Um novo retrato da Web brasileira

Marco Modesto[1], **Álvaro R. Pereira Jr**[1], **Nivio Ziviani**[1],
Carlos Castillo[2], **Ricardo Baeza-Yates**[2]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 – 31270-901 Belo Horizonte, MG

[2]Centro de Investigación de la Web – Departamento de Ciencias de la Computación
Universidad de Chile – Av. Blanco Encalada 2120, Tercer Piso – Santiago, Chile

{mabm, alvaro, nivio}@dcc.ufmg.br,{ccastill, rbaeza}@dcc.uchile.cl

Resumo. *O objetivo deste artigo é avaliar características quantitativas e qualitativas da Web brasileira, confrontando estimativas atuais com estimativas obtidas há cinco anos. Grande parte do conteúdo Web é dinâmico e volátil, o que inviabiliza a sua coleta na totalidade. Logo, o processo de avaliação foi realizado sobre uma amostra da Web brasileira, coletada em março de 2005. Os resultados são estimados de forma consistente, usando uma metodologia eficaz, já utilizada em trabalhos similares com Webs de outros países. Dentre os principais aspectos observados neste trabalho estão a distribuição dos idiomas das páginas, o uso de ferramentas abertas versus proprietárias para geração de páginas dinâmicas, a distribuição dos formatos de documentos, a distribuição de tipos de domínios e a distribuição dos links a Web sites externos.*

Abstract. *The objective of this paper is to evaluate quantitative and qualitative characteristics of the Brazilian Web, matching present estimatives with estimatives from five years ago. Most of the Web content is dynamic and volatile, becoming the crawling of the total Web content a impracticable task. Thus, the evaluation process was performed over a sample of the Brazilian Web, crawled on March 2005. The results are consistently estimated, using an effective methodology. Many statistical data are presented. Among the main aspects observed in this work are the distribution of idioms of the pages, usage of open source versus private development tools for generation of dynamic pages, distribution of document formats, distribution of types of domains and the distribution of the links to external Web sites.*

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 cresce com o passar do tempo.

Este trabalho apresenta um estudo recente realizado sobre a Web brasileira, através de estimativas consistentes. O estudo abrange análise quantitativa e qualitativa. Com relação à análise quantitativa, estima-se o atual tamanho de toda a Web brasileira, o número de páginas `html` existentes, o número médio de páginas por site e por domínio, o tamanho médio de arquivos multimídia por página, dentre outras informações relevantes. Com relação à análise qualitativa, busca-se compreender os países que são mais referenciados por páginas brasileiras, as tecnologias mais utilizadas, as linguagens mais usadas. Além dessa análise, o estudo apresenta importantes comparações da Web atual com a Web brasileira de cinco anos atrás [Veloso et al., 2000], avaliando sua evolução neste período.

Estudos de caracterização de Webs possuem diversas aplicações, nas quais destacam-se: avaliação de arquiteturas de softwares de coleta; melhoria das funções de ranqueamento de páginas em máquinas de busca; estudo de comportamentos sociais; estudos linguísticos; entre outros. O confronto de coletas realizadas em diferentes épocas fornece uma base para estimativas sobre o futuro da Web.

Alguns conceitos são importantes para o entendimento do restante do trabalho. *Documento* é o arquivo resultante de uma requisição HTTP correta (exemplo: `html`, `pdf`, `doc`). Uma *página* é um documento no formato `html`. Um *domínio* é qualquer nome da forma $x.y.z$, onde y é o domínio de primeiro nível regulamentado pelo Registro.br[1]. Um *Web site* ou *site* representa uma coleção de documentos referenciados por URLs que dividem o mesmo endereço de domínio. *Níveis* são contados através da estrutura de diretórios encontrada dentro dos servidores. O diretório raiz constitui o nível zero, os sub-diretórios do diretório raiz constituem o nível um e assim por diante. Por exemplo, `http://www.dcc.ufmg.br` e `http://www.ee.ufmg.br` são URLs de Web sites diferentes que pertencem ao mesmo domínio `ufmg.br`. A URL `http://www.ufmg.br/dcc/webbr.html` corresponde a uma página que pertencente ao Web site `www.ufmg.br` e está no nível um. `http://www.ufmg.br/webbr.html` está no nível zero.

Algumas distribuições que apresentamos nesse artigo seguem a lei de Zipf [Zipf, 1949], chamada lei quantitativa fundamental da atividade humana.Na lei de Zipf, a frequência de um evento é inversamente proporcional ao seu "rank". A frequência do *i*-ésimo evento mais frequente é proporcional a $1/i^{\Theta}$ vezes a do evento mais frequente, $\Theta \geq 1$.

