Aspectos Dinâmicos dos Documentos da Web brasileira

Nahur M. Fonseca Rodolfo S. F. Resende Clarindo I. P. S. Pádua
UFMG - ICEx - DCC
Av. Antônio Carlos, 6627 CEP 31270-010
Belo Horizonte, MG - Brazil
{nahur,rodolfo,clarindo}@dcc.ufmg.br

31 de maio de 2000

Resumo

Este artigo está dividido em duas partes. Na primeira parte, discutimos a experiência do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais na cooperação com empresas. Na segunda parte, mostramos um trabalho de investigação sobre os aspectos dinâmicos da Web brasileira, que deve seus resultados à infra-estrutura que é descrita na primeira parte.

Abstract

This paper consists of two parts. In the first part, we discuss an academic departament experience in cooperating with the industry. In the second part, we present a research about the dynamic aspects of the Brazilian Web. Our research results are based on the infra-structure described in the first part.

Introdução

Este artigo é composto de duas partes. A primeira parte discute o tema Empresa e Escola sem a pretensão de discutir aspectos sociológicos e econômicos do tema. Nossa contribuição consiste em um breve relato sobre como o Departamento de Ciência da Computação (DCC) da Universidade Federal de Minas Gerais (UFMG) tem conseguido boas soluções na conciliação dos interesses nem sempre convergentes da interação entre Empresa e Escola. Nosso relato revela que esta discussão pode apresentar um interessante viés relacionado com a inserção da empresa na escola.

A segunda parte consiste em um artigo acadêmico tradicional e serve como exemplo efetivo de como que as soluções do DCC/UFMG atendem a um amplo espectro de interesses com benefício da qualidade acadêmica. Esta segunda parte apresenta um estudo sobre aspectos dinâmicos da Web brasileira. Este estudo, assim como muitos outros trabalhos de investigação científica envolvendo alunos de graduação e pós-graduação do DCC/UFMG,

só foi possível em função das excelentes soluções que o departamento tem sido capaz de encontrar ao equacionar a interação entre Empresa e Escola.

Neste artigo, utilizamos o termo *escola*, em geral, no sentido mais amplo, correspondendo a qualquer instituição cuja missão seja a formação profissional superior. Apesar disso, entendemos que o sentido mais particular, correspondendo à *universidade*, é, em muitas das vezes, mais apropriado. A missão fundamental da universidade é a formação de pessoal. O espectro de perfis na formação de pessoal diferencia as instituições de ensino. A universidade se diferencia das demais instituições em função do fato de formar pessoas com capacidade de atuação no processo de geração do conhecimento. Nossa visão portanto, é que a geração do conhecimento também faz parte da missão da universidade, mas atrelada à missão fundamental que é a formação de pessoal.

I - Primeira Parte: A Empresa e a Escola

Conforme discutido mais a frente não é incomum as empresas tomarem para si mesmas as funções das escolas com relação às suas demandas de treinamento. Pretendemos contribuir na discussão do tema Empresa e Escola chamando a atenção para a possibilidade de trazermos a Empresa para a Escola. O espectro portanto passa a ser: a Escola dentro da Empresa, a interação de Escola e Empresa independentes e a Empresa dentro da Escola.

A Escola dentro da Empresa

Não iremos explorar os aspectos relacionados com a escola dentro da empresa. Mas cabe observar que o ensino tem dimensões de formação e treinamento. As empresas que assumem funções de escola, tradicionalmente, exploram apenas a dimensão de treinamento. Esta situação leva a crer que a decisão da sociedade é não delegar para esta forma de interação o papel de atividade de formação. No entanto, nos últimos anos temos assistido a um número crescente de projetos de ensino dentro de empresas e voltados não só para treinamento mas também para formação. Deve ser reforçado que aqui não estamos descrevendo atividades de ensino conveniadas entre escola e empresa, estas atividades são descritas a seguir, contextualizadas no DCC/UFMG.

