Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Combinando Heurísticas Locais e Global para Rotulagem de Dados Anônimos Extraídos da Web

Lisandra Santos da Costa

Manaus – Amazonas
Outubro de 2009

Lisandra Santos da Costa

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas na área de Recuperação de Informação e Banco de Dados, como requisito parcial para obtenção do Título de Mestre em Informática.

Orientador: Prof. João Marcos Bastos Cavalcanti, Ph.D.
Co-orientador: Prof. Dr. Altigran Soares da Silva

Lisandra Santos da Costa

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas na área de Recuperação de Informação e Banco de Dados, como requisito parcial para obtenção do Título de Mestre em Informática.

Banca Examinadora

Prof. João Marcos Bastos Cavalcanti, Ph.D. – Orientador
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. Altigran Soares da Silva – Co-orientador
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dra. Renata Galante
Departamento de Informática – UFRGS

Manaus – Amazonas
Outubro de 2009

Às grandes mulheres guerreiras, Sílvia Ramos e Miracy Godinho,

por toda dedicação e carinho.

Agradecimentos

Primeiramente à minha mãe *Sílvia Ramos* pelo constante apoio e compreensão.

Também, os meus sinceros agradecimentos aos meus orientadores *João M. Cavalcanti* e *Altigran Soares*. Obrigada pela oportunidade, paciência e pelos ensinamentos transmitidos durante toda minha formação.

Ao colega *Marco Sevalho* por toda ajuda e colaboração no meu trabalho.

Aos colegas de mestrado, *Andréa Giordanna, Francisca Sancha, Kelen Acquati, Isaac Bandeira, Geraldo Filho, Nick Diego, Ícaro de Oliveira, André Carvalho, Bruno Araújo e Karane Mariano* que me acompanharam durante o curso, por toda ajuda prestada e por tornarem a caminhada menos difícil.

Ao meu grande amigo *Klessius Berlt* pelo companheirismo, ajuda e carinho durante toda jornada.

A toda minha família e amigos, e em especial ao meu namorado *Ricardo Cabral* pelo amor, conforto e parceria até nas horas mais difíceis.

Resumo

Com o objetivo de manipular de forma eficiente dados de páginas web, muitos trabalhos têm sido desenvolvidos com o objetivo de extrair automaticamente esses dados. Porém, tais propostas reconhecem apenas a estrutura implícita dos dados e não sua semântica, produzindo como saída conjuntos de dados, ditos anônimos, por não possuírem nenhum rótulo que os descreva.

Neste trabalho apresentamos uma abordagem totalmente automática para o problema de rotulagem de dados anônimos, que consiste em encontrar termos que descrevam semanticamente um conjunto relacional de dados anônimos extraídos da Web, tais quais produzidos por extratores de dados da Web. Nossa abordagem destaca-se por explorar tanto as evidências locais quanto globais, combinando de forma eficiente os dois tipos de métodos (local e global). Na fase local contamos com 4 heurísticas locais que exploram cada uma aspectos distintos de cada atributo anônimo. A fase global tem duas funções básicas: encontrar um conjunto de rótulos candidatos para um dado atributo anônimo e atribuir o rótulo mais significativo a este atributo. Este método utiliza uma máquina de busca popular para encontrar termos que melhor descrevem os atributos anônimos em páginas web.

Ao final são apresentados os resultados dos experimentos realizados com conjuntos de dados extraídos de sites da Web e verificamos a eficácia da abordagem proposta em relação aos métodos utilizados de forma isolada.

Sumário

Resumo **vi**

1 Introdução **1**

1.1 Organização da Dissertação 4

2 Trabalhos Relacionados **6**

3 Rotulagem de Dados Anônimos Extraídos da Web **10**

3.1 Visão Geral da Nossa Abordagem 11

3.2 Rotulagem Local 13

3.2.1 Heurística 1 – Rótulos de Campos de Formulários 13

3.2.2 Heurística 2 – Tabelas de Páginas Web 15

3.2.3 Heurística 3 - Padrões de Texto 16

3.2.4 Heurística 4 – Prefixos 17

3.3 Rotulagem Global 18

3.3.1 Seleção Global de Rótulos 19

3.3.2 Associação Global de Rótulos 22

4 Experimentos **24**

4.1 Cenário experimental 24

4.2 Experimentos com a Rotulagem Local 26

4.2.1 Resultados dos Experimentos com a Rotulagem Local 28

4.3 Experimentos com a Rotulagem Global 31

4.3.1 Resultados dos Experimentos com a Rotulagem Global 32

5 Conclusão **35**

Referências Bibliográficas **37**

Lista de Figuras

3.1 Exemplo de conjunto de dados anônimos do domínio música, contendo uma relação R(A1,A2). 10
3.2 Visão Geral da Abordagem Proposta 11
3.3 Formulário de consulta e resultados de busca de *http://www.amazon.com/books*. 14
3.4 Exemplo de uma página contendo valores de dados numa estrutura de tabela. 15
3.5 Registro de dado com valores de dados cercados de rótulos de *www.wacthzone.com*. 16
3.6 Resultado da seleção local de rótulos para $R(A_i, A_j)$. 18
3.7 Padrões de consulta usados no método de Seleção Global de Rótulos. 21

Lista de Tabelas

4.1 Conjuntos de dados usados em nossos experimentos, seus respectivos domínios, número de tuplas (t) e atributos (a) de cada conjunto, e sua classificação quanto a popularidade na Web. 25

4.2 Número de campos de busca, consultas e páginas web usadas para construir os conjuntos de dados. 26

4.3 Valores de acurácia média das heurísticas locais para todas as 21 bases de dados. 29

4.4 Valor de acurácia média para cada base de dados usando **C1** and **C2**. 29

4.5 Valor de acurácia da rotulagem local para cada domínio. 30

4.6 Distribuição de atributos de acordo com o tamanho do conjunto de rótulos candidatos gerados a partir da fase local. 31

4.7 Valor de acurácia de *C1* e *C2* para $|L| = 1$. 31

4.8 Acurácia do método global usando $C1$ e $C2$. 32

4.9 Valores de acurácia para cada base de dados usando o método global e o ganho obtido utilizando *C1*. 33

4.10 Valores de acurácia para cada base de dados usando o método global e o ganho obtido utilizando *C2*. 34

Capítulo 1

Introdução

A Web é considerada uma grande e rica fonte de informações de diversos domínios. Seus dados são utilizados por inúmeras aplicações para auxiliar usuários em tarefas comuns do dia-a-dia. Mas para isso, alguns desafios são encontrados, visto que páginas web geralmente são construídas para serem navegadas e compreendidas por pessoas e não para serem computadas por aplicações. Muitas aplicações têm sido propostas envolvendo tarefas de coleta, indexação, busca de documentos, extração de dados, integração de esquemas, etc. Por isso torna-se tão interessante o estudo de métodos capazes de manipular de forma eficiente esses dados.

Geralmente, usuários buscam informações na Web submetendo consultas a máquinas de busca ou navegando através de *links* de páginas. Entretanto, tais estratégias possuem certas limitações para o acesso aos dados, por exemplo: o número de *links* encontrados navegando por páginas web pode dificultar o processo de busca por itens de dados e torná-lo mais complicado e demorado. Por outro lado, consultas em máquinas de busca, apesar de muitas vezes ser a opção mais eficiente, em geral trazem uma quantidade de resultados muito maior do que o usuário pode lidar. Dessa forma, apesar da enorme quantidade de dados disponíveis na Web, estes são dificilmente recuperados ou manipulados apropriadamente como é feito nos bancos de dados tradicionais [Laender et al., 2002].

Uma possível solução é extrair dados da Web para povoar bancos de dados, para que possam ser utilizados de forma mais eficiente. Por isso, o problema de extração de dados da

Web tem sido bastante discutido na literatura [Buttler et al., 2001, Crescenzi et al., 2001] e várias técnicas têm sido propostas, como os *wrappers*, que são programas que extraem dados da Web e os reorganizam num formato adequado para uso posterior. A geração automática de *wrappers* [Laender et al., 2002, Kuhlins and Tredwell, 2002] é uma alternativa para extrair de forma automática dados da Web, porém não elimina totalmente a necessidade de intervenção humana, pois essas técnicas reconhecem apenas a estrutura implícita dos dados e não sua semântica e por isso produzem como saída conjuntos de dados anônimos (sem rótulos descritivos associados a eles). Portanto, torna-se necessário o estudo de técnicas capazes de associar automaticamente rótulos que descrevam a semântica de cada item de dado extraído. A rotulagem automática de dados representa um grande passo para a extração e manipulação automática de dados extraídos da Web [Arlotta et al., 2003].

Outra área de pesquisa para a qual a rotulagem automática de dados torna-se tão interessante é a chamada integração de dados. Por exemplo, um usuário que deseja comparar diversos modelos de celular para então realizar a compra daquele que mais lhe agrada deverá preencher formulários de busca várias vezes em diversos sites de lojas virtuais, recuperar os dados relevantes retornados de cada um, compará-los manualmente para então efetuar a compra on-line. Surge então a necessidade de ferramentas de integração de dados que auxiliem o usuário em tarefas similares a essa, e para isso é preciso que dados da Web sejam extraídos e rotulados automaticamente.

