

# 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 [2], 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.

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

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

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

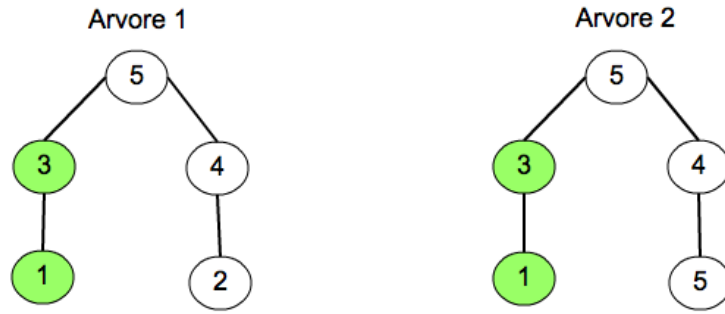


Figura 3: Maior floresta comum entre Arvore1 e Arvore2, utilizando o método de Valiente.

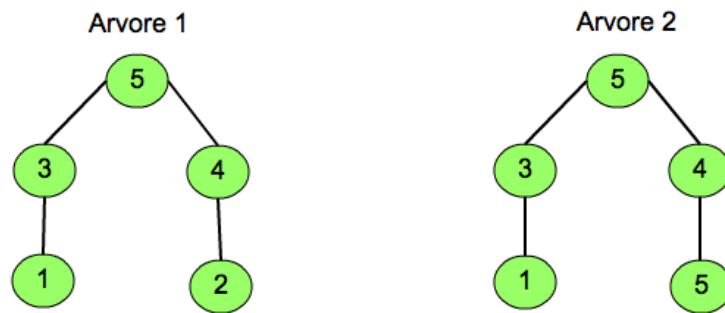


Figura 4: Maior floresta comum entre Arvore1 e Arvore2, utilizando o método de Valiente modificado.

## 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

O processo de extração proposto e implementado deve ser avaliado quanto à sua efetividade. A qualidade de um extrator é determinada pela capacidade de coletar o maior número possível de registros úteis, agregando o mínimo possível de informações irrelevantes.

### 3.1 Metodologia de avaliação

A avaliação experimental consiste na aplicação de métricas tradicionais para quantificar o resultado da extração de informações em páginas de dados reais. Cada página empregada nos experimentos possui um gabarito correspondente, contendo todos os registros relevantes que devem ser extraídos. A comparação entre o resultado da extração e o gabarito, por meio de métricas apropriadas, fornece uma medida de efetividade do processo de extração.

#### 3.1.1 Métricas

Para avaliar a qualidade do extrator, empregou-se as métricas precisão (*precision*) e revocação (*recall*) [1]. A primeira mede a quantidade de respostas corretas em relação ao total de respostas retornadas e a segunda determina a quantidade de respostas corretas obtidas em relação ao total de respostas esperadas (relevantes):

$$prec = \frac{\text{respostas corretas}}{\text{respostas retornadas}} \quad rev = \frac{\text{respostas corretas}}{\text{total de respostas}}$$

#### 3.1.2 Aplicação das métricas

O processo de extração pretende obter, de páginas html, registros relevantes. Cada registro, por sua vez, contém atributos diversos. É importante que a avaliação do extrator considere a efetividade do processo de obtenção dos registros e dos seus atributos.

Para tanto, empregou-se três métodos de avaliação:

- Cálculo da precisão e revocação sobre os registros extraídos. Ou seja, avaliação da habilidade do algoritmo de localizar e extrair os registros relevantes.
- Cálculo da precisão e revocação sobre os atributos extraídos. Neste caso, o total de atributos relevantes extraídos que interessa, sendo considerados de maneira independente dos registros.
- Cálculo da precisão e revocação média dos registros extraídos. Este cálculo considera, respectivamente, a média aritmética da precisão e da revocação da extração dos valores dos atributos de cada registro. Ou seja, a precisão da extração de um registro será dada pela média dos valores de precisão da extração de todos os seus atributos e a revocação por registro será dada de maneira análoga.


### 3.2 Experimentos e resultados

A metodologia descrita foi empregada na avaliação do extrator atuando sobre quatro coleções, contento, cada uma, três páginas reais. O detalhamento dos experimentos, os resultados e a análise dos dados, para cada coleção, são apresentados em seguida.

#### 3.2.1 Coleção “cdnow”

A coleção “cdnow” é composta por páginas de um site de venda de álbuns de música. Como pode ser visto na figura 6, a página contém, basicamente, um menu com estilos musicais, um campo para busca e uma relação de álbuns à venda. Neste caso, os registros de interesse são os álbuns e os atributos de cada registro são o artista, o título do álbum e os dois valores de preço.

