Classificação Semântica Automática de Documentos da WWW

Ana Paula Ribeiro
Rodrigo Fonseca
Wagner Meira Jr.
Virgílio Almeida

{anapaula, rfonseca, meira, virgilio}@dcc.ufmg.br
Departamento de Ciência da Computação
Instituto de Ciências Exatas, Sala 3004
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 CEP 31270-010, Belo Horizonte, MG

Resumo

A cada dia que passa a Internet vem sendo acessada por uma parcela cada vez maior da população. Esse fato vem reforçar a observação de que fatores sócio-culturais e geográficos influenciam fortemente o tráfego de documentos na World Wide Web. Desta forma, a caracterização dos acessos à rede e uma melhor compreensão do comportamento dos usuários são ferramentas essenciais para quaisquer tarefas de planejamento e pesquisa relacionadas à WWW. Um componente importante dessa caracterização é a classificação semântica os documentos acessados, ou seja, a divisão dos acessos em categorias, cada uma associada a um assunto de interesse dos usuários. Este trabalho apresenta uma proposta para a classificação semântica automática de páginas da World Wide Web. Esta proposta foi aplicada a logs do POP-MG (Ponto de Presença Internet em Minas Gerais) resultando em um universo de páginas classificadas correspondente a até 65% das páginas constantes dos logs.

Palavras Chave: **WWW, Internet, Classificação Semântica, Comportamento dos Usuários.**

1. INTRODUÇÃO

O crescimento e popularização da World Wide Web (WWW) tem tornado cada vez mais patente a influência de fatores sócio-culturais e econômicos no comportamento dos usuários no que se refere ao acesso à rede. Esta influência se manifesta principalmente com relação às preferências por assuntos inerentes às várias páginas.

A variação das preferências dos usuários a documentos na WWW é extremamente dinâmica, tendo em vista diversos fatores: (i) efeitos transientes: acontecimentos momentâneos que alteram o padrão de comportamento dos usuários, como por exemplo eleições e Copa do Mundo; (ii) renovação: páginas são criadas a uma taxa muito alta e observa-se um efeito migratório com páginas novas sendo muito acessadas e a frequência de acessos diminuindo ao longo do tempo; (iii) mudança de interesses: o próprio usuário tende a sempre procurar por novas fontes de informação, embora elas possam ser sobre o mesmo tema. Todo este dinamismo representa um grande problema para qualquer atividade de planejamento envolvendo a WWW. Assim, o dimensionamento de servidores e meios de interconexão têm de considerar a ocorrência de surtos de acessos e o comportamento dinâmico dos usuários. Da mesma forma, entidades comerciais que desejem utilizar a WWW para fins lucrativos necessitam de informações a respeito dos interesses e hábitos dos usuários, de forma a poder explorá-los mais eficientemente.

A execução dessas atividades de planejamento demanda, portanto, um acompanhamento criterioso dos padrões de acesso das várias classes usuários. No nosso caso, um padrão de acesso é uma distribuição de probabilidades de acessos para os vários assuntos de interesse. Essas probabilidades são atribuídas a classes que compõem uma taxonomia dos interesses dos usuários. Há várias formas de se determinar padrões de acesso, as quais se diferenciam pelo nível de exigência do usuário final e de privacidade. Pesquisas de caráter público, tais como a do Cadê?/Ibope[4], são um exemplo de mecanismo de obtenção de padrões de acesso no qual a privacidade do usuário é total (tendo em conta o caráter voluntário da pesquisa), mas há um nível de exigência alto, uma vez que o usuário é que deve ter a iniciativa de responder à pesquisa. Uma segunda abordagem é a instrumentação do *browser* utilizado para acessar a WWW, que passa a registrar todos os acessos feitos pelo usuário. Essa estratégia sem dúvida provê informações

