UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE COMPUTAÇÃO

Projeto da Disciplina de Banco de Dados I - Relatório Final

29 de junho de 2012

Modelagem e Construção de um Banco de Dados de Ofertas de Produtos

Aluno: Daniel Bastos Moraes (RA 123530)

Professor: Prof. Dr. Ricardo Torres

Sumário

Capítulo 1

Introdução

Em sistemas de recuperação de informação, o que dita a maneira pela qual os resultados são ordenados são os critérios de relevância, compostos pelas informações utilizadas para definir o quão relevante um item é para uma determinada consulta. Tais critérios têm sido extensivamente estudados. No entanto, muitos desses estudos foram conduzidos em períodos anteriores ao dos *sites* de comércio eletrônico e estiveram mais direcionados na definição de critérios de relevância para buscas gerais da *Web*, e não para as finalidades específicas do comércio eletrônico.

Com isso, boa parte dos *sites* de comércio eletrônico utilizam estratégias pouco eficazes para ranquear os seus resultados, fazendo com que usuários percam tempo e esforço na busca das ofertas que realmente lhes interessariam. Quando se busca por um determinado produto em um *site* de comércio eletrônico, por exemplo, frequentemente são apresentadas ao e-consumidor somente opções de ordenação das ofertas a partir de critérios isolados, tais como: preço (do mais baixo ao mais alto, ou vice-versa); produto mais comprado; mais recente; entre outros [1]. É plausível pressupor que os *sites* que recorrem a esse tipo de estratégia o fazem por não possuírem formas mais eficazes de definir quais seriam as ofertas relevantes para o cliente. Dessa forma, fica a cargo do e-consumidor optar por um critério específico para ponderar a relevância das ofertas que lhe são apresentadas. Todavia, a relevância de um determinado produto no contexto de compra e venda não se dá perante um único critério (e.g., preço, popularidade ou data da oferta) [1].

É interessante destacar que, dada a crescente importância do comércio eletrônico e da pesquisa *online*, pesquisas que busquem desenvolver novos métodos de ranqueamento para sistemas de recuperação de informação do comércio eletrônico são necessárias. Tais métodos têm implicações tanto para os clientes quanto para as organizações comerciais que oferecem seus produtos para busca e compra *online* [2]. No entanto, apesar de já haverem trabalhos divulgados em meios científicos que apresentam estratégias de ranqueamento para anúncios e ofertas de produtos, a maior parte desses estudos esteve direcionada para o ranqueamento de *links* patrocinados.

Talvez um dos grandes motivos para o baixo número de trabalhos de ranqueamento

voltados para o comércio eletrônico, seja a dificuldade de se encontrar bancos de dados públicos e de qualidade. Em geral, esses bancos de dados são construídos internamente por empresas privadas, e não são disponibilizados publicamente.

Um dos maiores sistemas de recuperação de informação do comércio eletrônico é o *Google Product Search*, o qual contém ofertas de diversas categorias de produtos e de diferentes países. Apesar de o seu banco de dados não estar disponível publicamente, boa parte das informações de produtos, lojas e ofertas podem ser acessados publicamente no próprio *site* do *Google Product Search*. Essas informações são suficientes para a desenvolvimento de diversos trabalhos que busquem o desenvolvimento de novas estratégias de ranqueamento para o comércio eletrônico.

Nesse sentido, o objetivo deste trabalho é a modelagem e a construção de um banco de dados de ofertas de produtos, alimentado com as informações referentes a produtos, lojas e ofertas que estão disponíveis publicamente no *site* do *Google Product Search*. Este, por sua vez, poderá auxiliar o desenvolvimento de pesquisas voltadas para o desenvolvimento de estratégias de ranqueamento para comércio eletrônico, bem como diversos outros trabalhos que possam tirar proveito de um banco de dados desse tipo.

O restante deste texto está organizado da seguinte forma: o Capítulo 2 apresenta o método utilizado para a contrução do banco de dados proposto neste trabalho, já o Capítulo 3 destaca algumas análises efetuadas nos dados coletados. Finalmente e o Capítulo 4 discute os resultados e a conclusão do trabalho.

Capítulo 2

Método

Devido à falta de conjuntos de dados públicos com quantidades significativas de ofertas de produtos, este trabalho teve por objetivo construir, por meio da implementação de um *spider*, um banco de dados de ofertas de produtos.

Após uma breve avaliação em relação à qualidade dos dados apresentados por alguns dos principais *sites* de comércio eletrônico, constatou-se que, para as necessidades deste trabalho, os *sites* mais propícios para serem utilizados pelo *spider* foram os *sites* da *Amazon*[1] e do *Google Product Search*[2]. O *site* da *Amazon* descreve qualitativamente as características dos produtos com poucas inconsistências. Já o *Google Product Search*, possui uma grande coleção de ofertas (para determinadas classes de produtos). Além disso, como ilustrado na Figura 2.1, o *Google Product Search* reúne, em uma única página, ofertas de diversos *sites* de comércio eletrônico para um mesmo produto, isto é, as ofertas já estão vinculadas ao seu produto correspondente. Isso, por seu turno, facilitará bastante a construção do banco de dados, não sendo necessária a implementação de estratérias que agrupem ofertas de lojas diferentes para um mesmo produto.

O *spider* foi implementado para funcionar de forma ininterrupta, a fim de manter a base de dados de ofertas incremental e consistente no decorrer do tempo. Dessa forma, o *spider* busca manter o banco de dados o mais sincronizado possível com as informações referentes às ofertas de *laptops* do *Google Product Search*. Além disso, os estados anteriores do banco de dados estão sendo mantidos com granularidade diária.

Em relação às tecnologias utilizadas, o *spider* foi implementado em *Python*, por possuir facilidades no que diz respeito às necessidades de um *Web crawler*. Já o banco de dados utilizado para o armazenamento dos dados coletados foi o PostgreSQL.

