Invited Lecture

Search Joins with the Web

Prof. Dr. Christian Bizer
A Search Join is a join operation which extends a local table with additional attributes based on the large corpus of structured data that is published on the Web.

1. Motivation and Definition
2. Profile of the available Web Data
3. Feasibility of Search Joins
4. Table Relevance
Deluge of Structured Data on the Web
In corpus of 14B raw tables, 154M are “good” relations (1.1%).

HTML-embedded Data on the Web

Several million websites semantically markup the content of their HTML pages.

**Markup Syntaxes**
- Microformats
- RDFa
- Microdata

**Data Consumers**
- Google
- Facebook
- Microsoft
- Yahoo!
- Yandex
Linked Data on the Web

~ 900 data sets (2014)
Several 100,000 datasets are available via data portals.
Table Search

Given some keywords describing the user’s information need, generate ranked list of relevant tables.

- Example: Google Table Search
  - http://research.google.com/tables

- Problem:
  The user is left alone with the integration.
  - using some tool
  - doing cut@paste

- Pimplikar, Sarawagi: Answering table queries on the web using column keywords. VLDB 2012.
## Goal 1: Extend Local Table with Single Attribute

Given a local table, a search attribute, and keywords describing the extension attribute, add the extension attribute to the table and fill it with data from the Web.

<table>
<thead>
<tr>
<th>No.</th>
<th>Region</th>
<th>Unemployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alsace</td>
<td>11 %</td>
</tr>
<tr>
<td>2</td>
<td>Lorraine</td>
<td>12 %</td>
</tr>
<tr>
<td>3</td>
<td>Guadeloupe</td>
<td>28 %</td>
</tr>
<tr>
<td>4</td>
<td>Centre</td>
<td>10 %</td>
</tr>
<tr>
<td>5</td>
<td>Martinique</td>
<td>25 %</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GDP per Capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>45.914 €</td>
</tr>
<tr>
<td>51.233 €</td>
</tr>
<tr>
<td>19.810 €</td>
</tr>
<tr>
<td>59.502 €</td>
</tr>
<tr>
<td>21.527 €</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
Goal 2: Extend Local Table with Many Attributes

Given a local table, a search attribute, add all attributes to the local table that can be filled beyond a density threshold.

No. Region Unemp. Rate
1 Alsace 11 %
2 Lorraine 12 %
3 Guadeloupe 28 %
4 Centre 10 %
5 Martinique 25 %

Region density >= 0.8

GDP per Capita | Population Growth | Overseas departments | ...
--- | --- | --- | ---
45.914 € | 0,16 % | No | ...
51.233 € | -0,05 % | No | ...
19.810 € | 1,34 % | Yes | ...
59.502 € | NULL | NULL | ...
NULL | 2,64 % | Yes | ...
... | ... | ... | ...
Prototype: RapidMiner Linked Open Data Extension

Features
1. Extend local table with additional attributes from a Linked Data source
2. Add more attributes by following RDF links into other Linked Data sources
3. Mine extended tables using all RapidMiner features

http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/
Resulting Correlations

- Overseas Department (positive)
- Population growth (positive)
- Fast food restaurants (positive)
- Police stations (positive)
- Hospital beds/inhabitants (negative)
- GDP (negative)
- Energy consumption (negative)

Linked Data Sources used: Eurostat and DBpedia
Definition: Search Join

A Search Join is a join operation which extends a local table with additional attributes based on the large corpus of structured data that is published on the Web.

■ Input:
  1. Corpus of heterogeneous Web tables
  2. Query table
  3. Search attribute definition
  4. Extension attribute(s) definition
    - Single attribute case: keyword describing extension attribute
    - Multiple attributes case: density threshold

■ Output:
  - Query table augmented with additional attribute(s)
**Search Joins in SQL**

```sql
SELECT city.*, web.population
FROM city SEARCH JOIN web ON city.name;
```

```sql
SELECT city.name, web.*(0.9)
FROM city SEARCH JOIN web ON city.name;
```
Elements of a Search Join

\[ R = c(m(q, s_{q,s,a}(T_{Web}))) \]

- **s()**: Search operator determines the set of the top-k relevant Web tables. (Web tables which are beneficial join partners)

- **m()**: MultiJoin operator performs a series of left-outer joins between the query table and all tables in the input set.