com maior nível de detalhe, mas com pouca privacidade para o usuário. Além disso, o usuário pode se comportar de forma diferenciada, ciente de que o seu comportamento está sendo observado. A terceira opção é a análise do Log de um servidor Proxy, que se tornou viável com a popularização de proxy caches. Um servidor Proxy WWW é um agente intermediário entre o cliente (usuário final) usando um browser e o servidor Web. Ele recebe as requisições dos clientes e repassa-as aos servidores de fato. Cada resposta recebida é enviada de forma transparente ao cliente que originou a requisição. Por centralizar vários pedidos de conexão, o servidor Proxy WWW se torna um ótimo ponto de armazenamento da informação que trafega entre os clientes e a Internet. Desta forma, pode-se fazer "caching" desta informação nos proxies através de servidores Proxy Cache WWW (ou simplesmente Proxy). Através de arquivos de log, que registram todos os acessos a um servidor proxy, sabemos quais documentos o grupo de usuários servidos por ele acessou. O custo de monitoração e determinação de padrões de acesso tanto para o usuário quanto para o servidor é inexistente, pois o software de cache já guarda o log naturalmente. Estudos recentes, como [3, 6] utilizam-se desta estratégia de coleta.

Neste trabalho, apresentamos uma proposta de metodologia para classificação semântica automática, ou seja, uma categorização por assuntos, de páginas constantes do log de um servidor proxy. É importante que tal tarefa seja automática por dois motivos principais: (i) uma classificação manual, na qual uma pessoa visita cada página do log, não é prática, pois demanda muito tempo e está muito sujeita a erros; e (ii) à medida em que uma pessoa vai adquirindo intimidade com um assunto, seus critérios de categorização vão mudando, ou seja, não se teriam classificações coerentes ao longo do tempo.

Também é importante que a amostra de dados seja significativa, para que os resultados sejam relevantes e amplamente aplicáveis. Se utilizarmos o log da UFMG, por exemplo, teremos uma taxonomia dos acessos da comunidade acadêmica, que certamente será diferente da taxonomia dos acessos de uma empresa.

Este artigo está organizado em 5 seções. Na próxima seção apresentamos a estratégia utilizada na classificação, na seção 3 apresentamos detalhes da implementação do algoritmo, na seção 4 descrevemos os resultados e na seção 5 apresentamos as conclusões e direções futuras do trabalho.

2. ESTRATÉGIA

Apresentamos nesta seção a estratégia utilizada para a classificação semântica automática de páginas da World Wide Web. Essa estratégia possui dois aspectos principais a serem considerados: a utilização de dados relevantes e o método utilizado na classificação.

2.1. Relevância dos Dados

Para desenvolvimento e avaliação dos algoritmos, utilizamos arquivos de log do software de cache Squid [2], retirados do servidor proxy do Ponto de Presença da Internet em Minas Gerais, POP-MG [7]. O POP-MG gerencia um dos maiores backbones do estado de Minas Gerais, sendo 70% dos provedores de acesso do estado seus clientes. É um dos maiores pontos de presença da RNP no Brasil, tendo um volume de tráfego próximo a 6 Mbps. Seu servidor de cache recebe aproximadamente 1,800,000 requisições diárias.

Nos arquivos de log do Squid são registrados os URLs de todos os documentos requisitados ao servidor, juntamente com outros dados como tamanho do documento e hora da requisição. Esses logs foram "filtrados" até que contivessem apenas URLs únicos, e foi associado, a cada uma deles, o número total de requisições por usuários. Os URLs que se referiam a imagens também foram excluídos para efeitos de classificação, assumindo-se que tais figuras têm a mesma semântica das páginas onde elas se encontram.

2.2. Método de Classificação

O primeiro método considerado foi pesquisar a procura de palavras significativas nos textos contidos em cada página da Web acessada (i.e., constante no log). Uma palavra chave, ou um grupo delas, poderia permitir a classificação de determinada página em uma categoria pré-definida, ou seja, dar àquela página uma classificação semântica.

