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

Usando Evidências Geográficas para Classificação Automática de Páginas Web

Maely da Silva Moraes

Manaus – Amazonas
Abril de 2005

Maely da Silva Moraes

Usando Evidências Geográficas para Classificação Automática de Páginas Web

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

Maely da Silva Moraes

Usando Evidências Geográficas para Classificação Automática de Páginas Web

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
Abril de 2005

Para Mateus.

Agradecimentos

A Deus, acima de tudo.

Aos meus pais, Miram Souza e Felipe França, pelo apoio incondicional.

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

Ao colega Pável Calado, pelas orientações imprescindíveis para a realização deste trabalho.

Ao professor e amigo Celso Rômulo, pelo incentivo e contribuição no decorrer de todo o trabalho.

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

Ao professor José Raimundo, pela força e apoio.

Aos colegas de curso: Daniel Fernandes, David Braga, Edson César, Eduardo Abinader, Juliana Marreiros, Ketlen Lucena, Keyla Ahnizeret, Leandro Galvão, Moisés Carvalho, Nívea Michelle Melo e Ville Caribas, pelo companheirismo.

Aos colegas de trabalho da Infraero: Paulo Cavalcante, Carlos Pereira e Harley Lima, pelo apoio e compreensão.

Ao meu namorado Ronaldo Francisco pela compreensão, apoio e incentivo na fase final deste trabalho.

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.

Eu pedi Força...
e Deus me deu Dificuldades para me fazer forte.
Eu pedi Sabedoria...
e Deus me deu Problemas para resolver.
Eu pedi Prosperidade...
e Deus me deu Cérebro e Músculos para trabalhar.
Eu pedi Coragem...
e Deus me deu Perigo para superar.
Eu pedi Amor...
e Deus me deu pessoas com Problemas para ajudar.
Eu pedi Favores...
e Deus me deu Oportunidades.
Eu não recebi nada do que pedi...
Mas eu recebi tudo de que precisava.

Autor desconhecido

Resumo

Neste trabalho é proposto um procedimento para determinação do escopo geográfico de páginas *Web* que leva em consideração diferentes fontes de informação a fim de permitir a classificação geográfica de páginas *web*, em vez de sites inteiros. Como as páginas muitas vezes apresentam pouquíssimos dados, foi proposto o uso do texto de páginas *Web* previamente classificadas como fonte de informação, combinado com a informação baseada na estrutura de *links* presentes nestas páginas. Experimentos com a coleção de páginas do sistema de busca *TodoBR* mostram que, para determinação do escopo geográfico de páginas *Web*, o ideal é que se combine a informação baseada no conteúdo textual com a informação baseada nos *links* das páginas *Web*.

Palavras-chave: Recuperação de Informação, Classificação Geográfica, Wold Wide Web.

Abstract

In this work, we present a procedure to determine the geographical scope of *Web* pages. The procedure uses different sources of information in order to obtain classification of individual pages, instead of whole sites. Since, in general, pages contain very little data, we propose, as a source of information, the use of text in pages previously classified. This is combined with the linked-based information contained in these pages. Experiments using the search engine *TodoBR* page collection have shown that, to determine the geographical scope of *Web* pages, to combining textual content-based information with linked-based information in *Web* pages lets to better results.

Sumário

1 Introdução **1**

1.1 Motivação e Objetivos 1

1.2 Trabalhos Relacionados 3

1.3 Principais Contribuições 5

1.4 Organização da Dissertação 5

2 Conceitos Básicos **6**

2.1 Escopo Geográfico de uma Página *Web* 6

2.2 Princípios da Classificação Automática 7

2.2.1 Coleção de Treino e Coleção de Teste 7

2.2.2 Formas de Avaliação 7

2.2.3 Seleção de Características 8

2.3 Entropia 9

2.4 Conceitos Estatísticos 9

2.4.1 Variáveis Aleatórias e Distribuição de Probabilidade 9

2.4.2 População e Amostra 10

2.4.3 Estimação de Parâmetros Populacionais 10

3 Escopo Geográfico de Páginas *Web* **11**

4 Classificação de Páginas *Web* **16**

4.1 Similaridade Baseada no Texto das Páginas *Web* 16

4.2 Combinando Fontes de Informação 18

5 Experimentos **20**

5.1 Coleções 20
5.1.1 Coleção de Treino 20
5.1.2 Coleção de Teste 23
5.2 Medidas de Avaliação 24
5.3 Resultados 24
5.3.1 Avaliação do uso das fontes de informação em separado e combinadas . . 25
5.3.2 Avaliação do uso do conteúdo textual completo das páginas *Web* da coleção de treino 28

6 Conclusão e Trabalhos Futuros **30**
6.1 Trabalhos Futuros 30

Referências Bibliográficas **32**

Lista de Figuras

3.1 Hierarquia geográfica considerada para a coleção de páginas da *Web* Brasileira. . . . 12

5.1 Distribuição do percentual de páginas *Web* pré-classifcadas por localidades. 23

5.2 Comparação das fontes de informação em separado e combinadas através da medida *F1* média em função do limiar r para a medida *Power*. 25

5.3 Exemplo 1 de onde o informação de CEP está disponível. 27

5.4 Exemplo 2 de onde o informação de CEP está disponível. 27

Lista de Tabelas

5.1 Estatísticas para o cálculo da proporção amostral de páginas corretamente classificadas usando a heurística de CEPs 22

5.2 Resumo dos resultados obtidos para as fontes de informação em separado e combinadas em função do limiar r para a medida $Power$. 26

5.3 Resumo dos resultados obtidos dentro da própria coleção de treino para as fontes de informação em separado e combinadas em função do limiar r para a medida $Power$. 28

Capítulo 1

Introdução

1.1 Motivação e Objetivos

Não é novidade a crescente popularização da *World Wide Web* (ou simplesmente *Web*) e o aumento do seu volume de dados nos últimos anos. No entanto, este vasto repositório de informação é bastante desorganizado, pois permite que seus usuários sejam também criadores de documentos, transformando a *Web* em um meio de publicação sem restrição.

A todo momento, páginas *Web* são incluídas, alteradas ou excluídas, e tudo isso sendo feito sem que haja qualquer padronização que facilite o acesso a informação provida por cada página. Essa volubilidade faz da *Web* um ambiente complexo, heterogêneo e dinâmico, tornando a tarefa de recuperar informações úteis, difícil e tediosa.

Hoje, os principais intermediários entre os usuários e a enorme quantidade de informação disponível na *Web* são sistemas de Recuperação de Informação (RI) conhecidos como máquinas de busca, onde os usuários submetem consultas através de palavras-chave, e obtêm como resposta documentos ordenados por relevância. Contudo, normalmente, os usuários não sabem como refinar suas consultas e a qualidade da busca pode ficar bastante comprometida. Daí a importância de se pesquisar recursos que possam melhorar o acesso à informação de interesse dos usuários sem exigir deles um esforço extra.

Existem várias relações entre os documentos da *Web* que podem ser de ordem prática para os usuários. A estrutura de *links*, por exemplo, é uma relação bastante explorada. Os usuários podem navegar através dos *links* das páginas *Web* e alcançar outras possivelmente relacionadas. No entanto, esta tarefa de navegação é muito cansativa. A análise da estrutura de *links* também

tem sido empregada para melhorar a ordenação por relevância das máquinas de busca .

Uma outra relação, menos explícita, é o tipo de assunto abordado nas páginas *Web*, principalmente das páginas que não possuem *links* entre si. Sendo possível, por exemplo, organizá-las em categorias, permitindo que os usuários naveguem sobre diversas páginas que discutem o mesmo assunto. Diretórios *Web*, tais como Yahoo![1], Open Directory Project[2] ou *Cadê?*[3], fazem extensivo uso deste tipo relação, porém, a classificação das páginas *Web* é realizada por especialistas e a um custo muito alto, devido à quantidade de páginas, variedade de assuntos e demanda de tempo.

Outro tipo de relação que pode ser interessante estudar é o *escopo geográfico*. Uma página *Web* pode estar relacionada a uma dada localização geográfica de diversas formas. Ela pode, por exemplo, conter informação sobre entidades do mundo real, situadas em um determinado local, ou prover informação de interesse para a comunidade que mora em uma localização ou na sua vizinhança. É fácil pensar em diversas aplicações envolvendo escopo geográfico de uma página *Web*. Os usuários da *Web* estão freqüentemente interessados em encontrar páginas de serviços, tais como restaurantes ou jornais, próximo à área residencial deles. A publicidade na *Web* também se beneficiaria, uma vez que as propagandas comerciais poderiam ser posicionadas em áreas previlegiadas das páginas *Web* cujo escopo geográfico fosse de interesse.

Explorar as relações existentes entre as páginas *Web* não seria tão difícil se diversos tipos de dados semânticos, que poderiam ser fornecidos facilmente pelos criadores das páginas, estivessem disponíveis. Um exemplo deste tipo de informação semântica poderia ser o escopo geográfico da página *Web*. A ausência de tal informação motivou o trabalho desta dissertação cujo objetivo é propor um método de como explicitar o escopo geográfico de páginas *Web* de maneira automática.