Destacam-se os seguintes trabalhos relacionados. Em 1999, Lawrence e Lee Giles levantaram dados gerais sobre a Web mundial com base em levantamentos estatísticos obtidos a partir de amostragens [Lawrence and Giles, 1999], metodologia diferente da utilizada neste trabalho. No âmbito da Web brasileira, [Veloso et al., 2000], também fizeram um trabalho estatístico, porém baseado em páginas coletadas através de caminhamento pelos *links* de um conjunto de "sementes" de páginas, utilizando o coletor COBWeb [Silva et al., 1999]. Lawrence e Lee Giles estimaram o tamanho da Web em 15 Terabytes e Veloso et al estimaram a Web brasileira em 121 Gigaby-

[1]Órgão que regulamenta os domínios da Internet no Brasil (http://www.registro.br).

tes. Tais valores correspondem ao espaço ocupado pelos documentos `html`, excluindo imagens e outros tipos de arquivos. Estudos realizados no Chile [Castillo, 2004], Grécia [Efthimiadis and Castillo, 2004] e Coréia do Sul [Baeza-Yates et al., 2004] utilizaram o mesmo coletor deste trabalho, o que dá uma validade maior aos resultados.

2. Definição da Web brasileira

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: os Web sites com páginas dinâmicas podem ter um número infinito de páginas [Baeza-Yates and Castillo, 2004]. Por isto, a tarefa de coleta normalmente é relaxada para que se recupere o maior subconjunto possível de documentos que atendem à especificação fornecida.

Para tornar o problema computável, a forma estudada para restringir a Web brasileira baseou-se no domínio dos Web sites, onde se considerou todas as páginas que possuem domínio com terminação `.br`. Sites com conteúdo brasileiro e hospedados em domínios não `.br` não foram coletados. Para evitar a coleta de um número infinito de páginas de um Web site, limitou-se o número de níveis das páginas estáticas e dinâmicas e o número de páginas do Web site.

O domínio `.br` é uma boa representação da Web brasileira. Não é difícil para um usuário Web perceber que a grande maioria dos Web sites brasileiros estão dentro do domínio `.br`. Os principais motivos devem ser a facilidade para registro do domínio e o custo: atualmente registrar um domínio `.br` possui um dos menores custos do mundo: 30 Reais (aproximadamente 11 dólares) anuais, ficando mais econômico do que domínios `.net` e `.com`. Em países como a Espanha, onde o domínio nacional é caro e difícil de se conseguir, a maioria dos sites usam domínios internacionais, como o `.com`.

2.1. Medindo a Web através de caminhada por *links*

Pode-se considerar 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 2 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.

Para realizar a coleta de documentos, escolhem-se alguns vértices como ponto de partida no grafo e visita-se todos os pontos que puderem ser atingidos a partir deste conjunto inicial (semente). Um bom conjunto inicial, como a lista completa dos domínios registrados no país é muito importante para alcançar uma grande parte do grafo. 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.

3. Experimentos

3.1. Coleta

O coletor utilizado foi o WIRE [WIRE, 2004], desenvolvido por Carlos Castillo [Castillo, 2004] no Centro de Pesquisa da Web[2] da Universidade de Chile. Seu ponto forte é o desempenho em coletas, pois utiliza algoritmos como o *Pagerank*: páginas mais relevantes podem ser coletadas antes. O WIRE começa a coleta por uma semente de domínios e alcança outros através da caminhada por *links*, adicionando-os à fila.

Como semente de domínios, usou-se a lista do TodoBR[3] de dezembro de 2003 (coleção WBR2003) com URLs com prefixo `www` e sufixo `.br`. Na época da coleta, haviam 657 mil domínios registrados com DNS válido no Registro.br. Estima-se que a coleta de todos estes sites dure várias semanas. Então se decidiu fazer um levantamento por amostragem limitando a coleta em uma semana. Limitou-se a profundidade das páginas estáticas em 6 níveis, das páginas dinâmicas em 5 níveis e o número de páginas por Web site em 10.000. Valores similares foram usados em outros trabalhos [Castillo, 2004]. Arquivos binários (.mp3, .avi, .wav, etc) não foram coletados.

A tabela 1 apresenta informações gerais obtidas na coleta. Detectou-se que 6% das páginas visitadas eram duplicadas: alguns Web sites mantém cópias de suas páginas em diferentes locais. Um Web site tem em média 85 páginas, 1,2 Megabytes de espaço e 3,36 níveis de profundidade. O tamanho médio de uma página é de 14,4 Kilobytes.

