5
Extraindo listas de produtos em sites de comércio eletrônico

Existem diversos trabalhos direcionadas à detecção de listas e tabelas na literatura como (Liu et. al., 2003, Tengli et. al., 2004, Krüpl et. al., 2006, Zhai et. al., 2005). No entanto, a capacidade de adaptação dos modelos existentes para que passem a identificar uma lista específica é uma dúvida que surge quando analisamos as abordagens propostas. Neste capítulo é apresentada a tarefa de identificação de listas de produtos em sites de comércio eletrônico. São utilizados os algoritmos que identificam listas, apresentados no Capítulo 3, junto a um pequeno conjunto de regras específicas para a resolução dessa tarefa.

As listas de produtos foram escolhidas pois os itens apresentados nessas listas são variados, ou seja, em cada site de comércio eletrônico é possível observar itens com diferentes informações. Essa variação na informação se reflete na estrutura HTML dos itens, o que dificulta a identificação desse tipo específico de lista. Porém, mesmo com toda essa variação é possível observar um padrão interessante, onde as listas, normalmente, apresentam um resumo das informações dos produtos como preço, disponibilidade e promoção. Com isso, ao identificar uma lista de produtos é possível obter todas as informações contidas no resumo, sem a necessidade de visitar a página de cada produto da lista. Por esse motivo, acreditamos que identificar as listas de produtos possa ser uma capacidade que além de agregar informação às tarefas de extração de informação, também proporcione benefícios como a diminuição do uso de rede e do volume de processamento destinado aos documentos HTML.

Para ilustrar uma lista de produtos, na Figura 5.1 é apresentada uma lista do site bestbuy.com, onde as informações são exibidas linha a linha. É interessante observar os elementos utilizados para a criação dessa lista, já que esses elementos serão utilizados para a identificação. Note que entre cada item, apresentado por uma elemento “div”, existem dois elementos, um “hr” e um “script”, criando conjuntos generalizadores de tamanho três (com três elementos). Já na lista de produtos do site Americanas.com, ilustrada na Figura 5.2, os itens são organizados lado a lado, não existindo nenhum elemento de separação. É interessante notar que nesse último site, é utilizado o elemento

