UNIVERSITY OF MANNHEIM

Team Project FSS 2018

Mining Product Data from the Web

Phase 2

Progress and Focus of these Phases

✓ 1. Which e-shops to consider? → **Data Selection and Crawling**

2. Which data to extract? → **Feature Extraction**

3. How to recognize identical products? → **Identity Resolution**

4. How to group similar products? → **Categorization / Cluster Analysis**

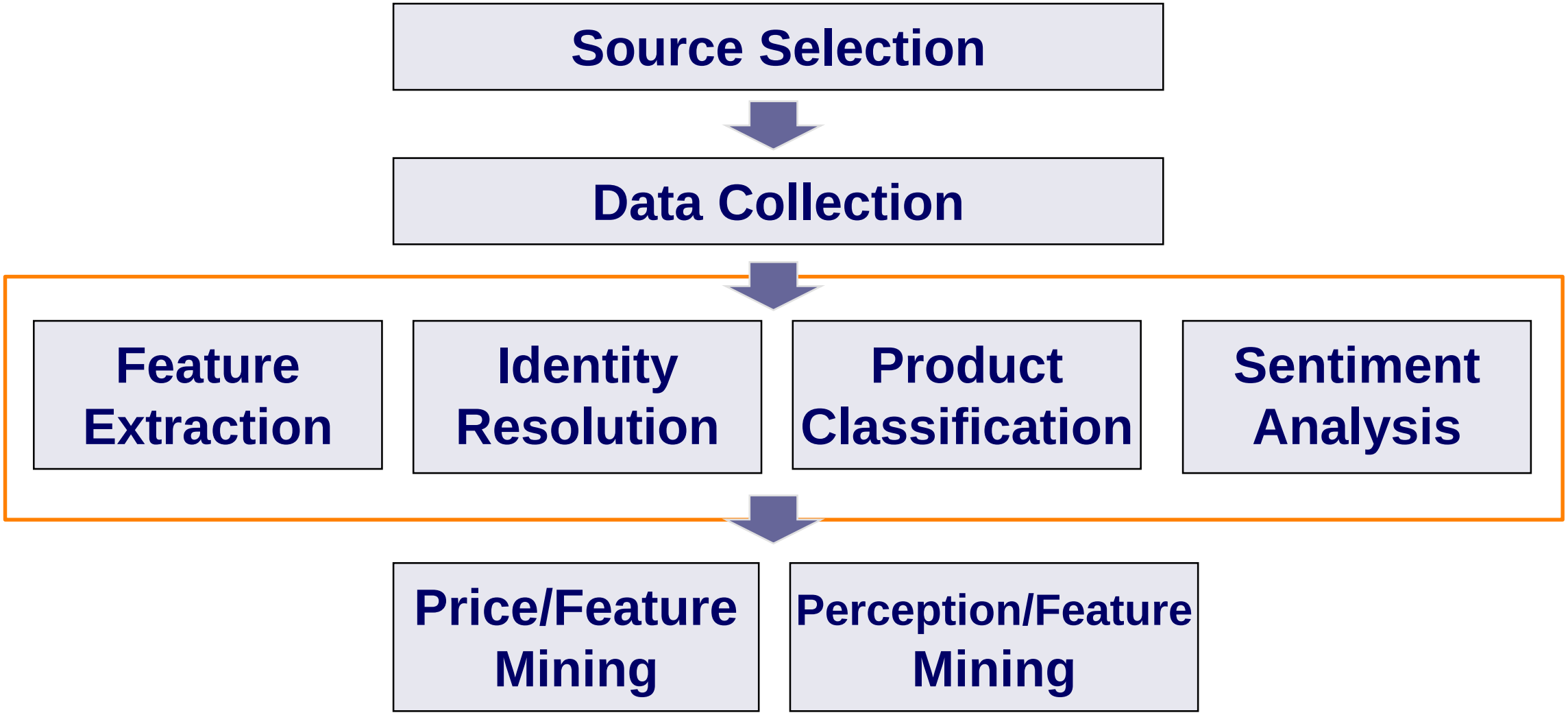
5. How to understand user perception? → **Sentiment Analysis**

Phase 3 will be a refinement of phase 2

6. How to combine extracted information? → **Data Fusion**

7. What patterns can be found in the data? → **Data Mining**

Main Steps of the Project



Detailed Schedule for Phase 2

Date	Session
Friday, 13.04.2018, 9:15am	Introduction to Phase 2, subgroup formation
Friday, 20.04.2018, 8:30am	Meet Anna and discuss plans
Friday, 27.04.2018, 9:15am	Meet Chris and Anna, report profiling results and specific goals for phase 2
Friday, 18.05.2018, 9:15am	**2nd Deliverable:** 10 minutes presentation from each subgroup, code & data

Results from Phase 1

What should you have from Phase 1?