Recentemente, alguns métodos têm sido propostos para rotulagem automática de dados extraídos da Web [Arlotta et al., 2003, Wang and Lochovsky, 2003, Song et al., 2004]. A maioria desses métodos utiliza informações contidas nas páginas que contém os dados extraídos ou nos formulários utilizados na geração dessas páginas. Chamamos esse tipo de informação de *evidências locais* e os métodos de rotulagem que as utilizam de *métodos locais*. Apesar de apresentarem altos níveis de acurácia, os métodos locais apresentam algumas falhas, por exemplo, a ausência de rótulos para alguns dos dados contidos nas páginas. Portanto, é necessário que outras fontes de evidências sejam exploradas.

Em [da Silva et al., 2007, Lu et al., 2007] são propostos métodos para rotulagem automática que utilizam evidências externas à fonte de origem dos dados, as quais chamamos de *evidências globais* e os métodos de rotulagem que as utilizam, chamamos de *métodos globais*. O trabalho apresentado em [da Silva et al., 2007] propõe uma nova abordagem para atribuir rótulos semanticamente representativos aos dados anônimos extraídos da Web. Ele destaca como uma de suas vantagens a não utilização das páginas de onde os dados foram extraídos, tentando assim amenizar as falhas encontradas na rotulagem local, como a ausência de rótulos nas páginas. Os rótulos selecionados são encontrados na Web através de consultas a máquinas de busca. Essa abordagem relata bons níveis de revocação e acurácia e se mostra como uma alternativa interessante para a rotulagem de dados. Porém, uma de suas desvantagens é o tempo de submissão das consultas e obtenção dos rótulos.

Acreditamos que as páginas de onde os dados foram extraídos possuem informações valiosas para encontrarmos rótulos semanticamente descritivos para os atributos anônimos, porém não devem ser utilizadas como única fonte de informação. Este trabalho propõe uma nova abordagem para rotulagem automática de dados que utiliza tanto *evidências locais* como *globais*. Para isso, a abordagem é dividida em duas fases. Dado um conjunto de dados anônimos divididos em atributos, a primeira fase, *Seleção Local de Rótulos*, consiste em selecionar um conjunto de rótulos candidatos para cada atributo utilizando 4 *heurísticas locais*. Na segunda fase utilizamos o *método global* para encontrar rótulos candidatos aos dados aos quais nenhum rótulo candidato foi selecionado na primeira fase, e para selecionar o rótulo adequado dentre o conjunto de rótulos candidatos obtidos na primeira fase para um determinado dado anônimo.

A abordagem global só é utilizada caso os atributos não sejam rotulados pelos *métodos locais* diminuindo o número de consultas submetidas à máquina de busca. No caso em que os *métodos locais* selecionam mais de um rótulo candidato para o mesmo atributo, a abordagem global seleciona de forma totalmente automática o rótulo adequado para esse atributo. E quando nenhum rótulo candidato for selecionado para um atributo na

primeira fase, a abordagem global seleciona um conjunto de rótulos candidatos e associa o melhor deles a este atributo. Experimentos realizados demonstram que nossa abordagem é mais eficiente que as existentes, pois quando as evidências locais falham, evidências globais são buscadas em outras fontes com o auxílio de uma máquina de busca de uso geral, e quando o método local traz mais de um rótulo para um dado atributo, o método global consegue identificar o rótulo adequado para esse atributo.

As duas abordagens foram adaptadas para trabalharem em conjunto tirando melhor proveito das vantagens apresentadas por cada uma delas. Atualmente, os métodos locais foram baseados nas heurísticas propostas em [Wang and Lochovsky, 2003], já o método global é baseado em [da Silva et al., 2007]. Nossa abordagem se destaca dentre as demais por explorar tanto as evidências locais quanto as globais e por ser totalmente automática, apresentando um novo método de atribuição de rótulos. Nossa principal contribuição é a forma de combinar os dois tipos de métodos, *local* e *global*, e de acordo com os resultados de nossos experimentos nossa abordagem é bastante eficaz.

1.1 Organização da Dissertação

Esta dissertação é composta por cinco capítulos organizados da seguinte forma.

No Capítulo 2 são discutidos os principais trabalhos relacionados. Para cada trabalho é apresentado o seu funcionamento básico e as evidências utilizadas no processo de rotulagem. São discutidos também os pontos positivos e negativos de cada trabalho e uma breve comparação com nossa abordagem.

No Capítulo 3 é apresentado de forma geral o funcionamento de nossa abordagem e os principais conceitos utilizados. No mesmo capítulo, na Seção 3.2 e 3.3 são detalhados os *métodos locais* e *globais* utilizados.

No Capítulo 4 são descritos o cenário experimental, a metodologia de avaliação, métricas e bases utilizadas e resultados obtidos. Este capítulo é dividido em duas partes principais: uma descrevendo os experimentos e resultados utilizando o *método local* e a outra avalia o *método global*.

Finalmente, o Capítulo 5 apresenta as conclusões deste trabalho com base nos resultados empíricos obtidos no Capítulo 4.

Capítulo 2

Trabalhos Relacionados

Para explorar de maneira mais eficiente dados da Web, muitas pesquisas têm sido focadas em extração de dados como [Arasu and Garcia-Molina, 2003, Zhai and Liu, 2005]. Vários trabalhos com o objetivo de facilitar o desenvolvimento de *wrappers* têm sido propostos. Ferramentas de extração são classificadas de acordo com o seu nível de automação em [Laender et al., 2002]: *manuais*, *semi-automáticas* e *automáticas*. Estudos recentes têm se concentrado na geração automática de *wrappers* com o objetivo de diminuir a interferência humana e conseqüentemente, os custos de manutenção e manipulação de *wrappers*. No entanto, muitas dessas propostas não eliminam a intervenção humana, pois os dados extraídos pelos *wrappers* são anônimos, sendo necessário que sejam rotulados manualmente. Para alcançar de fato a extração automática de dados na Web foram propostos alguns métodos que visam a rotulagem automática desses dados, estes são descritos abaixo.

O *Labeller* [Arlotta et al., 2003] é uma ferramenta proposta para rotular atributos anônimos de forma automática utilizando evidências locais. Por isso, conta com um método que parte de uma hipótese simples de que freqüentemente itens de dados vêm acompanhados de termos que descrevem o seus significados para auxiliar os usuários na compreensão de tais informações, supõe também que esses termos ocorrem próximos dos valores de dados nas páginas de onde os dados foram extraídos. O *Labeller* propõe algumas heurísticas baseadas nessa hipótese. Tais heurísticas levam em conta a disposição

dos rótulos fornecidos nas páginas em relação aos valores de dados. Por exemplo, rótulos geralmente são termos localizados bem próximos aos valores de dados, dispostos ou acima ou do lado esquerdo desses valores. Esse trabalho apresenta bons resultados, porém sabemos que nem todos os rótulos são fornecidos nas páginas o que torna esse método muito limitado.

O *DeLa* (*Data Extraction and Label Assignment*) [Wang and Lochovsky, 2003] é um sistema que reconstrói (parte de) uma base de dados da *Hidden Web* enviando consultas através de formulários HTML e extraindo objetos de dados a partir das páginas retornadas. Além disso, o sistema é composto por um rotulador de dados. Seu processo de rotulagem utiliza 4 heurísticas locais, nas quais nos baseamos. Diferente do *Labeller*, o *DeLa* não utiliza somente termos contidos nas páginas para rotular os dados, mas também baseia-se em outras evidências, por exemplo, formato dos dados e rótulos de campos de formulários. Estas heurísticas serão descritas mais adiante na Seção 3.2. Os resultados apresentados indicam alta acurácia, entretanto, acreditamos que o uso isolado de evidências locais não é a maneira mais apropriada de rotular dados anônimos. Pois nem sempre os valores de dados vêm acompanhados de termos descritivos nas páginas ou não possuem um formato padrão, como a abordagem sugere. Além disso, não é apresentado de forma clara como os resultados obtidos pelas heurísticas são combinados.

Em [Lu et al., 2007] é apresentado um método de rotulagem de dados extraídos da *Deep Web*. Dado um conjunto de dados anônimos a primeira etapa do método faz o agrupamento dos dados, tal que ao final cada grupo possui dados com o mesmo significado semântico (atributos). Para cada grupo formado, um conjunto de heurísticas é usado para rotular os valores de dados contidos nesse grupo. Essas heurísticas usam *evidências locais* e *globais*. As heurísticas locais são baseadas nas heurísticas propostas em [Wang and Lochovsky, 2003]. A principal diferença em relação à rotulagem é que enquanto o *DeLa* utiliza o formulário de busca utilizado na geração páginas dinâmicas de um site específico para rotular os atributos anônimos deste site, o método proposto utiliza vários formulários de diferentes sites do mesmo domínio para auxiliá-lo no processo de

rotulagem. De acordo com nosso levantamento bibliográfico, foi a primeira tentativa de rotular atributos anônimos buscando rótulos descritivos fora do domínio de onde os dados foram extraídos, ou seja, utilizando *evidências globais*. Apesar de apresentar uma melhora significativa em conseqüência desse conjunto de formulários, não resolve totalmente problemas específicos de rotulagem, para isso são necessárias novas técnicas. Cada heurística pode retornar um rótulo para um mesmo atributo, e por isso é usado um modelo probabilístico para selecionar o rótulo mais apropriado para o atributo anônimo. Essa técnica de seleção de rótulos apesar de apresentar bons resultados necessita de treino, enquanto que a nossa abordagem conta com uma técnica totalmente automática para selecionar o rótulo mais adequando para um atributo.