O método apresentado nesta dissertação atribui escopo geográfico a um conjunto inicial de páginas *Web* através de uma heurística simples baseada na ocorrência de Códigos de Endereçamento Postal (CEPs) no texto das páginas. A partir deste conjunto de páginas pré-classificadas, propaga-se a informação geográfica delas para outras páginas através da combinação de informação de *links* com informação textual das páginas *Web*. Experimentos realizados indicam que tal combinação pode ser útil neste caso.

[1] Veja `http://www.yahoo.com/`.
[2] Veja `http://dmoz.org/`.
[3] Veja `http://cade.com.br/`.

1.2 Trabalhos Relacionados

Embora muitos *Web sites* refiram-se a alguma localização geográfica as máquinas de busca atuais não estão bem preparadas para recuperar informação associada à localização geográfica. Este problema tem motivado várias pesquisas, já que muitas tarefas de RI da *Web* podem ser melhoradas através do conhecimento do escopo geográfico de uma página [9]. Nos próximos parágrafos desta seção será feita uma breve descrição de alguns trabalhos que procuram explorar informações geográficas de *sites*.

Em [3], é apresentado um estudo que classifica os *Web sites* como sendo de interesse local ou global. Um *site*, de interesse local é relevante para a comunidade dentro de um determindado limite geográfico como *sites* sobre restaurantes e teatros. Por outro lado, um *site* de interesse global é aquele relevante para toda comunidade *Web*, como *sites* de serviços bancários e lojas *on-line*. Neste trabalho os autores discutem várias heurísticas sobre como mapear um *Web site* a uma localização geográfica e também apresentam uma máquina de busca que leva em consideração evidências geográficas na ordenação por relevância dos seus resultados.

Um trabalho similar é apresentado em [11], onde diversas técnicas são propostas para reconhecer nos textos das páginas *Web* referências geográficas, tais como nomes de cidade, CEPs, números de telefone, entre outros. Esta informação é então transformada em coordenadas geográficas em um mapa. Os autores apresentam também um sistema experimental para navegação que permite o acesso aos *Web sites* através da proximidade geográfica e contexto espacial deles, onde o usuário inicia a busca pedindo para encontrar *sites* que sejam referentes a lugares na vizinhança de um *Web site* atualmente indicado.

Ding et al. [6] propuseram um algoritmo que classifica um dado *Web site* através do exame da distribuição geográfica de outros *sites* que apontam para ele e através do exame da distribuição de referências geográficas contidas no texto das páginas deste *site*. Esta dissertação faz uso de algumas das idéias propostas em [6] mas é diferente porque foi proposta a classificação automática de páginas *Web* individuais, em vez de *sites* completos. Além disso, é feita a combinação da informação derivada da estrutura de *links* com a informação do texto das páginas *Web*.

Das heurísticas dicutidas em [3, 11, 6], apenas a heurística de extração de CEPs foi empregada no trabalho desta dissertação. No entanto, a discussão variou desde o exame do texto das páginas

em busca de referências geográficas (nomes de cidades, CEPs, números de telefones, etc) a consulta de base de dados externas contendo informação sobre a localização dos servidores que hospedam *Web sites*. Estas heurísticas, embora eficazes, têm algumas limitações. Primeiro, elas permitem a classificação de um número limitado de páginas, já que muitas das páginas na *Web* contêm referências geográficas desconhecidas ou não são mencionadas em qualquer base de dados externa. Segundo, elas são geralmente limitadas a fornecerem informação sobre um *site* completo, em vez das páginas individuais deles. Isto é claramente inconveniente, visto que muitas páginas individuais dentro de um mesmo *site* podem pertencer a escopos geográficos diferentes.

Uma vez que a informação geográfica sobre páginas é conhecida, pode ser interessante também, determinar o escopo geográfico das consultas de usuários. Gravano et al. [7] propôs um algoritmo para determinar automaticamente a localização geográfica pretendida de uma dada consulta. Para alcançar isto, um conjunto de classificadores de documentos é aplicado sobre os resultados das consultas, baseando-se em referências a localizações geográficas encontradas no texto das páginas. Os resultados mostram que o método proposto pode determinar eficientemente o escopo geográfico de uma consulta, embora a maioria das consultas tenha poucas referências às localizações geográficas, sendo então consideradas do tipo global.

Vale a pena mencionar que existem hoje diversos serviços comerciais *on-line*, tais como MapQuest[4], Yahoo!Maps[5], ou InfoSpace[6], que fazem uso da informação geográfica. No entanto, são limitados a procurar por um conjunto predeterminado de serviços e dependem do trabalho humano para executar a classificação. O método proposto no trabalho desta dissertação pode contribuir para criação desse tipo de serviço com um custo muito menor.

Finalmente, o método apresentado nesta dissertação faz uso de uma combinação de fontes de informação, conteúdo textual e *links* de páginas *Web*. Notou-se que diversos trabalhos já têm mostrado que este tipo de combinação pode ser útil em diferentes tarefas de RI, tais como ordenção por relevância de documentos [2, 4, 10] e classificação [5, 8].

[4] disponível em `http://www.mapquest.com/`.
[5] disponível em `http://maps.yahoo.com/`.
[6] disponível em `http://www.infospace.com/`.

1.3 Principais Contribuições

Este trabalho apresenta um proposta de como determinar o escopo geográfico de páginas *Web* de maneira automática. As pricipais contribuições incluem os seguinte tópicos:

- Criação de uma coleção de treino utilizando-se uma heurística baseada na ocorrência de CEPs no texto das páginas *Web*, descrita na Seção 5.1.
- Estudo sobre como combinar as fontes de informação, conteúdo textual e *links*, na determinação do escopo geográfico de páginas *Web* de forma automática. A descrição da combinação é apresentada na Seção 4.2
- Realização de experimentos cujos resultados, apresentados na Seção 5.3, mostram que, para determinação do escopo geográfico de páginas *Web*, o ideal é que se combine a informação baseada no conteúdo textual com a informação baseada nos *links* das páginas *Web*, já que estas informações quando usadas em separado podem produzir resultados inferiores ao de sua combinação.

1.4 Organização da Dissertação

Esta dissertação está organizada em seis capítulos dos quais este é o primeiro. O capítulo 2 apresenta alguns conceitos necessários para o entendimento deste trabalho. O capítulo 3 descreve as principais idéias proposta por Ding et al. [6] que foram adotadas neste trabalho. O capítulo 4 apresenta o método proposto para classificar automaticamente páginas *Web* usando as evidências geográficas obtidas através da análise da estrutura de links e conteúdo textual das páginas individualmente. O capítulo 5 apresenta os experimentos realizados e análise dos resultados. Finalmente, no capítulo 6 são apresentadas as conclusões e sugestões de trabalhos futuros.

Capítulo 2

Conceitos Básicos

Neste capítulo é apresentada uma breve revisão dos conceitos básicos necessários para um melhor entendimento do método proposto nesta dissertação. Começando pela definição de escopo geográfico adotada neste trabalho. Prosseguindo com a discussão de alguns princícios básicos comuns a diferentes métodos de classificação automática e as métricas utilizadas para avaliação dos resultados. Continuando com o conceito de entropia empregado na estimativa do escopo geográfico. Finalizando com alguns conceitos estatísticos utilizados ao longo do trabalho.

2.1 Escopo Geográfico de uma Página *Web*

Quando alguém cria uma página *Web* tem em mente um público alvo (área geográfica) para ela. Este público pode ser bem limitado, geograficamente ou não. Por exemplo, o público alvo da página *Web* de uma pizzaria que faz entregas em domicílios até uma distância x da loja, é restrito à sua vizinhança. Agora, tome como exemplo, a página de um serviço de compra e venda de produtos pela Internet, o público alvo agora pode se estender por todo país, ou até pelo mundo. Partindo deste princípio, a noção de escopo geográfico de uma página *Web* pode ser extraída da distribuição geográfica do seu público alvo. Sendo assim, pode-se dizer que o escopo geográfico de uma página *Web* p é a área geográfica que o criador de p pretende alcançar [6].

Como a noção dada acima é muito subjetiva, tão subjetiva quanto a noção de relevância de documentos na área de Recuperação de Informação (comumente abreviada como RI), capturar o escopo geográfico de uma página *Web* com precisão, só é possível classificando-se manualmente

cada página *Web* de acordo com o escopo geográfico pretendido. O objetivo do método proposto nesta dissertação é estimar automaticamente o escopo geográfico de uma página *Web*, de forma que este se aproxime ao máximo da definição apresentada aqui. Esta tarefa de estimação pode ser considerada como uma tarefa de classificação automática de páginas *Web*, onde as categorias equivalem à localizações geográficas. Através das idéias propostas em [6] procedeu-se a classificação combinando duas fontes de informação: o conteúdo textual das páginas e a estrutura de *links* delas.

2.2 Princípios da Classificação Automática

O trabalho de classificação automática consiste em atribuir categorias pré-definidas a novos documentos, a partir de um conjunto de documentos de treino [20].