Total de páginas	Percentual
Únicas	93.60%
Duplicadas	6.40%
Estáticas	58.30%
Dinâmicas	41.70%

Web sites	Valor
Páginas	85,28
Profundidade máxima média	3,36
Tamanho médio (MB)	1,20

Tabela 1: Sumários da coleta em 2005

4. Resultados - Características das páginas

4.1. Tamanho

Para economizar tempo e banda de rede, foram baixados apenas os primeiros 300 Kilobytes de cada página, mesmo valor utilizado em outros trabalhos [Castillo, 2004]. O número de páginas que ultrapassam este tamanho é pequeno. O centro da distribuição do tamanho das páginas segue a Lei de Zipf com parâmetro -3,59, como mostrado na figura 1. Próximo a 300 Kilobytes o número de páginas é maior do que o esperado devido ao limite de *download*.

Abaixo de 20 Kilobytes a distribuição de páginas não segue a Lei de Zipf: existem poucas páginas com tamanho inferior a este. Isso ocorre devido à codificação das tags `html` que gastam um espaço considerável: em média 75% para um arquivo de 10 Kilobytes. Mesmo um texto pequeno ocupa um espaço grande quando é transformado

[2] http://www.ciw.cl

[3] http://www.todobr.com.br

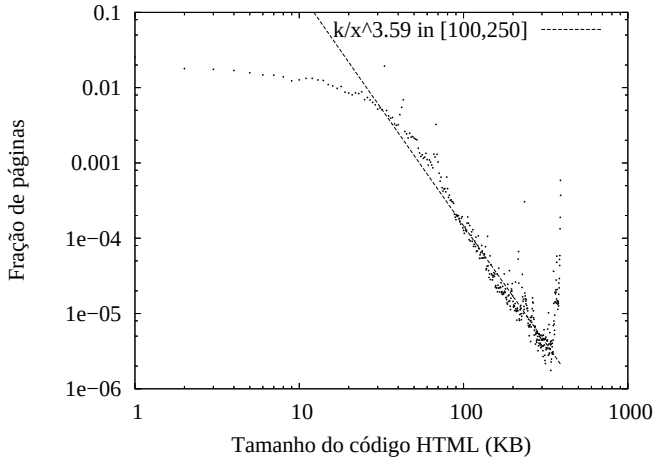


Figura 1: Tamanho das páginas em 2005 (Kilobytes).

para uma página `html`, principalmente quando são usados editores `html` automáticos. A distribuição dos tamanhos das páginas possui uma forte tendência: 50% das páginas contêm apenas 1% do espaço total ocupado por todas as páginas.

Em [Veloso et al., 2000] percebeu-se que cerca de 80% das páginas `html` têm tamanho entre 1 e 10 Kilobytes e quase todas elas têm tamanho entre 150 bytes e 100 Kilobytes. Na média, as páginas encontradas tinham um tamanho de 9,01 Kilobytes.

4.2. Idiomas

As metodologias para detectar o idioma da página são diferentes para as pesquisas de 2000 e da atual. Na atual, para avaliar o idioma das páginas, definiu-se um conjunto de amostra com 9.426 páginas. Excluíram-se as páginas com menos de 50 palavras. Suas listas de palavras foram comparadas com uma série de listas de "stop words" em vários idiomas, o que permitiu identificar o idioma de 3.366 páginas.

Idioma do documento	2000	2005
Português	75,25%	88,63%
Inglês	19,13%	11,20%
Espanhol	1,27%	1,16%
Italiano	–	0,24%
Francês	–	0,24%
Total	95,65%	99,47%

Tabela 2: Distribuição dos idiomas na Web brasileira.

Segundo a tabela 2, no ano 2000 tinhamos 75% das páginas em português e 19,13% em inglês. Em 2005, 87% das páginas estavam em português e 11% em inglês. O uso do inglês caiu aproximadamente pela metade em 2005 quando comparado com 2000. Outros idiomas aparecem com uma freqüência muito menor.

4.3. Domínios

A tabela 3 apresenta a distribuição dos domínios obtida por métodos diferentes. Em 2000 extraiu-se a distribuição das URLs das páginas coletadas. Os dados de 2005 baseiam-se na quantidade de domínios registrados no Registro.br. Os domínios `.br` e `.edu.br` são associados a Universidades e os `.net.br` são associados a serviços de telecomunicações. Ambos foram inclusos em *Outros* no trabalho de 2000. Em 2005 existiam 58 tipos de domínios da Web brasileira, mas 91,31% dos domínios concentravam-se

no tipo `.com.br`, destinado a organizações comerciais. O terceiro domínio mais comum (`.adv.br`) e o quarto (`.ind.br`) não estão listados na tabela: são destinados respectivamente a advogados e a indústrias. Os dois tipos mais comuns permanecem os mesmos nos dois momentos: `.com.br` e `.org.br` (entidades não governamentais sem fins lucrativos).

