Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Classificação de Documentos Web Utilizando Múltiplas Fontes de Evidência

Ketlen Karine Teles Lucena

Manaus – Amazonas
Agosto de 2004

Ketlen Karine Teles Lucena

Classificação de Documentos Web Utilizando Múltiplas Fontes de Evidência

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Orientador: Prof. Dr. Edleno Silva de Moura

Ketlen Karine Teles Lucena

Classificação de Documentos Web Utilizando Múltiplas Fontes de Evidência

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Banca Examinadora

Prof. Dr. Edleno Silva de Moura – Orientador
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. Altigran Soares da Silva
Departamento de Ciência da Computação – UFAM/PPGI

Prof. João Marcos Bastos Cavalcanti, Ph.D.
Departamento de Ciência da Computação – UFAM/PPGI

Manaus – Amazonas
Agosto de 2004

Para Alice.

Agradecimentos

A Deus, acima de tudo.

À minha mãe, Raimunda Teles, pelo apoio incondicional.

Ao meu marido, Walfredo Lucena, por ter estado sempre presente e pela paciência.

Ao meu orientador, Edleno Moura, pelo incentivo e insistência e por ter acreditado na minha capacidade.

Aos colegas Marco Antônio Cristo e Pável Calado, pelo material e orientações imprescindíveis para a realização deste trabalho.

Ao amigo Vilar Camara Neto, pela contribuição fundamental para a escrita desta dissertação e pela dedicação.

Ao amigo Fabrício D'Morison, pelo suporte aos experimentos e pela disponibilidade.

Às amigas Keyla Ahnizeret e Nívea Michelle Melo, pela força e apoio.

Aos amigos do GTI: Daniel Fernandes, David Braga, Edson César, Eduardo Abinader, Maely Moraes e Moisés Carvalho, pelo companheirismo.

À Elienai Nogueira e à Mary Jani Fontenelle, pelo apoio administrativo e pela amizade.

Ao PPGI, pela oportunidade.

A todos aqueles que ajudaram de alguma forma na realização deste trabalho, o meu mais profundo agradecimento.

Aprender é descobrir aquilo que você já sabe. Fazer é demonstrar que você sabe. Ensinar é lembrar aos outros que eles sabem tanto quanto você. Vocês são todos aprendizes, fazedores, professores.

Richard Bach

Resumo

A classificação automática de documentos *web* tem sido um tema amplamente pesquisado em Recuperação de Informação. A utilização do conteúdo textual de páginas *web*, bem como de sua estrutura de links tem trazido informações bastante relevantes para o processo de separação de documentos em categorias. A combinação dessas características também se apresenta como uma opção para melhorar a qualidade dos resultados de uma classificação. Entretanto, mais pesquisas ainda são necessárias para se obter resultados mais precisos e confiáveis.

O objetivo deste trabalho é analisar o impacto da utilização de duas fontes de evidências na categorização de páginas *web*: o Título HTML das páginas e o Conteúdo dos Links de páginas que apontam para a página a ser classificada. Estas duas novas fontes de informação foram incluídas em método de classificação previamente proposto para a combinação de diferentes fontes de evidências, através de um modelo de Rede Bayesiana.

O título das páginas mostrou-se eficaz como fonte de informação para classificação de páginas, quer usado isoladamente para a classificação, quer quando combinado a outras evidências. Isto significa que, mesmo sendo composto de poucas palavras, o título consegue caracterizar de forma relativamente precisa o assunto de que trata uma página. O uso do conteúdo textual dos links isoladamente não resultou em valores mais altos do que a classificação usando somente o texto das páginas, porém quando utilizado em conjunto com outras informações obteve ganhos consideráveis.

Assim sendo, o uso de evidências como o Título e o Conteúdo dos Links sugere benefícios para a tarefa de classificação de documentos *web*. Nos experimentos, os melhores resultados foram conseguidos pela utilização do Título, e dado que o custo para se computar uma classificação usando somente esta evidência é menor do que quando se utiliza o texto, o Título se apresenta com uma evidência textual importante para a classificação final.

Palavras-chave: Tecnologia da Informação, Recuperação de Informação, Classificação, Web, Link Analysis, Bayesian Networks.

Abstract

Automatic classification of Web documents is one of the most studied subjects in Information Retrieval. The use of text and links inside web pages has shown to be a good path, as it has returned good results and important information to help the process of separating these documents in categories.

Better than using text and links separately is to make use of the combination of both of them, and some other sources of evidence can be used among different hypertext features. However, further research is needed to reach more accurate and reliable results.

This work analyzes the impact of the use of two sources of evidence, with hypertext features, while performing the categorization: the HTML Title and the Link Content of the training pages that points to the submitted pages to be classified. These new sources of evidence were included in a previous proposed method of classification to combine different sources of information, through a Bayesian Model.

The HTML Title, used alone or combined, has shown to be an efficient method. This means that, even if it is composed by few words, the title is able to characterize in a relatively accurate way the subject of a page. The literal link content used alone has not produced values greater than those generated by the single use of the text of the pages. However, when combined with other sources of evidence, it obtained good profits.

Finally, these two sources of evidence comes not only to help, but to improve the classsification tasks. For comparative interests, the best experimental results were reached through the use of the HTML Title and, moreover, considering that the computational effort costs less when performing a classification by title, this source of evidence has proven to be important to achieve a good quality classification.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A tarefa de localizar, organizar e catalogar as informações disponibilizadas pela *World Wide Web* (WWW) consiste em um vasto campo para pesquisas na área de Recuperação de Informação (RI). Características da *World Wide Web* como o crescimento exponencial do volume de dados [34] e sua volatilidade dificultam bastante o acesso a informações. Uma das maneiras de se contornar este problema seria a criação de *sites* que classifiquem as páginas, organizando-as em hierarquias conhecidas como diretórios.

A divisão de páginas por categorias, classificadas de acordo com o assunto de que tratam, auxilia bastante a pesquisa e a navegação no *site*, principalmente quando o usuário ainda não tem certeza do que está buscando.

Estudos recentes demonstram que usuários freqüentemente preferem navegar e realizar pesquisas através de diretórios [34], pois a estrutura das categorias permite o refinamento da consulta, diminuindo o tempo gasto na busca de informações. O uso constante de diretórios de categorias é bastante observado em grandes portais de busca, tais como o Yahoo! (`http://www.yahoo.com`), o que demonstra a utilidade prática da classificação de documentos em hierarquias de categorias.

1.1 O problema da classificação

Atualmente, a classificação de documentos *web* é realizada de forma manual, onde especialistas catalogam as páginas de acordo com áreas de conhecimentos pré-estabelecidas. Esta tarefa possui alto custo, pois demanda muito tempo, tendo em vista a grande quantidade de páginas

e a diversidade de assuntos.

Assim sendo, a solução seria o desenvolvimento de algoritmos que gerem uma classificação automática de páginas. Isto proporcionaria um melhor desempenho na obtenção do resultado final de uma classificação a um custo baixo [22].

Uma abordagem bastante estudada tem sido a implementação de classificadores baseados no conteúdo textual da página *web*, onde o texto existente na página é a única fonte de informação utilizada para determinar a categoria de cada documento. Este método apresenta bons resultados, porém, falha em algumas situações encontradas na *World Wide Web*. Por exemplo, páginas de conteúdo ambíguo, ou seja, aquelas que contêm imagens, objetos *flash* ou referências em outros idiomas, são ignoradas na classificação final. Tal fato ocorre porque documentos *web* que contenham apenas imagens ou que não possuam nenhuma informação textual, levam o classificador por conteúdo a não incluir em uma categoria páginas que seriam importantes para a mesma. Deste modo, classificadores baseados em conteúdo têm um baixo desempenho quando aplicados a coleções com características de hipertexto, como a *World Wide Web* [3, 6].

Segundo A. Sun [25], páginas *web* são essencialmente hipertexto, pois possuem características de contexto como *hyperlinks*, *tags* HTML e meta-dados. Muitas pesquisas têm sido desenvolvidas assumindo que o conteúdo textual de páginas *web* pode ser utilizado para uma classificação prévia e seu resultado posteriormente submetido a uma combinação com outras características não-textuais a fim de melhorar o resultado da classificação final [3, 34]. Desta forma, pode-se estender as técnicas de classificação por conteúdo, aproveitando as características de hipertexto dos documentos *web*, como por exemplo as informações de *links*. A utilização dos *links* tem obtido resultados significativos na classificação de documentos, como pode ser observado em vários trabalhos publicados na literatura [3, 6, 25, 34] (vide Seção 1.2).

Uma taxonomia definida por A. Sun [25] resume abordagens para a utilização das informações de *links*. Nela são estabelecidas três categorias: *hipertexto*, *análise de links* e *vizinhança da categoria*. A abordagem de *hipertexto* considera o conteúdo textual da página e as características de contexto, como as *tags* HTML e o texto dos *links* (*anchor words ou link content*) das páginas vizinhas. Na *análise de links* são aplicados algoritmos que manipulam tanto o conteúdo das páginas quanto a estrutura de *links* entre elas — por exemplo o algoritmo HITS [16], que descobre e computa regularidades em um conjunto de páginas *web*. Já a abordagem de *vizinhança da categoria* procura definir um documento como pertencente a uma determinada categoria com

base em informações obtidas da classe a que pertençam suas páginas vizinhas.

1.2 Objetivos do trabalho

P. Calado e M. Cristo [3, 8] utilizaram três abordagens de informação de *links* (*hipertexto*, *análise de links* e *vizinhança da categoria*) para desenvolver uma metodologia para classificação automática de documentos *web*, que combinou resultados obtidos de classificadores baseados em texto com informações da estrutura de *links*. A combinação dos experimentos foi obtida através de um modelo de Rede Bayesiana. A coleção utilizada para os conjuntos de teste e treinamento foi uma base de dados extraída do diretório do *Cadê*[1], que possui 44.099 documentos [3, 8]. Realizaram-se experimentos com quatro diferentes medidas de similaridade, derivadas da estrutura de *links*: *Acoplamento*, *Amsler*, *Co-citacão* e *Companion*, cujas definições encontram-se na Seção 2.2.

Nosso trabalho consiste em estudar novas fontes de evidência para classificação de documentos *web*, utilizando a base do *Cadê*, e incorporá-las ao modelo proposto por P. Calado e M. Cristo. Como novas fontes de informação, em adição às sugeridas por [3], foram propostos o título HTML (*title*) da página e o conteúdo dos *links* (*anchor words*) das páginas *web* que faziam referência ao documento a ser classificado. Para os nossos experimentos foram utilizadas três fontes de evidência baseadas em conteúdo textual: o Texto da página, o Título e o Conteúdo de Links, que nesta dissertação será denominado *ContLinks*. Nosso objetivo é verificar se a inclusão destas novas fontes pode resultar em uma melhoria na qualidade da classificação gerada pelo modelo proposto.

A partir da base formada por essas três evidências foram aplicados três algoritmos tradicionais para classificação por conteúdo textual: *k*NN (*k-Nearest-Neighbor*), Naive Bayes e SVM (*Support Vector Machine*). Outra fonte de informação foi a classificação por estrutura de *links* das páginas onde, assim como no trabalho de P. Calado e M. Cristo, foram utilizadas as mesmas medidas de similaridade de *links*. O modelo proposto em [3] foi modificado para considerar os resultados das classificações usando as novas fontes de evidência. Desta forma, foi possível analisar a influência desta metodologia no resultado da classificação final.

[1] Disponível em `http://www.cade.com.br/`.

1.3 Trabalhos relacionados

A organização de documentos em categorias tem sido um problema extensivamente estudado em RI. Mais especificamente, a classificação de documentos explorando as informações de *links* coloca-se como um novo desafio de pesquisa, pois a conectividade entre os documentos *web* contribui imensamente para que os recursos de hipertexto, estrutura de *links* e vizinhança entre as páginas possam ser utilizados como bons parâmetros de classificação. Além disso, a combinação de várias fontes de evidência também tem sido bastante utilizada. Informações como o título e a URL[2] das páginas *web* e os chamados *anchor words* (informações de *links*, como o conteúdo, etc.) têm obtido bom desempenho na classificação quando usados isoladamente ou quando combinados. Recentemente muitos trabalhos têm explorado esse aspecto, obtendo, em sua maioria, sucesso em seus experimentos[3, 8, 25, 34, 35, 33] .

I-H. Kang e G. Kim [14] combinaram infomações de URL com informações de *links* para classificar consultas de usuários em três categorias. No entanto, os resultados não foram igualmente bons para todas as classes, de onde se concluiu que a combinação estática de múltiplas evidências pode, eventualmente, diminuir o desempenho da classificação.

A melhor forma de combinar múltiplas evidências para classificação e recuperação de informações também tem sido objeto de estudo. Em 1991, H. Turtle e W. Croft [29] descreveram um modelo de recuperação baseado em rede de inferências e sugeriram seu uso para representar informações de hipertexto, obtidas através de duas abordagens: o *vizinho mais próximo (nearest neighbor)* e a *co-citação*. Os *links* não foram representados explicitamente na rede e dependiam da estimação da dependência probabilística entre os nós. Já B. Ribeiro-Neto e R. Muntz [20] apresentaram um modelo que pode ser utilizado em várias áreas clássicas de RI, pois ele permite a combinação de características de modelos diferentes em um mesmo esquema de representação. Seus resultados, utilizando a coleção *Cystic Fibrosis* (CF), confirmaram que o modelo pode ser estendido para adicionar novas fontes de evidências e melhorar a recuperação de informação.