[1] http://www.amazon.com
[2] http://www.google.com/shopping

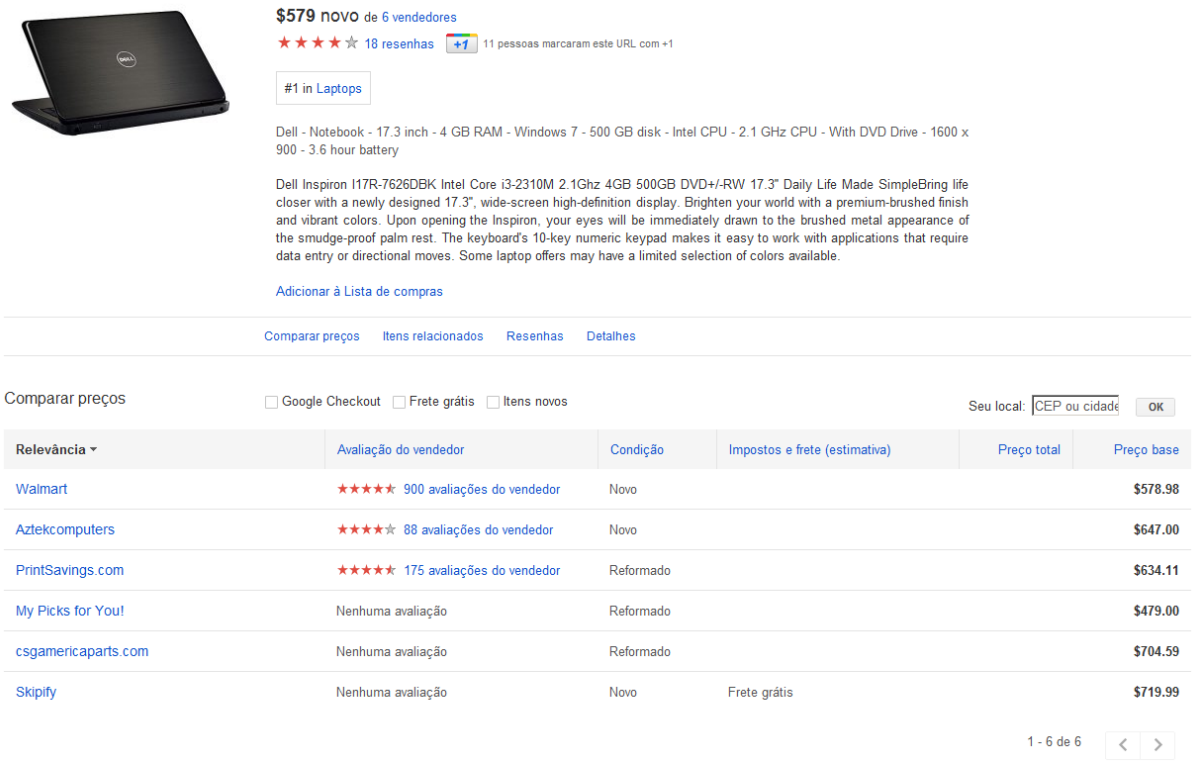


Figura 2.1: Um produto associado a diversas ofertas no *Google Product Search*.

2.1 Modelagem do banco de dados

O banco de dados foi modelado objetivando-se armazenar, de forma consistente, informações de produtos, lojas e ofertas contidas no *site* do *Google Product Search* que possam ser relevantes para pesquisas voltados para o desenvolvimento de novas estratégias de ranqueamento para o comércio eletrônico. Assim, alguns dos principais objetivos desse banco de dados são: (i) possuir uma quantidade espressiva de produtos, lojas e ofertas de produtos, (ii) possuir informações detalhadas dos produtos (características, popularidade, etc) e (iii) identificar produtos iguais associados à diferentes ofertas com um único identificador. Nesse sentido, foram modeladas as seguintes tabelas para o banco de dados:

- **product:** Armazena os identificadores dos produtos.
- **product_info:** Armazena informações relevantes dos produtos.
- **product_rating:** Armazena a classificação de cada produto em um determinado instante de tempo.
- **product_picture:** Armazena as figuras de cada produto.
- **product_rank:** Armazena o *rank* de um determinado produto (na lista de produtos de sua categoria) em um determinado instante de tempo.

- **category:** Armazena as categorias de produtos consideradas (ex: *laptop*).
- **feature_group:** Armazena os grupos de características de uma determinada categoria de produto (ex: "áudio e vídeo", para a categoria "*laptop*").
- **feature:** Armazena as características de um determinado grupo de características (ex: "processador gráfico", do grupo de caterísticas "áudio e vídeo").
- **product_spec:** Armazena as características de cada produto.
- **store:** Armazena informações relevantes das lojas.
- **store_rating:** Armazena a classificação de cada loja em um determinado instante de tempo.
- **offer:** Armazena informações relevantes das ofertas de produtos.

Um detalhe importante é o fato de as tabelas *product* e *product_info* terem sido separadas. Essa separação é devido ao processo de coleta. Como será mostrado mais adiante, o identificador do produto e as informações adicionais do produto são coletados por *spiders* diferentes, em instantes de tempo diferentes.

Na Figura (2.2), pode ser visto o Diagrama Entidade Relacionamento (DER) do banco de dados proposto.

2.2 Coleta dos dados

Inicialmente, o intuito foi coletar informações do *Google Product Search* com o uso de suas APIs. No entanto, as APIs disponibilizadas pelo *Google* se mostraram bastante limitadas, não trazendo grande parte das informações relevantes dos produtos. Diante disso, a implementação de um *spider*, que também coletasse tais informações relevantes, se mostrou necessária.

O *spider* visa coletar, de tempos em tempos, informações relevantes do *Google Product Search* e armazená-las em um banco de dados. As coletas são feitas por meio de consultas às páginas do *Google Product Search*, onde os dados são extraídos por meio de uma análise (*parsing*) do *HTML* resultante.

No caso deste trabalhos, os produtos coletados pelo *Spider* foram apenas os da categoria *laptop* do *Google Product Search*.

2.2.1 Google product search

Por ser essencialmente um motor de busca, no *Google Product Search* produtos só podem encontrados por meio de buscas baseadas em palavras-chave. Dessa maneira, para encontrar *laptops*, é necessário efetuar uma busca no *site* pelo termo "*laptop*". No caso específico de