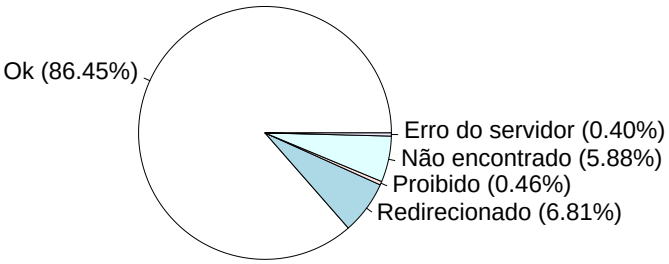
Domínio	2000	2005
.com.br	73%	91,31%
.org.br	5%	2,74%
.br, .edu.br	–	0,29%
.gov.br	4%	0,11%
.net.br	–	0,07%
Outros	18%	5,48%
Total	100%	100,00%

Tabela 3: Distribuição dos domínios

4.4. Código HTTP

A figura 2 mostra a distribuição do código de status HTTP. Para uma maior clareza, na figura juntou-se vários códigos:

- Ok: inclui as respostas `Ok` (200) e `Conteúdo parcial` (206).
- Redirecionado: inclui `Redirecionado` (301).
- Não encontrado: inclui `Não encontrado` (404).
- Erro do servidor: inclui `Erro interno do servidor` (500), `Gateway com problemas` (502), `Indisponível` (503), e `Sem conteúdo` (204).
- Proibido: inclui `Desautorizado` (401), `Proibido` (403) e `Inaceitável` (407).



Código HTTP	2000	2005
Ok	89,00%	86,45%
Redirecionado	–	6,81%
Não encontrado	7,09%	5,88%
Erro do servidor	3,22%	0,40%
Proibido	0,23%	0,46%
Outros	0,46%	–
Total	100,00%	100,00%

Figura 2: Resposta do servidor

A classificação do código de status recebidos durante o processo de coleta de documentos é apresentada na Tabela 2. A classificação do trabalho realizado em 2000 é diferente da atual. O valor de *Outros* está bem diferente nas duas coletas porque códigos como `Redirecionado` (301) foram classificados em grupos diferentes. O percentual de *links* quebrados (páginas não encontradas) é relativamente baixo: 6% em 2005. Isto mostra uma maior preocupação na consistência dos *links*, por parte de quem publica. A taxa das páginas processadas sem erros é basicamente igual nas duas épocas: 89% e 86%.

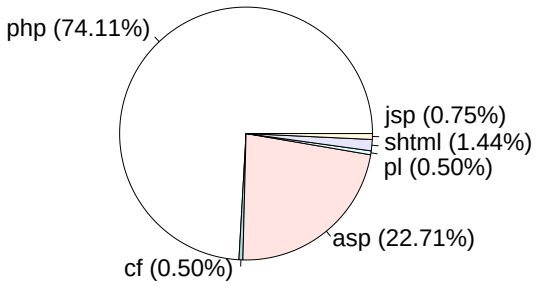
4.5. Tipos de arquivo

A análise dos *links* com arquivos oferece boa oportunidade de mensurar o uso de diversos padrões de arquivos. Entre as 20 extensões de arquivos mais encontradas há formatos

de imagens, arquivos relacionados a `html`, páginas dinâmicas, animações, documentos e códigos fonte de linguagens de programação. As próximas subseções analisam a distribuição de alguns dos tipos de arquivos encontrados.

4.6. Páginas dinâmicas

As páginas dinâmicas são páginas construídas usando um pré-processador de hipertexto, onde o conteúdo destas páginas é geralmente recuperado de um banco de dados no momento em que o usuário acessa a página. A classificação quanto ao tipo da página (estática/dinâmica) se deu pela extensão do arquivo da página.