2.2.1 Coleção de Treino e Coleção de Teste

Os classificadores automáticos trabalham com uma coleção pré-classificada, geralmente, por um processo de classificação manual, cujos documentos são divididos em dois conjuntos, chamados de coleção de treino e coleção de teste.

A coleção de treino é um conjunto de documentos usado pelo classificador para identificar as características das categorias existentes na coleção. Tais categorias são as mesmas nas quais os novos documentos poderão ser classificados. Nesta dissertação as categorias são localizações geográficas do Brasil (vide descrição da hierarquia no Capítulo 3).

A coleção de teste é um conjunto de documentos utilizados para estimar qualidade do classificador. A partir da análise do classificador os documentos são alocados em uma ou mais categorias.

2.2.2 Formas de Avaliação

O objetivo de todo classificador é alocar corretamente os documentos na categoria onde ele é relevante. Sendo assim, para avaliá-lo é necessária a comparação do resultado obtido pelo algoritmo com a classificação original dos documentos. Nesta dissertação os experimentos foram conduzidos de duas maneiras: a primeira foi adotando uma uma coleção de teste, descrita na Seção 5.1, posteriormente submetida ao classificador, para então estimar a qualidade da

classificação gerada através de medidas comumente usadas em sistemas de RI, a *precisão* e a *revocação*. A segunda maneira foi adotando o método de *validação cruzada*[13] sobre a própria coleção de treino, onde os documentos pré-classificados, foram divididos em n conjuntos de documentos e o experimento foi repetido n vezes, sendo que, em cada uma das iterações um destes conjunto era definido como sendo de teste e os demais como conjunto de treino, minimizando, com dessa maneira, a variabilidade das estimativas de qualidade.

Considerando N como o conjunto de documentos relevantes e R o conjunto de documentos retornados pelo sistema, pode-se determinar os seguintes conceitos:

A *precisão* consiste na proporção de documentos retornados pelo sistema que podem ser considerados relevantes [14], podendo ser estimada pela Equação (2.1).

$$\text{Precisão} = \frac{|N \cap R|}{|R|} \tag{2.1}$$

A *revocação* consiste na proporção de documentos relevantes que foi retornada pelo sistema [14], podendo ser estimada pela Equação (2.2).

$$\text{Revocação} = \frac{|N \cap R|}{|N|} \tag{2.2}$$

No trabalho apresentado nesta dissertação foram utilizadas métricas de *precisão* e *revocação* adaptadas ao problema de estimação do escopo geográfico, descritas em [6] e apresentadas na Seção 5.2. Para simplificar a interpretação dos experimentos foi combinado *precisão* com *revocação* em uma única métrica definida como:

$$\text{F1} = \frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.3}$$

2.2.3 Seleção de Características

Muitos termos ou palavras de um documento não são suficientemente significativos para descrever o assunto de um texto, já alguns termos "trazem consigo"mais significado que muitos outros. A seleção de características tenta eliminar termos que não são representativos na caracterização do texto, isso porque considerar todos os termos de um documento pode prejudicar a eficência da categorização e, também, aumentar o custo computacional.

2.3 Entropia

O conceito de entropia empregado neste trabalho é baseado na idéia da teoria da informação [22]. Também chamada de entropia informacional, a entropia é a média da quantidade de informação contida em um conjunto de símbolos. De modo geral, a entropia informacional H de um conjunto de símbolos é dada pela Equação (2.4)

$$H = \sum_{i=1}^{m} p_i \times \log\left(\frac{1}{p_i}\right) \tag{2.4}$$

onde p_i é a probabilidade de ocorrência de cada um dos m símbolos diferentes em uma sequência de N símbolos, dada pela razão entre a frequência de ocorrência do símbolo i e o total de símbolos da sequência. Suas propriedades são:

- A entropia informacional de um conjunto de símbolos é sempre um número positivo ou nulo: $H \geq 0$. Ela é igual a zero se e somente se um dos valores da sequência tiver probalidade 1 e os demais tiverem probabilidades nulas.
- O valor máximo da entropia ocorre quando a distribuição dos m valores na sequência de N símbolos for uniforme, isto é, quando $p_i = \frac{1}{m}$. O valor máximo de H é $\log(m)$.

2.4 Conceitos Estatísticos

Os conceitos estatísticos empregados na estimação da proporção de acerto obtida na classificação de páginas *Web* através da heurística de extração de CEPs são definidos abaixo.

2.4.1 Variáveis Aleatórias e Distribuição de Probabilidade

A maioria das variáveis que se tem interesse de estudar é aleatória, ou seja, apresenta valores diferentes quando observadas em repetições feitas mesmo sob condições idênticas. Daí a importância de se estudar o conceito de variável aleatória antes de considerar resolução de problemas com o auxílio de técnicas estatísticas. A palavra aleatória indica que a variável está sujeita a variabilidade e exige para seu estudo o conhecimento de seus possíveis valores e da correspondente freqüência. Estes dois identificam a distribuição de probabilidade da variável aleatória. As variáveis aleatórias podem ser classificadas como discretas ou contínuas. No caso

das variáveis discretas, elas podem assumir um número finito de valores, cujas probabilidades de ocorrência são conhecidas. A variável contínua pode assumir qualquer valor em um intervalo e as probabilidades necessárias são calculadas como a área abaixo da curva da distribuição, chamada de função densidade de probabilidade.

2.4.2 População e Amostra

Sempre que se trabalha com técnicas estatísticas deve-se ter clara a definição dos termos "população"e "amostra". A população diz respeito a um conjunto de todos os elementos que apresentam uma ou mais características em comum. Quando se extrai um conjunto de observações da população, ou seja, toma-se parte desta para a realização do estudo do comportamento de uma variável aleatória, tem-se a amostra. De posse de uma amostra representativa, ou seja, onde cada elemento da população tem a mesma probabilidade de ser escolhido, é possível se obter inferências válidas em torno dos parâmetros populacionais.

2.4.3 Estimação de Parâmetros Populacionais

A estimação dos parâmetros populacionais pode ser pontual ou intervalar. Na estimação pontual é fornecido um único valor para o parâmetro populacional em estudo. Na estimação intervalar é fornecido um intervalo de valores possíveis de serem assumidos pelo parâmetro em questão, bem como a probabilidade dela estar correta. As estimativas pontuais quase sempre diferem dos verdadeiros parâmetros populacionais. Desse fato decorre a necessidade de construção de intervalos ao redor das estimativas pontuais, que são determinados com base nas probabilidades dos intervalos, em questão, conterem os parâmetros populacionais procurados. Esses intervalos são conhecidos como intervalos de confiança. O centro do intervalo é a estimativa pontual. A principal característica da estimativa intervalar é que sua probabilidade de acerto é diretamente proporcional ao tamanho do intervalo onde se espera encontrar o parâmetro. Por outro lado, a significância da estimativa é inversamente proporcional a esse mesmo intervalo. O aumento do intervalo tem a vantagem de aumentar a chance de acerto da estimativa e a desvantagem de torná-la menos significativa.

Capítulo 3

Escopo Geográfico de Páginas *Web*

Nesta dissertação a definição de escopo geográfico de páginas *Web* é derivada da definição de escopo geográfico de *Web sites* apresentada por Ding et al. [6], por esse motivo algumas de suas idéias são empregadas aqui, começando pela organização das localizações geográficas em uma estrutura hierárquica.

Nos experimentos realizados para este trabalho foi utilizada uma coleção brasileira de páginas *Web* e, devido a isso, as localizações geográficas foram organizadas em uma estrutura de 4 níveis: país, região, estado e cidade, correspondendo a divisão geográfica existente no Brasil. A Figura 3.1 ilustra parte da hierarquia considerada. Assim, no nível de país tem-se o Brasil, de quem os descendentes, no nível de região são as diferentes regiões geográficas brasileiras: Norte, Nordeste, Sudeste, Sul e Centro-Oeste. Cada região é dividida em diversos estados. Por exemplo, a região Norte contém os estados Amazonas, Pará, Amapá, e assim por diante. Finalmente, cada estado é dividido em cidades, as quais constituem o nível das folhas na hierarquia.

Definida a estrutura herárquica são admitidas as suposições abaixo, devidamente adaptadas para o contexto de páginas *Web*:

- Considere que o escopo geográfico de uma página *Web* p seja o Brasil inteiro, provavelmente ela atrai o interesse das várias localidade do país. Esse interesse pode se mostrar através de páginas de todas as localidades que contenham *links* para esta página p. Mais especificamente, duas condições têm que ser satisfeitas para que uma dada localização faça parte do escopo geográfico de uma página p:
  - Uma fração significativa das páginas *Web* da localização ℓ contem *links* para p.

