Um Retrato da Web Brasileira*

Eveline A. Veloso [1] Edleno S. de Moura [2,1] Paulo B. Golgher [1]
Altigran S. da Silva [1,2] Rodrigo B. Almeida [1]
Alberto H. F. Laender [1] Berthier Ribeiro-Neto [1]
Nivio Ziviani [1]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte MG Brasil 31270-010
{eveline,edleno,golgher,alti,barra,
laender,berthier,nivio}@dcc.ufmg.br

[2] Departamento de Ciência da Computação
Universidade do Amazonas
Manaus AM Brasil 69077-000
{edleno,alti}@dcc.fua.br

Resumo

Este trabalho apresenta um retrato do conteúdo e da estrutura da Web brasileira obtido a partir de dados coletados em abril de 2000. O estudo foi realizado com o auxílio de agentes distribuídos implementados para coletar documentos da Web brasileira (páginas HTML e arquivos PostScript, PDF, Word e texto). Durante o processo de coleta, foram armazenados diversos dados relacionados com a estrutura e organização da Web brasileira . Algumas características importantes da Web brasileira são apresentadas, tais como o tamanho médio dos documentos, o número de servidores e o número médio de *links* por página, e, a partir dos dados coletados, estimamos o tamanho da Web brasileira.

Abstract

This paper presents a snapshot of the content and structure of the Brazilian Web based on data collected in April 2000. The study was carried out with the help of a page crawler implemented to collect documents of the Brazilian Web (HTML pages, PostScript, PDF, Word and text files). During the collecting process, data related to the organization and structure of the Brazilian Web was stored. We present some important characteristics of the Brazilian Web, such as the average file size, the number of hosts, the average number of links per page, and we estimate the size of the Brazilian Web based on the data collected.

*Este trabalho foi realizado com financiamento parcial do PRONEX, processo 76.97.1016.00.

1 Introdução

O surgimento da *World Wide Web* (ou simplesmente Web) tem causado uma revolução não só na área de ciência da computação, mas também em toda a sociedade contemporânea. Hoje em dia, milhões de usuários publicam e têm acesso à informação livremente na Internet através da Web, fazendo uso da rede com os mais diversos objetivos. Além disso, a Web deve tornar-se um veículo de comunicação ainda mais importante no futuro, visto que o número de usuários e de aplicações tende a crescer com o passar do tempo.

A obtenção de dados gerais sobre a Web, tais como o número de páginas HTML existente, o número de servidores e outros dados globais podem ser extremamente úteis para um melhor entendimento deste novo meio de comunicação. Estudos recentes têm procurado levantar dados gerais sobre a Web mundial com base em levantamentos estatísticos obtidos a partir de amostragens [LG99]. No âmbito da Web brasileira, apesar de sua importância para a sociedade, há pouca informação disponível.

Este trabalho apresenta um estudo sobre o conteúdo e a estrutura da Web brasileira. O estudo foi realizado com o auxílio de agentes implementados especialmente para coletar documentos da Web brasileira (páginas HTML e arquivos PostScript, PDF, Word e texto). Durante o processo de coleta, foram armazenados também diversos outros dados, tais como o número de *links* que apontam para um determinado documento e o número de *links* inconsistentes. Os dados coletados foram analisados e a partir deles são tecidas considerações sobre diversas características importantes. Até onde sabemos, os resultados apresentados aqui representam a mais completa fonte de informação disponível sobre a Web brasileira.

Este artigo está organizado da seguinte forma. Na Seção 2 o problema da coleta de documentos na Web é discutido. A Seção 3 descreve como foi feita a coleta de documentos utilizada neste artigo. Na Seção 4 são apresentados os resultados obtidos sobre a estrutura e o conteúdo da Web brasileira. Na Seção 5 são tecidas considerações a respeito do tamanho da Web brasileira. Finalmente, na Seção 6 são apresentadas conclusões sobre o trabalho realizado.

2 O Problema de Coleta de Documentos na Web

Dada uma especificação dos documentos a serem coletados, por exemplo uma especificação caracterizando a parte da Web que diz respeito ao Brasil, um processo ideal de coleta teria que obter todos os documentos que satisfizessem essa especificação. Contudo, esta tarefa é extremamente complexa e não pode ser realizada na forma como a Web funciona atualmente [MM98]. Por isto, a tarefa de coleta normalmente é relaxada para que se recupere o maior subconjunto possível de documentos que atendem à especificação fornecida.