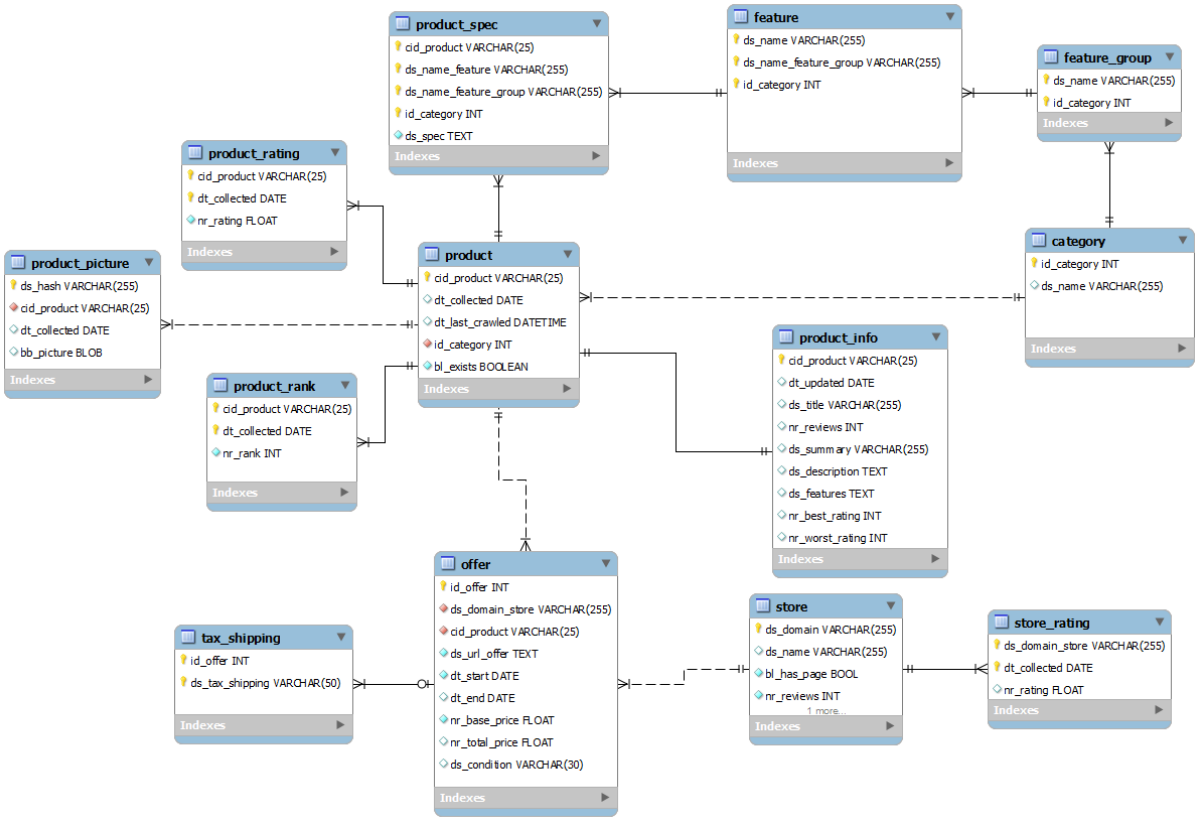


Figura 2.2: Modelagem do banco de dados utilizado pelo *Spider*.

laptop, por existir uma categoria com esse mesmo nome, os resultados provenientes da busca são exatamente os produtos que pertencem à categoria de mesmo nome.

A Figura 2.3 mostra um exemplo de uma página contendo os resultados de uma busca no *Google Product Search*. Estes resultados, por sua vez, são exibidos em páginas, com dez produtos por página e com um limite de mil produtos (ou 99 páginas) por busca. Nessas páginas, são apenas exibidas informações resumidas do produto, tais como título, foto e um preço médio de suas ofertas.

Além dessas informações, cada produto possui um *link* para uma página individual, contendo informações detalhadas do produto. Geralmente, algumas informações interessantes que podem ser encontradas nessa página individual são: um texto descritivo para o produto, a lista das lojas que possuem ofertas para esse produto e as características detalhadas do produto.

2.3 Spider

Um dos principais objetivos do *spider* é manter a base da dados o mais sincronizada possível com os dados contidos no *Google Product Search*, sendo importante que a estratégia de coleta

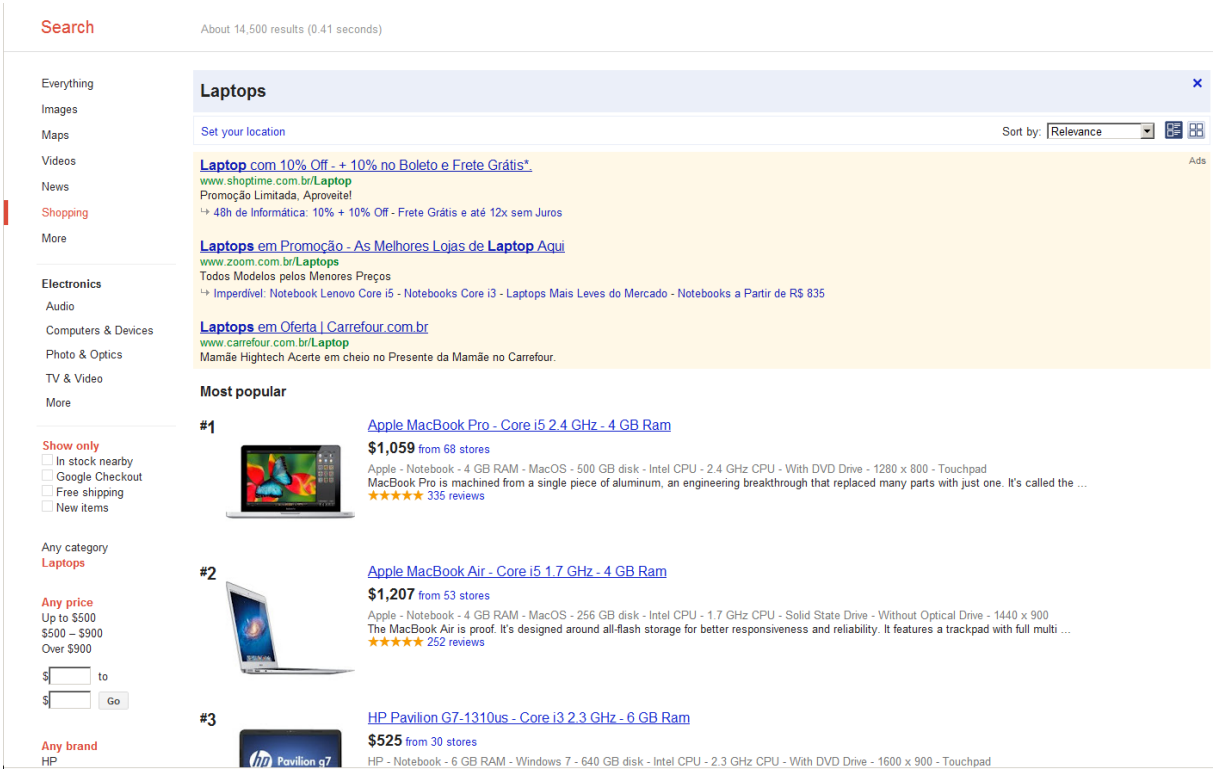


Figura 2.3: Resultado de uma consulta pelo termo *laptop* no *Google Product Search*.

seja eficiente. Assim, o intuito é coletar os produtos considerados em um curto espaço de tempo, reduzindo as chances de o *spider* não coletar algum produto que surgiu no intervalo entre duas coletas distintas.

Visando aumentar a eficiência desse processo e garantir que boa parte das ofertas que surgirem no *Google Product Search* sejam coletadas, o *spider* foi decomposto em dois processos independentes: (i) *Product ID Spider* e (ii) *Product Data Spider*. O primeiro módulo visa coletar apenas os identificadores dos produtos, disponíveis nas páginas dos resultados da consulta no *Google Product Search*. Já o segundo, tem por objetivo coletar informações detalhadas de cada produto que foi previamente identificado na etapa (i). Essa separação garante uma maior eficiência na coleta dos idenficadores dos produtos, pois, o *Product ID Spider* se restringe apenas a navegar nos resultados de busca *Google Product Search*, sem acessar as páginas individuais de cada produto. Essas páginas, por sua vez, são apenas acessadas mais tarde pelo *Product Data Spider*.