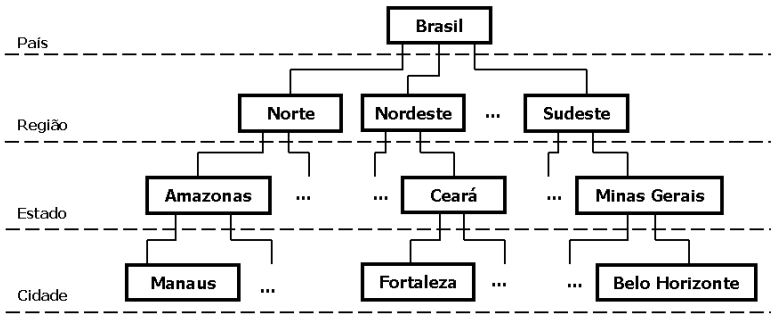

Figura 3.1: Hierarquia geográfica considerada para a coleção de páginas da *Web* Brasileira.

  - As páginas *Web* em ℓ que contem *links* para p estão uniformente distribuídas através de ℓ.

- Agora considere uma página p cujo escopo geográfico seja o estado do Amazonas, provavelmente o texto de p mencione mais frequentemente as cidades do estado do Amazonas do que localizações correspondentes a outros estados ou países. Mais especificamente, duas condições têm que ser satisfeitas para que uma dada localização faça parte do escopo geográfico de uma página p:

  - Uma fração significativa de todas as localizações mencionadas nas páginas *Web* são a própria ℓ ou sublocalizações de ℓ.
  - As referências de localizações no texto das páginas *Web* p estão uniformemente distribuídas através de ℓ.

Para medir quão relacionada uma página *Web* p está a uma dada localização ℓ, duas medidas de associação são definidas: *Power* e *Spread*. A função $Power(p, \ell)$ mede o grau de associação entre a página p e a localização ℓ. Quanto maior o valor de $Power(p, \ell)$ maior a certeza de que a página p pertence à localização ℓ. A função $Spread(p, \ell)$ mede quão uniforme é esta associação entre as sublocalizações de ℓ. Se, por exemplo, uma página p pertence ao estado do Amazonas, espera-se que o grau de associação entre cada cidade no Amazonas e p seja similar. Caso contrário, se uma cidade, digamos Manaus, tiver uma associação muito mais forte do que

as cidades restantes, a página não pertenceria ao estado, mas somente à cidade de Manaus. Tais medidas foram definidas para explorar o conteúdo textual e para explorar a estrutura de *links* dos *Web sites*.

A medida *Power* baseada na estrutura de *links* necessita de uma coleção inicial de páginas pré-classificadas, ou melhor, cujo escopo geográfico seja conhecido, a fim de que se possa propagar esta informação para as páginas vizinhas através da informação de *links*. De posse de tal coleção a medida de *Power* pode ser computada como segue:

$$Power(p, \ell) = \frac{Links(p, \ell)}{Pages(\ell)} \tag{3.1}$$

onde $Links\,(p, \ell)$ retorna o número de páginas na localização ℓ que contêm um *link* para a página p e $Pages(\ell)$ retorna o número de páginas pertencentes a localização ℓ. Assim, quanto mais páginas na localização ℓ apontarem para a página p, maior a possibilidade de que p esteja também na localização ℓ.

A medida *Spread* pode ser definida usando o conceito de entropia, da teoria da informação, a fim de medir a uniformidade da associação entre as sublocalidades $\ell_1, ..., \ell_n$ da localidade ℓ e a página *Web* p. Para tanto, supõe-se que há "uma fonte de informação"associada com uma página *Web* p e a localização ℓ. A fonte de informação gera símbolos que representam as diferentes sublocalidades de uma dada localidade que faz parte do possível escopo geográfico da página *Web* p. Seja:

$$P_i = \frac{Power(p, \ell_i) + 1}{\sum_{j=1}^{N} Power(p, \ell_j) + N} \tag{3.2}$$

onde cada ℓ_i é uma sublocalização de ℓ e N é o número total de sublocalizações de ℓ. Se cada ℓ_i for interpretado como sendo um símbolo gerado por ℓ e P_i tem probabilidade de ocorrência de ℓ_i[1], a entropia máxima ($\log N$) é alcançada quanto mais uniforme for o *Power* através de todas ℓ_i de ℓ. Assim, a medida *Spread* é definida como:

$$Spread(p, \ell) = \frac{-\sum_{i=1}^{N} P_i \times \log(P_i)}{\log N} \tag{3.3}$$

A medida de *Power* baseada no conteúdo textual de *Web sites* consiste em contar as referências geográficas que aparecem no texto das páginas *Web* pertencentes ao *site*. E pode ser

[1] Normalizada para evitar valores de probabilidade iguais a zero.

definida como segue:

$$Power(p, \ell) = \frac{\text{Referências}(p, \ell)}{\text{Localizações}(p)} \tag{3.4}$$

onde Referências(p, ℓ) é o número de referências à uma localização ℓ que podem ser encontradas no texto da página p e Localizações(p) é o número total de referências à localizações na página p. A medida *Spread* pode ser computada usando a Equação (3.3), simplesmente aplicando a Equação (3.4) na Equação (3.2).

A medida *Power* baseada no conteúdo textual, definidida acima, exige adaptação para o contexto de páginas *Web*, já que estas possuem muito menos informação textual do que um *site* inteiro, no capítulo 4 é feita apresetação desta adaptação.

Em ambas as abordagens, conteúdo textual e informação de *links* das páginas, a função $Spread(p, \ell)$ alcançará o seu valor máximo (igual à 1) nos dois seguintes casos especiais:

- ℓ é um nó folha na hierarquia de localizações: neste caso, por definição, a distribuição de *Power* através de ℓ é completamente uniforme, já que ℓ está no último nível da hierarquia que corresponde às cidades do Brasil.

- $Power(p, \ell) = 0$: neste caso, não há associação entre a localização ℓ e a página *Web* p. Como o *Spread* mede a uniformidade desta associação através de ℓ, $Spread(p, \ell)$ é trivialmente máxima neste caso.

Definidas as medidas de *Power* e *Spread*, o processo de classificação consiste em encontrar o conjunto de localizações $\mathcal{L}_p$ tal que:

$$\ell \in \mathcal{L}_p \iff \begin{cases} Spread(p, \ell) \geq t \\ Spread(p, \ell') < t, \text{ para todo ancestral } \ell' \text{ de } \ell \end{cases} \tag{3.5}$$

onde t é um limiar fixado. O conjunto $\mathcal{L}_p$ pode ser ajustado a fim de manter somente localizações com valores de *Power* altos. Isto é alcançado através da consideração de um limiar r definindo como segue:

$$\mathcal{L'}_p = \{\ell \in \mathcal{L}_p | Power(p, \ell) \geq \max_{\ell' \in \mathcal{L}_p}(Power(p, \ell')) \times r\} \tag{3.6}$$

onde $0 \leq r \leq 1$. Este conjunto final de localizações é o escopo geográfico da página p.

Explorar somente a informação de *links* das páginas *Web* possui duas limitações principais. A primeira, diz respeito a como se obter a coleção inicial de páginas e a segunda, refere-se ao

fato de que a maioria das páginas *Web* contêm pouca informação de *links*. Quanto a exploração somente do conteúdo textual a principal limitação refere-se à necessidade de bases de dados externas detalhadas sobre: pontos de referências geográficas e normalização dos nomes das localidades (como apelidos e ambiguidades). Além disso, confiar somente em uma única fonte de informação, pode impedir a aplicação eficiente das medidas, definidas acima, já que algumas páginas *Web* podem não permitir captura do seu conteúdo e outras podem ter um número muito pequeno de *links*. A fim de superar tais limitações este trabalho propõe:

- Adaptação da medida *Power* baseada texto, para que ela use a informação textual não somente da própria página *Web* a ser classificada, mas também de outras páginas já classificadas, dispensando a necessidade de bases de dados externas.

- Uma forma de combinar a informação extraída da estrutura de *links* com a informação obtida a partir do texto da página, a fim de fazer uso de toda informação possível, disponível na página *Web*.

- O uso de uma heurística baseada na localização de CEPs no texto das páginas para obtenção de uma coleção inicial de páginas pré-classificadas de forma automática, usando apenas uma base de CEPs.

Capítulo 4

Classificação de Páginas *Web*

O interesse de estimar o escopo geográfico de páginas *Web* individuais surgiu principalmente por duas razões. A primeira diz respeito ao fato de ser vago o conceito de *Web site*, o que pode tornar difícil a determinação automática de quais páginas fazem parte de um mesmo *site*. A segunda está fundamentada na hipótese de que dentro de um mesmo *site* podem haver páginas que possuem escopos geográficos distintos. No entanto, essa mudança de foco, de *sites* para páginas *Web*, traz consigo dificuldades adicionais. A maioria das páginas *Web* tem conteúdo textual muito pequeno o que implica em possuírem pouca ou nenhuma referência textual a localizações geográficas e, ainda, possuem pouquíssimos *links* ou simplesmente nenhum.

Neste capítulo são detalhados os melhoramentos proposto, sugeridos para tentar superar tais dificuldades. Para tanto serão estudadas as modificações da medida *Power* baseada em texto e uma forma de combinação da informação proveniente da estrutura de *links* com a informação obtida a partir do texto das páginas *Web*.