Em [da Silva et al., 2007] é apresentada uma nova abordagem automática que identifica e atribui rótulos semanticamente representativos a um conjunto de dados anônimos extraídos da Web. O método proposto utiliza somente evidências globais, ou seja, não necessita das páginas de onde os dados foram extraídos no processo de rotulagem. Ao invés de buscar os rótulos nas fontes de origem dos dados, o método utiliza uma máquina de busca popular para encontrar esses rótulos em documentos da Web que contenham os valores de dados extraídos. Essa abordagem é dividida em duas etapas principais: a primeira seleciona um conjunto de rótulos candidatos para o conjunto de atributos anônimos, enviando consultas para a Web e buscando termos que ocorram próximos aos valores de dados. A segunda etapa seleciona para cada atributo anônimo o rótulo mais adequado dentre o conjunto de rótulos candidatos também enviando consultas formadas pelo rótulo candidato e alguns valores de dados. A partir do número de respostas retornadas o método associa o rótulo com mais afinidade ao atributo. Essa abordagem apresenta altos níveis de acurácia em seus experimentos, entretanto podemos observar uma desvantagem no seu processo de rotulagem que é justamente o uso de consultas a máquinas de busca. Tais consultas são operações caras em termos de tempo de processamento e são executadas para uma amostra do conjunto de dados para cada atributo anônimo.

Nossa abordagem partiu da observação de que as heurísticas locais apresentam bons

níveis de acurácia, mas ainda assim é possível melhorar esses resultados, aplicando a rotulagem global nos atributos onde os métodos locais falham. Dessa forma, as consultas só são submetidas às máquinas de busca nos casos realmente necessários. Os métodos locais, mesmo apresentando bons resultados, falham em alguns casos pois são muito dependentes de rótulos encontrados nas páginas de onde os dados foram extraídos e do formato desses dados. Além disso, os métodos locais mais conhecidos podem retornar mais de um rótulo canditado para um dado atributo anônimo, porém a seleção do rótulo mais apropriado a este não é feita de maneira totalmente automática como em nossa abordagem. Já o método global que utilizamos como uma etapa da nossa abordagem apresenta bons reusultados, entretanto para cada atributo anônimo é necessário buscar evidências submentendo consultas a uma máquina de busca a partir de uma amostra dos valores de dados. O uso da máquina de busca na Web é uma operação custosa em termos de tempo de execução. Na nossa abordagem esse número de consultas é reduzido devido ao uso de evidências locais.

A seguir as evidências utilizadas por cada um dos trabalhos descritos nessa seção:

- *Labeller*: termos que ocorrem nas páginas (*evidências locais*).
- *DeLa*: termos que ocorrem nas páginas, formato dos valores de dados e rótulos de formulários de consulta (*evidências locais*).
- *[Lu et al., 2007]*: termos que ocorrem nas páginas, formato dos valores de dados, rótulos de formulários de consulta que geraram o conjunto de páginas (*evidências locais*) e rótulos de formulários de consulta de outros sites do mesmo domínio (*evidências globais*).
- *[da Silva et al., 2007]*: termos que ocorrem em documentos da Web que possuem os mesmos valores de dados anônimos (*evidências globais*).

Capítulo 3

Rotulagem de Dados Anônimos Extraídos da Web

Ferramentas de extração automática de dados geralmente produzem conjuntos de dados com estrutura e domínios dos seus atributos bem definidos, apesar de não explícitos. Os dados produzidos por tais ferramentas são ditos anônimos, pois não possuem nenhum rótulo que lhe dê alguma propriedade semântica. Por exemplo, a Figura 3.1 mostra um conjunto de dados anônimos sobre música. Este conjunto de dados é um exemplo de uma saída produzida pela maioria das ferramentas de extração automática de dados existentes na literatura.

Note que os conjuntos de dados têm um esquema implícito bem definido que pode ser visto como uma relação $R(A1, A2)$. Ambos atributos também têm os domínios bem definidos ($A1$ contém nomes de artistas e $A2$ contém nomes de álbuns). O esquema implícito do conjunto de dados extraídos também representa aspectos do domínio do

R	A_1	A_2
	Miles Davis	Kind of Blue
	John Coltrane	A Love Supreme
	John Coltrane	My Favorite Things
	...	

Figura 3.1: Exemplo de conjunto de dados anônimos do domínio música, contendo uma relação R(A1,A2).

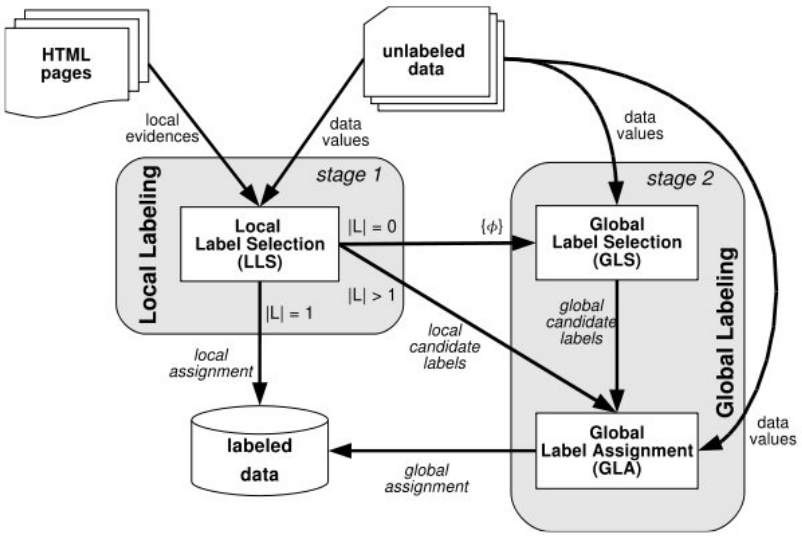

Figura 3.2: Visão Geral da Abordagem Proposta

conjunto de dados (música nesse caso), que é também conhecido por nós. O problema de rotulagem consiste em encontrar rótulos descritivos para um conjunto relacional de dados anônimos.

3.1 Visão Geral da Nossa Abordagem

A busca e associação de rótulos descritivos para esses atributos anônimos é o tema abordado por esta dissertação. Nesta seção nós introduzimos uma abordagem automática para solucionar este problema que consiste de dois métodos complementares: um *método de rotulagem local* e um *método de rotulagem global*. Cada método considera um grupo de evidências distintas para encontrar e associar rótulos candidatos: as *evidências locais* e as *evidências globais*. A principal hipótese é que nós podemos encontrar descritores semânticos tanto em partes de páginas HTML de onde os dados foram extraídos quanto em páginas HTML indexadas por máquinas de busca populares que contêm os mesmos valores de dados. A abordagem proposta é ilustrada pela Figura 3.2.

Em nossa abordagem o processo de rotulagem é composto de duas etapas. A primeira etapa implementa os métodos de rotulagem local que encontra rótulos candidatos para cada atributo anônimo A_j utilizando evidências contidas nas páginas de onde foram ex-

traídos os dados. Para encontrar esses rótulos o método usa quatro heurísticas locais, cada uma explorando propriedades distintas do domínio do atributo. Essas heurísticas contam com alguns aspectos da coleção de páginas web usadas para extração dos dados. A hipótese é que estas páginas HTML contêm termos que descrevem e estão relacionados ao domínio dos atributos. Estes termos formam um conjunto com um grande potencial para rótulos candidatos e chamamos de evidências locais.

A primeira etapa *Seleção Local de Rótulos* recebe como entrada um conjunto de valores de dados de um dado atributo anônimo e um conjunto de evidências locais encontrado na coleção de páginas HTML de onde o conjunto de dados anônimos foi extraído. Essa etapa produz um conjunto L de rótulos candidatos para cada atributo anônimo A_j. O número de rótulos candidatos selecionados ($|L|$) nessa etapa determina a necessidade e o comportamento do método de rotulagem global, como descrito abaixo:

- $|L| = 1$: Se apenas um rótulo candidato é selecionado a *Associação Local* é realizada considerando este rótulo.
- $|L| = 0$: Se nenhuma heurística produziu um rótulo candidato, o método de rotulagem global é executado (*Seleção Global de Rótulos*).
- $|L| > 1$: Se mais de um rótulo é selecionado, o método global é executado (*Associação Global de Rótulos*).

A *Rotulagem Global* implementa um método baseado em [da Silva et al., 2007]. Este método utiliza evidências globais e possui duas etapas principais: *Seleção Global de Rótulos* e *Associação Global de Rótulos*. O valor de $|L|$ é que vai determinar qual a próxima etapa para um atributo A_j. Se mais de um rótulo candidato é selecionado para A_j ($|L| > 1$) a *Rotulagem Global* executa a *Associação Global de Rótulos* passando o conjunto de rótulos candidatos locais. Nesse passo, o rótulo com mais afinidade semântica com os valores de dados é associado ao atributo anônimo. Se nenhum rótulo é selecionado ($|L| = 0$) a *Seleção Global de Rótulos* é executada para buscar um conjunto de rótulos

candidatos na Web e então a fase de *Associação Global de Rótulos* é usada para decidir qual dos rótulos candidatos selecionados é o melhor descritor para o dado atributo.

3.2 Rotulagem Local

Esta seção detalha o método de rotulagem local que corresponde ao primeiro estágio de nossa abordagem. Aqui é descrito como encontramos rótulos candidatos locais para o conjunto de dados anônimos usando informações disponíveis nas páginas web. O método usa 4 heurísticas baseadas em [Wang and Lochovsky, 2003].

