



Deep Web

Ronaldo S. Mello

GBD/INE/CTC/UFSC



Novembro, 2010



Roteiro

1. Introdução

2. Principais Tópicos de Pesquisa

i. Crawling

ii. Extração

iii. Matching

iv. Consulta

3. Algumas Iniciativas

4. Tendências

Referências

Roteiro

1. **Introdução**

2. Principais Tópicos de Pesquisa

- i. Crawling
- ii. Extração
- iii. Matching
- iv. Consulta

3. Algumas Iniciativas

4. Tendências

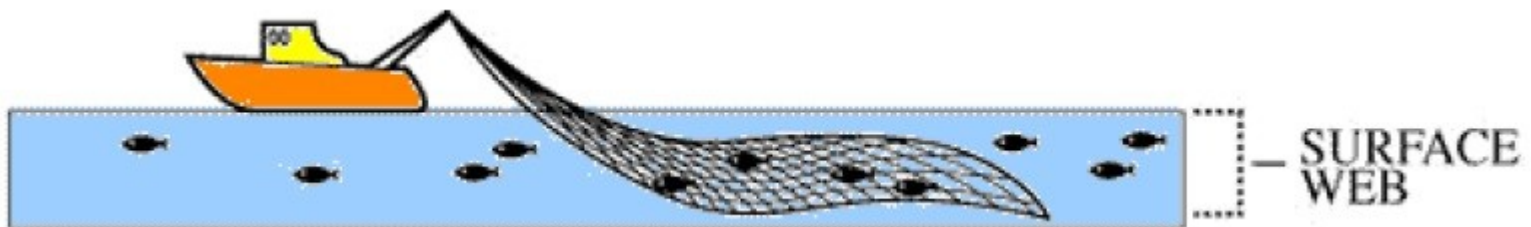
Referências

Dados na Web

- Um “oceano” de conteúdo
- Pesquisa a dados na Web
 - “Atirar uma rede neste oceano”
- Oceano
 - Tem uma **superfície**
 - Alguns animais (dados) são facilmente visíveis
 - Fácil capturá-los com a rede (pesquisá-los)
 - É **profundo**
 - Animais que não são visíveis
 - Difícil encontrá-los e capturá-los

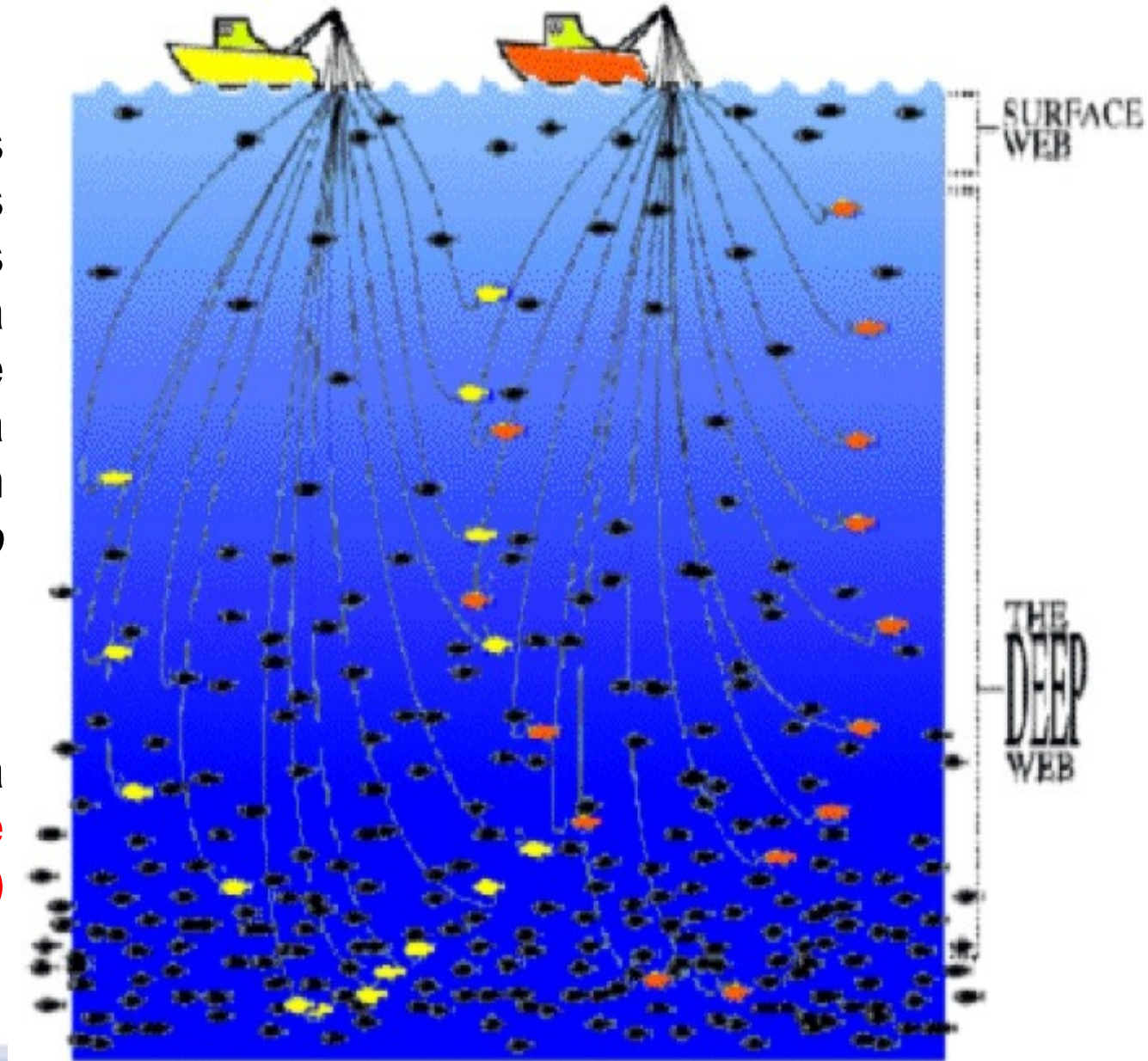
Surface Web (ou Web Visível)

- Dados em páginas Web **estáticas**
 - Dados alcançados pelas máquinas de busca - *search engines* - “barcos pesqueiros”
 - Google, Yahoo!, Bing, ...
- Processo de pesquisa
 - Não é focado em domínio (***keywords***)
 - Dados são facilmente localizados no conteúdo da página ou através de seus *links*