Para que pudéssemos iniciar o trabalho, fizemos uma avaliação das 60 páginas mais acessadas de um log, o que corresponde a 14% do total de acessos. Estas páginas foram analisadas manualmente e observamos que apenas 20% delas continham textos passíveis de pesquisa e categorização. Nestes textos, encontramos algumas palavras-chave que nos informam sobre o tipo de assunto a que a página se refere. Entretanto, essas palavras normalmente

constavam, também, do endereço URL associado a página. Em 45% das páginas, havia apenas algumas palavras, todas relacionadas a imagens que eram parte dessas páginas. Mais uma vez, estas palavras quase sempre constavam, também, nos respectivos URLs. Os 35% restantes das páginas não puderam ser classificadas pela estratégia descrita em virtude de diversos motivos.

Tendo em vista estas observações, concluímos que seria válido, e mais simples, a utilização dos próprios URLs para pesquisa de palavras significativas e consequente classificação das páginas em categorias. Desenvolvemos um programa, descrito na próxima seção, que foi utilizado no log inicial e em outros três logs para avaliação da aplicabilidade do algoritmo.

3. IMPLEMENTAÇÃO

Nesta seção, descrevemos o algoritmo, assim como detalhes do programa utilizado para classificação semântica das páginas da Web. O programa recebe dois arquivos de entrada:

Dicionário :Contém uma lista de palavras agrupadas em categorias semânticas.

URLs :Contém uma lista de URLs. A cada URL é associado o número de vezes que ele foi acessado pelos usuários, em um determinado espaço de tempo.

O programa inicia com a montagem do dicionário a partir do arquivo correspondente. O dicionário é armazenado em uma tabela *hash*, que associa cada palavra à sua respectiva categoria.

A seguir, os URLs são lidos um a um do arquivo correspondente. O programa verifica, para cada URL lido, a ocorrência das palavras do dicionário. Quando alguma palavra é encontrada no URL, a página correspondente é classificada de acordo com a categoria a qual a palavra pertence. O número de acessos feitos pelos usuários a essa página é então somado ao total de acessos a páginas daquela categoria.

Entretanto, quando da utilização desta estratégia, podem ocorrer dois problemas: (i) nenhuma das palavras do dicionário é encontrada em determinado URL ou (ii) duas ou mais palavras do dicionário, associadas a categorias diferentes, são encontradas num mesmo URL (colisão).

No primeiro caso, onde nenhuma das palavras é encontrada no URL, o programa não tem como definir uma categoria, sendo a página considerada como “não classificada”. O URL é gravado em um arquivo separado, onde estarão, ao final da execução do programa, todos os URLs não classificados. Este arquivo poderá servir para posterior avaliação e melhoria do dicionário.

Para o segundo problema, avaliamos duas hipóteses para solução. A primeira consistia em estabelecer uma ordem de prioridades das categorias, onde uma categoria prevaleceria sobre outra. Esta hipótese não se mostrou muito eficaz pois não conseguimos chegar a uma ordem de prioridades única para todos URLs, uma vez que a ordem das prioridades das categorias teria de ser alterada dependendo da natureza da página. Além disso, como haviam categorias com o mesmo nível de prioridade, colisões eram frequentes.

A segunda hipótese, que foi adotada como solução, faz uso de uma noção intuitiva: é notório o fato de que, quando se caminha por um URL, sua especificidade aumenta, ou seja, palavras encontradas no final de um URL podem classificar mais especificamente determinada página, se forem comparadas a palavras encontradas no começo do URL. Esta observação foi utilizada para solucionar o problema das colisões.

Assim, o programa procura, em cada URL, cada uma das palavras do dicionário. A palavra encontrada e sua posição relativa são, então, armazenadas. Estas informações somente serão atualizadas se alguma outra palavra do dicionário, com uma posição relativa maior, ou seja, mais próxima do fim do URL, for encontrada. Ao final da pesquisa em determinado URL, estará armazenada a palavra, contida no dicionário, que mais se aproxima do final do URL. Esta palavra é utilizada para conferir à página sua categoria.