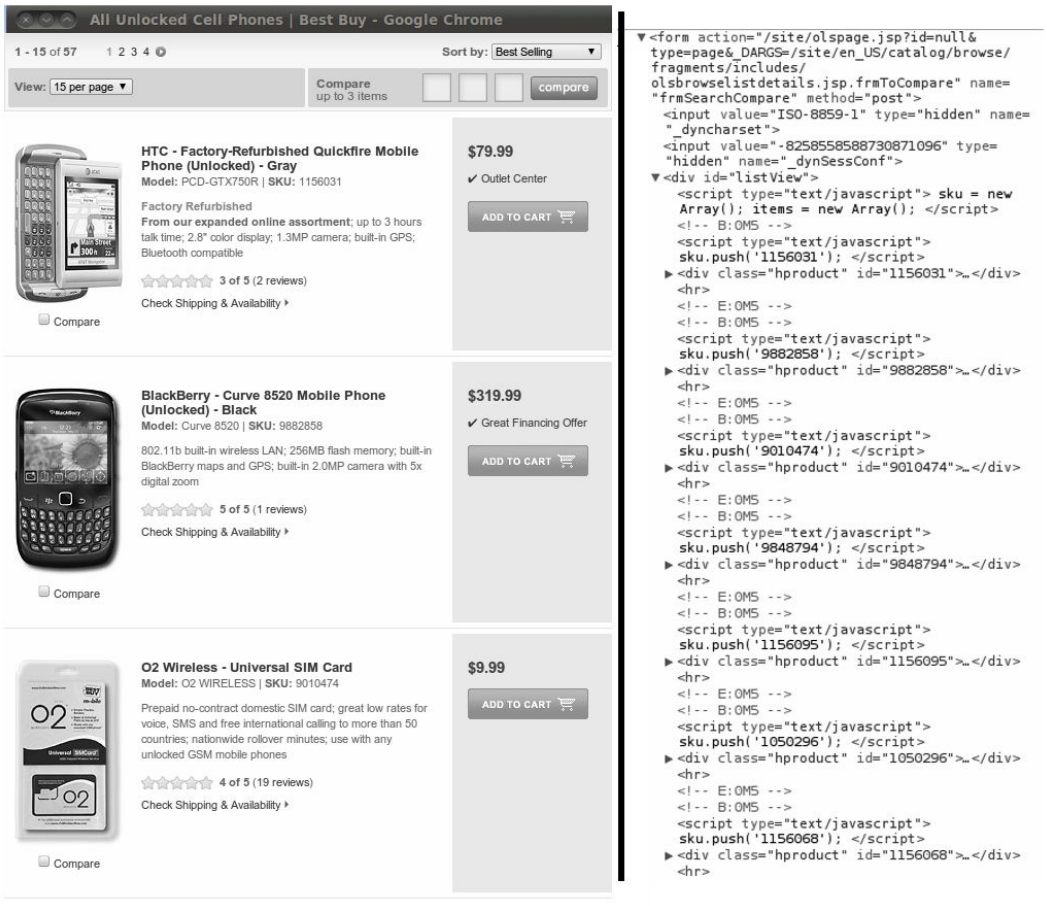

Figura 5.1: Exemplo de uma lista de produtos do site bestbuy.com

HTML "li", que é destinado à criação de listas.

5.1 Trabalhos existentes

A falta de trabalhos com o objetivo de identificar listas de produtos dificulta a comparação de nossa abordagem. Por esse motivo, utilizamos os trabalhos que têm como objetivo a identificação de listas genéricas para realizar algumas comparações. Utilizamos esses trabalhos pois eles também diminuem o domínio de busca, assim como a abordagem estrutural implementada nesta dissertação. Além disso, ambas as abordagens permitem a criação de pós-processamentos para serem aplicados a um conjunto reduzido de elementos.

Os trabalhos (Liu et. al., 2003) e (Zhai et. al., 2005) são interessantes do ponto de vista da redução do domínio de busca, pois nessas referências são identificadas regiões visuais que apresentam algum tipo de lista. Como pode ser observado, os procedimentos descritos no Capítulo 3 são baseados nesses dois trabalhos, sugerindo que as etapas específicas utilizadas neste capítulo para identificar as listas de produtos poderiam ser replicadas nos dois trabalhos

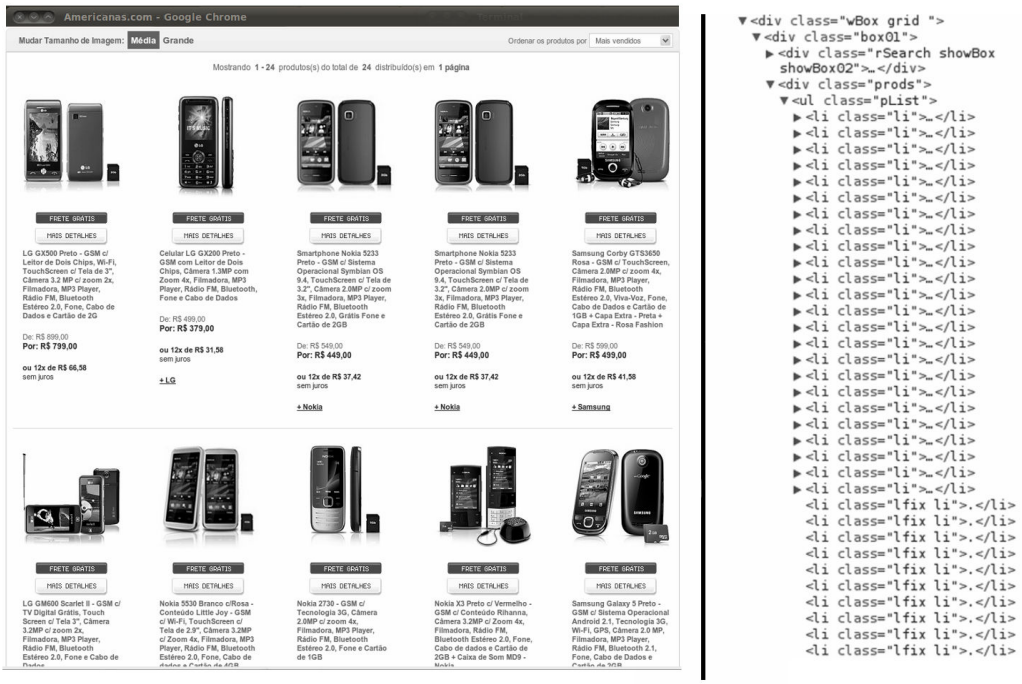

Figura 5.2: Exemplo de uma lista de produtos do site americanas.com

citados. No entanto, ao modificar os algoritmos que criam as listas, podem existir variações nos resultados, pois a forma com que as listas são geradas influencia diretamente nos resultados. Essa variação pode ser notada durante nossa experimentação e será descrita na Seção 5.4.