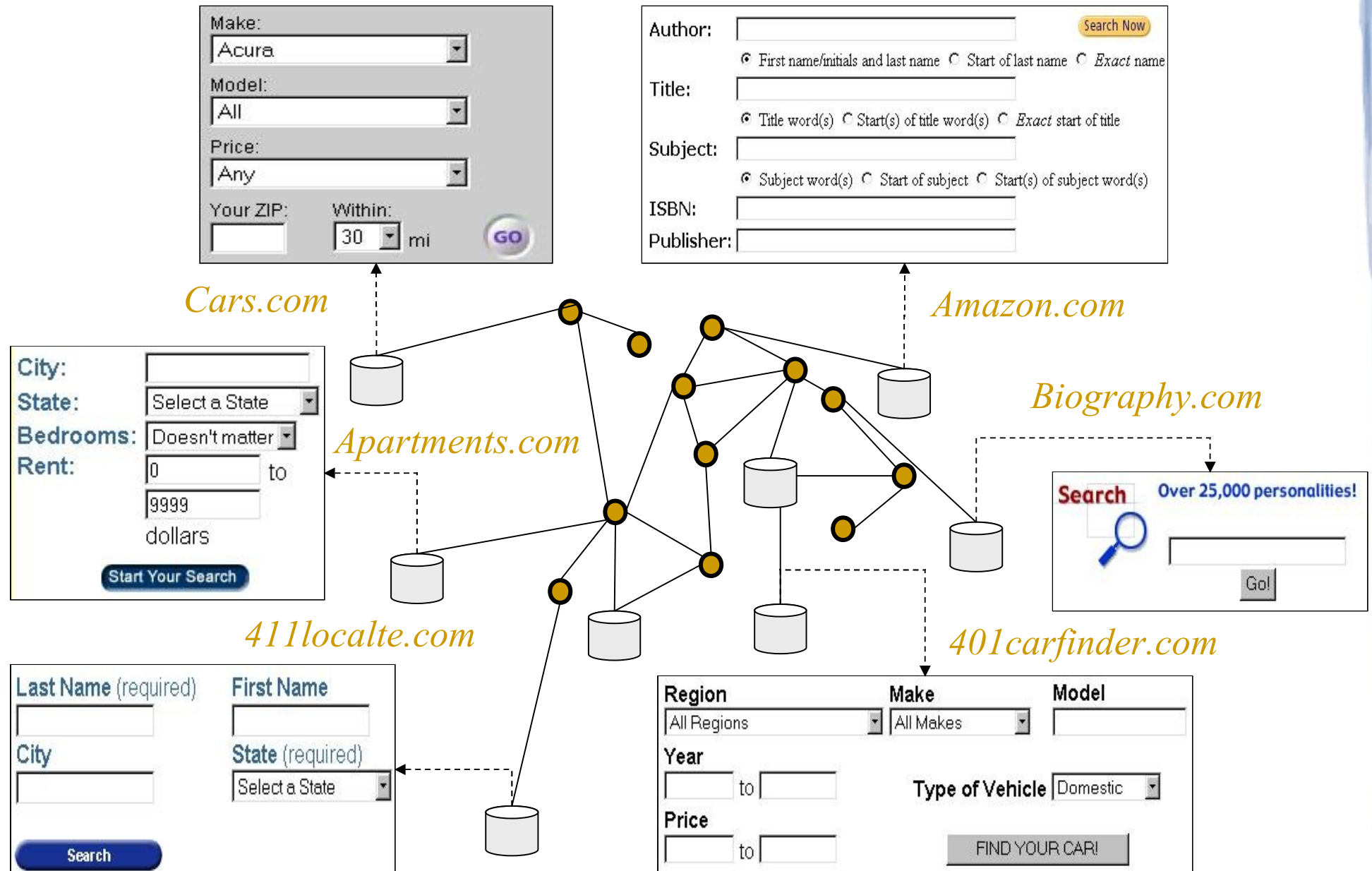


Deep Web (ou Web Escondida)

- Dados **invisíveis**
 - Visíveis apenas quando mostrados em páginas dinâmicas criadas a partir do resultado de uma pesquisa definida sobre um formulário - **Web Form**
- **Web Form (WF)**
 - Interface de pesquisa para um **Banco de Dados (BD)** “escondido” na Web







Deep Web: "Oceano" de Dados & Domínios



Deep Web vs. Surface Web

- Similaridade
 - Ambas crescem rapidamente em diversos domínios
- Diferenças
 - Quadrantes de Kevin Chang

| | <i>Surface Web</i> | <i>Deep Web</i> |
|------------------|--|---|
| <u>Access</u> |  |  |
| <u>Structure</u> |  |  |

Por quê o Interesse pela Deep Web?

- Principal fonte de dados estruturados na Web a disposição
 - **Serviços úteis em diversos domínios!**
 - Companhias aéreas, concessionárias e revendas de veículos, hotéis, classificados, acervos bibliográficos e científicos, ...
 - Exemplo:
 - *Vou mudar de cidade por motivos pessoais e preciso investigar opções baratas para a viagem, aluguel de carro e casa, bem como ofertas de emprego no novo local*
 - Não considerá-los (descobrir e utilizar) é um desperdício!

Por quê o Interesse pela Deep Web?

- Principais **Aplicações**

- Diretórios/catálogos de BDs escondidos (**BDs na Web**) por domínio
- Sistemas de busca de BDs na Web baseados em seus dados/metadados
 - *Preciso comprar um carro. Onde encontro revendas online? Desejo consultar informações de marca, modelo, ano e preço*
- Sistemas integrados de busca/prestação de serviços baseados em BDs na Web
 - *Quero consultar valores de diárias de hotel em Florianópolis (num único site, de preferência...)*
- Busca por WFs similares
 - *Esse formulário de busca de ofertas de emprego é um pouco limitado ou tem poucos dados. Quero acessar outros...*

Deep Web – Algumas Informações

- Deep Web ~2000x maior que Surface Web
 - Não há estimativas atualizadas do seu tamanho...
 - 2001: 7500 Tb (Deep Web) vs. 19 Tb (Visible Web)
- #WFs ~= 25 milhões
- #Deep Web (#BDs na Web) ~= **2.6 milhões**
- 95% da Deep Web é estimada como **pública**
 - Não está sujeita a taxas e registros
- Grandes domínios em ordem de frequência
 - 1) **Serviços** (hotéis, veículos, empregos, previsão do tempo, ...)
 - 2) **Ciência & educação** (bases científicas, sites educacionais, instituições, ...)
 - 3) **Arte & cultura** (cinema, música, eventos, tickets, ...)
 - 4) **Acervo bibliográfico** (conferências, periódicos, ...)

Roteiro

1. Introdução

2. **Principais Tópicos de Pesquisa**

i. Crawling

ii. Extração

iii. Matching

iv. Consulta

3. Algumas Iniciativas

4. Tendências

Referências

Deep Web - Tópicos de Pesquisa

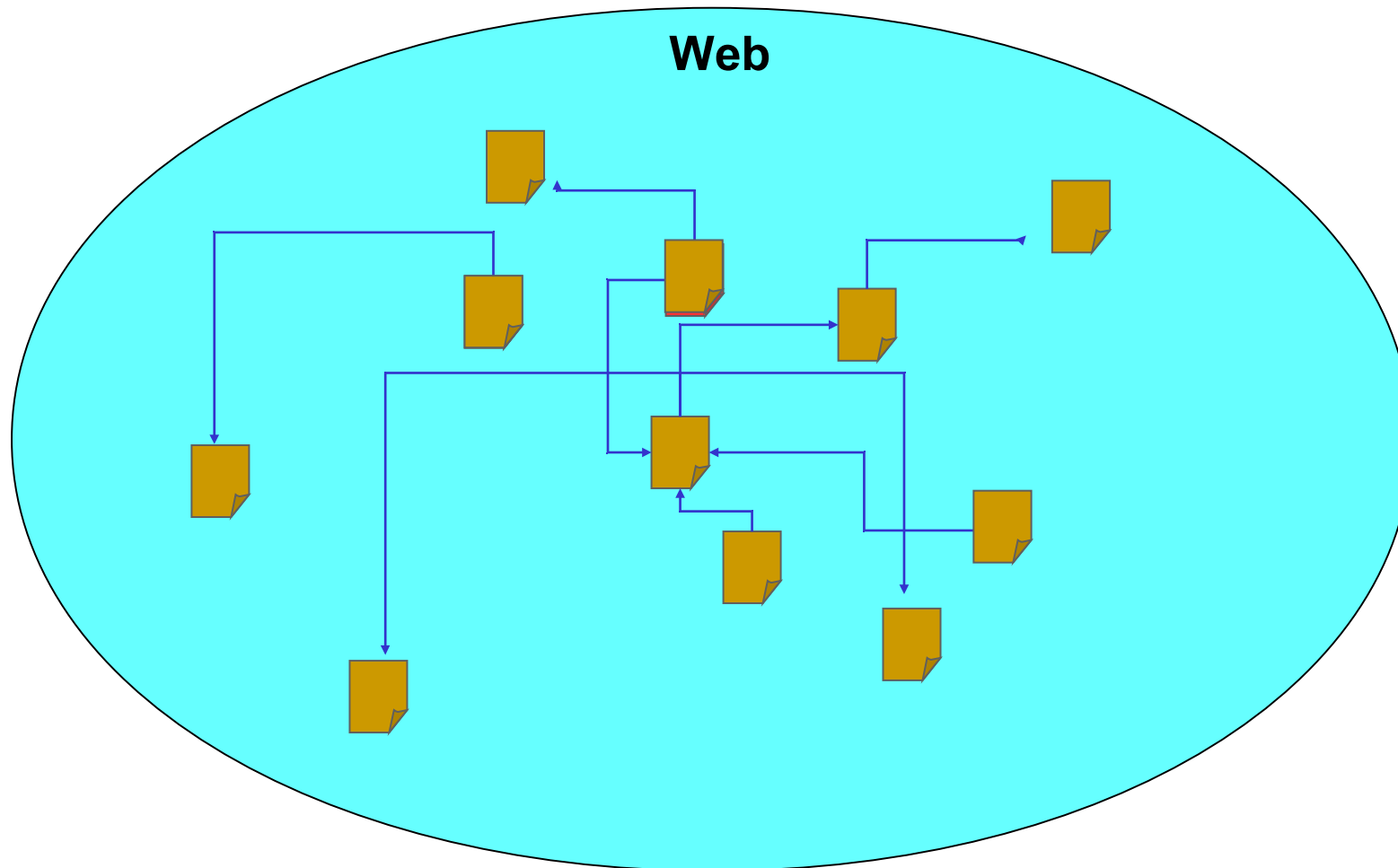
- Implementar aplicações para Deep Web traz desafios de pesquisa para a comunidade de BD
 - Como descobrir onde existem BDs na Web?
 - Deep Web crawling
 - Como descobrir a estrutura/dados dos BDs na Web?
 - Extração de dados da Deep Web
 - Como prover catálogos e serviços integrados por domínio para BDs na Web?
 - Matching (casamento) de dados da Deep Web
 - Como acessar dados de interesse em BDs na Web?
 - Consulta a dados na Deep Web