### 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

Find It

**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 <a href="#">Songs Of Sanctuary</a>	\$16.97	<del>\$11.88</del>
Air (Techno) <a href="#">Moon Safari</a>	\$16.97	<del>\$11.88</del>
Alabama <a href="#">For The Record-41 Number One Hits</a>	\$28.97	<del>\$20.28</del>
Ally Mcbeal <a href="#">TV Soundtrack</a>	\$17.97	<del>\$12.58</del>
Armageddon (Soundtrack) <a href="#">Soundtrack</a>	\$17.97	<del>\$12.58</del>
J.S. Bach <a href="#">Son &amp; Partitas (6)</a>	\$20.97	<del>\$14.68</del>
Backstreet Boys <a href="#">Backstreet Boys</a>	\$17.97	<del>\$12.58</del>
Barenaked Ladies <a href="#">Stunt</a>	\$16.97	<del>\$11.88</del>
Barenaked Ladies <a href="#">Rock Spectacle</a>	\$16.97	<del>\$11.88</del>
Cecilia Bartoli <a href="#">Live In Italy</a>	\$17.97	<del>\$12.58</del>
Beastie Boys <a href="#">Hello Nasty</a>	\$17.97	<del>\$12.58</del>
Beatles <a href="#">Beatles (White Album/Limited Ed.)</a>	\$32.97	<del>\$23.08</del>

Figura 6: Página exemplo da coleção “cdnow”

A tabela 1 contém os valores obtidos com o cálculo das métricas considerando apenas os registros extraídos (independente dos atributos estarem completos).

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”

Os dados da tabela 1 sugerem que o algoritmo foi capaz de extrair todos os registros relevantes (revocação igual a 1.0). Entretanto, como a precisão não foi alta, o extrator também retornou dados irrelevantes. Observando a página e analisando o conteúdo retornado, constata-se que o menu de estilos musicais também foi coletado. Isso ocorreu devido ao fato do menu ser bem estruturado e conter muitas entradas, sendo facilmente confundido com uma relação de registros relevantes.

A tabela 2 contém o cálculo de precisão e revocação considerando a coleta de atributos.

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”

Neste caso, cada registro contém exatamente 4 atributos. Todos os atributos dos registros relevantes foram coletados pelo extrator (o que é representado pelo valor de revocação 1.0). Os atributos adicionais

coletados, responsáveis pela redução da precisão, são os da listagem de estilos musicais (considerou-se um atributo por item).

A tabela 3 contém a média dos valores por registro para a coleção.

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”

A coleção “monster” contém páginas com uma listagem de registros formados pelos atributos: data, localização, cargo e companhia. A figura 7 é um exemplar desta coleção:

**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	<a href="#">Web Developer</a>	StarTech Staffing
Jun 8	US-TN-Nashville	<a href="#">Programmer Analyst</a>	OA0
Jun 8	US-CA-Sacramento	<a href="#">Lotus Notes/Domino Developer</a>	Dynamic Staffing
Jun 7	US-CA-San Francisco	<a href="#">Development Manager</a>	LookSmart
Jun 7	US-VA-FallsChurch	<a href="#">Internet Consultant</a>	AppNet, Inc.
Jun 7	US-IL-Chicago	<a href="#">OpenStep Opportunity</a>	Technisource
Jun 7	US-CO-Broomfield	<a href="#">Oracle Database Administrator</a>	Level 3 Communications
Jun 7	US-CA-San Francisco	<a href="#">Programmer/Analyst - COBOL</a>	Boeing
Jun 7	US-NY-New York City	<a href="#">SR. NETWORK PLANNER</a>	Forsoft
Jun 7	South Korea	<a href="#">Senior Software Developer</a>	01 Inc
Jun 7	US-CT-Wallingford	<a href="#">project lead</a>	Maxim Group
Jun 7	US-GA-Atlanta	<a href="#">Network Verification Testers</a>	Butler International
Jun 7	US-IL-Chicago	<a href="#">Internet/Intranet Developer</a>	Hall Kinion
Jun 7	US-CA-Los Angeles	<a href="#">Project Manager</a>	Search West
Jun 7	US-WA-Bothell	<a href="#">Systems/Business Analyst</a>	Parity
Jun 7	US-IN-Indianapolis	<a href="#">Management Information Systems Manager</a>	Haggard & Stocking
Jun 7	US-GA-Atlanta	<a href="#">NETWORK ADMINISTRATORS</a>	Equifax
Jun 7	South Korea	<a href="#">Web Engineer</a>	01 Inc
Jun 7	US-MA-Boston	<a href="#">Web Systems Administrator</a>	Aquent Partners
Jun 7	US-GA-Atlanta	<a href="#">INTERNET ARCHITECTS</a>	Equifax

Figura 7: Página exemplo da coleção “monster”

A tabela 4 apresenta o cálculo das métricas considerando apenas os registros.

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”

Neste caso, não foi coletado nenhum registro adicional. Portanto, a precisão foi 1.0. Entretanto, para a segunda página da coleção, alguns registros foram perdidos, ocasionando uma redução na revocação. A justificativa para isso é o fato de algumas entradas conterem um número menor de atributos, o que modifica a estrutura padrão dos registros.