Essas heurísticas contam com a hipótese de que valores de dados do mesmo domínio e atributo tem estruturas e propriedades semânticas comuns. Por exemplo, um atributo que corresponde ao preço de um dado produto tem valores de dados numéricos formatados com "." (ponto), "," (vírgula) e "$" (cifrão). Outra propriedade comum é a presença de termos descritivos ao redor de valores de dados, como o termo "preço:" frequentemente localizado antes do valor do preço.

Cada heurística analisa um grupo diferente de evidências sobre um atributo anônimo em uma coleção de páginas web. Nosso método implementa 4 heurísticas baseadas em evidências locais que trabalham de forma independente, cada uma retornando um conjunto de rótulos candidatos. Além disso, mais heurísticas podem ser adicionadas ou adaptadas. Neste trabalho nós usamos as seguintes heurísticas locais: (1) *Rótulos de Campos de Formulários*, (2) *Tabelas de Páginas Web*, (3) *Padrões de Texto* e (4) *Prefixos*. A seguir detalhamos cada uma delas.

3.2.1 Heurística 1 – Rótulos de Campos de Formulários

Bases de dados da *Deep Web* geram páginas web através de interface de consulta. Essas interfaces são formulários de consulta com campos de entrada rotulados usados para submeter palavras-chave. As páginas de resultados têm registros de dados com um conjunto de unidades de dados, cada uma correspondendo a um conceito semântico (atributo). A

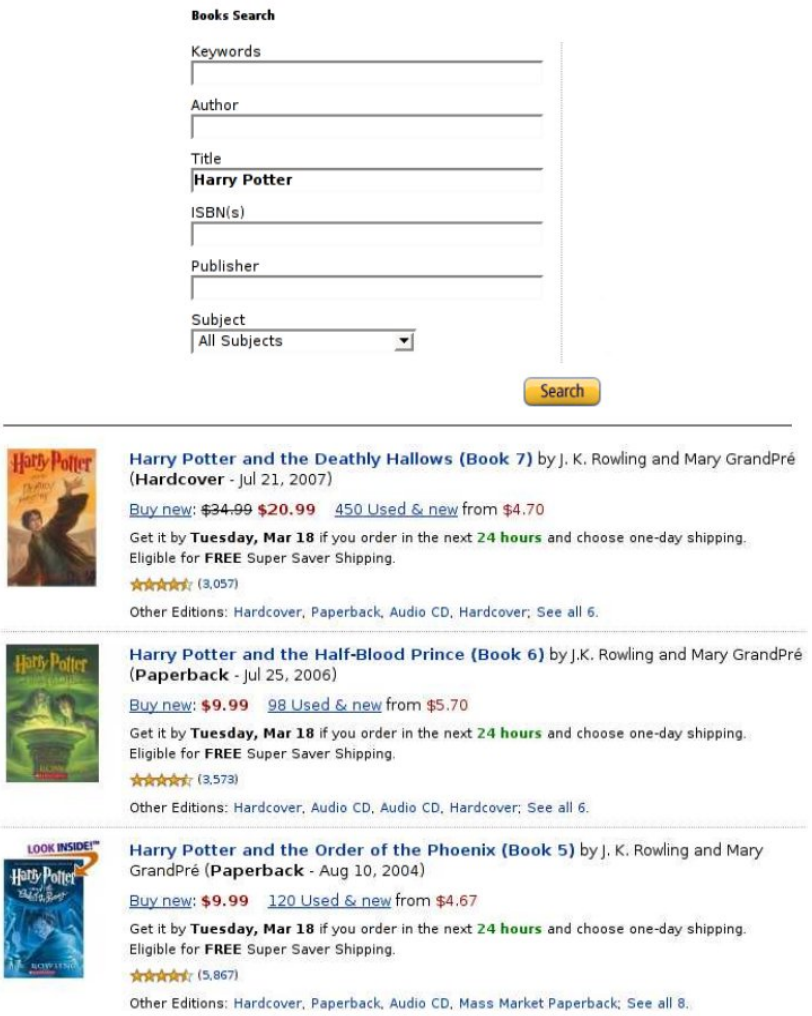

Figura 3.3: Formulário de consulta e resultados de busca de *http://www.amazon.com/books*.

intuição é que os termos da consulta provavelmente aparecerão nos valores de dados contidos nas páginas de resultado e na mesma unidade de dados para mais de um registro. Se um grupo de termos de consulta submetidos num mesmo campo do formulário frequentemente aparecer nos valores de dados em uma unidade de dado específica nos registros, nós podemos inferir que o rótulo do campo do formulário é um descritor significativo para estes valores de dados.

Dado o conjunto de campos de formulários utilizados na geração das páginas, para cada campo o conjunto de termos submetidos são buscados nos valores de cada atributo. O atributo que apresentar o maior número de ocorrência desses termos é associado ao

Year	Title	Genre	Style
1999	Capcom Generations *European*	Action	Multi-Game Compilation
2002	Capcom vs. SNK Pro	Fighting	2D Fighting
1997	Car and Driver Presents: Grand Tour Racing '98	Racing	Miscellaneous Racing
2001	Card Games	Traditional	Card Game
1998	Cardinal Syn	Fighting	3D Fighting
1998	Cardinal Syn *European*	Fighting	3D Fighting

Figura 3.4: Exemplo de uma página contendo valores de dados numa estrutura de tabela.

título do campo de formulário correspondente, ou seja, esse título é selecionado como um rótulo candidato para esse atributo.

As evidências locais neste caso é sobre a correspondência entre os formulários de consulta da *Deep Web* e unidades de dados nas páginas de resultados. Por exemplo, na Figura 3.3 existem resultados de busca para a consulta *"Harry Poter"* em uma loja virtual de livros. Essa consulta foi submetida através do campo de formulário de consulta chamado *"Title"*. Os resultados mostram três registros de dados representando livros compostos por uma conjunto de unidades de dados (título, autor, formato, etc.). Veja que o termo *"Harry Poter"* aparece na mesma unidade de dado em todos os objetos retornados. Se isso acontece com outros valores de uma mesmo domínio isso torna uma forte evidência que o rótulo *"Title"* é um descritor significativo para este atributo anônimo. Uma observação importante sobre essa heurística é que podemos usá-la somente com conjuntos de dados que foram extraídos de páginas geradas a partir consultas a formulários, e quando temos as seguintes informações: rótulos dos campos do formulário e termos usados nas consultas.

3.2.2 Heurística 2 – Tabelas de Páginas Web

Muitos web sites usam estruturas tabulares para organizar seus valores de dados como podemos observar na Figura 3.4. Nestes casos, as colunas representam valores de atributos e as linhas os registros de dados. Os cabeçalhos das tabelas, frequentemente localizados no topo, descrevem cada coluna. De acordo com esta evidência os termos de um cabeçalho de

Figura 3.5: Registro de dado com valores de dados cercados de rótulos de *www.wacthzone.com*.

uma tabela tornam-se bons rótulos candidatos para seus atributos anônimos. A evidência local para esta heurística é sobre a existência de uma estrutura tabular com valores de dados de atributos distintos.

Dado o conjunto de valores anônimos e as páginas de onde foram extraídos, os dados são buscados nas páginas e caso esses dados estajam numa estrutura tabular os termos que ocorrem no topo da tabela são identificados. Se esses termos ocorrem nos valores de dados eles são desconsiderados, caso não sejam encontrados nos dados eles são selecionados como rótulos candidatos para os atributos correspondentes a cada coluna.

3.2.3 Heurística 3 - Padrões de Texto

Alguns valores de dados tem algum padrão em comum ou um domínio de valor bem conhecido. Por exemplo, na Figura 3.5 os preços vêm sempre acompanhados do símbolo "$". Outros padrões são comumente encontrados em páginas web, como datas que são frequentemente formatadas usando o padrão "dd/mm/aaaa", e-mails devem ter o símbolo "@" e valores de descontos podem ter o sufixo "% off". Alguns conceitos como meses do ano, dias da semana, nomes de países e outros são facilmente reconhecidos. Conceitos como estes podem ter um rótulo pré-definido, um conjunto de padrões ou domínios de

valores relacionados a eles. Por exemplo, os conceitos de dados tem um rótulo "data" e um conjunto de padrões como "dd/mm/aaaa" e "dd-mm-aa". Como outro exemplo, o conceito mês tem um rótulo "mês" e o domínios de valores composto por "janeiro", "fevereiro", etc. Dessa forma, dado um conjunto de dados e um conceito reconhecido, se todos os valores de dados casam seu formato padrão ou seu domínios de valores conhecidos, esta heurística seleciona o rótulo do conceito como um rótulo candidato.

3.2.4 Heurística 4 – Prefixos

Páginas web ricas de dados frequentemente apresentam valores de dados cercados de termos descritivos. Esses termos geralmente estão localizados na esquerda ou acima dos valores mas podem também estar a direita acima. Por exemplo, na Figura 3.5 mostra um registro de dados de uma loja virtual de relógios onde todos os valores do domínio modelo são comumente prefixados pelo rótulo "Model". Então este rótulo pode ser associado a todos os valores de dados desse atributo. Esta heurística busca por prefixos comuns em registros de dados da mesma unidade de dados e seleciona seu rótulo como descritor para o atributo.

Em nossa abordagem o método de seleção local de rótulos aplica as quatro heurísticas mencionadas para cada atributo. Essas heurísticas recebem a coleção de páginas web como entrada e buscam por essas evidências. Note que dependendo da estrutura da página web e seus valores de dados e formatos, nem todas as heurísticas podem ser aplicadas. Por exemplo, se uma coleção de páginas web não foi gerada por um formulário de consulta, a heurística Rótulo de campos de formulário não pode ser executada. Outro exemplo é com a heurística Tabela de Páginas web quando as páginas web não têm uma estrutura tabular. Nestes casos a saída da heurística é um conjunto vazio. Um exemplo de resultado da seleção local de rótulo para uma relação anônima (A_i, A_j) é ilustrada na Figura 3.6.

