

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

18 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 9, 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.

Home & Kitchen › Kitchen & Dining › Coffee, Tea & Espresso

Showing 1 - 24 of 10,689 Results Sort by Popularity

			
Krupps 20342 Electric Coffee and Spice grinder with stainless steel blades, Black <u>Buy new:</u> \$24.00 \$19.99 <u>10 new</u> from \$17.98 <u>1 used</u> from \$20.97 Get it by Tuesday, Oct 18 if you order in the next 10 hours and choose one-day shipping. ★★★★☆ (483) Eligible for FREE Super Saver Shipping.	Cuisinart DCC-RWF1 Replacement Coffeemaker Water Filters, Set of 2 <u>Buy new:</u> \$7.75 <u>31 new</u> from \$4.19 Get it by Tuesday, Oct 18 if you order in the next 10 hours and choose one-day shipping. ★★★★☆ (164) Eligible for FREE Super Saver Shipping.	Keurig My K-Cup Reusable Coffee Filter <u>Buy new:</u> \$14.95 \$14.15 <u>35 new</u> from \$9.99 <u>1 used</u> from \$15.99 Get it by Tuesday, Oct 18 if you order in the next 10 hours and choose one-day shipping. ★★★★☆ (859) Eligible for FREE Super Saver Shipping.	Senseo Dark Roast Coffee, 18-Count Pods (Pack of 6) <u>Buy new:</u> \$33.26 \$27.02 <u>Subscribe & Save:</u> \$22.97 <u>14 new</u> from \$27.02 Get it by Tuesday, Oct 18 if you order in the next 10 hours and choose one-day shipping. ★★★★☆ (104) Eligible for FREE Super Saver Shipping.

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 9, 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.

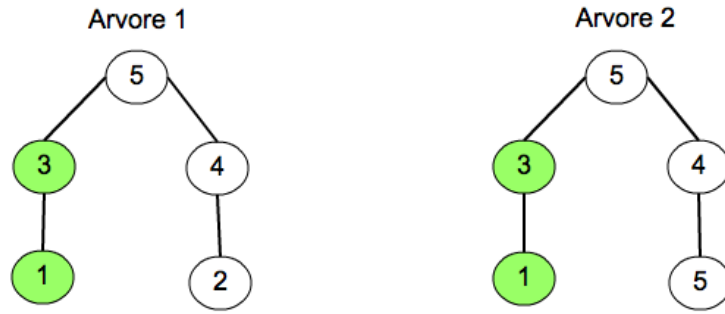


Figura 3: Maior floresta comum entre Arvore1 e Arvore2, utilizando o método de Valiente.

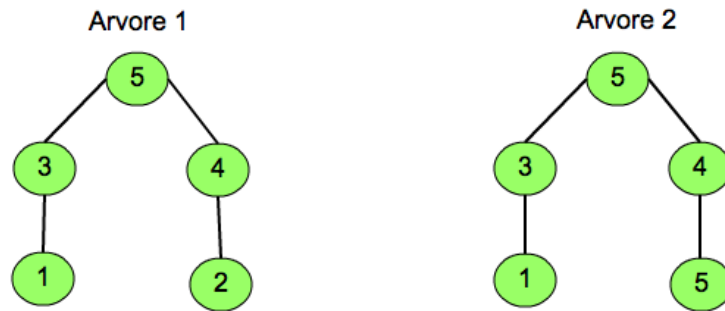


Figura 4: Maior floresta comum entre Arvore1 e Arvore2, utilizando o método de Valiente modificado.

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 contêm 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:    Extraí 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.

3 Avaliação Experimental

A qualidade de um extrator é determinada por sua capacidade de coletar o maior número possível de registros úteis, agregando o mínimo possível de informações irrelevantes.

3.1 Metodologia

3.1.1 Métricas

Para medir a... (*recall*) e precisão *precision*.

3.1.2 Como medir

As métricas serão aplicadas de duas maneiras...
por registro por atributo...
por média dos valores de precisão e revocação de cada registro


3.2 Experimentos e resultados

Os experimentos foram aplicados sobre 4 coleções, considerando a metodologia explicada.

