcular o valor da média $microF_1$ e $macroF_1$ para cada uma das porcentagens de características eleitas em cada uma das abordagens investigadas. Desta forma, foram realizados 39 experimentos para cada uma das abordagens da Tabela 4.4, além do resultado obtido com a expansão sem a seleção de características, ou seja, utilizando 100% das características candidatas.

A significância estatística dos resultados de $microF_1$ e $macroF_1$ foi obtida por meio do teste estatístico Wilcoxon [66] bicaudal tendo como amostras pareadas todos os 39 diferentes valores de $microF_1$, e posteriormente de $macroF_1$, com relação à linha base, obtidos para cada k de uma mesma abordagem.

As tabelas comparativas apresentadas nas seções seguintes contém a coluna s.e. a qual informa a significância estatística dos resultados apresentados, expressos pelas Figuras ▲ ■ e ●, as quais significam, respectivamente, que o ganho ou perda apresentado foi fortemente significativo ($\geq 98\%$), significativo ($90\% \leq x < 98\%$) ou não significativo ($< 90\%$).

Os resultados apresentados para as coleções Reuters-21578 e Ohsumed foram obtidos a partir da divisão fixa de conjunto de treino e teste. Entretanto, para a coleção 20Newsgroups utilizou-se 5 particionamentos diferentes, de modo os resultados apresentados são calculados a partir da média obtida entre as partições, conforme explanado na Seção 4.2.

A estabilidade de cada abordagem é calculada pelo desvio padrão dos valores de $microF_1$ e $macroF_1$, e é expresso nas tabelas comparativas apresentadas nas seções seguintes (coluna d.p.).

4.5.1 Expansão com w-conceitos

Primeiramente, as abordagens propostas foram aplicadas utilizando w-conceitos como características de expansão de documentos. A Figura 4.4 mostra os resultados em termos de $microF_1$ para a coleção Reuters. Cada quadrante da Figura corresponde a uma medida de seleção de característica utilizada de forma comparativa com as demais. O gráfico de cada quadrante contém duas curvas, cada uma correspondendo a um dos métodos de expansão de termos investigados: SRC e CRC.

Comparativo entre CRC e SRC

A primeira comparação que pode ser realizada é quanto ao desempenho das metodologias de expansão de termos sem a utilização de restrição de classe (SRC) e com a aplicação de restrição de classe (CRC).

A metodologia SRC se mostrou superior à metodologia CRC para todas as medidas de seleção de características aplicadas à w-conceitos da coleção Reuters, como pode ser visto na Figura 4.4 para $microF_1$ e na Figura 4.5 para $macroF_1$.

O método de expansão CRC apresentou um comportamento estável somente quando utilizado com a medida de FT1C de seleção de características, sendo que para as demais medidas de seleção de características o método prejudica sensivelmente a CAT quando comparado com a classificação sem expansão tanto em $microF_1$ quanto em $macroF_1$.

Apesar de nenhum método de expansão de características ter alcançado ganhos consideráveis em termos de $microF_1$ na coleção Reuters, o método de

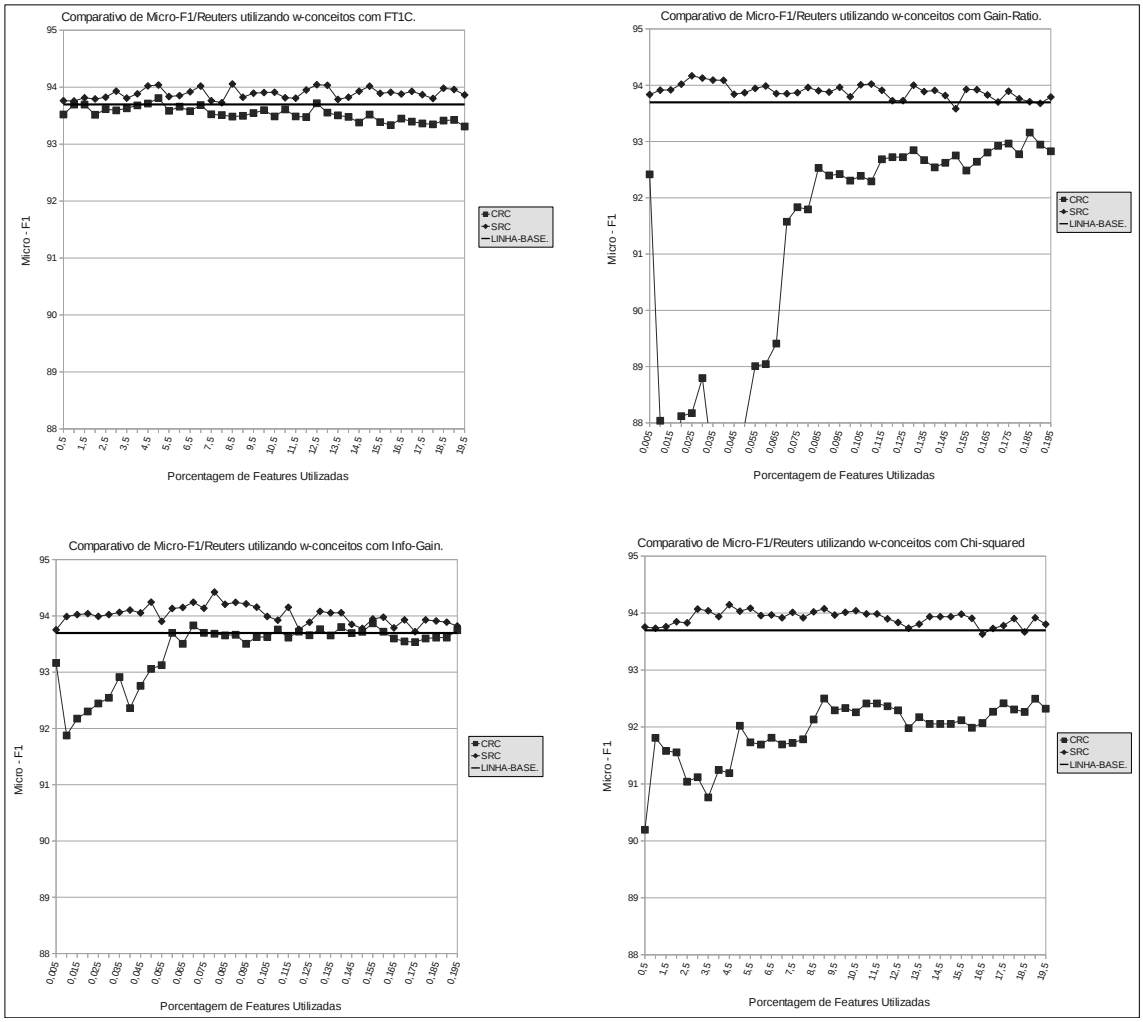

Figura 4.4: *Resultados de $microF_1$ para coleção Reuters com w-conceitos e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

expansão de características SRC demonstrou ser o mais estável tanto em relação à variação de percentagem de inclusão de w-conceitos, quanto à variação das medidas de seleção de características.

É importante ressaltar que apesar da contribuição do método de expansão SRC ser pequena em termos de $microF_1$, este resultado é importante, haja vista que o valor da linha-base de $microF_1$ para a Reuters já se encontra em um patamar elevado $(93,69\%)$.

Nas Figuras 4.6 e 4.7, pode-se verificar que a metodologia com restrição de classe CRC melhora os resultados de classificação da coleção Ohsumed, em comparação com a linha base, tanto em $microF_1$ quanto em $macroF_1$ somente quando utilizada em conjunto com FT1C, sendo que para $macroF_1$ (Figura 4.7)

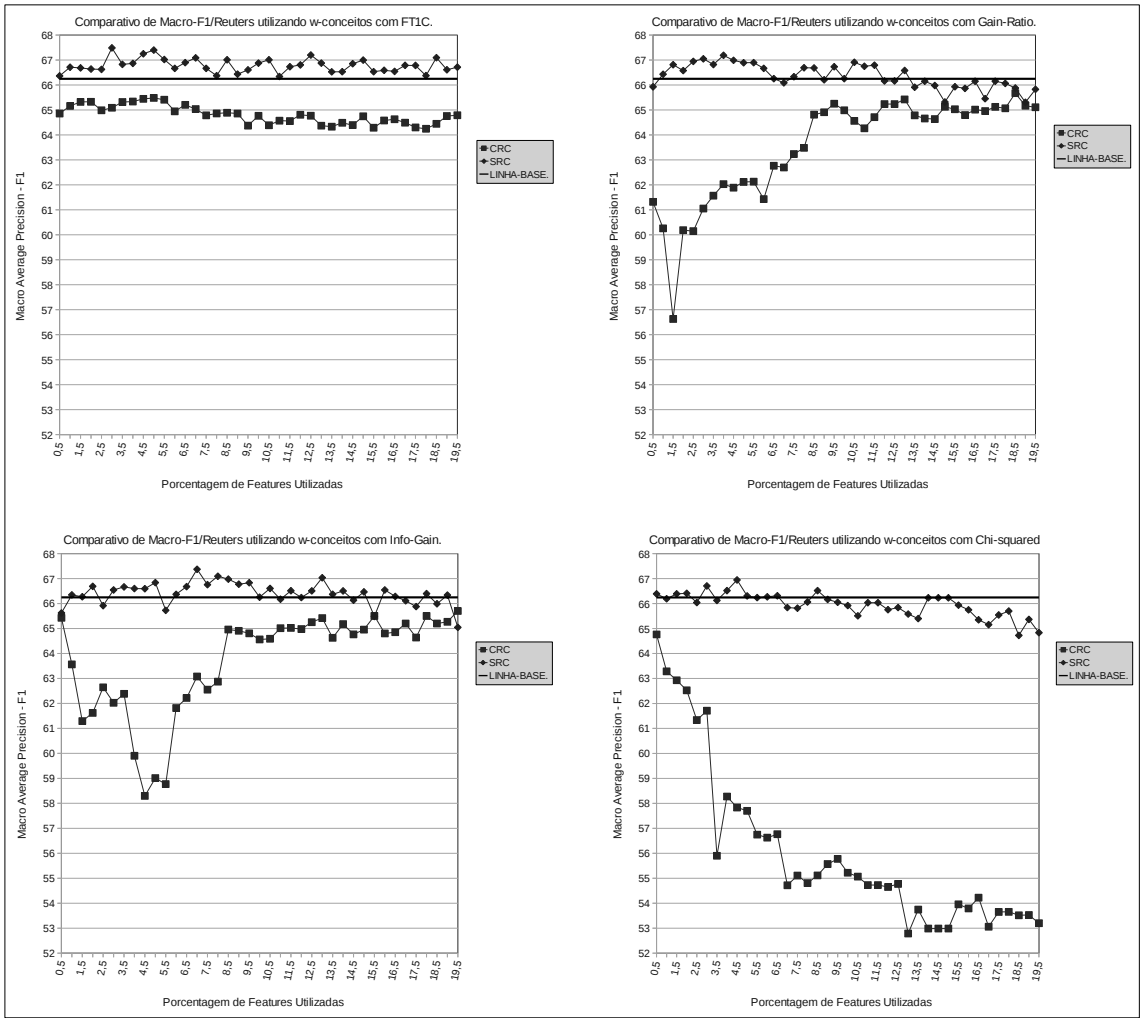

Figura 4.5: *Resultados de $macroF_1$ para coleção Reuters com w-conceitos e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

esta metodologia consegue uma pequena superação em relação à abordagem sem restrição de classe SRC, apresentando o maior valor de $macroF_1$ dentre todas as abordagens, alcançando $52,16\%$ nesta média, o que representa $8,16\%$ de ganho se comparado à linha base, um pouco superior à metodologia SRC a qual atinge $51,80\%$, o que representa $7,41\%$ de ganho. Com as demais medidas de seleção de característica, o método SRC visivelmente se mostra superior ao método CRC.

Os resultados mostrados nas Figuras 4.8 e 4.9, para a coleção 20Newsgroups, demostram que também para esta coleção o método CRC apresenta ganhos somente quando utilizado com a medida FT1C, na expansão com w-conceitos.

Apesar dos bons resultados apresentados pela expansão de documentos utilizando CRC em conjunto com a medida FT1C nas coleções Ohsumed e 20News-

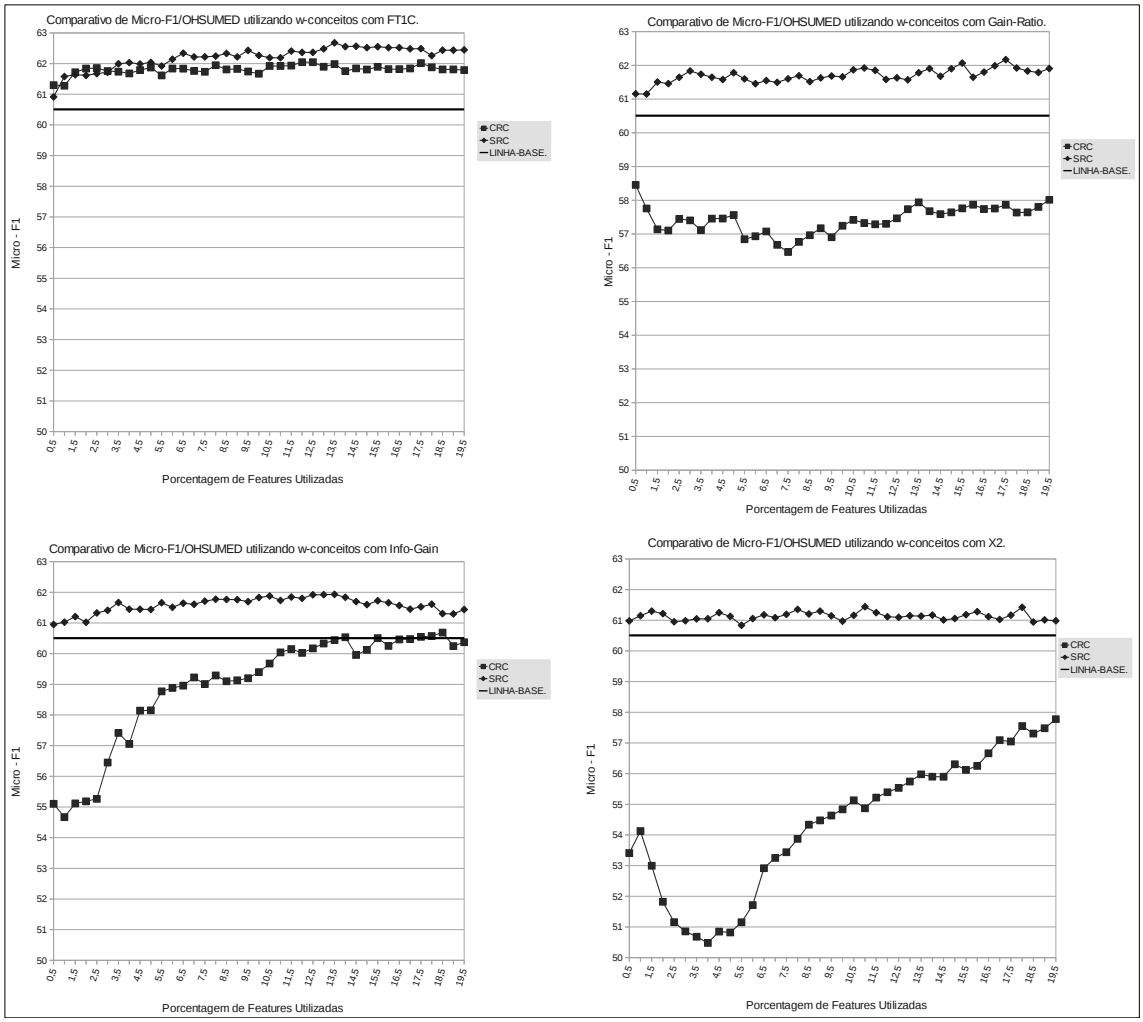

Figura 4.6: *Resultados de microF$_1$ para coleção Ohsumed com w-conceitos e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

Groups, o baixo desempenho desta restrição ao ser utilizada com outras medidas de seleção de características, além da degradação da classificação da coleção Reuters, mesmo em conjunto com a medida FT1C, nos leva a concluir que o método CRC é instável, variando muito o resultado da classificação e que portanto não é recomendado como um método geral de expansão de características. O método SRC, por sua vez é o que apresentou resultados mais estáveis em todas as coleções, e principalmente, quando utilizado com a medida FT1C .