Escola e Empresa Independentes

O DCC/UFMG tem, de forma similar a vários departamentos de universidades de todo o mundo, interagido com várias empresas através de atividades de formação e treinamento. O DCC/UFMG desde os anos setenta, tem interagido com empresas, não só através de atividades de treinamento e formação mas também através de atividades de consultoria. A atividade de consultoria tem sido uma forma tradicional de interação entre Empresa e Escola e funciona não só como meio de suplementar recursos do professor e da escola, mas também como agente de associação das demandas sociais e do ensino. As atividades de consultoria do DCC/UFMG apresentam um perfil tradicional e não têm características especiais a serem ressaltadas.

O DCC/UFMG participou de novas formas de interação através da execução de projetos de interesse de empresas. Isto aconteceu de forma incipiente nos anos oitenta e de forma bastante mais organizada nos anos noventa. No caso do DCC/UFMG devem ser destacados os projetos acertados com a TELEMIG, hoje sucedida pela TELEMAR. A execução de tais projetos tem um carater inovador na medida em que, via de regra, são projetos envolvendo alta tecnologia e contribuem em uma escala maior que as consultorias, na associação das demandas sociais e do ensino. Além disso, agregam também a possibilidade de atividades de pessoal formado e em formação. A execução de projetos dentro do DCC/UFMG permite ainda hoje que o departamento possa formar excelentes alunos e expor os professores a relevantes demandas apresentadas pelos diversos setores da sociedade.

A característica comum dos projetos não só com a TELEMAR, mas também com a Engetron, Batik sucedida pela Lucent, Bosh, COPASA, MPAS, PRODABEL e tantos outros é a geração de possibilidades nas outras áreas de atuação do departamento. Ou seja, atividades de extensão induzindo possibilidades para as atividades de ensino e pesquisa. Nestas cooperações, sempre houve espaço para trabalhos de final de curso de graduação, dissertações de mestrado e até mesmo teses de doutorado.

Apesar das várias direções que têm sido colocadas em termos da problemática da cooperação entre Empresa e Escola, temos convicção de que todas as demandas usuais e mais ainda, a chamada *educação continuada*, irão colocar pressões ainda maiores para que este tipo de cooperação se fortaleça.

A Empresa dentro da Escola

O DCC/UFMG se integrou à Rede Nacional de Pesquisa (RNP) desde o início dos anos noventa. Tratava-se, então, de um projeto acadêmico tradicional. Em 1996, o Ministério da Ciência e Tecnologia implementou uma política de incentivo ao desenvolvimento da Internet no Brasil. Essa política abriu a RNP à atividade comercial, permitindo a venda de acesso a empresas e provedores de acesso. Foi aberta a possibilidade de que os pontos de acesso, até então somente acadêmicos, pudessem também operar comercialmente como provedores de acesso. Ocorreu uma grande inovação com a inauguração, dentro do DCC/UFMG, de uma entidade para prestação de serviços relacionados com a Internet, em particular serviços de provimento de acesso. Esta entidade, denominada POP-MG [10] não tem os aspectos usuais de outras atividades que têm início e fim bem determinados.

A operação comercial do POP-MG exigiu um grande grau de profissionalização que não é usualmente encontrado em universidades. Esta profissionalização se desdobrou em uma série de benefícios para o DCC/UFMG como um todo. Além de ser auto-suficiente em termos financeiros, o POP-MG gera recursos para o DCC e para a universidade. A equipe do POP-MG também auxilia na gerência da rede interna do departamento e cuida do acesso da rede da UFMG à Internet. Outro importante desdobramento para o POP-MG foi a criação da Rede Internet Minas. A Rede Internet Minas é uma rede acadêmica cuja espinha dorsal (backbone) encontra-se distribuída no Estado de Minas Gerais e conecta-se à Internet através do POP-MG.