4.1 Similaridade Baseada no Texto das Páginas *Web*

Para computar a medida $Power(p, \ell)$, proposta por Ding et al. [6], é necessário contar quantas referências geográficas aparecem no texto da página *Web* p. Para esta tarefa, bases de dados detalhadas de pontos de referência geográfica são requeridas. Ter acesso a tais bases de dados pode ser difícil, além de todos os problemas relacionados a normalização dos nomes de localidades (apelidos e ambiguidades). Referências comuns, como nomes de cidade ou CEPs, podem até estar disponíveis, no entanto, muitas páginas conterão somente pontos de referência locais,

geralmente desconhecidos fora de sua vizinhança e pouco prováveis de estarem presentes em qualquer base de dados. Por outro lado, é possível assumir que outras páginas *Web* da mesma localidade ℓ mencionarão os mesmos pontos de referências. Baseado nesta suposição, foi feita a proposta de uma medida *Power* baseada no conteúdo textual das páginas *Web*, adaptada para o contexto de páginas *Web* individuais. Esta medida usa a informação textual não somente da própria página, mas também de outras páginas já classificadas, sendo definida como segue:

Se um dado termo t aparece em todas as páginas de uma localização ℓ, este termo é provavelmente um ponto de referência geográfica em ℓ. Genericamente, quanto maior for o número de páginas de ℓ que o termo t aparecer, maior a possibilidade de que t refira-se a um ponto de referência geografica em ℓ. Se o termo t ocorre em muitas páginas na coleção, entretanto, é provável que seja genérico demais para estar associado a uma dada localização. Sendo assim foi definido o peso de um termo t como:

$$w(t) = \text{idf}_{\text{coleção}}(t) \times \text{idf}_{\text{localidade}}(t) \times \text{Concentração}(t, \ell) \times \text{idf}_{\text{coleção}}(\ell) \tag{4.1}$$

Onde:

- o idf do termo t na coleção, é definido como segue

$$\text{idf}_{\text{coleção}}(t) = \frac{\log\left(\frac{N_{\text{páginas}}}{NP(t)}\right)}{\log N_{\text{páginas}}},$$

representa a importância do termo t na coleção, onde $N_{páginas}$ é número de páginas da coleção e $NP(t)$ é o número de páginas onde o termo t ocorre.

- o idf do termo t na localidade ℓ, é definido como segue

$$\text{idf}_{\text{localidade}}(t) = \frac{\log\left(\frac{N_{\text{localidade}}}{NL(t)}\right)}{\log N_{\text{localidade}}},$$

representa a importância do termo t na localidade ℓ, onde $N_{\text{localidade}}$ é número de localidades da hierarquia adotada e $NL(t)$ é o número de localidades onde o termo t ocorre.

- a concentração do termo t na localidade ℓ, é definido como segue

$$\text{Concentração}(t, \ell) = \frac{NP(t, \ell)}{NP(t)}$$

representa a concentração do termo t na localidade ℓ, onde $NP(t, \ell)$ é número de páginas onde o termo t ocorre na localidade ℓ e $NP(t)$ é o número de páginas onde o termo t ocorre.

- o idf da localidade ℓ na coleção, é definido como segue

$$\text{idf}_{\text{coleção}}(\ell) = \frac{\log\left(\frac{N_{\text{páginas}}}{N_{localidade}}\right)}{\log N_{\text{páginas}}},$$

representa a importância da localidade ℓ na coleção, onde $N_{\text{páginas}}$ é número de páginas da coleção e $N_{\text{localidade}}$ é o número de páginas da localidade ℓ.

Usando a Equação (4.1), é possivel agora definir:

$$Power(p, \ell) = \frac{\sum_{i=1}^{T} f(t_i, p) \times w(t_i, \ell)}{\sum_{i=1}^{T} f(t_i, p)} \tag{4.2}$$

onde cada t_i é um termo da página p, e $f(t_i, p)$ a *frequência do termo* t_i, isto é, o número de vezes que o termo t_i aparece na página p. Assim, o valor de associação entre as páginas p e a localização ℓ é a soma da importância dos termos em p, normalizados pelo seu tamanho. Como antes, os valores de *Spread* podem ser computados usando a Equação (3.3), pela aplicação da Equação (4.2) na Equação (3.2).

Uma vez definidas as medidas de *Power* e *Spread*, o processo de classificação adotado aqui é o mesmo descrito no Capítulo 3.

4.2 Combinando Fontes de Informação

Como a maioria das páginas *Web* contém pouquíssima informação, fazer uso de toda informação disponível na página *Web* é imprescindível. Para conseguir isto, será proposta, a seguir, a combinação, em um único valor, da medida *Power* baseada na estrutura de *links* com a medida *Power* baseada no conteúdo textual em um único valor. Para que se possa realizar tal combinação é necessário uma função que respeite as seguintes exigências:

1. Se uma página tiver somente uma fonte da informação disponível, esta fonte deve ser usada. Assim, se uma página não possuir nenhum *link* mas tiver no seu texto referências a uma dada localização geográfica, este texto deve ser levado em consideração.

2. Se a página possuir as duas fontes de informação, texto e *links*, para considerar a sua

associação a uma dada localização, apenas uma deve ser levada em consideração. A suposição é que se a página possui ambos, textos e *links*, e estes façam referência a uma localização ℓ, é mais provável que tal página seja de ℓ do que uma página que tenha somente o texto ou somente *links* referentes a ℓ;

3. A função de combinação deve afetar de maneira mínima a medida de *Spread*. Se os valores de *Power* derivados dos *links* ou do texto estiverem distribuídos uniformemente através de todos as sublocalizações ℓ_i de ℓ, os valores combinados devem também ser uniformemente distribuídos através de todas as ℓ_i.

4. Tanto a informação textual quanto a informação de *links* têm a mesma importância. Isto é assumido, uma vez que não há nenhum conhecimento, a priori, sobre se *links* são mais confiáveis do que o texto, ou vice-versa, ao determinar o espaço geográfico de uma página *Web*;

A fim de satisfazer tais condições, assumindo que $Power_L(p, \ell)$ e $Power_T(p, \ell)$ sejam as medidas de *Power* baseadas em *links* e conteúdo textual, definidas na Equação (3.1) e Equação (4.2), respectivamente. A medida de *Power* combinada $Power_C$ foi definida como:

$$Power_C(p, \ell) = \max\left(\frac{Power_L(p, \ell)}{\sum_{i=1}^{N} Power_L(p, \ell_i)}, \frac{Power_T(p, \ell)}{\sum_{i=1}^{N} Power_T(p, \ell_i)}\right) \tag{4.3}$$

onde cada ℓ_i é uma sublocalização da mesma localização pai de ℓ e $\max(x, y)$ é a função que retorna o valor máximo entre x e y.

Usando o valor máximo de *Power* para uma página p, garantido pela Equação (4.3), se uma fonte de informação estiver disponível na página p, a fonte que fornece mais certeza será sempre utilizada. Além disso, o valor de *spread* não deve ser afetado negativamente, uma vez que a função da combinação sempre escolherá um dos valores de *Power* computados para uma única fonte da informação e o normalizará de modo que os valores de *links* e de texto estejam na mesma escala.

Capítulo 5

Experimentos

Neste capítulo são apresentados os resultados dos experimentos para avaliação da eficiência do método de determinação do escopo geográfico de páginas *Web* proposto no presente trabalho. Primeiramente, são apresentadas as coleções de páginas *Web* utilizadas nos experimentos e a metodologia adotada na obtenção das mesmas. Depois são apresentados os resultados obtidos.

5.1 Coleções

Os experimentos foram realizados com páginas da *Web* brasileira. Tais páginas, aproximandamente 12 milhões com cerca de três milhões de termos, foram extraídas e indexadas pela máquina de busca *TodoBR*[1] em 2002. A avaliação do método proposto foi realizada através do uso de uma coleção de treino e uma coleção de teste.

5.1.1 Coleção de Treino

A coleção de treino foi obtida automaticamente utilizando-se uma heurística baseada na ocorrência de CEPs no conteúdo textual das páginas *Web*, com a qual foi possível classificar um conjunto de pouco mais de 150 mil páginas. A heurística proposta utiliza as seguintes regras para associar uma página p a uma localização ℓ:

- Se todos os CEPs que apareceram em p são de uma mesma cidade, p é associada a esta cidade;

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

- Se os CEPs que apareceram em p pertencem a diferentes cidades, então p é associada ao maior nó na hierarquia que inclui todas as cidades;

- Páginas que não possuem ocorrências de CEPs não são incluídas na coleção de treino.

A classificação foi realizada em quatro níveis (cidade, estado, região e país), de tal forma que uma página associada a um nível não estaria associada aos outros três níveis. Assim, se uma página p contém CEPs de duas cidades do estado do Amazonas, p será associada ao estado do Amazonas. Se p contém CEPs de cidades de diferentes estados da região Norte, p é associada à região Norte. Finalmente, se p contém CEPs de cidades de diferentes regiões, ela é associada ao nível de país. A lista de cidades e seus CEPs correpondentes foi obtida do Correio Brasileiro[2].