Cada heurística tem como saída um conjunto de rótulos candidatos. Esses conjuntos são agrupados criando um conjunto final de rótulos candidatos para o atributo anônimo. No exemplo, 3 rótulos foram selecionados para A_j, cada rótulo de uma heurística diferente,

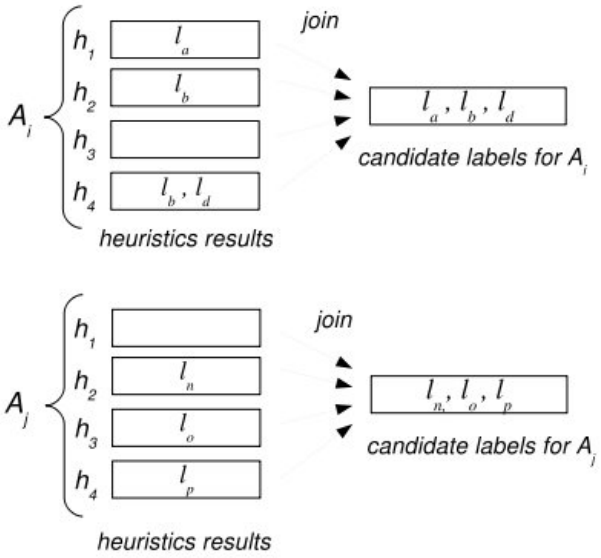

Figura 3.6: Resultado da seleção local de rótulos para $R(A_i, A_j)$.

exceto de $h1$ que não produziu nenhum. O número final de rótulos candidatos selecionados ($|L|$) para um dado atributo anônimo A_j controla o próximo passo da abordagem. Se somente um rótulo candidato é selecionado ($|L| = 1$), este rótulo é logo associado ao atributo anônimo correspondente. Se nenhum ou mais de um rótulo candidato foi retornado ($|L| = 0$ ou $|L| > 1$) o estágio 2, *Rotulagem Global*, deverá ser executado para o dado atributo anônimo.

A próxima seção apresenta o método de rotulagem global baseada nos resultados do método local.

3.3 Rotulagem Global

Nesta seção descrevemos o método de rotulagem global que utilizamos como parte de nossa abordagem híbrida. O método aqui detalhado é baseado no trabalho [da Silva et al., 2007, Sevalho, 2007] com algumas pequenas adaptações ao nosso cenário. De acordo com a Figura 3.2 o método de rotulagem global é dividido em duas etapas: a *Seleção Global de Rótulos* e a *Associação Global de Rótulos*. A partir do conjunto de rótulos candidatos gerado na primeira fase da rotulagem (*Rotulagem Local*) a etapa seguinte é então deter-

minada. A primeira etapa, *Seleção Global de Rótulos*(Seção 3.3.1), é utilizada quando a *Seleção Local de Rótulos* falha ao encontrar termos descritivos utilizando as evidências locais, ou seja, quando o conjunto de rótulos candidatos gerado na primeira fase é vazio, essa etapa é responsável por selecionar um bom conjunto de rótulos candidatos para um dado atributo utilizando evidências globais. Se o conjunto de rótulos candidatos selecionados na fase local tiver mais de um elemento a próxima etapa para o dado atributo anônimo será a *Associação Global de Rótulos*, na qual cada rótulo candidato do conjunto será avaliado e o rótulo com maior afinidade com os valores do dado atributo anônimo será selecionado como o melhor descritor para este atributo.

Esse método é baseado na hipótese de que se os dados anônimos foram extraídos da Web, certamente existem outros documentos na Web que tratam desse domínio específico. Dessa forma, esses documentos podem apresentar informações valiosas que indiquem bons descritores para esses dados anônimos (evidências globais).

A seguir é feita a descrição do funcionamento das duas etapas do método global utilizado nesse trabalho.

3.3.1 Seleção Global de Rótulos

Esta etapa é necessária somente para os atributos que não tiveram nenhum rótulo candidato selecionado na Fase Local. Essa primeira etapa do método global tem como objetivo gerar uma boa lista de rótulos candidatos para um conjunto de dados anônimos que serão utilizados na próxima etapa (*Associação Global de Rótulos*). Totalmente automática, essa técnica busca por rótulos candidatos em documentos web que contenham os valores de um dado atributo anônimo que ocorrem juntamente com certos padrões de texto comumente usados para enumerar instâncias de classes de objetos. Tais padrões de texto expressam o relacionamento entre termos e de acordo com trabalhos já publicados, quando esses padrões são usados em técnicas para encontrar instâncias de uma dada classe apresentam altos níveis de acurácia.

Neste trabalho, utilizamos os padrões léxico sintáticos conhecidos como padrões *Hearst*

que expressam uma relação de hiponímia entre termos [Hearst, 1992]. Por exemplo, o padrão "NP1 such as (NP2)" pode ser usado para encontrar uma ou mais instâncias de NP2 de uma classe NP1 na língua inglesa. Utilizamos uma abordagem similar para encontrar rótulos candidatos, já que nosso problema pode ser considerado o inverso do identificado acima. Na Figura 3.7 são apresentados os 4 tipos de padrões utilizados.

Nossa técnica de seleção de rótulos candidatos é baseada na seguinte hipótese. Dado um atributo A_i de uma relação R, assumimos que:

a) Rótulos possíveis para A_i ocorrem frequentemente em documentos web que contêm valores de A_i.

b) Tais rótulos geralmente aparecem próximos dos valores de A_i nesses documentos.

c) Essas ocorrências próximas expressam relacionamentos de hiponímia entre o rótulo e seu valor.

d) Esta relação é freqüente na Web, de modo que provavelmente esses documentos são coletados por máquinas de busca populares.

Baseado nos padrões léxico-sintáticos utilizados, o método considera três estratégias de busca por rótulos (*forward*, *backward* e *bidirectional*). Para cada estratégia, formulamos uma expressão de consulta definida como $q(p, v)$, onde q é uma instanciação do padrão p usando o valor $v \in A_j$. A consulta q é submetida à máquina de busca que retorna um conjunto de documentos da Web contendo q. Os rótulos candidatos são substantivos que ocorrem próximos a q nos resultados retornados e o padrão utilizado na consulta determina onde procurar os termos mais adequados:

a) *forward search*: busca à direita da expressão de consulta.

b) *backward search*: busca à esquerda da expressão de consulta.

c) *bidirectional search*: busca antes e depois da expressão de consulta.

p	region(s_i,p)	query(p,v)
bwd	terms before query expression	"such as v"
$perm$	terms before query expression	"v_1, v_2"
fwd	terms after query expression	"v is "
val	all terms in answer	"v"

Figura 3.7: Padrões de consulta usados no método de Seleção Global de Rótulos.

Além disso, o método restringe o espaço de busca do documento ao *snippet* que são os pequenos fragmentos de texto retornados pela máquina de busca como parte da resposta, onde geralmente encontram-se os termos usados na consulta. Dessa forma, ao invés de procurar os rótulos candidatos no documento inteiro, ele faz a busca somente nos *snippets*.

Para determinar quais rótulos candidatos são os mais adequados ao atributo anônimo, é calculado um escore para cada termo encontrado que avalia o seu grau de representatividade para o domínio do atributo. E de acordo com esse grau de representatividade os rótulos candidatos são ordenados e é gerado um *ranking*. Um bom rótulo candidato deve ocorrer com freqüência e o mais próximo possível da expressão de consulta nas respostas retornadas pela máquina de busca. Esse escore é chamado fator de coincidência e é associado a cada rótulo candidato. Seja um valor v e um termo t no conjunto de *snippets* S, o fator de coincidência de t e v em D é dado pela Equação 3.1:

$$\alpha(t, v, S) = \sum_{s_i \in S_t} \frac{W_i^t}{d(t, v, s_i)^2} \quad , \tag{3.1}$$

onde S_t é o subconjunto de *snippets* $s_i \in S$ que contêm o termo t na região de busca do rótulo, W_i é a quantidade de termos no *snippet* e $d(t, v, s_i)$ é a distância (número de termos) entre t e v neste *snippet*.

A Figura 3.7 mostra os padrões utilizados em nosso método, onde p é o nome dado ao padrão, $region(s_i, p)$ é a região de busca quando utilizado o padrão p e no *snippet* s_i e $query(p, v)$ é a frase de consulta submetida a máquina de busca com o padrão p e valor v.

Ambos padrões *bwd (backward)* e *perm (permutation)* são usados com buscas *backward*,

enquanto o padrão *forward* é usado com busca *forward* e o padrão *val (value)* é usado com uma busca *bidirectional*. Utilizamos dois valores de atributos com o padrão *permutation* na construção da consulta. Depois de associado o fator de coincidência para cada rótulo candidato é gerado um *ranking* de acordo com esse escore para cada atributo, ao final, temos uma lista de rótulos candidatos para cada atributo anônimo.

3.3.2 Associação Global de Rótulos

Nessa etapa supomos que o atributo anônimo tenha um conjunto de rótulos candidatos que pode ter sido selecionado tanto na *Seleção Local de Rótulos* quanto na *Seleção Global de Rótulos*, ou seja, dado um conjunto de rótulos candidatos $L, |L| > 1$. Ela consiste em associar para a um dado atributo anônimo um rótulo do conjunto de rótulos candidatos L, tal que esse rótulo seja considerado o melhor descritor para o domínio do atributo anônimo de L.