3.1. Dicionário

O dicionário é criado incrementalmente a partir de logs de acesso da comunidade alvo. De forma a maximizar a cobertura das palavras do dicionário, fez-se um *ranking* de todas as palavras existentes nesse log e uma avaliação das palavras que poderiam servir para caracterizar o URL que as continha, como “playboy”, “franca98” e “jimihendrix”. Excluimos, obviamente, palavras sem relevância para a caracterização como “com”, “br”, “www”. As categorias, nas quais as palavras selecionadas foram agrupadas, foram sendo criadas à medida que palavras com novas semânticas eram encontradas, tendo como referência categorias existentes em diversos sites de busca.

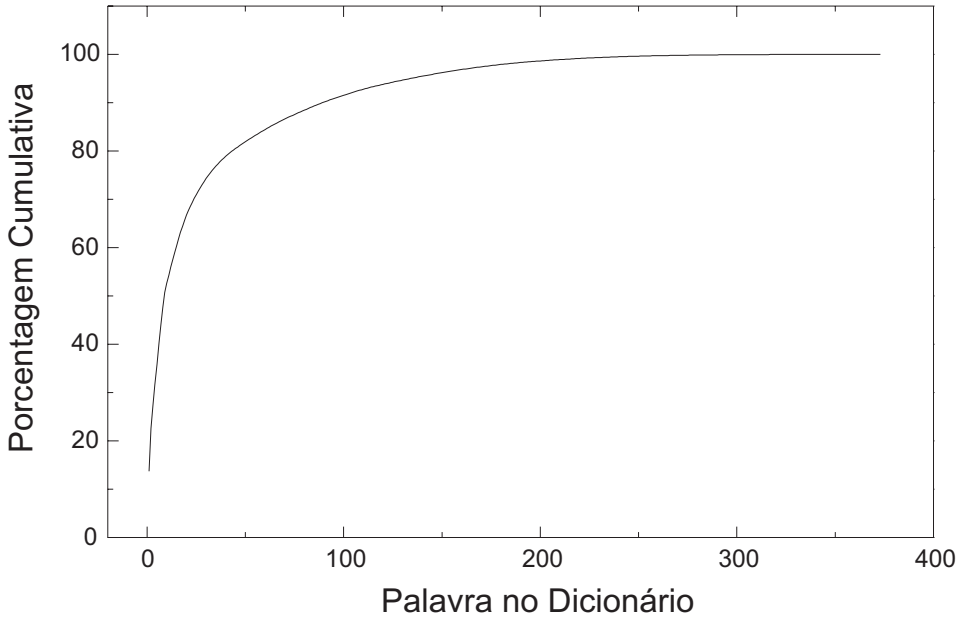


Figura 1: Ranking das Palavras do Dicionário

Após criado, analisamos a cobertura do dicionário em relação à classificação de URLs. O gráfico 1 ilustra esta análise. O dicionário consiste de 373 palavras. Somando-se os acessos correspondentes a cada palavra, conforme descrito acima, obtivemos aproximadamente 1,440,000 acessos *às palavras*. No gráfico, mostramos o crescimento do número de acessos a palavras do dicionário à medida que cada uma das palavras vai sendo incluída, em relação ao total obtido com o dicionário completo. Podemos observar que até aproximadademte a 200ª palavra mais frequente, há um aumento significativo da porcentagem do número de acessos e neste ponto torna-se praticamente constante. Isso é um indicativo de que não há um aumento significativo do número de URLs classificados com o acréscimo de palavras no dicionário a partir de um determinado ponto, servindo, assim, como um bom parâmetro para controlar o tamanho do mesmo.

Na tabela 1, apresentamos, a título de exemplo, a descrição das categorias do dicionário utilizado para os nossos experimentos.

4. RESULTADOS

Para avaliação da eficiência do algoritmo proposto, utilizamos o programa descrito na seção 3 em quatro diferentes conjuntos de URLs, sendo três retirados do POP-MG, em datas diferentes, e um retirado do servidor de cache do provedor português Esoterica. Um mesmo arquivo de entrada, dicionário, criado a partir de um dos logs brasileiros (a que denominamos *log base*), foi utilizado para classificar os outros três logs. A classificação dos logs brasileiros objetiva avaliar o impacto do fator tempo sobre a validade do dicionário, uma vez que há uma diferença de 4 meses entre o log base, usado para construir o dicionário, e os outros dois. Já o log do servidor português permitiu uma avaliação do efeito da localização geográfica na aplicabilidade do dicionário. Os resultados das classificações podem ser vistos na tabela 2 e na figura 2.