Os alunos de graduação e pós-graduação podem complementar sua formação em atividades do POP-MG. Um destes ramos de atividades compreende tecnologias relacionadas

com *caching*.

As novas experimentações de interação do DCC/UFMG com empresas caracterizam-se pela criação de empresas de alta-tecnologia e todas elas têm se beneficiado da infra-estrutura criada pelo POP-MG.

O trabalho descrito na segunda parte deste artigo teve apoio do projeto SIAM. O SIAM é um projeto financiado pelo programa PRONEX do Ministério da Ciência e Tecnologia. Uma boa parte das atividades do projeto SIAM se beneficia da infra-estrutura proporcionada pelo POP-MG. O trabalho descrito a seguir ilustra a importância da conciliação das atividades de interação Empresa e Escola com a missão de formação de pessoal e geração de conhecimento.

II - Segunda Parte: Estudo da Web Brasileira

Nesta segunda parte, descrevemos como a infra-estrutura do POP-MG e do projeto SIAM foi usada em uma investigação sobre os aspectos dinâmicos dos documentos da Web brasileira.

A World Wide Web, ou simplesmente Web, é um sistema distribuído que utiliza o paradigma cliente/servidor para acessar documentos (hipertexto) e cuja rede de comunicação é a Internet [21]. A Web é também um sistema dinâmico, em que documentos são inseridos, modificados e removidos.

O tamanho da Web continua a crescer rapidamente, segundo o Inktomi [9], o número de documentos da Web, em janeiro de 2000, ultrapassou 1 bilhão de documentos. Portanto, é inviável pesquisar esses documentos manualmente. Ao invés disso, existem máquinas de busca para a Web que lidam com esse problema. AltaVista [1], Google [7] e TodoBR [12] são alguns exemplos.

Existem diversas implementações de máquinas de buscas. Uma primeira divisão que é feita considera distintos os diretórios e as máquinas de busca propriamente ditas [11]. Nos diretórios, uma equipe se encarrega de categorizar os documentos em tópicos de interesse, e os usuário podem pesquisar manualmente sua estrutura, ou pesquisá-la por palavra-chave. O Yahoo [15] e o Cadê [3] ilustram esse tipo de ferramenta. Nas máquinas de busca propriamente ditas, as consultas dos usuários são respondidas baseado em uma referência estática de toda ou parte da Web. Uma descrição detalhada de um sistema desse tipo foi feita por Brin e Page [17].

Atualmente, existem outras variações de máquinas de busca. Máquinas de busca especializadas limitam o escopo de assuntos das consultas que o sistema se propõe a responder, por exemplo, assuntos sobre medicina, direito ou mercado financeiro [8, 5, 13]. Existem máquinas de busca que usam técnicas de inteligência artificial para tratar as consultas dos usuários [2]. Por fim, máquinas de busca distribuídas colocam o servidor de arquivos e a interface de busca nas máquinas dos clientes espalhados pelo mundo, criando uma rede para compartilhar informação [6].

Nós iremos nos ater às máquinas de busca propriamente ditas, como descrito por Brin e Page. Nesse tipo de sistema, robôs [14] se encarregam de manter a referência estática da Web sobre a qual são feitas as consultas do usuário. Entretanto, a Web é um sistema dinâmico, criando, portanto, uma inconsistência entre a referência estática e o estado da

Web.

Nós mostramos que o estudo dos aspectos dinâmicos da Web, sejam a inserção, remoção e modificação de documentos, do ponto de vista das máquinas de busca, influencia as decisões de qualidade na construção desse tipo de ferramenta.

Este estudo discute nossos resultados preliminares sobre a análise da Web brasileira, baseado no trabalho de Douglis e outros [19]. Nós focamos em como melhorar a qualidade de serviço das máquinas de busca. Nós também descrevemos a implementação de ferramentas e o arcabouço que foram necessários para suportar esse trabalho.