Comparativo entre medidas de seleção de características

Conforme explicitado na subseção anterior, o método CRC aplicado a w-conceitos somente apresentou resultados satisfatórios quando utilizado em conjunto

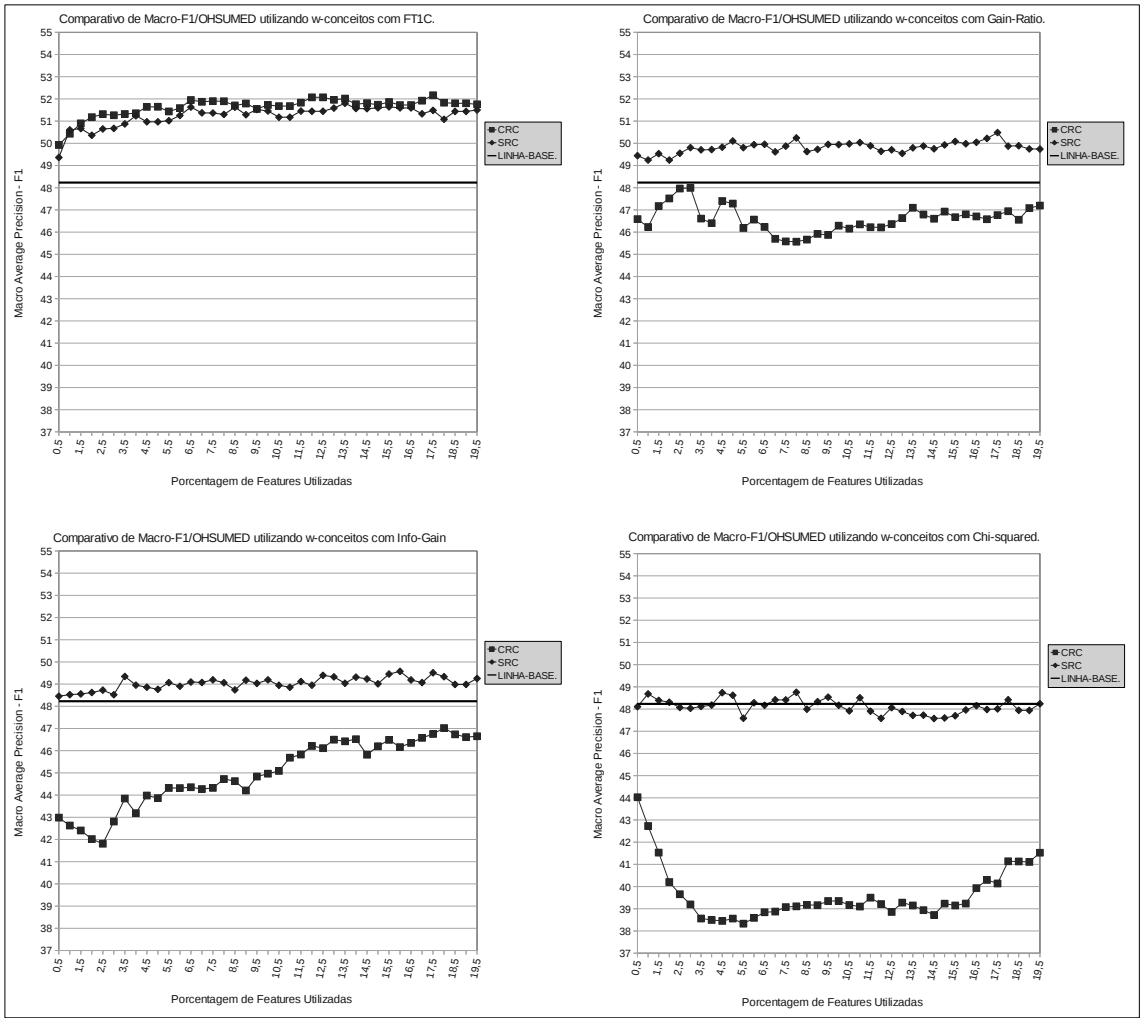

Figura 4.7: *Resultados de $macroF_1$ para coleção Ohsumed com w-conceitos e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

com a medida FT1C. Diante do exposto, utilizamos apenas o método sem restrição de classe SRC nos experimentos de comparação entre as 4 medidas de seleção de características utilizadas.

Apesar do método de seleção de características *Information Gain* ter apresentado o pico mais alto de $microF_1$ para a expansão de w-conceitos na coleção Reuters com SRC, a medida FT1C se mostrou competitiva com as medidas já consagradas na literatura *Information Gain*, *Gain Ratio* e *Chi-squared*, visto que apresentou a menor queda, estando sempre acima da linha base, conforme pode ser visualizado na Tabela 4.5.

A Figura 4.5 mostra os resultados em termos de $macroF_1$ para a coleção Reuters. A medida de seleção de características FT1C apresentou o maior ganho

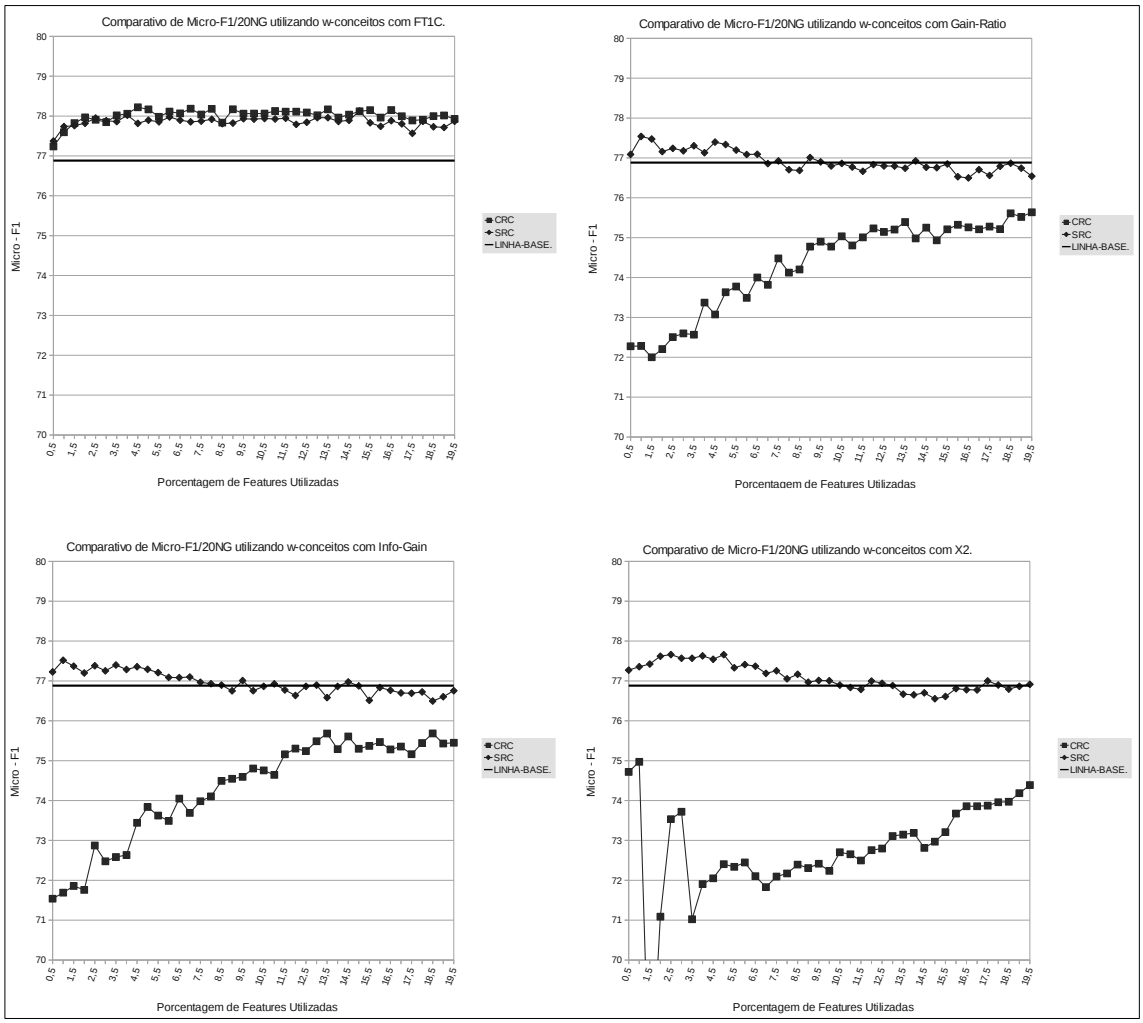

Figura 4.8: *Resultados de $microF_1$ para coleção 20NG com w-conceitos e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

Resultados de $microF_1$ para Reuters expandida com w-conceitos/SRC							
Med. de Seleção	Linha Base	$microF_1$ MÁX	ganho	$microF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	93,69	94,05	0,38%	93,72	0,03%	0,092	▲
Gain Ratio		94,16	0,50%	93,58	-0,12%	0,130	▲
Info Gain		94,42	0,77%	93,71	0,02%	0,160	▲
Chi-Squared		94,14	0,47%	93,63	-0,06%	0,124	▲
100% dos Candidatos		93,77	0,09%	93,77	0,09%		

Tabela 4.5: *Resultados máximos e mínimos de $microF_1$ para Reuters expandida com w-conceitos e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

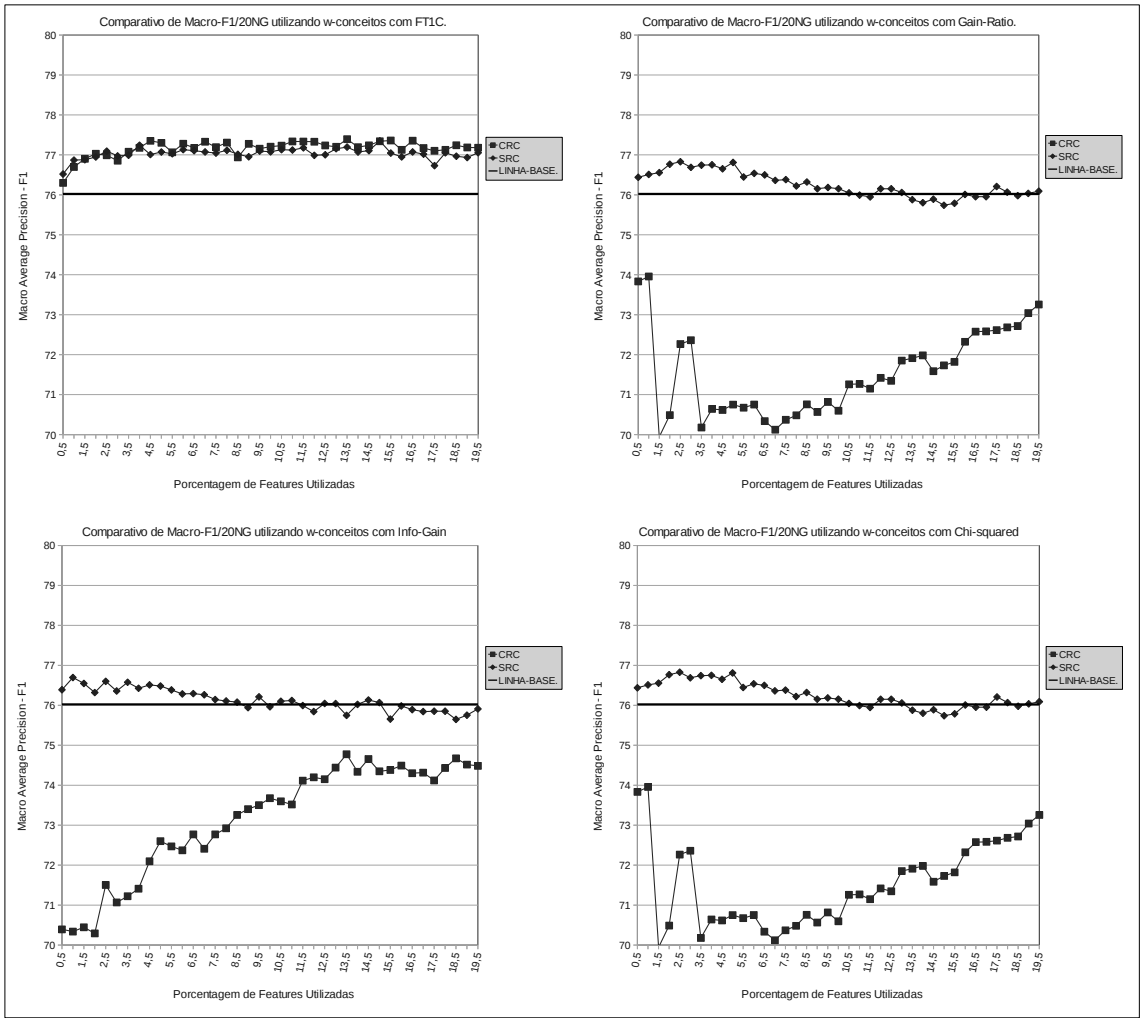

Figura 4.9: *Resultados de $macroF_1$ para coleção 20NG com w-conceitos e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

de $macroF_1$, utilizando o método de expansão SRC, como pode se observado na Tabela 4.6, com valor máximo de $67,48\%$, mostrando maior estabilidade durante as variações de porcentagens de características, sendo a única a se manter acima da linha base em toda a faixa de valores analisada.

Nesse sentido, apesar de apresentar um ganho máximo de $macroF_1$ de apenas $1,86\%$, a medida FT1C com método SRC não gera degradação da classificação da coleção Reuters em nenhuma porcentagem de característica.

As Figuras 4.6 e 4.7 apresentam os resultados de $microF_1$ e $macroF_1$, respectivamente, para a coleção Ohsumed ao ser expandida com w-conceitos. Ao se analisar os gráficos das Figuras pode-se constatar uma maior estabilidade do método FT1C se comparado às demais medidas de seleção de características, também para

Resultados de $macroF_1$ para Reuters expandida com w-conceitos/SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	66,24	67,48	1,86%	66,33	0,13%	0,283	▲
Gain Ratio		67,18	1,42%	65,30	-1,42%	0,485	●
Info Gain		67,37	1,69%	65,04	-1,81%	0,404	▲
Chi-Squared		66,94	1,05%	64,72	-2,29%	0,447	▲
100% dos Candidatos		65,68	-0,85%	65,68	-0,85%		

Tabela 4.6: *Resultados máximos e mínimos de $macroF_1$ para Reuters expandida com w-conceitos e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

esta coleção. Os valores máximos e mínimos de todas as medidas de seleção de características para a coleção Ohsumed são apresentados nas Tabelas 4.7 e 4.8. Por

Resultados de $microF_1$ para Ohsumed expandida com w-conceitos/SRC							
Med. de Seleção	Linha Base	$microF_1$ MÁX	ganho	$microF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	60,50	62,67	3,58%	60,91	0,67%	0,364	▲
Gain Ratio		62,17	2,74%	61,14	1,06%	0,212	▲
Info Gain		61,93	2,35%	60,95	0,73%	0,254	▲
Chi-Squared		61,44	1,54%	60,83	0,54%	0,133	▲
100% dos Candidatos		61,14	1,06%	61,14	1,06%		

Tabela 4.7: *Resultados máximos e mínimos de $microF_1$ para Ohsumed expandida com w-conceitos e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

meio destas tabelas é possível verificar que a medida FT1C apresentou os melhores resultados tanto em $microF_1$ para o qual obteve $3,58\%$ de ganho em relação à linha base, quanto de $macroF_1$ para a qual o ganho foi de $7,41\%$.