O método utiliza um modelo probabilístico simples que avalia a afinidade de um rótulo candidato com um atributo anônimo, ou seja, mede a probabilidade de um rótulo descrever bem um dado atributo. Para isso, o método envia determinados tipos de consultas a uma máquina de busca na Web e a partir do número de respostas retornadas mede o quão bom é um rótulo candidato para um conjunto de valores de um atributo.

A rotulagem ocorre em um processo de três etapas: construção de consultas especulativas, submissão das consultas para a máquina de busca na Web e cálculo dos rótulos que possuem maior afinidade com os valores usados nas consultas especulativas.

Uma consulta especulativa é uma expressão gerada a partir de um rótulo candidato l_i e um valor vx de um atributo anônimo A_j que formula a hipótese de que um dado termo é um bom rótulo para o dado atributo. Submetida essa consulta especulativa, a máquina de busca é usada como um oráculo que determina o quão plausível é essa hipótese de acordo com o número de respostas retornadas. Ou seja, para cada consulta especulativa submetida, o método de rotulagem apenas considera a quantidade de documentos *(Document Count)* respondidos pela máquina de busca para determinar qual rótudo candidato é mais

adequado para descrever o atributo anônimo, sem a necessidade de verificar o conteúdo dos documentos.

Dada a relação $R(A1, A2, ..., An)$ com n atributos anônimos $A1, ..., An$ onde cada A_j pertence a um dominio D_j, assumimos que dada uma instância de R com t tuplas, e dado $L = l1, ..., lm$ com m rótulos candidatos ($m > n$). O objetivo do método é associar cada A_j a um rótulo $l_i \in L$ o qual é o melhor descritor para o atributo A_j.

Para isso é utilizada a métrica chamada *LAA (Label-Attribute Affinity)* Afinidade Rótulo-Atributo que é definida como a equação 3.2

$$LAA(A_j, l_i) = \left(\frac{1}{\parallel [A_j] \parallel}\right) \sum_{x=1}^{\parallel [A_j] \parallel} \frac{DC(l_i \wedge v_x)}{\sum_{y=1}^{m} DC(l_y \wedge v_x)} \tag{3.2}$$

onde $[A_j]$ é o conjunto de valores distintos do atributo A_j e vx E $[A_j]$. $DC(q)$ é o *Document Count* de uma consulta q, definida como a quantidade de documentos relevantes para q de acordo uma determinada máquina de busca na Web. É importante destacar que somente uma amostra das tuplas no conjunto de dados anônimos é necessária para gerar as consultas especulativas obtendo bons resultados.

Capítulo 4

Experimentos

Este capítulo apresenta os experimentos realizados para avaliar nossa abordagem, assim como o cenário utilizado e os resultados obtidos.

4.1 Cenário experimental

Para medir a qualidade de nossa abordagem, foi construída uma coleção composta de 21 bases de dados extraídas da Web, listadas na Tabela 4.1. Com o objetivo de maximizar a diversidade dessa coleção, as bases de dados foram selecionadas de 12 domínios distintos e foram escolhidos atributos com diferentes características (com muitos valores distintos, com poucos valores distintos, numéricos, textuais, etc.), totalizando 114 atributos.

Para verificar o impacto da popularidade dos diferentes domínios da coleção na Web nos resultados, cada domínio foi classificado como popular ou não-popular. Sendo que domínios que contém muitas bases de dados encontradas na Web foram classificados como popular e domínios com poucas bases de dados, como não-popular.

Para executar a fase local são necessárias as seguintes informações: base de dados da coleção extraída de páginas Web, as próprias páginas coletadas, conjunto de consultas formado pelos títulos dos campos de busca utilizados e o conjunto de termos que foram submetidos a este campo de busca.

Para cada conjunto de páginas do mesmo site e domínio, foi selecionado um conjunto

domínio	popular	não-popular	site	t	a
address book		x	www.yahoo.com	100	3
books	x		www.ebay.com	98	5
			www.amazon.com	445	7
files		x	www.democrats.org	327	4
finance	x		www.yahoo.com	157	4
games	x		www.allgames.com	144	5
			www.allgame.com	470	5
job	x		www.monster.com	394	4
			www.jobbankusa.com	15	8
			www.careerbuilder.com	11	6
			www.softwarejobs.com	25	5
medicines		x	www.vitacost.com	351	7
movies	x		www.imdb.com	350	5
			www.allmovie.com	697	5
music	x		www.cdnow.com	219	4
			www.sonybmg.com.br	20	6
			www.amazon.com	44	5
			www.allmusic.com	100	4
posters		x	www.postershop.com	299	8
teams		x	www.fifaworldcup.com	32	10
wacthes		x	www.wacthzone.com	405	4
12	6	6	21	4703	114

Tabela 4.1: Conjuntos de dados usados em nossos experimentos, seus respectivos domínios, número de tuplas (t) e atributos (a) de cada conjunto, e sua classificação quanto a popularidade na Web.

de atributos que foi extraído para formar uma base de dados. A extração dos dados foi realizada utilizando a ferramenta semi-automática de extração de dados da Web, DESANA [de Sá Júnior et al., 2006]. As páginas coletadas foram geradas de duas maneiras. Na primeira, *links* de sites que não continham formulários, como *"All Products"*, que levam a páginas que contém uma grande quantidade de valores de dados, foram seguidos e as páginas coletadas. Na segunda, os sites que possuem formulários HTML para geração de páginas dinâmicas foram escolhidos e termos relacionados ao domínio específico foram submetidos manualmente para consulta, afim de gerar as páginas. Neste caso, foram selecionados 9 sites com formulários com elementos do tipo SELECTED e TEXTBOX. Para cada campo de busca de um formulário utilizado, foram coletados o título do campo t_j (texto visível ao usuário) e os termos utilizados como consulta neste campo (T_j). A Tabela 4.2 mostra o número de campos de busca e de termos que foram utilizados nas

consultas em 9 formulários. Mostra também o número total de páginas Web usadas para construir as 21 bases de dados.

campos de busca	19
consultas	897
paginas na Web	505

Tabela 4.2: Número de campos de busca, consultas e páginas web usadas para construir os conjuntos de dados.

Para medir a qualidade dessa nova abordagem adotamos a acurácia como métrica de avaliação. Esta é definida como a porcentagem da quantidade de atributos rotulados corretamente sobre a quantidade total de atributos. Como o objetivo deste trabalho é mostrar o ganho obtido com uso da rotulagem local e global em conjunto, foi necessário primeiramente executar e avaliar o desempenho da rotulagem local em nossa coleção. Depois disso, a rotulagem global foi executada somente com os atributos que não foram devidamente rotulados na primeira fase. Com isso, a combinação das duas abordagens pode ser avaliada em relação ao uso isolado das mesmas. Dessa forma, os experimentos foram executados em duas etapas. Na primeira, avaliamos a rotulagem local, ou seja, as 4 heurísticas locais (vide Seção 3.2) foram executadas e duas combinações destas foram avaliadas. Na segunda etapa, executamos a rotulagem global somente nos atributos aos quais nenhum ou mais de um rótulo foi atribuído na primeira etapa. Isso nos permitiu determinar o ganho obtido através dessa combinação (rotulagem global e local).

4.2 Experimentos com a Rotulagem Local

A primeira etapa de nossos experimentos tem como objetivo a avaliação da rotulagem local. Nesta seção são apresentados os resultados da avaliação de cada heurística local quando usadas de maneira isolada e duas combinações destas.

A primeira combinação, chamada de $C1$, inclui todas as 4 heurísticas locais. Na segunda combinação, chamada de $C2$, as heurísticas $h2, h3$ e $h4$ foram usadas. Experimentos usando a combinação $C2$ foram conduzidos baseados na seguinte observação: os

formulários de busca, em muitos casos, não são necessários para a extração completa de uma base de dados da Web. Embora esses formulários ajudem o usuário na busca de dados mais rapidamente, retornando somente dados relevantes para sua consulta, para uma completa extração dos dados é mais interessante seguir links fornecidos pelos sites, que listam todos os dados, agrupados por categorias por exemplo. Além disso, submeter consultas a formulários tem um custo maior que seguir links, pois para isso é necessário executar um grande quantidade de consultas para que possa ser retornada uma grande quantidade de dados, e outra questão importante é quais os termos que serão usados nas consultas. E mesmo assim, não há nenhuma garantia que a base será completamente extraída. Por isso avaliamos uma combinação das heurísticas locais que não considera heurística $h1$, que supõe que consultas foram submetidas aos formulários para geração das páginas Web. Os resultados obtidos através dessa combinação podem ser usados para estimar o ganho obtido através da nossa abordagem quando é usada em bases de dados que não foram extraídos de páginas Web dinâmicas. Ou seja, a combinação $C1$ reflete um cenário com páginas geradas dinamicamente, enquanto que a combinação $C2$ reflete um cenário onde as páginas usadas para extrair os dados não foram geradas através de consultas a formulários.

As heurísticas locais foram executadas em todos os atributos. Cada heurística pode retornar um rótulo candidato para um atributo anônimo. Na combinação das heurísticas, os rótulos candidatos retornados foram unidos em um único conjunto de rótulos candidatos para um dado atributo. Dessa forma, para cada atributo são retornados dois conjuntos de rótulos candidatos, cada um referente a uma combinação, $C1$ e $C2$. Esse conjunto pode ter um único elemento, mais de um elemento ou pode ser vazio.