1 Motivação

Este trabalho está inserido no âmbito do projeto SIAM - Sistemas de Informação em Ambientes Móveis. Este projeto envolve a integração de duas áreas: sistemas avançados de recuperação de informação e computação móvel.

Na área de sistemas de recuperação de informação, um dos projetos do SIAM é a implementação de uma máquina de busca para a Web brasileira, o TodoBR, que já está disponível via Web[12].

As publicações sobre as máquinas de busca em geral descrevem poucos detalhes em consequência das políticas de sigilo adotadas pelos seus autores. Os estudos sobre máquinas de busca podem revelar segredos comerciais envolvendo importantes vantagens competitivas. A criação de uma máquina de busca operacional desenvolvida dentro do DCC/UFMG permite que sejam feitos diversos estudos visando a divulgação científica.

A posição privilegiada do DCC/UFMG, sob o aspecto de prestação de serviços relacionados ao acesso à Internet, graças ao POP-MG, viabilizou a realização de nosso estudo, principalmente no que concerne à seleção e à coleta dos documentos que foram utilizados neste trabalho.

2 Seleção de Documentos da Web Brasileira

De acordo com estatísticas da máquina de busca TodoBR, a Web brasileira possuía, em dezembro de 1999, cerca de 6 milhões de documentos. Esse número a torna a maior parte da WWW na América do Sul. Nós escolhemos estudar a Web brasileira principalmente devido à infra-estrutura disponível. Conforme discutido na primeira parte deste artigo um aspecto importante deste trabalho foi poder contar com a infra-estrutura do POP-MG e do projeto SIAM.

Neste trabalho, um documento HTML foi considerado como sendo da Web brasileira se o domínio de sua URL terminava em *.br*, independente de sua língua ou conteúdo.

Nós separamos os documentos em duas categorias, MA e AL.

A categoria MA (Mais Acessados) corresponde aos documentos mais acessados segundo o servidor *proxy-cache* do POP-MG, que intercepta grande parte dos acessos aos documentos da Web brasileira. Nós ordenamos os documentos pelo número de acessos. A linha de corte foi determinada no primeiro documento em que o número de acessos fosse apenas um. Usando esse processo, selecionamos 1258 documentos. A Figura 1 mostra os documentos

Figura 1: Documentos da Categoria MA por Sub-domínio

Figura 2: Documentos da Categoria AL por Sub-domínio

dessa categoria separados por sub-domínio. O sub-domínio *.com* compreende o maior número de documentos, seguido pelos sub-domínios *.gov* e *.br*.

A categoria AL (Aleatórios) corresponde aos documentos escolhidos aleatoriamente de uma população de mais de 2 milhões de documentos fornecida pelo TodoBR. Nós fixamos um número em torno de cinco mil documentos baseado no espaço de armazenamento disponível e no tamanho médio estimado dos documentos. A Figura 2 mostra os documentos dessa categoria separados por sub-domínio. Novamente, há a predominância do sub-domínio *.com*, seguido pelos sub-domínios *.br*, *.gov* e *.org*.

3 Ferramentas

Para dar suporte à nossa análise, nós utilizamos algumas ferramentas já disponíveis e desenvolvemos outras que são descritas a seguir.

3.1 Coletor de Documentos

Nós utilizamos o *wget* para Unix para coletar os documentos Web. Nós usamos os seus arquivos de *log* para coletar informação dos cabeçalhos HTTP das requisições [22], por exemplo, o código de status retornado pelo servidor. O coletor de documentos é um *script* Perl que dispara vários processos do *wget* diariamente para coletar os dois conjuntos de documentos, MA e AL.