Extensão	Percentual
`php`	74,11%
`asp`	22,71%
`shtml`	1,44%
`jsp / jhtml`	0,75%
`Cold Fusion`	0,50%
`Perl`	0,50%
Total	100,00%

Figura 3: Tipos de geradores de páginas dinâmicas em 2005

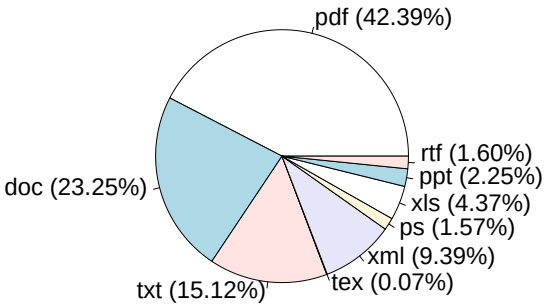
Aproximadamente 3,2 milhões de páginas dinâmicas foram coletadas, ou em termos relativos 41% do total de páginas. Ao percorrer as páginas coletadas, encontrou-se 130 milhões de *links* para páginas dinâmicas, cuja distribuição do tipo de tecnologia utilizada é mostrada na figura 3. A aplicação mais usada foi o `php`, seguida por `asp`, `ssi` (`.shtml`) e `java` (`.jhtml/.jsp`). O `php` é uma tecnologia de código aberto com a vantagem de ser bastante flexível com milhares de módulos especialmente desenvolvidos para páginas Web. Isso sugere o motivo porque `Perl`, uma linguagem com propósitos mais amplos e poucos módulos, é pouco usada na Web. O `php` teve 3 vezes mais ocorrências que o `asp` da Microsoft.

4.7. Documentos

Encontrou-se 39 milhões de *links* para arquivos com extensões usadas para documentos. O formato `html` é líder seguido por `pdf`, Microsoft Word e texto plano. A distribuição é mostrada na figura 4. A terceira coluna da tabela e o gráfico mostram a distribuição em 2005 dos documentos excluindo-se os `html`. A Microsoft domina o mercado de Sistemas Operacionais para PCs, porém a quantidade de referências para documentos gerado por seus produtos “Office” é inferior a outros formatos. [Castillo, 2004] afirma que o motivo deste fato possa ser as preocupações com vírus ou com perda de formatação.

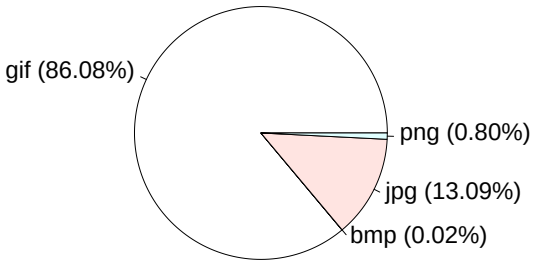
4.8. Multimídia

Foram encontrados diversos *links* para arquivos multimídia, sendo a grande maioria para imagens: 160 milhões. *Links* para arquivos de áudio compreendem 160 mil e 46 mil *links* para arquivos de vídeo. A distribuição dos *links* para tipos de imagens é mostrada na figura 5. O formato `gif` da Compuserve é o mais usado para imagens, seguido por `jpeg`. O formato `gif` é geralmente utilizado para desenhos e o formato `jpeg` para fotos. Como as páginas Web contém muito mais desenhos que fotos, o formato `gif` é mais comum. O formato de código aberto `png`, que foi concebido para substituir o formato `gif`, ainda é pouco usado.



Extensão	2000	2005	Não `html`
`html`	97,18%	97,92%	–
`pdf`	0,33%	0,88%	42,39%
`doc`	0,41%	0,48%	23,25%
`txt`	–	0,31%	15,12%
`xml`	–	0,19%	9,39%
`xls`	–	0,09%	4,37%
`ppt`	–	0,05%	2,25%
`rtf`	–	0,03%	1,60%
`ps`	0,13%	0.03%	1,57%
`tex`	–	0.00%	0,07%
Outros	1,95%	–	–
Total	100,00%	100,00%	100,00%

Figura 4: Distribuição dos tipos de documentos

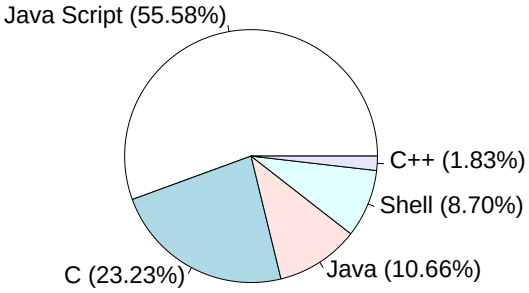


Extensão	Percentual
`gif`	86,08%
`jpg`	13,09%
`png`	0,80%
`bmp`	0,02%
Total	100,00%

Figura 5: Distribuição do formato de imagens em 2005.

4.9. Código fonte

Foram encontrados 6.500 *links* com extensões associadas a códigos fontes e 182.000 arquivos com extensões associadas a softwares. A distribuição dos códigos fonte é mostrada na figura 6. Contabilizou-se o número de *links* para códigos fonte e não o número de códigos baixados. O formato mais comum de código fonte é o JavaScript com 56% das ocorrências. Como JavaScript é basicamente utilizado na Web é normal que seja o mais comum. O segundo formato de código mais encontrado são os escritos em C.