- **c()**: Consolidation operator merges corresponding attributes and fuses attribute values in order to return a concise result table containing high-quality data.
The Search Operator

The Search operator determines the set of relevant Web tables.

\[ T_r = s_{q,s,a}(T_{\text{Web}}) \]

- **Input**
  - \( T_{\text{Web}} = \) set of Web tables
  - \( q = \) query table
  - \( s = \) search attribute
  - \( a = \) attribute description

- **Output**
  - \( T_r = \) set of relevant Web Tables
The MultiJoin operator performs a series of left-outer joins between the query table and all tables in the input set.

**Input**

- $q = \text{query table}$
- $T_r = \text{set of relevant Web tables}$

**Output**

- $t_e = \text{extended query table}$

\[
t_e = m(q, T_r)
\]
Consolidation Operator

The consolidation operator merges corresponding attributes and fuses attribute values in order to return a concise result table containing high-quality data.

\[ t_r = c(t_e) \]

- **Input**
  \[ t_e = \text{extended query table} \]

- **Output**
  \[ t_r = \text{result table} \]

- Might employ various sources of attribute correspondences.

- Might employ various conflict resolution functions.

<table>
<thead>
<tr>
<th>No</th>
<th>Region</th>
<th>Unemploy</th>
<th>GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alsace</td>
<td>11 %</td>
<td>45.914 €</td>
</tr>
<tr>
<td>2</td>
<td>Lorraine</td>
<td>12 %</td>
<td>51.233 €</td>
</tr>
<tr>
<td>3</td>
<td>Guadeloupe</td>
<td>28 %</td>
<td>19.000 €</td>
</tr>
<tr>
<td>4</td>
<td>Centre</td>
<td>10 %</td>
<td>59.500 €</td>
</tr>
</tbody>
</table>
Data Model for Representing Web Data

- **Entity-Attributes-Tables**
  - One entity per row

- **Subject Attribute**
  - Name of the entity
  - String, no number or other data type
  - Relatively unique values

<table>
<thead>
<tr>
<th>Rank</th>
<th>Film</th>
<th>Studio</th>
<th>Director</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Star Wars –Episode 1</td>
<td>Lucasfilm</td>
<td>George Lucas</td>
<td>121 min</td>
</tr>
<tr>
<td>2.</td>
<td>Alien</td>
<td>Brandwine</td>
<td>Ridley Scott</td>
<td>117 min</td>
</tr>
<tr>
<td>3.</td>
<td>Black Moon</td>
<td>NEF</td>
<td>Louis Malle</td>
<td>100 min</td>
</tr>
</tbody>
</table>
Data Model Details

■ Table Meta-Information
  – Provenance: Source URL
  – Table Context: Text around the table

■ Attribute Headers
  – Attribute name
  – Unit of measurement
  – Date

■ Data types
  – String, Number, Date, Geo Coordinates
  – Lists
  – URI Reference
2. Profile of the available Web Data
In corpus of 14B raw tables, 154M are “good” relations (1.1%).

Classification Precision: 70-80%

- Crestan, Pantel: Web-Scale Table Census and Classification. WSDM 2011.
Subject Attribute Detection (Ventis)

- Simple heuristic approach (Accuracy: 83%)
  - scan columns from left to right
  - take first column that is not a number or a date
- SVM Classifier (Accuracy: 94%)
  - fraction of cells with unique content
  - variance in the number of tokens in each cell
  - column index from the left
  - ....

Header Detection (Pimplikar)

- one header row: 60%
- two or more header rows: 22%
- no header: 18%

- Pimplikar, Sarawagi: Answering table queries on the web using column keywords, VLDB 2012.
Common Crawl is a non-profit foundation dedicated to building and maintaining an open crawl of the web, thereby enabling a new wave of innovation, education and research.
Web Data Commons – Web Tables Corpus

- Large corpus of relational Web tables for public download
- extracted from Common Crawl 2012 (3.3 billion pages)
- 147 million relational tables
  - selected out of 11.2 B raw tables (1.3%)
  - download includes the HTML pages of the tables (1TB zipped)

Table Statistics

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes</td>
<td>2</td>
<td>2,368</td>
<td>3.49</td>
<td>3</td>
</tr>
<tr>
<td>Data Rows</td>
<td>1</td>
<td>70,068</td>
<td>12.41</td>
<td>6</td>
</tr>
</tbody>
</table>

- Heterogeneity: Very high.