3.2 HtmlDiff

Esta ferramenta, disponível via WWW[4], aprimora as característica do *diff* do Unix, fazendo comparações palavra por palavra e marcando com cores as diferenças entre dois documentos HTML. A saída do HtmlDiff é uma página HTML com uma sintaxe visual clara para destacar o que foi removido da página e o que foi acrescentado.

```
          Total of words and bytes per level:

         L0    L1    L2    L3
WORD    196   130   173    11
BYTE   1845   696  3211   969

       SAME    DELETED   CHANGED  APPENDED  UPDATED
    WD%  BT%  WD% BT%   WD%  BT%  WD% BT%   WD%  BT%
L0 81.1 77.5  0.0 0.0   4.1  7.1  0.0 0.0  14.8 15.4
L1 63.8 68.7  0.0 0.0  21.5 16.1  0.0 0.0  14.6 15.2
L2 96.5 95.2  1.2 0.3   1.2  1.7  0.0 0.0   1.2  2.8
L3 27.3 20.2  0.0 0.0  36.4 39.0  0.0 0.0  36.4 40.8
Tt 80.8 76.8  0.4 0.1   8.2 10.1  0.0 0.0  10.6 13.0
```

Figura 3: Exemplo de Saída do HTMLDiff-modificado

Nós não usamos a versão original dessa ferramenta. Ao invés disso, nós criamos uma nova versão a qual chamamos de HtmlDiff-modificado.

3.3 HTMLDiff-modificado

O HtmlDiff-modificado foi feito para receber duas páginas HTML como entrada e fazer a comparação das duas, palavra por palavra, classificando as diferenças em níveis e resumindo os resultados no formato mostrado na Figura 3.

As modificações podem ocorrer em quatro níveis. As tags HTML que não possuem informação (por exemplo, <HTML> e <HEAD>) compõe o nível zero (L0). As tags HTML que podem conter informação (por exemplo, <P>, <TD> e <LI>) compõe o nível um (L1). O nível dois (L2) é composto pelos tags que representam *links*. Finalmente, no nível três (L3), estão as imagens e outros objetos.

As diferenças são contadas usando duas métricas, o percentual do tamanho em bytes (%BT), e o percentual do total de palavras (%WD), de acordo com o conceito de palavra usado no HtmlDiff.

A tabela de saída possui cinco colunas, SAME, DELETED, CHANGED, APPENDED e UPDATED. Elas têm significado semelhante à sintaxe usada pelo comando *diff* do Unix. A coluna SAME refere às palavras que foram mantidas iguais nos dois documentos. A coluna DELETED refere às palavras que foram removidas do documento 1. A coluna APPENDED refere às palavras que foram adicionadas ao documento 2. As colunas CHANGED e UPDATED referem às palavras que foram modificadas da forma no documento 1 para a nova forma no documento 2, respectivamente.

3.4 HtmlDiff-stats

O HtmlDiff-stats é um *script* Perl que deve ser utilizado em conjunto com o HtmlDiff-modificado. O HtmlDiff-stats executa o HtmlDiff-modificado para todos os documentos coletados em um dia, comparando-os com os documentos da coleta imediatamente anterior.

A síntese dos resultados são mostrados em um quadro como o apresentado acima.

Outras informações apuradas pelo HtmlDiff-stat são o número de páginas coletadas no dia, o número de páginas não coletadas, o tamanho médio das páginas e a lista das páginas que foram modificadas no dia por nível.

3.5 Exibidor Gráfico

Nós criamos um sítio [23] para visualizar os resultados desse estudo. O sítio usa as tecnologias *CGI* e *Javascript* para gerar os gráficos de resultados.

É possível ver uma caracterização mais detalhada sobre os documentos Web selecionados, seus domínios, bem como seus servidores. Também é possível ver a frequência de modificações para diferentes períodos. O quadro do HtmlDiff-stats foi transformado em um gráfico de barras.

4 Resultados

Nós coletamos diariamente todos os documentos das categorias MA e AL durante 54 e 42 dias respectivamente. Aqui, apresentamos alguns de nossos resultados sobre os aspectos dinâmicos dos documentos da Web brasileira. Nós indicamos como melhorar a qualidade de serviço das máquinas de busca usando a análise dos aspectos dinâmicos da Web.

