UNIVERSIDADE FEDERAL DO AMAZONAS – UFAM
PRO REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DEPARAMENTO DE APOIO A PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

IDENTIFICAÇÃO DE REGIÕES REGULARES EM PÁGINAS WEB

Bolsista: Tiago Pinho da Silva, FAPEAM

MANAUS
2012

UNIVERSIDADE FEDERAL DO AMAZONAS – UFAM
PRO REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DEPARAMENTO DE APOIO A PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

RELATORIO PARCIAL
PIB-E/0103/2011

IDENTIFICAÇÃO DE REGIÕES REGULARES EM PÁGINAS WEB

Bolsista: Tiago Pinho da Silva, FAPEAM
Orientador: David Braga Fernandes de Oliveira

MANAUS
2012

Sumário

1. Resumo

Regiões regulares são blocos das páginas Web cujo conteúdo é apresentado seguindo uma regularidade ou recorrência. Exemplos típicos de blocos com conteúdo regular são os menus, que são normalmente formados por um conjunto de links dispostos segundo uma regularidade estrutural. Informações sobre a localização das regiões regulares das páginas Web são muito úteis para diversas aplicações, tais como extratores de informações e registros de páginas Web, sistemas de recuperação de informação, segmentadores de páginas, etc. Através deste projeto de iniciação científica, nós idealizamos e desenvolvemos um algoritmo automático de identificação das regiões regulares de um conjunto de páginas Web dadas de entrada. Conforme os critérios iniciais deste projeto, nosso algoritmo provê certa tolerância à pequenas irregularidades dentro destas regiões seja através do uso de limiares. Os resultados experimentais atestam a qualidade do método desenvolvido, apresentando uma precisão de identificação das páginas regulares próxima a 100%.

2. Introdução

As páginas Web podem ser subdivididas em diferentes segmentos ou blocos (tais como menus, títulos, parágrafos, notas de rodapé etc), cada qual com uma função específica dentro das páginas [1, 2, 3, 4]. Muitos métodos de recuperação de informação [7, 8, 9] e mineração de dados [5, 6] têm procurado tirar proveito desta característica das páginas Web, buscando reconhecer em cada bloco uma fonte de informação independente e de importância variável.

Na literatura existem inúmeros métodos de identificação dos blocos das páginas, sendo que um dos mais eficientes foi proposto por Fernandes et. al [1]. Um dos passos propostos por Fernandes et al para fazer essa segmentação é a identificação de porções de código HTML das páginas dispostos de forma recorrente ou regular. Por exemplo, os menus presentes nas maiorias das páginas Web possuem um conjunto de links dispostos em uma barra vertical ou horizontal. Analisando a estrutura DOM desses menus, é possível perceber que o conjunto de *tags* que separam dois links adjacentes são sempre os mesmos, de forma que o código do menu na árvore DOM é formado por sucessivas repetições das mesmas *tags*. Diz-se, portanto, que tais regiões possuem estrutura regular. Além de ser útil para o método de segmentação proposto por Fernandes et al, a tarefa de identificação destes elementos é muito importante para diversas atividades de recuperação de informação e Web mining.

No entanto, de acordo com os experimentos mostrados em Fernandes et al, a identificação das regiões de estrutura recorrente pode ser afetada pela presença de pequenas irregularidades dentro destas regiões. Por exemplo, o fato de apenas um dos links de um dado menu estar em negrito já quebra a regularidade da região da árvore DOM correspondente, impedindo a identificação deste segmento como uma região de estrutura recorrente. No entanto, embora o elemento em negrito descaracterize aquilo que poderia ser considerado como uma região de perfeita regularidade, o menu em questão ainda possui conteúdo extremamente regular.

Desta forma, o objetivo deste trabalho foi desenvolver e implementar um algoritmo capaz de identificar todas as regiões recorrentes de um grupo de páginas Web. Nosso algoritmo provê certa tolerância à pequenas irregularidades dentro destas regiões através do uso de limiares.

3. Revisão Bibliográfica

Com base em um levantamento feito na literatura, foi possível identificar prós e contras de alguns métodos já existentes para o problema de identificação de estruturas regulares em páginas Web. Tais métodos, que serviram de base para a idealização do algoritmo proposto neste trabalho, são o *Mining Data Records* [10] e o *Somtree* [1]. Esses métodos são descritos a seguir.

