Integrating Product Specifications from the Web
Prof. Dr. Christian Bizer
Professor for Information Systems V
Research Interests:
  • Web Data Integration
  • Data and Web Mining
  • Adoption of Data Web Technologies
  • Knowledge Base Construction
Room: B6 - B1.15
eMail: chris@informatik.uni-mannheim.de
Consultation: Wednesday, 13:30-14:30
Hallo

− Anna Primpeli
− Graduate Research Associate
− Research Interests:
  • Data Extraction
  • Web Data Integration
  • Active Learning
  • Structured Data on the Web
− Room: B6, 26, C 1.04
− eMail: anna@informatik.uni-mannheim.de
Agenda of Today’s Kickoff Meeting

1. You and Your Experience
2. Motivation and Project Goals
3. The WDC Training Dataset for Large-Scale Product Matching
4. Organization
5. Specific Subtasks
6. Schedule
7. Formal Requirements
Motivation of the Team Project

The Web is a rich source of product information
- same product is described by 100s of websites
  - merchants, producer, consumers
- different websites describe different aspects of a product
  - technical spec vs consumer experience
- there are plenty of offers for a product online
  - we can collect price information on global scale
- many websites point us at similar products

Using information about products from the Web, we can
- build comprehensive product catalogues and search engines
- construct global price comparison engines
- understand consumer and market behavior
Groups product offers from different sources that describe the same product entity.

Aggregates the offers’ features and maps them to a category-specific schema.

Presents a structured product catalog to the user.
Project Goal

... drive a well structured, comprehensive product catalogue from the Web

... similar to the Google Shopping catalogue

... by:

1. Find clusters of offers from e-Shops that refer to the same product

2. Extract and normalize product features (key value pairs) from these clusters

3. Assemble a clean product catalogue for two product categories
   1. Single integrated schema per category
   2. Clean set of values per attribute
How to find many offers of the same product?

Not an easy task!

1. Which sources to consider?
2. Which data to extract?
3. How to recognize identical products?
4. How to categorize products?

OR...

Use the WDC Training Dataset for Large-Scale Product Matching
http://webdatacommons.org/largescaleproductcorpus/index.html
How to determine product features and extract feature values?

- Specification Tables and Lists
  - Detailed features as key/value pairs

- Product Titles
  - Product name plus selected features

- Product Descriptions
  - long free texts

Before we can use these features:
- values need to be cleansed and normalized
- we might want to apply information extraction in order to increase the structuredness and density of the data
The WDC Training Dataset for Large-Scale Product Matching

1. Semantic Annotations in HTML Pages
2. Web Data Commons Project
3. Web Data Commons – Training Dataset for Large-Scale Product Matching
Semantic Annotation of HTML Pages: Schema.org

- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Place, Local Business, Product, Review, Person
- Encoding: Microdata, RDFa, JSON-LD
Example: Microdata Annotations in HTML

```html
<div itemtype="http://schema.org/Product">
  <span itemprop="name">Sony GTK-XB5L Audiosystem</span>
  <span itemprop="gtin13">04048945021687</span>
  <span itemprop="description">high-power home audio system with Bluetooth technology</span>
</div>

<div itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">
  <span itemprop="ratingValue">4</span> stars-based on
  <span itemprop="reviewCount">250</span> reviews.
</div>
```
schema.org Annotations: Most Popular Classes

Development of Selected Classes by #PLDs

http://webdatacommons.org/structureddata/
### Top 15 Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>PLDs</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>schema:Product/name</td>
<td>535,625</td>
<td>92%</td>
</tr>
<tr>
<td>schema:Offer/price</td>
<td>462,444</td>
<td>80%</td>
</tr>
<tr>
<td>schema:Product/offers</td>
<td>462,233</td>
<td>79%</td>
</tr>
<tr>
<td>schema:Offer/priceCurrency</td>
<td>430,556</td>
<td>74%</td>
</tr>
<tr>
<td>schema:Product/image</td>
<td>419,391</td>
<td>72%</td>
</tr>
<tr>
<td>schema:Product/description</td>
<td>377,639</td>
<td>65%</td>
</tr>
<tr>
<td>schema:Offer/availability</td>
<td>337,876</td>
<td>58%</td>
</tr>
<tr>
<td>schema:Product/url</td>
<td>263,720</td>
<td>45%</td>
</tr>
<tr>
<td>schema:AggregateRating/ratingValue</td>
<td>184,004</td>
<td>32%</td>
</tr>
<tr>
<td>schema:Product/sku</td>
<td>126,696</td>
<td>22%</td>
</tr>
<tr>
<td>schema:AggregateRating/reviewCount</td>
<td>112,408</td>
<td>19%</td>
</tr>
<tr>
<td>schema:Product/aggregateRating</td>
<td>101,434</td>
<td>17%</td>
</tr>
<tr>
<td>schema:Product/brand</td>
<td>73,934</td>
<td>13%</td>
</tr>
<tr>
<td>schema:Product/productID</td>
<td>35,211</td>
<td>6%</td>
</tr>
<tr>
<td>schema:Product/manufacturer</td>
<td>21,967</td>
<td>4%</td>
</tr>
</tbody>
</table>