4.1 Frequência de Modificação

Esta classe de gráficos mostra quantas versões dos documentos HTML são criadas ao longo de um período. Uma nova versão é criada sempre que um documento sofre alguma modificação. Os gráficos estão separados pelas categorias MA ou AL. O eixo das abscissas dá o número de versões enquanto o eixo das ordenadas, o número de documentos que tiveram um dado número de versões.

Os gráficos mostram que muitos documentos são pouco modificados, muitos são modificados frequentemente e poucos estão nos níveis intermediários. Além disso, esse comportamento dos documentos da categoria MA e os da categoria AL é semelhante (Figuras 4 e 5); e o período, uma semana ou um mês, não altera a forma em U do gráfico de frequência de modificações dos documentos (Figuras 6 e 7).

Um trabalho relacionado com a observação acima, realizado por Coffman, Jr., Liu e Weber [20], propõe uma política ótima para escalonamento de robôs de máquinas de busca, baseado na frequência de modificação dos documentos. Segundo demostrações e algumas suposições daquele trabalho, com as informações dos gráficos acima seriam criados duas classes de documentos para coleta. Uma classe composta pelos documentos muito frequentemente modificados e outra composta pelos documentos raramente modificados. Os acessos aos documentos devem ser o mais igualmente espaçados possível, e a taxa de remoção e inserção de documentos devem ser mantidos baixos entre cada período de coleta.

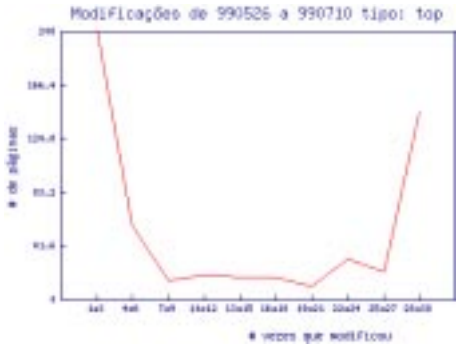

Figura 4: Frequência de Modificação Categoria MA em 1 Mês

Figura 5: Frequência de Modificação Categoria AL em 1 Mês

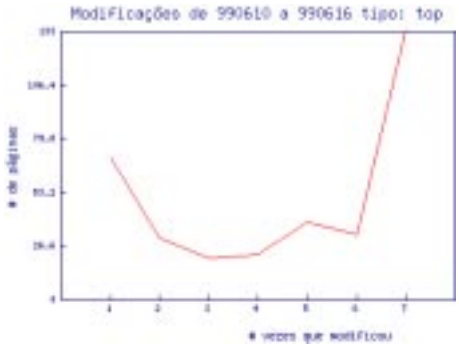

Figura 6: Frequência de Modificação Categoria MA 1 Semana

Figura 7: Frequência de Modificação Categoria AL 1 Semana

Figura 8: Documentos Categoria MA Modificados em 1 Mês

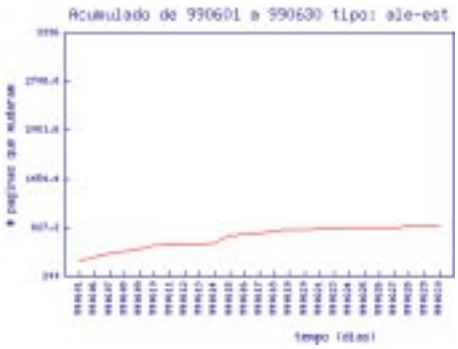

Figura 9: Documentos Categoria AL Modificados em 1 Mês

4.2 Documentos Mais Modificados

As Figuras 8 e 9 mostram que os documentos da categoria MA são mais modificados do que os documentos da categoria AL. Essa classe de gráficos mostra que, em um período de um mês, o número de documentos que são modificados ao menos uma vez cresce quase linearmente com o tempo. Para os documentos da categoria MA, 40% deles são modificados até o final do período, enquanto apenas 25% dos documentos da categoria AL foram modificados no mesmo período. Ao todo, 30% dos documentos foram modificados.