Normalmente a coleção de treino utilizada em um processo de classificação é pequena e composta de documentos classificados manualmente, contudo o método que é proposto aqui necessita de uma coleção grande e confiável, uma vez que sua finalidade é propagar a classificação inicial através da combinação da informação derivada da estrutura de *links* com a informação do texto das páginas *Web*. Através da heurística baseada na ocorrência de CEPs, foi possível classificar uma coleção relativamente grande de páginas, no entanto a tarefa de verificação manual da corretude das associações geradas se tornou inviável. Para contornar esta dificuldade foi adotado o procedimento estatístico de modelar o problema segundo uma distribuição de probabilidade e fazer inferências em torno dos seus parâmetros, assumindo um certo grau de confiança para as estimativas a serem obtidas.

O interesse aqui é obter inferências em torno da proporção de páginas que foram corretamente associadas a um dos níveis na hierarquia adotada, sendo que uma vez associada a um dos níveis a página não estaria associada a nenhum outro. Neste caso, a variável de interesse é binária (do tipo "sucesso"ou "fracasso", ou seja, "associação correta"ou "associação incorreta"), cujas observações são independentes e a probalidade de sucesso é a mesma para cada uma delas. Dado este cenário, o modelo adotado assume uma variável aleatória Z tendo distribuição de Bernoulli com parâmetro Π, tal que:

$$Z = \begin{cases} 1 & \text{se } \ell = \ell' \\ 0 & \text{se } \ell \neq \ell' \end{cases} \tag{5.1}$$

[2]See `http://www.correios.com.br/`.

onde ℓ é a localização automaticamente associada à página e ℓ' é a localização correta da página. Para estimar a proporção populacional de páginas que foram corretamente associadas $\Pi = P(Z = 1)$ é necessário tomar uma amostra de n páginas e computar:

$$\hat{\Pi} = \frac{1}{n} \sum_{i=1}^{n} Z_i \tag{5.2}$$

onde cada Z_i é o resultado da variável Z para uma página na amostra. A estimativa intervalar da proporção populacional de páginas corretamente associadas foi dada pelo intervalo de Wald:

$$\hat{\Pi} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\Pi}(1 - \hat{\Pi})}{n}} \tag{5.3}$$

Os intervalos foram estimados com um coeficiente de confiança aproximado $1 - \alpha = 0.95$, onde $z_{\frac{\alpha}{2}} = 1.96$. Para computar $\hat{\Pi}$, uma amostra[3] de páginas de cada nível da hierarquia adotada foi coletada aleatoriamente. A Tabela 5.1 mostra o tamanho das amostras, tamanho da população, número de classificações corretas e estimativa intervalar para Π, para cada nível da hierarquia.

Dados os valores na Tabela 5.1, pode-se dizer que, com aproximadamente 95% de confiança, a proporção de páginas corretamente classificadas está no intervalo $[0.902, 0.951]$ para o nível de cidades, $[0.968, 0.991]$ para o nível de estado, $[0.919, 0.976]$ no nível de região, e $[0.914, 0.961]$ no nível de país. Uma explicação da avaliação estatística adotada aqui pode ser encontrada em [1].

Nível	População	Tamanho da Amostra	Margem de Erro das Amostras	Páginas corretamente classificadas	Estimativa Intervalar de Π
Cidade	132 757	435	0.0245	403	$[0.902, 0.951]$
Estado	8 164	596	0.0113	584	$[0.968, 0.991]$
Região	4 604	229	0.0288	217	$[0.919, 0.976]$
País	11 342	402	0.0236	377	$[0.914, 0.961]$

Tabela 5.1: Estatísticas para o cálculo da proporção amostral de páginas corretamente classificadas usando a heurística de CEPs .

Dado que as estimativas intervalares da proporção de páginas corretamente associadas para cada um dos níves da hierarquia são boas, principalmente se considerado o fato de que trata-se de um processo simples e sem interferência humana, assumiu-se o conjunto de páginas obtidas com a heurística baseada em CEPs como coleção de treino, já que quanto maior essa coleção for,

[3] O cálculo do tamanho da amostra [12] para cada nível da hierarquia foi feito assumindo que a população é infinita com probabilidade de sucesso igual a de fracasso, e margem de erro fixadas em 0.50 e 0.05, respectivamente.

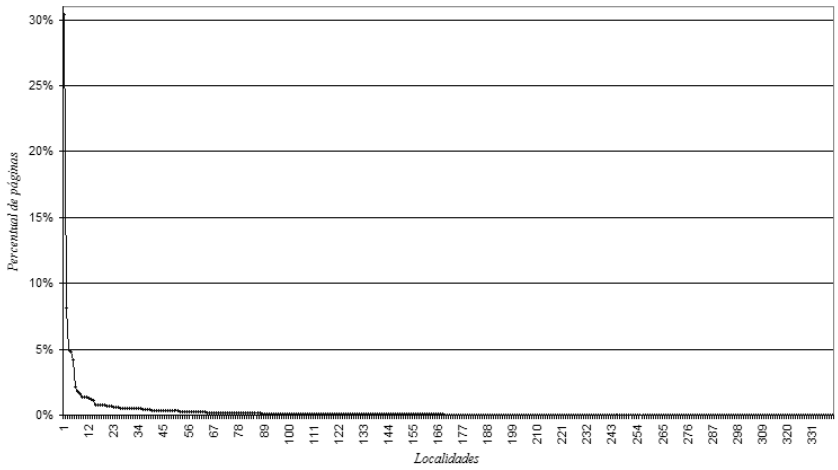

Figura 5.1: Distribuição do percentual de páginas *Web* pré-classifcadas por localidades.

maior a chance de ampliarmos essa classificação inicial. No entanto, como pode ser observado na Figura 5.1, apesar de se ter alcaçando um número grande de páginas, não houve uniformidade na distribuição de páginas por localidade, ou seja, houve uma concentração muito grande de páginas de exemplo em algumas localidades e na maioria das localidades uma quantidade muito pequena. Por exemplo, das 340 localidades da hierarquia adotada só a cidade de São Paulo possui mais de 30% das páginas pré-classificadas. O impacto disto na classificação será discutido na Seção 5.3.

5.1.2 Coleção de Teste

A coleção de teste foi obtida a partir do diretório *Cadê?*, de onde extraiu-se um conjunto de 330 URLs de páginas *Web* geograficamente classificadas de forma manual por especialistas e tomando-se o cuidado de não selecionar páginas previamente incluídas na coleção de treinamento. Através de tais URLs obteve-se o conteúdo textual e a informação de *links* destas páginas na coleção do *TodoBR*.

5.2 Medidas de Avaliação

A qualidade dos resultados obtidos nos experimentos foi avaliada utilisando-se as medidas *Precisão*, *Revocação* e *F1*, como descrita em [6].

Seja R o conjunto de localizações geográficas que fazem parte do verdadeiro escopo geográfico de uma página p, isto é, conjunto correto de localizações que o método proposto deve associar à página p. Seja A o conjunto de localizações geográficas associadas à página p pelo método proposto nesta dissertação. Para computar a interseção entre ambos os conjuntos, primeiro deve ser feita uma expansão destes conjuntos para que sejam incluídas todas as suas sublocalizações. Sendo assim, segue a definição:

$$R' = \{\ell | \ell \in R \text{ ou } \ell \text{ é uma sub-região de qualquer } \ell' \in R\}$$
$$A' = \{\ell | \ell \in A \text{ ou } \ell \text{ é uma sub-região de qualquer } \ell' \in A\}$$

A *Precisão* e a *Revocação* pode agora ser definida como:

$$\text{Precisão} = \frac{|A' \cap R'|}{A'}$$
$$\text{Revocação} = \frac{|A' \cap R'|}{R'}$$

A medida *F1* que combina *Precisão* e *Revocação* com igual peso é definida como:

$$\text{F1} = \frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

5.3 Resultados

Nesta seção são apresentados os resultados dos experimentos realizados a fim de avaliar o método proposto para determinar o escopo geográfico de páginas *Web* através do uso do conteúdo textual e *links* destas como fontes de informação. As informações sobre *links* referen-se às páginas que apontam para uma página p (*in-links* de p) e também para quais páginas p aponta(*out-links* de p).

Em todos os experimentos descritos nesta seção foi fixado o limiar de poda para a medida *Spread* ($t = 0.9$).

5.3.1 Avaliação do uso das fontes de informação em separado e combinadas

A Figura 5.2 e a Tabela 5.2 mostram o impacto do limiar de poda r sobre as medidas de qualidade usadas para avaliar a eficiência do método proposto. Quando considerada somente a informação textual das páginas *Web* os resultados foram melhores do que quando considerada somente a informação de *links*. Embora a combinação das fontes de informação, conteúdo textual e *links* das páginas, tenha apresentado uma melhora nos resultados, estes ainda são insuficientes, já que a precisão máxima obtida é de aproximadamente 34%. Isso se deve, provavelmente, às características da coleção de treino, baixa conectividade entre as páginas desta coleção com as páginas do restante da coleção e poucas páginas de exemplo na maioria das localidades, fazendo com que as medidas da similaridade introduzam muito ruído no processo da classificação.