**Samsung Galaxy S4**

Verizon AT&T T-Mobile GSM

Unlocked Smartphone SRF

Das Samsung Galaxy S4 ist der unterhaltsame und hilfreiche Begleiter für Ihr mobiles Leben. Es verbindet Sie mit Ihren Liebsten. Es lässt Sie gemeinsam unvergessliche Momente erleben und festhalten. Es vereinfacht Ihren Alltag.

**UPC** 610214632623

![Samsung Galaxy S4](image)
The WDC Training Dataset for Large-Scale Product Matching

- Training corpus grouping schema.org product/ offer annotations by identifier value.
  - all WDC 2017 product data is included that
  - provides some sort of product ID (gtin, mpn, sku, identifier)

- Initial cleaning steps are performed

- Clustering of product descriptions from different PLDs (web sites) that share identifier values.

Details and Download:
- [http://webdatacommons.org/largescaleproductcorpus/index.html](http://webdatacommons.org/largescaleproductcorpus/index.html)
Distribution of Offers per Category in the English Training Set

source: http://webdatacommons.org/categorization/index.html
Number of Key Value Pairs per Specification Table
Organization

Duration: 6 months (27.09.2019 – 29.03.2020)
Participants: 8 people
Type of work: Team and subgroup based
Milestones: 4 project phases
ECTS Points: 12

Evaluation
- Intermediate presentations
- Final report
- Individual contribution to the deliverables
1. Which categories to consider? → Data Selection and Profiling
2. How to integrate the schemata of the different sources? → Schema matching
3. How to increase the amount of extracted feature-values? → Advanced Feature Extraction
4. How to fuse multiple values of the same product and same feature? → Data Fusion
5. How to evaluate the quality of the product catalog? → Catalog Profiling and Evaluation
Phase 1: Data Selection, Data Profiling

Participants: all team members

Duration: 27.09. – 25.10.

Deliverables: 20 min. presentation, code & data

Input: WDC English Training Set, Categorization of Offers V2, Specification tables

provided at http://webdatacommons.org/largescaleproductcorpus/
and http://webdatacommons.org/categorization/

Tasks

1. Consider the offers with specification tables and
   1a. Profile amount of specification tables per category (cluster and offer level)
   1b. Profile amount of feature-value pairs per category (cluster and offer level)

3. Based on the profiling results decide on two product categories (or subcategories)

4. Profile the features and the feature values of the offers of the selected categories
Phase 1: How to get started?

- Get the updated categorization of the corpus
  
  http://webdatacommons.org/categorization/index.html
  
  File: categories_offers_en_clusters.csv.gzip

- Get the offers of the english corpus
  
  http://webdatacommons.org/largescaleproductcorpus/index.html
  
  File: offers_english.json.gz

- Get the specification tables per offer
  
  http://webdatacommons.org/largescaleproductcorpus/index.html
  
  File: specTables.json.gz
Phase 1: How to get started?

Example cluster categorization entry
source: categories_offers_en_clusters.csv.gz

How many offers per category?

Example offer entry
source: offers_english.json.gz

How many offers have specification tables?

Example specification table entry
source: specTables.json.gz

How many offers have how many key value pairs?

Aggregate on the cluster level to answer:
• How many clusters have specification tables?
• How many clusters have how many key value pairs?