Nosso trabalho inspirou-se basicamente em três pesquisas: Y. Yang [34] realizou classificações baseadas em hipertexto, utilizando algoritmos classificadores tradicionais. A. Sun [25] usou o *Support Vector Machine* (SVM) para classificar documentos, acrescentando características de contexto. P. Calado e M. Cristo [3, 8] combinaram informações de *links* com métodos de

[2] *Uniform Resource Locator*

classificação por conteúdo para a divisão dos documentos *web* em categorias. Os resultados destes trabalhos foram interessantes e trazem boas sugestões de utilização de diversas evidências a fim de melhorar de classificação de documentos *web*. Na seção abaixo serão descritos cada um desses trabalhos e a relação que cada um teve com nossa proposta.

1.3.1 Classificação baseada em hipertexto

Y. Yang [34] publicou em 2002 um resumo de várias propostas de classificação de documentos *web* baseadas em hipertexto. De uma página *web* foram investigados *hyperlinks*, tags HTML, a distribuição em categorias das páginas apontadas pela página a ser classificada, etc. As coleções usadas para testes foram *Hoovers* (`http://www.hoovers.com`), com 4.285 páginas, e *Univ-6*, com 4.165 páginas.

Seus experimentos consistiam em aplicar, em um conjunto de páginas a serem categorizadas, três tipos de classificadores por conteúdo: *Naive Bayes*, FOIL e kNN. A partir deste primeiro resultado utilizaram-se novos critérios para uma nova classificação. Foram selecionadas todas as palavras do texto dos *hyperlinks* da página; os nomes (ou identificadores), o título HTML das páginas, as *meta-tags* HTML, entre outros. Em seguida foi aplicado novamente um classificador por conteúdo.

Para a primeira coleção, os melhores resultados foram obtidos pela combinação do kNN com o texto da página (58,1 de medida F_1[3]) e das *meta-tags* HTML das páginas, (com 49,8 pontos), também com o kNN. Para a *Univ-6*, a dupla kNN e texto dos *hyperlinks* conseguiu o maior valor: 87,2 pontos em F_1. A diferença de valores entre as duas coleções foi significativa, de quase 20 pontos, para várias combinações.

Apesar dos resultados serem muito específicos a esses conjuntos de dados, este trabalho foi um dos que serviram de base para esta dissertação porque apresentou sugestões de combinação de diversas fontes de informação com métodos tradicionais de classificação.

1.3.2 Classificação combinando características de contexto

O trabalho apresentado por A. Sun [25] estudou os efeitos da utilização de características de contexto na classificação de páginas utilizando o SVM. A base dos experimentos foi a coleção

[3] Medida apresentada na 2.4.2

WebKB[4], que possui 4.159 documentos coletados e classificados manualmente em 7 categorias. Além do texto da página, foram considerados o título e palavras obtidas de páginas vizinhas, denominadas *anchor words*. Como exemplo de *anchor words*, pode-se citar o texto dos *hyperlinks* e os parágrafos ou as linhas ao redor deles.

Seus resultados foram bastante significativos. O experimento utilizando somente o texto obteve 49,2 de medida F_1. Quando utilizado o título e as *anchor words* foram alcançados os valores 53,2 e 58,2, respectivamente. Quando combinados o título e as *anchor words* o resultado foi 59,9. Isto mostrou que a inclusão dessas características de contexto melhorou a qualidadade da classificação significativamente.

As idéias de utilização de *anchor words* e o *título* das páginas, bem como seus resultados nos levaram e a experimentar e testar essas mesmas fontes de evidência na classificação de páginas *web* e a combiná-las com outras possíveis fontes.

1.3.3 Classificação utilizando estruturas de *links*

O trabalho de P. Calado e M. Cristo [3, 8] representou o ponto de partida para esta dissertação, pois foi estudada a influência das informações de *links* na classificação de uma coleção *web* e a combinação desses resultados com uma classificação baseada em conteúdo textual. A coleção *web* usada como conjunto de treinamento e teste foi extraída do diretório do *Cadê* [10].

Foi realizada uma classificação utilizando informação de *links* e utilizando-se quatro medidas de similaridade entre páginas *web*, derivadas da estrutura de *links*: *Acoplamento*, *Amsler*, *Co-citação* e *Companion*. Em seguida, esses resultados foram combinados à classificação obtida por três algoritmos classificadores por conteúdo: kNN, Naive Bayes e SVM. Os resultados desta combinação foram representados por um modelo de Rede Bayesiana.

Quando testadas isoladamente, as medidas de similaridade *Co-citação*, *Amsler* e *Companion* apresentaram os melhores resultados, com desempenho muito parecido. Isso vem confirmar a importância da informação de *links* para a classificação de documentos *web* e justificar a sua utilização neste trabalho. Por sua vez, a combinação das medidas baseadas em *links* com os classificadores por conteúdo resultou em valores bem heterogêneos. As informações de *links* classificaram corretamente um grande número de documentos, contudo introduziram algumas informações incorretas nas categorias. A aplicação dos classificadores por conteúdo filtraram

[4] `http://www-2.cs.cmu.edu/~webkb`

esses ruídos, porém removeram alguns documentos colocados corretamente nas classes. Para se evitar esse problema foi sugerida a inclusão de pesos a cada uma das fontes de evidências.

No contexto geral, a combinação das medidas de similaridade e classificadores baseados em conteúdo textual alcançou um resultado considerável na distribuição final dos documentos nas categorias, pois foi obtido um ganho de até 46 pontos em relação a uma classificação realizada individualmente com cada classificador. Este foi o principal motivo que inspirou a realização dos experimentos desta dissertação, pois, acrescentando-se à proposta de P. Calado e M. Cristo uma combinação de resultados de classificação com outras fontes de informação, foi possível analisar a influência das mesmas na classificação final.

1.4 Organização da dissertação

Esta dissertação está organizada em cinco capítulos. No capítulo 2 são apresentados os principais conceitos necessários à compreensão deste trabalho.

O capítulo 3 detalha o método utilizado por P. Calado e M. Cristo [3] para a combinação da classificação de documentos *web* por algoritmos classificadores por conteúdo e por estruturas de *links*. Descreve como as novas fontes de evidências foram tratadas e como foram incluídas no método proposto em [3] e o modo como foram combinadas. Também apresenta e discute a nossa proposta de modificação do modelo de Rede Bayesiana, que estendeu o modelo apresentado por P. Calado e M. Cristo [3] para suportar a inclusão de novas fontes de informação.

O capítulo 4 descreve detalhes sobre os experimentos realizados, tais como o ambiente de desenvolvimento, características da coleção que serviu como base de dados, as ferramentas e *softwares* utilizados. Apresenta e analisa ainda, através de tabelas e gráficos, todos os resultados alcançados.

Finalmente, o capítulo 5 traz o resumo dos resultados e as conclusões obtidas, discute as contribuições do trabalho desenvolvido e apresenta sugestões de trabalhos futuros.

Capítulo 2

Conceitos Básicos

Neste capítulo serão apresentados e discutidos de forma sucinta alguns conceitos básicos necessários à compreensão dos experimentos deste trabalho.

2.1 Algoritmos Classificadores por Conteúdo

Para os experimentos utilizando classificadores por conteúdo serão usados três algoritmos bastante conhecidos e extensivamente aplicados na tarefa de classificação de texto: *k-Nearest-Neighbor* (kNN), *Naive Bayes* e *Support Vector Machine* (SVM). Para os dois primeiros algoritmos foram utilizadas as implementações encontradas em um pacote de programas chamado *Bow* [17] e disponíveis em `http://www.cs.cmu.edu/~mccallum/bow`. A implementação do SVM foi retirada de *LIBSVM* [7], acessível em `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

2.1.1 k-Nearest-Neighbor (kNN)

O algoritmo de classificação kNN é baseado em memória e não requer nenhum modelo adaptado a ele [28]. Consiste em um método de aprendizagem baseado em instâncias e foi extensamente utilizado por Y. Yang [32] em seu trabalho de categorização e recuperação de texto, devido à sua simplicidade e eficência, e por fazer uso direto da informação de similaridade. Não possui processamento na fase de treinamento, pois não é necessário estimar as distribuições de probabilidade das classes.

O kNN calcula a similaridade entre as instâncias de um conjunto de teste e um conjunto de treinamento e considera as k primeiras instâncias do *ranking* mais próximas a uma categoria

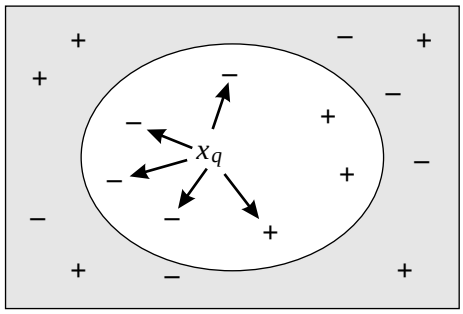

Figura 2.1: kNN classificando a instância para $k = 5$

como as mais similares à mesma. Em resumo, ele classifica uma instância de acordo com a classe a que pertençam as k instâncias vizinhas mais próximas. Dado um ponto x_0 da consulta, encontra-se os k pontos de treinamento $x_{(r)}$, $r = 1, \ldots, k$ o mais perto da distância a x_0 e classifica-se usando a categoria com maior quantidade de k vizinhos [28].

Assim pode-se resumir o algoritmo kNN nos seguintes passos:

1. Calcula-se a distância entre a instância a classificar e todos os padrões de treinamento;
2. Verificam-se a quais classes pertencem os k padrões mais próximos;
3. A classificação é feita associando-se a instância à classe mais freqüente entre os padrões mais próximos.

A Figura 2.1 exemplifica como o kNN classificaria um instância x_q utilizando $k = 5$. Neste exemplo, x_q pertenceria à categoria *negativo* $(-)$.

O valor de similaridade entre duas instâncias é a distância métrica entre elas. Normalmente usa-se a *distância Euclidiana* (ver equações (2.2) e (2.3)). Seja uma instância t com n características, representada pelo vetor

$$\langle v_1(t), v_2(t), \ldots, v_n(t) \rangle \tag{2.1}$$

onde $v_i(t)$ é o valor da característica de índice i da instância t. Desta forma, a distância entre duas instâncias t_i e t_j é

$$d_{(t_i,t_j)} = \|t_i - t_j\| \tag{2.2}$$

ou

$$d_{(t_i,t_j)} = \sqrt{\sum_{m=1}^{n} \left(v_m(t_i) - v_m(t_j)\right)^2} \tag{2.3}$$

A similaridade final entre dois documentos é cosseno do ângulo entre os vetores que representam os documentos. Denominando-se os vetores t_i e t_j de X e Y, obtém-se:

$$\cos\theta(\overrightarrow{X}, \overrightarrow{Y}) = \frac{\sum_i X_i Y_i}{\sqrt{\sum_j X_j^2 \sum_l Y_l^2}} \tag{2.4}$$

Apesar de sua simplicidade, o kNN tem sido amplamente usado em problemas de classificação, pois é freqüentemente bem sucedido onde cada classe tem muitas características que a identifiquem, e o limite de decisão é muito irregular [28].

2.1.2 Naive Bayes (NB)

O classificador conhecido como *Naive Bayes* (NB) é uma simplificação do teorema de Bayes (vide equação 2.5). Este método probabilístico é uma técnica de aprendizagem de máquina (*Machine Learning*), comumente usado na tarefa de classificação de documentos [3, 34]. O NB parte da hipótese de que a probabilidade da ocorrência de cada atributo do vetor entrada é independente da ocorrência de qualquer outro atributo. Em alguns domínios, seu desempenho tem sido comparável a técnicas mais sofisticadas de apredizagem de máquina, tais como redes neurais [18] e árvores de decisão [18].

A equação a seguir representa o Teorema de Bayes:

$$P(h \mid X) = \frac{P(h)\,P(X \mid h)}{P(X)} \tag{2.5}$$

Na equação (2.5) h é a hipótese e X é o evento observado (ou a entrada). $P(h)$ é a probabilidade de que a hipótese h ocorra sem nenhuma entrada. $P(X)$ é a probabilidade do evento X ocorrer sem que se conheça h. $P(X \mid h)$ é a probabilidade de X dado h, ou seja, é a probabilidade de X ocorrer quando h for verdadeiro. $P(h \mid X)$ é a probabilidade que a hipótese h espera, dado que um evento X é observado.

Para o problema de classificação, pode-se interpretar a hipótese h como "o documento de entrada X pertencente à categoria C_i" e X como o documento de entrada. O valor $P(h \mid X)$ é então a probabilidade de um documento de entrada pertencer à classe C_i.