Extensão	Percentual
JavaScript	55,58%
C	23,23%
Java	10,66%
Shell	8,70%
C++	1,83%
Total	100,00%

Figura 6: Distribuição do tipo de código fonte em 2005.

5. Resultados - Características dos Web sites

5.1. Número de páginas

Na Web brasileira cada site contém em média 85 páginas. O valor mais freqüente é muito menor pois a distribuição é bastante tendenciosa, como mostrado na figura 7. Esta é uma distribuição de Zipf com parâmetro 1,61. Existem Web sites muito grandes: os 10% dos Web sites com mais páginas contêm mais de 80% das páginas Web. A distribuição dos tamanhos das páginas também é tendenciosa: os 10% dos Web sites do topo contêm 80% do total do tamanho em bytes. O coletor deve escalonar seu acesso aos Web sites com cuidado, pois como a maioria das páginas são encontradas em poucos sites, deve-se evitar o congestionamento no acesso a esses Web sites.

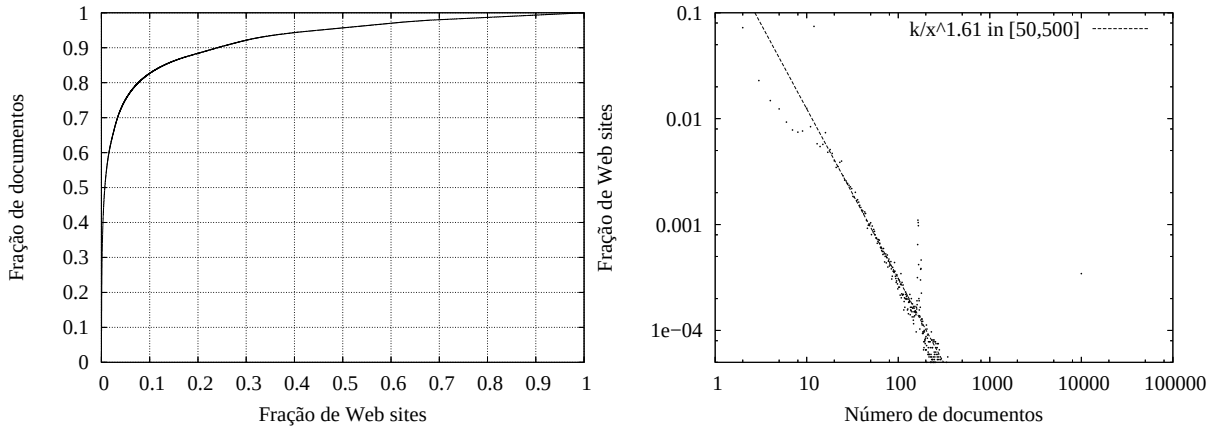


Figura 7: Número de páginas por Web site em 2005.

5.2. Links para domínios externos

Foram encontrados 137 milhões de *links* externos para outras páginas, sendo que 92% eram para páginas com domínios `.br`. O restante são para domínios externos. A distribuição dos *links* para os 10 domínios externos mais referenciados é mostrada na tabela 4.

Domínio	Percentual
`.com`	55,81%
`.org`	16,91%
`.net`	15,19%
Argentina	1,67%
Reino Unido	1,11%
`.info`	0,66%
`.edu`	0,63%
Alemanha	0,53%
Portugal	0,52%
Chile	0,49%
Outros	6,46%
Total	100,00%

Tabela 4: Domínios de primeiro nível dos *links* externos em 2005

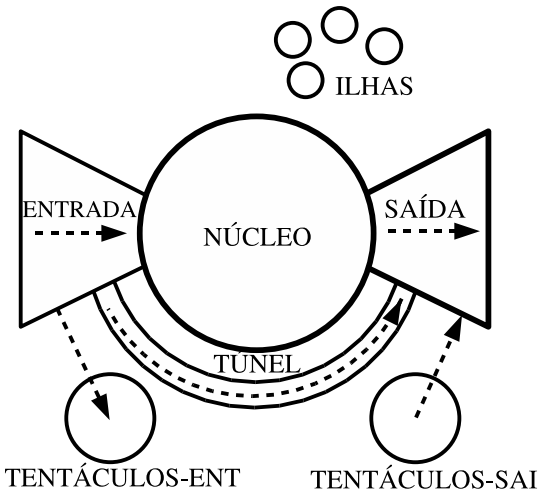
Como esperado, o maior domínio da Web mundial também é domínio externo mais referenciado pela Web brasileira: `.com`. Os `.com`, `.org` e `.net` geralmente são associados aos EUA, porém devido a razões econômicas, culturais e históricas também