Roteiro

1. Introdução
2. **Principais Tópicos de Pesquisa**
 - i. **Crawling**
 - ii. Extração
 - iii. Matching
 - iv. Consulta
3. Algumas Iniciativas
4. Tendências
- Referências

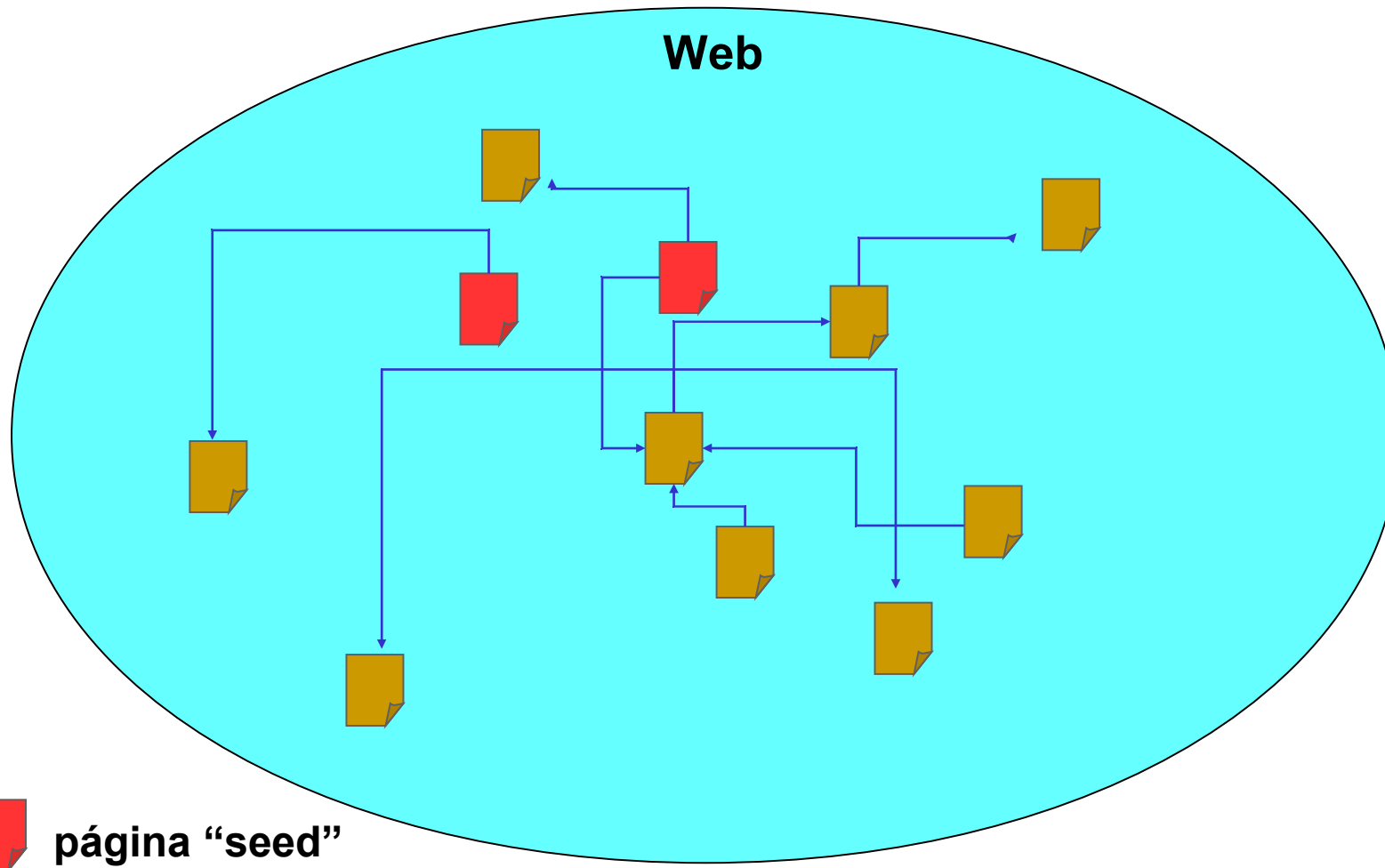
Web Crawling (tradicional)

- Descubierta de datos (estáticos) na Web pelas máquinas de busca



Web Crawling (tradicional)

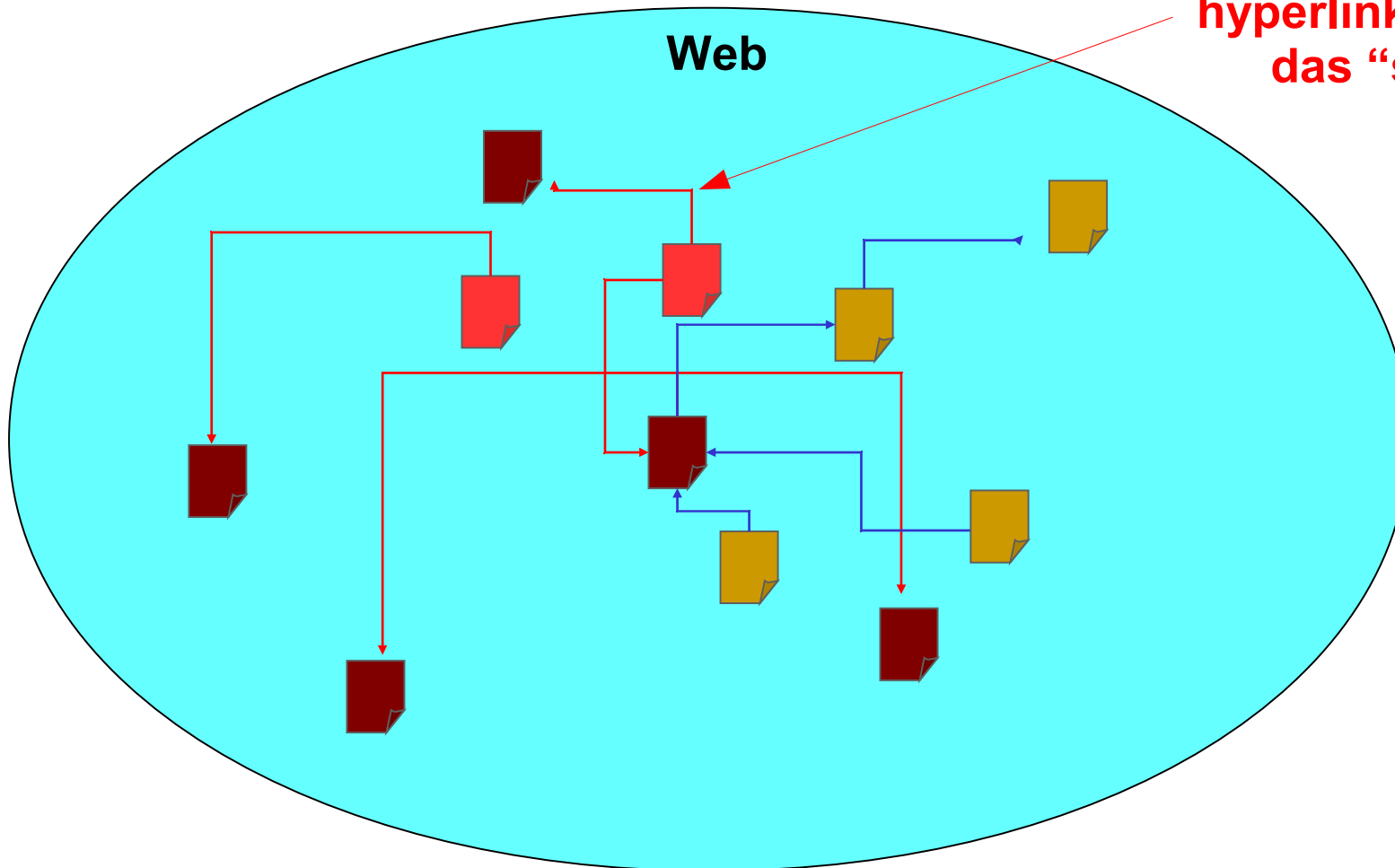
- Descubierta de datos (estáticos) na Web pelas máquinas de busca



Web Crawling (tradicional)

- Descoberta de dados (estáticos) na Web pelas máquinas de busca

Navegação em
hyperlinks a partir
das “seeds”



Web Crawling (tradicional)

- Descoberta de dados (estáticos) na Web pelas máquinas

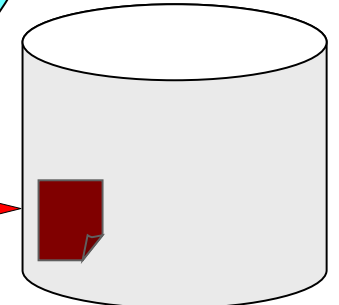


The screenshot shows the website for INE (Informática e Estatística) at CTC-UFSC. The main content is a news article titled "SECCOM 2010" published on October 10, 2010. The article describes the SECCOM event, which is a week of courses and seminars in computer science organized by the PET Computing Group. The event is scheduled for October 18-21 at UFSC in Florianópolis. The article also mentions that the event is open to the public and aims to promote mini-courses and seminars related to computer science.