Designando-se os valores de atributo X por $\langle a_1, a_2, \ldots, a_n \rangle$ e a hipótese h por v_j, pode-se reescrever o valor da probabilidade mais provável (v_B) para cada instância.

$$v_B = \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n \mid v_j)\, P(v_j)}{P(a_1, a_2, \ldots, a_n)} \tag{2.6}$$

O Naive Bayes é baseado na suposição de que os atributos são condicionalmente independentes dado um valor alvo [18], ou seja, a hipótese é que, dado um valor alvo da instância, a probabilidade de observação do conjunto $\langle a_1, a_2, \ldots, a_n \rangle$ é o produto das probabilidades dos atributos individualmente. Substituindo-se essa definição na equação (2.6), obtém-se:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i (P \mid v_j) \tag{2.7}$$

O método do NB envolve uma etapa de aprendizagem em que os vários termos de $P(v_j)$ e de $P(a_i \mid v_j)$ são estimados, baseados em suas freqüências sobre os dados do treinamento [18]. O conjunto destas estimativas corresponde à hipótese instruída. Esta hipótese é usada então para classificar cada novo exemplo aplicando-se a equação (2.7).

Uma diferença interessante entre o NB e outros métodos de aprendizagem é que ele considera que não há nenhuma busca explícita através do espaço de hipóteses possíveis (neste caso, o espaço dos valores possíveis que podem ser atribuídos aos vários termos de $P(v_j)$ e de $P(a_i \mid v_j)$). Ao invés disso, a hipótese é dada simplesmente contando-se a freqüência de várias combinações dos dados dentro dos exemplos do conjunto de treinamento [18].

2.1.3 Support Vector Machine (SVM)

O Support Vector Machine foi desenvolvido por V. Vapnik em 1982 [30] e é baseado no *Princípio Estrutural da Minimização do Risco* (proposto também por V. Vapnik [31]) da teoria computacional do aprendizado. Tem encontrado aplicação em tarefas como reconhecimento digital de imagens e classificação de texto [27], e seu algoritmo do SVM tem custo quadrático.

O SVM consiste na definição, em um espaço vetorial, da "melhor" superfície de decisão que separa os pontos de dados em duas classes [33], como pode ser visto na Figura 2.2 (um exemplo para duas dimensões). Quando a superfície de decisão é linearmente separável, é chamada de hiperplano. Para n dimensões, os pontos de dados não serão linearmente separáveis por classes.

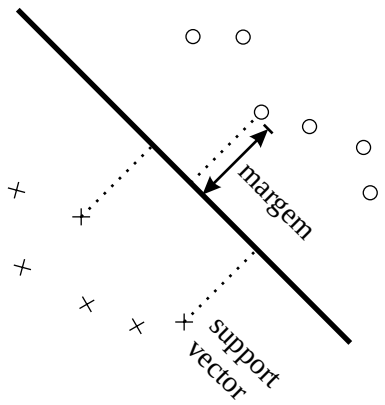

Figura 2.2: Superfície de decisão separando os pontos de dados em duas classes

Portanto, o objetivo do SVM é descobrir um hiperplano que maximize a margem de separação entre as classes do conjunto de treinamento. Um hiperplano pode ser descrito da seguinte maneira:

$$\overrightarrow{w} \cdot \overrightarrow{x} + b = 0 \tag{2.8}$$

onde $\overrightarrow{x}$ é a instância a ser classificada, e o vetor $\overrightarrow{w}$ e a constante b têm valores retirados do conjunto de treinamento linearmente separável [5]. A classificação é realizada aplicando-se a seguinte função de decisão:

$$f(\overrightarrow{s}) = \text{sign}\,((\overrightarrow{w} \cdot \overrightarrow{x}) + b) \tag{2.9}$$

Uma propriedade importante do SVM é que a superfície de decisão é determinada somente pelas instâncias que possuem a distância exata de $\frac{1}{\|\overrightarrow{w}\|}$ do plano de decisão, como pode ser observado na Figura 2.2. Estas instâncias (ou pontos de dados) são chamados de *support vectors* [33]. Esses pontos são os que realmente contam para o conjunto de treinamento. Isto significa que o SVM pode ignorar os outros pontos, pois a função de decisão permanece a mesma.

Quando se tratar de um espaço não-separável, o SVM mapeia os vetores de dados orginais para um outro espaço dimensional onde as instâncias sejam linearmente separáveis. A Figura 2.3 ilustra esse caso.

Apesar de ser um método de uso relativamente recente em RI, o SVM tem sido bastante utilizado para solucionar problemas de classificação de texto [3, 13, 27, 33], obtendo bons resultados quando comparados a outros classificadores convencionais, como o kNN e o Naive Bayes. Esse desempenho deve-se, em grande parte, à habilidade do SVM de manipular largos espaços de características sem a necessidade de uma seleção prévia das mesmas.

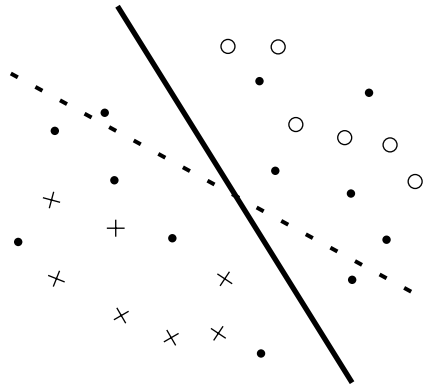

Figura 2.3: Classes linearmente não-separáveis

2.2 Medidas de Similaridade de *Links*

Para a classificação utilizando estruturas de *links*, foram utilizadas em nossos experimentos quatro medidas de similaridade de *links*: *Co-citação, Acoplamento, Amsler* e *Companion.* Essas medidas são derivadas das informações de *links* e objetivam determinar a similaridade entre páginas *web* através do "assunto" de que trata cada documento *web.* Na área da Ciência Bibliométrica, as citações bibliográficas em artigos científicos têm sido bastante estudadas para se estabelecer relações entre documentos. *Co-citação*, *Acoplamento* e *Amsler* são baseadas na premissa de que documentos co-relacionados através de suas citações bibliográficas possuem graus de similaridade entre si [1, 15, 24].

2.2.1 Co-citação

Se um artigo cita outros dois artigos, esses dois são chamados de *co-citados*, pois trabalha-se com a hipótese de que se um autor cita um artigo é porque este tem alguma relação com o seu próprio trabalho. Este conceito foi proposto em 1973 por H. Small [24] como uma medida de similaridade entre artigos científicos. Quando se aplica esta definição para a estrutura de *links* da *World Wide Web*, conclui-se que se uma página possui *links* para outras duas, seus assuntos podem ser considerados relacionados.

A medida de co-citação de uma página pode ser estabelecida da seguinte forma:

$$\text{cocit}(d_1, d_2) = \frac{\mid P_{d_1} \cap P_{d_2} \mid}{\mid P_{d_1} \cup P_{d_2} \mid} \tag{2.10}$$

onde d é um documento *web* e P_d um conjunto de páginas que apontam para d (pais de d). Se

P_{d_1} e P_{d_2} são vazios, então a co-citação também será igual a 0. O valor deve ser normalizado para que a medida varie entre 0 e 1.

2.2.2 Acoplamento

Conhecido como *bibliographic coupling* e introduzido por M. Kessler [15], o acoplamento também determina a similaridade entre dois artigos ou documentos *web*. Sua idéia é a inversa da *co-citação*: se dois artigos citam um mesmo artigo, os dois podem ser considerados uma unidade de acoplamento. No ponto de vista da *World Wide Web*, se duas páginas possuem *links* para um terceira, elas estarão relacionadas, pois os *links* tendem a apontar para páginas que tratam do mesmo assunto.

O acoplamento pode ser definido como:

$$\text{acop}(d_1, d_2) = \frac{\mid C_{d_1} \cap C_{d_2} \mid}{\mid C_{d_1} \cup C_{d_2} \mid} \tag{2.11}$$

onde d é um documento *web* e C_d um conjunto de páginas apontadas por d (filhas de d). Assim como na co-citação, caso C_{d_1} e C_{d_2} sejam vazios, então o valor de acoplamento será igual a 0. O resultado do acoplamento deve ser normalizado para que a medida esteja entre 0 e 1.

2.2.3 Amsler

Segundo R. Amsler [1], dois artigos A e B são similares se:

1. A e B são citados pelo mesmo artigo;
2. A e B citam o mesmo artigo;
3. A cita um terceiro artigo C que cita B.

Ou seja, a medida Amsler é a combinação dos conceitos de co-citação e acoplamento. Escrevendo-se mais formalmente, tem-se:

$$\text{amsler}(d_1, d_2) = \frac{(P_{d_1} \cup C_{d_1}) \cap (P_{d_2} \cup C_{d_2})}{\mid (P_{d_1} \cup C_{d_1}) \cup (P_{d_2} \cup C_{d_2}) \mid} \tag{2.12}$$

onde d é um documento *web* e Pd C_d um conjunto de páginas que apontam para d (pais de d) e Cd um conjunto de páginas apontadas por d (filhas de d). Caso d_1 e d_2 não possuam pais e nem filhos, a medida será igual a 0. O valor deve ser normalizado pelo número total de *links*.

2.2.4 Companion

Em 1999, J. Kleinberg [16] propôs um algoritmo chamado *HITS (Hyperlink Induced Topic Search)* que a partir de um grafo de vizinhanças com páginas *web*, determina valores de *Autoridade* e *Hub* para cada página. Os vértices deste grafo representam os documentos *web* e as arestas indicam os *links* entre eles. Os conceitos de *Autoridade* e *Hub* são recursivos. *Autoridade* é um índice utilizado para determinar se uma página é uma boa referência sobre determinado assunto: por exemplo, páginas populares têm alto valor de *Autoridade*. Já uma página que aponta para páginas com alto valor de *Autoridade* possuem alto valor de *Hub*.

O *Companion* foi apresentado por J. Dean e M. Henzinger [9] como um algoritmo baseado no HITS e com o objetivo de encontrar páginas relacionadas através de sua estrutura de *links*. Para isso ele utiliza os valores de *Autoridade* e *Hub* para determinar um grau de relação entre documentos *web* do conjunto dado com outros documentos de sua vizinhança. O *Companion* foi usado pela primeira vez para classificação de documentos *web* por D. Mota em 2001 [19].

As etapas seguintes resumem basicamente o algoritmo do Companion:

1. Monta um grafo de vizinhança de cada página (nó) da categoria;
2. Elimina as páginas repetidas para cada nó;
3. Aplica o HITS para computar os valores de *Autoridade* e *Hub* para cada página do grafo e de sua vizinhança.

Após estes passos, o *Companion* retorna um *ranking* com as páginas que obtiveram um alto valor de *Autoridade*. Estas páginas podem ser consideradas as mais relacionadas a uma página inicial. Para se obter uma classificação de páginas através deste algoritmo basta utilizar o resultado do *ranking* para determinar a classe do documento. Isto é, quando um documento D_1 obtiver um alto índice de relacionamento com um documento D_2, ele poderá ser considerado pertencente à mesma classe de D_2.

2.3 Redes Bayesianas

Uma Rede Bayesiana é baseada no modelo de rede básica de inferência e consiste em um grafo acíclico de dependências, cujos nós representam variáveis proposicionais ou constantes, e as arestas indicam as relações de dependência entre as proposições. Esse tipo de rede permite combinar características de diferentes modelos em um mesmo esquema representacional, e por isso consegue modelar os eventos e a interdependência de três componentes básicos em Recuperação de Informação (RI): palavras-chaves, documentos e consultas.

H. Turtle e W. Croft [29] propuseram o primeiro modelo de Rede Bayesiana para RI, chamado *inference network model*, onde demonstravam que estender o modelo básico de rede de inferência com representações booleanas das consultas de usuário pode render uma melhoria na qualidade da resposta. Uma variação deste modelo, denominada *belief network model for IR*, foi apresentada por B. Ribeiro-Neto e R. Muntz [20] para introduzir evidências de consultas anteriores, com o objetivo de melhorar a recuperação de informação. I. Silva [23] propôs um modelo que representava ao mesmo tempo os três modelos clássicos de RI (Vetorial, Booleano e Probabilístico), os ciclos de realimentação de relevantes e alternativas de similaridade consulta-consulta.

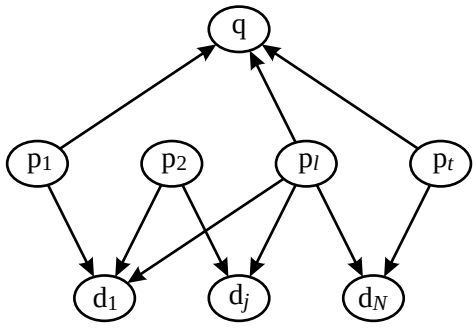

Figura 2.4: Modelo de Rede Bayesiana

A Figura 2.4 representa um exemplo de Modelo de Rede Bayesiana. Se uma proposição representada por um nó p implica em um proposição q, uma aresta é colocada de p para q. O nó q contém um *link* que especifica $P(q \mid p)$ para todos os possíveis valores das duas variáveis.