Para o log base, conforme esperado, obtivemos a melhor cobertura, com 65,1% dos acessos do log classificados, apesar da simplicidade do algoritmo. Em números absolutos, isso corresponde a 940,000 acessos classificados, de um total de 1,450,000. As categorias com mais acessos foram Internet e Busca, com aproximadamente 1/4 dos acessos. Confirmando resultados de [4], há grande interesse por assuntos como notícias, sexo e ciências (ensino/pesquisa).

Nos outros dois logs retirados de um dos servidores do POP-MG, após um intervalo de 4 meses, e utilizando o mesmo dicionário, conseguimos uma cobertura muito semelhante, classificando 60% e 63% dos acessos constantes destes logs. Ainda observando a tabela 2, podemos ver que as porcentagens para cada categoria não variaram mais do que 5% entre o log base e os novos logs. Consideramos esta consistência um forte indicativo de que as classificações

Internet	Provedores de Acesso, Informação e Email
Busca	Páginas de busca (Cadê, Yahoo, etc)
Sexo	Tema Adulto
Notícias	Jornais, Revistas, Agências
Informática	Empresas de Software, Hardware, Download, Programação
BatePapo	Salas de chat
Pessoais	Páginas pessoais
Ensino/Pesquisa	Ensino Superior, Básico, Instituições de Pesquisa
Anúncios	Páginas de empresas de publicidade via Web
Esportes	Notícias esportivas, páginas de atletas
Informações/Serviços	Páginas com informações sobre empresas e serviços prestados por elas
Radios/Televisões	Páginas de rádios e televisões
Música	Venda de CDs e Páginas de Bandas e Cantores
Diversão	Piadas, Quadrinhos, Jogos
Automóveis	Páginas de produtos e empresas automobilísticas
Finanças	Bancos, Seguradoras, Bolsas, Factoring
Livrarias	Vendas de Livros
Turismo	Agências de Turismo e Páginas de Lugares Turísticos
ComprasOnline	Compra de produtos diversos
Cinema	Páginas de filmes
Alimentação	Páginas de empresas e produtos alimentícios
Artes/Cultura	Museus, Monumentos e Atividades Culturais
Religião	Páginas de Igrejas e Seitas Religiosas
Política	Partidos Políticos

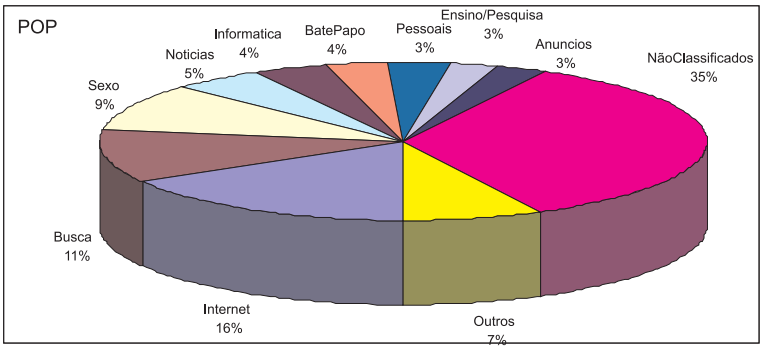
Tabela 1: Categorias Semânticas e seus significados