Existem também trabalhos de segmentação como (Cai et. al., 2003) e (Chakrabarti et. al., 2008), que têm como objetivo dividir um documento HTML em vários segmentos visuais, sendo uma abordagem popular dentre os trabalhos de extração de informação. Tais abordagens também poderiam ser utilizadas para a identificação de listas de produtos, sendo necessária apenas a criação de um classificador com o objetivo de identificar as listas de produtos dentre os segmentos. No entanto, tanto em (Chakrabarti et. al., 2008) quanto em (Cai et. al., 2003) são utilizadas diversas informações visuais, como tamanho de um bloco de informação e a posição (x,y) de um elemento na tela, indo em direção contrária à desejada nesta dissertação que é obter informações da estrutura do documento. A utilização de informações visuais torna necessário mais tempo de processamento por documento, já que, para a obtenção de informações visuais, é necessário renderizar o documento, como foi discutido no Capítulo 2.

Um ponto importante a ser observado é que em (Chakrabarti et. al., 2008) o conjunto de atributos utilizados é separado em dois. O primeiro grupo é destinado aos atributos visuais, que necessitam da renderização. O segundo é

o grupo chamado de *content-based*, que corresponde aos atributos obtidos a partir da árvore DOM. No entanto, a referência não apresenta os resultados de forma que seja possível saber qual a contribuição de cada conjunto de atributos.

5.2 Corpus de exploração

Para o desenvolvimento de técnicas de identificação de listas de produtos e a avaliação da qualidade, foi criado um conjunto com 1114 documentos de 8 sites de comércio eletrônico, sendo:

- 118 documentos do site da americanas.com;
- 147 documentos do site da bestbuy.com;
- 150 documentos do site circuitcity.com;
- 168 documentos do site dealextream.com;
- 168 documentos do site ebay.com;
- 116 documentos do site mysimon.com;
- 108 documentos do site submarino.com.br;
- 139 documentos do site target.com.

Para obter esses documentos foi utilizado um procedimento automático, o que tornou o processo mais ágil. Basicamente foram criadas regras de XPath[1] junto a um *crawler*, criando um anotador automático, onde o nó que enraíza a lista de produtos ganha um atributo e é salvo. Em cada site de comércio eletrônico escolhido, foi avaliado manualmente qual nó DOM deveria ser classificado como lista de produtos. Foi definido que o nó classificado deve conter todos os itens da lista de produtos. Esse nó foi marcado como "product_list". Além disso, esse nó recebeu um atributo chamado "proof_annotation" com valor "product_list".

Um fato importante sobre a construção do conjunto de experimentação é que não foi feito nenhum tipo de pós-processamento para excluir listas que contêm apenas um produto, ou mesmo vazias. Por esse motivo, em alguns casos a abordagem proposta não identificou a lista de produtos, já que os algoritmos de detecção de lista assumem que uma lista é formada por pelo menos dois elementos.

Para permitir a observação de alguns padrões e também realizar a experimentação de forma que os algoritmos não fiquem viciados no corpus,

[1] http://www.w3.org/TR/xpath/

o conjunto de documentos foi dividido. Foram separados os documentos dos sites americanas.com, target.com e dealextream.com para o ajuste de alguns parâmetros necessários e avaliação da corretude dos algoritmos. Esse conjunto de treino também foi utilizado para a geração das regras especializadas em listas de produtos. O restante dos documentos foram separados para serem realizados os testes.

5.3 Métrica

Utilizamos métricas iguais às apresentadas na Seção 4.2, sendo feitas apenas algumas modificações que são descritas a seguir. Essas métricas foram escolhidas, pois elas são usualmente adotadas dentre os trabalhos que têm o objetivo de identificar um elemento na árvore DOM.

Na Tabela 5.3, são exemplificados os casos que podem ocorrer durante a classificação de um nó. **VP** é o número de nós que são classificados como lista de produtos corretamente. **FN** é o número de nós que são classificados como "outro" e deveriam ser lista de produtos. **FP** é o número de nós que são classificados como lista de produtos e deveriam ser classificados como "outro". Finalmente, **VN** é o número de nós "outro" que são classificados corretamente.

Classe correta	Classificado como lista de produto	Classificado como "outro"
lista de produto	VP	FN
"outro"	FP	VN

Tabela 5.1: Classificações possíveis de uma tabela para o cálculo das métricas

Com isso, é possível calcular as métricas *recall*, *precision* e F_1 apresentadas na Seção 4.2. O cálculo das métricas é realizado no final do processamento, sendo utilizado os valores totais de VP, FN e FP e não a média das métricas página a página.