Um modelo de Rede Bayesiana para RI pode ser visto na Figura 2.5. Nele cada nó d_j modela um documento, o nó q modela a consulta do usuário e os nós k_i representam os termos da coleção.

A similaridade entre um documento d_j e uma consulta q é a probabilidade do documento d_j

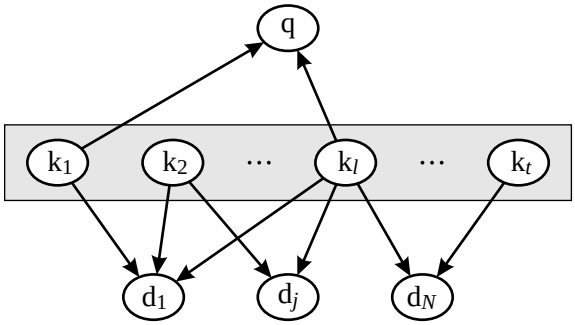

Figura 2.5: Modelo de Rede Bayesiana para RI

ser observado dado que q ocorre. Deste modo, a similaridade $P(d_j \mid q)$ pode ser definida como a seguinte expressão genérica:

$$P(d_j \mid q) = \eta \sum_{m=1} P(d_j \mid k)\, P(q \mid k)\, P(k) \tag{2.13}$$

sendo $\eta = \frac{1}{P(q)}$ uma constante de normalização. Assim sendo, para se representar qualquer modelo tradicional de RI pode-se usar a Equação (2.13), bastando que sejam definidas os valores de $P(d_j \mid k)$, $P(q \mid k)$ e $P(k)$.

Neste trabalho foi adotado o modelo de Rede Bayesiana proposto por B. Ribeiro-Neto e R. Muntz [20] para representar a combinação dos resultados da classificação de páginas *web* utilizando múltiplas evidências.

2.4 Métodos de avaliação

Para a avaliação os resultados dos experimentos, foram escolhidas medidas de avaliação de desempenho comumente usadas em sistemas de RI. Essas medidas baseiam-se na coleção de testes e no conjunto de respostas corretas fornecido por especialistas.

2.4.1 Precisão e Revocação

Precisão e *Revocação* são estimativas de qualidade bastante utilizadas para avaliação de respostas de métodos de classificação automática de páginas *web* [3, 8, 19, 34]. Considerando N como o conjunto de documentos relevantes identificados pelos especialistas e R o conjunto de documentos respondidos pelo sistema que foram examinados, pode-se determinar os seguintes conceito:

A *Precisão* consiste na proporção de documentos relevantes (N) dentro da resposta retornada pelo sistema em comparação com o resultado dos especialistas (R). É a fração de documentos recuperados que podem ser considerados relevantes [2]. A *Precisão* pode ser definida pela Equação (2.14):

$$p = \frac{\mid N \cap R \mid}{\mid R \mid} \tag{2.14}$$

Já a *Revocação* é a aproximação do total de relevantes do resultado dos especialistas, ou seja, quanto maior a *Revocação*, mais perto do total exato de relevantes. É a fração de documentos relevantes que foi recuperada pelo sistema [2]. A Equação (2.15) a seguir, define a revocação:

$$r = \frac{\mid N \cap R \mid}{\mid N \mid} \tag{2.15}$$

2.4.2 Medida F_1

A *média harmônica* ou *medida* F_1 é útil quando se deseja combinar os valores de Precisão (p) e Revocação (r) em um único valor de medida de qualidade [2, 5]. A medida F_1 pode ser definida como:

$$F_1 = \frac{(\alpha^2 + 1)\, p\, r}{\alpha^2\, p + r} \tag{2.16}$$

onde α indica o peso dos valores de precisão (p) e revocação (r). Quando considerado $\alpha = 1$, a medida F_1 pode ser reescrita como:

$$F_1 = \frac{2\, p\, r}{p + r} \tag{2.17}$$

A função F_1 assume valores entre 0 e 1. Quando o resultado for 0, significa que nenhum documento foi recuperado e quando for 1, o sistema conseguiu recuperar todos os relevantes com precisão máxima. Conseqüentemente, o máximo valor de F_1 pode ser interpretado como o melhor resultado da combinação de precisão e revocação [2].

Capítulo 3

Método de Classificação Proposto

Nossa proposta é estudar o impacto da inclusão de duas fontes de evidência (Título e ContLinks), submetê-las a uma classificação utilizando classificadores tradicionais baseados em texto e combinar os resultados com os obtidos por uma classificação por estruturas de *links*. Essa combinação é obtida através da extensão do modelo de Redes Bayesianas proposto por P. Calado e M. Cristo [3]. Este modelo foi modificado para adicionar as novas fontes de informação e permitir a análise do desempenho das mesmas. Assim sendo, neste capítulo será discutido como as informações baseadas em *links* podem ser combinadas com as informações baseadas em conteúdo textual, de acordo com a proposta de P. Calado e M. Cristo [3]. E ainda, como foram incluídas as novas fontes de evidência na extensão do modelo de Rede Bayesianas proposto em [3].

3.1 Fontes de evidências

3.1.1 Informações baseadas no Conteúdo Textual

Documentos *web* incluem uma variedade de informações textuais que podem fornecer indicações importantes sobre o assunto de que tratam. Entretanto, devido à volatilidade da *World Wide Web*, o conteúdo textual de uma página, quando visto como um todo, não fornece necessariamente uma descrição apropriada de seu assunto. Isto pode comprometer a utilização de somente uma parte específica de texto na página como um indicativo confiável de sua categoria.

Uma solução possível é considerar cada parte do texto como uma fonte de evidência independente. Neste caso, foram utilizadas três possíveis fontes obtidas do conteúdo de páginas: (i)

texto do documento *(body)*, (ii) título HTML *(title)*, e (iii) conteúdo dos *links* das páginas que apontam para o documento a ser classificado *(anchor text)*.

O texto ou *body* de uma página é formado pelas palavras que estão no documento *Web* entre as *tags* `<BODY>` e `</BODY>`. Geralmente é esperado que o texto de uma página contenha uma grande quantidade de informações sobre o provável assunto desta. O título de uma página é composto por termos localizados entre `<TITLE>` e `</TITLE>` de um documento HTML. Estes termos são normalmente usados para descrever o conteúdo do documento e pode ser visto como um conjunto de palavras-chaves relacionadas ao assuntos da página. O conteúdo dos *links* ou *anchors text* são as palavras encontradas entre as *tags* `<A>` e `</A>` de páginas que apontam para a página a ser classificada. Apesar de ser uma informação reduzida, o conteúdo dos *links* pode fornecer uma descrição sucinta da página para onde fazem referência.

Em todas as fontes de evidências usadas para classificar documentos foram aplicados os três algoritmos baseados em conteúdo, discutidos no Capítulo 2: kNN, Naive Bayes e SVM.

3.1.2 Informações baseadas na Estrutura de *Links*

Como no trabalho de P. Calado e M. Cristo [3], para determinar a similaridade de um assunto entre duas páginas *web*, foram usadas quatro medidas de similaridade derivadas da estrutura de *links*: *Co-citação*, *Acoplamento*, *Amsler* e *Companion*. As três primeiras foram propostas como medidas de similaridade entre artigos científicos. Embora *web*links sejam diferentes das citações em artigos, é válido estabelecer uma analogia entre eles, pois geralmente *links* e citações possuem o mesmo significado. Neste caso, pode-se supor que o autor da página *web* sempre irá colocar *links* para páginas relacionadas a sua própria página.

O Companion é uma abordagem diferente das anteriores. Dada uma página *web* p, o algoritmo encontra um conjunto de páginas relacionadas a p examinando seus *links*. Ele retorna um grau de similaridade de p, que significa o quão relacionada esta página é com o assunto de uma categoria. Para isso, o Companion considera as páginas da vizinhança de p e os *links* entre ela e as outras páginas da categoria, formando um grafo. Esse grafo é processado por um algoritmo chamado HITS [16], o qual retorna os graus de *autoridade* e *hub* para cada página. Intuitivamente, uma boa *autoridade* é uma página que fornece informações importantes sobre um dado assunto e um bom *hub* é uma página que aponta para boas *autoridades*. Neste trabalho foi usado o grau de *autoridade* como uma medida de similaridade entre p e cada página do grafo.

Para se obter uma classificação de documentos a partir dessas medidas de similaridade de *links*, foi aplicado o algoritmo kNN. Este classificador atribui uma categoria a um documento de teste baseado nas categorias dos k vizinhos mais similares pertencentes ao conjunto de treinamento. A similaridade entre os documentos foi encontrada, portanto, substituindo-se a convencional distância euclidiana[1] pelos pelas medidas de similaridade de *links* descritas acima.

O resultado desta classificação foi combinado ao resultado da classificação utilizando o conteúdo textual (já incluindo o Título e o ContLinks) dos documentos através de um modelo de Rede Bayesiana [3], descrito na próxima seção.

3.2 Combinando múltiplas evidências

Para suportar a inclusão de novas fontes de informação, e, baseando-se na proposta de P. Calado [3] para combinar informações de *links* e texto, este trabalho também propõe a utilização de um modelo de Rede Bayesiana [20] para combinar evidências. As Redes Bayesianas têm sido aplicadas com sucesso em muitas tarefas de recuperação de informação (RI) [5, 20, 23], pois oferecem uma base formal tanto para simulação de modelos de RI tradicionais quanto para a combinação de informações de fontes diferentes. Além disso, segundo P. Calado [5], as Redes Bayesianas permitem também uma visão uniforme, flexível, intuitiva e formal de diversos problemas de análise de *links*.

O modelo proposto neste trabalho consiste em uma extensão direta do modelo de P. Calado, a fim de permitir a combinação de um número maior de fontes de evidências (neste caso, o Título e o ContLinks), como mostrado na Figura 3.1.

Assim como o modelo de P. Calado, cada nó corresponde a uma variável aleatória. Os nós-raiz, denominados $D_1 \ldots D_N$, representam o conhecimento prévio sobre o problema, ou seja, um conjunto de documentos já classificados (conjunto de treinamento). O nó C indica a categoria ou classe. A categoria é constituída por um conjunto de documentos classificados. Assim, há arestas dos nós D_j ao nó C, representando as características dos documentos classificados que irão influenciar na definição das características representativas da categoria.

Cada conjunto de nós $E_{i_1} \ldots E_{i_K}$ representa as fontes de informação associadas aos documentos a ser classificados (conjunto de teste). Cada nó E_{i_k} representa a informação da i-ésima

[1] Medida normalmente utilizada pelo kNN para definir a similaridade entre documentos

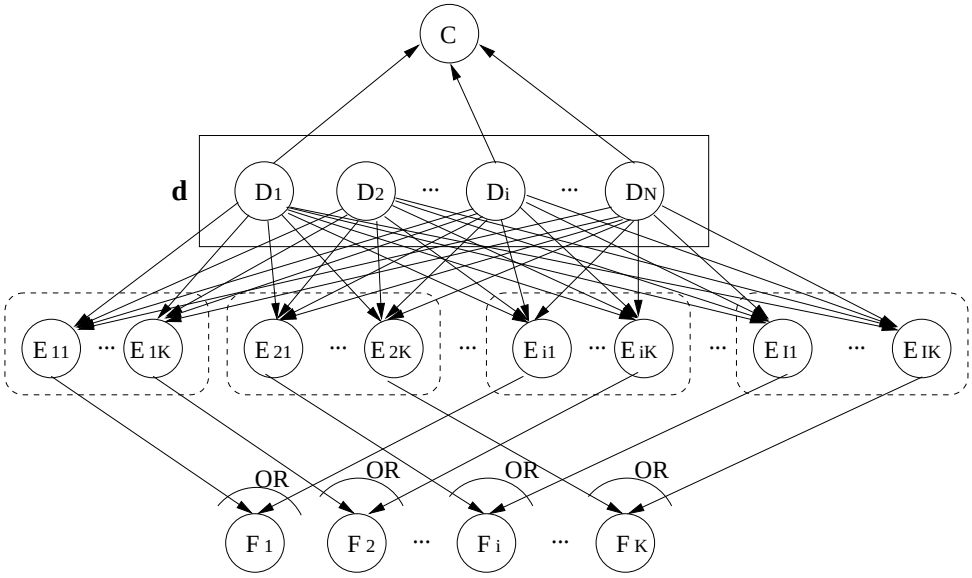

Figura 3.1: Modelo de Rede Bayesiana proposto

fonte de evidência, indicando que o documento de teste k pertence à categoria C. Desde que essa evidência dependa do conjunto de treinamento, existirão arestas de cada nó D_j a cada nó E_{i_k}. Desta forma, considerando-se o conjunto de treino, pode-se inferir se o documento de teste k pertence ou não à categoria C.

Finalmente, os nós F_1 até F_K representam a evidência final que coloca cada documento de teste como pertecente à categoria C. Esta evidência depende de todas as fontes de informação, como mostrado pelas arestas que chegam a F.