Índice de termos

| | |
|------------------|---------------------|
| ... | |
| <i>Graduação</i> | (URL1, freq y), ... |
| ... | |
| <i>SECCOM</i> | (URL1, freq x), ... |
| ... | |

BD de páginas
com alto acesso



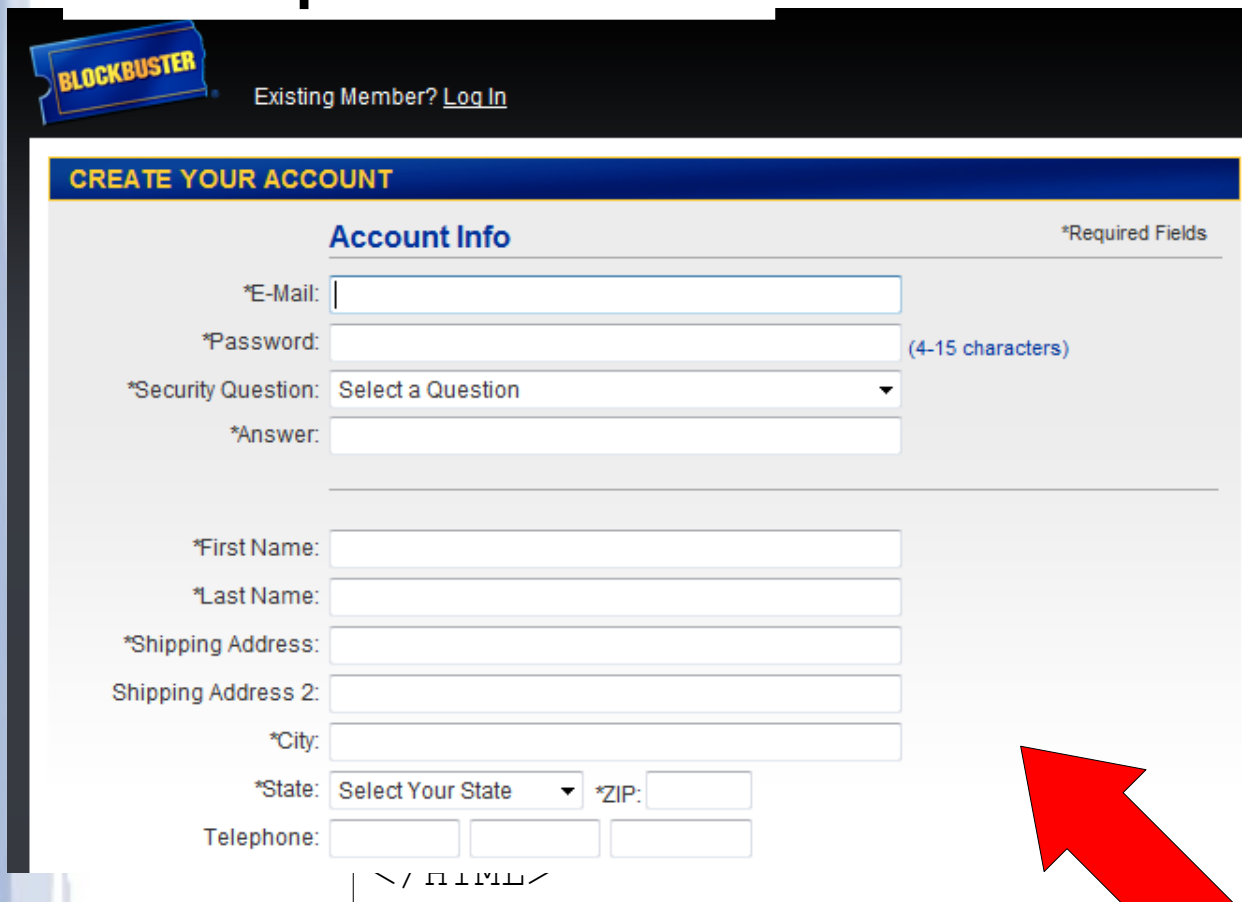
Deep Web Crawling

- O que procurar?
 - Páginas que possuam *forms* para BDs!
- Abordagem mais simples
 - Encontrar páginas HTML com tag `<form>`