No trabalho de Douglis e outros [19], os documentos mais acessados são também os mais frequentemente modificados, o que também foi observado em nossa análise. A diferença entre o trabalho dele e o nosso é que a coleta dos documentos naquele estudo foi feita apenas via *Proxy*, portanto, os documentos observados tinha sido acessado pelo menos uma vez pela comunidade de usuários. No nosso trabalho, nós separamos os documentos em duas categorias, MA e AL, e fizemos a coleta dos documentos diretamente dos servidores.

As máquinas de busca podem melhorar a ordem de relevância dos documentos retornados a uma consulta a partir da observação do fato de que os documentos mais frequentemente acessados são também os mais acessados, e provavelmente, os mais relevantes. Tal informação deve ser imbutida nos processos de construção dos índices e processamento da consulta das máquinas de busca. Ele deve também ser combinado com outros modelos para determinar a relevância dos documentos, como o modelo vetorial [16], ou uma de suas variações usadas nas máquinas de busca correntes [18].

4.3 Categorização das Modificações

O gráfico da Figura 10 mostra em que parte dos documentos ocorrem as modificações para a categoria AL. Os documentos foram subdivididos entre estáticos, que referem a documentos HTML estáticos, e dinâmicos que são documentos HTML gerados dinamicamente pelo servidor (p.e. um *script* CGI).

Esse gráfico foi obtido do quadro estatístico da ferramenta HTMLDiff-modificado. Os quatro níveis explicados na Seção 3.2 são representados por números (de 0 a 3) e um nível foi adicionado para totalizar as modificações (representado pela letra T). Os cinco percentuais de modificação por nível foram transformados em barras, e uma nova barra foi

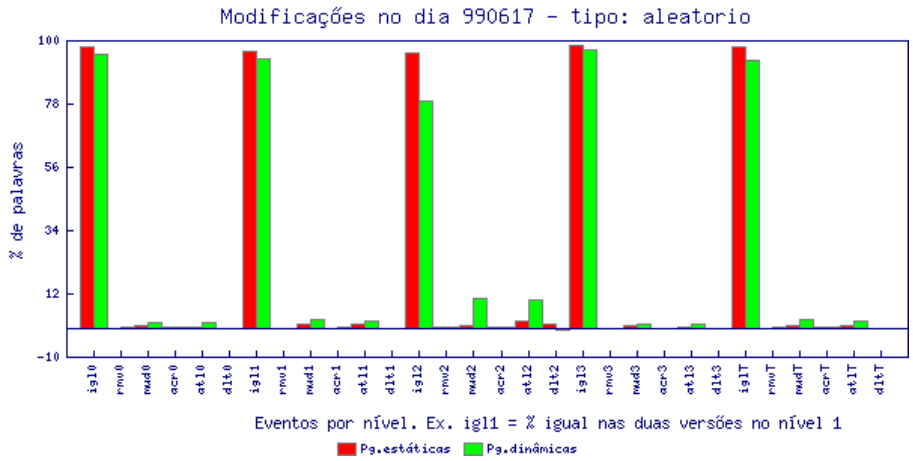

Figura 10: Categorização das Mudanças

acrescentada (dlt) para mostrar a diferença resultante. Se esse valor for maior que zero, o documento aumentou. Se for menor que zero, então ele diminuiu.