- http://webdatacommons.org/webtables/
Attribute Statistics

<table>
<thead>
<tr>
<th>Attribute</th>
<th>#Tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>4,600,000</td>
</tr>
<tr>
<td>price</td>
<td>3,700,000</td>
</tr>
<tr>
<td>date</td>
<td>2,700,000</td>
</tr>
<tr>
<td>artist</td>
<td>2,100,000</td>
</tr>
<tr>
<td>location</td>
<td>1,200,000</td>
</tr>
<tr>
<td>year</td>
<td>1,000,000</td>
</tr>
<tr>
<td>manufacturer</td>
<td>375,000</td>
</tr>
<tr>
<td>country</td>
<td>340,000</td>
</tr>
<tr>
<td>isbn</td>
<td>99,000</td>
</tr>
<tr>
<td>area</td>
<td>95,000</td>
</tr>
<tr>
<td>population</td>
<td>86,000</td>
</tr>
</tbody>
</table>

28,000,000 different attribute labels

Subject Attribute Values

<table>
<thead>
<tr>
<th>Value</th>
<th>#Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>usa</td>
<td>135,000</td>
</tr>
<tr>
<td>germany</td>
<td>91,000</td>
</tr>
<tr>
<td>greece</td>
<td>42,000</td>
</tr>
<tr>
<td>new york</td>
<td>59,000</td>
</tr>
<tr>
<td>london</td>
<td>37,000</td>
</tr>
<tr>
<td>athens</td>
<td>11,000</td>
</tr>
<tr>
<td>david beckham</td>
<td>3,000</td>
</tr>
<tr>
<td>ronaldinho</td>
<td>1,200</td>
</tr>
<tr>
<td>oliver kahn</td>
<td>710</td>
</tr>
<tr>
<td>twist shout</td>
<td>2,000</td>
</tr>
<tr>
<td>yellow submarine</td>
<td>1,400</td>
</tr>
</tbody>
</table>

1.74 billion rows
253,000,000 different subject labels
More and more Websites semantically markup the content of their HTML pages.

- Microformats
- RDFa
- Microdata
Schema.org

- Ask site owners to embed data to enrich search results.
- 200+ Classes: Product, Review, LocalBusiness, Person, Place, Event, …
- Encoding: Microdata or RDFa
Usage of Schema.org Data @ Google

Data snippets within search results

Data snippets within info boxes
Websites containing Structured Data (2012)

Web Data Commons - Microformat, Microdata, RDFa Corpus
- 7 billion RDF triples from Common Crawl 2012
- Winter 2013 release upcoming

369 million of the 3 billion pages contain Microformat, Microdata or RDFa data (12.3%).

2.29 million websites (PLDs) out of 40.6 million provide Microformat, Microdata or RDFa data (5.65%)

Google, October 2013:
15% of all websites provide structured data.
## Top Classes Microdata (2012)