A necessidade deste relaxamento pode ser compreendida se considerarmos a Web como um grafo direcionado, onde cada *URL* é um vértice e cada *link* de uma *URL* p_1 para uma *URL* p_2 é uma aresta do grafo saindo do vértice correspondente a p_1 e chegando no vértice

correspondente a p_2. O grafo que representa a Web pode ser não conexo, pois há diversas situações que podem ocasionar sua ruptura como, por exemplo, um usuário que publique uma *URL* sem que haja *links* apontando para a mesma. Neste caso, a referida *URL* passaria a fazer parte da Web, mas o vértice que a representa não poderia ser atingido a partir de outro vértice. Para um usuário qualquer visitar tal *URL*, seria necessário que ele a conhecesse previamente. É fácil ver que o mesmo raciocínio também é válido para a Web brasileira.

Uma solução para realizar a coleta de documentos é escolher alguns vértices como ponto de partida no grafo e visitar todos os pontos que puderem ser atingidos a partir deste conjunto inicial. Esta solução não garante que todos os vértices do grafo sejam visitados, visto que o grafo pode ser desconexo e que o conteúdo pode mudar enquanto o grafo está sendo percorrido. Contudo, essa idéia pode ser utilizada para que se obtenha uma aproximação do conjunto de documentos que deseja-se coletar. Esta aproximação pode ser usada para estimar as características do conjunto completo. Esta idéia é utilizada na Seção 5 para estimar o tamanho da Web brasileira.

3 Coleta de Documentos na Web Brasileira

O processo de coleta de onde se originaram os resultados apresentados neste artigo foi realizado utilizando o CoBWeb [SVG+99], um coletor automático de documentos Web que opera de forma distribuída objetivando principalmente eficiência, robustez e parcimônia na utilização dos recursos compartilhados. O CoBWeb foi implementado com o objetivo de alimentar a base de documentos do TodoBR (www.todobr.com.br), uma máquina de busca para a Web brasileira, e também com o objetivo de realizar levantamentos sobre sua estrutura e conteúdo. O CoBWeb coleta somente documentos Web indexáveis encontrados em servidores HTTP cujos domínios DNS sejam sub-domínios do domínio ".br". Dessa forma, considera-se a Web brasileira como sendo composta por esse conjunto de servidores, embora existam servidores fora desse escopo que também poderiam ser incluídos por hospedarem páginas relacionadas ao Brasil.

Em uma sessão típica de coleta, vários coletores são executados simultaneamente coordenados por um *escalonador* central. Os coletores são distribuídos por diversas máquinas aumentando a velocidade da tarefa de coleta de documentos.

4 Resultados Obtidos

Nesta seção são apresentados os resultados obtidos, a partir do processo de coleta de documentos da Web brasileira realizado em abril de 2000, sobre sua estrutura e conteúdo.

Foram encontrados 164.948 servidores Web, espalhados em 142.196 domínios distintos, com uma média de 70 documentos por servidor, perfazendo um total de 11.546.360 documentos, que ocupam mais de 99,2 Gbytes.

A Tabela 1 apresenta dados gerais sobre os tipo de documentos encontrados durante a coleta. São apresentados o percentual que cada tipo representa em relação ao número

Formato	Documentos (%)	Tamanho Médio(bytes)		Tamanho Total(Gbytes)	
		forma original	só texto	forma original	só texto
HTML	97,18%	7405	2141	77,38	22,37
PDF	0,33%	251220	57460	8,91	2,04
PS	0,13%	220699	15970	3,08	0,22
MS-WORD	0,41%	106136	19087	4,68	0,84
Outros	1,95%	24951	-	5,23	-

Tabela 1: Distribuição dos documentos por tipo

total de documentos, o tamanho médio dos arquivos quando armazenados no formato original e o tamanho médio dos arquivos considerando apenas o texto encontrado nos mesmos. Quase todos os documentos encontrados são do tipo HTML e o segundo tipo mais encontrado é o de documentos no formato MS-Word, da Microsoft. Os documentos marcados como *outros* incluem todos os demais tipos de documento encontrados, tais como documentos texto e documentos em formato RTF. Para estes documentos não foi realizada a extração do texto. Os arquivos no formato PDF são os que apresentam maior tamanho na média, inclusive quando consideramos apenas o texto neles encontrado. Observa-se ainda que, na média, dois terços do tamanho total dos arquivos HTML são gastos com *tags*.