Group clusters by category to answer:
• How many specification tables are there per category?
• How many key value pairs are there per category?
Phase 2: Schema Matching

Participants: two subgroups (one subgroup per selected category)

Duration: 25.10 – 29.11.

Deliverables: 30 min. Presentation, code & data

Input: Feature-value pairs from the specification tables of product offers for two categories

Tasks:

1. Closed schema matching (N→1)
   1a. Retrieve a dictionary from top attributes using the profiling results of phase 1
   1b. Perform and evaluate label / instance-based schema matching

2. Open schema matching (N→N)
   2a. Consider the clustered offers, perform and evaluate duplicate-based schema matching

3. Compile dictionary of possible values for each attribute
Phase 3: Data Normalization and Advanced Feature Extraction

**Participants:** two subgroups (one subgroup per selected category)

**Duration:** 29.11 – 24.01

**Deliverables:** 30 min. presentation, code & data

**Input:** Dictionary from phase 2, all offers from the WDC English Training set

**Tasks**

Using the dictionary from phase 2:

1. Design/learn data normalization rules considering their datatypes  
   - e.g.: convert all dates to years, normalize values of categorical attributes

2. Increase the completeness/accuracy of the catalog by extracting feature-value pairs from offers containing no specification tables using their titles and descriptions.
Phase 4: Data Fusion and Catalog Curation

Participants: two subgroups (one subgroup per selected category)

Duration: 24.01 – 06.03.

Deliverables: 30 min. presentation, code & data

Tasks

1. Compile integrated catalog for each category
   1a. Map the feature-values of every product to the catalog
   1b. Fuse values (voting-based, source-based, …)
   1c. Evaluate fusion

2. Profile attribute density in the integrated catalog
<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday, 27.09.2019</td>
<td>Kickoff meeting (today)</td>
</tr>
<tr>
<td></td>
<td>Phase 1 (all members): Data Selection, Data Profiling</td>
</tr>
<tr>
<td>Monday, 14.10.2019, 17:30</td>
<td>Meet Anna and report plan/current results</td>
</tr>
<tr>
<td>Friday, 25.10.2019, 9:00</td>
<td>1\textsuperscript{st} Deliverable: 20 minutes presentation, code &amp; data - Subgroup formation</td>
</tr>
<tr>
<td></td>
<td>Phase 2 (in 2 subgroups): Schema Matching</td>
</tr>
<tr>
<td>Friday, 08.11.2019, 17:30</td>
<td>Meet Anna and report plan/current results</td>
</tr>
<tr>
<td>Friday, 29.11.2019, 09:00</td>
<td>2\textsuperscript{nd} Deliverable: 30 minutes presentation, code &amp; data</td>
</tr>
<tr>
<td></td>
<td>Phase 3 (in 2 subgroups): Advanced Feature Extraction and Data Normalization</td>
</tr>
<tr>
<td>Friday, 13.12.2019, 17:30</td>
<td>Meet Anna and report plan/current results</td>
</tr>
<tr>
<td>Friday, 31.01.2020, 9:00</td>
<td>3\textsuperscript{rd} Deliverable: 30 minutes presentation, code &amp; data</td>
</tr>
<tr>
<td></td>
<td>Phase 4 (in 2 subgroups): Data Fusion and Catalog Curation</td>
</tr>
<tr>
<td>Friday, 07.02.2020, 17:30</td>
<td>Meet Anna and report plan/current results</td>
</tr>
<tr>
<td>Friday, 06.03.2020, 9:00</td>
<td>4\textsuperscript{th} Deliverable: 30 minutes presentation, code &amp; data</td>
</tr>
<tr>
<td>Sunday, 29.03.2020</td>
<td>Final Report Submission</td>
</tr>
</tbody>
</table>
Deliverables

1. On the deliverable dates provide us via e-mail with:
   • **Presentation slides**
   • **Task to member report**: excel sheet stating which team member conducted which subtask
   • **Code/ Data**: link or zipped folder with your code and data

2. **Final Report**
   • **15 pages** including appendices, not including the bibliography
   • every additional page reduces your grade by 0.3

All e-mails should be sent to Chris & Anna!
Formal Requirements & Consultation

Final grade
- 20% for every phase
- 20% for final report
- Late submission: -0.3 per day

Consultation
- Send one e-mail per team or subgroup stating your questions to Anna
Related Work (1/2)


Related Work (2/2)


Questions?