- ✓ Crawled corpus of Bag and Camera products
  - Min. 20 e-shops located in 2 countries
  - Estimated high product overlap
  - Different subcategories
  - Pages rich in annotations and specification tables
  - Pages describing different products with similar attributes ex: Polaroid vs Polaroid Kit
- ✓ Basic Feature Extraction and Profiling
  - Identify and profile product specifications, prices, category information, product IDs, and reviews
- ✓ Product Catalogs
  - Min 50 bag products
  - Min 50 camera products

Phase 2 – Subgroup 1: Feature Extraction

Goal: Extract clean feature - value pairs from the product pages and perform schema maching

How?

1. Identify where features are located: tables, lists, free text following certain patterns, schema.org annotations
2. Extract product features and map them to the catalog
   - Generic Approach: Consider table and list structure, schema.org annotations and DOM structure
     1. Create a gold standard for schema matching
     2. Perform schema matching
        - Label based
        - Instance based
   - Catalog-oriented Approach: Apply regex expressions exploiting the knowledge in the catalog
3. Compare the two approaches

Phase 2 – Subgroup 2: Identity Resolution

Goal: Match entities between your product corpus and the product catalog

How?

1. Create your gold standard
   - Annotate manually min. 100 product pairs (product page – catalog entry)
   - Make sure you include good negative examples!
2. Consider Bag of Words models from different parts of the product page, e.g. Tables, annotations, free text
3. Preprocess the data
4. Apply basic IR techniques
5. Apply machine learning methods
   - Learn one model for all product pairs
   - Learn multiple models – one for each product (you need a big gold standard)
   - Feature vectors: tokens (binary or tfidf weights) , similarity scores computed with different measures
6. Evaluate and if necessary refine your gold standard

Phase 2 – Subgroup 3: Categorization

Goal: Learn a model to assign the correct category to every product

How?

1. Define an initial hierarchy of product categories
2. Select a hierarchical classification method [1]
3. Create a gold standard considering the requirements of your method
   - Min. 200 annotated products in the form of <product_a : action camera>
4. Use product features and apply hierarchical classification
   - Your features should be simply induced, e.g. Bag of Words model.
5. Evaluate and if necessary refine your gold standard

[1] Silla, Carlos N., and Alex A. Freitas. "A survey of hierarchical classification across different application domains." Data Mining and Knowledge Discovery 22.1-2 (2011): 31-72.

Phase 2 – Subgroup 4: Sentiment Analysis

Goal: Perform aspect based sentiment analysis on product reviews

How?

1. Profile the review information and if necessary crawl more reviews from the product pages
   - How many reviews?
   - Are there predefined aspects/ information about reviewers?
2. Identify reviewed features and subfeatures
   - e.g. „The *display* is of great quality“ VS „Although the *display screen* is big, I am not satisfied with its *resolution*“
3. Extract the sentiment for each feature and overall for the product
   - Usage of linguistic patterns, e.g. Adjective + Noun : This bag is of great material
   - Consideration of negation and degree words
   - Usage of polarity dictionaries
4. Evaluate your scoring against the extracted score

Phase 2 Results & Deliverable

Duration: 13.04.2018 – 15.05.2018

Deliverables:

1. **A 10 min presentation from each subgroup**

   The presentation slides should be provided by the end of the meeting.

2. **Data and Code**

   Add your data and code in a zipped folder and send (URL) via e-mail

3. **Member to subtask report**

   Send one excel sheet per team explaining who did what together with the deliverables.

All deliverables should be sent to Chris & Anna!