O processo de coleta do *Spider* é ilustrado na Figura 2.4, onde o *Product ID Spider* coleta, de tempos em tempos, os identificadores dos produtos disponíveis do *Google Product Search* e o *Product Data Spider* coleta, também de tempos em tempos, informações detalhadas dos produtos que foram previamente identificados pelo *Product ID Spider*.

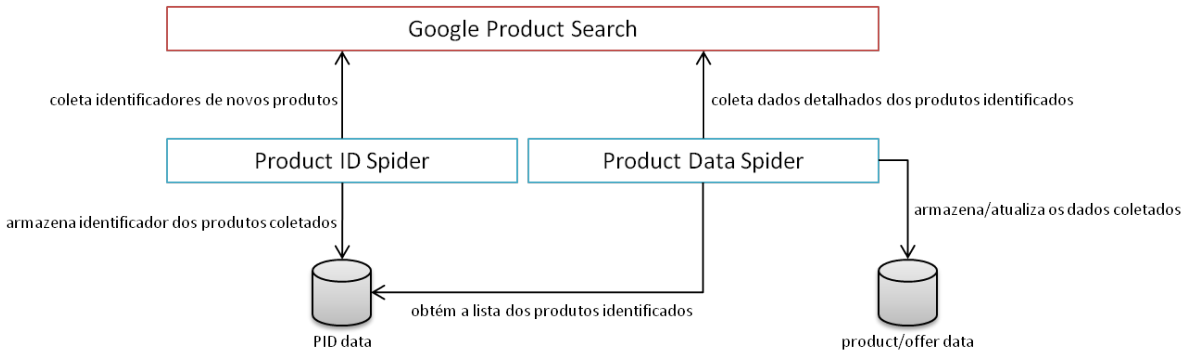


Figura 2.4: Ilustração do processo de coleta do Product ID Spider e do Product Data Spider.

2.3.1 Product ID *spider*

O *Product ID Spider* tem por objetivo coletar os identificadores dos produtos que estiveram disponíveis no *Google Product Search* em um determinado instante de tempo. Como neste trabalho os produtos considerados são essencialmente os da categoria *laptop*, os identificadores são coletados com base em consultas no *Google Product Search* pelo termo *laptop*.

O primeiro desafio enfrentado na implementação do *Product ID Spider* foi devido ao limite superior de mil resultados por consulta imposto pelo *Google Product Search*. Com a existência dessa limitação, não é possível coletar os identificadores de todos os produtos por meio de uma simples consulta pelo termo *laptop*.

Contudo, esse problema foi resolvido por meio da utilização de filtros de pesquisa, que são disponibilizados pelo *Google Product Search* para filtrar os resultados de uma consulta com base em uma determinada característica. Na Figura 2.3 são mostrados alguns filtros que estão disponíveis para a categoria *laptop*. Assim, o *Product ID Spider* faz uso de filtros baseados em intervalos de preço, onde se define limitantes de preço inferior e superior para os resultados da consulta. O intuito disso é definir uma sequência de intervalos de preços válidos que contenham todos os resultados da consulta, onde cada intervalo não ultrapasse mil resultados. Assim, uma vez definido tais intervalos, basta que o *Product ID Spider* efetue uma sequência de consultas para cada um dos intervalos definidos e colete os identificadores de *laptops* retornados em cada um deles.

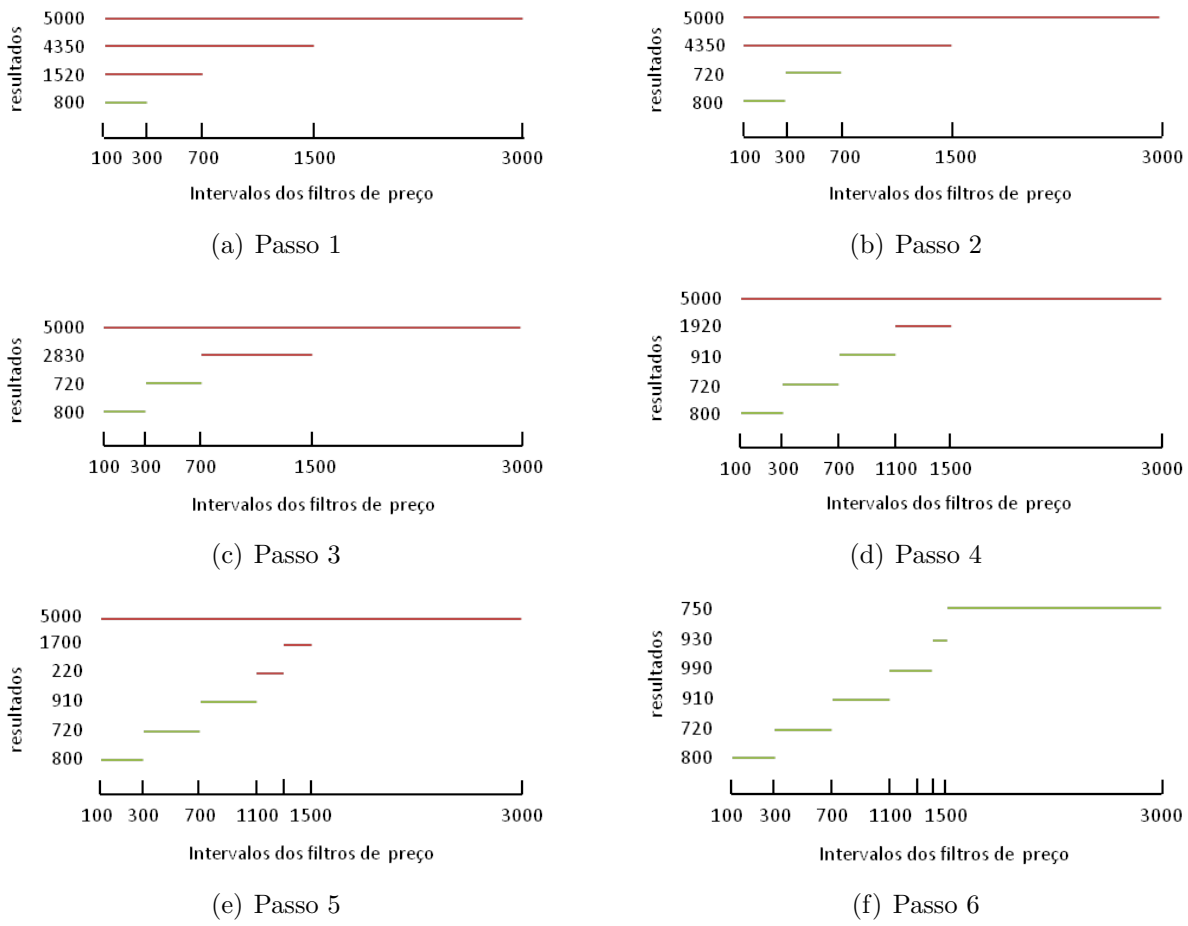
Dessa forma, o *Product ID Spider* faz uso de um algoritmo para construção desses intervalos. O algoritmo faz uso de dois parâmetros: (i) o limitante superior de resultados que, no *Google Product Search* deve ser de mil resultados; (ii) um limitante inferior que define o número mínimo de resultados permitido em um intervalo de preço, objetivando-se definir intervalos que contenham quantidades significativas de resultados.