3.2.1 Coleção “cdnow”

Esta coleção, possui...

Make your own custom Valentine's CD

CDnow Home
 Shipping to Europe or the Middle East?
[Rock/Pop](#)
[Jazz/Blues](#)
[Urban/Electronic](#)
[Classical](#)
[Country/Folk](#)
[World/New Age](#)
[Children](#)

[MTV CD Lounge](#)
[VH1 Music Shop](#)

Artist

Album Advisor™—Discover music based on your taste.

Hundreds of our best-selling albums now 30% off!

Stock up on your favorite music.

	List Price	Add to Cart
Adiemus Songs Of Sanctuary	\$16.97	\$11.88
Air (Techno) Moon Safari	\$16.97	\$11.88
Alabama For The Record-41 Number One Hits	\$28.97	\$20.28
Ally Mcbeal TV Soundtrack	\$17.97	\$12.58
Armageddon (Soundtrack) Soundtrack	\$17.97	\$12.58
J.S. Bach Son & Partitas (6)	\$20.97	\$14.68
Backstreet Boys Backstreet Boys	\$17.97	\$12.58
Barenaked Ladies Stunt	\$16.97	\$11.88
Barenaked Ladies Rock Spectacle	\$16.97	\$11.88
Cecilia Bartoli Live In Italy	\$17.97	\$12.58
Beastie Boys Hello Nasty	\$17.97	\$12.58
Beatles Beatles (White Album/Limited Ed.)	\$32.97	\$23.08

Figura 6: Página exemplo da coleção “cdnow”

Entrada	Total	Extraídos	Corretos	Precisão	Revocação
cd1	30	39	30	0.77	1.0
cd2	30	39	30	0.77	1.0
cd3	30	39	30	0.77	1.0

Tabela 1: Resultados da avaliação por registro para coleção “cdnow”

Entrada	Total	Extraídos	Corretos	Precisão	Revocação
cd1	120	129	120	0.93	1.0
cd2	120	129	120	0.93	1.0
cd3	120	129	120	0.93	1.0

Tabela 2: Resultados da avaliação por atributo para coleção “cdnow”

4 atributos por registro. Os registros adicionais possuíam poucos atributos.

Entrada	Precisão	Revocação
cd1	0.77	0.77
cd2	0.77	0.77
cd3	0.77	0.77

Tabela 3: Resultados da avaliação considerando a média dos valores por registro para a coleção “cdnow”

Os 30 registros do gabarito foram extraídos corretamente (com todos os atributos). Entretanto, foram obtidos 9 registros adicionais, que não são relevantes para o objetivo da extração. Para cada um destes registros, a precisão e a revocação assumem valor 0. Portanto, os valores de precisão e revocação são calculados pela média aritmética $\frac{30 \times 1.0 + 9 \times 0.0}{39} = 0.77$.

3.2.2 Coleção “monster”

Esta coleção, possui...

Entrada	Total	Extraídos	Corretos	Precisão	Revocação
monster1	50	50	50	1.0	1.0
monster2	50	44	44	1.0	0.88
monsters4	50	50	50	1.0	1.0

Tabela 4: Resultados da avaliação por registro para coleção “monster”

Não foi coletado nenhum registro adicional

Entrada	Total	Extraídos	Corretos	Precisão	Revocação
monster1	300	300	300	1.0	1.0
monster2	298	274	274	1.0	0.92
monster4	299	299	299	1.0	1.0

Tabela 5: Resultados da avaliação por atributo para coleção “monster”

monster1 contém 6 atributos por registro. monster2 contém 6 atributos por registro, exceto por uma tupla que contém apenas 4. monster4 contém 6 atributos por registro, exceto por uma tupla que contém apenas 5. Todos os atributos dos registros extraídos foram coletados corretamente.

Os atributos foram coletados corretamente, entretanto alguns registros inteiros foram perdidos no caso de registros com formatação diferente..