Dadas estas definições, pode-se usar a rede para determinar a probabilidade do documento k pertencer à categoria C. Traduzindo-se para uma equação de probabilidade, tem-se:

$$P(f_k \mid c) = \eta \sum_{\forall \mathbf{d}} \Big(1 - \prod_{\forall i} \big(1 - P(e_{i_k} \mid \mathbf{d})\big)\Big) P(c \mid \mathbf{d})\ P(\mathbf{d}) \tag{3.1}$$

onde $\eta = \frac{1}{P(c)}$ é uma constante de normalização e $\mathbf{d}$ é o estado possível de cada variável D_j. Necessita-se somente definir as probabilidades $P(e_{i_k} \mid \mathbf{d})$, $P(c \mid \mathbf{d})$, e $P(\mathbf{d})$.

Seja $\mathscr{C}$ um conjunto de documentos definidos como pertencentes à categoria C e seja $\overline{\mathscr{C}}$ o conjunto de documentos não-pertencentes a C. Define-se:

$$P(e_{ik} \mid \mathbf{d}) = \text{class}_i(k, \mathscr{C}, \overline{\mathscr{C}}) \tag{3.2}$$

onde $\text{class}_i(k, \mathscr{C}, \overline{\mathscr{C}})$ é a função que retorna um valor da associação entre o documento de teste k e a categoria C, baseado no conjunto de documentos. A função class_i representa o valor dado pelo mecanismo de classificação associado à i-ésima fonte de evidência. Para nossos experimentos, esse valor é fornecido por: (i) Resultado dos classificadores kNN, Naive Bayes e SVM, para as fontes de evidências baseadas em conteúdo textual; (ii) Pelas fontes de evidência baseada na estrutura dos *links*.

Assume-se que este valor é normalizado para $0 \leq \text{class}_i(\mathscr{C}, \overline{\mathscr{C}}) \leq 1$.

A probabilidade $P(c \mid \mathbf{d})$ é então usada para selecionar somentes os documentos de treino que pertencem à categoria que se deseja processar. Define-se $P(c \mid \mathbf{d})$ como:

$$P(c \mid \mathbf{d}) = \begin{cases} 1 & \text{se } \forall_i, d_i = 1 \Leftrightarrow i \in \mathscr{C} \\ 0 & \text{para o restante} \end{cases} \tag{3.3}$$

onde $\mathscr{C}$ é o conjunto de documentos de treino que pertencem à categoria C.

Finalmente, visto que não se possui uma preferência inicial como o conjunto de treinamento mais provável de ser observado, pode-se considerar *a priori* a probabilidade $P(\mathbf{d})$ como uma constante para todo $\mathbf{d}$. Aplicando-se as Equações (3.2) e (3.3) na Equação (3.1), obtém-se a equação final para se computar a probabilidade de que o documento i pertença à classe C:

$$P(f_k \mid c) = \rho\Big(1 - \prod_i \big(1 - \text{class}_i(k, \mathscr{C}, \overline{\mathscr{C}})\big)\Big) \tag{3.4}$$

onde $\rho = \frac{P(\mathbf{d})}{P(c)}$ é uma constante de normalização e $\mathbf{d}$ é o estado onde somente os documentos definidos como pertencentes à categoria C estão ativos.

Capítulo 4

Experimentos

Para a realização dos experimentos, foi seguida uma metodologia adotada por P. Calado e M. Cristo [3]. Foram usadas as mesmas ferramentas (*Rainbow* [17] e *LIBSVM* [7]) que disponibilizaram os programas executáveis dos algoritmos kNN, Naive Bayes e SVM. Também foi utilizada a base de dados do *Cadê* [10].

4.1 A coleção de teste

Para avaliar o desempenho das fontes de evidência na classificação e os efeitos de sua combinação, os experimentos foram executados utilizando-se um conjunto de páginas *web* já classificadas em categorias, extraído do diretório do *Cadê*.

O *Cadê* é um diretório formado por páginas brasileiras que foram classificadas manualmente por especialistas. O conteúdo dessas páginas foi obtido a partir do banco de dados de páginas coletadas pela máquina de busca do *TodoBR*[1]. P. Calado e M. Cristo construíram duas sub-coleções a partir dos dados disponíveis neste diretório: *Cade12* e *Cade188*. Essa mesma divisão foi utilizada nos experimentos. A primeira sub-coleção é formada pelas 12 categorias do primeiro nível do *Cadê* (serviços, sociedade, lazer, informática, saúde, educação, internet, cultura, esportes, notícias, ciências e compras *on line*), perfazendo um total de 44.099 documentos.

Já o subconjunto *Cade188* consiste nas 188 categorias do segundo nível (Biologia, Química, Dança, Música, Escolas, Universidades, etc.), com 42.123 páginas. Cada página *web* foi classificada como fazendo parte de uma única categoria. As duas sub-coleções têm um vocabulário

[1]Disponível em `http://www.todobr.com.br/`.

de 191.962 (*Cade12*) e 168.257 (*Cade188*) palavras-chave, depois de removidas as *stop words*. As Figuras 4.1 e 4.2 mostram a distribuição de documentos nas classes para as sub-coleções *Cade12* e *Cade188*, respectivamente. A Figura 4.3 exibe um comparativo da distribuição dos documentos em ambas as sub-coleções.

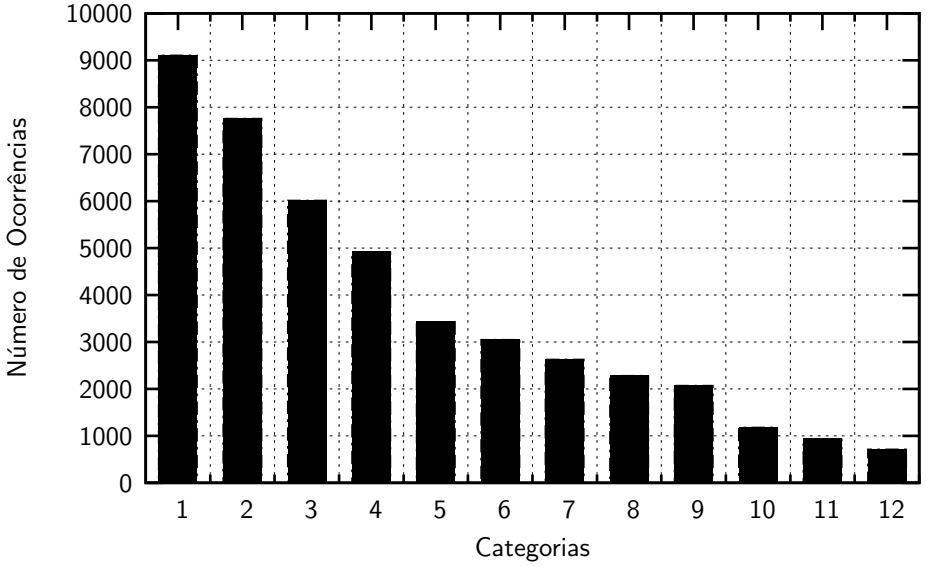

Figura 4.1: Distribuição de documentos nas classes da sub-coleção *Cade12*

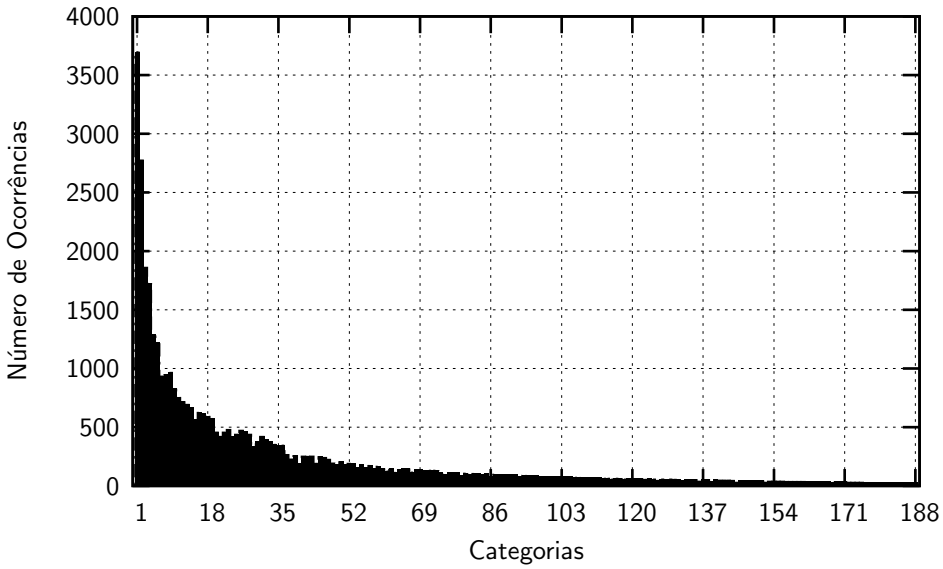

Figura 4.2: Distribuição de documentos nas classes da sub-coleção *Cade188*

As informações sobre os *links* relacionados às páginas do *Cadê* foram também coletadas da coleção do *TodoBR*. O *TodoBR* fornece 40.871.504 *links* entre as páginas *web* (em média 6,9 *links* por página), sendo que os *links* que conectavam diferentes páginas de um mesmo site foram

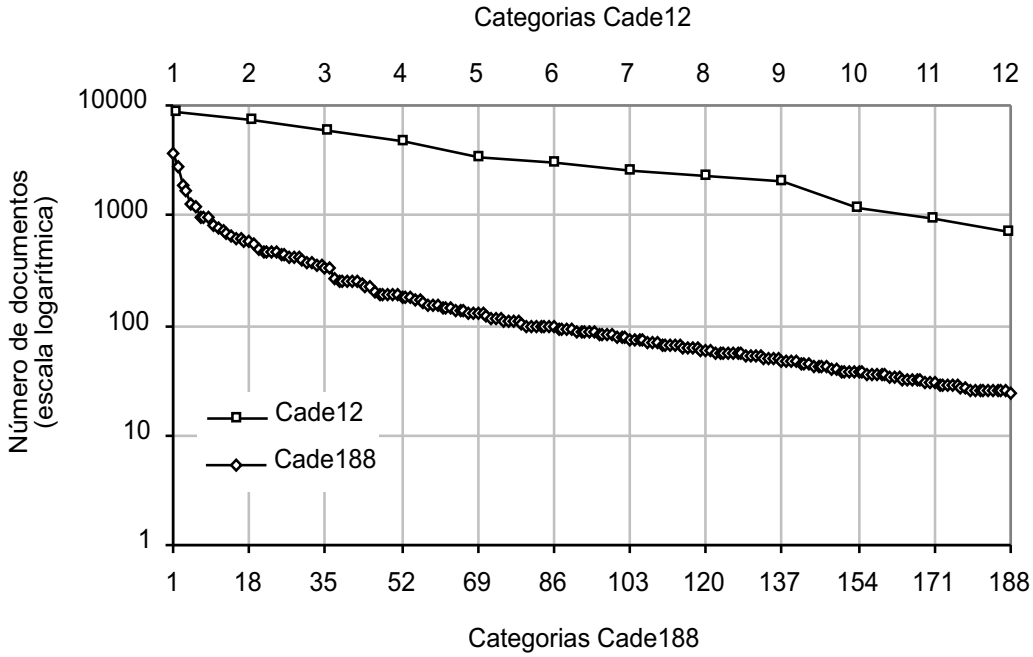

Figura 4.3: Comparação da distribuição de documentos nas classes das sub-coleções *Cade12* e *Cade188*.

descartados. Os *links* foram divididos em dois tipos: *internos* e *externos*. Os *links internos* foram definidos como os *hyperlinks* para páginas dentro do diretorio do *Cadê*, enquanto que os *links externos* foram aqueles que apontavam para as páginas externas ao *Cadê*, porém pertencentes à coleção do *TodoBR*. Esta divisão é importante para verificar se a informação externa fornecida pelo *TodoBR* pode ser usada para melhorar os resultados da classificação.

É interessante ainda discutir os conceitos de *in-links* e *out-links*. Os *in-links* de uma página p são todas as outras páginas que possuem *links* para p. Já os *out-links* de p são as páginas para as quais p aponta, ou seja, são os documentos para onde p possui link. Vale ressaltar que as páginas externas ao diretório *Cadê* consistem em uma rica fonte de informação de *links*. Cerca de 96% das páginas do *Cadê* são apontadas por páginas externas (*in-links* externos) enquanto que menos de 4% apontam para páginas externas (*in-links* internos). Este foi um dos motivos para a escolha desta coleção para a realização dos nossos experimentos. Como o *Cadê* era um subconjunto do *TodoBR* e já dispunha destas informações de *links* externos, não houve necessidade de se coletar páginas *web* e nem de acessar bancos de dados de outras máquinas de busca. Desta forma, o processo de classificação utilizando estrutura de *links* foi bastante facilitado. A Tabela 4.1 resume os dados sobre os *links*.