A coleção 20Newsgroups tem seus resultados de $microF_1$ e $macroF_1$ apresentados nos gráficos das Figuras 4.8 e 4.9. A partir destes gráficos é possível constatar a superioridade da medida de seleção FT1C na expansão por w-conceitos, a qual apresentou maior estabilidade que as demais, mostrando um bom suporte às alterações na porcentagem w-conceitos utilizados.

As demais medidas *Gain Ratio*, *Information Gain* e *Chi-squared* apresentaram quedas acentuadas de desempenho com o aumento da porcentagem de características utilizadas na expansão por w-conceitos.

Por meio das Tabelas 4.9, 4.10 pode-se verificar que os maiores valores de $microF_1$ e $macroF_1$ também foram alcançados pela medida FT1C com ganhos de

Resultados de $macroF_1$ para Ohsumed expandida com w-conceitos/SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	48,22	51,80	7,41%	49,36	2,34%	0,464	▲
Gain Ratio		50,48	4,67%	49,24	2,11%	0,252	▲
Info Gain		49,57	2,78%	48,45	0,47%	0,281	▲
Chi-Squared		48,75	1,08%	47,57	-1,36%	0,331	■
100% dos Candidatos		48,15	-0,15%	48,15	-0,15%		

Tabela 4.8: *Resultados máximos e mínimos de $macroF_1$ para Ohsumed expandida com w-conceitos e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

$1,61\%$ e $1,74\%$ respectivamente. Esta medida também foi a única que não trouxe

Resultados de $microF_1$ para 20Newsgroups expandida com w-conceitos/SRC							
Med. de Seleção	Linha Base	$microF_1$ MÁX	ganho	$microF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	76,88	78,12	1,61%	77,37	0,63%	0,125	▲
Gain Ratio		77,53	0,85%	76,49	-0,49%	0,263	●
Info Gain		77,51	0,82%	76,49	-0,50%	0,273	●
Chi-Squared		77,66	1,01%	76,55	-0,42%	0,335	▲
100% dos Candidatos		76,67	-0,27%	76,67	-0,27%		

Tabela 4.9: *Resultados máximos e mínimos de $microF_1$ para 20Newsgroups expandida com w-conceitos e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

degradação para a CAT dentro da faixa de porcentagens analisada, mesmo no pior caso de $microF_1$ e $macroF_1$.

Além das 4 medidas de seleção de características, também testou-se o desempenho da CAT ao se expandir as coleções Reuters, Ohsumed, e 20Newsgroups utilizando 100% das características candidatas, ou seja, sem realizar qualquer seleção de características ou restrição de expansão. A inserção de 100% dos w-conceitos é utilizada por Wang et al. [64], cujas conclusões defendem que a utilização de w-conceitos (relações de sinonímia) no processo de expansão de documentos não consegue trazer melhorias à CAT. A análise das Tabelas 4.5, 4.6, 4.7, 4.8, 4.9 e 4.10, especificamente as linhas referenciadas como "100% de candidatos" confirmam a baixa qualidade dos resultados de $microF_1$ e $macroF_1$ para esta abordagem. Entretanto, ao se analisar as mesmas Tabelas, foi possível verificar que a aplicação de seleção de características (do inglês, *feature selection*), conseguiu selecionar w-

Resultados de $macroF_1$ para 20Newsgroups expandida com w-conceitos/SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	76,02	77,35	1,74%	76,51	0,65%	0,139	▲
Gain Ratio		76,68	0,86%	75,66	-0,47%	0,263	▲
Info Gain		76,69	0,88%	75,64	-0,49%	0,276	■
Chi-Squared		76,82	1,05%	75,73	-0,37%	0,319	▲
100% dos Candidatos		75,86	-0,21%	75,86	-0,21%		

Tabela 4.10: *Resultados máximos e mínimos de $macroF_1$ para 20Newsgroups expandida com w-conceitos e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

conceitos que potencializam o processo de CAT das coleções testadas, alcançando excelentes resultados, em especial para a coleção Ohsumed, a qual é reconhecida por sua dificuldade em obter bons resultados de classificação. Diante do que foi exposto, é possível responder positivamente ao Problema de Pesquisa 1, confirmando a hipótese relacionada ao mesmo, quando utiliza-se w-conceitos como características de expansão.

O desempenho superior da medida FT1C em todas as coleções, quando enriquecidas com w-conceitos, responde positivamente ao Problema de Pesquisa 2, auxiliando na confirmação da hipótese relacionada a este problema.

4.5.2 Expansão com categorias diretas

Assim como realizado para a Seção 4.5.1, avalia-se os resultados de $microF_1$ e $macroF_1$ ao se enriquecer uma coleção com categorias diretas dos w-conceitos sem a utilização de nenhuma medida de seleção de característica ou método de expansão. Tais resultados devem ser confrontados com a utilização das diversas abordagens propostas, de forma a auxiliar na confirmação ou não das hipóteses relacionadas aos Problemas de Pesquisa 1, 2 e 3.

Comparativo entre CRC e SRC

A metodologia SRC também se mostrou mais estável que a metodologia CRC para todas as medidas de seleção de características aplicadas às categorias em todas as coleções testadas, se adaptando bem às diversas medidas de seleção de características. A metodologia CRC, ao proibir o enriquecimento dos documentos de uma classe com características da Wikipédia que obtém valor global nas medidas de avaliação menor que qualquer uma das k características selecionadas, leva o

classificador a aprender tal comportamento, o qual não se repete ao se enriquecer o conjunto de teste, de modo que tal fato gera erros de classificação ao se utilizar o método CRC.

Nos experimentos com expansão por categorias, o método CRC conseguiu sobrepor os ganhos do método SRC somente para a média $macroF_1$ da coleção Ohsumed, utilizando a medida FT1C, como pode ser visto nas Figuras 4.10 a 4.15.

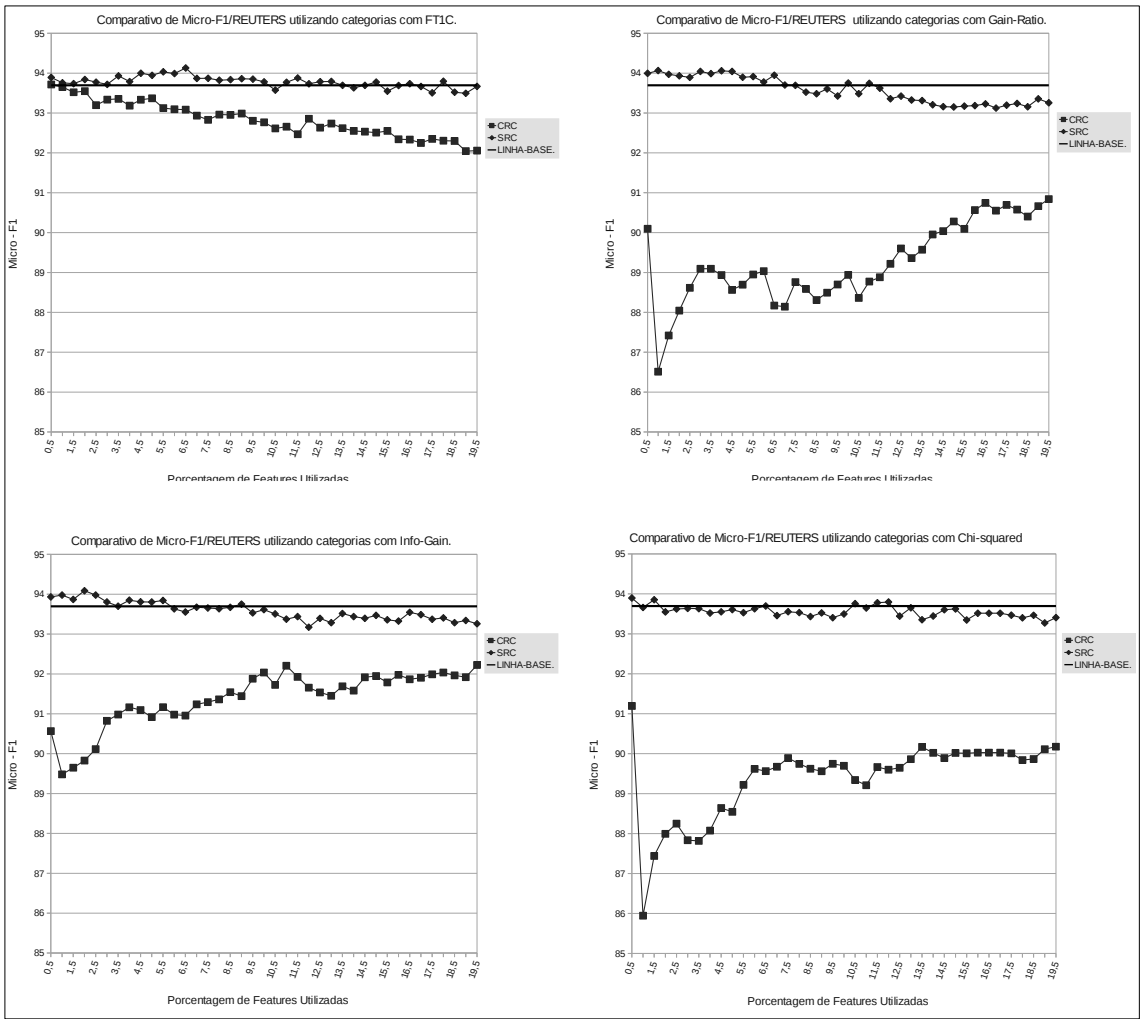

Figura 4.10: *Resultados de $microF_1$ para coleção Reuters com categorias e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

A degradação dos resultados de $microF_1$ e $macroF_1$ ao se utilizar o método de restrição CRC com as medidas de seleção de características *Information Gain*, *Gain Ratio* e *Chi-squared*, juntamente com o baixo desempenho desta metodologia de restrição na média $microF_1$ com a medida FT1C, contribuem para se refutar a

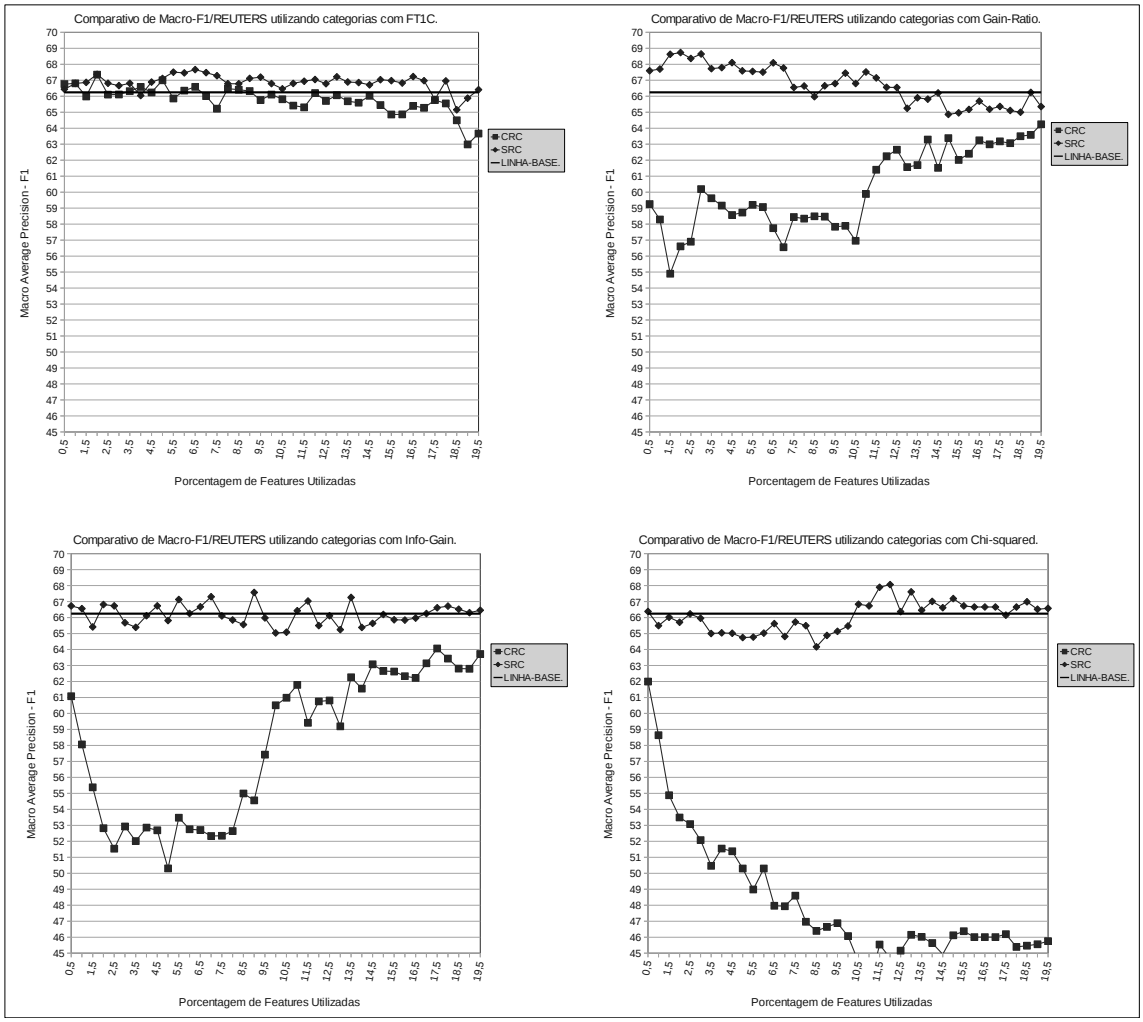

Figura 4.11: *Resultados de macroF_1 para a coleção Reuters com categorias e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

hipótese de que esta restrição traria melhoria à CAT quando comparado com a não utilização desta restrição, como levantado pelo Problema de Pesquisa 3.

Comparativo entre medidas de seleção de características

Conforme explicitado na subseção anterior, o método CRC aplicado às categorias somente apresentou resultados satisfatórios quando utilizado em conjunto com a medida FT1C. Diante do exposto, utilizamos apenas o método sem restrição de classe, SRC, nos experimentos de comparação entre as 4 medidas de seleção de características utilizadas.