Em nossa abordagem, quando o conjunto de rótulos candidatos contém mais de um elemento, a seleção do rótulo candidato mais adequado ao atributo é feita somente na fase global. Por isso, a avaliação nesse caso é feita de maneira diferente, pois ainda não podemos afirmar se a rotulagem local falhou ou obteve sucesso, já que o conjunto possui mais de um rótulo candidato e nenhum desses ainda fora selecionado. Para isso, criamos

uma maneira de medir a acurácia do método local em função da probabilidade do rótulo correto ser escolhido a partir de um seleção aleatória.

Depois de executar a rotulagem local, foi retornado para cada atributo um conjunto de rótulos candidatos. Cada rótulo candidato foi avaliado por um especialista na área e classificado manualmente como correto ou incorreto para um dado conjunto de valores. A probabilidade de um rótulo candidato classificado como correto ser atribuído a um atributo A_i a partir de uma seleção aleatória é calculada da seguinte forma: dado um atributo A_i ao qual foi retornado um conjunto de rótulos candidatos com m rótulos, onde n rótulos foram classificados como correto, então a probabilidade de um rótulo correto ser escolhido é $p(A_i) = n/m$. Por exemplo, supondo que em um conjunto de rótulos candidatos com dois rótulos, um dos rótulos foi classificado como correto e o outro como incorreto, a probabilidade do rótulo correto ser selecionado aleatoriamente é 0.5. A acurácia média para cada base de dados D com conjunto de atributos A e n atributos rotulados é definido na equação seguinte 4.1

$$Acuracia(D) = \sum_{a \in A} \frac{p(a)}{|A|} \tag{4.1}$$

4.2.1 Resultados dos Experimentos com a Rotulagem Local

A Tabela 4.3 mostra os valores de acurácia para cada heurística local e as combinações $C1$ e $C2$. Os resultados mostram que a acurácia de cada heurística local quando usada de forma isolada é baixa, mas quando todas elas são usadas em conjunto a acurácia alcançada é 88.35%. Quando comparamos ambas combinações, $C2$ obteve o menor valor de acurácia porque a evidência usada na heurística 1 não foi considerada. O alto valor de acurácia de $C1$ mostra que o melhor resultado é obtido combinando todas as heurísticas.

A Tabela 4.4 apresenta os valores de acurácia de cada base de dados para as combinações $C1$ e $C2$. Ela nos mostra que a acurácia média de $C1$ é alta, mas acreditamos que os resultados podem ser melhorados através da combinação com o método global. Nossa hipótese é baseada na seguinte observação: nas bases de dados em que a rotulagem

heurísticas	acurácia média
h1	19.63%
h2	39.69%
h3	26.15%
h4	28.73%
C1	88.15%
C2	75.66%

Tabela 4.3: Valores de acurácia média das heurísticas locais para todas as 21 bases de dados.

local falhou, as páginas web não continham rótulos explícitos e os campos dos formulários disponíveis não foram suficientes para rotular corretamente todos os atributos. Por exemplo, na base de dados *www.allmovie.com*, um atributo não foi rotulado porque as páginas não continham nenhum rótulo para esse atributo que corresponde ao gênero de filmes e não havia nenhum campo de busca correspondente no formulário utilizado para gerar as páginas.

domínio	base de dados	**acurácia média**	
		C1	C2
Adress book	people.yahoo.com	100%	100%
books	www.ebay.com	100%	20.00%
	www.amazon.com	92.86%	71.43%
Files	www.democrats.org	100%	100%
finance	finance.yahoo.com	100%	100%
games	www.allgames.com	40.00%	20.00%
	www.allgame.com	100%	80.00%
job	www.monster.com	100%	100%
	www.jobbankusa.com	100%	100%
	www.careerbuilder.com	100%	100%
	www.softwarejobs.com	100%	100%
medicines	www.vitacost.com	50.00%	50.00%
movies	www.imdb.com	100%	100%
	www.allmovie.com	80.00%	80.00%
music	www.cdnow.com	87.50%	37.50%
	www.sonybmg.com.br	50.00%	50.00%
	www.amazon.com	90.00%	70.00%
	www.allmusic.com	75.00%	25.00%
posters	www.postershop.com	100%	100%
teams	www.fifaworldcup.com	90.00%	90.00%
watches	www.wacthzone.com	100%	75.00%
Média		88.35%	75.66%

Tabela 4.4: Valor de acurácia média para cada base de dados usando **C1** and **C2**.

A Tabela 4.5 mostra os valores de acurácia média de cada domínio usando as combinações $C1$ e $C2$. Podemos ver que 6 domínios alcançaram o valor máximo em $C1$: *address book, books, files, finance, job, posters* e *watches*. O domínio *medicines* é o domínio que contém o menor valor de acurácia, dos 7 atributos no total, somente 3 foram rotulados corretamente, um rotulado incorretamente, 2 não foram rotulados e um atributo recebeu dois rótulos candidatos (um correto e um incorreto). Dois atributos não foram rotulados porque não havia rótulos nas páginas, nenhuma evidência de formulário e possuía valores sem padrão.

domínios	**acurácia**	
	C1	C2
address book	100%	100%
books	96.43%	67.00%
files	100%	100%
finance	100%	100%
games	70.00%	50.00%
job	100%	100%
medicines	50.00%	50.00%
movies	90.00%	90.00%
music	75.63%	50.00%
poster	100%	100%
teams	90.00%	90.00%
watches	100%	75.00%

Tabela 4.5: Valor de acurácia da rotulagem local para cada domínio.

Para mostrar a importância do uso da fase global após a fase local, analisamos a distribuição do tamanho dos conjuntos de rótulos candidatos obtidos com a combinação das heurísticas locais. A Tabela 4.6 mostra esta distribuição em três classes para as duas combinações. A situação mais interessante para aplicação do método global é quando nenhum rótulo candidato é retornado para um atributo. Isto implica que o uso das evidências locais não foram suficientes para rotular estes atributos e o método global usará evidências globais para fazer isto. Em $C1$ esta situação representa 7.89% e em $C2$ ela aumenta pra 19.30%. Para 73.68% dos atributos em $C1$ e 70.17% em $C2$, somente um rótulo candidato foi selecionado. Para 18.42% e 10.52% dos atributos em $C1$ e $C2$ respectivamente mais de um rótulo foi retornado, sendo necessária a segunda etapa do

método global para selecionar o rótulo mais apropriado deste conjunto. A Tabela 4.7 apresenta os valores de acurácia para as combinações *C1* e *C2* quando somente um rótulo foi selecionado na fase local. Os altos valores de acurácia apresentados mostram que nesses casos não há necessidade do método global ser executado para validar o resultado, pois a grande maioria dos rótulos retornados são corretos para os atributos. Em *C1* 96.42% dos rótulos que foram retornados são adequados para os atributos correspondentes e em *C2* isso representa 96.25%.

|L|	**C1**	**C2**
0	7.89%	19.30%
1	73.68%	70.17%
>1	18.42%	10.52%

Tabela 4.6: Distribuição de atributos de acordo com o tamanho do conjunto de rótulos candidatos gerados a partir da fase local.

|L|	**C1**	**C2**
1	96.42%	96.25%

Tabela 4.7: Valor de acurácia de *C1* e *C2* para $|L| = 1$.

4.3 Experimentos com a Rotulagem Global

Na segunda parte de nossos experimentos, verificamos a eficácia de nosso método de rotulagem global no conjunto de atributos anônimos que não foram rotulados na fase local. Além disso, calculamos o ganho obtido com o uso das duas abordgens quando usadas em conjunto sobre o uso somente da abordagem local.

Nestes experimentos, todas as consultas foram submetidas à máquina de busca Yahoo! através de uma API de busca pública. O método de rotulagem global foi executado 30 vezes para cada atributo contido na coleção. Também, executamos o método de rotulagem global em dois diferentes conjuntos de rótulos candidatos, um criado por $C1$ e o outro por $C2$. Os rótulos que foram retornados através desse método foram avaliados manualmente.

4.3.1 Resultados dos Experimentos com a Rotulagem Global

Avaliamos a desempenho do método de rotulagem global nas situações em que ele é aplicado: ($|L| = 0$ e $|L| > 1$).

A Tabela 4.8 mostra os valores de acurácia obtidos através da aplicação do método global nos conjuntos de rótulos candidatos criados através de $C1$ e $C2$. Podemos observar que em ambas combinações quando o método global é aplicado no conjunto de rótulos candidatos com mais que um rótulo ($|L| > 1$), o rótulo correto é sempre selecionado. Na primeira situação ($|L| = 0$), com a combinação $C1$, o método rotulou 61.16% dos atributos anônimos que correspondem a 7.89% da coleção completa. Em $C2$, a acurácia é 70.46% para 19.30% da coleção.

|L|	**acurácia média**	
	C1	C2
0	61.16%	70.46%
>1	100%	100%

Tabela 4.8: Acurácia do método global usando $C1$ e $C2$.

As Tabelas 4.9 e 4.10 apresentam a acurácia obtida após a fase de rotulagem global de cada base de dados que obteve um valor de acurácia menor que 100% na fase local, e o ganho obtido com a execução do método global sobre o método local com a combinação $C1$ e $C2$ respectivamente.