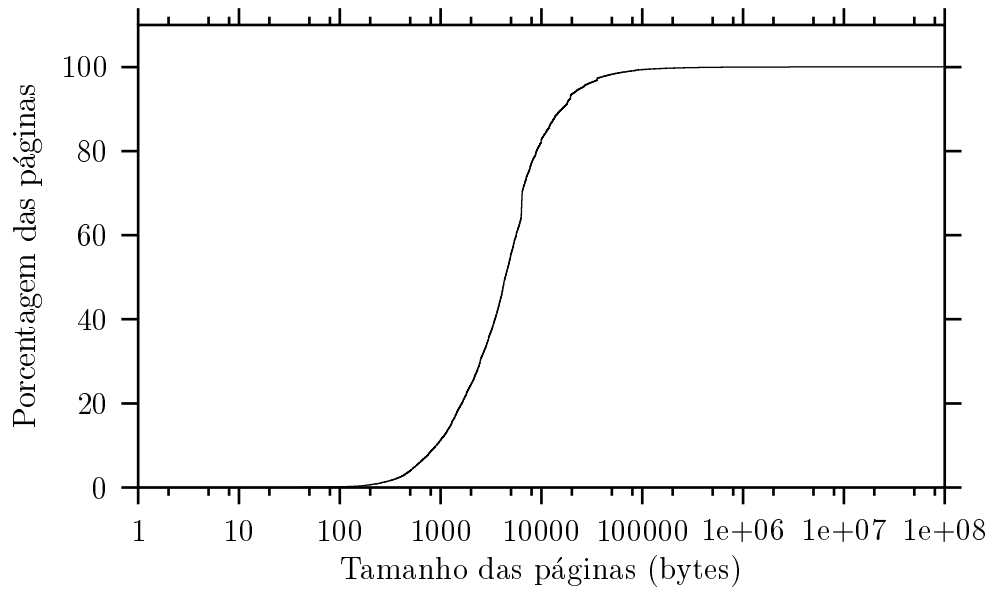

Figura 1: Gráfico acumulativo com a distribuição dos documentos por tamanho

A Figura 1 mostra a distribuição dos documentos por tamanho. Percebe-se que cerca de 80% dos documentos têm tamanho entre 1 e 10 Kbytes e quase todos os documentos têm tamanho entre 150 bytes e 100 Kbytes. Na média, os documentos encontrados têm

um tamanho de 9,01 Kbytes.

Idioma	Documentos (%)
Português	75,25%
Inglês	19,13%
Espanhol	1,27%
Outros	4,35%

Tabela 2: Distribuição dos documentos por idioma

A Tabela 2 apresenta os três idiomas mais usados nos documentos da Web brasileira. Como pode ser visto, 75,25% estão em língua portuguesa e a língua estrangeira mais utilizada na Web brasileira é o inglês, com um total de 19,13% do total dos documentos.

Palavra	Documentos (%)
copyright	16,23%
site	15,87%
mail	15,85%
page	14,75%
brasil	13,73%
paulo	12,25%
pagina	11,63%
internet	11,60%
home	10,85%
links	10,49%

Tabela 3: Substantivos mais encontrados nos documentos da Web brasileira

A Tabela 3 apresenta as dez palavras mais encontradas. A lista inclui apenas substantivos, desconsiderando outros tipos de palavra que também aparecem com muita freqüência, tais como artigos, adjetivos e preposições.

Domínio	Documentos (%)
COM	73%
ORG	5%
GOV	4%
Outros	18%

Tabela 4: Distribuição dos documentos por domínio

A Tabela 4 apresenta os três tipos de domínio com maior número de documentos. A maior parte dos documentos está no DPN (domínio de primeiro nível). O domínio *com*, hospeda 73% dos documentos encontrados. A porção incluída em *outros* é composta principalmente pelos documentos de universidades, que possuem domínios próprios.