Potentially Useful Software

- Crawling
  - Scrapy : https://scrapy.org/
  - Any23
- Data Integration
  - Winte.r Framework : https://github.com/olehmberg/winter
  - Silk Framework : https://github.com/silk-framework/silk
- Data Mining, Machine Learning
  - RapidMiner : https://rapidminer.com/
- Natural Language Processing
  - Stanford NLP: https://nlp.stanford.edu/software/
  - RiTa library: http://rednoise.org/rita/download.php

Related Work for Feature Extraction

- Qiu, Disheng, et al. "Dexter: large-scale discovery and extraction of product specifications on the web." *Proceedings of the VLDB Endowment* 8.13 (2015): 2194-2205
- Petar Petrovski, Christian Bizer: Extracting Attribute-Value Pairs from Product Specifications on the Web. International Conference on Web Intelligence (WI2017), pp. 558-565, Leipzig, Germany, August 2017.
- Petrovski, Petar, Volha Bryl, and Christian Bizer. "Learning regular expressions for the extraction of product attributes from e-commerce microdata." Proceedings of the Second International Conference on Linked Data for Information Extraction-Volume 1267. CEUR-WS. org, 2014
- Dalvi, Nilesh, Ravi Kumar, and Mohamed Soliman. "Automatic wrappers for large scale web extraction." *Proceedings of the VLDB Endowment* 4.4 (2011): 219-230
- Probst, Katharina, et al. "Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions." *IJCAI*. Vol. 7. 2007
- Petar Petrovski, Anna Primpeli, Robert Meusel, Christian Bizer: The WDC Gold Standards for Product Feature Extraction and Product Matching. 17th International Conference on Electronic Commerce and Web Technologies (EC-Web 2016), Porto, Portugal, September, 2016.

Related Work for Identity Resolution

- Petrovski, Petar, Volha Bryl, and Christian Bizer. "Integrating product data from websites offering microdata markup." *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014.
- Kannan, Anitha, et al. "Matching unstructured product offers to structured product specifications." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011
- Köpcke, Hanna, Andreas Thor, and Erhard Rahm. "Evaluation of entity resolution approaches on real-world match problems." *Proceedings of the VLDB Endowment* 3.1-2 (2010): 484-493
- Christen, Peter. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012
- Isele, Robert, and Christian Bizer. "Learning linkage rules using genetic programming." *Proceedings of the 6th International Conference on Ontology Matching-Volume 814*. CEUR-WS. org, 2011
- Petar Petrovski, Anna Primpeli, Robert Meusel, Christian Bizer: The WDC Gold Standards for Product Feature Extraction and Product Matching. 17th International Conference on Electronic Commerce and Web Technologies (EC-Web 2016), Porto, Portugal, September, 2016.
- Petar Petrovski, Christian Bizer: Learning Expressive Linkage Rules from Sparse Data. Under review at the Semantic Web Journal, 2018.

Related Work for Categorization

- Meusel, Robert, et al. "Exploiting microdata annotations to consistently categorize product offers at web scale." *International Conference on Electronic Commerce and Web Technologies*. Springer International Publishing, 2015
- Navigli, Roberto, Paola Velardi, and Stefano Faralli. "A graph-based algorithm for inducing lexical taxonomies from scratch." IJCAI. Vol. 11. 2011.
- Ristoski, Petar, et al. "Large-scale taxonomy induction using entity and word embeddings." Proceedings of the International Conference on Web Intelligence. ACM, 2017.
- Silla, Carlos N., and Alex A. Freitas. "A survey of hierarchical classification across different application domains." Data Mining and Knowledge Discovery 22.1-2 (2011): 31-72.
- Melo, André, Heiko Paulheim, and Johanna Völker. "Type prediction in rdf knowledge bases using hierarchical multilabel classification." Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics. ACM, 2016.
- Poon, Hoifung, and Pedro Domingos. "Unsupervised ontology induction from text." Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

Related Work for Sentiment Analysis

- Liu, Bing. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.
- Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.
- Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005.
- Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.
- Hu, Minqing, and Bing Liu. "Mining opinion features in customer reviews." *AAAI*. Vol. 4. No. 4. 2004.
- Eirinaki, Magdalini, Shamita Pisal, and Japinder Singh. "Feature-based opinion mining and ranking." *Journal of Computer and System Sciences* 78.4 (2012): 1175-1184.

Questions?

Subtask Assignment

- Subgroup 1: Feature Extraction
  - Members: Chung, Adela
  - Timeslot for 27.04: 9:15 – 9:35
- Subgroup 2: Identity Resolution
  - Members: Larissa, Ersejda
  - Timeslot for 27.04: 9:35 – 9:55
- Subgroup 3: Categorization
  - Members: Bengi, Anjeza
  - Timeslot for 27.04: 10:15 – 10:35
- Subgroup 4: Sentiment Analysis
  - Members: Matthias, Murad
  - Timeslot for 27.04: 10:35 – 10:55