<table>
<thead>
<tr>
<th>Class</th>
<th>PLDs Total</th>
<th>PLDs in Alexa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 schema:BlogPosting</td>
<td>25,235</td>
<td>1,502</td>
</tr>
<tr>
<td>2 datavoc:Breadcrumb</td>
<td>21,729</td>
<td>5,244</td>
</tr>
<tr>
<td>3 schema:PostalAddress</td>
<td>19,592</td>
<td>1,404</td>
</tr>
<tr>
<td>4 schema:Product</td>
<td>16,612</td>
<td>3,038</td>
</tr>
<tr>
<td>5 schema:LocalBusiness</td>
<td>16,383</td>
<td>845</td>
</tr>
<tr>
<td>6 schema:Article</td>
<td>15,718</td>
<td>3,025</td>
</tr>
<tr>
<td>7 datavoc:Review-aggregate</td>
<td>8,517</td>
<td>2,376</td>
</tr>
<tr>
<td>8 schema:Offer</td>
<td>8,456</td>
<td>1,474</td>
</tr>
<tr>
<td>9 datavoc:Rating</td>
<td>7,711</td>
<td>1,726</td>
</tr>
<tr>
<td>10 schema:AggregateRating</td>
<td>7,029</td>
<td>1,791</td>
</tr>
<tr>
<td>11 schema:Organization</td>
<td>7,011</td>
<td>1,270</td>
</tr>
<tr>
<td>12 datavoc:Product</td>
<td>6,770</td>
<td>1,156</td>
</tr>
<tr>
<td>13 schema:WebPage</td>
<td>6,678</td>
<td>2,112</td>
</tr>
<tr>
<td>14 datavoc:Organization</td>
<td>5,853</td>
<td>654</td>
</tr>
<tr>
<td>15 datavoc:Address</td>
<td>5,559</td>
<td>654</td>
</tr>
<tr>
<td>16 schema:Person</td>
<td>5,237</td>
<td>890</td>
</tr>
<tr>
<td>17 schema:GeoCoordinates</td>
<td>4,677</td>
<td>312</td>
</tr>
<tr>
<td>18 schema:Place</td>
<td>4,131</td>
<td>488</td>
</tr>
<tr>
<td>19 schema:Event</td>
<td>4,102</td>
<td>659</td>
</tr>
<tr>
<td>20 datavoc:Person</td>
<td>2,877</td>
<td>523</td>
</tr>
<tr>
<td>21 datavoc:Review</td>
<td>2,816</td>
<td>783</td>
</tr>
</tbody>
</table>

- schema = Schema.org
- datavoc = Google’s Rich Snippet Vocabulary
Example: Microdata Local Business

```html
<div itemscope itemtype="http://schema.org/LocalBusiness">
  <h1><span itemprop="name">Beachwalk Beachwear & Giftware</span></h1>
  <span itemprop="description">A superb collection of fine gifts and clothing to accent your stay in Mexico Beach.</span>
  <div itemprop="address" itemscope itemtype="http://schema.org/PostalAddress">
    <span itemprop="streetAddress">3102 Highway 98</span>,
    <span itemprop="addressLocality">Mexico Beach</span>,
    <span itemprop="addressRegion">FL</span>
  </div>
  <span itemprop="telephone">850-648-4200</span>
</div>
```
### Microdata Product (2012)

<table>
<thead>
<tr>
<th>Property</th>
<th>PLDs Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
</tr>
<tr>
<td>1 dc:title</td>
<td>16,488</td>
</tr>
<tr>
<td>2 schema:Product/name</td>
<td>14,342</td>
</tr>
<tr>
<td>3 schema:Product/description</td>
<td>10,297</td>
</tr>
<tr>
<td>4 schema:Product/image</td>
<td>8,093</td>
</tr>
<tr>
<td>5 schema:Product/offers</td>
<td>7,545</td>
</tr>
<tr>
<td>6 schema:Offer/price</td>
<td>6,894</td>
</tr>
<tr>
<td>7 schema:AggregateRating</td>
<td>4,308</td>
</tr>
<tr>
<td>8 schema:AggregateRating/ratingValue</td>
<td>3,990</td>
</tr>
<tr>
<td>9 schema:PostalAddress/streetAddress</td>
<td>3,723</td>
</tr>
<tr>
<td>10 schema:PostalAddress/addressRegion</td>
<td>3,502</td>
</tr>
</tbody>
</table>

### Example Name:
- Apple MacBook Air 11-in, Intel Core i5 1.60GHz, 64 GB, Lion 10.7

### Example Description:
- Configured with Intel Core 2 Duo processor, faster Flash Storage with 64 GB Solid State Drive and USB 3.0 ...
Identity Resolution for Electronic Products

- We trained parser for product descriptions on data from Amazon.
- We analyzed 1.9 million product offers from 9200 e-shops

<table>
<thead>
<tr>
<th>Product</th>
<th>Offers</th>
<th>Reviews</th>
<th>Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple iPod Nano 8GB</td>
<td>829</td>
<td>84</td>
<td>0</td>
</tr>
<tr>
<td>Apple iPhone 4 16GB</td>
<td>624</td>
<td>35</td>
<td>52</td>
</tr>
<tr>
<td>Sony Ericsson Xperia Mini</td>
<td>450</td>
<td>31</td>
<td>12</td>
</tr>
<tr>
<td>Apple iPad 2 16GB</td>
<td>423</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>Motorola XOOM 32GB</td>
<td>270</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Samsung Galaxy SII</td>
<td>142</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