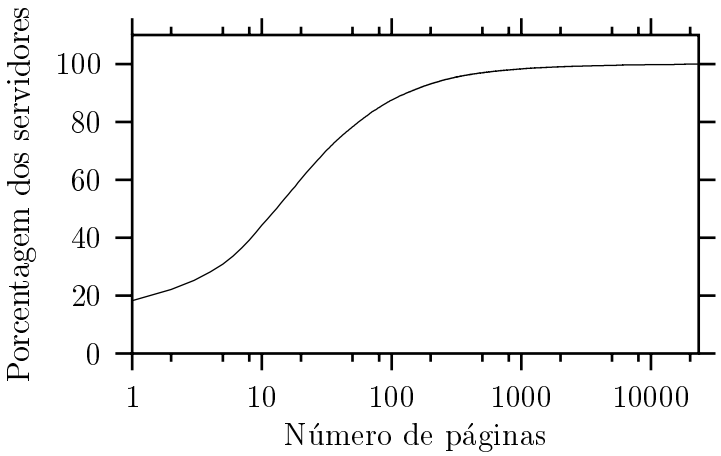

Figura 2: Gráfico acumulativo com o número de documentos por servidor

A Figura 2 mostra a distribuição do número de documentos por servidor. Cerca de 20% dos servidores possuem apenas 1 documento e mais de 80% dos servidores têm menos de 100 documentos. O servidor com maior número de documentos encontrado durante a coleta tem 23.284 documentos. Na média, foram encontrados 70 documentos por servidor.

As Figuras 3 e 4 apresentam dados organizados por níveis. Os níveis são contados através da estrutura de diretórios encontrada dentro dos servidores. O diretório raiz constitui o nível 1, os sub-diretórios do diretório raiz constituem o nível 2 e assim por diante. Por exemplo, a página http://www.dcc.ufmg.br/index.html está no nível 1, enquanto a página http://www.dcc.ufmg.br/pesquisa/index.html está no nível 2.

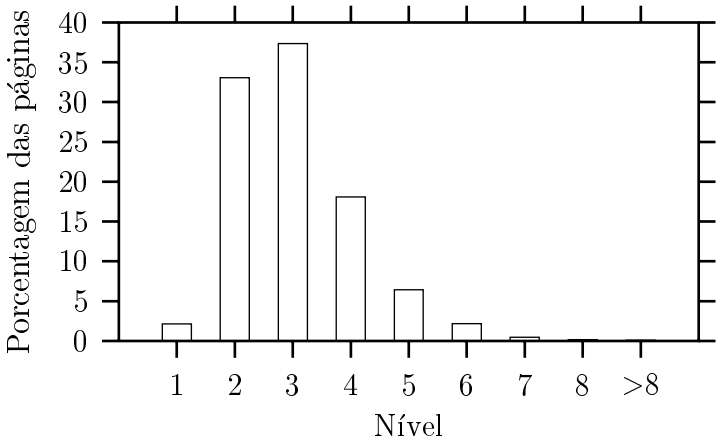

Figura 3: Número de documentos por nível

A Figura 3 mostra a distribuição do número de documentos por nível. A maior parte dos documentos encontra-se nos níveis 2, 3 e 4, que somados possuem 87% do total. Menos de 2% dos documentos estão acima do nível 6.

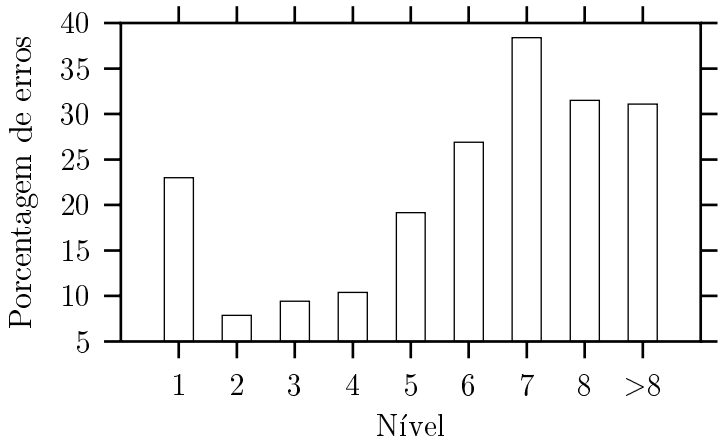

Figura 4: Percentual de erros por nível

A Figura 4 mostra o percentual, para cada nível, de erros encontrados pelo coletor em *links* seguidos. O percentual de erros em cada nível foi medido em relação ao total de *links* seguidos no nível. Por exemplo, 38% dos *links* para documentos do nível 7 encontrados na Web brasileira resultaram em erro. A quantidade de erros encontradas no nível 1 é alta porque nesse nível são incluídos os erros causados por servidores que estão inativos ou temporariamente desativados.