A Figura 2.5 ilustra um exemplo de execução do algoritmo para a construção deintervalos de preço válidos. Nesse exemplo, os preços dos produtos variam entre $100 a $3000 e os limitantes inferior e superior de resultados são de 300 e 1000, respectivamente. O algoritmo se baseia basicamente em dois procedimentos: (i) *break*, que visa dividir ao meio um deter-

minado intervalo que excede o limite superior de 1000 resultados; (ii) *join*, que visa ampliar, por meio da junção com parte do intervalo subsequente, um determinado intervalo que possui menos de 300 (limite inferior) resultados. Inicialmente, no Passo 1, o algoritmo detectou um total de 5000 resultados e executou uma sequência de *breaks* até que chegou no intervalo \$100-\$300, com 800 resultados. Como esse intervalo também possui mais de 300 resultados, ele é definido como um intervalo válido. A partir desse ponto, o algoritmo buscará definir o próximo intervalo válido, com início em \$301. Por meio da sequência de *breaks* efetuada anteriormente, o intervalo \$100-\$700 já foi conhecido, com um total de 1520 resultados. No entanto, como o intervalo \$100-\$300 também já foi definido, o intervalo \$301-\$700 pode ser facilmente computado por meio da subtração entre os dois. Assim, no Passo 2, o verificou-se que o intervalo \$301-\$700 possui um total de $1520 - 800 = 720$ resultados, tornando-o, então, o segundo intervalo válido da sequência. No Passo 3, o intervalo \$701-\$1500 foi definido de forma semelhante ao que foi feito anteriormente no Passo 2. No entanto, como esse intervalo excede o limite superior de resultados, no Passo 4 ele foi dividido ao meio (*break*) nos intervalos \$701-\$1100 e \$1101-1500, onde o primeiro foi definido como o terceiro intervalo válido, com um total de 910 resultados. No Passo 5, o intervalo \$1101-\$1300 não excede o limite superior. No entanto, esse não é um intervalo válido por possuir menos de 300 resultados. Nesse caso, o algoritmo efetua um *join* com metado do intervalo subsequente. Por fim, após a junção, no Passo 6, os intervalo \$1101-\$1400, \$1401-\$1500 e \$1501-\$5000 são os últimos intervalos válidos da sequência.

2.3.2 Product data *spider*

O *Product Data Spider* coleta informações detalhadas dos produtos identificados pelo *Product ID Spider*. Dentre as informações coletadas pelo *Product Data Spider*, estão imagens do produto, lojas que possuem ofertas para o produto e algumas de suas características. Essas e outras informações do produto são coletadas por meio da página individual de cada produto no *Google Product Search*. Um exemplo de uma página contendo a listagem das lojas e das características de um produto podem ser vistas nas Figuras 2.6 e 2.7, respectivamente. Todas as informações relevantes contidas na página individual do produto serão coletadas, por meio da análise (*parsing*) do *HTML* resultante.



(a) Passo 1

(b) Passo 2

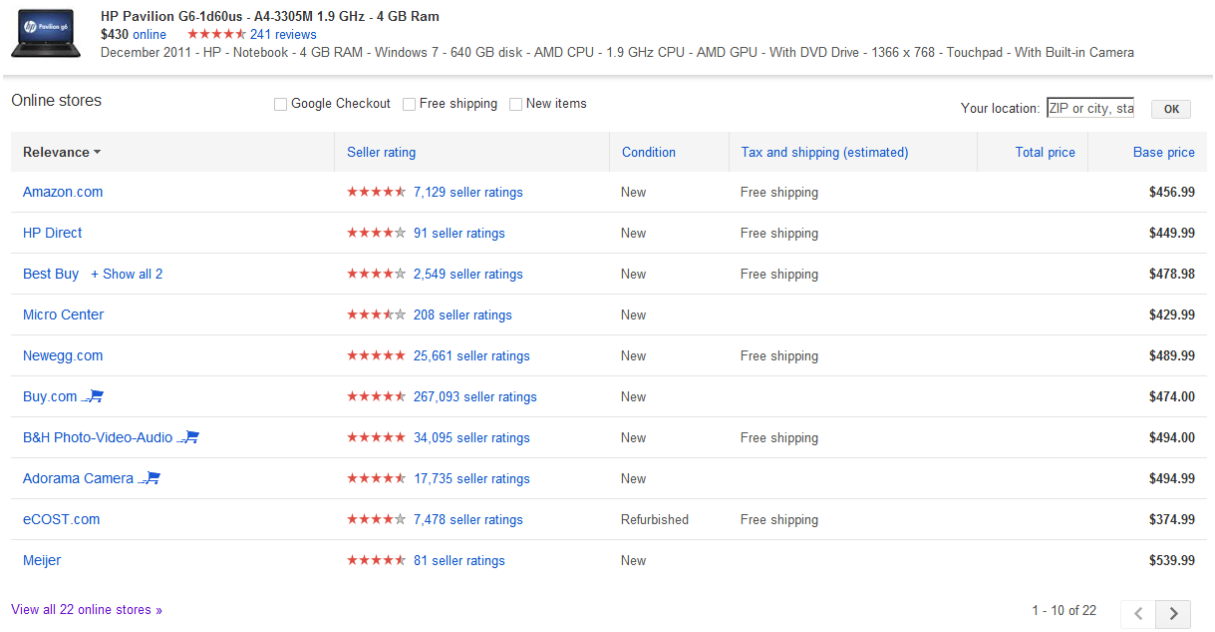
(c) Passo 3

(d) Passo 4

(e) Passo 5

(f) Passo 6

Figura 2.5: Ilustração do processo de definição dos filtros de preço.