- Petar Petrovski, et al.: Integrating Product Data from Websites offering Microdata Markup. DEOS workshop @ WWW 2014.
Table Generation
- Represent each class per website as separate table.
- subject column: naming convention itemprop="name"

Resulting tables
- several million tables
- mostly 2-10 attributes wide
- up to 100,000s of rows

Heterogeneity
- low as data providers use vocabularies recommended by Google, Microsoft, Yahoo, and Facebook

Linked Data

Extends the Web with a single global data graph

1. by using RDF to publish structured data on the Web
2. by setting links between data items within different data sources.
Data Graph including Integration Hints

- **Data Providers**
  - set instance-level and schema-level RDF links
  - reuse terms from common vocabularies
Effort Distribution between Publisher and Consumer

Publisher reuses vocabularies

Consumer calculates links and correspondences

Publisher or third party publishes links and correspondences

---

Application Layer

Data Access, Integration and Storage Layer

Web of Linked Data

Publication Layer

- Web Data Access Module
- Vocabulary Mapping Module
- Identity Resolution Module
- Quality Evaluation Module
- Integrated Web Data

SPARQL

HTTP

HTTP

LD Wrapper

LD Wrapper

RDFa

RDF/XML

Database A

Database B

Legacy App C
Linked Data on the Web

- 900 data sets (2014)
- 500 million RDF links (2011)
Wikipedia Data available as Linked Data

Diagram showing Wikipedia connected to DBpedia, Yago, Freebase, Universität Mannheim, Max Planck Institut für Informatik, and Universität Leipzig.
Most Linked Data sources have a rather regular structure
- as they are generated from relational databases

Table generation
- generate one table per class and data source
- use RDF property labels as attribute names
- subject attribute: naming convention rdfs:label, foaf:name

Representation of RDF Links
- Add ID attribute containing original URI of each entity
- For each link predicate type add two attributes
  1. URI reference to target entity
  2. rdfs:label of target entity
Profile of the Resulting Tables

- Billion Triples Challenge 2012 Dataset
  - 53,000 tables
    - Attributes: Min 3, Max 1,479, Avg 9
    - Data Rows: Min 1, Max 372,000, Avg 180

- DBpedia as Tables
  - 365 tables
    - Attributes: Min 7, Max 730, Avg 498
    - Data Rows: Min 1, Max 577,000, Avg 12,000

Heterogeneity
- Low in some domains: People, Publications
- High in other domains: Life Science, eGovernment

- http://km.aifb.kit.edu/projects/btc-2012/
- http://wiki.dbpedia.org/DBpediaAsTables
Data Portals

- datacatalogs.org lists 377 data portals world-wide
  - open government data portals
  - international organizations and NGOs
  - scientific data portals

Challenges
- syntax heterogeneity: Excel, CSV, XML, HTML, PDF
- data values are often time dependent
- table context understanding and header-unfolding necessary

Profile of 7600 Tables from PublicData.eu

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes</td>
<td>2</td>
<td>488</td>
<td>8.88</td>
<td>9</td>
</tr>
<tr>
<td>Data Rows</td>
<td>1</td>
<td>5,600,000</td>
<td>3,160</td>
<td>66</td>
</tr>
</tbody>
</table>

There is lots of data available that
- we can fit into our data model.
- use for search join experiments.

A wide range of topics is covered.

The size of the tables varies widely.