Tipo de Erro	% do total de erros	% do total de requisições
BAD REQUEST (400)	0,12%	0,03%
FORBIDDEN (403)	2,02%	0,23%
NOT FOUND (404)	64,73%	7,09%
INTERNAL SERVER ERROR (500)	29,28%	3,19%
NOT IMPLEMENTED (501)	0,66%	0,09%
SERVICE TEMP. UNAVAILABLE (503)	0,18%	0,03%
OUTROS	3,01%	0,34%

Tabela 5: Tipos de erro encontrados

A classificação dos erros mais encontrados durante o processo de coleta de documentos é apresentada na Tabela 5. A nomenclatura utilizada para a classificação dos erros segue a descrita em [FGM+97]. O percentual de erros está representado tanto em relação ao total de erros encontrados (coluna do meio) quanto em relação ao total geral de requisições (coluna da esquerda). O erro mais freqüênte corresponde ao de documentos não encontrados, o que indica que há uma grande quantidade de *links* inconsistentes. Este fato pode ser atribuído à velocidade com que os dados publicados na Web são alterados.

A Figura 5 mostra um gráfico com a distribuição dos documentos conforme o número de *links* que possuem. O gráfico mostra que os documentos que possuem até 30 *links* correspondem a mais de 96% do total e que 22% dos documentos não apresentam *links*

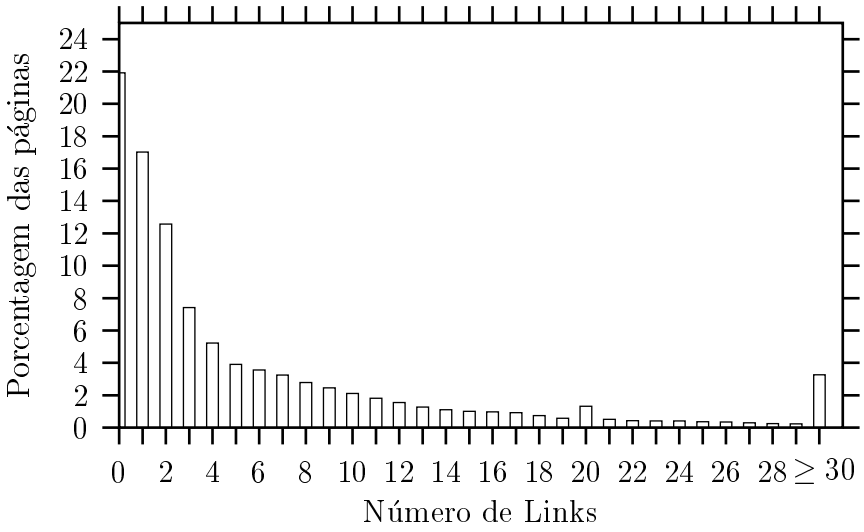

Figura 5: Distribuição das páginas por número de *links*

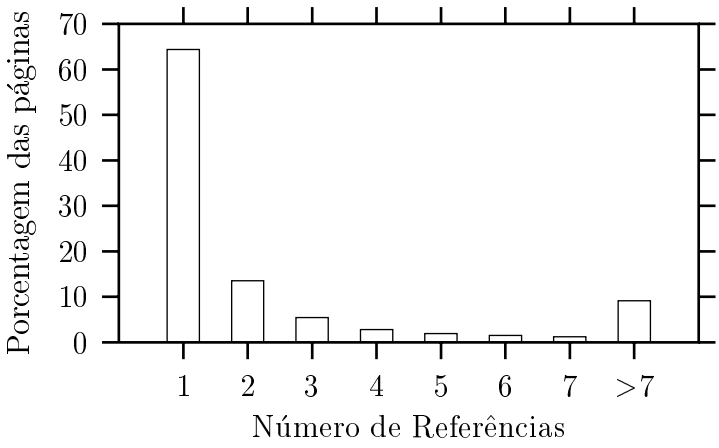

Figura 6: Distribuição das páginas pelo número de referências

para nenhum outro documento. Na média, foram encontrados 6,74 *links* por documento.

A Figura 6 mostra a distribuição das páginas de acordo com o número de *links* que as citam (referências). Como pode ser visto, mais de 63% das páginas possuem apenas 1 referência. Resultados semelhantes também são encontrados em estudos sobre a estrutura de *links* da Web mundial [PBMW98].