1. igl =3D conteúdo igual à versão anterior
2. rmv =3D conteúdo removido da versão anterior
3. mud =3D conteúdo foi atualizado na versão nova
4. acr =3D conteúdo acrescentado à versão nova
5. alt =3D conteúdo novo que substituiu o conteúdo em *mud*
6. dlt =3D alt - mud

A figura acima serve para mostrar que os *links* (nível 2) representam a parte dos documentos mais sujeitos a modificações. Uma máquina de busca que considere essa informação nos seus processos deve, portanto, adotar uma política de atualização mais frequente. Outra análise feita sobre o gráfico acima é que a variação do tamanho dos documentos estáticos (coluna dlt) manteve-se praticamente nulo.

Conclusão

Neste artigo, apresentamos em uma primeira parte um relato sobre a evolução do DCC/UFMG com relação à questão da cooperação Empresa e Escola. Mostramos que esta questão envolve, além dos aspectos usuais de cooperação entre escola e empresas independentes, a possibilidade de uma interação ainda mais forte tanto de um lado quanto do outro. Especificamente, fazemos o relato da criação do POP-MG, uma experiência inovadora de relacionamento de um departamento acadêmico com empresas.

Na segunda parte deste artigo, mostramos como que a infra-estrutura descrita na primeira parte deu subsídios a uma investigação sobre aspectos dinâmicos da Web brasileira.

Nós usamos documentos HTML coletados diariamente de maio de 1999 até junho de 1999. Foram desenvolvidas algumas ferramentas para dar suporte a este estudo. Um resultado particularmente interessante é a relação de auto-similaridade no comportamento da frequência de mudança nos documentos da Web brasileira, além de resultados mais voltados para máquinas de busca [24]. Existem poucas publicações sobre os aspectos discutidos, em particular o problema da caracterização da inserção de documentos na WWW continua sendo um desafio.

Referências

[1] Alta Vista Search Engine. http://www.altavista.com.

[2] Autonomy Home Site. http://www.autonomy.com.

[3] Cade Search Engine. http://www.cade.com.br.

[4] CPAN. http://www.perl.com/cpan.

[5] FindLaw Search Engine. http://www.findlaw.com.

[6] Gnutella Home Site. http://gnutella.wego.com.

[7] Google Search Engine. http://www.google.com.

[8] HealthFinder Search Engine. http://www.healthfinder.gov.

[9] Inktomi Home Site. http://www.inktomi.com/.

[10] Ponto de Presença Internet em Minas Gerais. http://www.pop-mg.com.br.

[11] Search Engine Watch. http://www.searchenginewatch.com/.

[12] TodoBR Search Engine. http://www.todobr.com.br.

[13] TradingDay Search Engine. http://www.tradingday.com.

[14] WebCrawler site. http://info.webcrawler.com/mak/projects/robots/robos.html.

[15] Yahoo Search Engine. http://www.yahoo.com.

[16] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman, 1999.

[17] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of WWW7 International Conference*, 1998.

[18] Andrei Broder and Monika Henzinger. Information-Retrival on the Web. Tools and algorithms issues. apresentado em Compaq Systems Research Center, 1998.

[19] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of Change and other Metrics: a Live Study of the World Wide Web. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, 1997.

[20] Jr. E. G. Coffman, Zhen Liu, and Richard R. Weber. Optimal Robot Scheduling for Web Search Engines. Technical Report 3317, INRIA, Dezembro 1997.

[21] Tim Berners-Lee et al. The World-Wide Web. *Communications of the ACM*, 32:76–82, 1994.

[22] R. Fielding, J. Gettys, J. Mogul, and L. Berners et al. RFC 2068: HyperText Transfer Protocol — HTTP/1. 1, 1997. http://info.internet.isi.edu:80/in-notes/rfc/files.

[23] N. Fonseca. Aspectos Dinâmicos dos Documentos da Web Brasileira. http://www.lbd.dcc.ufmg.br/∼nahur/WSS3/index.html.

[24] N. Fonseca, R. Resende, and W. Meira Jr. Dynamic Aspects of Documents of the Brazilian Web. In *Proceedings of the 1st Internation Conference on Web Information Systems Engineering (WISE 2000)*, June 2000.