Internet

All Jobs Subsearch

[[New Search](#)]

Monster Technology Jobs 1 to 50 of more than 1,000

Date	Location	Job Title	Company
Jun 8	US-TX-fort worth	Web Developer	StarTech Staffing
Jun 8	US-TN-Nashville	Programmer Analyst	OAQ
Jun 8	US-CA-Sacramento	Lotus Notes/Domino Developer	Dynamic Staffing
Jun 7	US-CA-San Francisco	Development Manager	LookSmart
Jun 7	US-VA-FallsChurch	Internet Consultant	AppNet, Inc.
Jun 7	US-IL-Chicago	OpenStep Opportunity	Technisource
Jun 7	US-CO-Broomfield	Oracle Database Administrator	Level 3 Communications
Jun 7	US-CA-San Francisco	Programmer/Analyst - COBOL	Boeing
Jun 7	US-NY-New York City	SR. NETWORK PLANNER	Forsoft
Jun 7	South Korea	Senior Software Developer	01 Inc
Jun 7	US-CT-Wallingford	project lead	Maxim Group
Jun 7	US-GA-Atlanta	Network Verification Testers	Butler International
Jun 7	US-IL-Chicago	Internet/Intranet Developer	Hall Kinion
Jun 7	US-CA-Los Angeles	Project Manager	Search West
Jun 7	US-WA-Bothell	Systems/Business Analyst	Parity
Jun 7	US-IN-Indianapolis	Management Information Systems Manager	Haggard & Stocking
Jun 7	US-GA-Atlanta	NETWORK ADMINISTRATORS	Equifax
Jun 7	South Korea	Web Engineer	01 Inc
Jun 7	US-MA-Boston	Web Systems Administrator	Aquent Partners
Jun 7	US-GA-Atlanta	INTERNET ARCHITECTS	Equifax

Figura 7: Página exemplo da coleção “monster”

Entrada	Precisão	Revocação
monster1	1.0	1.0
monster2	1.0	0.88
monster4	1.0	1.0

Tabela 6: Resultados da avaliação considerando a média dos valores por registro para a coleção “monster”

3.2.3 Coleção “music_musicall”

Entrada	Total	Extraídos	Corretos	Precisão	Revocação
001	50	44	44	1.0	0.88
005	25	23	23	1.0	0.92
006	50	46	46	1.0	0.92

Tabela 7: Resultados da avaliação por registro para coleção “music_musicall.com”

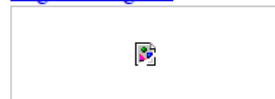
Não foi coletado nenhum registro adicional

Gabarito apresenta 3 Atributos por registro, entretanto, a primeira coluna deveria ser computada... cálculos consideraram esta coluna

» [Music Videos](#) » [Writers' Bloc](#)
 » [Top Searches](#) » [Whole Note](#)
 » [Classical Corner](#) » [Artist Spotlight](#)
 » [Top Composers](#) » [Classical Reviews](#)

You are not logged in.

[Login or Register](#)



explore by.../jazz/
Top Albums

[Send to Friend](#)

[Genre](#)

1 2 3 4 5

[Next page>](#)

Tier	Artist	Title	Year
1st	Miles Davis	'Round About Midnight [Bonus Tracks]	2001
1st	Miles Davis Quintet	'Round About Midnight [Japan]	2001
1st	Miles Davis	'Round About Midnight	1955
1st	Jelly Roll Morton	1923-1924 [Milestone]	1923
1st	Chick Webb	1929-1934	1929
1st	Stéphane Grappelli	1935-1940	1935
1st	Willie "The Lion" Smith	1938-1940	1938
1st	Bobby Short	50 by Bobby Short	1986
1st	Original Dixieland Jazz Band	75th Anniversary	1917
1st	John Coltrane Quartet	A Love Supreme [Japan 2001 Reissue]	1964
1st	John Coltrane	A Love Supreme	1964
1st	Fletcher Henderson	A Study in Frustration/Thesaurus of Classic	1923

Figura 8: Página exemplo da coleção “music_musicall.com”

Entrada	Total	Extraídos	Corretos	Precisão	Revocação
001	200	176	176	1.0	0.88
005	100	92	92	1.0	0.92
006	200	184	184	1.0	0.92

Tabela 8: Resultados da avaliação por atributo para coleção “music_musicall.com”

Entrada	Precisão	Revocação
001	1.0	0.88
005	1.0	0.92
006	1.0	0.88

Tabela 9: Resultados da avaliação considerando a média dos valores por registro para a coleção “music_musicall.com”

3.2.4 Coleção “wines”

O menu foi capturado por ser bem estruturado e possuir mais registros do que a tabela principal (que contém as informações de interesse).