Additional types of data sources not considered
- Web 2.0 APIs, Deep Web via HTML forms
- HTML Lists, Excel files somewhere on the Web

- Furche: The Ontological Key: Automatically Understanding and Integrating Forms to access the Deep Web. VLDB-J 2013.
2. Feasibility of Search Joins
### Search Join Systems

<table>
<thead>
<tr>
<th>Developer</th>
<th>Octopus</th>
<th>InfoGather</th>
<th>WikiTables</th>
<th>MSJ Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Research University Washington</td>
<td>Microsoft Research Purdue University</td>
<td>Northwestern University</td>
<td>University of Mannheim</td>
<td></td>
</tr>
<tr>
<td>Extend Operation</td>
<td>Single Attribute</td>
<td>Single Attribute</td>
<td>Multiple Attributes</td>
<td>Single Attribute Multiple Attributes</td>
</tr>
<tr>
<td>Corpus</td>
<td>Google Web crawl via Search API</td>
<td>HTML tables from Bing Web crawl</td>
<td>Tables from Wikipedia</td>
<td>WDC Web Tables Linked Data</td>
</tr>
<tr>
<td>Use Case</td>
<td>Data Gathering</td>
<td>Data Gathering</td>
<td>Data Gathering Data Mining</td>
<td>Data Gathering Data Mining</td>
</tr>
</tbody>
</table>

Infogather

- Prototype developed by Microsoft Research
- Operation: Extend with Single Attribute
- Corpus: 573 million Web tables from Bing Crawl (2011)
- Split HTML Tables into Binary-Entity-Attribute Tables

Subject Attribute | Value Attribute

<table>
<thead>
<tr>
<th>Region</th>
<th>Unemployment</th>
<th>Region</th>
<th>GDP per Capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alsace</td>
<td>11 %</td>
<td>Alsace</td>
<td>45.914 €</td>
</tr>
<tr>
<td>Lorraine</td>
<td>12 %</td>
<td>Lorraine</td>
<td>51.233 €</td>
</tr>
<tr>
<td>Guadeloupe</td>
<td>28 %</td>
<td>Guadeloupe</td>
<td>19.810 €</td>
</tr>
</tbody>
</table>

Matching Graph

- Pre-compute matching graph between BEA tables

- Features used for matching:
  1. Attribute label similarity
  2. Attribute values similarity
  3. Key values overlap
  4. Textual context around tables similarity
  5. Table to context similarity
  6. Table to table as bag of words similarity
  7. URL similarity
  8. Column width similarity

- Accuracy of the resulting correspondences:
  - Cameras and movies: 0.95
  - Governors and members of parliament: 0.5
Query Processing

- **Input:** Query table, extension attribute

- **Web tables considered relevant for query**
  - share subject attribute value with query table
  - and directly match extension attribute
  - or are connected to directly matching tables via matching graph

- **Matching score**
  - directly matching tables: entity overlap / min( |t_q|, |t_w| )
  - indirectly matching tables: propagate score along edges of matching graph

- **Predict values**
  - cluster by value
  - sum matching scores per cluster
  - choose centroid of cluster with highest score or top-k centroids
Experimental Setting

Queries

<table>
<thead>
<tr>
<th>Query</th>
<th>Subject Attribute</th>
<th>Extension Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameras</td>
<td>Camera model</td>
<td>Brand</td>
</tr>
<tr>
<td>Movies</td>
<td>Movie name</td>
<td>Director</td>
</tr>
<tr>
<td>Baseball</td>
<td>Team name</td>
<td>Player</td>
</tr>
<tr>
<td>Albums</td>
<td>Musical band</td>
<td>Album</td>
</tr>
<tr>
<td>UK-pm</td>
<td>UK political party</td>
<td>Member of parliament</td>
</tr>
<tr>
<td>US-gov</td>
<td>US state</td>
<td>Governor</td>
</tr>
</tbody>
</table>

Ground Truth

- Camera, movies: Bing shopping database

Size of query table

- 12 to 6000 rows
## Evaluation Results

<table>
<thead>
<tr>
<th>Query</th>
<th>Subject attribute</th>
<th>Extension Attribute</th>
<th>Precision</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameras</td>
<td>Camera model</td>
<td>Brand</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>Movies</td>
<td>Movie name</td>
<td>Director</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>Baseball</td>
<td>Team name</td>
<td>Player</td>
<td>0.72</td>
<td>1.00</td>
</tr>
<tr>
<td>Albums</td>
<td>Musical band</td>
<td>Album</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>UK-pm</td>
<td>UK political party</td>
<td>Member of parliament</td>
<td>0.60</td>
<td>0.91</td>
</tr>
<tr>
<td>US-gov</td>
<td>US state</td>
<td>Governor</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>AVG</strong></td>
<td></td>
<td></td>
<td><strong>0.79</strong></td>
<td><strong>0.97</strong></td>
</tr>
</tbody>
</table>

Response times: around 100 milliseconds
The Search operator determines the set of relevant Web tables.

\[ T_r = S_{q,s,e}(T_{Web}) \]
Information Provision on the Web

Everything on the Web is a claim by somebody.