Embora não tenha sido realizada a coleta de imagens, foi computado o número de referências a imagens existente em cada página HTML. A partir destes dados, foi gerado o gráfico da Figura 7, que apresenta a distribuição das páginas HTML de acordo com o número de imagens referenciadas. Como pode ser visto, 24% das páginas da Web brasileira não faz referência a imagens e 53% do total faz referência a menos de três imagens. Na

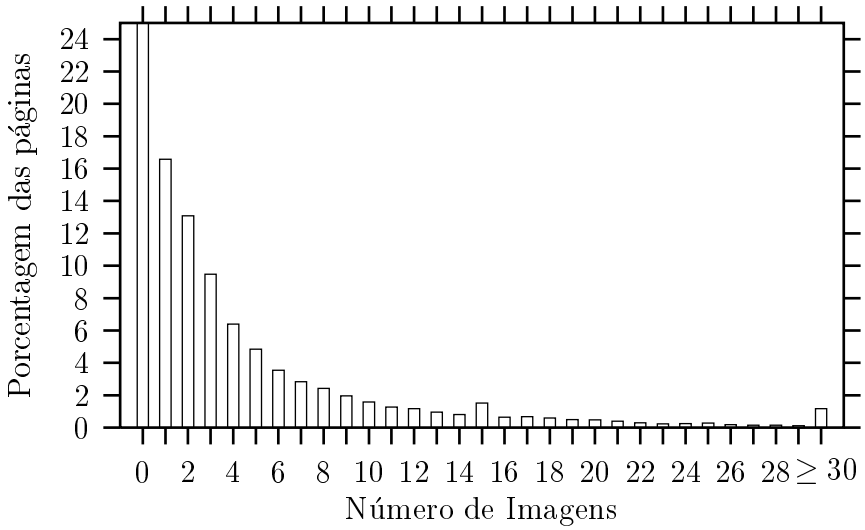

Figura 7: Distribuição das páginas pelo número de imagens

média, há 4,6 referências a imagens por página.

5 Qual é o Tamanho da Web Brasileira ?

Como mencionado na Seção 2, não é possível determinar o número exato de documentos da Web brasileira. Contudo, uma consulta à página de estatísticas de domínios DNS da Web brasileira cadastrados na FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo (http://www.fapesp.br), na ocasião da coleta dos documentos, revelou que havia cerca de 220.000 domínios cadastrados com DNS válido. Isto significa que a coleta deixou de visitar cerca de 78 mil domínios dentre os cadastrados na FAPESP. Considerando-se as médias de documentos por servidor (70) e de servidores por domínio (1,16) obtidas no processo de coleta, pode-se estimar que há $1,16 \times 70 \times 220.000 \cong 17.870.000$ documentos na Web brasileira. Isso significa que a coleta realizada obteve cerca de 64% do total de páginas.

No entanto, é importante observar que esta estimativa considera que os 78 mil domínios que não foram encontrados durante a coleta possuem em média a mesma quantidade de servidores e de documentos encontrados nos 142 mil domínios visitados. Na realidade, muitos desses domínios podem não ter sido considerados na coleta por não possuírem servidores ou ainda por não estarem ativos. Além disso, há muitos servidores que não tiveram suas páginas coletadas por serem protegidos contra acesso não autorizado.

6 Conclusões e Trabalhos Futuros

Os dados apresentados neste artigo constituem um retrato da situação atual da Web brasileira. Estes dados podem servir de fonte de informação para inúmeras pesquisas relacionadas à Web.

Como trabalho futuro, pretende-se estudar métricas para avaliar o quão próximo uma dada coleta está do total de documentos real existente. Também pretende-se estudar mecanismos para incluir na coleta páginas de outros domínios que digam respeito ao Brasil. Esta inclusão poderia dar uma idéia melhor sobre o conteúdo dos documentos colocados pela comunidade brasileira à disposição dos usuários de todo o mundo. Outra possível aplicação para os mecanismos de coleta desenvolvidos seria a busca e identificação de sites ilegais dentro da Internet, tais como cópias, páginas com conteúdo proibido, etc. Além disto, pretende-se realizar estudos sobre a evolução do conteúdo da Web brasileira ao longo dos anos, para que seja possível traçar no futuro um histórico sobre as mudanças ocorridas com o passar dos anos.

Referências

[FGM+97] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. Technical Report RFC 2068, January 1997.

[LG99] S. Lawrence and C. L. Giles. Accessibility of infomation on the web. *Nature*, 400(8):107–109, 1999.

[MM98] A. Mendelzon and T. Milo. Formal models of web queries. *Information Systems*, 23(8):615–637, 1998.

[PBMW98] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[SVG+99] A. Silva, E. Veloso, P. Golgher, B. Ribeiro, A. Laender, and N. Ziviani. CoBWeb: A Crawler For the Brazilian Web. In *Proc. of the 6th International Symposium on String Processing and Information Retrieval (SPIRE'99)*, pages 184–191. Carleton University Press, 1999.