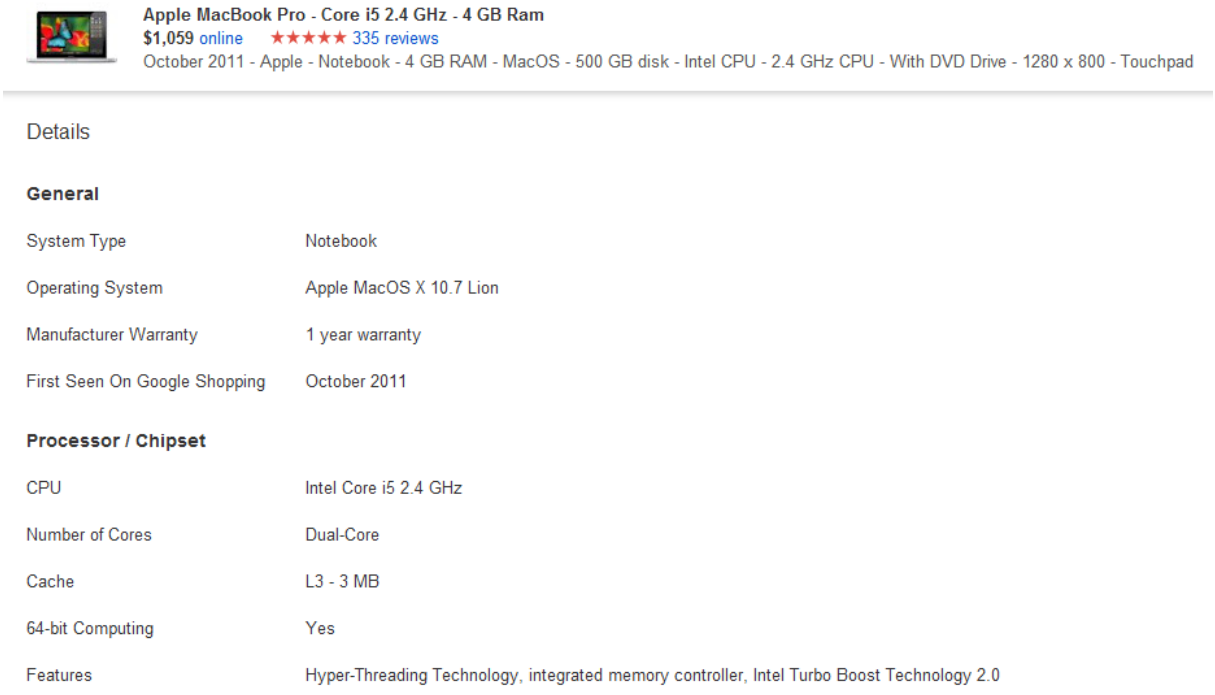
HP Pavilion G6-1d60us - A4-3305M 1.9 GHz - 4 GB Ram
$430 online ★★★★½ 241 reviews
December 2011 - HP - Notebook - 4 GB RAM - Windows 7 - 640 GB disk - AMD CPU - 1.9 GHz CPU - AMD GPU - With DVD Drive - 1366 x 768 - Touchpad - With Built-in Camera

Online stores ☐ Google Checkout ☐ Free shipping ☐ New items Your location: ZIP or city, sta OK

Relevance	Seller rating	Condition	Tax and shipping (estimated)	Total price	Base price
Amazon.com	★★★★½ 7,129 seller ratings	New	Free shipping		$456.99
HP Direct	★★★★☆ 91 seller ratings	New	Free shipping		$449.99
Best Buy + Show all 2	★★★★☆ 2,549 seller ratings	New	Free shipping		$478.98
Micro Center	★★★½☆ 208 seller ratings	New			$429.99
Newegg.com	★★★★★ 25,661 seller ratings	New	Free shipping		$489.99
Buy.com	★★★★½ 267,093 seller ratings	New			$474.00
B&H Photo-Video-Audio	★★★★★ 34,095 seller ratings	New	Free shipping		$494.00
Adorama Camera	★★★★½ 17,735 seller ratings	New			$494.99
eCOST.com	★★★★☆ 7,478 seller ratings	Refurbished	Free shipping		$374.99
Meijer	★★★★½ 81 seller ratings	New			$539.99

View all 22 online stores » 1 - 10 of 22

Figura 2.6: Lojas que possuem ofertas para o produto “HP Pavilion G6-1d60us”.



Apple MacBook Pro - Core i5 2.4 GHz - 4 GB Ram
$1,059 online ★★★★★ 335 reviews
October 2011 - Apple - Notebook - 4 GB RAM - MacOS - 500 GB disk - Intel CPU - 2.4 GHz CPU - With DVD Drive - 1280 x 800 - Touchpad

Details

General

System Type	Notebook
Operating System	Apple MacOS X 10.7 Lion
Manufacturer Warranty	1 year warranty
First Seen On Google Shopping	October 2011

Processor / Chipset

CPU	Intel Core i5 2.4 GHz
Number of Cores	Dual-Core
Cache	L3 - 3 MB
64-bit Computing	Yes
Features	Hyper-Threading Technology, integrated memory controller, Intel Turbo Boost Technology 2.0

Figura 2.7: Características do produto “Apple MacBook Pro”.

Capítulo 3

Análise dos resultados

O *Product ID Spider* têm coletado identificadores de *laptop* do *Google Product Search* nos meses de maio e junho. Neste momento, já foi coletado um total de 23500 identificadores únicos de *laptops*. A Figura 3.1 mostra a variação no total de identificadores coletados por período, nos últimos 60 dias. O grande salto no total de identificadores coletados no dia 21 de junho foi devido a uma otimização efetuada no *Product ID Spider* nesse mesmo dia.