A tabela 5 contém os valores de precisão e revocação considerando atributos.

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”

Todos os atributos coletados são relevantes. Mas, como já foi observado, alguns registros (e seus atributos) não foram coletados. Entretanto, a porcentagem de registros relevantes coletados foi alta.

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”

A tabela 6 contém os valores das médias das métricas por registro. A ausência de dados irrelevantes na coleta e a falha ao coletar todos os registros ficam evidentes neste cálculo.


### 3.2.3 Coleção “music\_allmusic.com”

A coleção “music\_allmusic.com” contém páginas com informações de artistas e álbuns. Cada detalhe de um álbum é um registro que contém como atributos informações como o ano de lançamento e o artista. A figura 8 é um exemplo de página da coleção “music\_allmusic.com”:



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

Figura 8: Página exemplo da coleção “music\_allmusic.com”

A tabela 7 apresenta os valores para as métricas considerando registros, enquanto em 8 estão os dados tratando atributos.

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\_allmusic.com”

O valor 1.0 de precisão evidencia que não foi coletado nenhum registro adicional (não relevante). Entretanto, alguns registros foram perdidos. A análise da página mostrou que alguns registros possuem uma estrutura irregular, prejudicando a qualidade da coleta. Apesar disso, os valores de precisão e revocação foram altos.

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\_allmusic.com”

O gabarito para esta coleção contém somente 3 atributos por registro. Entretanto, a primeira coluna de descrição de cada registro (*Tier*), também compõe o registro e deve ser considerada no cálculo.

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\_allmusic.com”

A tabela 9 contém os valores calculados da média de precisão e revocação por registro. Como nenhum registro irrelevante foi coletado e todos os registros de interesse foram extraídos, a precisão obtida é 1.0. Os valores de revocação foram elevados, já que, poucos atributos foram negligenciados pelo extrator.

### 3.2.4 Coleção “wines”

A coleção “wines” contém páginas como a da figura 9. Cada registro é formado pelos atributos que descrevem uma variedade de vinho (título, descrição, preço, etc).

The screenshot shows a web interface for a wine collection. On the left is a sidebar with a tree view of categories: 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, and Wine Team. The main content area has a 'Sort by:' dropdown set to 'Producer'. Below it is a pagination bar for 'Page 1 of 5' with links 1, 2, 3, 4, 5. The content is organized by producer: Barberani, Bartenura, and Bolognani. Each producer section lists wines with a small image icon, the wine name (year, producer, name, origin), a description, and the price. For example, under Barberani, there is '1998 Barberani Orvieto Classico Castagnolo, Umbria, Italy' for \$13.00 and '1998 Barberani Grechetto, Italy' for \$19.95. Each entry has 'Add to My' and 'Wish List' links. Under Bartenura, there is 'Bartenura Asti Spumante, Piedmont, Italy' for \$14.00. Under Bolognani, there is '1998 Bolognani Muller-Thurgau, Trentino, Italy' for \$13.00. At the bottom of the sidebar, there are two customer service links: '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' and 'To make your shopping easier, select your shipping destination in advance.'

Figura 9: Página exemplo da coleção “wines”

A tabela 10 apresenta os valores para as métricas considerando registros. Assim como ocorreu para a primeira coleção analisada, a precisão foi reduzida, enquanto a revocação foi elevada. O motivo foi o mesmo da falha da outra coleção: o menu lateral também foi extraído pelo algoritmo. Neste caso, o número de itens neste menu é superior ao número de registros na página, o que aumenta a probabilidade de sua extração. Por isso, a precisão obtida foi muito prejudicada (a página contém apenas 10 registros e 16 itens no menu lateral).

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”

Os dados da tabela 11 são referentes aos atributos e a tabela 12 contém o cálculo das métricas para a média por registro.

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”

O extrator teve seu pior desempenho para esta coleção. Além do menu conter mais itens que a lista de registros, os dados não estão bem formatados na página. Um exemplo disso é o fato de somente alguns registros possuírem um título adicional (acima de seu *link*).

## 4 Conclusão

O trabalho prático 1 consistiu no desenvolvimento de um método para extração automática de dados presentes em coleções de páginas HTML. Além da implementação, foi realizada uma análise experimental para validar a efetividade da proposta. Os dados da tabela 13 resumem os resultados obtidos:

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 por registro	0.77	1.0	1.0	0.38	0.79
Revocação pela média por registro	0.77	0.96	0.89	0.34	0.74

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

A tabela 13 contém a média dos valores obtidos para cada coleção, considerando cada métrica empregada. Também é apresentado um valor de média geral, que sumariza os resultados de todas as coleções. Os valores elevados de precisão e revocação, exceto para ocasiões específicas e justificáveis, sugerem que o método implementado é genérico e efetivo.

## Referências

- [1] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. In *Broadcast News Workshop 99 Proceedings*, pages 249–252. Morgan Kaufmann Pub, 1999.
- [2] 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.