5.4 Abordagem proposta

A abordagem utilizada nesta dissertação foi incentivada pelo trabalho (Zhai et. al., 2005), pois é possível notar que a estrutura formada por uma lista de produtos é identificada como uma lista genérica. Durante os experimentos iniciais, foi possível observar que, além da lista de produtos, existem também diversas outras listas em um documento HTML como menus, comentários, listas de seleções (list box), dentre outras.

Na Tabela 5.2, são apresentados os resultados da exploração inicial, realizada em 24 documentos de cada site do conjunto de treino. Essa exploração teve o objetivo de ajudar na escolha dos algoritmos que seriam utilizados durante o processo de experimentação. Além disso, também foram ajustados os parâmetros dos algoritmos, como a distância máxima permitida entre cada conjunto generalizador.

Site	Métrica %	CS+DC	CS+DG	CS+DT	**CA**
Americanas.com.br	Recall	79.16	75.00	79.16	75.00
	Precision	2.76	2.76	2.75	3.18
	F1	5.33	5.70	5.32	**6.11**
target.com	Recall	91.66	95.83	100	100
	Precision	3.57	3.85	3.51	5.32
	F1	6.88	7.40	6.78	**10.10**
dealextream.com	Recall	100	100	100	100
	Precision	5.73	6.36	6.36	8.79
	F1	9.65	11.97	11.97	**16.16**
Médias	F1	7.29	8.36	8.02	**10.79**

Tabela 5.2: Testes iniciais com 24 documentos do corpus de treino

Optamos por dar continuidade aos experimentos apenas com dois algoritmos, escolhendo os que apresentaram o maior F_1 médio. O primeiro algoritmo escolhido foi o Casamento de Árvores (CA), que utiliza a distância de árvores, pois apresentou o melhor resultado F_1 médio. O segundo algoritmo foi o Casamento Simples com distância de *tag* (CS+DG). A existência de outras listas em um documento HTML resulta na baixa precisão de todos os algoritmos, como pode ser observado na coluna *precision* da Tabela 5.2. Nessa tabela também é possível notar a capacidade de identificar as listas de produtos, comprovado pelo recall.

Na Tabela 5.3, são exibidos os resultados da segunda fase de experimentação, onde os dois algoritmos escolhidos foram executados sobre todo o conjunto de treino. Nessa tabela pode ser observado que o *recall* dos algoritmos se manteve próximo ao da primeira fase de experimentação, confirmando que os algoritmos são capazes de identificar as listas de produtos. Também, é possível notar que a *precision* se manteve baixa, reforçando a necessidade de regras específicas para separar as listas de produtos das demais listas.

É interessante relembrar que a primeira fase de experimentos foi direcionada à observação do comportamento dos algoritmos de detecção de listas genéricas, avaliando a capacidade desses em identificar as listas de produtos.

Finalmente, realizamos um experimento sobre o conjunto de teste, cujos resultados são apresentados na Tabela 5.4. Nessa tabela, pode ser observado o

Site	Método	Recall %	Precision %	F1 %
americana.com	CS+DG	75.00	2.96	5.70
	CA	73.72	3.15	**6.05**
target.com	CS+DG	95.83	3.85	7.40
	CA	97.84	5.18	**9.85**
dealextream.com	CS+DG	100	6.36	11.97
	CA	100	8.70	**16.01**

Tabela 5.3: Escolha do melhor método com todos os documentos do corpus de treino

número de nós retornados por cada algoritmo, ficando evidente a diminuição do domínio de busca, já que o número de nós foi reduzido de aproximadamente 1 milhão para menos de 40 mil, o que resulta em uma média de 36 listas por documento.

Método	Recall %	Precision %	F_1 %	nós retornados
CS+DG	89.81	2.06	4.04	36675
CA	96.88	3.78	7.27	17260
Todos os nós	100	0.00	0.00	1028016

Tabela 5.4: Resultados no corpus de teste

Ainda na Tabela 5.4, é possível notar que o recall melhorou no conjunto de teste. Essa diferença pode ser atribuída à dificuldade que foi imposta pelo site americanas.com, onde as listas do conjunto de teste são, aparentemente, melhor estruturadas, o que torna o processo de identificação mais fácil.

Criando regras específicas

Durante as primeiras fases de experimentação, foi possível observar a grande diminuição no domínio de busca. No entanto, a existência de diversas listas nos documentos HTML mostrou que apenas os algoritmos genéricos não são suficientes para identificar as listas de produtos. Por esse motivo, iniciamos a segunda fase de experimentação buscando informações úteis para separar as listas de produtos das demais.