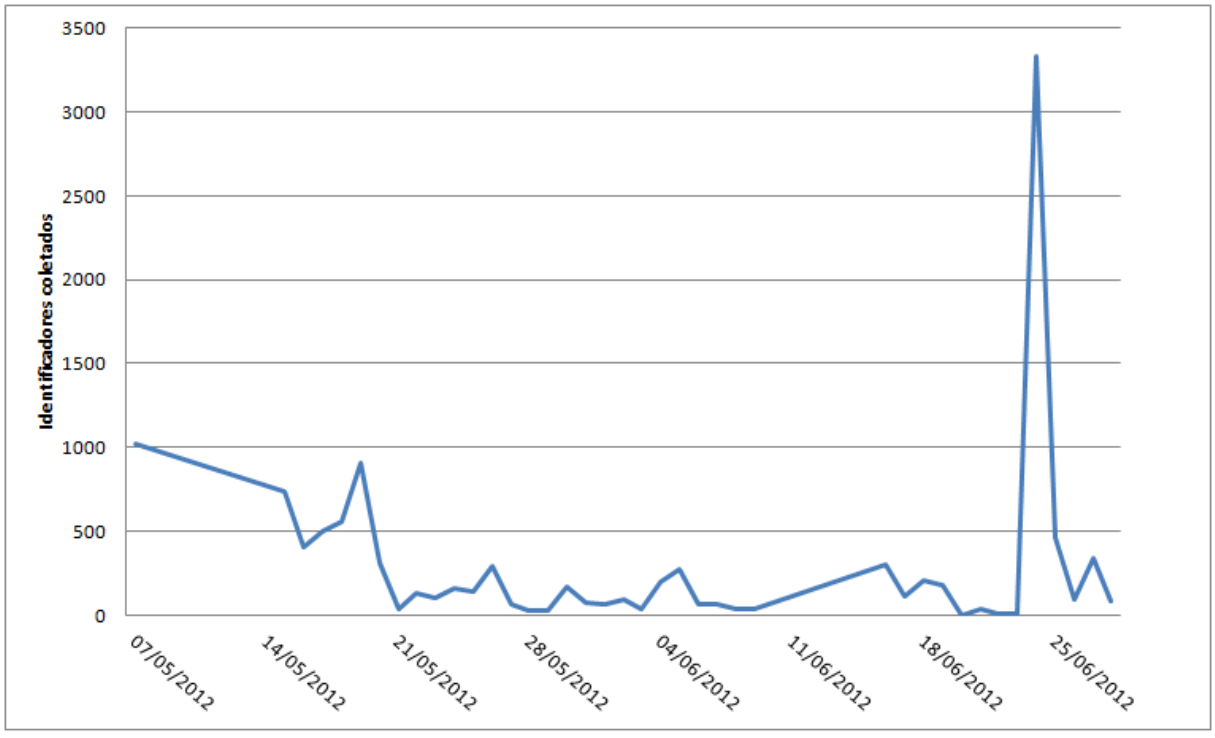


Figura 3.1: Total de identificadores de *laptops* coletados por período.

A Figura 3.2 mostra a variação no total de ofertas coletadas por período, desde a primeira execução do *Product Data Spider*. O salto no total de identificadores coletados no dia 25 de junho foi devido a uma otimização efetuada no *Product Data Spider*. Desconsiderando

o pico inicial, o *Product Data Spider* têm identificado diariamente uma média de 400 novas ofertas.

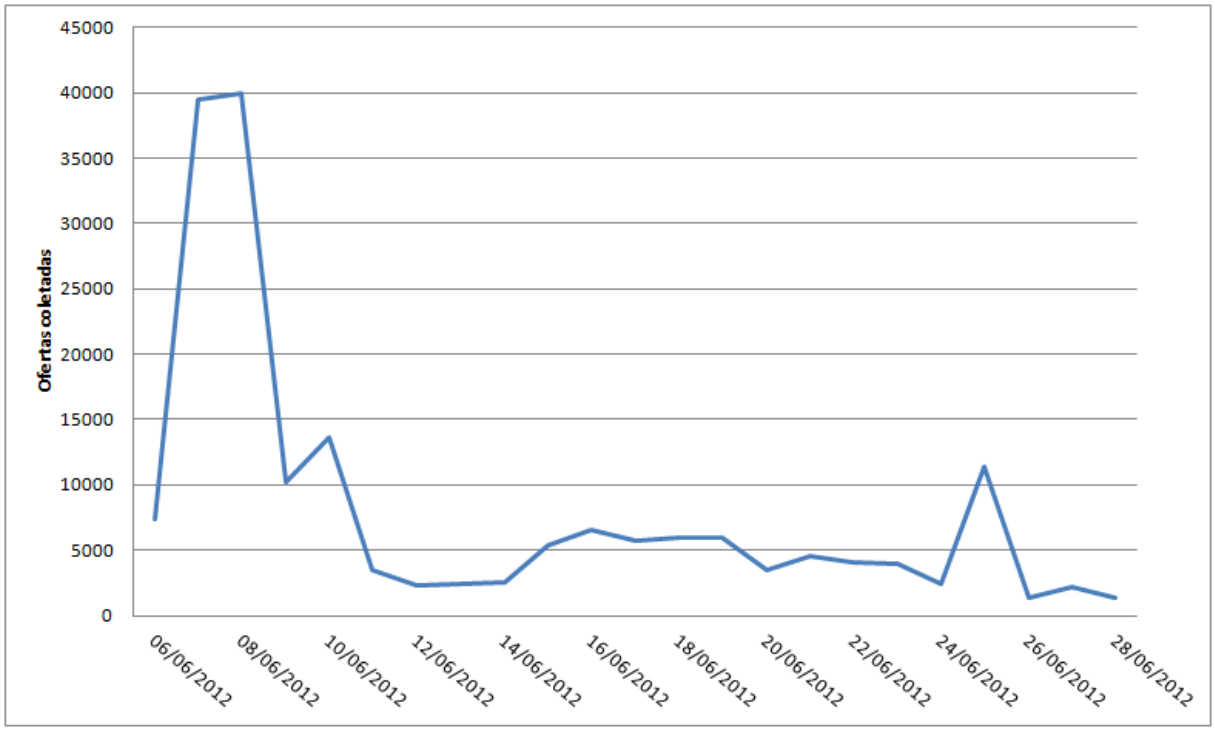


Figura 3.2: Total de ofertas coletadas por período.

3.1 Especificações dos *laptops*

Foram observados diversos problemas nas especificações dos produtos coletados. Em geral, os maiores problemas podem ser definidos em três tipos: (i) *laptops* com especificações iguais, mas escritas de maneira diferente; (ii) especificações com informações multivaloradas; (iii) especificações escritas de maneira incorreta.

A Figura 3.1 lista as vinte marcas mais presentes nos produtos coletados. Nessa caso, o problema do tipo (i) aconteceu em relação à marca *HP*, escritas de duas maneiras diferentes: "Hewlett-Packard" e "HEWLETT PACKARD / HP". Um outro problema é em relação à marca *Shopforbattery*. Como essa é uma marca de baterias, pôde-se perceber que baterias de *laptops* também estavam presentes na categoria de *laptop* do *Google Product Search*. No caso dessas baterias, as suas categorias no banco de dados foram ajustadas para a categoria correspondente.

Nas especificações referentes aos processores (Figura 3.2), pôde ser observado problemas do tipo (ii). Nesse caso, os processadores possuem informações mulivaloradas, incluindo marca, família, modelo e frequência. Um problema semelhante foi observado nas especificações de memória (Figura 3.3), de sistema operacional 3.4, de processador gráfico 3.5 e

de capacidade de disco rígido 3.6. Esse tipo de problema pode ser facilmente resolvido por meio da decomposição do atributo. Os problemas do tipo (iii), apareceram de forma menos frequentemente.

Tabela 3.1: Marcas mais frequentes nos dados coletados.

Nome	**Total**
Toshiba	2510
Hewlett-Packard	2452
Lenovo	2304
Sony	1814
HP	1095
Shopforbattery	946
Acer	801
Dell	743
ASUSTeK COMPUTER	621
Panasonic	413
DELL	263
TOSHIBA	244
Fujitsu	243
Apple	240
HEWLETT PACKARD / HP	239
Gateway	238
Samsung	217
MSI	195
ASUS	187
IBM	179

Tabela 3.2: Processadores mais frequentes nos dados coletados.