Dizer que a coleção de treino possui baixa conectividade siginifica dizer que há uma quantidade insuficiente de *links*, tanto chegando quanto partindo das páginas *Web* pré-classificadas. Isso porque, frequentemente, a informação de CEP, utilizada para criar a coleção de treino, não está presente nas páginas principais de *Web sites*, mas, normalmente, em uma página isolada do *site*, sendo, muitas vezes, apontadas somente por páginas do próprio *site* do qual fazem parte. Esta situação compromete a propagação da classificação inicial da coleção de treino para

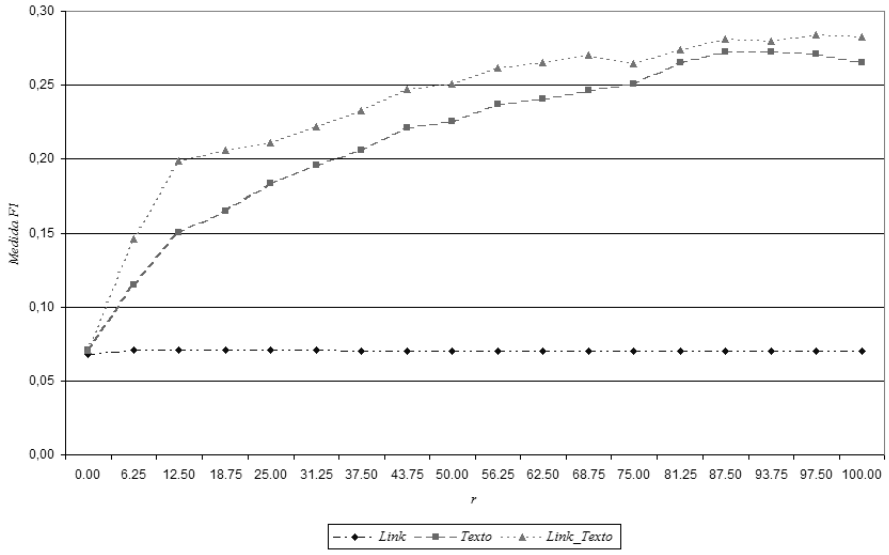

Figura 5.2: Comparação das fontes de informação em separado e combinadas através da medida *F1* média em função do limiar r para a medida *Power*.

Limiar r	Precisão			Revocação		
	Link	Texto	Link-Texto	Link	Texto	Link-Texto
0.00	0.035150	0.036751	0.036737	0.990598	0.981496	0.981239
6.25	0.037106	0.064348	0.085139	0.809785	0.533235	0.514411
12.50	0.037038	0.089779	0.128201	0.809439	0.455966	0.444307
18.75	0.036962	0.105492	0.142746	0.809073	0.373454	0.367757
25.00	0.036894	0.126380	0.154915	0.808885	0.337085	0.332508
31.25	0.036850	0.140317	0.171232	0.808830	0.324351	0.314111
37.50	0.036711	0.154489	0.190291	0.807994	0.308334	0.300647
43.75	0.036698	0.174040	0.212162	0.807934	0.304711	0.297216
50.00	0.036698	0.182356	0.221814	0.807934	0.296002	0.289004
56.25	0.036588	0.201389	0.245072	0.797829	0.287682	0.280699
62.50	0.036595	0.210852	0.256791	0.797829	0.280158	0.273910
68.75	0.036599	0.225074	0.273503	0.797777	0.273245	0.267388
75.00	0.036599	0.239155	0.282097	0.797777	0.264210	0.248648
81.25	0.036599	0.267288	0.305562	0.797777	0.263944	0.248412
87.50	0.036599	0.282209	0.323197	0.797777	0.263758	0.248352
93.75	0.036599	0.285404	0.326677	0.797777	0.260111	0.244802
97.50	0.036599	0.295151	**0.339260**	0.797777	0.250362	0.244802
100.00	0.036599	0.289477	0.339465	0.797777	0.244638	0.242046

Tabela 5.2: Resumo dos resultados obtidos para as fontes de informação em separado e combinadas em função do limiar r para a medida *Power*.

outras páginas, porque quando uma página é submetida ao método de determinação do escopo geográfico, ela pode até ter muitos *links*, mas se estes não forem de páginas pré-classificadas, ou seja, de páginas associadas a algumas localidade geográfica, não será possível tirar conclusão sobre o seu escopo geográfico utilizando os *links*. A Figura 5.3 mostra um exemplo desta situação. A situação acima pode piorar se, além da ausência de *links*, a página cujo escopo geográfico está sendo estimado possuir pouca ou nenhuma referência geográfica no seu texto.

Se cada localidade da hierarquia tivesse tantos exemplos quantos fossem necessários para o classificador entender quais as características de cada localidade, a determinação do escopo seria facilitada. Porém, a pequena quantidade de páginas pré-classificadas por localidade é fato na coleção de treino adotada. Isso é ruim porque a maioria das localidades possui pouca informação para caracterizá-las. Considere, por exemplo, a seguinte situação: uma determinada localidade ℓ possui apenas 10 páginas de exemplo, em cada uma delas existem referências geográficas (como nomes de rua, cidade, códigos de área, etc) que não se repetem entre elas. Tais informações são muito importantes no processo de determinação do escopo geográfico, no entanto, mesmo tendo escopo geográficos iguais, as 10 páginas podem tratar de assuntos diferentes (como pizza, serviço imobiliário, animais, remédio, livros, CDs, etc.) fazendo com que as referências geográficas sejam tratadas pelo método como quaisquer outros termos das páginas, ou seja, muita informação que não é relevante geograficamente acaba "poluindo" o treino. A Figura 5.4 mostra um exemplo desta situação.

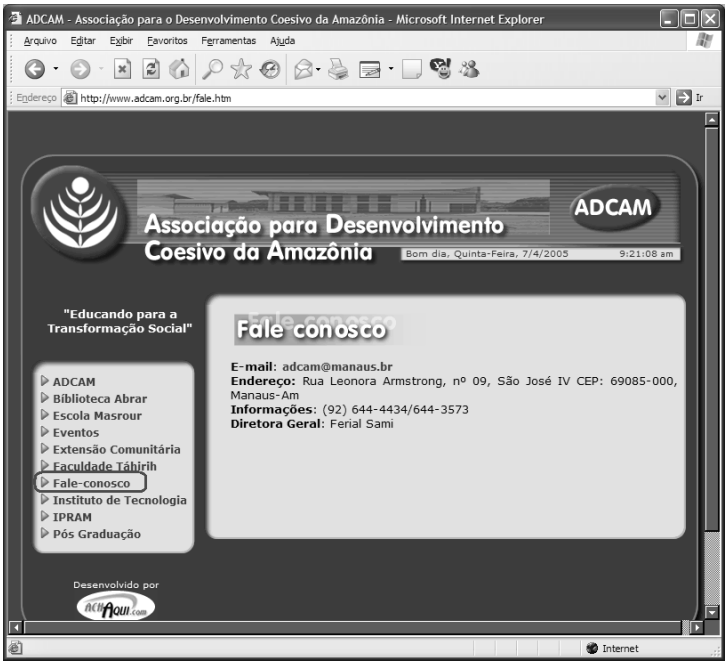

Figura 5.3: Exemplo 1 de onde o informação de CEP está disponível.

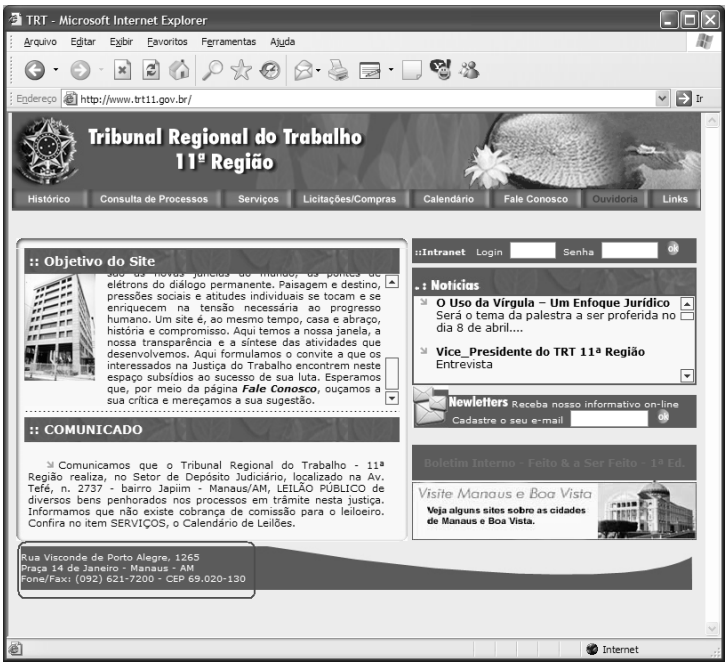

Figura 5.4: Exemplo 2 de onde o informação de CEP está disponível.