são utilizados por todos os outros países. Os domínios nacionais mais referenciados são os com terminação `.ar` (Argentina), `.uk` (Reino Unido) e `.de` (Alemanha). Pode-se notar que há relações desta lista com os países em que o Brasil possui fortes relações econômicas (EUA, Argentina), influências da proximidade (Chile) e de características culturais (Portugal). Os *links* para o Reino Unido e Alemanha são comuns em todas Webs nacionais estudadas através do WIRE: estes países possuem uma forte presença na Web mundial.

5.3. Estrutura macroscópica da Web

Como dito no início do trabalho, é possível considerar a Web como um grafo direcionado. O grafo é fortemente conectado se cada dois vértices quaisquer são alcançados a partir de um outro. Os Componentes Fortemente Conectados de um grafo são os conjuntos de vértices mutuamente alcançáveis [Ziviani, 2004].

As considerações sobre a estrutura macroscópica da Web baseiam no trabalho de Broder et al [Broder et al., 2000]. Os sites são classificados em algum grupo de acordo com suas ligações (*links*) entre outros sites. O componente fortemente conectado da Web é chamado de *núcleo*. Os sites que são referenciados pelo *núcleo* mas não o referenciam são agrupados em *saída*. Os sites que referenciam o *núcleo* mas não são referenciados por ele são agrupados em *entrada*. Outros sites que são alcançados por *entrada* ou somente podem alcançar *saída* fazem parte dos *tentáculos*. Os sites que estão entre o caminho de *entrada* para *saída* fazem parte do *túnel*. Os sites desconexos, que não possuem ligação aos conjuntos são chamadas de *ilhas*.



Componente	Percentual
núcleo	25,27%
entrada	12,95%
saída	45,33%
tentáculos-ent	2,14%
tentáculos-sai	1,73%
túnel	0,23%
ilhas	12,35%

Figura 8: Estrutura macroscópica da Web em 2005

A figura 8 mostra os conjuntos e a proporção de cada um na Web brasileira. O maior componente é o *saída*, seguido pelo *núcleo*. A Web é dinâmica. Geralmente os Web sites começam em *entrada*, passam pelo *núcleo* e vão para a *saída*, onde permanecem até serem excluídos. Os Web sites pertencentes às ilhas são geralmente sites que reservam um domínio para uso futuro, sites novos ou sites dedicados a um grupo específico e restrito de usuários. Em estudos similares [Castillo, 2004], utilizou-se a lista completa de domínios registrados no país. Em tais estudos o componente das ilhas é muito maior que na medição brasileira, porque a semente deles é composta por qualquer tipo de domínios. Já a semente desta coleta de 2005 é composta por domínios de Web sites populares, que foram indexados pelo coletor do TodoBR.

6. Comparação com outras Webs

A tabela 5 compara dados demográficos e econômicos das Webs de três países: Brasil, Chile [Castillo, 2004] e Portugal [Gomes and Silva, 2003]. O IDH [Nations, 2003], índice criado pela ONU, mede o nível de desenvolvimento humano dos países a partir de indicadores de educação (alfabetização e taxa de matrícula), longevidade (expectativa de vida ao nascer) e renda (PIB per capita). A tabela 5 compara a classificação do IDH dos três países. Por exemplo, o Brasil possui o 65^o melhor desenvolvimento humano entre os 175 países onde o índice é calculado.

O número de domínios e de páginas per capita não está diretamente relacionado à riqueza ou ao IDH do país. A taxa de resposta correta do servidor HTTP ficou próxima para as três Webs. O número de domínios do Chile é maior que o número de Web sites. Muitos domínios chilenos podem estar fora da Web pública indexável ou podem estar inativos. A influência do inglês é muito menor na Web brasileira do que na Web chilena. A taxa do uso do `php` é próxima no Brasil e no Chile.

	Brasil	Chile	Portugal
População [Nations, 2004]	186,4 M	16,3 M	10,5 M
PIB [Economist, 2002]	US$ 461 bi	US$ 66 bi	US$ 147 bi
PIB per capita [Economist, 2002]	US$ 7.643	US$ 10.373	US$ 18.323
IDH [Nations, 2003]	65^o	43^o	23^o
Domínios registrados	657 K[a]	100 K[b]	47 K[c]
Domínios / 1000 habitantes:	3,5	6,1	4,5
Páginas com HTTP normal	87%	78%	84%
Uso do inglês	11%	27%	–
Uso do `php`	74%	72%	–

[a] Segundo o Registro.br existiam 736 K domínios registrados, porém 10% estavam irregulares.
[b] Dado de abril de 2004.
[c] Estimativa. Em agosto de 2004 existiam 40,6 mil domínios .PT registrados.