Para a coleção Reuters, a metodologia FT1C demonstrou maior estabilidade, apresentando menor depreciação dos resultados ao se aumentar a porcentagem

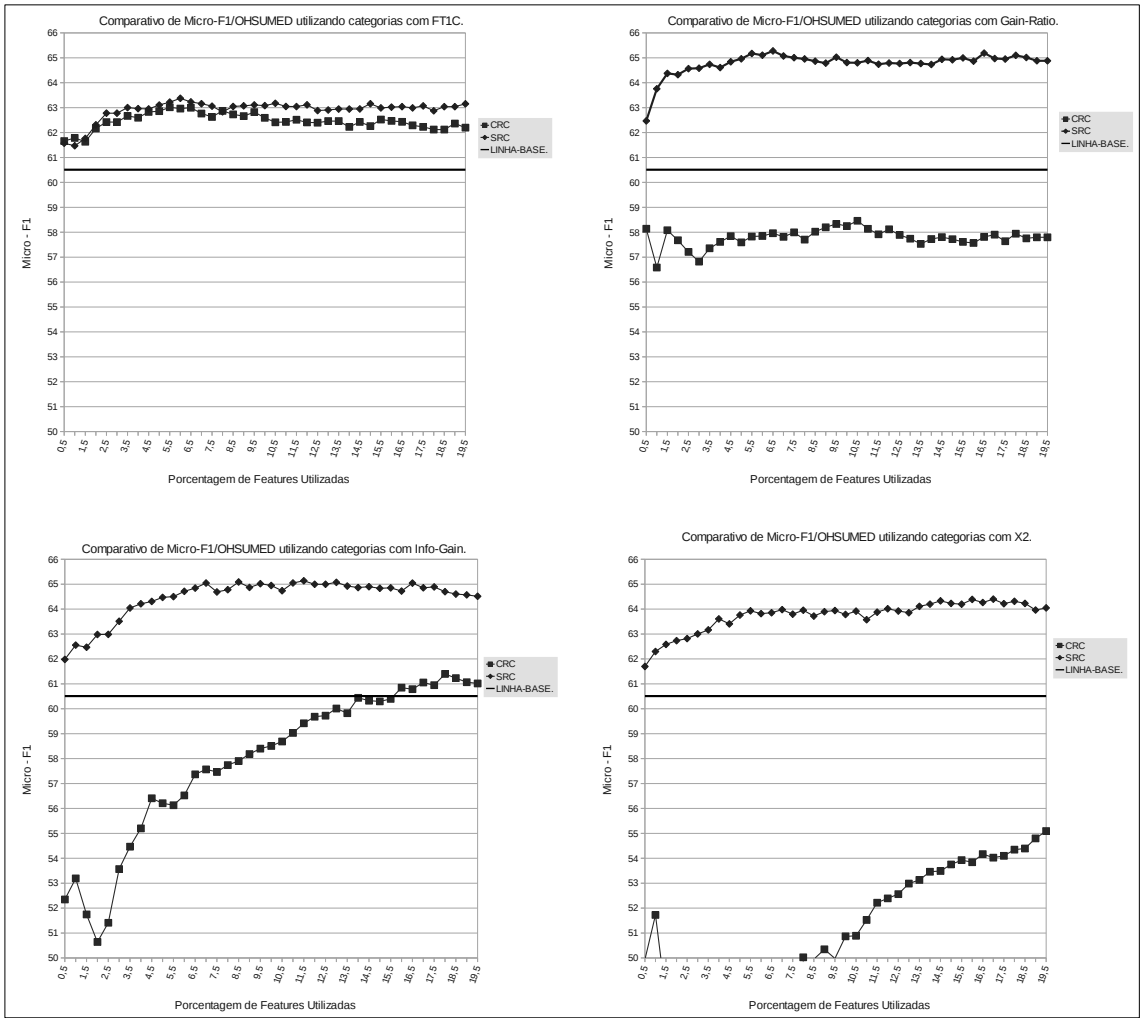

Figura 4.12: *Resultados de $microF_1$ para coleção Ohsumed com categorias e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

de categorias utilizadas, como pode ser constatado nos gráficos das Figuras 4.10 e 4.11.

As Tabelas 4.11 e 4.12 apresentam os melhores e piores resultados de $macroF_1$ e $microF_1$, respectivamente, para as diversas medidas de seleção de características.

A partir destas Tabelas é possível visualizar que nenhuma medida conseguiu ganhos expressivos em $microF_1$. Ao se analisar os ganhos de $macroF_1$, por meio da Tabela 4.12, pode-se perceber que o maior ganho foi atingido pela medida de seleção de característica *Gain Ration*, com $3,75\%$. Entretanto, nesta mesma Tabela pode-se perceber que esta medida não demostrou estabilidade com o aumento da quantidade de características utilizadas, ao passo que a medida FT1C obteve ganhos menores,

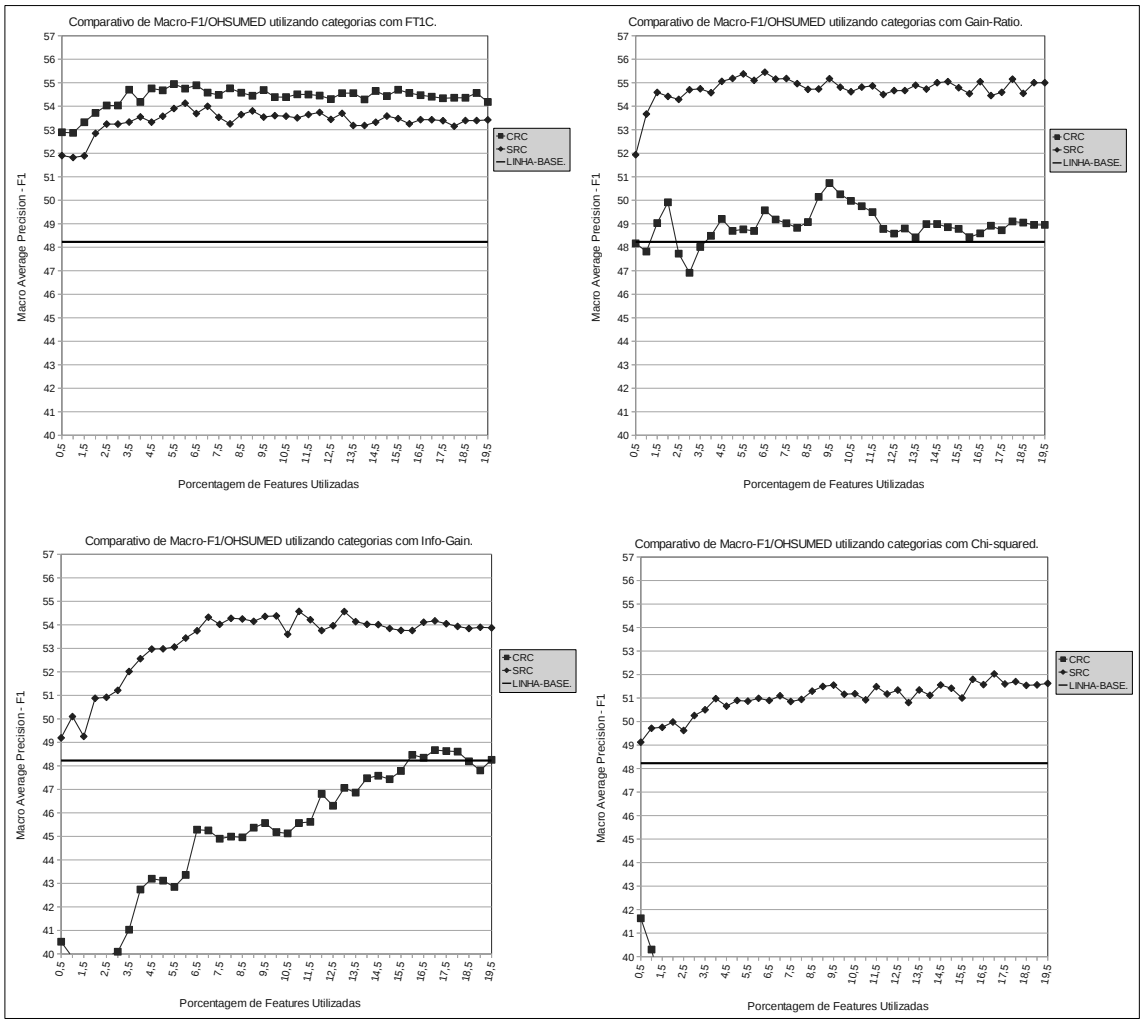

Figura 4.13: *Resultados de macroF_1 para coleção Ohsumed com categorias e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

Resultados de microF_1 para Reuters expandida com categorias/SRC							
Med. de Seleção	Linha Base	microF_1 MÁX	ganho	microF_1 MÍN.	ganho	d.p.	s.e.
FT1C	93,69	94,12	0,46%	93,49	-0,21%	0,144	▲
Gain Ratio		94,06	0,39%	93,12	-0,61%	0,327	■
Info Gain		94,08	0,41%	93,17	-0,55%	0,227	▲
Chi-Squared		93,89	0,21%	93,27	-0,45%	0,139	▲
100% dos Candidatos		92,86	-0,89%	92,86	-0,89%		

Tabela 4.11: *Resultados máximos e mínimos de microF_1 para Reuters expandida com categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

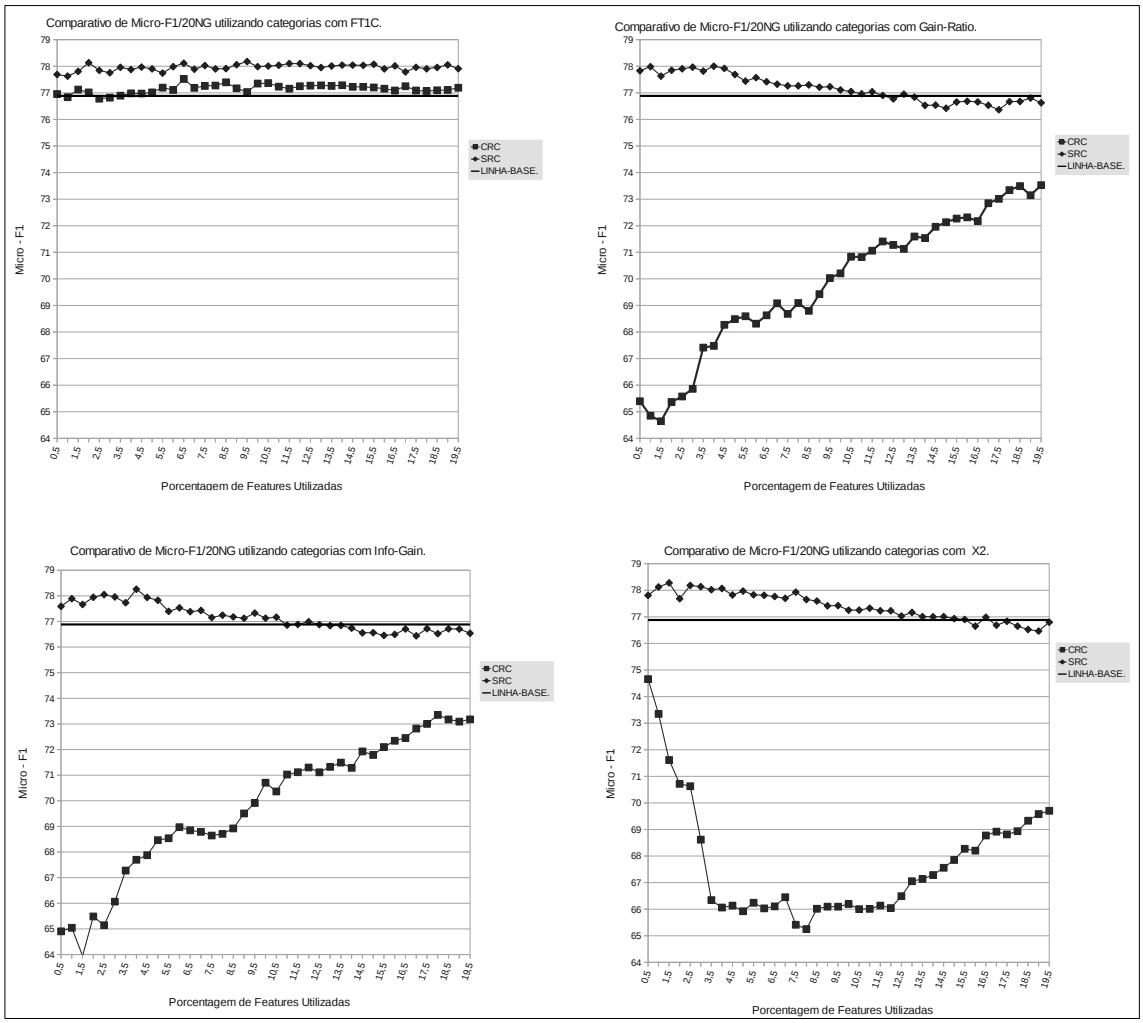

Figura 4.14: *Resultados de $microF_1$ para coleção 20NG com categorias e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

Resultados de $macroF_1$ para Reuters expandida com categorias/SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	66,24	67,66	2,14%	65,15	-1,65%	0,492	▲
Gain Ratio		68,73	3,75%	64,86	-2,08%	1,170	▲
Info Gain		67,57	2,00%	65,03	-1,82%	0,648	●
Chi-Squared		68,06	2,74%	64,16	-3,13%	0,946	●
100% dos Candidatos		65,24	-1,51%	65,24	-1,51%		

Tabela 4.12: *Resultados máximos e mínimos de $macroF_1$ para Reuters expandida com categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

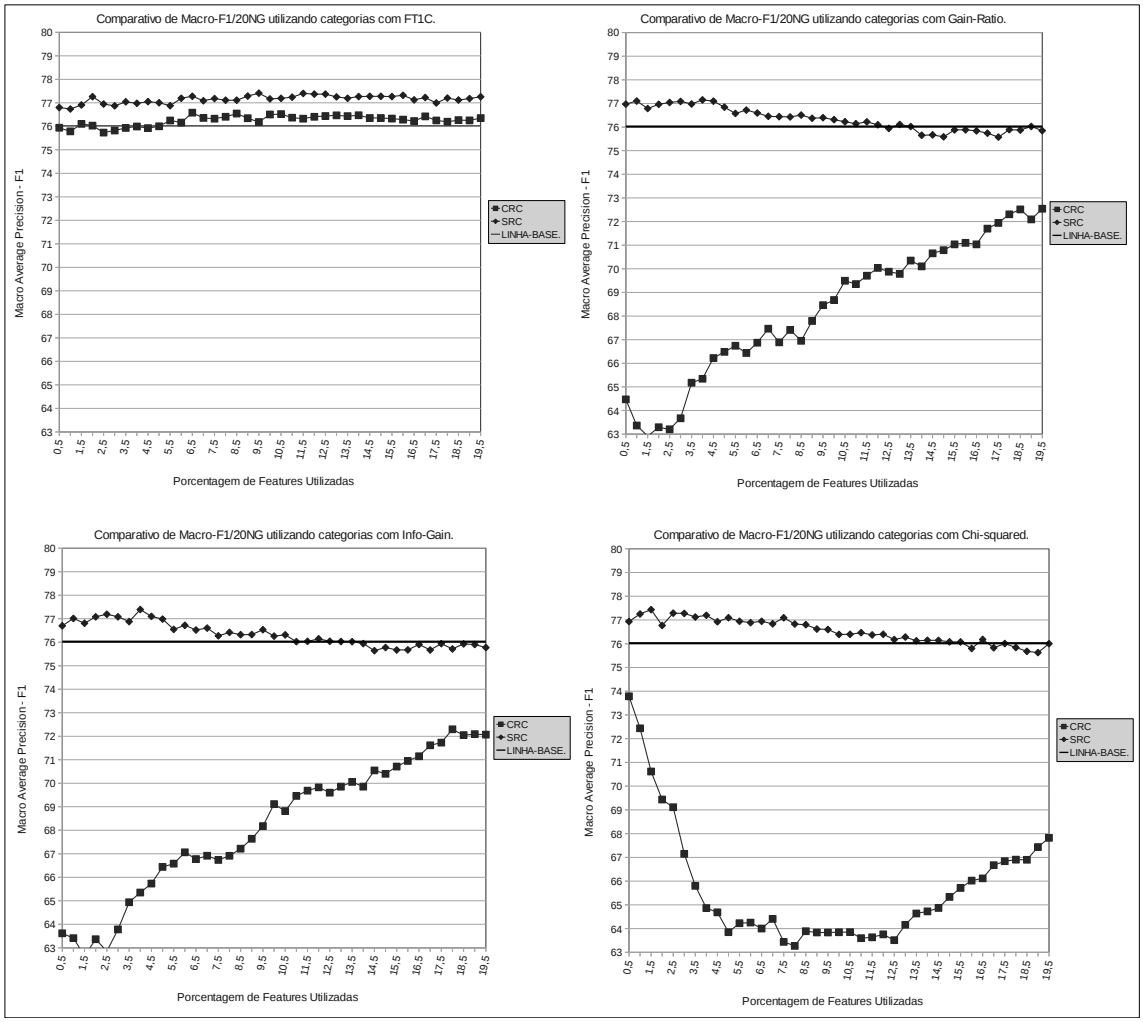

Figura 4.15: *Resultados de $macroF_1$ para coleção 20NG com categorias e medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

porém menos susceptíveis à degradação a medida que se aumenta a porcentagem de características utilizadas.

A expansão da coleção Ohsumed, utilizando categorias, apresentou ganhos consideráveis utilizando a metodologia SRC, como pode ser visto nos gráficos das Figuras 4.12 e 4.13, as quais apresentam os resultados de $microF_1$ e $macroF_1$ respectivamente.

A medida de seleção de característica *Gain Ratio* apresentou boa estabilidade e ótimos ganhos, tanto em $microF_1$ quanto $macroF_1$. A medida *Gain Ratio* alcançou ganho de $7,87\%$ em $microF_1$, explicitado na Tabela 4.13, e $14,97\%$ em $macroF_1$, visualizado na Tabela 4.14, ao passo que a medida FT1C conseguiu apenas $4,73\%$ e $12,23\%$ respectivamente.

Resultados de $microF_1$ para Ohsumed expandida com categorias/SRC							
Med. de Seleção	Linha Base	$microF_1$ MÁX	ganho	$microF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	60,50	63,37	4,73%	61,46	1,58%	0,420	▲
Gain Ratio		65,27	7,87%	62,46	3,23%	0,470	▲
Info Gain		65,13	7,65%	61,98	2,43%	0,815	▲
Chi-Squared		64,40	6,43%	61,69	1,96%	0,315	▲
100% dos Candidatos		63,13	4,35%	63,13	4,35%		

Tabela 4.13: *Resultados máximos e mínimos de $microF_1$ para Ohsumed expandida com categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

Resultados de $macroF_1$ para Ohsumed expandida com categorias/SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	48,22	54,12	12,23%	51,81	7,44%	0,505	▲
Gain Ratio		55,44	14,97%	51,93	7,68%	0,572	▲
Info Gain		54,57	13,15%	49,19	1,99%	1,459	▲
Chi-Squared		52,02	7,87%	49,12	1,85%	0,654	▲
100% dos Candidatos		51,77	7,36%	51,77	7,36%		

Tabela 4.14: *Resultados máximos e mínimos de $macroF_1$ para Ohsumed expandida com categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

Os experimentos com categorias no enriquecimento da Ohsumed tornam explícita a excelente capacidade da Wikipédia em fornecer bons candidatos para esta coleção. Mesmo sem a utilização de seleção de características, ou seja 100% das categorias candidatas, tais elementos proporcionaram ganhos de $4,35\%$ de $microF_1$ e $7,36\%$ de $macroF_1$, o que não ocorre para as outras coleções ao se utilizar o mesmo tipo de características (categorias).

Para o enriquecimento por categorias, a coleção 20Newsgroups tem seus resultados de $microF_1$ e $macroF_1$ apresentados nos gráficos das Figuras 4.14 e 4.15. A partir destes gráficos é possível constatar a boa estabilidade do método FT1C para esta coleção. As Tabelas 4.15 e 4.16 expõem os valores mínimos e máximos de $microF_1$ e $macroF_1$ para esta coleção. Apesar de não apresentar os melhores valores de $microF_1$ e $macroF_1$, a medida FT1C não apresenta degradação na classificação para a faixa de porcentagens analisadas. Por outro lado, as medidas *Information Gain*, *Gain Ratio* e *Chi-squared* apresentam valores mínimos abaixo da

Resultados de $microF_1$ para 20Newsgroups expandida com categorias/SRC							
Med. de Seleção	Linha Base	$microF_1$ MÁX	ganho	$microF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	76,88	78,17	1,68%	77,62	0,96%	0,125	▲
Gain Ratio		78,00	1,45%	76,36	-0,67%	0,507	▲
Info Gain		78,25	1,79%	76,43	-0,57%	0,519	▲
Chi-Squared		78,27	1,81%	76,46	-0,54%	0,519	▲
100% dos Candidatos		75,93	-1,24%	75,93	-1,24%		

Tabela 4.15: *Resultados máximos e mínimos de $microF_1$ para 20Newsgroups expandida com categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

Resultados de $macroF_1$ para 20Newsgroups expandida com categorias/SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	76,02	77,41	1,82%	76,73	0,93%	0,168	▲
Gain Ratio		77,15	1,48%	75,57	-0,58%	0,490	▲
Info Gain		77,39	1,79%	75,64	-0,50%	0,500	▲
Chi-Squared		77,43	1,85%	75,62	-0,52%	0,509	▲
100% dos Candidatos		75,15	-1,14%	75,15	-1,14%		

Tabela 4.16: *Resultados máximos e mínimos de $macroF_1$ para 20Newsgroups expandida com categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

linha base. É importante salientar que todas as medidas de seleção de características possuem desempenhos melhores que os obtidos com a expansão utilizando 100% das características candidatas.

Além das 4 medidas de seleção de características, também testou-se o desempenho da CAT ao se expandir as coleções Reuters, Ohsumed, e 20Newsgroups utilizando 100% das categorias candidatas, ou seja, sem realizar qualquer seleção de características ou restrição de expansão. A inserção de 100% das características do tipo categoria também foi utilizada por Wang et al. [64], onde os autores reportam os bons resultados encontrados.

Neste contexto, o presente trabalho demonstra que a utilização de métodos de seleção de características para as categorias, juntamente com a expansão sem restrição de classe (SRC), potencializam a eficácia da CAT quando comparada ao método de expansão com 100% dos candidatos. Como pode ser visto, a coleção Ohsumed alcançou ganhos expressivos de $microF_1$, que saltaram de $4,35\%$ para

$7,87\%$, e de $macroF_1$ que passaram de $7,36\%$ para $14,97\%$, após a utilização de seleção de característica *Gain Ratio* (Tabelas 4.13 e 4.14).

Diante do que foi exposto, novamente é possível responder positivamente ao Problema de Pesquisa 1, também no uso de categorias, contribuindo para a confirmação da hipótese relacionada a este problema quanto à melhoria do uso de categorias provindas da Wikipédia na expansão de documentos.

Apesar da medida FT1C não obter os resultados mais altos na coleção Ohsumed ao se utilizar categorias na expansão de documentos, esta medida demostrou resultados competitivos com as demais medidas, além de uma maior estabilidade ao se aumentar a porcentagem de características utilizadas no processo de expansão. Desse modo, os resultados com categorias também contribuem para a confirmação da hipótese relacionada ao Problema de Pesquisa 2.

4.5.3 Expansão com w-conceitos + categorias diretas

Assim como foi realizado nas Subseções 4.5.1 e 4.5.2 para as abordagens com w-conceitos e com categorias, esta subseção analisa o efeito derivado da união dos conjuntos de w-conceitos candidatos juntamente com o conjunto de categorias candidatas. As próximas linhas discutem os resultados obtidos para esta metodologia. Nesta abordagem, um conjunto de características eleitas contém somente os w-conceitos e/ou categorias mais bem valoradas pelas medidas de seleção de características FT1C, *Information Gain*, *Gain Ratio* ou *Chi-squared*.

Comparativo entre CRC e SRC

Assim como nas outras abordagens, a metodologia SRC se mostrou superior à metodologia CRC para todas as medidas de seleção de características aplicadas à união de w-conceitos com categorias diretas, provindos da coleção Reuters, como pode ser visto na Figura 4.16 para $microF_1$ e na Figura 4.17 para $macroF_1$. O mesmo ocorreu também para a coleção 20Newsgroups, como pode ser visto nas Figuras 4.18 para $microF_1$ e 4.19 para $macroF_1$.

A metodologia de expansão CRC obteve ganhos estáveis apenas para a metodologia FT1C, com destaque para a $macroF_1$ da coleção Ohsumed, como pode ser visto na Figura 4.20. Entretanto, a medida CRC não se mostrou portável para outras metodologias de seleção de características, além de baixos ganhos em $microF_1$, como visto na Figura 4.12. Dessa forma, assim como nas abordagens com w-conceitos e com categorias, a hipótese do Problema de Pesquisa 3 não pôde ser confirmada também para utilização de w-conceitos e categorias.

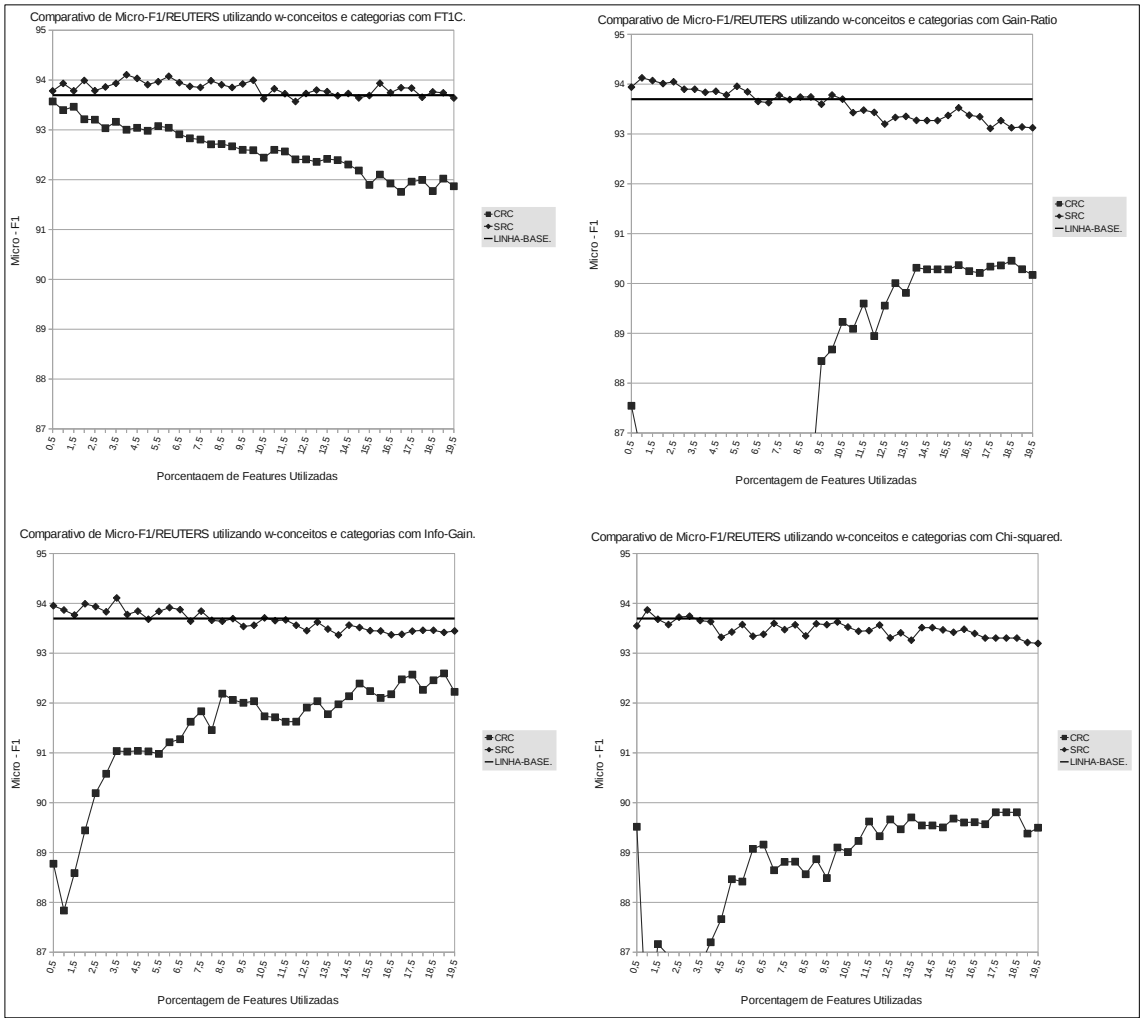

Figura 4.16: *Resultados de $microF_1$ para coleção Reuters com w-conceitos e categorias utilizando medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

Comparativo entre medidas de seleção de características

Seguindo a mesma abordagem das Subseções 4.5.1 e 4.5.2 o método CRC aplicado a w-conceitos unidos com categorias, somente apresentou resultados satisfatórios quando utilizado em conjunto com a medida FT1C. Diante do exposto, utilizamos apenas o método sem restrição de classe SRC nos experimentos de comparação entre as 4 medidas de seleção de características utilizadas.

Para a coleção Reuters não houve melhorias significativas em termos de $microF_1$. Entretanto, é possível notar uma melhora nos valores de $macroF_1$, como pode ser visto nos gráficos das Figuras 4.17 e 4.16 e nas Tabelas 4.17 e 4.18.

Como exemplo, pode-se comparar o ganho obtido ao se aplicar a medida *Gain Ratio* utilizando apenas categorias, sendo que nesta configuração o ganho foi

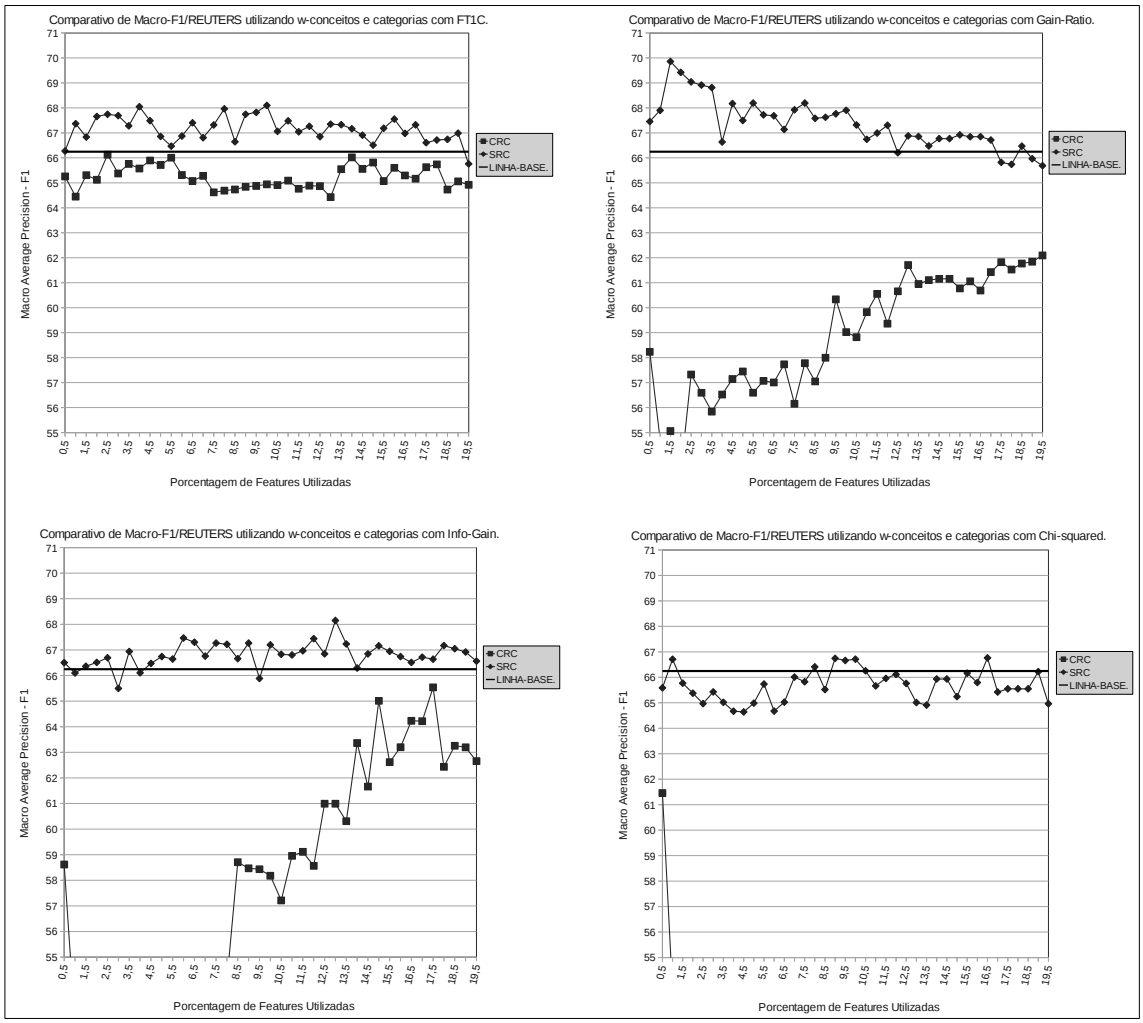

Figura 4.17: *Resultados de macroF_1 para coleção Reuters com w-conceitos e categorias utilizando medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

Resultados de microF_1 para Reuters expandida com w-conc. e cat. /SRC							
Med. de Seleção	Linha Base	microF_1 MÁX	ganho	microF_1 MÍN.	ganho	d.p.	s.e.
FT1C	93,69	94,10	0,43%	93,56	-0,13%	0,129	▲
Gain Ratio		94,12	0,45%	93,11	-0,62%	0,297	■
Info Gain		94,10	0,43%	93,36	-0,35%	0,197	●
Chi-Squared		93,86	0,18%	93,19	-0,53%	0,150	▲
100% dos Candidatos		93,01	-0,73%	93,01	-0,73%		

Tabela 4.17: *Resultados máximos e mínimos de microF_1 para Reuters expandida com w-conceitos + categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

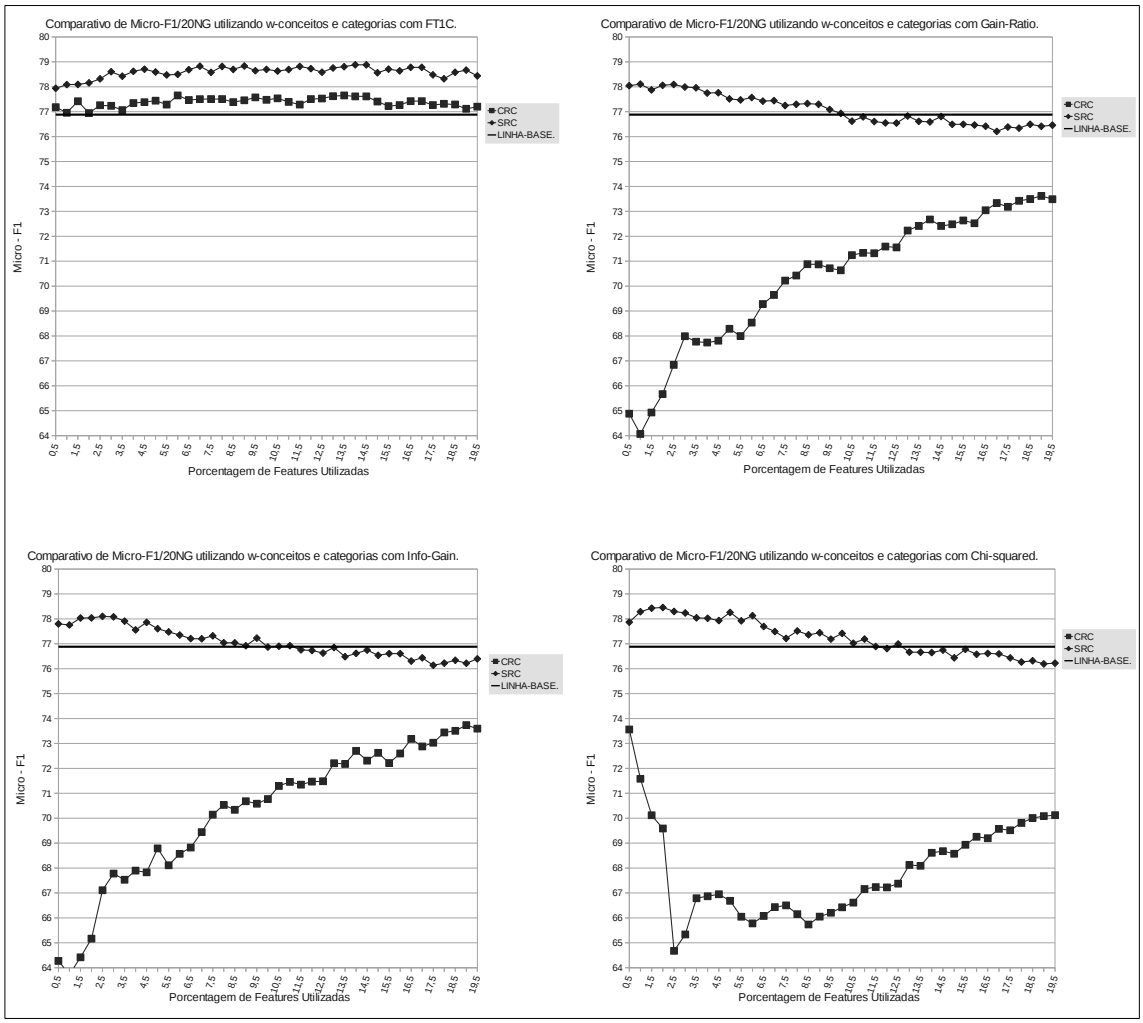

Figura 4.18: *Resultados de $microF_1$ para coleção 20NG com w-conceitos e categorias utilizando medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

Resultados de $macroF_1$ para Reuters expandida com w-conc. e cat./SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	66,24	68,09	2,79%	65,75	-0,73%	0,459	▲
Gain Ratio		69,85	5,45%	65,68	-0,84%	0,966	▲
Info Gain		68,14	2,87%	65,49	-1,13%	0,485	▲
Chi-Squared		66,76	0,77%	64,64	-2,42%	0,611	▲
100% dos Candidatos		64,28	-2,96%	64,28	-2,96%		

Tabela 4.18: *Resultados máximos e mínimos de $macroF_1$ para Reuters expandida com w-conceitos + categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

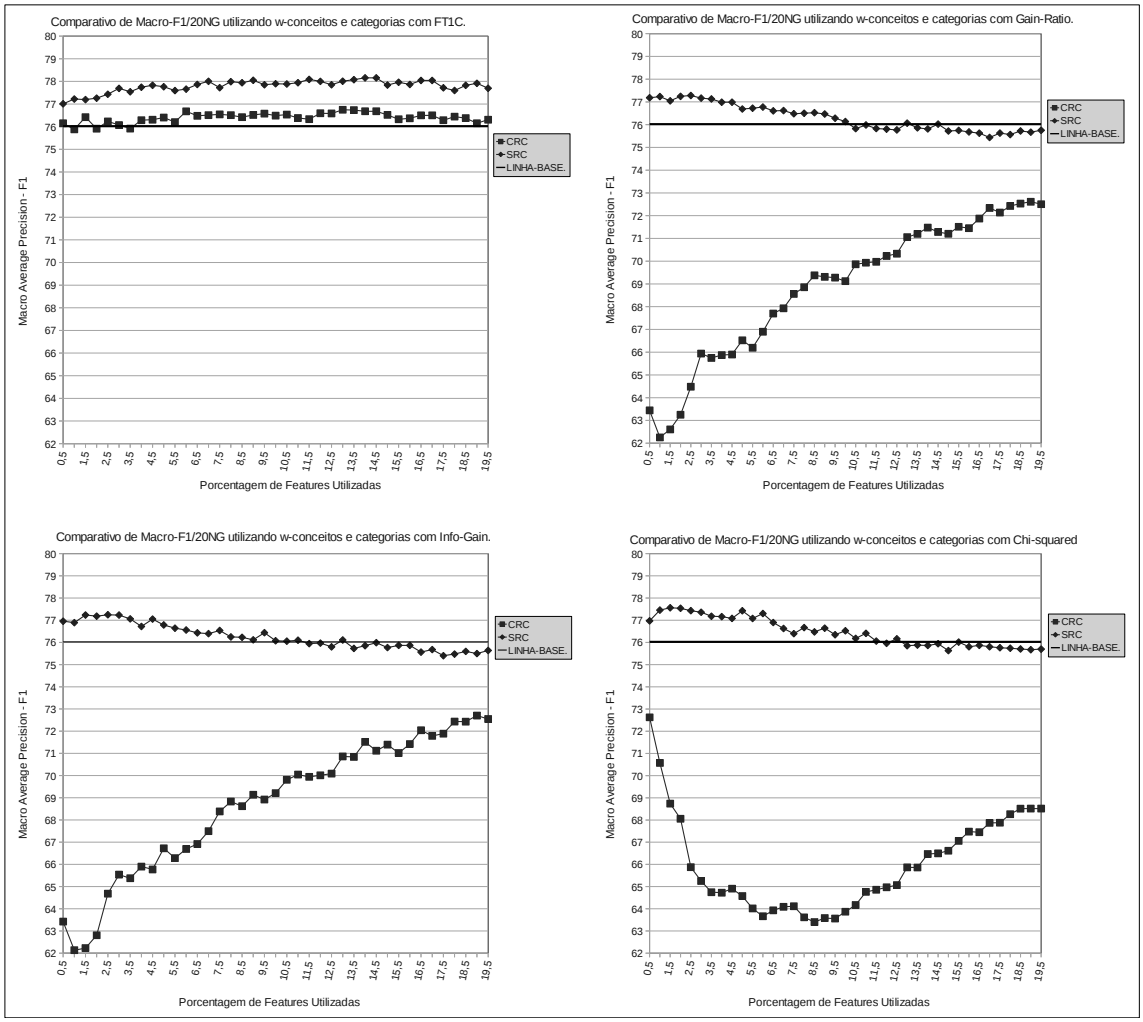

Figura 4.19: *Resultados de $macroF_1$ para coleção 20NG com w-conceitos e categorias utilizando medidas de seleção de características FT1C, Gain-Ratio, Info-Gain e Chi-Squared.*

de $3,75\%$, ao passo que a mesma medida com a utilização da união de w-conceitos e categorias, obteve ganho de $5,45\%$.

Os gráficos das Figuras 4.20 e 4.21 apresentam os resultados da expansão da coleção Ohsumed a partir de w-conceitos e categorias. Por meio destes gráficos é possível observar que, com a união destes dois tipos de características, obteve-se a melhoria da métrica $microF_1$ apenas para a medida FT1C, e em $macroF_1$ obtiveram-se melhorias para as medidas FT1C e *Gain Ratio*. O melhor resultado de $macroF_1$ obtido para esta coleção foi de $15,08\%$ pela medida *Gain Ratio*, como pode ser visto na Tabela 4.19. A mesma medida também obteve o melhor resultado em $microF_1$ na coleção Ohsumed, com um valor de $7,67\%$, como pode ser visto na Tabela 4.20 . Além disso, a medida FT1C também se mostrou estável às variações

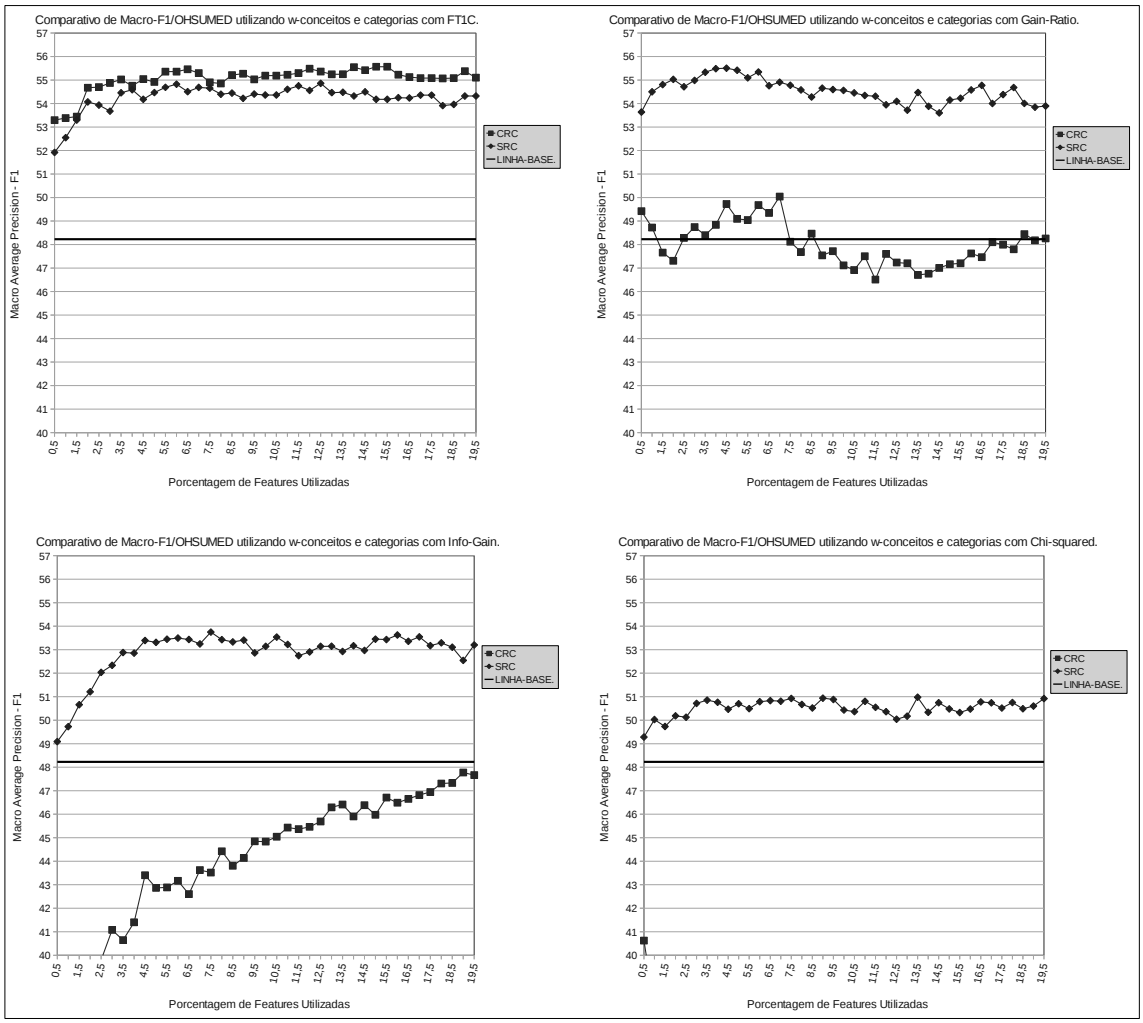

Figura 4.20: *Resultados de $macroF_1$ para coleção Ohsumed com w-conceitos e categorias utilizando 4 medidas de medidas de seleção de características.*

Resultados de $macroF_1$ para Ohsumed expandida com w-conc. e cat./SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	48,22	54,85	13,74%	51,91	7,65%	0,574	▲
Gain Ratio		55,50	15,08%	53,59	11,13%	0,517	▲
Info Gain		53,74	11,44%	49,09	1,78%	1,030	▲
Chi-Squared		50,97	5,70%	49,28	2,18%	0,354	▲
100% dos Candidatos		50,64	5,02%	50,64	5,02%		

Tabela 4.19: *Resultados máximos e mínimos de $macroF_1$ para Ohsumed expandida com w-conceitos + categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

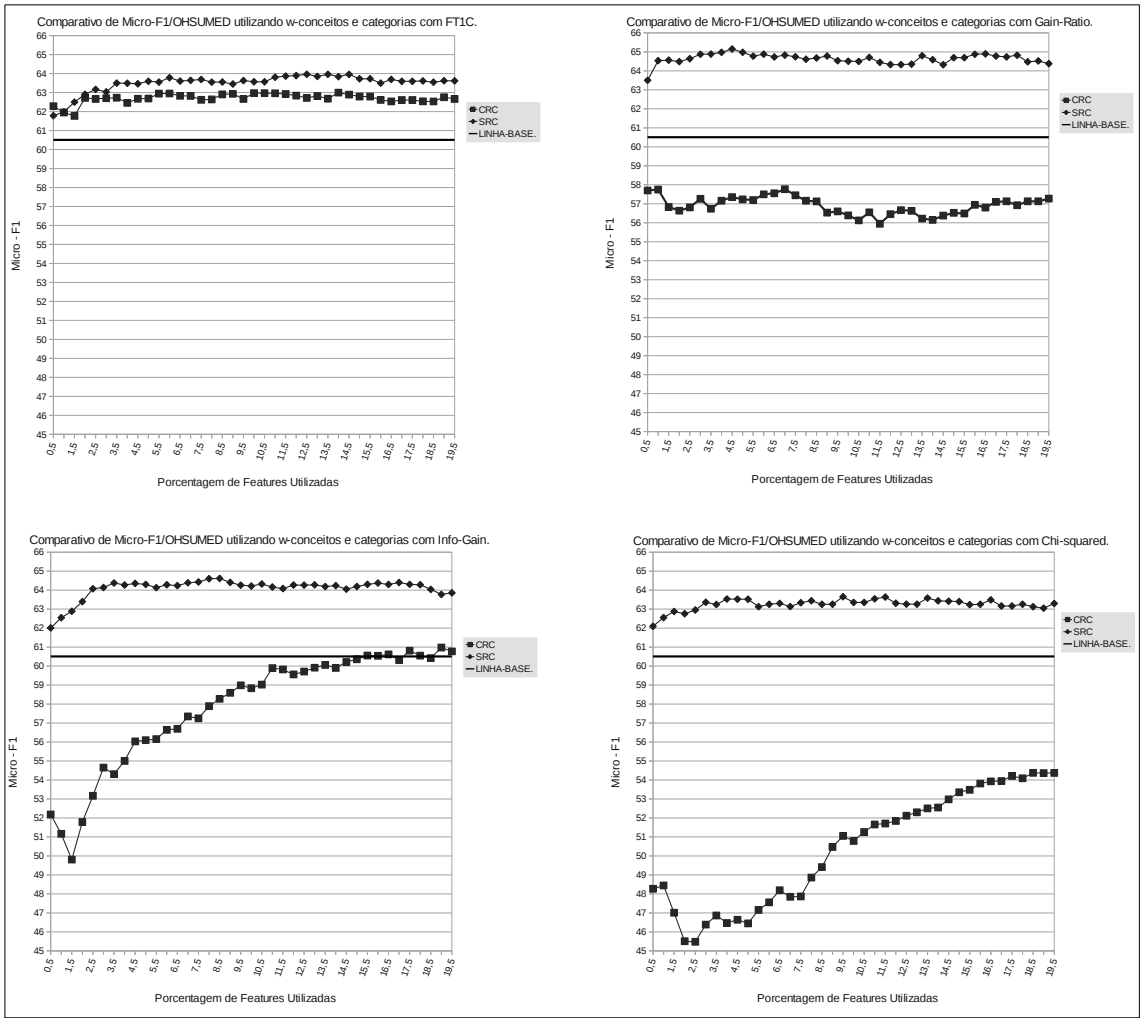

Figura 4.21: *Resultados de $microF_1$ para coleção Ohsumed com w-conceitos e categorias utilizando 4 medidas de medidas de seleção de características.*

de porcentagens de características utilizadas no enriquecimento desta coleção.

Para a coleção 20Newsgroups, a medida de seleção de característica mais estável foi a FT1C, sendo que as demais medidas apresentaram maior queda de desempenho com o aumento do número de características utilizadas, como pode ser visto nos gráficos das Figuras 4.18 e 4.19. O melhor valor de $microF_1$ para esta coleção foi de $2,59\%$, obtido por meio da medida FT1C, como exposto na Tabela 4.21.

A medida FT1C também obteve o maior valor de $macroF_1$ para esta coleção, $2,80\%$.(Tabela 4.22).

Neste contexto, o presente trabalho demonstra que a utilização de métodos de seleção de características aplicadas à w-conceitos em união com categorias, juntamente com a expansão sem restrição de classe (SRC), potencializam a eficácia

Resultados de $microF_1$ para Ohsumed expandida com w-conc. e cat./SRC							
Med. de Seleção	Linha Base	$microF_1$ MÁX	ganho	$microF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	60,50	63,96	5,71%	61,78	2,10%	0,480	▲
Gain Ratio		65,15	7,67%	63,50	4,95%	0,275	▲
Info Gain		64,61	6,79%	62,00	2,48%	0,532	▲
Chi-Squared		63,65	5,19%	62,09	2,62%	0,300	▲
100% dos Candidatos		62,27	2,93%	62,27	2,93%		

Tabela 4.20: *Resultados máximos e mínimos de $microF_1$ para Ohsumed expandida com w-conceitos + categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

Resultados de $microF_1$ para 20Newsgroups expandida com w-conc. e cat./SRC							
Med. de Seleção	Linha Base	$microF_1$ MÁX	ganho	$microF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	76,88	78,87	2,59%	77,93	1,37%	0,230	▲
Gain Ratio		78,10	1,59%	76,20	-0,87%	0,612	■
Info Gain		78,09	1,58%	76,13	-0,96%	0,591	●
Chi-Squared		78,45	2,04%	76,19	-0,89%	0,701	▲
100% dos Candidatos		76,27	-0,79%	76,27	-0,79%		

Tabela 4.21: *Resultados máximos e mínimos de $microF_1$ para 20Newsgroups expandida com w-conceitos + categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

Resultados de $macroF_1$ para 20Newsgroups expandida com w-conc. e cat./SRC							
Med. de Seleção	Linha Base	$macroF_1$ MÁX	ganho	$macroF_1$ MÍN.	ganho	d.p.	s.e.
FT1C	76,02	78,15	2,80%	77,00	1,29%	0,275	▲
Gain Ratio		77,28	1,65%	75,43	-0,77%	0,591	▲
Info Gain		77,24	1,60%	75,40	-0,82%	0,558	■
Chi-Squared		77,55	2,02%	75,63	-0,51%	0,642	▲
100% dos Candidatos		75,47	-0,72%	75,47	-0,72%		

Tabela 4.22: *Resultados máximos e mínimos de $macroF_1$ para 20Newsgroups expandida com w-conceitos + categorias e SRC, com 4 medidas de feature selection confrontadas com a utilização de 100% dos candidatos.*

da CAT quando comparada ao método de expansão que utiliza 100% dos candidatos, respondendo positivamente ao Problema de Pesquisa 1.

Os resultados apresentados para todas as coleções confirmam a hipótese de competitividade da medida FT1C na seleção de características, relativa ao Problema de Pesquisa 2, também para a expansão de documentos utilizando w-conceitos em união com categorias.

4.5.4 Análise geral dos resultados

Em linhas gerais, as medidas de seleção de características, combinadas com o método (SRC), se mostraram fundamentais para o processo de expansão de documentos com características extraídas da Wikipédia.

Como visto na Seção 4.5.1, a expansão das coleções de dados utilizando 100% dos w-conceitos trouxe degradação dos resultados na maioria dos casos. Entretanto, os experimentos puderam comprovar que o uso de medidas de seleção das características boas discriminadoras de classe conseguem mudar este quadro de degradação dos resultados criando um cenário em que tais características se tornam importantes para a melhoria da CAT, contrariando o que foi exposto por [64].

Assim como ocorre para os w-conceitos, o uso de categorias da Wikipédia também se mostrou propenso a melhorias com o uso das medidas de seleção de características, como exposto na Seção 4.5.2. Apesar dos bons resultados apresentados por [64] no uso deste tipo de característica provinda da Wikipédia, os experimentos conseguiram comprovar que tais resultados podem ser melhorados, sendo que para nosso ambiente experimental conseguimos elevar consideravelmente os patamares de ganhos para todas as coleções, tanto em $microF_1$ quanto em $macroF_1$. O mesmo desempenho é obtido para as características compostas da união entre w-conceitos e categorias, tratado na Seção 4.5.2. Os experimentos apresentados nas Subseções 4.5.1, 4.5.2 e 4.5.3, respondem positivamente ao Problema de Pesquisa 1, demostrando a importância da etapa de seleção de características durante o processo de expansão de documentos.

É importante observar que o uso das medidas de seleção de características contribuiu de maneira mais significativa para a coleção Ohsumed do que para as demais coleções analisadas. Para compreender tal comportamento deve-se observar que a Ohsumed é uma coleção reconhecidamente mais difícil de se classificar, o que pode ser comprovado observando-se a linha-base desta coleção, bem abaixo das demais coleções. Coleções que já apresentam uma linha-base elevada estão menos susceptíveis a melhorias maiores no processo de classificação.

A média $macroF_1$ notadamente apresentou os melhores ganhos quando comparados aos obtidos pela média $microF_1$. Ao se analisar os resultados apresentados

nas Subseções 4.5.1, 4.5.2 e 4.5.3 pode-se verificar que as coleções mais desbalanceadas Reuters e Ohsumed apresentaram as maiores diferenças de ganhos entre $macroF_1$ e $microF_1$, enquanto que a coleção mais balanceada 20Newsgroups apresentou ganhos mais equilibrados nas duas médias. Sabendo que a média $macroF_1$ estabelece igual importância para todas as categorias de uma coleção, como explanado na Subseção 2.4.4, coleções desbalanceadas tendem a apresentar menor desempenho em categorias menores por possuírem uma menor quantidade de exemplos para o aprendizado do classificador. Neste contexto, a expansão de documentos de categorias menores utilizando-se de características mais discriminativas possibilita a geração de melhores classificadores para estas classes, impactando positivamente de maneira mais elevada na média $macroF_1$ do que na média $microF_1$.

Os experimentos com a medida de seleção proposta FT1C comprovam sua capacidade em selecionar características boas descriminadoras de classe, se apresentando, portanto, como opção competitiva quando comparada com medidas já consagradas na literatura como *Information Gain*, *Gain Ratio* e *Chi-squared* (X^2).

A Tabela 4.23 relaciona as medidas de seleção de características que obti-

Melhores desempenhos das medidas de Seleção de Características						
Tipo de Característica	Reuters-21578		Ohsumed		20 newsgroups	
	MICRO	MACRO	MICRO	MACRO	MICRO	MACRO
w-conceitos	Info. Gain	FT1C	FT1C	FT1C	FT1C	FT1C
categorias	FT1C	Gain Ratio	Gain Ratio	Gain Ratio	FT1C	FT1C
wc + cat.	Gain Ratio	Gain Ratio	Gain Ratio	Gain Ratio	FT1C	FT1C

Tabela 4.23: *Comparativo entre os melhores desempenhos das medidas de seleção de características para todas as abordagens.*

veram melhor desempenho para cada uma das abordagens analisadas, tanto para $microF_1$ como para $macroF_1$. Esta Tabela reflete os bons resultados apresentados pela medida de seleção de característica FT1C, proposta pelo presente trabalho.

Dentre as medidas analisadas, a FT1C apresentou o melhor desempenho ao se trabalhar com w-conceitos em todas as coleções, mesmo a medida *Information Gain* tendo apresentado o maior valor de $microF_1$ na coleção Reuters, a FT1C obteve valores competitivos e maior estabilidade.

A medida FT1C também demonstrou estabilidade ao se trabalhar com categorias, ao passo que a medida *Gain Ratio* demostrou ganhos mais elevados para a coleção Ohsumed ao se utilizar este tipo de característica. O mesmo não ocorreu para a coleção 20Newsgroups, para a qual a medida FT1C demostrou maior estabilidade e ganhos mais elevados. Ainda utilizando categorias, a medida FT1C não apresentou os maiores ganhos de $microF_1$, mas também se mostrou competitiva apresentando

menores quedas nos ganhos à medida que utiliza-se uma maior porcentagem de características.

O mesmo comportamento apresentado para categorias se repetiu para o tipo de característica formada pela união de w-conceitos e categorias. Este fato se deve à forte capacidade das categorias em fornecer características boas discriminadoras de classes, influenciando substancialmente na formação dos conjuntos de características eleitas.

Os experimentos demostraram que mesmo para as abordagens em que a medida FT1C não apresentou os melhores resultados, a mesma se portou de forma competitiva. Com isso a hipótese relacionada ao Problema de Pesquisa 2 é confirmada para todas as coleções analisadas, tanto utilizando w-conceitos quanto categorias ou a combinação de ambos.

Todavia, ao se analisar o comportamento da metodologia CRC, pode-se verificar que se em uma determinada categoria c_j, uma característica eleita t_i não alcança valor local para $f(t_i, c_j)$ maior ou igual aos valores de $f_{global}(t_i)$ das k características eleitas, então esta característica não será usada para enriquecer nenhum documento de treino em c_j. Entretanto, como t_i está entre as características eleitas, a mesma será utilizada para enriquecer todos os documentos de teste em que ela ocorra, conforme metodologia estabelecida para este conjunto, como pode ser visto na Figura 3.3. O uso de metodologias diferentes entre treino e teste acaba por gerar um modelo de classificador que não se adéqua bem ao conjunto de teste, provocando erros na classificação. O problema descrito acima não ocorre para a metodologia SRC já que tanto no conjunto de treino quanto no de teste as características eleitas são utilizadas em qualquer que seja o documento em que a mesmas apareçam.

Os experimentos demostraram que a utilização da metodologia de restrição CRC não possibilitou melhorias estáveis ao processo de seleção de características, apresentando apenas ganhos isolados. Portanto, a hipótese relacionada ao Problema de Pesquisa 3 é refutada para expansão de w-conceitos, categorias e união destes, provindos da Wikipédia.

No próximo capítulo são apresentadas as conclusões gerais a cerca do presente trabalho, além de expor os possíveis trabalhos futuros relacionados ao tema da pesquisa.

CAPÍTULO 5

Conclusão

Neste trabalho estudou-se a expansão de documentos utilizando-se de relações de sinonímia de conceitos (w-conceitos) e categorias extraídos da Wikipédia. Analisou-se a melhoria deste processo adicionando uma etapa de seleção de características boas discriminadoras de classes. Ademais, foi analisado o desempenho da adição de uma restrição de classe para a utilização das relações de sinonímia e categorias selecionadas na etapa anterior.

Primeiramente, foi avaliado se a aplicação de um método de seleção de características consegue melhorar a eficácia da utilização das relações de sinonímia (w-conceitos) e de categorias provindas da Wikipédia durante o processo de expansão de documentos, reduzindo a inserção de ruídos e potencializando a adição de características boas discriminadoras de classes.

Durante a expansão de documentos utilizando apenas w-conceitos da Wikipédia, o uso de medidas de seleção de características demonstrou ser de fundamental importância. Como pôde ser visto na Subseção 4.5.1, os experimentos comprovam que o uso de medidas de seleção de w-conceitos bons discriminadores de classe conseguem mudar o cenário de degradação apresentado por [64], tornando tais características importantes para a melhoria da CAT. Os resultados para esta abordagem obtiveram ganhos máximos de $3,58\%$ na medida $microF_1$ e $7,41\%$ na medida $macroF_1$, para a coleção Ohsumed.

Foi constatado que o uso de medidas de seleção de características também é útil para a melhoria da abordagem de expansão de documentos, utilizando apenas as categorias diretas dos w-conceitos da Wikipédia. Na Subseção 4.5.2 são apresentados os resultados para esta abordagem, na qual os ganhos máximos obtidos foram de $7,87\%$ na medida $microF_1$ e $14,97\%$ na medida $macroF_1$, para a coleção Ohsumed.

Ao unir as características candidatas provindas de w-conceitos e categorias, a utilização de medidas de seleção de características se mostrou particularmente interessante, visto que a técnica conseguiu selecionar elementos bons discriminadores de classes, independentemente de sua origem (w-conceitos ou categorias), criando um grupo de w-conceitos eleitos de melhor qualidade ainda, com ganho na Ohsumed de

$15,08\%$ na medida $macroF_1$.

Os bons resultados nas três abordagens supramencionadas e discutidas nas Seções 4.5.1, 4.5.2, 4.5.3, solucionam o Problema de Pesquisa 1 e confirmam a hipótese relacionada a este problema.

A estratégia de extração de características da Wikipédia apresentada na Seção 3.1, a qual divide o documento em trechos de textos se mostrou importante para o processo de identificação de w-conceitos, uma vez que conseguiu-se diminuir a inserção de elementos ruidosos.

A medida proposta para a seleção de características denominada Fator de Tendência a uma Categoria (FT1C) demonstrou desempenho e estabilidades superiores na maioria das abordagens propostas. Como pode ser visto nas Subseções 4.5.1, 4.5.2 e 4.5.3, a medida FT1C apresenta-se como opção competitiva para o processo de seleção de w-conceitos e categorias provindas da Wikipédia, visto que em grande parte dos experimentos esta medida obteve maiores ganhos e melhor estabilidade que as medidas já consagradas na literatura *Information Gain*, *Gain Ratio* e *Chi-squared*. O Problema de Pesquisa 2 é, portanto, resolvido e a hipótese relacionada ao mesmo foi confirmada.