Em [1] é proposto um novo modelo de representação do conteúdo de Web sites em sistema de recuperação de informação que leva em consideração a estrutura interna das páginas. Neste trabalho, é apresentado um método de identificação automática dos blocos presentes nas páginas Web, além de um conjunto de nove funções capazes de distinguir o impacto de ocorrências de termos dentro dos blocos das páginas. Em [1], a identificação dos blocos foi feita a partir da construção de uma estrutura em árvore chamada SOMtree, que é capaz de aglutinar as páginas HTML e representá-las em uma única estrutura.

Uma vez que a estrutura SOMtree é utilizada por nosso algoritmo de identificação de regiões regulares, vamos rapidamente introduzir os procedimentos de construção desta estrutura. A Figura 1 mostra um exemplo de uma SOMtree gerada a partir de duas páginas HTML, ρ_1 e ρ_2. Nesta figura, vemos que as páginas ρ_1 e ρ_2 possuem como raiz a tag HTML, e por causa disso a SOMtree resultante da aglutinação dessas páginas também possuirá a tag HTML, mas com um contador indicando que esta tag ocorreu nas duas páginas de origem. Da mesma forma, todas as tags que ocorrem em ambas as páginas serão copiadas para a SOMtree, todas com o valor de contador (frequência) igual a 2. Olhando as páginas, verificamos que elas divergem abaixo do nó div à direita. Abaixo desse nó vemos que a página ρ_1 tem os nós h1 e pre enquanto a ρ_2 tem os nós h1 e p. Como os nós h1 têm os mesmos caminhos em ambas as páginas ρ_1 e ρ_2 este resulta em um único nó na SOMtree com frequência 2. Por outro lado os nós pre (em ρ_1) e p (em ρ_2) ocorrem somente uma vez em suas respectivas páginas. Sendo assim cada um deles resulta em um nó diferente na SOMtree. A contagem, ou frequência, desses dois nós é um em ambos os casos.

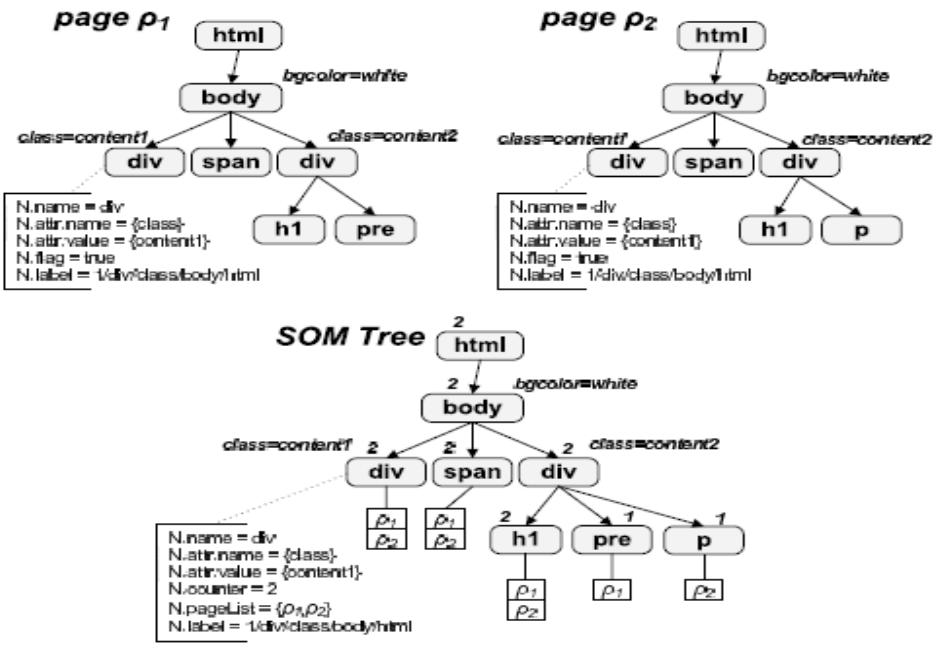

Figura 1 Exemplo da SOMtree gerada a partir das páginas p1 e p2.

O MDR (*Mining Data Records*) é um algoritmo utilizado para encontrar registros de dados em páginas WEB, sendo assim um pouco mais específico que o nosso problema. Entretanto, antes de encontrar os registros de dados de uma página, o algoritmo em questão necessita encontrar as regiões regulares das páginas, sendo, portanto, uma possível solução para o problema tratado neste trabalho.

O MDR possui um conceito chamado *nós generalizados*, que são conjuntos de nós que possuem o mesmo pai, possuem o mesmo tipo, e são adjacentes. Com base neste conceito, os autores de [10] definem que uma *região regular* é um conjunto de nós generalizados que possuem subárvores de mesmo tamanho.

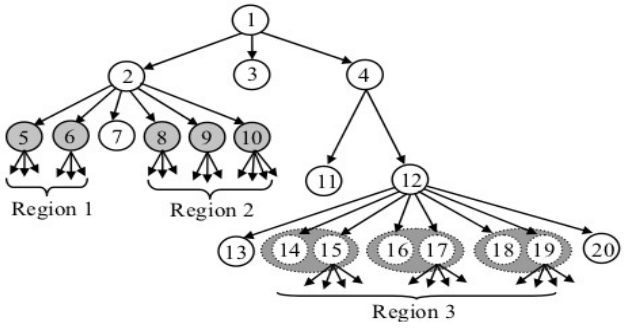

Figura 2: Exemplo de árvore de tags. Os nós dentro pintados de cinza, bem como os nós dentro de regiões cinzas, são sub-árvores das regiões regulares definidas pelas Regiões 1, 2 e 3.

Devemos enfatizar que uma região regular inclui as subárvores dos nós, e não somente os nós sozinhos.

4. Metodologia Utilizada

Para atingir os objetivos deste projeto, uma série de etapas foram seguidas, a saber:

- Levantamento dos métodos da literatura para identificação de regularidades em árvores genéricas.

- Idealização de um novo método de identificação de regiões regulares em árvores DOM tolerante a pequenas irregularidades. A idealização do novo método foi feita a partir de comparações e testes nestes dois algoritmos afim de encontrarmos prós e contras para o desenvolvimento do novo algoritmo.

- Implementação do método, que deverá ser feita usando a linguagem Perl.

- Experimentação e avaliação do método proposto. Para que se possa verificar a eficiência e eficácia do algoritmo, é necessário um conjunto de testes e comparações com outras técnicas para o problema de identificação de regiões regulares.

- Reuniões periódicas com o orientador: as reuniões são fundamentais para o direcionamento, auxílio e acompanhamento do andamento do projeto.

- Apresentação dos resultados: após pesquisas, análises e execução do projeto, serão apresentados os resultados finais obtidos com a solução proposta, embasado em gráficos e resultados medidos durante o decorrer da implementação.

5. Resultados e Discussões

Conforme falado anteriormente, o método proposto neste trabalho combina um algoritmo chamado MDR e uma estrutura de dados em árvore chamada SOMtree. O algoritmo possui os seguintes passos: 1) encontrar os conjuntos de nós generalizados presentes na página Web dada de entrada para nosso algoritmo, 2) gerar uma SOMtree para cada conjunto de nós generalizados, e 3) calcular o coeficiente de regularidade de cada região.

Segundo [10] toda região regular é criada a partir de um conjunto de nós generalizados, embora nem toda região que possua nós generalizados seja regular. Sendo assim a busca dos nós generalizados é feita para encontrar possíveis regiões regulares de forma a diminiur o escopo das possíveis soluções do problema. Depois de encontrados cada nó generalizado é guardado em um *array* para que futuramente a subárvore que este é raiz possa ser analisada e criada a SOMtree da mesma .

O segundo passo é criar a SOMtree de cada conjunto de nós generalizados encontrados no passo anterior. Conforme discutimos na seção 2, o número de nós de uma SOMtree será tão menor quanto maior for a regularidade das páginas incluídas na SOMtree. Desta forma, ao inserir as sub-árvores dos nós generalizados em uma única SOMtree, o número de nós da SOMtree resultante será um indício da regularidade dessas sub-arvores. Com base neste fato, neste trabalho nós inferimos a regularidade ou não de uma região através da razão entre a quantidade de nós presentes nas sub-árvores dos nós generalizados e a quantidade de nós encontrados na SOMtree formada a partir destes nós. A esta razão damos o nome de *coeficiente de regularidade*.

Através de uma série de experimentos, concluímos que uma região pode ser dita regular quando seu coeficiente de regularidade está entre 0 e 0.25. Para avaliarmos a eficácia de nosso método, selecionarmos um conjunto de 18 páginas e identificamos manualmente todas as suas regiões regulares. Desta forma, a eficácia de nosso método será medido através da comparação entre o conjunto de regiões regulares identificadas manualmente, e o conjunto de regiões regulares identificadas por nosso método. Como o método é um procedimento heurístico, em alguns exemplos, devido a casos especiais que fogem ao padrão observado, o resultado fica próximo do ideal. Entretanto pelo fato do resultado não se distanciar muito do ideal o método criado possui uma grande variedade de aplicações para ser utilizado.

Na tabela a seguir mostramos os resultados de alguns experimentos que demonstraram a eficácia do método idealizado. Como é possível notar o método não possui exatidão em alguns casos como por exemplo a página da Saraiva que a apesar de haver pouca regiões regulares o método não pôde encontrar todas a regiões, pelo fato do site não esta estruturado de forma organizada fazendo com que algumas regiões

regulares passem desapercebidas pelo método. Entretanto alguns sites como o Submarino possuíram 100% de exatidão isso se deve ao fato de que a página estava bem estruturada e o método pode detectar todos as regiões encontradas manualmente no site.

Como foi dito anteriormente o método deveria admitir pequenas irregularidades o que de fato ele faz, e pelo fato de ser um método automático possui ainda alguns ajustes que precisam ser tratados para deixá-lo cada vez mais eficaz.

Sites	Regiões encontradas pelo método	Não regiões encontradas pelo método	Não encontradas pelo método	Regiões Encontradas Manualmente
MercadoLivre	89%	0%	11%	9
Ebay	91%	0%	9%	11
Amazon	100%	3%	0%	13
Saraiva	80%	0%	20%	5
ShopTime	100%	9%	0%	10
Submarino	100%	0%	0%	8
Americanas	90%	9%	10%	11
Extra	100%	6%	0%	14
NetShoes	100%	10%	0%	10
Walmart	100%	9%	0%	11
Magazine Luiza	100%	0%	0%	15
Casas Bahia	95%	10%	0%	20
PoliShop	100%	7%	0%	14
CompraFacil	100%	0%	0%	16
Dealextreme	100%	0%	0%	39
Hot Toys	100%	0%	0%	6
Mania Virtual	100%	0%	0%	10
B&H	100%	27%	0%	8

Figura 4: Tabela de alguns resultados do método

6. Conclusão

Neste PIBIC propomos um algoritmo para identificação de regiões regulares de uma determinada página Web, analisando na árvore DOM desta página a frequência que alguns nós se repetem.

Os experimentos feitos com o algoritmo mostraram que este produz resultados bastante aproximados do reais, o que abre oportunidades para futuras aplicações deste método em problemas que envolvam regiões regulares.

Como trabalho futuro planejamos aplicar este método em um algoritmo de geração de resumos de páginas que será tratado no próximo PIBIC.

7. Referências Bibliográficas

[1] Fernandes de Oliveira, David. S. de Moura, Edleno, S. da Silva, Altigran. Ribeiro-Neto, Berthier. A Site Oriented Method For Segmenting Web Pages. In: Conference on Research and Development in Information Retrieval, ACM SIGIR'11, Pequim, China. 2011.

[2] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Vips: a vision-based page segmentation algorithm. Microsoft Technical Report, 2003.

[3] D. Chakrabarti, R. Kumar, and K. Punera. A graph-theoretic approach to webpage segmentation. In WWW'08, pages 377–386, 2008.

[4] C. Kohlschutter and W. Nejdl. A densitometric approach to web page segmentation. In CIKM '08, pages 1173–1182, New York, NY, USA, 2008. ACM.

[5] Y. Cao, Z. Niu, L. Dai, and Y. Zhao. Extraction of informative blocks from web pages. In ALPIT '08, pages 544–549, Washington, DC, USA, 2008. IEEE Computer Society.

[6] J. Kang and J. Choi. Detecting informative web page blocks for efficient information extraction using visual block segmentation. In ISITC '07, pages 306–310, Washington, DC, USA, 2007. IEEE Computer Society.

[7] K. Ahnizeret, D. Fernandes de Oliveira, J. M. Cavalcanti, E. S. de Moura, and A. S. da Silva. Information retrieval aware web site modelling and generation. In ER '04, pages 402–419, Berlin, Heidelberg, 2004. Springer.

[8] E. S. de Moura, D. Fernandes de Oliveira, B. Ribeiro-Neto, A. S. da Silva, and M. A. Goncalves. Using structural information to improve search in web collections. JASIST, 61:2503–2513, December 2010.

[9] D. Fernandes de Oliveira, E. S. de Moura, B. Ribeiro-Neto, A. S. da Silva, and M. A. Goncalves. Computing block importance for searching on web sites. In CIKM '07, pages 165–174, New York, NY, USA, 2007. ACM.

[10]. Liu, B., Grossman, R. and Zhai, Y. "Mining data records from Web pages." KDD-03, 2003

[11] Davi de Castro Reis, Paulo Braz Golgher, Altigran Soares da Silva, Alberto H. F. Laender:Automatic web newsextraction using tree edit distance. WWW 2004: 502- 511.

[12] S.M.Selkow. The tree-to-tree editing problem. Information Processing Letter, 6:184-186, Dec 1997.

[13] W.Yang. Identifying sintaxe differences between tow programs. Softw. Pract. Exper,. 21(7):739-755,1991.