Para os experimentos, as sub-coleções foram divididas em três bases de dados de acordo

Estatísticas	Número de *links*
Links Internos	3.830
Links de páginas externas	570.337
Links para páginas externas	5.894
Páginas do *Cadê* sem *in-links*	1.625
Páginas do *Cadê* sem *out-links*	40.723

Tabela 4.1: Estatísticas de *links* da coleção *Cadê*.

com critérios de evidência: *CadeTexto*, *CadeTitulo* e *CadeLink*. A *CadeTexto* contém o texto dos documentos, sem as *tags* HTML e sem acentuação. A *CadeTitulo* contém o título dos documentos, ou seja, o conteúdo encontrado entre `<TITLE>` e `</TITLE>`, retirados também a acentuação gráfica. A *CadeLink* contém as palavras entre as *tags* `<A>` e `</A>` das páginas que apontam para a página a ser classificada, também tratada da mesma forma que as duas anteriores.

4.1.1 Ferramentas utilizadas

A implementação dos algoritmos utilizados para a classificação por conteúdo foi obtida através do pacote de ferramentas *Bow* [17] (kNN e *Naive Bayes*) e LIBSVM [7] (SVM). Esta última é uma biblioteca implementada em liguagem C, contendo a implementação do SVM e todos os programas de suporte a ele. A *Bow* ou *libbow* é uma biblioteca em linguagem C usada para a análise estatística do conteúdo textual e programas de recuperação de informação. Para os experimentos, foi utilizada a *Rainbow*, que é uma biblioteca de rotinas, disponibilizadas pela *Bow* [17], para trabalhar especificamente com classificação de documentos.

Rainbow é um pacote de programas que executam uma classificação estatística de texto e funcionam basicamente em dois passos: (i) leitura dos documentos de treinamento e criação de um modelo com estatísticas sobre esse conjunto de dados; (ii) utilização deste modelo para executar a classificação.

4.1.2 Formas de avaliação dos resultados

O desempenho dos métodos apresentados foi avaliado usando-se as medidas convencionais de precisão, revocação e média F_1. Os resultados de cada classificação foram resumidos em uma matriz que contém a distribuição dos documentos em cada classe. Nesta matriz estão determinados o número da execução ou *run* (*trial*) do experimento e a quantidade de documentos

corretamente classificados nas categorias.

4.1.3 Descrição dos experimentos

Os experimentos foram divididos em quatro partes: (i) Classificação das bases de dados, divididas de acordo com critérios de evidências (texto, título e conteúdo dos *links*) utilizando os três algoritmos (kNN, Naive Bayes e SVM); (ii) Uma outra classificação das três bases só que agora utilizando a estrutura de *links* das páginas; (iii) Combinação das classificações de todas as fontes de evidências; (iv) Avaliação dos resultados.

4.1.3.1 Classificação utilizando kNN, Naive Bayes e SVM

Não é objetivo deste trabalho avaliar o desempenho dos três algoritmos de classificação por conteúdo (kNN, Naive Bayes e SVM), entretanto, ao longo da apresentação dos resultados algumas considerações interessantes sobre o comportamento de cada algoritmo serão discutidas.

Antes de se iniciar o processo de classificação foi necessário criar um modelo com estatísticas dos documentos para cada coleção. Para isso, os documentos foram submetidos a um tratamento que consistiu em:

1. Transformar todos os caracteres para *lowercase* (minúsculo);
2. Ignorar todos os caracteres entre "<" e ">" dos arquivos HTML;
3. Retirar todos os acentos e caracteres da tabela ASCII estendida;
4. Montar uma *stoplist*, que é uma lista com palavras comuns (*stopwords*), como "the", "of", "is", etc;
5. Criar um vocabulário com as palavras do conjunto de treinamento, com tamanho reduzido utilizando-se o método de *information gain*[2] [18].

Os conjuntos de treinamento e de testes foram determinados através do processo de *cross-validation* ou *validação cruzada* [11]. O processo de *10-fold-cross-validation* é um método estatístico que visa minimizar o erro inerente a uma comparação. Neste método, a coleção é dividida aleatoriamente em dez partes. Seguem-se então dez rodadas de classificação em que, a

[2]Medida usada para se selecionar o melhor atributo de cada classe dada. É uma propriedade estatística que mede a relevância de um atributo na classificação do conjunto de treinamento.

cada rodada, uma parte diferente da coleção é usada como conjunto de teste enquanto que as nove outras restantes são usadas como conjunto de treino. O resultado final de cada experimento representa a média das dez rodadas.

Foram mantidos os mesmos parâmetros dos experimentos de P. Calado e M. Cristo [3] para a classificação utilizando os algoritmos kNN, Naive Bayes e SVM. Para o kNN foram usados os 15.000 primeiros termos considerados pelo método *information gain* como características importantes para o modelo. O k escolhido foi 30, pois foi o valor testado, dentro do intervalo $\langle 5, 10, 30, 50, 80, 100 \rangle$, que melhor se adaptou ao tamanho da coleção. Para o Naive Bayes utilizou-se os 8.000 primeiros termos dados pelo *information gain* [18]. Já o SVM trabalhou com o valor de 5.000 termos para a seleção de características. Esses parâmetros foram definidos através de experimentos com vários valores e escolhidos aqueles que geraram os melhores resultados para cada algoritmo.

Realizaram-se três classificações para cada algoritmo, uma para cada base de textual: Texto, Título e ContLinks. Os documentos das classes continham ou somente o texto, ou somente o título, ou somente o conteúdo dos *links* das páginas.

As Tabelas 4.2 e 4.3 resumem os resultados da média F_1 da classificação obtida pelos algoritmos kNN, Naive Bayes e SVM, usados isoladamente para as três bases de dados (Texto, Título e ContLinks). Os resultados que alcançaram melhor valor estão mostrados em negrito.

Métodos	Texto	Título	ContLinks
kNN	39,72	42,20	**31,72**
NB	38,96	46,55	30,98
SVM	**39,96**	**46,71**	31,07

Tabela 4.2: Classificação por conteúdo textual – *Cade12*

Métodos	Texto	Título	ContLinks
kNN	22,40	**32,17**	**19,94**
NB	22,19	30,98	18,61
SVM	**23,78**	32,10	18,77

Tabela 4.3: Classificação por conteúdo textual – *Cade188*

A classificação por conteúdo textual, como esperado, mostrou resultados fracos, indicando que o texto das páginas *web* não fornecem informação suficientemente boa para uma classificação confiável de documentos. Todavia, é interessante atentar para os resultados utilizando-se o Título e o ContLinks (conteúdo dos *links*). Os resultados para a classificação utilizando Título

foram sempre melhores do que o Texto e o ContLinks, para os três algoritmos. Isto significa que as palavras do Título caraterizam melhor a classe a que pertence cada documento do que as outras fontes de evidência. Um motivo para tal fato ocorrer é que, apesar de que nem todas as páginas possuírem títulos, aquelas que o tem trazem entre suas *tags* `<TITLE>` e `</TITLE>` termos que resumem o assunto tratado pela página.

Os valores foram muito mais baixos para a coleção *Cade188*, pois os classificadores tendem a não ter um desempenho muito bom em coleções onde a distribuição das classe é mais espalhada. O melhor resultado foi alcançado pelo SVM na coleção *Cade12*, com o Título, obtendo-se o valor para média F_1 de 46,71. Já para o diretório *Cade188*, o melhor valor foi conseguido pelo Naive Bayes, com F_1 de 32,17, também utilizando o Título.

4.1.3.2 Classificação utilizando Estrutura de *Links*

Os resultados das classificações utilizando estrutura de *links* foram fornecidos por P. Calado e M. Cristo [3]. Os arquivos contendo informações *links*, baseados nas quatro medidas de similaridade (Acoplamento, Amsler, Co-citação e Companion), foram utilizados na classificação dos documentos feita novamanente pelo kNN. O kNN realizou a classificação utilizando a estrutura de *links*, com $k = 30$, e estabelecendo o *ranking* de classificação da classe utilizando como medida de similaridade o Acoplamento, Amsler, Co-citação e o Companion.

As Tabelas 4.4 e 4.5 apresentam os resultados da classificação por estrutura de *links*, usando-se os *links* internos e externos. Os maiores valores de cada fonte de evidência estão enfatizados em negrito.

Medidas de Similaridade	Links Int	Links Ext
Acoplamento	21,81	22,05
Amsler	**23,03**	40,12
Co-citação	22,42	**86,80**
Companion	22,66	84,32

Tabela 4.4: Classificação por estrutura de *links* – *Cade12*

Medidas de Similaridade	Links Int	Links Ext
Acoplamento	8,33	8,55
Amsler	**9,36**	85,71
Co-citação	9,04	**85,83**
Companion	8,99	81,95

Tabela 4.5: Classificação por estrutura de *links* – *Cade188*

Para as medidas de similaridade de *links*, quando somente os *links* internos são considerados, as informações são claramente insuficientes, trazendo assim valores muito baixos de F_1. Considerando somente os *links* internos, muito das informações da estrutura de *links* da coleção é perdida. Como mostra a tabela 4.1, aproximadamente 98% da informação de *links* da coleção é oriunda de páginas externas.

Entretanto, quando os *links* externos são considerados, as informações de *links* sozinhas já são suficientes para se obter resultados de classificação bem acima daqueles conseguidos pelos classificadores baseados em texto. Para a coleção *Cade12*, os melhores resultados foram alcançados usando a Co-citação e o Companion, com 86,80 e 84,32 pontos na média F_1, respectivamente. Para a coleção *Cade188*, o Companion teve o melhor desempenho, com 69,6 pontos em F_1.

O Acoplamento gerou os valores mais baixos de F_1 comparando-se com as outras medidas restantes. Tal fato não surpreende, pois esta medida baseia-se unicamente na informação de *out-links* e, como mostrado na Seção 4.1, mais de 90% das páginas não possuem nenhum *out-links*. Desta forma, visto que a maioria dos *links* é de páginas externas às páginas na coleção, pode-se esperar que as medidas que empregam os *in-links* obtenham um melhor desempenho na classificação.

4.2 Resultados das combinações

Nesta seção serão analisados os efeitos das combinações de todas as fontes de evidências, utilizando o modelo de Rede Bayesiana proposto na Seção 3.2. Os resultados das combinações serão divididos em duas partes.

Primeiramente, realizou-se a combinação das bases de dados (Texto, Título e ContLinks) para cada classificador baseado no conteúdo. As Tabelas 4.6 e 4.7 exibem o resultado da média F_1 para estas combinações. Novamente, os melhores valores, para cada combinação, são exibidos em negrito.

Combinação	*k*NN	NB	SVM
Texto + Título	45,27	**51,86**	42,35
Texto + ContLinks	**46,67**	44,71	33,58
Título + ContLinks	47,78	**48,86**	36,76
Texto + Título + ContLinks	46,81	**53,73**	37,76

Tabela 4.6: Combinação dos métodos baseados no conteúdo textual – *Cade12*

Combinação	*k*NN	NB	SVM
Texto + Título	34,64	**33,07**	26,48
Texto + ContLinks	**33,32**	28,45	18,14
Título + ContLinks	37,85	**33,19**	22,93
Texto + Título + ContLinks	**36,45**	35,72	21,90

Tabela 4.7: Combinação dos métodos baseados no conteúdo textual – *Cade188*

A segunda parte traz todas combinações realizadas entre as diferentes fontes de evidência, para as duas coleções: *Cade12* e *Cade188*. Cada tabela equivale às combinações por classificador. O maior valor conseguido por todas as combinações está apresentado em negrito.

Combinação	Acoplamento		Amsler		Co-citação		Companion	
	int	ext	int	ext	int	ext	int	ext
Texto	35,66	35,67	36,39	83,16	36,43	83,36	35,93	83,30
Título	41,32	41,45	41,86	79,42	41,67	79,66	41,58	81,12
ContLinks	30,92	31,08	31,75	85,90	31,31	**85,96**	31,44	84,59
Texto + Título	42,81	42,85	43,28	75,19	43,25	75,41	42,92	78,18
Texto + ContLinks	40,77	40,76	41,29	82,36	41,35	82,54	40,90	83,07
Título + ContLinks	43,92	44,03	44,39	78,63	44,25	78,86	44,10	80,79
Texto + Título + ContLinks	44,72	44,77	45,17	74,61	43,13	74,83	44,83	77,80

Tabela 4.8: Combinação *k*NN e medidas de similaridade de *links* – *Cade12*

Combinação	Acoplamento		Amsler		Co-citação		Companion	
	int	ext	int	ext	int	ext	int	ext
Texto	21,19	21,19	21,85	80,45	21,97	80,66	21,96	79,64
Título	31,59	31,65	31,99	78,03	31,95	78,22	31,58	78,93
ContLinks	19,85	20,00	20,45	84,85	20,32	**85,00**	20,12	82,61
Texto + Título	32,34	32,40	32,68	73,09	32,75	73,30	32,26	75,54
Texto + ContLinks	27,86	27,83	28,26	79,58	28,39	79,81	27,85	79,63
Título + ContLinks	34,69	34,70	34,97	77,12	34,95	77,33	34,59	78,61
Texto + Título + ContLinks	34,72	34,77	35,00	72,32	35,08	72,56	34,64	75,29

Tabela 4.9: Combinação *k*NN e medidas de similaridade de *links* – *Cade188*

Novamente, é observado que os resultados que usam somente os *links* internos são geralmente fracos. Embora a combinação mostre uma melhoria sobre o uso dos *links* internos, os valores de F_1 continuam abaixo daqueles que foram conseguidos pela classificação baseada em conteúdo. Devido à falta das ligações internas, as medidas da similaridade introduzem muito ruído no processo da classificação.