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<HTML>    ...
  <FORM ...>
    <LABEL for="firstname">First name: </LABEL>
      <INPUT type="text" id="firstname"><BR>    ...
  </FORM>    ...
</HTML>
```

Deep Web Crawling

Forms para cadastro



BLOCKBUSTER Existing Member? [Log In](#)

CREATE YOUR ACCOUNT

Account Info *Required Fields

*E-Mail:

*Password: (4-15 characters)

*Security Question: Select a Question

*Answer:

*First Name:

*Last Name:

*Shipping Address:

Shipping Address 2:

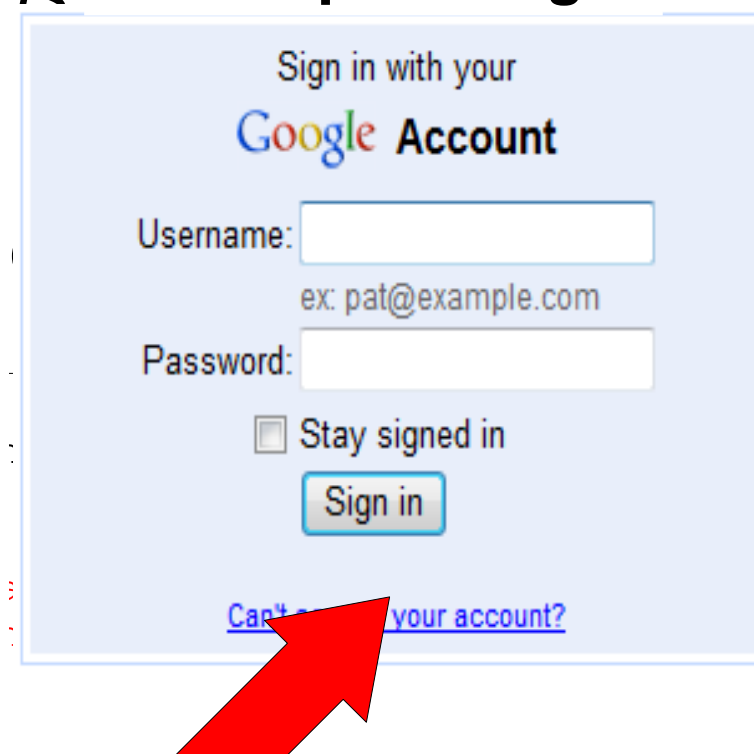
*City:

*State: Select Your State *ZIP:

Telephone:

[\ / | | | | | /](#)

Forms para Login



Sign in with your
Google Account

Username:
ex: pat@example.com

Password:

Stay signed in

[Can't sign in? Create your account?](#)

PROBLEMA: Nem toda WF é uma “porta de entrada” para pesquisa um BD na Web!

Deep Web Crawling

- Necessita-se de *“focused-crawlers”* !
 - *Crawlers* especializados na busca de WFs para BDs (geralmente focados em um domínio)
- Algumas Abordagens
 - Comparação com termos relevantes do domínio
 - Comparação com templates estruturais de WFs
 - Técnicas de aprendizado de máquina (*machine learning*)
 - Aprendizado de características de WFs relevantes a partir de amostras de páginas
 - ...

Abordagem de (Barbosa & Freire, 2007)

- Um **classificador de páginas** de WFs de BDs em um certo domínio
- Aplica *machine learning* (classificação)
 - Aprende **características relevantes** existentes nas WFs
 - Aprimora, a cada *crawling*, estas características
 - Incorpora novas características encontradas
 - Aprende a classificar melhor a relevância das WFs

Passos da Abordagem

1) Treinamento

- Análise do conteúdo e dos *links* (navegação *backward*) que conduzem a páginas "seeds" (amostra)
- Seleção manual de características do domínio
 - atributos (aspectos estruturais relevantes)
 - termos (valores mais significativos do domínio) encontrados nas WFs, âncoras e links

Passos da Abordagem

2) Aprendizado

- Analisa *links* para outras páginas, comparando termos na **âncora** e **palavras próximas** a ela (até uma certa distância) com termos do domínio
 - Aplica *stemming* (“radicalização”) e remoção de *stop words* (artigos, preposições, ...) para facilitar a comparação
 - Exemplo: 'make **of** cars', 'car makes' → 'car make'
- Analisa a **URL** do *link*, verificando se termos significativos aparecem como substrings nela
- Analisa o conteúdo da página apontada pelo *link*, comparando os **atributos** e **valores** na WF com atributos e termos já aprendidos no domínio
- Caso a página seja considerada relevante:
 - Registra a sua URL
 - Cataloga novos aspectos estruturais e termos aprendidos

Exemplo – Domínio de Veículos

- Características conhecidas
 - Termos: *buy, rent, new, used, car, make, model, year, price, from, to*
 - Estruturas:

Make ▾ Model ▾ Year ▾

Make ▾ Model ▾ Price: From ▾ To ▾

URL:

<http://www.cars.com>



Search For a

New Cars

Build or find your car

Make

All Makes ▾

Model

All Models ▾

Maximum Price

No Max ▾

Search Within

30 miles ▾ of

Your ZIP

Search New

Exemplo – Domínio de Veículos

- Características conhecidas

- Termos: *buy, rent, new, used, car, make, model, year, price, from, to*

- Estruturas:

Make ▾ Model ▾ Year ▾

Make ▾ Model ▾ Price: From ▾ To ▾

URL:

<http://www.cars.com>



Search For a

New Cars

Build or find your car.

Make

All Makes ▾

Model

All Models ▾

Maximum Price

No Max ▾

Search Within

30 miles ▾ of

Your ZIP

Search New

Exemplo – Domínio de Veículos

- Características conhecidas

- Termos: *buy, rent, new, used, car, make, model, year, price, from, to*

- Estruturas:

Make Model Year

Make Model Price: From To

Aprendizado:

Termos: *ZIP, miles, ...*

Estrutura: ... Maximum Price , ..., Your Zip

URL:

<http://www.cars.com>



Search For a

New Cars

Build or find your car

Make

All Makes

Model

All Models

Maximum Price

No Max

Search Within

30 miles of

Your ZIP

Search New

Exemplo – Domínio de Veículos

- Características conhecidas

- Termos: *buy, rent, new, used, car, make, model, year, price, from, to*

- Estruturas:

Make ▾ Model ▾ Year ▾

Make ▾ Model ▾ Price: From ▾ To ▾

URL:

<http://www.cars.com>



Search For a

New Cars

Build or find your car

Make

All Makes ▾

Model

All Models ▾

Maximum Price

No Max ▾

Search Within

30 miles ▾ of

Your ZIP

Search New

PROBLEMAS:

- Não considera sinônimos na comparação
Ex.: *brand ~ make* e *manufacturer ~ make*
- Domínios com WFs muito heterogêneas
(muitos *templates* possíveis...)

Roteiro

1. Introdução

2. **Principais Tópicos de Pesquisa**

i. Crawling

ii. **Extração**

iii. Matching

iv. Consulta

3. Algumas Iniciativas

4. Tendências

Referências

Deep Web - Extração

- Aquisição e catalogação de informações relevantes sobre os BDs na Web
- Tipos de extração
 - **Metadados** (WFs)
 - Atributos e restrições (valores e dependências)
 - **Dados**
 - Conteúdo “escondido” nos BDs

Extração de Metadados

- Abordagem mais simples
 - Analisa a tag `<label>` das WFs nas páginas HTML, extraíndo informação delas

```
<HTML> ...
  <FORM> ...
    <LABEL for="example_text_1">Name:</label>
      <INPUT type="text" name="text_1" id="text_1" /> ...
    <LABEL for="example_select_1">State:</label>
      <SELECT name="select_1" id="select_1">
        <option>AK</option>
        <option>AL</option>
        <option>AR</option> ...
      </SELECT> ...
  </FORM> ...
</HTML>
```

restrições de valor de atributos

rótulos de atributos

Extração de Metadados

- Abordagem mais simples
 - Analisa a tag `<label>` das WFs nas páginas HTML, extraíndo informação delas

```
<HTML> ...
  <FORM> ...
    <LABEL for="example_text_1">Name:</label>
      <INPUT type="text" name="text_1" id="text_1" /> ...
    <LABEL for="example_select_1">State:</label>
      <SELECT name="select_1" id="select_1">
        <option>AK</option>
        <option>AL</option>
        <option>AR</option> ...
      </SELECT>
```

PROBLEMAS:

- Nem todo campo de uma WF possui `<label>`
- Nem sempre é fácil descobrir o rótulo e as restrições de um campo em um código HTML!

Extração de Metadados

```
<HTML> ...  
<FORM ...> ...  
  Name: <INPUT type="text" name="customerName"> ...  
  eMail: <script language="JavaScript" ...> ... </script>  
    ... (required field) ...  
    <INPUT type="text" name="email"> ...  
</FORM> ...  
</HTML>
```