Cada registro possui 6 atributos. Foram coletados 16 atributos a mais (referentes ao menu) e 2 registros ficaram com apenas um atributo coletado por causa da ausência do título

$$16 \cdot 0 + 9 + 1.0 + 1 \cdot$$

☐ [Red Wines](#)
[Cabernet](#)
[Merlot](#)
[Zinfandel](#)
☐ [White Wines](#)
[Chardonnay](#)
[Bubbly Wines](#)
☐ [Rare Wines](#)
☐ [What's New](#)
[Samplers](#)
☐ [Specials](#)
☐ [Peter's Picks](#)
☐ [Bang for the Buck](#)
[Personalized Wine](#)
[Wineries](#)
[Wine Team](#)

Have a question for Customer Service? Use [Live Help](#) to get it answered without going offline. Monday-Sunday: 8 a.m. to 5 p.m. Pacific time

 To make your shopping easier, select your shipping destination in advance.

Sort by:

Page 1 of 5 1 2 3 4 5

Barberani

[1998 Barberani Orvieto Classico](#) ☐ **\$13.00**
[Castagnolo, Umbria, Italy](#)
A lovely, dry vineyard-designated Orvieto Classico from one of the region's best producers. Delicious flavors of fig, green melon, pear, and light herb offer a refreshing respite when you've had one too many oaky whites. [Add to My Wish List](#)

[1998 Barberani Grechetto, Italy](#) ☐ **\$19.95**
Grechetto is the most distinctive and interesting of the five grape varieties used in Orvieto Classico. This wine offers a nice clean, refreshing taste of melon, pear, and anise. Pairs beautifully with fresh trout, calamari, or Caesar salad. [Add to My Wish List](#)

Bartenura

[Bartenura Asti Spumante, Piedmont, Italy](#) ☐ ☐ **\$14.00**
This wonderfully aromatic Italian bubbly is bright, crisp, and just sweet enough to pair well with ripe fruit. Kosher for Passover. [Add to My Wish List](#)

Bolognani

[1998 Bolognani Muller-Thurgau, Trentino, Italy](#) ☐ **\$13.00**
Muller-Thurgau (Mew-lair Toor-gau) sounds suspiciously like a [Add to My](#)

Figura 9: Página exemplo da coleção “wines”

Entrada	Total	Extraídos	Corretos	Precisão	Revocação
winesByProducer01	10	26	10	0.38	1.0
winesByProducer02	10	26	10	0.38	1.0
winesByProducer03	10	26	10	0.38	1.0

Tabela 10: Resultados da avaliação por registro para coleção “wines”

Entrada	Total	Extraídos	Corretos	Precisão	Revocação
winesByProducer01	60	71	55	0.92	0.91
winesByProducer02	60	66	50	0.75	0.83
winesByProducer03	60	71	55	0.92	0.91

Tabela 11: Resultados da avaliação por atributo para coleção “wines”

Entrada	Precisão	Revocação
winesByProducer01	0.38	0.35
winesByProducer02	0.38	0.32
winesByProducer03	0.38	0.35

Tabela 12: Resultados da avaliação considerando a média dos valores por registro para a coleção “wines”

4 Conclusão

asasd

Métricas/Coleções	cdnow	monster	music_allmusic.com	wines	Média geral
Precisão por registro	0.77	1.0	1.0	0.38	0.79
Revocação por registro	1.0	0.96	0.91	1.0	0.97
Precisão por atributo	0.93	1.0	1.0	0.86	0.95
Revocação por atributo	1.0	0.97	0.91	0.88	0.94
Precisão pela média	0.77	1.0	1.0	0.38	0.79
Revocação pela média	0.77	0.96	0.89	0.34	0.74

Tabela 13: Média dos resultados por coleção e média geral de todos os resultados

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.