1. Claims use different surface forms.
   - entity name
   - attribute labels
   - data value

2. Claims refer to a specific point in time.

3. The trustworthiness of claims varies widely.
1. Entity Coverage
   - Web table should cover many entities in the query table.

2. Attribute Relevance
   - Web table should contain relevant attributes.

3. Timeliness
   - The data should refer to the desired point in time.

4. Trustworthiness
   - The data should be trustworthy.
Identity Resolution Approaches

1. Exact matching on subject attribute
2. Approximate matching on subject attribute
3. Matching using external knowledge about surface forms
4. Matching using multiple attributes from both tables
5. relying on owl:sameAs links

$$c = \frac{\text{entity overlap}}{|t_q|}$$
Exact Matching

Exact matching of normalized values against WDC table corpus.

20 country names

15 names of mayor cities
Collective Disambiguation

Query Table

<table>
<thead>
<tr>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madison</td>
</tr>
<tr>
<td>Berlin</td>
</tr>
<tr>
<td>Chatham</td>
</tr>
<tr>
<td>Perth</td>
</tr>
</tbody>
</table>

Web Table 1

<table>
<thead>
<tr>
<th>City</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Munich</td>
<td>Germany</td>
</tr>
<tr>
<td>Berlin</td>
<td>Germany</td>
</tr>
<tr>
<td>Mannheim</td>
<td>Germany</td>
</tr>
<tr>
<td>Frankfurt</td>
<td>Germany</td>
</tr>
<tr>
<td>Karlsruhe</td>
<td>Germany</td>
</tr>
</tbody>
</table>

\[ s = \frac{1}{4} \]

Web Table 2

<table>
<thead>
<tr>
<th>City</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madison</td>
<td>USA</td>
</tr>
<tr>
<td>Berlin</td>
<td>USA</td>
</tr>
<tr>
<td>Chatham</td>
<td>USA</td>
</tr>
<tr>
<td>Fort Reed</td>
<td>USA</td>
</tr>
<tr>
<td>Sunville</td>
<td>USA</td>
</tr>
</tbody>
</table>

\[ s = \frac{3}{4} \]
Expansion with additional Surface Forms

Examples of surface forms
- Berlin, 柏林, Berlijn, Berlín, Berlino, Берлин, Berlim, ベルリン
- FC Bayern München, Bayern Munich, FC Bayern, Bayern München

Community-generated sources of surface forms
- Wikipedia
  - Redirects, Cross-Language Links
  - Easy accessible via DBpedia
- owl:sameAs Links
  - rdfs:labels of interlinked resources
  - 500 million links (2011)
Identity Resolution via owl:sameAs Links

Query table: 20 country names

1. Names are matched to DBpedia
2. owl:sameAs links are followed from DBpedia
The Web table should contain relevant attributes.

Attribute relevance depends on query type

1. Extend with single attribute
   1. tables that have attribute directly matching the keywords
   2. tables that have corresponding attributes

2. Extend with all attributes above density threshold
   1. take all attributes
   2. prefer attributes that are related to attributes in query table
Extend with Single Attribute

Simplest approach: Exact matching of normalized values.

Query table: 20 country names
Search attribute: „population“

Query table: 15 names of mayor cities
Search attribute: „country“

Median of values differs on average 4% from Wikipedia value.
Include Corresponding Attributes

- Approaches:
  1. attributes having a synonymous name
  2. attributes that correspond according to matching
     1. tables with each other
     2. tables against a mediated schema (knowledge base)

- Experimental Results: Yakout et al.
  - Attribute Synonyms
    - Extension attribute value precision: 70 %
  - Attributes corresponding via Schema Matching
    - Extension attribute value precision: 79 %

Mediated Schemata

- Cross-domain knowledge bases
  - **DBpedia**
    - Classes: 259; attributes 1,373; entities: 4 million
  - **Freebase**
    - Classes 1,450; attributes 3,500; entities: 15 million