```
<HTML> ...  
<FORM ...> ...  
  <LABEL for="makeid">Make:</label>  
    <SELECT name="makeid" id="makeid"  
      onchange="popMakes ();">  
    </SELECT> ...  
</FORM> ...  
</HTML>
```

PROBLEMAS:

- Nem todo campo de uma WF possui <label>
- Nem sempre é fácil descobrir o rótulo e as restrições de um campo em um código HTML!

Extração de Metadados

- Algumas abordagens
 - Comparação com **termos** do domínio
 - Análise do **layout** da WF e inferência do nome do atributo
 - Proximidade rótulo-campo, tamanho e estilo de fonte, ...
 - Técnicas de **machine learning**
 - Aprendizado da estrutura e da terminologia de atributos de WFs em um certo domínio
 - Análise de **dependências** entre atributos
 - *Make* e *Model* são atributos que em geral aparecem juntos, pois *Make* → *Model*. Se encontrei um (1) deles, provavelmente encontrarei o outro
 - ...

Abordagem de (Alvarez et al., 2007)

- Abordagem para extração de atributos de WFs
- Estratégias:
 - análise do *layout* da WF
 - comparação com termos do domínio
- Compara o conteúdo de cada WF com um *template do domínio* que descreve
 - atributos (nomes, sinônimos e peso)
 - valores (termos mais comuns em buscas)

Exemplo de *Template* de Domínio

Domain "Books"

Attributes: $A = \{a_1, a_2, a_3, a_4, a_5\}$

| Attribute | Name | Aliases | s_i (specificity index) |
|-----------|--------|--------------------------------|---------------------------|
| a_1 | TITLE | 'entitle', 'title of the book' | 0.6 |
| a_2 | AUTHOR | 'writer', 'written by' | 0.7 |
| a_3 | ISBN | | 1 |
| a_4 | FORMAT | 'binding type' | 0.25 |
| a_5 | PRICE | 'cost of book' | 0.05 |

Queries: $Q = \{q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8\}$

| |
|--|
| $q_1 = \{ (TITLE, 'java'), (FORMAT, 'hardcover') \}$ |
| $q_2 = \{ (TITLE, 'xml'), (AUTHOR, 'Priscilla Walmsley') \}$ |
| $q_3 = \{ (TITLE, 'j2ee') \}$ |
| $q_4 = \{ (TITLE, 'concurrent programming') \}$ |
| $q_5 = \{ (TITLE, 'ejb3') \}$ |
| $q_6 = \{ (TITLE, 'java server faces') \}$ |
| $q_7 = \{ (AUTHOR, 'Herbert Schildt') \}$ |
| $q_8 = \{ (TITLE, 'web services') \}$ |

Relevance threshold: $\mu = 0.9$

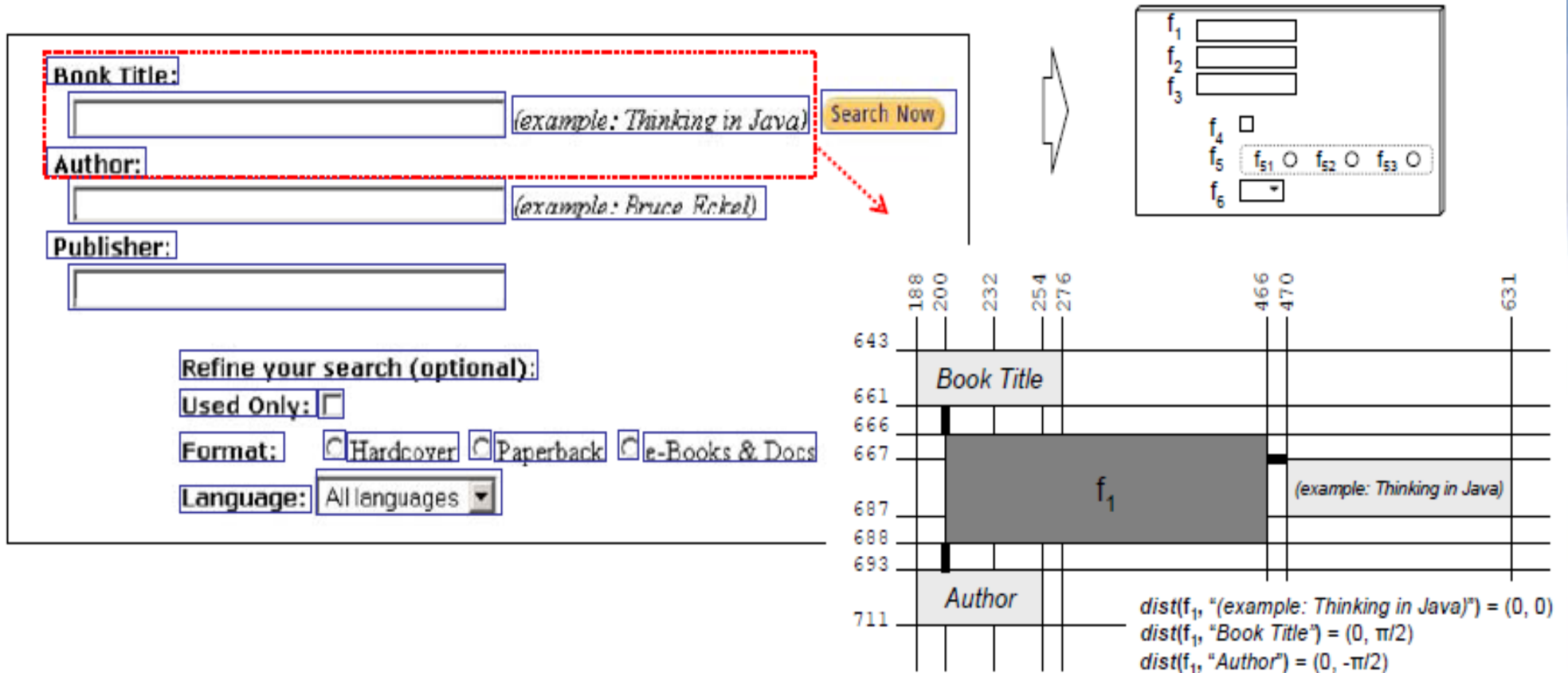
Descoberta de Atributos de WFs

1) Determina a relevância de palavras próximas a campos com base em

- **Distância mínima** no *layout* da WF
- **Ângulo** no *layout* da WF (preferência por posições *left / top*)
- **Comparação de nomes** e sinônimos no *template*
- **Comparação de valores** válidos (caso existam)
- **Similaridade** das palavras
 - Métricas de similaridade: TF-IDF + Jaro-Winkler
 - TF-IDF: importância de um termo nas WFs de mesmo domínio (sua frequência)
 - Jaro-Winkler: similaridade de *strings*

2) Ranking de associações candidatas e “poda” com base em um *threshold*

Descoberta de Atributos - Análise de *Layout*



Campo 1 (f_1): apesar da string "(example: Thinking in Java)" estar ligeiramente mais próxima que a string "Book Title", a string "Book Title" está posicionada em um ângulo mais adequado (posicionamento mais usual)

Descoberta de Atributos - Resultados

Domain "Books"

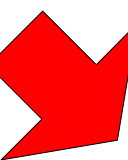
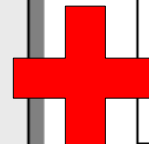
Attributes: $A = \{a_1, a_2, a_3, a_4, a_5\}$

| Attribute | Name | Aliases | s_i (specificity index) |
|-----------|--------|--------------------------------|---------------------------|
| a_1 | TITLE | 'entitle', 'title of the book' | 0.8 |
| a_2 | AUTHOR | 'writer', 'written by' | 0.7 |
| a_3 | ISBN | | 1 |
| a_4 | FORMAT | 'binding type' | 0.25 |
| a_5 | PRICE | 'cost of book' | 0.05 |

Queries: $Q = \{q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8\}$

| |
|--|
| $q_1 = \{(TITLE, 'java'), (FORMAT, 'hardcover')\}$ |
| $q_2 = \{(TITLE, 'xml'), (AUTHOR, 'Priscilla Walmsley')\}$ |
| $q_3 = \{(TITLE, 'j2ee')\}$ |
| $q_4 = \{(TITLE, 'concurrent programming')\}$ |
| $q_5 = \{(TITLE, 'ejb3')\}$ |
| $q_6 = \{(TITLE, 'java server faces')\}$ |
| $q_7 = \{(AUTHOR, 'Herbert Schildt')\}$ |
| $q_8 = \{(TITLE, 'web services')\}$ |

Relevance threshold: $\mu = 0.9$



Book Title: (example: Thinking in Java)

Author: (example: Bruce Rockel)

Publisher:

Refine your search (optional):

Used Only:

Format: Hardcover Paperback e-Books & Docs

Language: All languages

| Assignment | Form Field | Domain Attribute | c_i (confidence) |
|------------|------------|-----------------------|--------------------|
| A_1 | f_1 | $a_1 = \text{TITLE}$ | 0.71 |
| A_2 | f_2 | $a_2 = \text{AUTHOR}$ | 1 |
| | f_3 | (unassigned) | |
| | f_4 | (unassigned) | |
| A_3 | f_5 | $a_4 = \text{FORMAT}$ | 1 |
| | f_6 | (unassigned) | |

Descoberta de Atributos - Resultados

Domain "Books"

Attributes: $A = \{a_1, a_2, a_3, a_4, a_5\}$

| Attribute | Name | Aliases | s_i (specificity index) |
|-----------|--------|--------------------------------|---------------------------|
| a_1 | TITLE | 'entitle', 'title of the book' | 0.6 |
| a_2 | AUTHOR | 'writer', 'written by' | 0.7 |

Book Title:

(example: Thinking in Java)

Author:

(example: Bruce Rockel)

Publisher:

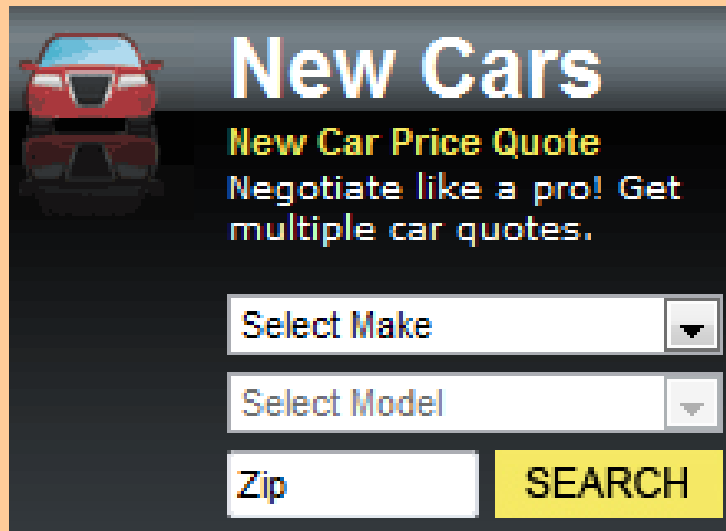
optional):

Format: Paperback E-Books & Docs

Pages:

PROBLEMAS:

- Dependência do *template* de domínio (que é fixo!)
- A abordagem falha em casos como este (rótulo do atributo dentro dos valores permitidos!):



| Main Attribute | c_i (confidence) |
|----------------|--------------------|
| TITLE | 0.71 |
| AUTHOR | 1 |
| (unassigned) | |
| (unassigned) | |
| FORMAT | 1 |
| | f_6 |
| | (unassigned) |

Extração de Dados

- Problema: não se sabe quantos dados existem, pois o BD está “escondido”!
 - Principal questão:
 - Qual o conjunto adequado e mínimo de consultas a serem submetidas nas WFs para cobrir todos os dados do BD?
 - Compromisso: cobertura vs. #consultas
- Tema ainda em aberto
 - Poucas soluções na literatura...

Abordagem de (Halevy et al., 2008)

- Abordagem utilizada em experimentos para indexação da Deep Web pela Google
- Estratégia:
 - Define *templates de consulta (TCs)*
 - Subconjunto de atributos da WF
 - Verifica se um TC é “**informativo**”
 - Testam combinações de valores e verificam a cobertura dos resultados
 - TC é informativo se $D / S > \text{threshold}$, onde
D = #conjuntos “fortemente” distintos de resultado
(leva em conta também volume de dados)
S = #consultas submetidas (≤ 200)

Abordagem de (Halevy et al., 2008)

- Estratégia: (cont.)
 - Inicia com TCs de tamanho 1 (1 atributo) e vai **incrementando** até tamanho 3 (3 atributos)
 - TCs superiores a 3 são muito restritivos
 - Um TC pouco informativo é descartado
 - Exemplo: (*make, year*) é mais informativo que (*make, price*), pois existem mais veículos de um certo ano do que veículos com um certo preço
 - Inicia com valores “seeds” (considerados relevantes) para cada atributo
 - Novos valores relevantes descobertos vão sendo incorporados aos testes
 - Exemplo: inicia-se com '*Ford*' e '*Fiat*' para *make*, mas posteriormente, ao testar outros atributos, descobre-se que há muitos veículos '*Toyota*'. Novos testes são feitos então com *make* = '*Toyota*'

Abordagem de (Halevy et al., 2008)

- Estratégia: (cont.)

- Iniciam com TCs de tamanho 1 (1 atributo) e vão **incrementando** até tamanho 3 (3 atr.)

PROBLEMAS:

- Difícil avaliar a cobertura de qualquer abordagem...
- Valores “seeds” ruins reduzem a cobertura, pois retornam poucos dados no resultado (**chute “ruim” !!**)