Notamos que as listas de produtos são, normalmente, compostas por diversos itens. Note que cada item é um conjunto generalizador formado durante o processo de identificação da lista pelos algoritmos CS ou CA. Utilizamos essa observação para criar uma primeira regra, chamada de R1, classificando a lista que contém maior número de itens de cada documento como lista de produtos.

Os resultados da aplicação da primeira regra gerada podem ser observados na Tabela 5.5. Nessa tabela, é possível perceber que ambos os algoritmos (CS+DG e CA) obtiveram resultados idênticos. Ainda na Tabela 5.5, pode ser observado que o *recall* e a *precision* são iguais. Isso significa que em 48,23% das vezes a lista com mais elementos é a lista de produtos. O fato curioso da igualdade do *recall* e *precision* será discutido posteriormente. No entanto, é importante notar que há ainda muitos casos onde a lista de produtos não é a maior lista, sendo uma fração de 51,77%. Por esse motivo, voltamos a examinar os resultados, em busca de mais características para identificar as listas de produtos.

Durante esse processo, foi possível observar que as listas de produtos, em grande parte dos sites, contêm imagens. Por essa razão, criamos mais uma regra, a regra R2, onde a lista com maior número de imagens é classificada como lista de produtos. Além disso, foi adicionada uma condição à primeira regra, criando a regra R3. A regra R3 classifica como lista de produtos a lista com maior número de itens com pelo menos duas imagens. Esse parâmetro foi escolhido, pois, por hipótese, assumimos que as listas são formadas por pelo menos dois itens. É natural esperar que existam pelo menos duas imagens nas listas de produtos, já que em um grande número de vezes a lista de produtos apresenta uma imagem para cada produto.

Regra	Algoritmo	Recall %	Precision %	F_1 %
R1	CS+DG	48.23	48.23	48.23
	CA	48.23	48.23	48.23
R2	CS+DG	44.70	44.70	44.70
	CA	87.05	87.95	87.95
R3	CS+DG	48.23	91.11	63.07
	CA	48.23	89.12	62.59
R2+R3	CS+DG	58.82	58.27	58.54
	CA	87.05	85.45	86.24

Tabela 5.5: Regras específicas sobre o corpus de treino

Ao unir as regras R2 e R3 (R2+R3), utilizando o resultado de ambas, é possível classificar duas listas de produtos em um único documento. Com isso, dois nós podem fazer parte da mesma subárvore, criando listas aninhadas, ou seja, uma lista faz parte da outra. Para evitar esse cenário, quando a regra R2 e R3 são utilizadas simultaneamente, é utilizada a seguinte condição: se a lista que contém mais imagens é ancestral da lista que contém maior número de itens, a lista com mais imagens não é classificada como uma lista de produtos.

A contribuição de cada uma das regras pode ser observada na Tabela 5.5. Note que, em grande parte dos resultados, a métrica *recall* é igual à métrica

precision. Isso acontece, pois a maior parte das regras classifica apenas um nó, e existe exatamente um nó correto a ser classificado em cada documento. Com isso, as duas métricas, *recall* e *precision*, se tornam iguais, podendo ser chamada de acurácia. A acurácia é obtida dividindo o número de acertos (listas que foram classificadas corretamente) pelo número de listas. As regras que geram resultados diferentes de *recall* e *precision* são a R3, permitindo que nenhum nó seja classificado como lista de produtos, e a R2+R3, que permite que dois nós sejam classificados como lista de produtos.

Site	CS+DG	CA
bestbuy.com	91.78	93.15
circuitcity	87.91	87.24
ebuy.com	0.0	100
mysimon	100	100
submarino.com	95.32	94.39
Média	75.00	94.95

Tabela 5.6: Resultado F_1 das regras específicas no corpus de teste

Finalmente, aplicamos as regras específicas R2+R3 sobre o conjunto de teste. Como pode ser observado na Tabela 5.6, os resultados se mostraram estáveis, repetindo a melhora apresentada entre o corpus de treino da primeira fase da experimentação.

Podemos concluir que os resultados da identificação de listas de produtos em sites de comércio eletrônico são satisfatórios, mesmo sabendo que podem existir diversos sites com listas de produtos mais complexas. Acreditamos que o pré-processamento, que identifica todas as listas genéricas, demostrou grande valor, facilitando a criação de regras específicas como era esperado. Então, estruturas mais complexas, não identificadas pelas regras específicas, devem demandar pouco esforço para a criação de novas regras.