- Advantage
  - matching against large KBs is likely easier than matching small tables
  - Zhang: Web tables against DBpedia matching accuracy: 85%

- Disadvantage
  - Restricted to attributes contained in the knowledge base

Probase and Biperpedia

- Build comprehensive KBs using
  - KBs like DBpedia and Freebase as seeds
  - information extraction from Web text
    - taxonomy and instances: Hearst patterns “Y such as X”
    - Attributes: Patterns “What is the A of I?”
  - search engine query logs “New York City inhabitants”

- Probase (Microsoft)
  - 2.7 million classes, a set of attributes for each class

- Biperpedia (Google)
  - 10,000 classes and 67,000 attributes.
  - table annotation accuracy: 51%

Approaches

1. take all attributes

2. prefer attributes that are related to attributes in query table
Take all Attributes

- Query table: 20 country names
- WDC Web tables (Top 200 tables)
  - 920 additional attributes
  - 346 after attribute consolidation
- Linked Data (BTC dataset)
  - 1131 additional attributes
  - 403 after attribute consolidation

Problem: Number of attributes might overwhelm the user (even if he only looks at correlating attributes)
What should be considered “consistent”?
Answer: Combinations of local and Web attributes that often occur together on the Web.

Approach
- Generate frequent item set database from Web table schema corpus.
- Retrieve frequency of all combinations of one local and Web attribute.
- Aggregate frequencies for Web attributes.

The data in the Web tables should refers to the desired point in time.

- Requires meta-information about the intended point in time.

- **HTML Tables**
  - Time reference in table
    - Problem: only present in a few tables
  - Time reference on page
    - Problem: difficult to hard to understand

- **Linked Data**
  - W3C Data Cube Vocabulary

### Table: Unemployment Data (2013)

<table>
<thead>
<tr>
<th>Region</th>
<th>Unemployment (2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alsace</td>
<td>11 %</td>
</tr>
<tr>
<td>Lorraine</td>
<td>12 %</td>
</tr>
<tr>
<td>Guadeloupe</td>
<td>28 %</td>
</tr>
</tbody>
</table>

The table below provides unemployment data for 2013.

...
Label Propagation

- Approach: Infer time reference by matching to other tables that contain this information.

- Evaluation Results:
  - Accuracy: 89%
  - Recall: 50%

Dimension: Trustworthiness

The data in the Web table should be trustworthy.

- Not considered by Search Join systems yet.

- Approach exploiting only the data itself
  - Consider value consistency between query table and Web table if the tables contain overlapping attributes

- Approaches exploiting external knowledge
  - Put preference on specific sources
    - prefer data from .gov websites
  - Exploit hyperlink structure of the Web
### Web Data Commons - Hyperlink Graph

- Covers 3.5 billion web pages and 128 billion hyperlinks
- Extracted from Common Crawl 2012

#### The Common Crawl WWW Ranking

Here you can browse a ranking of more than 100 million sites of the World Wide Web. Every single step leading to this ranking is open and accessible. Enjoy!

![Common Crawl Ranking](image)

<table>
<thead>
<tr>
<th>Harmonic centrality</th>
<th>Indegree centrality</th>
<th>Katz's index</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. youtube.com</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2. en.wikipedia.org</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>3. twitter.com</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>4. google.com</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>5. wordpress.org</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6. flickr.com</td>
<td>8</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>7. facebook.com</td>
<td>19</td>
<td>18</td>
<td>17</td>
</tr>
</tbody>
</table>

- [http://webdatacommons.org/hyperlinkgraph/](http://webdatacommons.org/hyperlinkgraph/)
Wrap-up: Dimensions of Table Relevance

1. Entity Coverage
   - Web table should cover many entities in the query table.

2. Attribute Relevance
   - Web table should contain relevant attributes.

3. Timeliness
   - The data should refer to the desired point in time.

4. Trustworthiness
   - The data should be trustworthy.
Conclusion

Search Joins bring together Web Search and DB Joins via the concept of table relevance.

- Simple queries are feasible with “acceptable”(?) precision.
- The Web is one application domain for search joins, corporate intranets are the other.