Tabela 5: Características demográficas e das Webs.

Sumário executivo

Os dados abaixo resumem os principais resultados estimados neste trabalho:

Sobre as características quantitativas:

- Coletou-se uma amostra com mais de 132 mil sites, com aproximadamente 7,7 milhões de páginas que ocupam mais de 91 Gigabytes de espaço.
- Cada domínio `.br` possui em média 1,1 Web sites.
- Os países mais referenciados são a Argentina, Reino Unido, Alemanha, Portugal e Chile, desconsiderando as referências para os Estados Unidos.

Sobre as características qualitativas:

- Os 10% dos Web sites com mais páginas contém mais de 80% do total de páginas.
- Os 10% dos Web sites do topo contém mais de 80% do tamanho total em bytes.
- Existem aproximadamente 6% de *links* quebrados.
- Mais de 86% das páginas são escritas em português e 11% em inglês.

Sobre as tecnologias:

- O formato de páginas dinâmicas mais utilizado é o `php`, encontrado em mais de 74% das páginas coletadas, enquanto o `asp` é encontrado em apenas 23%.
- Os formatos de documento não `html` mais usados são o `pdf`, com 42% e o `doc`, com 23%.

7. Conclusões e trabalhos futuros

Este trabalho caracterizou alguns aspectos da Web brasileira a partir de um conjunto de amostra e comparou com as estimativas de [Veloso et al., 2000]. Os resultados gerais desta amostragem se mostraram consistentes com medições completas realizadas no Chile e em Portugal.

A Web brasileira segue a tendência do uso de ferramentas com tecnologias abertas, como o `php` para gerar páginas dinâmicas. O formato para documentos `pdf` é mais utilizado que o `doc`. A influência de outros idiomas é pequena. O número de *links* quebrados é pequeno, mostrando uma certa preocupação com a qualidade das páginas.

Como trabalhos futuros pretende-se realizar uma coleta completa da Web brasileira, analisando sua evolução, e uma medição de outras características qualitativas como por exemplo o conteúdo semântico. Na linha de caracterizações de Web nacionais, outro trabalho relacionado seria uma correlação das características da Web de um país com seu poder econômico e suas características sociais.

Referências

Baeza-Yates, R. and Castillo, C. (2004). Crawling the infinite web: Five levels are enough. In *Workshop of Algorithms on Web Graphs (WAW), Springer LNCS*, pages 156–167.

Baeza-Yates, R., Lalanne, F., Castillo, C., and Dupret, G. (2004). Comparing the characteristics of the korean and the chilean web. Technical report, ITCC, DCC, University of Chile.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 309–320. North-Holland Publishing Co.

Castillo, C. (2004). *Effective Web Crawling*. PhD thesis, Department of Computer Science, University of Chile.

Economist, T. (2002). *Country Profiles*.

Efthimiadis, E. and Castillo, C. (2004). Charting the greek web. ASIST Conference (Poster), Providence, Rhode Island, USA.

Gomes, D. and Silva, M. J. (2003). A characterization of the portuguese web. *3rd ECDL Workshop on Web Archives, Trondheim, Norway*.

Lawrence, S. and Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400(6740):107–109.

Nations, U. (2003). *Human Development Report 2003*. New York: United Nations.

Nations, U. (2004). *Population Division, World Population Propects: The 2004 Revision Population Database*.

Silva, A. S., Veloso, E. A., Golgher, P. B., Ribeiro-Neto, B., and Ziviani, N. (1999). Cobweb - um coletor automático de documentos web. *Proc. XXVI Seminário Integrado de Software e Hardware(SEMISH 99), Rio de Janeiro, Brasil*, pages 233–247.

Veloso, E., Moura, E., Golgher, P., Silva, A., Almeida, R., Laender, A., Ribeiro, B., and Ziviani, N. (2000). Um retrato da web brasileira. *Anais do XXI Seminário Integrado de Hardware e Software (SEMISH 00), Curitiba, Paraná, Brasil*.

WIRE, W. I. R. E. (2004). http://www.cwr.cl/projects/wire/.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA.

Ziviani, N. (2004). *Projeto de Algoritmos: Com implementações em Pascal e C*. Pioneira Thomson Learning, 2^a edição, Belo Horizonte, MG.