Nome	Total
Intel Core i5 2520M / 2.5 GHz	376
Intel Core i5 520M / 2.4 GHz	323
Intel Core 2 Duo P8400 / 2.26 GHz	185
Intel Core 2 Duo P8600 / 2.4 GHz	170
Intel Core i5 2410M / 2.3 GHz	167
Intel Core i3 370M / 2.4 GHz	159
Intel Core i3 2310M / 2.1 GHz	153
Intel Core i3 350M / 2.26 GHz	149
Intel Core i5 2540M / 2.6 GHz	142
Intel Core i5 540M / 2.53 GHz	141
Intel Core i5 2450M / 2.5 GHz	132
Intel Core i5 2430M / 2.4 GHz	128
Intel Core i7 2630QM / 2 GHz	121
Intel Pentium M 740 / 1.73 GHz	118
Intel Core i7 2670QM / 2.2 GHz	117
Intel Core 2 Duo P8700 / 2.53 GHz	112
Intel Core i3 2350M / 2.3 GHz	111
Intel Core 2 Duo T7300 / 2 GHz	109
Intel Core i3 330M / 2.13 GHz	108
Intel Core i7 2620M / 2.7 GHz	108

Tabela 3.3: Capacidades de memória mais frequentes nos dados coletados.

Nome	Total
4 GB (2 x 2 GB)	1583
4 GB	1141
2 GB (1 x 2 GB)	998
512 MB	792
4 GB (1 x 4 GB)	766
3 GB (1 x 1 GB + 1 x 2 GB)	650
1 GB (1 x 1 GB)	630
2 GB	624
2 GB (2 x 1 GB)	502
1 GB	497
256 MB	456
1 GB (2 x 512 MB)	345
6 GB (1 x 4 GB + 1 x 2 GB)	228
8 GB (2 x 4 GB)	178
3 GB	176
6 GB	129
512 MB (2 x 256 MB)	115
512 MB (1 x 512 MB)	101
128 MB	90
8 GB	89

Tabela 3.4: Sistemas operacionais mais frequentes nos dados coletados.

Nome	Total
Microsoft Windows 7 Home Premium 64-bit Edition	2318
Microsoft Windows XP Professional	1325
Microsoft Windows 7 Professional 64-bit Edition	1250
Microsoft Windows Vista Home Premium	883
Microsoft Windows XP Home Edition	750
Microsoft Windows 7 Professional	611
Microsoft Windows Vista Business	545
Microsoft Windows Vista Business / XP Professional downgrade	348
Microsoft Windows 7 Professional / XP Professional downgrade	332
Microsoft Windows 7 Starter	265
Microsoft Windows Vista Home Premium 64-bit Edition	248
Microsoft Windows 7 Home Premium	221
Microsoft Windows Vista Home Basic	171
Microsoft Windows 7 Professional (32/64 bits)	111
Microsoft Windows XP Media Center Edition 2005	82
Microsoft Windows 2000	69
Microsoft Windows 7 Professional 32-bit	59
Microsoft Windows 98 Second Edition	50
Microsoft Windows XP Media Center Edition	50
Microsoft Windows Vista Ultimate	47

Tabela 3.5: Processadores gráficos mais frequentes nos dados coletados.

Nome	Total
Intel HD Graphics	1300
Intel HD Graphics 3000	1225
Intel GMA 4500MHD	931
Intel GMA X3100	630
Intel GMA 950	561
Intel GMA 900	256
Intel GMA 3150	245
ATI Mobility Radeon HD 4250	224
Intel GMA 4500M	216
Intel Extreme Graphics 2	197
AGP 4x - ATI Mobility Radeon 7500 - 32 MB DDR SDRAM	89
ATI Radeon Xpress 200M	86
NVIDIA Quadro NVS 150M - 256 MB	63
AMD Radeon HD 6310	59
ATI Radeon HD 3200	59
PCI Express x16 - NVIDIA NVS 4200M / Intel HD Graphics 3000 - 1 GB	56
ATI Radeon X1200	54
NVIDIA GeForce 8200M G	52
ATI Radeon X1250	50
ATI Mobility Radeon HD 3200	49

Tabela 3.6: Capacidades de disco rígido mais frequentes nos dados coletados.

Nome	Total
320 GB HDD / 5400 rpm	1047
320 GB HDD / 7200 rpm	1043
250 GB HDD / 5400 rpm	936
500 GB HDD / 5400 rpm	852
160 GB HDD / 5400 rpm	752
500 GB HDD / 7200 rpm	497
80 GB HDD / 5400 rpm	406
120 GB HDD / 5400 rpm	377
640 GB HDD / 5400 rpm	303
160 GB HDD / 7200 rpm	289
128 GB SSD	286
160 GB HDD	229
250 GB HDD / 7200 rpm	193
60 GB HDD / 5400 rpm	189
40 GB HDD / 4200 rpm	186
40 GB HDD	183
60 GB HDD / 4200 rpm	178
40 GB HDD / 5400 rpm	175
750 GB HDD / 5400 rpm	174
100 GB HDD / 5400 rpm	158

Capítulo 4

Conclusão

Foi possível perceber que o *Google Product Search* é uma ótima fonte com informações de lojas, produtos e ofertas, onde os dados são bastante dinâmicos, surgindo diariamente dezenas de novos *laptops* e centenas de novas ofertas. Além disso, o *Google Product Search* possui algoritmos que agrupam ofertas de um mesmo produto em uma única página, o que facilitou bastante a coleta dos dados. A quantidade de informações presentes na página de cada *laptop* e a qualidade dos dados impressiona, sendo bastante superior à maioria dos demais *sites* de comércio eletrônico investigados.

Após uma análise mais detalhada dos dados, foram constatados alguns problemas: (i) especificações iguais, mas escritas de maneira diferente; (ii) especificações com informações multivaloradas; (iii) especificações escritas de maneira incorreta. No entanto, esses problemas podem ser facilmente corrigidos por meio de algumas alterações no banco de bancos, discutidas na Seção 3.

Nesse sentido, o banco de dados construído neste trabalho poderá auxiliar o desenvolvimento de pesquisas voltadas para o desenvolvimento de estratégias de ranqueamento para comércio eletrônico, bem como diversos outros trabalhos que possam tirar proveito de um banco de dados com quantidades significativas de informações referentes a produtos, lojas e ofertas.

Referências

[1] Marco Gori and Augusto Pucci. A random-walk based scoring algorithm with application to recommender systems for large-scale e-commerce. In *ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, 2006.

[2] Bernard J. Jansen and Paulo R. Molina. The effectiveness of web search engines for retrieving relevant ecommerce links. *Information Processing and Management*, pages 1075–1098, July 2006.