5.3.2 Avaliação do uso do conteúdo textual completo das páginas *Web* da coleção de treino

Normalmente as páginas possuem informação textual que não tem nada haver com informação geográfica. Para comprovar o impacto disso no método proposto, foi realizado um experimento sobre a própria coleção de treino utilizando-se um procedimento de validação cruzada, o qual consistiu em dividir, aleatoriamente, tal coleção em dez partes . Procedendo depois 10 repetições do método, onde a cada rodada uma parte diferente da coleção é usada como conjunto de teste e as nove outras restantes são usadas como conjunto de treino. O resultado final do experimento representa a média das dez rodadas para as medidas de qualidade usadas na avaliação do método proposto.

Limiar r	*Precisão*			*Revocação*			*F1*		
	Link	Texto	Link-Texto	Link	Texto	Link-Texto	Link	Texto	Link-Texto
0.00	0.012956	0.013270	0.013270	0.999386	0.990887	0.990887	0.025580	0.026189	0.026189
6.25	0.072622	0.186775	0.251596	0.977837	0.822007	0.822172	0.135203	0.304388	0.385288
18.75	0.072658	0.375558	0.415465	0.977655	0.765730	0.766474	0.135263	0.503950	0.538849
50.00	0.072583	0.589658	0.605696	0.977282	0.706727	0.711074	0.135130	0.642907	0.654168
81.25	0.072581	0.683541	0.689403	0.976903	0.672021	0.678606	0.135123	0.677732	0.683962
87.50	0.072581	0.694175	0.699186	0.976903	0.667127	0.674638	0.135123	0.680382	0.686693
97.50	0.072581	0.708262	0.712750	0.976903	0.657270	0.666421	0.135123	0.681814	0.688807
100.00	0.072581	0.711315	**0.717464**	0.976903	0.655583	0.664980	0.135123	0.682313	**0.690226**

Tabela 5.3: Resumo dos resultados obtidos dentro da própria coleção de treino para as fontes de informação em separado e combinadas em função do limiar r para a medida *Power*.

Como se pode observar na Tabela 5.3, que mostra o impacto do limiar de poda r sobre as medidas de qualidade *Precisão*, *Revocação* e *F1*, a classificação obtida neste experimento também não foi boa. A precisão máxima obtida quando se usa apenas o conteúdo textual como fonte de informação foi de aproximandamente 72%. Isso pode ser uma evidência de que, além de haver poucas páginas de exemplo em cada localidade da hierarquia adotada, o texto das páginas pode conter muita informação que não é útil neste processo de classificação geográfica.

Através dos resultados mostrado na Tabela 5.3, a impressão que se tem é que se for levado em consideração o texto inteiro das páginas da coleção de treino pode-se obter um resultado mais confuso do que apenas os CEPs fossem observados, ou seja, o resultado parece ser melhor quando se descarta o resto do texto e observa-se somente a informação de CEP. Isso faz com que se acredite que poderia ser útil, levando a uma melhora nos resultados, uma "limpeza" do texto que representa o conteúdo das páginas da coleção de treino. Sendo necessária a retirada dos termos que possam vir a poluir a caracterização das localidades confundindo a determinação

do escopo geográfico das páginas que são submetidas ao método proposto.

Assim como no experimento anterior, quando somente a informação de *links* é considerada os resultados são claramente insatisfatórios, trazendo valores muito baixos de *F1*. Evidenciando uma baixa conectividade entre as próprias páginas *Web* da coleção de treino.

Capítulo 6

Conclusão e Trabalhos Futuros

Este trabalho realizou estudos sobre um procedimento automático para determinação do escopo geográfico de páginas *Web*. Diferentemente de outros métodos existentes, mais direcionados para classificação de *Web sites*, o método aqui apresentado tenta valer-se de diferentes fontes de informação para possibilitar a classificação de páginas *Web*.

Uma vez que as páginas *Web* contêm, normalmente, pouca informação, o que se propôs foi usar o texto de páginas *Web* previamente classificadas como fonte de informação, combinado com a informação baseada na estrutura de *links* presentes nestas páginas. O objetivo foi tentar tomar vantagem destas duas fontes de informação para garantir com mais seguraça quanto uma página está relacionada a uma determinada localidade. Apesar dos resultados experimentais indicarem que esta combinação pode ser útil, eles não foram satisfatórios. Isto se deveu, provavelmente, à baixa conectividade e à pequena quantidade de páginas de exemplos na maioria das localidades da coleção de treino, o que torna necessárias algumas melhorias no método ora proposto, as quais não puderam ser realizadas ainda neste trabalho.

6.1 Trabalhos Futuros

Nesta seção são apresentadas algumas sugestões de trabalhos futuros, que poderão dar continuidade aos experimentos realizados para esta dissertação. As propostas surgiram de necessidades observadas durante a fase de avaliação e análise dos seus resultados.

A heurística de CEPs, usada para criar a coleção de treino, se mostrou útil e poderia ser melhorada com a inclusão de termos (como nomes de cidade, pontos notáveis, nomes de rua,

códigos de área, etc) que descrevam melhor as localidades geográficas que fazem parte da hierarquia adotada. Para isso seria necessário uma base sobre referências geográficas em conjunto com a base de CEPs para montar a coleção de treino. Essa sugestão talvez leve a uma coleção de treino maior e, talvez, mais conectada, uma vez que tais termos podem ter mais chance de serem econtrados nas páginas principais dos *Web sites*. Além disso, o ideal seria usar como conteúdo textual das páginas da coleção de treino somente o texto em volta dos CEPs e das referências geográficas, diminuindo com isso a quantidade de termos que possam vir a "poluir" a descrição das localidades.

Para se tentar minimizar o problema da baixa conectividade da coleção de treino uma idéia pode ser agrupar as páginas desta coleção, para só então analisar a informação de *links* dos grupos de páginas ao invés de páginas individuais. Tais grupos devem atender algumas restrições a fim de garantir que as páginas dos grupos tenham o mesmo escopo geográfico. Como essa garantia é difícil, faz-se necessário o uso de heurísticas para tentar agrupar páginas que tenham o mesmo escopo geográfico, ainda que não se saiba qual o escopo geográfico destas, observando-se, por exemplo, a similaridade entre páginas que façam parte do mesmo *site*, assumindo-se com isso que elas possuem o mesmo escopo geográfico. Sendo assim, as páginas de um grupo não poderiam pertencer a sites diferentes.

Referências Bibliográficas

[1] P. J. Bickel e K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, Volume 1. Prentice Hall, Segunda edição, 2000.

[2] S. Brin e L. Page. *The anatomy of a large-scale hypertextual web search engine*. Em Proceedings of the 7th International World wide Web Conference, páginas 107-117, Abril 1998.

[3] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano e N. Shivakumar. *Exploiting geographical location information of web pages*. Em Proceedings of the ACM SIGMOD Workshop on Web Databases (WebDB'99), páginas 914-96, Junho 1999.

[4] P. Calado, B. Ribeiro-Neto, N. Ziviani, E. Moura e I. Silva. *Local versus Global Links Information*. ACM Transactions on Information Systems, 21(1):42-63, Janeiro 2003.

[5] D. Cohn e T. Hofmann. *The missinglink - a probabilistic mode of document content and hypertext connectivity*. Em T. K. Leen, T. G. Dietterich, and V. Tresp, editores, Advances in Neural Information Processing System 13, páginas 430-436. MIT Press, 2001.

[6] J. Ding, L. Gravano e N. Shivakumar. *Computing geographical scopes of web resources*. Em Proceedings of the 26th VLDB Conference, páginas 545-556, Setembro 2000.

[7] L. Gravano, V. Hatzivassiloglou e R. Lichtenstein. *Categorizing web queries according to geographical locality*. Em Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, páginas 325-333, Novembro 2003.

[8] T. Joachims, N. Cristianini e J. Shane-Taylor. *Composite kernels for hypertext categorisation*. Em Proceedings of the 18th International Conference on machine Learning, ICML-01, páginas 250-257, Junho 2001.

[9] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld e R. Weibel. *Spatial information retrieval and geographical ontologies: An overview of the SPIRIT profect.* Em Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, páginas 387-388, Agosto 2002.

[10] J. M. Kleinberg. *Authoritative sources in a hyperlinked enviroment.* Em Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, páginas 668-677, Janeiro 1998.

[11] K. S. McCurley. *Geospatial mapping and navigation of the web.* Em Proceedings of the Tenth International World Wide Web Conference, WWW'10, páginas 221-229, Maio 2001.

[12] W. G. Cochran. *Sampling Techniques.* Wiley Series in Probability and Statistics, 1977.

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

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

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

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

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

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

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

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

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

[22] W. Weaver e C. E. Shannon. *The Mathematical Theory of Communication.* Urbana, Illinois: University of Illinois Press, 1949.

[23] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Statistics, Terceira edição.

[24] TodoBR: Todo o Brasil na Internet. `http://www.todobr.com.br`, último acesso em abril/2005.

[25] Diretório do Cadê. `http://www.cade.com.br`, último acesso em abril/2005.