- Abordagem fica “pesada” para WFs com muitos atributos
- Dependências entre atributos não foi considerada, no sentido de evitar o teste de combinações inválidas

Exemplo: *make* = 'Fiat' e *model* = 'Focus'

incorporados aos testes

- Exemplo: inicia-se com 'Ford' e 'Fiat' para *make*, mas posteriormente, ao testar outros atributos, descobre-se que há muitos veículos 'Toyota'. Novos testes são feitos então com *make* = 'Toyota'

Roteiro

1. Introdução

2. **Principais Tópicos de Pesquisa**

i. Crawling

ii. Extração

iii. **Matching**

iv. Consulta

3. Algumas Iniciativas

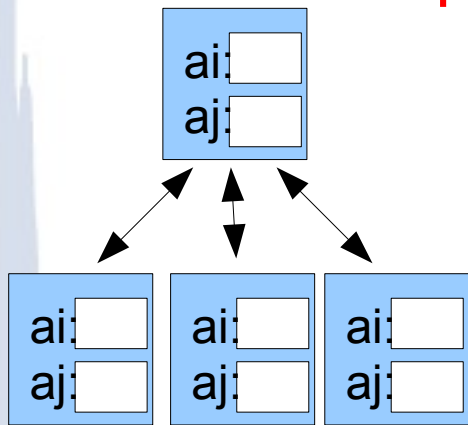
4. Tendências

Referências

Deep Web - Matching

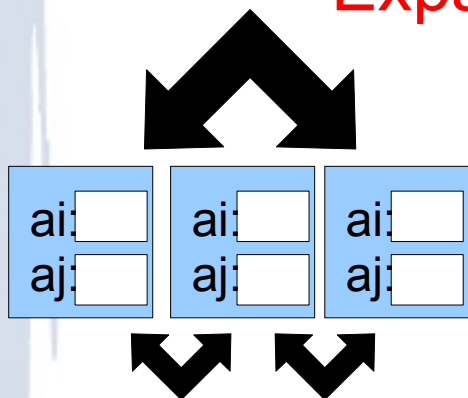
- Foco: casamento de esquemas de WFs
- Enfoques tradicionais de *matching* de BDs são usados

– Esquema global



- Visão integrada (centralizada) de esquemas de WFs
- Distribuição da consulta e integração do resultado
- Vantagens: usuário interage sobre uma visão mais ampla do domínio; mapeamentos apenas 1-N
- Desvantagem: manutenção do esquema global

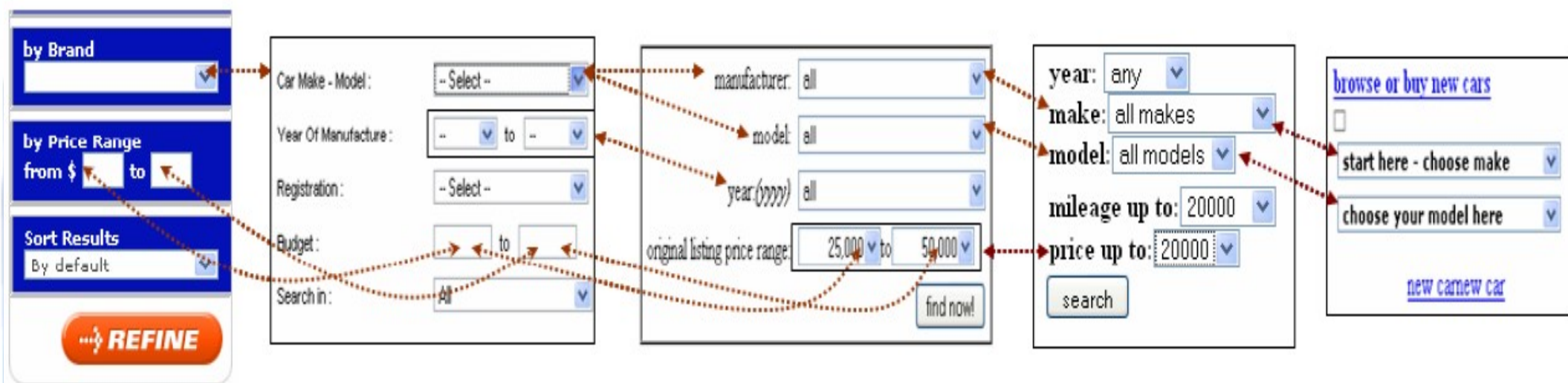
– Expansão de consultas (WFs interoperáveis)



- Consulta em uma WF é propagada para outras WFs
- Resultados apresentados segundo o esquema local
- Vantagem: evita a gestão de um esquema global
- Desvantagens: mapeamentos do tipo M-N

Deep Web - Matching

- Problema fundamental de *matching* em WFs
 - Definição de afinidades entre atributos



- Problema complicado devido a alta heterogeneidade das WFs
 - Mapeamentos 1-N
 - Diversidade de nomenclaturas e de restrições de valores
 - Dicotomia atributo-valor
 - ...

Deep Web - *Matching*

- Técnicas aplicadas (não-exclusivas)
 - **Dicionários** gerais (*Wordnet*) e de domínio
 - Funções de **similaridade** para atributos (rótulos, tipos de dados inferidos e valores permitidos)
 - **Correlações**
 - Análise da co-ocorrência / proximidade de conjuntos de atributos em WFs
 - Exemplo: *make* e *model*

Abordagem de (Nguyen et. al, 2008)

- Definição de grupos (*clusters*) de atributos similares em WFs distintas
- Principais passos:
 - a) Pré-processamento: *stemming* e eliminação de "stop words"
 - Ex.: "select a make", "select makes" → "select make"
 - b) Definição da similaridade entre pares de atributos:
$$\mathbf{Sim}(a_i, a_j) = \mathbf{f}(\text{labelSim}(a_i, a_j), \text{valueSim}(a_i, a_j), \text{correlNeg}(a_i, a_j), \text{correlPos}(a_i, a_j))$$

Abordagem de (Nguyen et. al, 2008)

- $\text{LabelSim}(a_i, a_j) = \cos(a_i.\text{label}, a_j.\text{label})$
 - Converte as *strings* de rótulos em vetores e calcula a distância cosseno entre eles
 - Considera a frequência de cada termo do rótulo sozinho ou em conjunto com outros termos nas WFs no cálculo da similaridade
 - Fato muito comum em WFs
 - Exemplo:
 $\text{LabelSim}(a_i.\text{'departure'}, a_j.\text{'departure date'}) = 0.76$
 - *strings* parecidos
 - *date* aparece com alta frequência associada a *departure* no domínio aéreo

Abordagem de (Nguyen et. al, 2008)

- ValueSim (ai, aj) = $\cos ($
concat (sort(ai.valor-1, ..., ai.valor-n)),
concat (sort(aj.valor-1, ..., aj.valor-n)))
 - Converte as *strings* representativas da concatenação ordenada dos valores permitidos para os atributos em vetores e calcula a distância cosseno entre eles
 - Considera também a frequência dos valores na amostra no cálculo da similaridade
 - Exemplo:
ValueSim (ai.{'fiat ford toyota'},
aj.{'fiat ford volkswagen'}) = 0.67

Abordagem de (Nguyen et. al, 2008)

- $\text{correlNeg}(a_i, a_j) = \begin{cases} 0, & \text{se } a_i, a_j \text{ estão na mesma WF} \\ \text{freq}(a_i) \cdot \text{freq}(a_j) / \text{freq}(a_i) + \text{freq}(a_j), & \text{caso contrário.} \end{cases}$
 - Considera que 2 atributos que aparecem juntos em WFs jamais terão correlação negativa
 - A fórmula gera valores mais altos para termos sinônimos (não aparecem juntos em WFs)
- $\text{correlPos}(a_i, a_j) = \text{freq}(a_i + a_j) / \min(\text{freq}(a_i), \text{freq}(a_j))$
 - 2 atributos que aparecem juntos em WFs terão alta correlação positiva

Abordagem de (Nguyen et. al, 2008)

- Exemplos (situações extremas)

Alta correlação

| | | Model | |
|------|---|-------|------|
| | | 0 | 1 |
| Make | 0 | 50 | 50 |
| | 1 | 100 | 5000 |

$$\text{correlNeg}(\text{Make}, \text{Model}) = 0$$

$$\text{correlPos}(\text{Make}, \text{Model}) = 5000 / \min(50, 100) = 100$$

Sem correlação

| | | Brand | |
|------|---|-------|----|
| | | 0 | 1 |
| Make | 0 | 10 | 20 |
| | 1 | 5000 | 0 |

$$\text{correlNeg}(\text{Make}, \text{Brand}) = 5000 \cdot 20 / (5000 + 10) = 100000 / 5020 \approx 20$$

$$\text{correlPos}(\text{Make}, \text{Brand}) = 0 / \min(20, 5000) = 0$$

Abordagem de (Nguyen et. al, 2008)

- Exemplos (situações extremas)

Alta correlação

| | | Model | |
|------|---|-------|------|
| | | 0 | 1 |
| Make | 0 | 50 | 50 |
| | 1 | 100 | 5000 |

$\text{correlNeg}(\text{Make}, \text{Model}) = 0$

$\text{correlPos}(\text{Make}, \text{Model}) = 5000 / \min(50, 100) = 100$

PROBLEMAS:

- Requer uma amostra de WFs pré-computada com frequências de rótulos e valores válidos
- Não trata a dicotomia atributo-valor, pois não define uma estratégia para comparação de rótulos e valores
- Não trata correspondências 1-N entre atributos (testa apenas pares de atributos em WFs diferentes)

$\text{correlNeg}(\text{Make}, \text{Brand}) = 5000 \cdot 20 / (5000 + 10) = 10000 / 5020 = 20$

$\text{correlPos}(\text{Make}, \text{Brand}) = 0 / \min(20, 5000) = 0$

Roteiro

1. Introdução

2. **Principais Tópicos de Pesquisa**

i. Crawling

ii. Extração

iii. Matching

iv. **Consulta**





3. Algumas Iniciativas

4. Tendências

Referências

Deep Web - Consultas

- Crawling, extração e matching visam disponibilizar dados para consultas integradas
- Quadrantes

| | <i>Consultas Não-Estruturadas</i> | <i>Consultas Estruturadas</i> |
|-----------------|--|---|
| Metadados (WFs) |  |  |
| Dados |  |  |

Deep Web - C




- Crawling, extração e disponibilizar dados para c
- Quadrantes

- Principal foco de pesquisa e desenvolvimento
 - algumas iniciativas
- Máquinas de busca baseadas em *keywords*
 - *indexam domínios, rótulos e valores de atributos das WFs*
- Dados de WFs geralmente mantidos em BDs relacionais

| | <i>Consultas Não-Estruturadas</i> | <i>Consultas Estruturadas</i> |
|-----------------|--|---|
| Metadados (WFs) |  |  |
| Dados |  |  |

Deep Web - C

- Crawling, extração e disponibilizar dados para c
- Quadrantes

| | <i>Consultas Não-Estruturadas</i> | |
|-----------------|--|---|
| Metadados (WFs) |  | |
| Dados |  |  |

- Pouca iniciativa, devido à dificuldade de extração
- Informações sobre *strings* de dados extraídos de WFs são mantidas em BDs relacionais
- Abordagens (protótipos)
 - retorna as WFs onde o dado (*keyword* de entrada) se encontra
 - preenche WFs com os dados (*keywords*) de entrada, retornando os resultados ao usuário (Ex.: 'Fiat' → infere que é uma marca de carro e busca informações em WFs de veículos) - *Google*

- Problema ainda em aberto

- Inexistência de sistemas e linguagens de consulta para WFs e para dados dos BDs na Web

- Carência de BDs tradicionais (visíveis) com esquemas e dados bem definidos sobre a Deep Web (por domínio) que permitam consultas a seus dados e metadados

Exemplos:


```
SELECT * FROM WebForms  
WHERE LABEL = 'Make' (filtro por metadado)
```

```
SELECT Model FROM Veiculos  
WHERE Make = 'Ford' (filtro por dado)
```

* Para formular estas consultas, preciso saber que 'LABEL' é um metadado passível de consulta e que 'Make' é um atributo do esquema de um BD na Web (*Veículos*) cujos dados também posso consultar

Consultas

matching visam
consultas integradas

| | <i>Consultas Estruturadas</i> |
|------------|---|
| <i>das</i> | |
| |  |
| |  |

Roteiro

1. Introdução

2. Principais Tópicos de Pesquisa

i. Crawling

ii. Extração

iii. Matching

iv. Consulta

3. **Algumas Iniciativas**

4. Tendências

Referências

Deep Web – Algumas Iniciativas

- Catálogos / Diretórios de *sites* / metadados
 - <http://metaquerier.cs.uiuc.edu/repository/>
 - <http://www.completeplanet.com>
 - ...
- Busca Integrada em um ou mais Domínios
 - <http://www.expedia.com>
 - <http://www.travelocity.com> (hotéis, carros, aéreo)
 - <http://apartments.cazoodle.com/> (locações)
 - ...
- Máquinas de Busca
 - <http://www.deeppeep.org>
 - <http://turbo10.com>
 - ...

Roteiro

1. Introdução

2. Principais Tópicos de Pesquisa

i. Crawling

ii. Extração

iii. Matching

iv. Consulta

3. Algumas Iniciativas

4. **Tendências**

Referências

Deep Web – Tendências de Pesquisa

- Mecanismos **eficientes** de **extração** e **matching**
 - Lidar com a alta heterogeneidade de metadados nas WFs
 - Boa cobertura de dados recuperados de BDs na Web
 - Conhecimento da **semântica** da Deep Web
 - Preencher automaticamente WFs a partir de *keywords* (Ex.: 'Toyota corolla 2010' → preencher WFs de Veículos)
 - Contextualizar *keywords* como atributos ou valores em determinados domínios
 - Extrair esquemas/dados de BDs na Web e organizá-los em BDs relacionais
 - Viabiliza **consultas estruturadas** e **por similaridade**
(Edital Universal CNPq – Projeto: 481569/2010-3)
 - Identificar relacionamentos entre dados na Web
 - Viabiliza consultas inter-BDs na Web
- (Ex.: livros a venda (BD e-commerce) de atores famosos (BD cinema))

Roteiro

1. Introdução

2. Principais Tópicos de Pesquisa

i. Crawling

ii. Extração

iii. Matching

iv. Consulta

3. Algumas Iniciativas

4. Tendências

Referências

Algumas (Outras) Referências Web

- www.press.umich.edu/jep/07-01/bergman.html
(Deep Web – artigo introdutório)
- <http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php#query=deepweb&qp=H1.20.21:W1.4:F1.4:F2.4:F3.4>
(Deep Web – artigos acadêmicos)
- http://en.wikipedia.org/wiki/Deep_Web
- <http://www.inf.ufsc.br/~ronaldo/deepWeb>



Deep Web

Ronaldo S. Mello

GBD/INE/CTC/UFSC

Novembro, 2010