Os resultados mostram alguma melhoria quando são empregados os *links* externos. Entretanto, e embora haja uma melhoria grande sobre o uso isolado dos algoritmos para classificação baseada em conteúdo, nem sempre a combinação pode substituir os resultados das medidas da similaridade de *links*. Para a coleção *Cade12*, Amsler e a Co-citação obtiveram um ganho

Combinação	Acoplamento		Amsler		Co-citação		Companion	
	int	ext	int	ext	int	ext	int	ext
Texto	39,11	39,07	39,45	78,72	39,76	78,86	39,52	80,55
Título	46,58	46,61	47,02	87,22	46,93	**87,37**	46,65	86,97
ContLinks	30,77	30,99	31,43	80,93	31,07	80,98	31,26	80,43
Texto + Título	51,65	51,61	51,93	79,59	51,95	79,73	51,76	82,03
Texto + ContLinks	44,82	44,78	45,17	76,17	45,18	76,30	44,96	78,64
Título + ContLinks	48,52	48,56	48,76	81,90	48,70	82,06	48,47	82,87
Texto + Título + ContLinks	53,58	53,55	53,76	77,19	53,77	77,33	53,59	79,86

Tabela 4.10: Combinação Naive Bayes e medidas de similaridade de *links* – *Cade12*

Combinação	Acoplamento		Amsler		Co-citação		Companion	
	int	ext	int	ext	int	ext	int	ext
Texto	22,23	22,22	22,82	70,29	22,82	70,40	22,67	73,13
Título	30,04	30,03	30,48	82,99	30,50	**83,18**	30,13	82,54
ContLinks	18,72	18,86	19,34	77,97	19,21	78,12	19,08	76,87
Texto + Título	32,72	32,71	33,05	69,87	33,05	69,98	32,96	73,74
Texto + ContLinks	28,46	28,45	28,82	66,92	28,84	67,02	28,71	70,46
Título + ContLinks	32,34	32,34	32,66	76,14	32,70	76,32	32,39	77,20
Texto + Título + ContLinks	35,47	35,46	35,70	66,72	35,71	66,84	35,62	70,91

Tabela 4.11: Combinação Naive Bayes e medidas de similaridade de *links* – *Cade188*

de menos de um 1 ponto na média F_1, quando combinados com o classificador kNN. Para a *Cade188*, e quando combinadas com o SVM e Naive Bayes, essas medidas renderam uma perda de até 2,72 pontos. O algoritmo Companion, quando usado isoladamente teve um desempenho sempre melhor do que quando combinado com um classificador.

Em relação às bases para classificação textual (Texto, Título e ContLinks), o Título mantém uma maior regularidade para os valores alcançados. Seus resultados são sempre maiores do que a combinação usando somente texto e, na sua maioria, melhor que a combinação utilizando o ContLinks. Com exceção do kNN, cujos melhores resultados de combinação são obtidos por ContLinks+Co-citação, todos os outros algoritmos (Naive Bayes e SVM) trazem como melhor combinação Título+Co-Citação. No caso do SVM, o Título combinado ao Texto mais a

Combinação	Acoplamento		Amsler		Co-citação		Companion	
	int	ext	int	ext	int	ext	int	ext
Texto	24,49	24,71	25,69	87,49	25,16	87,59	25,38	87,41
Título	24,59	24,85	27,75	87,51	25,18	87,54	25,44	85,37
ContLinks	22,36	22,64	23,56	86,66	22,92	86,72	23,28	84,62
Texto + Título	24,45	27,67	28,58	87,99	28,12	**88,08**	28,26	86,31
Texto + ContLinks	23,84	24,10	25,01	86,78	24,40	86,83	24,75	85,19
Título + ContLinks	24,66	24,90	25,77	86,94	25,15	86,97	25,49	85,33
Texto + Título + ContLinks	27,16	27,35	28,26	87,27	27,72	87,35	27,96	86,05

Tabela 4.12: Combinação SVM e medidas de similaridade de *links* – *Cade12*

Combinação	Acoplamento		Amsler		Co-citação		Companion	
	int	ext	int	ext	int	ext	int	ext
Texto	8,34	8,57	9,39	85,70	9,06	85,82	9,03	82,09
Título	8,34	8,57	9,38	85,72	9,06	**85,84**	9,02	82,15
ContLinks	8,34	8,56	9,38	85,70	9,05	85,83	9,01	82,04
Texto + Título	8,34	8,58	9,38	85,70	9,05	85,82	9,02	82,15
Texto + ContLinks	8,34	8,56	9,38	85,68	9,05	85,80	9,02	82,09
Título + ContLinks	8,33	8,56	9,38	85,70	9,05	85,82	9,02	82,11
Texto + Título + ContLinks	8,34	8,56	9,38	85,68	9,06	85,80	9,02	82,13

Tabela 4.13: Combinação SVM e medidas de similaridade de *links* – *Cade188*

Co-citação obteve 88,08 de média F_1, sendo este o melhor resultado de todas as combinações possíveis.

Na coleção *Cade188*, as melhorias sobre os resultados da classificação baseada em estrutura de *links* são mais evidentes. Para o Acoplamento, os ganhos foram significativos, como mostra a Tabela 4.14, devido ao desempenho mais fraco quando usado sozinho. Os ganhos foram de até 26,91 pontos na média F_1. Para o Companion, o ganho chegou a ser de 5,02 na média F_1, quando combinado com o Naive Bayes.

A combinação das evidências, em geral, melhorou os resultados isolados. O ganho é considerável em relação à classificação baseada em texto. Por exemplo, considerando a combinação SVM (Texto+Título) + Co-citação, o valor de F_1 conseguido foi 88,08 na média F_1. Enquanto que o melhor resultado individual do SVM (Título) obteve 46,71, conseguindo-se então um ganho de 41,37 pontos.

Já para as medidas de similaridade os ganhos são discretos. Para a Co-citação, por exemplo, tem-se como valor individual 86,80 na média F_1. A sua melhor combinação foi Co-citação+SVM (Texto+Título) com 88,08, e isso oferece um ganho de 1,28 pontos.

O Tabela 4.15 mostra a combinação dos classificadores kNN, Naive Bayes e SVM com Título, Texto e ContLinks, para a Co-citação, usando *links* externos (*Cade12*). Essa medida de similaridade foi escolhida por apresentar os melhores resultados, tanto isoladamente quanto submetida a combinações. Na Figura (4.4), montado com os valores da Tabela 4.15, percebe-se claramente

Classificador	Combinação	Valor	Ganho
Estrutura de Links (Acoplamento)	Classificação individual	8,55	—
kNN	Texto + Título + ContLinks	34,77	26,22
NB	Texto + Título + ContLinks	**35,46**	26,91
SVM	Texto + Título	8,58	0,03

Tabela 4.14: Melhores resultados para o Acoplamento – *Cade188* (Link Ext)

Algoritmo	*k*NN			Naive Bayes			SVM		
Coleção	$\overline{x}$	min	max	$\overline{x}$	min	max	$\overline{x}$	min	max
Texto	82.0	80.7	83.4	74.6	70.4	78.9	86.7	85.8	87.6
Título	78.9	78.2	79.7	85.3	83.2	87.4	86.7	85.8	87.5
ContLinks	85.5	85.0	86.0	79.6	78.1	81.0	86.3	85.8	86.7
Texto + Título	74.4	73.3	75.4	74.9	70.0	79.7	86.9	85.8	88.1
Texto + ContLinks	81.2	79.8	82.5	71.7	67.0	76.3	86.3	85.8	86.8
Título + ContLinks	78.1	77.3	78.9	79.2	76.3	82.1	86.4	85.8	87.0
Texto + Título + ContLinks	73.7	72.6	74.8	72.1	66.8	77.3	86.6	85.8	87.3

Tabela 4.15: Média aritmética dos resultados, para os três algoritmos, com a Co-citação e *links* externos – *Cade12*

que o SVM obtém os melhores valores das combinações. Isso já era esperado, pois esse algoritmo sempre manteve um desempenho muito bom na classificação individual e na maior parte das combinações. O Naive Bayes fica em segundo lugar com os melhores valores. Cabe aqui uma análise interessante sobre esses resultados: apesar do SVM ter alcançado os melhores valores, sua diferença em pontos para os resultados do Naive Bayes não é tão significativa, principalmente, quando combinado com o ContLinks. O problema é que o SVM possui um alto custo de processamento, o que não ocorre com o Naive Bayes, e isso faz com que a relação *custo* × *benefício* seja mais favorável às classificações realizadas por esse algoritmo. Para exemplificar esta questão, pode-se considerar o tempo gasto nos experimentos com ambos os algoritmos: o SVM levou cerca de 30 horas para computar a classificação com as três bases (Título, Texto e ContLinks), utilizando 1.024MB de memória RAM. Já o Naive Bayes, utilizando somente 512MB de RAM, demorou menos de 10 minutos para realizar a mesma classificação.

A Tabela 4.16 apresenta os melhores resultados conseguidos pelos experimentos de P. Calado e M. Cristo [3], em comparação com os valores obtidos pelo nosso trabalho. Considera-se a seguinte notação:

- *Experimento A (Experim A)*: resultado dos experimentos de P. Calado e M. Cristo, para a coleção *Cade12*, usando *links* externos, para a combinação dos três algoritmos com a medida Co-citação, utilizando o texto das páginas.

- *Experimento B (Experim B)*: resultados dos experimentos realizados neste trabalho, com a mesma combinação do Experimento B, só que utilizando como base o Título e o Conteúdo dos *Links*.

Nota-se que a maioria dos resultados do Título foi melhor do que os experimentos somente

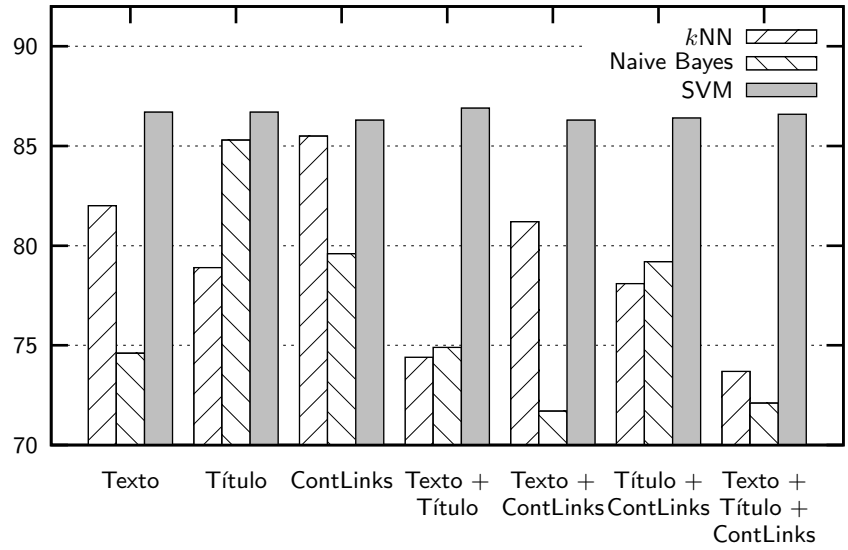

Figura 4.4: Gráfico: Média aritmética dos resultados, para os três algoritmos, com a Co-citação e *links* externos – *Cade12*

com o Texto, com exceção do kNN. O ContLinks alcançou, para todos os algoritmos consideráveis melhorias. O ganho foi considerável para as duas fontes de evidência, sendo que o maior valor para o Título foi de 87,54 pontos em F_1, para o SVM, com um ganho de 9,65 pontos. O ContLinks alcançou o maior valor também para o SVM com 86,72 pontos com um ganho de 8,83 pontos em relação aos valores conseguidos pelo Experimento A.

A utilização das fontes de evidências (Título e ContLinks) introduzidas nesta dissertação sugere alguns benefícios para a tarefa de classificação de documentos *web*. Nos experimentos, os resultados com o Título das páginas foram sempre melhores do que os testes utilizando somente o Texto. Além disso, o custo para se computar a classificação usando somente o Título é menor do que quando se utiliza todo o Texto de uma página, pois a quantidade de informação manipulada é bem menor. Essas análises colocam o Título como uma evidência de conteúdo textual importante para a classificação final.

Já o ContLinks apresentou individualmente valores inferiores ao Texto. Entretanto, na

Algoritmo	Experim A	Experim B	
	Texto	Título	ContLinks
kNN	81,55	79,66	85,96
Naive Bayes	59,03	87,37	80,98
SVM	77,89	87,54	86,72

Tabela 4.16: Comparação entre os Experimentos A e B – *Cade12*

combinação de alguns casos, o ContLinks conseguiu elevar os valores, trazendo relativa melhoria aos resultados. Isso pode ser verificado na Tabela 4.9, para a combinação Título + ContLinks. Tais conclusões confirmam que o título HTML das páginas e o conteúdo dos *links* das páginas que apontam para a página a ser classificada apresentam-se como evidências que podem trazer melhorias no resultado da classificação final, se forem utilizadas somente informações de conteúdo textual.

Capítulo 5

Conclusão e Trabalhos Futuros

Este trabalho realizou estudos sobre os efeitos da combinação de diferentes fontes de informação na classificação de documentos *web*. P. Calado e M. Cristo [3] propuseram um método para combinar classificação por estrutura de *links* com classificação obtidas por classificadores tradicionais baseados em informações de conteúdo textual. Tomando-se como base estas pesquisas, a proposta desta dissertação foi incluir fontes de informação já testadas em outros trabalhos de classificação [34, 25], realizar uma classificação utilizando como conteúdo textual estas novas fontes e, por fim, combiná-las aos resultados de uma classificação baseada em informações de *links*. Para isso, foi seguido o modelo de Redes Bayesianas proposto por [3], que foi estendido nesta proposta para suportar a inclusão dessas novas evidências.

A principal diferença deste trabalho para as pesquisas de P. Calado e M. Cristo [3] foi a inclusão de duas novas fontes de informação, além do Texto, para serem tratadas como conteúdo textual: Título e Conteúdo dos *Links* (das páginas que apontam para a página a ser classificada). O objetivo desta dissertação foi analisar o impacto da inclusão destas novas fontes de evidência na combinação com os resultados de uma classificação baseada na estrutura de *links* de páginas *web*.

Para a classificação baseada na estrutura dos *links*, assim como em [3], foram utilizadas quatro medidas de similaridade: Acoplamento, Amsler, Co-citação e Companion. Foram utilizados três classificadores tradicionais que usaram as informações baseadas em texto: kNN (k–Nearest-Neighbor), Naive Bayes e Support Vector Machine (SVM). Os experimentos foram realizados com a base de dados do diretório do *Cadê* [10].

A utilização do Título e do Conteúdo dos *links* (ContLinks) como evidências adicionais para

a classificação por conteúdo textual obteve resultados interessantes. O Título se mostrou um parâmetro de classificação superior ao Texto, que por sua vez teve um desempenho melhor do que o ContLinks. Em praticamente todas as combinações, e inclusive nos resultados individuais, o Título alcançou os melhores resultados. A exceção foi quando usado com o algoritmo kNN, cujos maiores valores foram determinados pelas combinações com o ContLinks. Isso vem confirmar a suposição de que as palavras contidas no Título HTML de páginas *web* podem fornecer ótimas informações sobre o assunto de que trata o referido documento.

Todas as medidas de similaridade de *links* testadas alcançaram um bom resultado quando usadas isoladamente, sendo que estes valores foram sempre melhores que os obtidos pelos classificadores tradicionais. Isto vem confirmar a importância das informações de *links* para a classificação de páginas *web*.

A Co-citação é a medida com os melhores resultados, sendo assim uma boa candidata para as soluções de RI na *Web* que pretendem usar os *links* para comparar páginas. A combinação das medidas baseadas em *links* com os classificadores baseados em conteúdo rendeu resultados heterogêneos. Embora tenha alcançado um ligeiro ganho sobre o uso isolado das medidas de similaridade de *links*, é questionável se o custo de se computar a combinação compensa realmente, pois os ganhos conseguidos são pequenos, com uma média de 2% em F_1. Somente as informações da estrutura de *links*, usando as medidas de similaridade já são capazes de classificar corretamente a maioria dos documentos. Assim, já eram esperadas melhorias discretas obtidas pela combinação com outras fontes de evidência.

O melhor desempenho dos três algoritmos tradicionais (kNN, Naive Bayes e SVM) foi alcançado pelo SVM, sendo que a melhor combinação (SVM (Texto+Título) + Co-citação (*Cade12*, *links* externos)) obteve 88,08 pontos em F_1. Entretanto, o Naive Bayes também alcançou bons resultados. A sua combinação Título + Co-citação (*Cade12*, *links* externos) conseguiu 87,37 pontos em F_1. Como o custo de se computar o Naive Bayes é bastante inferior ao do SVM (quase 300 vezes menor, para os experimentos realizados) e a diferença entre os ganhos não chegou a 1 ponto em F_1, pode-se concluir que a classificação utilizando este algoritmo oferece a melhor relação *custo* $\times$ *benefício*, independente da evidência textual escolhida (Título, Texto ou ContLinks).

Em resumo, pode-se concluir deste trabalho que a utilização de novas fontes de informação como Título e o ContLinks acrescentaram uma melhoria na classificação baseada em conteúdo

textual e um ligeiro ganho na combinação com as medidas de similaridade. Além disso, a estrutura de *links* apresenta-se como uma fonte de informação preciosa e eficaz para a classificação de documentos, trazendo sempre resultados melhores que os alcançados pelos algoritmos classificadores tradicionais. Nossa extensão do modelo de Rede Bayesiana proposto por P. Calado e M. Cristo [3] mostrou-se eficiente para gerar a combinação de todas as evidências testadas. As combinações trouxeram resultados melhores que os das evidências quando utilizadas isoladamente. Os experimentos com a sub-coleção *Cade12* resultaram em valores maiores que os executados com a sub-coleção *Cade188*, e isso pode gerar a seguinte hipótese: em quanto mais classes se dividirem as amostras, mantendo-se os mesmos documentos, menos as características selecionadas para identificar cada categoria serão capazes de descrevê-las. Tal suposição precisa ainda ser testada com outras coleções *web*.

5.1 Trabalhos futuros

Nesta seção serão apresentadas algumas sugestões de trabalhos futuros, que poderão dar continuidade aos experimentos realizados para esta dissertação. As propostas surgiram de idéias e necessidades observadas durante a fase de levantamento bibliográfico e por ocasião da realização dos testes e da análise de seus resultados.

Sugere-se o uso de outras coleções de referência para novas avaliações do desempenho da combinação de classificação utilizando informações de *links* e evidências baseadas em conteúdo textual. Isso seria interessante para se comprovar a eficácia do nosso trabalho, verificando-se o comportamento do método de classificação para outras coleções *web*.

Um outra sugestão seria a inclusão de pesos no modelo de Rede Bayesiana para cada evidência, pois, de acordo com as conclusões aqui obtidas, a presença de muitos *links* que não são relacionados diretamente ao assunto da página acabam categorizando vários documentos de modo errôneo. Por outro lado, utilizando-se a combinação com um classificador baseado em conteúdo, muitas destas páginas incorretamente classificadas foram rejeitadas, porém, muitos documentos corretamente classificados também foram descartados. A inclusão de pesos específicos para cada evidência pode sanar este problema. O desafio será, portanto, encontrar o peso ideal para cada fonte de informação. Uma forma seria a investigação de métodos que determinem automaticamente esses pesos. P. Calado [5] estudou os efeitos de cada peso nos resultados da

combinação de classificações e sugeriu que os pesos podem fornecer uma avaliação diferente de quando uma evidência deve ser favorecida em lugar de outra, dependendo da coleção testada.

Para a coleção Cade, utilizada neste trabalho, pode-se sugerir a realização de experimentos de classificação hierárquica, com o objetivo de verificar seu desempenho dos métodos aqui propostos na criação de diretórios de categorias, testando assim seu comportamento em diversas situações. Pode-se avaliar o comportamento das combinações de diferentes fontes de informação na construção de árvores hierárquicas de diretórios. Para isso utilizaremos a hierarquia das categorias do *Cadê*. Uma solução para esta classificação utilizando hierarquias, é realizar os experimentos em domínios mais reduzidos de páginas e categorias. Em cada nível, as subcategorias serão encaradas como um novo conjunto a ser classificado, com a diferença de que conterá um menor número de páginas, como sugerido em por F. Sebastiani [22]. A partir dos resultados deste experimento, pode-se avaliar a performance da combinação da classificação por conteúdo textual com a fornecida pelas estruturas dos *links*, quando aplicada ao conceito de classificação hierárquica.

Referências Bibliográficas

[1] Amsler, R. *Application of citation-based automatic classification.* Technical report, The University of Texas at Austin, Linguistics Research Center, 1972.

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

[3] Calado, P.; Cristo, M. A.; Moura, E. *et al. Combining link-based and content-based methods for Web document classification.* CIKM International Conference on Information and Knowledge Management, 2003.

[4] Calado, P.; Moura, E.; Ribeiro-Neto, B. *et al. Local versus Global Links Information.* ACM Transactions on Information Systems (TOIS), 2003.

[5] Calado, P. *Using Link Structure for Information Retrieval in the World Wide Web.* Tese de Doutorado, Departamento de Ciência da Computação – UFMG, 2004.

[6] Chakrabarti S.; Dom, B. & Indyk, P. *Enhanced hypertext categorization using hyperlinks.* Proceeding of the ACM SIGMOD International Conference on Management of Data, 1998.

[7] Chang, C-C. & Lin, C-J. *LIBSVM: a library for support vector machines.* Disponível em `http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf`, 2001.

[8] Cristo, M.; Calado P.; Moura, E. *et al. Link Information as a Similarity Measure in Web Classification.* 10th Symposium On String Processing and Information Retrieval, SPIRE 2003.

[9] Dean J. & Henzinger, M. R. *Finding related pages in the World Wide Web.* Computer Networks, 1999.

[10] Diretório do Cadê, `http://www.cade.com.br`, último acesso em agosto/2004.

[11] Friedl, H. & Stampfer E. *Cross-Validation.* Encyclopedia of Environmetrics, Volume I, Editores: A. El-Shaarawi, W. Piegorsch, Wiley Chichester. pp. 452–460, 2002.

[12] Joachims, T. *A statistical learning model of text classification for Support Vector Machines.* Annual ACM Conference on Research and Development in Information Retrieval, ACM Press, 2001.

[13] Joachims, T. *Text categorization with support vector machines.* 10*th* European Conference on Machine Learning, Springer-Verlag, 1998.

[14] Kang, I-H. & Chang, G. C. *Integration of multiples evidences based on a query type for web search.* Information Processing & Management, volume 40, Elsiever Science, 2004.

[15] Kessler, M. M. *Bibliographic coupling between scientific papers.* American Documentation, 1963.

[16] Kleinberg, J. M. *Authoritavite sources in a hyperlinked enviroment.* Journal of the ACM (JACM), 1999.

[17] McCallum, A. K. *BOW: a toolkit for statistical language modeling, text retrieval, classification and clustering.* Disponível em `http://www.cs.cmu.edu/~mccallum/bow`, 1996.

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

[19] Mota, D. F. *Classificação Automática de Documentos.* Dissertação de Mestrado. Departamento de Ciência da Computação – UFMG, 2001.

[20] Ribeiro-Neto, B. & Muntz, R. *A Belief Network Model for IR.* Annual ACM Conference on Research and Development inf Information Retrieval. Proceedings of the 19*th* Annual International ACM SIGIR, ACM Press, 1996.

[21] Rish, I. *An Empirical study of the naive Bayes classifier.* T.J. Watson Research Center. Workshop on "Empirical Methods in AI", IJCAI-01, 2001.

[22] Sebastiani, F. *Machine learning in automated text categorization.* ACM Surveys, 2002.

[23] Silva, I.; Ribeiro-Neto, B. *et al. Links-based and content-based evidential information in a belief network model.* 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000.

[24] Small, H. *Co-citation in the scientific literature: A new measure of relationship between two documents.* Journal of the American Society for Information Science, 1973.

[25] Sun A.; Lim, E-P. & Ng, W-K. *Web Classification using support vector machine.* Proceeding of the fourth international workshop on Web information and data management. ACM Press, 2002.

[26] TodoBR: Todo o Brasil na Internet, `http://www.todobr.com.br`, último acesso em agosto/2004.

[27] Tong, S. & Koller, D. *Support Vector Machine Active Learning with applications to text classification.* Journal of Machine Learning Research, 2002.

[28] Trevor, H. Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer Series in Statistics, 2001.

[29] Turtle, H. & Croft, W. B. *Evaluation of an Inference Network-Based Retrieval Model* ACM Transactions on Information Systems (TOIS).Volume 9, Issue 3, ACM Press, 1991.

[30] Vapnik, V. *Estimation of Dependences based on Empirical Data.* Springer, Verlag, 1982.

[31] Vapnik, V. *Statistical Learning Theory.* Wiley, Chichester, GB.; 1998.

[32] Yang, Y. *Expert network: Effective and efficient learning from human decision in text categorization and retrieval.* 17*th* Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 1994.

[33] Yang, Y. & Liu, Xin. *A re-examination of text categorization methods.* Proceedings of the 22*nd* Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 1999.

[34] Yang, Y.; Slattery, S. & Ghani, R. *A Study of approaches to hypertext categorization.* Journal of Inteligence Information Systems. Special Issue on Automated Text Categorization, 2002.

[35] Yang, Y.; Slattery, S. & Ghani, R. *Hypertext Categorization Using Hyperlink Patterns and Meta Data.* 18*th* International Conference on Machine Learning (ICML 2001), 2001.

[36] Yavuz, T. & Guvenir, H. A. *Application of k-Nearest Neighbor on feature projections classifier to text categorizartion.* Proceeding of ISCIS-98, 13*th* International Symposium on Computer and Information Sciences, 1998.