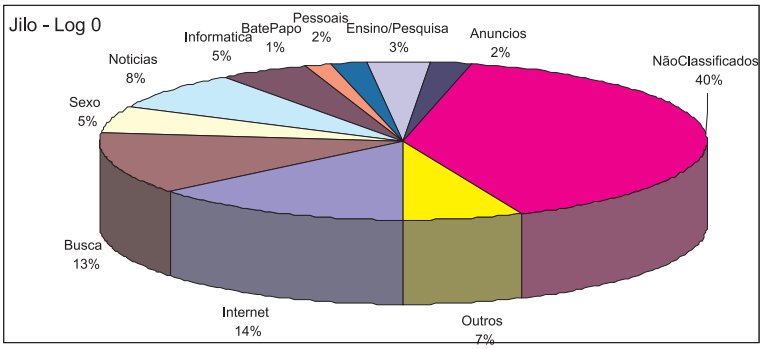
obtidas podem ser tomadas como representativas da real distribuição dos acessos por assuntos, e de que nossa metodologia pode ser aplicada para um dado servidor independentemente do tempo.

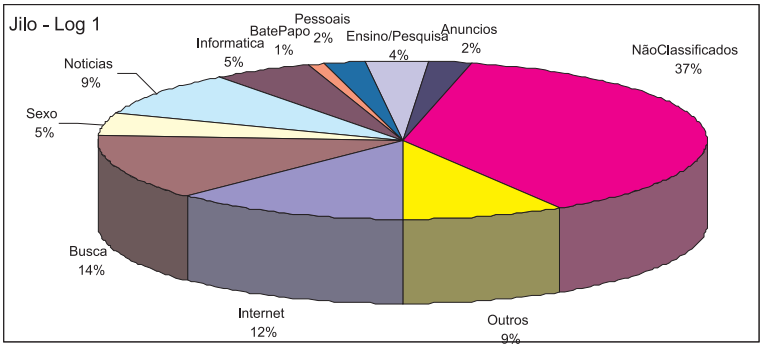
Por outro lado, utilizando o mesmo dicionário para classificar URLs do log de Portugal, com o intuito de avaliar o impacto da variação de contexto geográfico e sócio-cultural, obtivemos resultados bem diversos dos anteriores. Apenas 37% dos acessos foram classificados, e houve grande variação nas porcentagens referentes a cada categoria. Com base em dados de servidores cache em diversos países, Almeida et alli [1] observam que os padrões de acessos devem variar com o lugar. Isso explica, em parte, algumas diferenças entre os logs brasileiros e português. No entanto, os resultados obtidos na classificação deste último log nos apontam os principais motivos da baixa eficiência do algoritmo neste caso. Nos logs brasileiros, a categoria que obteve maior número de acessos foi a que se referia a páginas de provedores de acesso, provedores de informação e Email, cobrindo cerca de 15% do total das páginas analisadas. No log português essa proporção caiu drasticamente, ficando em torno de 1% do total de páginas. É interessante notar que este resultado já deveria ser esperado, pois o dicionário foi feito com base num log brasileiro e as referências a provedoras deveriam ser diferentes de um país para outro, já que se tratam de empresas locais.

Além da categoria Internet, houve, também, outras categorias que se mostraram bastante significativas nos logs brasileiros e que pareciam irrelevantes no log português. A categoria “Notícias”, referente a páginas de jornais e revistas, e “Ensino e Pesquisa”, referente a páginas de instituições de pesquisa e de ensino básico e superior, exemplificam esta aparente diferença no comportamento entre brasileiros e portugueses. Podemos observar que a maioria das páginas acessadas no Brasil, relacionadas a categoria “Notícias”, são de jornais e revistas brasileiros e como a identificação delas foi feita através dos nomes de tais jornais e revistas, como por exemplo “istoe” e “fsp”, é de se esperar que estes nomes não sejam encontrados nos logs portugueses. O mesmo pode ser dito em relação às instituições de “Ensino e Pesquisa” (e.g., ufmg, ufrj).

Quando observamos categorias que podem ser classificadas com base nas mesmas palavras tanto no Brasil quanto em Portugal, como “Sexo” (ex.:“sex”, “erotic” e “playboy”) ou categorias onde as páginas são, em sua maioria, internacionais como “Busca”, podemos constatar a proximidade dos valores referentes a quantidade de acessos a tais categorias. Uma última observação é que nenhuma das categorias teve seu respectivo número de acessos aumentado na avaliação do log português em relação aos logs brasileiros.