Em $C1$, seis bases de dados obtiveram ganho na acurácia. O menor ganho foi obtido na base de dados *www.allgames.com com* 4.00%. Essa base de dados contém 5 atributos, na fase local foram rotulados 3 atributos, 2 corretamente e 1 incorretamente, então com acurácia 40.00% usando somente a rotulagem local. Já na fase global, 2 atributos devem ser rotulados ($A1$ e $A5$). O atributo $A1$ corresponde ao título de jogos de videogame, entretanto somente 0.2% dos rótulos selecionados em 30 execuções estavam corretos. Em 73.00% das execuções, este atributo foi rotulado erroneamente como *review*, isso pode ser explicado pelo uso freqüente deste termo junto a títulos de jogos de videogame, pois é comum na Web usuários destes jogos fazerem comentários e avaliações a respeitos de tais jogos, e isto é identificado comumente de *review*. O atributo $A2$ corresponde ao sub-gênero

dos jogos. Este atributo foi rotulado incorretamente em todas as execuções na fase global. Todas as vezes ele foi rotulado como *sports* já que os valores de dados deste atributo são em sua maioria nomes de esportes *(football, volleyball, golf, etc)*.

Quatro bases de dados alcançaram 100% de acurácia com a aplicação do método global: *www.amazon.com* no domínio *books* e *music, www.cdnow.com* e *www.fifaworldcup.com*. O maior ganho (46.96%) foi obtido através da base de dados *www.sonybgm.com.br*. A base de dados *www.vitacost.com* apresentou um valor de 64.57%, mas é importante notar que este domínio foi classificado como não-popular e mesmo assim obteve um ganho de 14.57%. O ganho da acurácia média obtida sobre todas as bases de dados é 5.93% (Média 1). Quando analisamos somente as bases de dados com o valor de acurácia menor que 100% na fase local, ou seja, as bases de dados em que o método global foi de fato utilizado, a média obtida é de 13.83% (Média 2).

site	acurácia	ganho
www.amazon.com	100%	7.14%
www.allgames.com	44.00%	4.00%
www.vitacost.com	64.57%	14.57%
www.allmovie.com	99.33%	19.33%
www.cdnow.com	100%	12.50%
www.sonybmg.com.br	96.96%	46.96%
www.amazon.com	100%	10.00%
www.allmusic.com	75.00%	0%
www.fifaworldcup.com	100%	10.00%
Média 1	94.28%	5.93%
Média 2	86.65%	13.83%

Tabela 4.9: Valores de acurácia para cada base de dados usando o método global e o ganho obtido utilizando *C1*.

Em $C2$ o valor de acurácia obtido na fase local é de 75.66% para todas as bases e quando foram utilizadas as duas abordagens a acurácia alcançou 91.68%, ou seja, tivemos um ganho de 16.02% (Média 1). Esse resultado mostra que nossa abordagem teve um bom desempenho na base utilizada mesmo quando não temos informações referentes aos formulários utilizados na geração das páginas dinâmicas ou quando as páginas não foram geradas através de formulários. Das 21 bases de dados utilizadas, 12 tiveram um ganho

site	acurácia	ganho
www.ebay.com	66.64%	26.64%
www.amazon.com	84.66%	13.33%
www.allgames.com	51.43%	31.43%
www.allgame.com	90.00%	10.00%
www.vitacost.com	64.57%	14.57%
www.allmovie.com	99.33%	19.33%
www.cdnow.com	99.00%	61.50%
www.sonybmg.com.br	96.96%	46.96%
www.amazon.com	99.33%	29.33%
www.allmusic.com	74.17%	49.17%
www.fifaworldcup.com	100%	10.00%
www.wacthzone.com	99.17%	24.17%
Média 1	91.68%	16.02%
Média 2	85.51%	30.62%

Tabela 4.10: Valores de acurácia para cada base de dados usando o método global e o ganho obtido utilizando *C2*.

nos resultados em relação ao uso somente da rotulagem local. O menor ganho obtido foi de 10.00% na base de dados *www.allgame.com* mas mesmo assim obteve 90.00% de acurácia com o método global, o que representa um bom resultado. Quando analisamos somente as bases de dados onde o método global foi aplicado temos um ganho de 30.62% (Média), o que reforça nossa hipótese de que o uso das duas abordagens em conjunto, e do uso da abordagem global para selecionar o rótulo mais adequando dentre os rótulos candidados selecionados na fase local é válido e viável.

Capítulo 5

Conclusão

Neste trabalho apresentamos uma nova abordagem para rotulagem automática de dados extraídos da Web que utiliza tanto evidências locais quanto globais. Nossos experimentos demonstram que a proposta é eficiente ao atribuir rótulos descritivos a atributos anônimos.

Recentemente muitas ferramentas de extração de dados de páginas web têm sido propostas. Porém, a maioria destas identifica somente a estrutura implícita dos dados e não atribuem nenhuma semântica a eles. Isso limita severamente o uso desses dados por aplicações, onde o significado dos atributos é essencial para a manipulação correta por parte de tais aplicações, e a rotulagem manual desses dados gera um esforço humano que compromete o desempenho de dessas ferramentas. Portanto, é necessário o uso de técnidas de rotulagem de dados que torne o processo de extração e rotulagem tarefas totalmente automática.

Para avaliar o desempenho de nossa abordagem, realizamos experimentos utilizando 21 conjuntos de dados anônimos extraídos de sites da Web. Utilizamos 12 domínios distintos classificados como populares e não-populares. Os experimentos foram divididos em duas etapas: a primeira avalia o desempenho de heurísticas locais quando usadas de forma isolada na segunda fase de nosso método (Método Global) e a segunda etapa avalia o uso de nossa abordagem, ou seja, os métodos local e global de forma combinada. Dessa forma, podemos medir o ganho obtido com o uso de nossa combinação em relação aos principais métodos que utilizam apenas evidências locais.

Foram usadas 4 heurísticas locais na primeira etapa dos experimentos e avaliados dois tipos de combinações destas (*C1* e *C2*). A acurácia média da rotulagem local, utilizando todos os conjuntos de dados, foi de 88.15% para *C1* e 75.66% para *C2*, mostrando que esses métodos são eficientes quando exploram as evidências contidas nas páginas de onde os dados extraídos, nos formulários de busca utilizados para gerar tais páginas e nos próprios valores de dados. Entretanto, essas heurísticas falham para atributos que não possuem nenhuma dessas evidências ou é retornado mais de um rótulo para o mesmo. Dessa forma, o método global é usado para tentar cobrir esses casos.

Nossos melhores resultados com o método global aplicado de acordo com nossa abordagem é de 94.68% de acurácia média utilizando os 21 conjuntos de dados e a combinação *C2*, ou seja, um ganho de 16.02% para a combinação *C2*. E quando analisamos somente os conjuntos de dados que tiveram a acurácia média na fase local inferior a 100%, ou seja, quando o método global não é necessário, o ganho é de 30.62%. Com esses resultados é possível afirmar que nossa abordagem é totalmente válida e eficiente para rotular atributos anônimos de forma totalmente automática.

Referências Bibliográficas

[Arasu and Garcia-Molina, 2003] Arasu, A. and Garcia-Molina, H. (2003). Extracting structured data from web pages. In *Proc. of SIGMOD*, pages 337–348, San Diego, CA, USA.

[Arlotta et al., 2003] Arlotta, L., Crescenzi, V., Mecca, G., and Merialdo, P. (2003). Automatic annotation of data extracted from large web sites. In *WebDB*, pages 7–12.

[Buttler et al., 2001] Buttler, D., Liu, L., and Pu, C. (2001). A Fully Automated Object Extraction System for the World Wide Web. In *International Conference on Distributed Computing Systems*, pages 361–370.

[Crescenzi et al., 2001] Crescenzi, V., Mecca, G., and Merialdo, P. (2001). Roadrunner: Towards automatic data extraction from large web sites. In *Proc. of VLDB*, pages 109–118, Rome, Italy.

[da Silva et al., 2007] da Silva, A. S., Barbosa, D., Cavalcanti, J. M. B., and Sevalho, M. (2007). Labeling data extracted from the web. In *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS*, volume 4803/2007, pages 1099–1116. Springer Berlin / Heidelberg.

[de Sá Júnior et al., 2006] de Sá Júnior, S. A. L. F., da Silva, A. S., and Oliveira, D. (2006). DESANA: Efficiently publishing relational databases on the web by using keyword-based query interfaces. Sessão de Demonstração.

[Hearst, 1992] Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. pages 539–545.

[Kuhlins and Tredwell, 2002] Kuhlins, S. and Tredwell, R. (2002). Toolkits for generating wrappers. In *LNCS*, pages 184–198. Springer.

[Laender et al., 2002] Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., and Teixeira, J. S. (2002). A brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93.

[Lu et al., 2007] Lu, Y., He, H., Zhao, H., Meng, W., and Yu, C. (2007). Annotating structured data of the deep web. *icde*, 0:376–385.

[Sevalho, 2007] Sevalho, M. (2007). Rotulagem automática de dados anônimos extraídos da web. Dissertação de Mestrado, Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação, UFAM.

[Song et al., 2004] Song, H., Giri, S., and Ma, F. (2004). Data extraction and annotation for dynamic web pages. In *EEE '04: Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04)*, pages 499–502, Washington, DC, USA. IEEE Computer Society.

[Wang and Lochovsky, 2003] Wang, J. and Lochovsky, F. H. (2003). Data extraction and label assignment for web databases. In *Proceedings of the international conference on the World Wide Web*, pages 187–196.

[Zhai and Liu, 2005] Zhai, Y. and Liu, B. (2005). Web data extraction based on partial tree alignment. In *Proceedings of the international conference on the World Wide Web*, pages 76–85, Chiba, Japan.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo