Trabalho Prático 1:
Extração Automática de Dados

Disciplina: Gerência de Dados da Web

Professor: Alberto Laender

Alunos: Marco Túlio Correia Ribeiro
Lucas Cunha de Oliveira Miranda

{marcotcr,lucmir}@dcc.ufmg.br

17 de outubro de 2011

1 Introdução

Este trabalho consiste na implementação de um método para extração automática de dados presentes em uma coleção de páginas HTML fornecida como entrada.

Para o processo de extração, consideramos as seguintes hipóteses:

1. Levando em consideração o trabalho proposto por Valiente [1], os dados de interesse a serem extraídos das páginas estão nas folhas das sub-árvores que compõem a maior classe de equivalência dessas páginas. Essa hipótese, em geral, é válida - uma vez que cada registros presente nas páginas geralmente é composto por um único bloco, gerado automaticamente. Porém, em algumas páginas, os registros podem estar divididos em dois ou mais blocos, com classes de equivalência de tamanhos semelhantes. Portanto, a hipótese sob a qual implementamos o trabalho foi **"Os dados de interesse a serem extraídos das páginas estão nas folhas das sub-árvores que compõem as classes de equivalência mais frequentes dessas páginas"**, ou seja, consideramos a possibilidade de registros aparecerem em duas ou mais classes de equivalência frequentes.

2. **As folhas devem ser desconsideradas ao se calcular classes de equivalência das sub-árvores.** Essa hipótese é simples de ser verificada. A figura 1, por exemplo, retirada de uma página da amazon, contém 4 registros. Se as folhas forem consideradas ao se assinalar classes de equivalência, cada sub-árvore contendo um registro será assinalada a uma classe de equivalência. Se as folhas não forem consideradas, por outro lado, todas as sub-árvores contendo os registros serão assinalados à mesma classe de equivalência, e essa classe de equivalência será considerada frequente, facilitando a extração dos registros.

Figura 1: Exemplo de página da amazon

3. **Folhas compostas por tags HTML não são dados de interesse a serem extraídos** - Uma vez que o nosso extrator só lida com texto, folhas compostas por tags com imagens ou similares não são de nosso interesse - e portanto foram descartadas.

2 Implementação

Nesta seção, detalhamos a metodologia utilizada para a implementação do extrator, em especial visando as três hipóteses levantadas na seção anterior.

2.1 Determinando as sub-árvores de interesse

Para a determinação das classes de equivalência, utilizamos o método de Valiente [1] levemente modificado, tal que as folhas não sejam consideradas na determinação de classes de equivalência. Essa modificação é descrita em maiores detalhes na subseção seguinte. Foi fornecido o código fonte de uma implementação do método original.

Considerando a hipótese **"Os dados de interesse a serem extraídos das páginas estão nas folhas das sub-árvores que compõem as classes de equivalência mais frequentes dessas páginas"**, existe um compromisso entre a frequência de uma classe de equivalência e o tamanho da sub-árvore que é representada por essa classe. Em geral, as folhas compõem as maiores classes de equivalência (embora as mesmas não sejam consideradas por nosso extrator, devido à hipótese 2). As sub-árvores de tamanho 2, compostas geralmente por tags HTML envolvendo as folhas, são geralmente as próximas classes de equivalência mais frequentes. Na figura 1, por exemplo, a sub-árvore contendo a tag ¡a¿ (anchor), que envolve todos os links demonstrados em azul na página, é a que possui a classe de equivalência mais frequente. Porém, muitas das folhas encapsuladas por essa tag não contém dados de interesse ("Home & Kitchen", por exemplo).

Utilizando vários exemplos de teste, chegamos à conclusão que é interessante avaliar apenas sub-árvores de tamanho mínimo 3. Uma vez que esse parâmetro é determinado, precisamos decidir quais classes de equivalência devem ser consideradas "de interesse". Para tal, decidimos considerar as 10 classes de equivalência mais frequentes como interessantes, desde que o decaimento de frequência entre uma classe e a próxima menos frequente seja menos do que um limiar, que fixamos em 90%. A determinação de classes de equivalência que representam sub-árvores de interesse é apresentada em pseudo-código na Figura 2, a seguir.

```
1: Dada uma página html:
2: Utiliza o método de valiente modificado para determinar as classes de equivalência
3: Conta a frequência das classes de equivalência que representam as sub-árvores de tamanho mínimo 3
4: Ordena as classes de equivalência por frequência
5: i = 1
6: for all Classe de equivalência frequente j do
7:    if Frequencia_j/Frequencia_{j-1} < 0.1 OR i = 10 then
8:       break
9:    end if
10:   Adiciona j à lista das classes de interesse
11:   i++
12: end for
```

Figura 2: Determinação de classes de equivalência que representam sub-árvores de interesse.

2.2 Modificação no método de Valiente

Dada a hipótese **"As folhas devem ser desconsideradas ao se calcular classes de equivalência das sub-árvores."**, foi necessário modificar pontualmente o método proposto por Valiente [1], de forma que os *mappings bottom-up* não levem em consideração as folhas. Por exemplo, a Figura 3 mostra através de nós coloridos de verde a maior floresta comum entre a Arvore 1 e a Arvore 2, utilizando o método de valiente. Vale notar que o que impediu que uma floresta comum maior fosse identificada é a diferença entre as folhas 2 (da Arvore 1) e 5 (da Arvore 2). A figura 4, por outro lado, mostra a maior floresta comum entre as duas árvores utilizando o método de Valiente modificado, de forma que as folhas não são consideradas. Nesse caso, as duas árvores são consideradas equivalentes.

Desta forma, sub-árvores contendo os registros serão consideradas equivalentes, mesmo se os registros em si contenham dados diferentes.

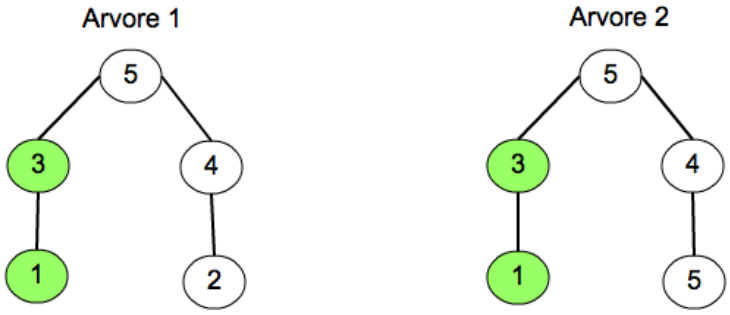

Figura 3: Maior floresta comum entre Arvore1 e Arvore2, utilizando o método de Valiente.

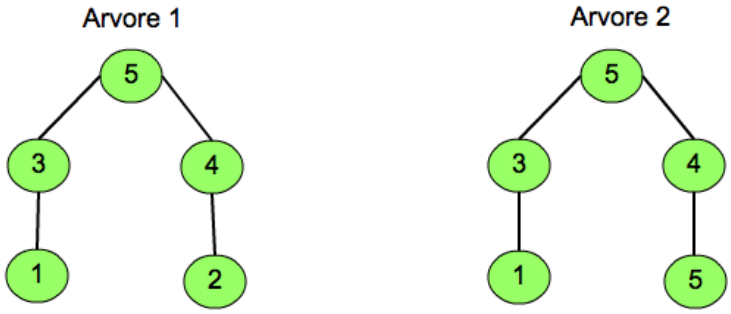

Figura 4: Maior floresta comum entre Arvore1 e Arvore2, utilizando o método de Valiente modificado.

2.3 A extração

Utilizando a metodologia descrita na subseção 2.1, um conjunto de no máximo 10 classes de equivalência consideradas interessantes é extraído. Dado esse conjunto, é possível que existam classes de equivalência que contém outras classes de equivalência, e isso não é desejável. Portanto, removemos do conjunto qualquer classe de equivalência que represente uma sub-árvore de outra classe de equivalência.

Dadas o conjunto atualizado das classes de equivalência de interesse, as folhas que não são compostas por tags html (segundo a hipótese 3 apresentada na introdução deste relatório) são extraídas. Todas as extrações são consideradas parte do mesmo registro, até que se uma classe de equivalência se repita. O pseudo-código do processo de extração é apresentado na Figura 5, a seguir:

1: Dado o conjunto de classes de equivalencia c interessantes e a página html p
2: Remove de c todas as classes de equivalência que estão contidas em outras classes de equivalência
3: $primeira\_classe = NULL$
4: **for all** Nodos n de p **do**
5: **if** n pertence a c **then**
6: **if** $primeira\_classe = NULL$ **then**
7: $primeira\_classe = n$
8: **end if**
9: **if** $n = primeira\_classe$ **then**
10: Inicia um registro novo
11: **end if**
12: Extrai as folhas de n que não sejam tags html
13: **end if**
14: **end for**

Figura 5: Determinação de classes de equivalência que representam sub-árvores de interesse.

Referências

[1] Gabriel Valiente. An efficient bottom-up distance between trees. In *Proceedings of the 8th International Symposium of String Processing and Information Retrieval*, pages 212–219. Press, 2001.