A verificação da eficácia da CAT ao se utilizar as medidas de avaliação de termos FT1C, *Information Gain*, *Gain Ratio*, *Chi-squared*, em conjunto com as metodologias CRC e SRC na expansão de documentos, como mostrado nos resultados experimentais das Subseções 4.5.1, 4.5.2, 4.5.3 permitem concluir que a utilização da restrição de classe CRC, apresentada na Seção 3.3 expõe resultados satisfatórios apenas quando utilizada em conjunto com a medida de seleção de características FT1C, e mesmo assim somente em casos específicos. A expansão de documentos sem a aplicação de restrição de classe (metodologia SRC) demonstrou maior flexibilidade e adaptação às diferentes medidas de seleção de características e não somente à FT1C, e melhores resultados na maioria das abordagens. Dessa forma, mesmo a metodologia CRC tendo sido útil ao trabalho de [14], a mesma não se mostrou como sendo a melhor opção na utilização com características provindas da Wikipédia. Diante do exposto, a hipótese relacionada ao Problema de Pesquisa 3 foi refutada pelos experimentos realizados.

A alta restritividade da medida FT1C se mostrou mais eficaz que as demais medidas ao se trabalhar com w-conceitos, notadamente mais ruidosos. A seleção de características em ambientes com muitos ruídos apresenta-se como um desafio a mais, visto que neste contexto, erros de seleção podem comprometer sensivelmente os resultados. O melhor desempenho da FT1C para este tipo de abordagem aponta que esta medida demonstra maior exatidão ao selecionar características boas discriminadoras de classes, embora não tenha apresentado os melhores ganhos com

categorias na coleção Ohsumed.

Os bons resultados obtidos para a Ohsumed, utilizando 100% das categorias, exibem que a Wikipédia possui excelente capacidade de fornecer este tipo de característica para esta coleção, apresentando uma baixa adição de ruídos. Neste contexto, pode-se concluir que o alto poder restritivo da medida FT1C acaba por limitar a inserção não só de elementos ruidosos, mas também de algumas características que poderiam auxiliar na classificação.

Trabalhos Futuros

Com o presente trabalho é possível visualizar novas possibilidades de pesquisas voltadas para a classificação de documentos. Em virtude disso, como trabalhos futuros, pretendemos realizar os seguintes estudos:

1. Propor um método de utilização dos textos âncoras dos artigos da Wikipédia, na expansão de documentos, visto que tais elementos também representam importantes relações de sinonímia contidas nesta enciclopédia.
2. Propor um método de utilização dos *links* existentes entre os artigos da Wikipédia,*in-links* e *out-links*, na determinação de relacionamento entre conceitos desta enciclopédia, e estudar seu uso na expansão de documentos.
3. Investigar se a utilização de medidas de seleção de características podem também potencializar os resultados obtidos com as abordagens acima.
4. Investigar se a medida FT1C apresenta-se como opção competitiva também para as futuras abordagens descritas acima.
5. Investigar o uso da medida de seleção de características FT1C como medida geral de seleção de características, podendo ser utilizada em diferentes contextos.
6. Propor melhorias na medida FT1C visando a melhorar seu desempenho em ambientes pouco ruidosos.

Referências Bibliográficas

[1] Amati, G.; D'Aloisi, D.; Giannini, V.; Ubaldini, F. **A Framework for Filtering News and Managing Distributed Data**. *Journal Of Universal Computer Science*, 3(8):1007–1021, 1997.

[2] Apté, C.; Damerau, F.; Weiss, S. M. **Automated learning of decision rules for text categorization**. *ACM Transactions on Information Systems*, 12(3):233–251, July 1994.

[3] Baeza-Yates, R.; Ribeiro-Neto, B. **Modern information retrieval**. ACM Press, New York, New York, USA, 1999.

[4] Bekkerman, R.; Allan, J. **Using Bigrams in Text Categorization**. *Department of Computer Science, University of Massachusetts, Amherst*, 1003(IR-408):1–10, 2003.

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

[6] Burges, C. J. C. **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[7] Carmel, D.; Roitman, H.; Zwerdling, N. **Enhancing cluster labeling using wikipedia**. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, p. 139, 2009.

[8] Chandrinos, K. V.; Androutsopoulos, I.; Paliouras, G.; Spyropoulos, C. D. **Automatic Web Rating: Filtering Obscene Content on the Web**. In: Borbinha, J. L.; Baker, T., editors, *Proceedings of ECDL00 4th European Conference on Research and Advanced Technology for Digital Libraries*, p. 403–406. Springer Verlag, Heidelberg, DE, 2000.

[9] Cheng, H.; Yan, X.; Han, J.; Hsu, C.-W. **Discriminative Frequent Pattern Analysis for Effective Classification**. *2007 IEEE 23rd International Conference on Data Engineering*, p. 716–725, 2007.

[10] Couto, T.; Ziviani, N.; Calado, P.; Cristo, M.; Gonçalves, M.; Moura, E. S.; Brandão, W. **Classifying documents with link-based bibliometric measures**. *Information Retrieval*, 13(4):315–345, 2009.

[11] Couto, T.; Cristo, M.; Gonçalves, M. A.; Calado, P.; Ziviani, N.; Moura, E.; Ribeiro-Neto, B. **A comparative study of citations and links in document classification**. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries - JCDL '06*, p. 75, 2006.

[12] Debole, F.; Sebastiani, F. **Supervised term weighting for automated text categorization**. In: *Proceedings of the 2003 ACM symposium on Applied computing - SAC '03*, p. 784, New York, New York, USA, 2003. ACM Press.

[13] Du, R.; Safavi-Naini, R.; Susilo, W. **Web filtering using text classification**. *The 11th IEEE International Conference on Networks ICON 2003*, p. 325–330, 2003.

[14] Figueiredo, F.; Rocha, L.; Couto, T.; Salles, T.; Gonçalves, M. A.; Meira Jr., W. **Word co-occurrence features for text classification**. *Information Systems*, 36(5):843–858, July 2011.

[15] Forman, G. **An extensive empirical study of feature selection metrics for text classification**. *The Journal of Machine Learning Research*, 3:1289–1305, 2003.

[16] Forman, G.; Rajaram, S. **Scaling up text classification for large file systems**. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08*, p. 239, 2008.

[17] Furnkranz, J. **A Study Using n-gram Features for Text Categorization**. *Austrian Research Institute for Artifical Intelligence*, 3(1998):1–10, 1998.

[18] Gabrilovich, E.; Markovitch, S. **Feature Generation for Text Categorization Using World Knowledge**. *Artificial Intelligence*, 19:1048, 2002.

[19] Gabrilovich, E.; Markovitch, S. **Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge**. *Proceedings of The 21st National Conference on Artificial Intelligence (AAAI)*, p. 1301–1306, 2006.

[20] Gantner, Z.; Schmidt-Thieme, L. **Automatic content-based categorization of Wikipedia articles**. *Proceedings of the 2009 Workshop on The People's Web Meets NLP Collaboratively Constructed Semantic Resources - People's Web '09*, (August):32–37, 2009.

[21] Hammami, M.; Tsishkou, D. **Adult content Web filtering and face detection using data-mining based kin-color model**. In: *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, p. 403–406. IEEE, 2004.

[22] Hu, J.; Fang, L.; Cao, Y.; Zeng, H.-J.; Li, H.; Yang, Q.; Chen, Z. **Enhancing text clustering by leveraging Wikipedia semantics**. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, p. 179, New York, New York, USA, 2008. ACM Press.

[23] Hu, X.; Zhang, X.; Lu, C.; Park, E. K.; Zhou, X. **Exploiting Wikipedia as external knowledge for document clustering**. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, p. 389, 2009.

[24] Ito, M.; Nakayama, K.; Hara, T.; Nishio, S. **Association thesaurus construction methods based on link co-occurrence analysis for wikipedia**. *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, p. 817, 2008.

[25] Joachims, T. **Text categorization with support vector machines: Learning with many relevant features**. *Machine Learning ECML98*, 1398(23):137–142, 1998.

[26] Joachims, T. **A support vector method for multivariate performance measures**. In: *Proceedings of the 22nd international conference on Machine learning*, p. 377–384. ACM, 2005.

[27] Joachims, T. **Training linear SVMs in linear time**. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 217–226. ACM, 2006.

[28] Lan, M.; Tan, C. L.; Su, J.; Lu, Y. **Supervised and traditional term weighting methods for automatic text categorization.** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):721–735, 2009.

[29] Lewis, D. D. **Representation quality in text classification: An introduction and experiment**. In: *Proceedings of Workshop on Speech and Natural Language. Hidden Valley, PA*, p. 288–295, 1990.

[30] Lewis, D. D. **An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task**. *Proceedings of the 15th annual international ACM*

SIGIR conference on Research and development in information retrieval SIGIR 92, pages(ACM Press):37–50, 1992.

[31] Lewis, D. D. **Feature selection and feature extraction for text categorization**. *Proceedings of the workshop on Speech and Natural Language - HLT '91*, p. 212, 1992.

[32] Lewis, D.; Yang, Y.; Rose, T.; Li, F. **Rcv1: A new benchmark collection for text categorization research**. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[33] Li, Y.; Luk, W. P. R.; Ho, K. S. E.; Chung, F. L. K. **Improving weak ad-hoc queries using wikipedia as external corpus**. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, p. 797, 2007.

[34] Liu, T.-Y. **Learning to Rank for Information Retrieval**. *Media*, 3(3):60558–60558, 2010.

[35] Manning, C. D.; Raghavan, P.; Schütze, H. **An Introduction to Information Retrieval**, volume 1. Cambridge University Press, Cambridge, England, Apr. 2009.

[36] McCallum, A.; Nigam, K. **A comparison of event models for naive bayes text classification**. *AAAI-98 workshop on learning for text*, p. 41–48, 1998.

[37] Metzler, D.; Novak, J.; Cui, H.; Reddy, S. **Building enriched document representations using aggregated anchor text**. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, p. 219, 2009.

[38] Michie, D.; Spiegelhalter, D. **Machine learning, neural and statistical classification**, volume 37. Ellis Horwood, Nov. 1994.

[39] Milne, D.; Medelyan, O.; Witten, I. **Mining Domain-Specific Thesauri from Wikipedia: A Case Study**. *2006 IEEEWICACM International Conference on Web Intelligence WI 2006 Main Conference ProceedingsWI06*, p. 442–448, 2006.

[40] Milne, D. N.; Witten, I. H.; Nichols, D. M. **A knowledge-based search engine powered by wikipedia**. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*, p. 445, 2007.

[41] Mladenic, D.; Grobelnik, M. **Word sequences as features in text-learning**. In: *Proc of ERK98 7th Electrotechnical and Computer Science Conference*, p. 145–148, 1998.

[42] Nakayama, K.; Hara, T.; Nishio, S. **A Thesaurus Construction Method from Large Scale Web Dictionaries**. *21st International Conference on Advanced Networking and Applications AINA 07*, (Aina):932–939, 2007.

[43] Nakayama, K.; Hara, T.; Nishio, S. **Wikipedia Mining for an Association Web Thesaurus Construction**. *Construction*, 4831:322–334, 2007.

[44] Page, L.; Brin, S.; Motwani, R.; Winograd, T. **The PageRank citation ranking: Bringing order to the web.** *World Wide Web Internet And Web Information Systems*, p. 1–17, 1999.

[45] Rosa, T. C. **Uso de Apontadores na Classificação de Documentos em Coleções Digitais**. PhD thesis, Universidade Federal de Minas Gerais, 2007.

[46] Salton, G.; Wong, A.; Yang, C. **A vector space model for automatic indexing**. *Communications of the ACM*, 18(11):613–620, 1975.

[47] Salton, G.; Buckley, C. **Term-weighting approaches in automatic text retrieval**. *Information Processing & Management*, 24(5):513–523, 1988.

[48] Schapire, R. E.; Singer, Y. **A boosting-based system for text categorization**. *Machine Learning*, 39(2/3):135–168, 2000.

[49] Schölkopf, B.; Smola, A. J. **Learning with Kernels**, volume 64 de **Adaptive Computation and Machine Learning**. MIT Press, 2002.

[50] Schonhofen, P. **Identifying Document Topics Using the Wikipedia Category Network**. In: *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, p. 456–462. IEEE, Dec. 2006.

[51] Sculley, D.; Wachman, G. M. **Relaxed online SVMs for spam filtering**. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 07*, p. 415, 2007.

[52] Sebastiani, F. **Machine Learning in Automated Text Categorization**. *Computing*, 34(1):1–47, 2002.

[53] Senellart, P.; Blondel, V. D. **Automatic discovery of similar words**. *Discovery*, p. 20 pp, 1913.

[54] Shen, D.; Sun, J.-T.; Yang, Q.; Chen, Z. **Text classification improved through multigram models**. *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*, p. 672, 2006.

[55] Slattery, S.; Craven, M. **Combining Statistical and Relational Methods for Learning in Hypertext Domains**. In: Page, D., editor, *Inductive Logic Programming 8th International Workshop ILP98 Madison Wisconsin USA July 2224 1998*, volume 1446 de **Lecture Notes in Computer Science**, p. 38–52. Springer, 1998.

[56] Smith, A. G. **Web links as analogues of citations**. *Information Research*, 9(4):net/ir/9–4/paper188, 2004.

[57] Srivastava, A. N.; Sahami, M. **Text mining: Classification, clustering, and applications**. Chapman & Hall/CRC, Minneapolis, Minnesota, U.S.A, 2009.

[58] Supreethi, K. P.; Prasad, E. V. **A Novel Document Representation Model for Clustering**. *International Journal of Computer Science Communication*, 1(2):243–245, 2010.

[59] Tan, C. **The use of bigrams to enhance text categorization**. *Information Processing & Management*, 38(4):529–546, 2002.

[60] Van Rijsbergen, C. J. **Information Retrieval**, volume 30 de **The Kluwer International Series on information retrieval**. Butterworths, 1979.

[61] Vapnik, V. N. **The Nature of Statistical Learning Theory**, volume 8 de **Statistics for Engineering and Information Science**. Springer, 1995.

[62] Völkel, M.; Krötzsch, M.; Vrandecic, D.; Haller, H.; Studer, R. **Semantic Wikipedia**. *Proceedings of the 15th international conference on World Wide Web - WWW '06*, (January 2001):585, 2006.

[63] Wang, P.; Domeniconi, C. **Building semantic kernels for text classification using wikipedia**. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08*, p. 713, 2008.

[64] Wang, P.; Hu, J.; Zeng, H.-J.; Chen, Z. **Using Wikipedia knowledge to improve text classification**. *Knowledge and Information Systems*, 19(3):265–281, Sept. 2008.

[65] Wikipedia. **Wikipedia, the free encyclopedia**, 2011.

[66] Wilcoxon, F. **Individual comparisons by ranking methods**. *Biometrics Bulletin*, 1(6):80–83, 1945.

[67] Wu, F.; Weld, D. S. **Autonomously semantifying wikipedia**. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*, p. 41, 2007.

[68] Yang, Y.; Liu, X. **A re-examination of text categorization methods**. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, p. 42–49. ACM, 1999.

[69] Yang, Y.; Pedersen, J. **A comparative study on feature selection in text categorization**. In: *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, p. 412–420. MORGAN KAUFMANN PUBLISHERS, INC., 1997.

[70] Zaïane, O. R.; Antonie, M.-L. **Classifying text documents by associating terms with text categories**. *Australian Computer Science Communications*, 5:215–222, 2002.

[71] Zhang, L.; Zhu, J.; Yao, T. **An evaluation of statistical spam filtering techniques**. *Acm Transactions On Asian Language Information Processing*, 3(4):243–269, 2004.