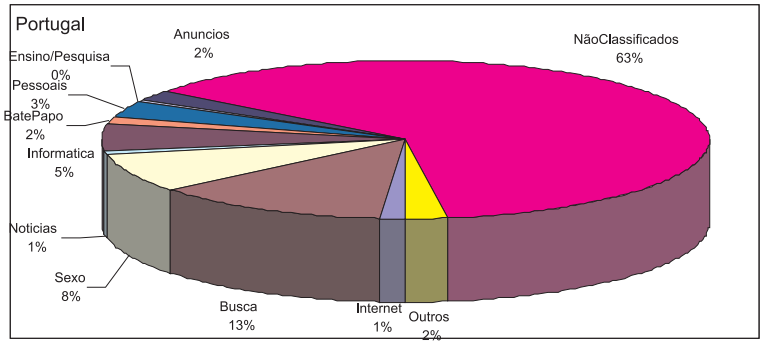


Figura 2: Taxonomia dos acessos

Categorias	log Base	Log 0		Log 1		Portugal
		%real	%var	%real	%var	%real
Internet	16	14	-2	12	-4	1
Busca	11	13	+2	14	+3	13
Sexo	9	5	-4	5	-4	8
Notícias	5	8	+3	9	+4	1
Informática	4	5	+1	5	+1	5
Bate Papo	4	1	-3	1	-3	2
Pessoais	3	2	-1	2	-1	3
Ensino/Pesquisa	3	3	0	4	+1	0
Anúncios	3	2	-1	2	-1	2
Outros	7	7	0	9	+2	2
Total Classificados	65	60	-5	63	-3	37
Não Classificados	35	40	+5	37	+2	63

Tabela 2: Classificação por Categorias e Desvios do base

5. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho avaliamos a aplicabilidade de um novo algoritmo de classificacão semântica de páginas da WWW. No que se refere à variação temporal, observamos que o método é robusto, tendo apresentado resultados bastante semelhantes em logs obtidos com um intervalo de até quatro meses. No que se refere à utilização do algoritmo em diferentes localidades geográficas, ele ainda poderá ser aplicado devendo, no entanto, ser adaptado, via dicionário, ao contexto que se deseja analisar.

Pretendemos continuar este trabalho aplicando o algoritmo para logs coletados em dias e horários específicos de forma a compreender melhor as mudanças no comportamento dos usuários, no que se refere ao acesso à Internet, de acordo com a hora do dia e ao dia da semana, ou por influência de algum evento de impacto social.

Agradecimentos

Os autores gostariam de agradecer a Márcio Anthony G. Cesário pelo auxílio na obtenção dos logs utilizados, e ao POP-MG[7] e ao provedor de acesso à Internet Esoterica – Novas Tecnologias de Informação, S.A.[5] – pela permissão para a obtenção de estatísticas a partir de seus respectivos logs.

REFERÊNCIAS

[1] V. Almeida, M. Cesário, R.Fonseca, W. Meira Jr. , and C. Murta. The influence of geographical and cultural issues on the cach e proxy server workload. In *Proceedings of WWW7*, 1998.

[2] National Laboratory for Applied Network Research. Squid Internet Object Cache. http://squid.nlanr.net/Squid/.

[3] Huberman, B.,et al. Strong regularities in world wide web surfing. Xerox Palo Alto Reseach Center, 1998.

[4] IBOPE. 2a. Pesquisa Cadê?/IBOPE. http://www.ibope.com.br/cade97/welcome.htm.

[5] S.A. Internet Esoterica Novas Tecnologias de Informação. Home Page. http://www.esoterica.pt/home.html.

[6] C. Kehoe and J. Pitkow. Surveying the territory: GvÚs five www user surveys. *The World Wide Web Journal*, 1(3), 1996.

[7] POP-MG. Ponto de Presença da Internet em Minas Gerais. http://www.pop-mg.rnp.br.