UNIVERSITY OF MANNHEIM

Team Project FSS 2018

Mining Product Data from the Web

Phase 3

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**

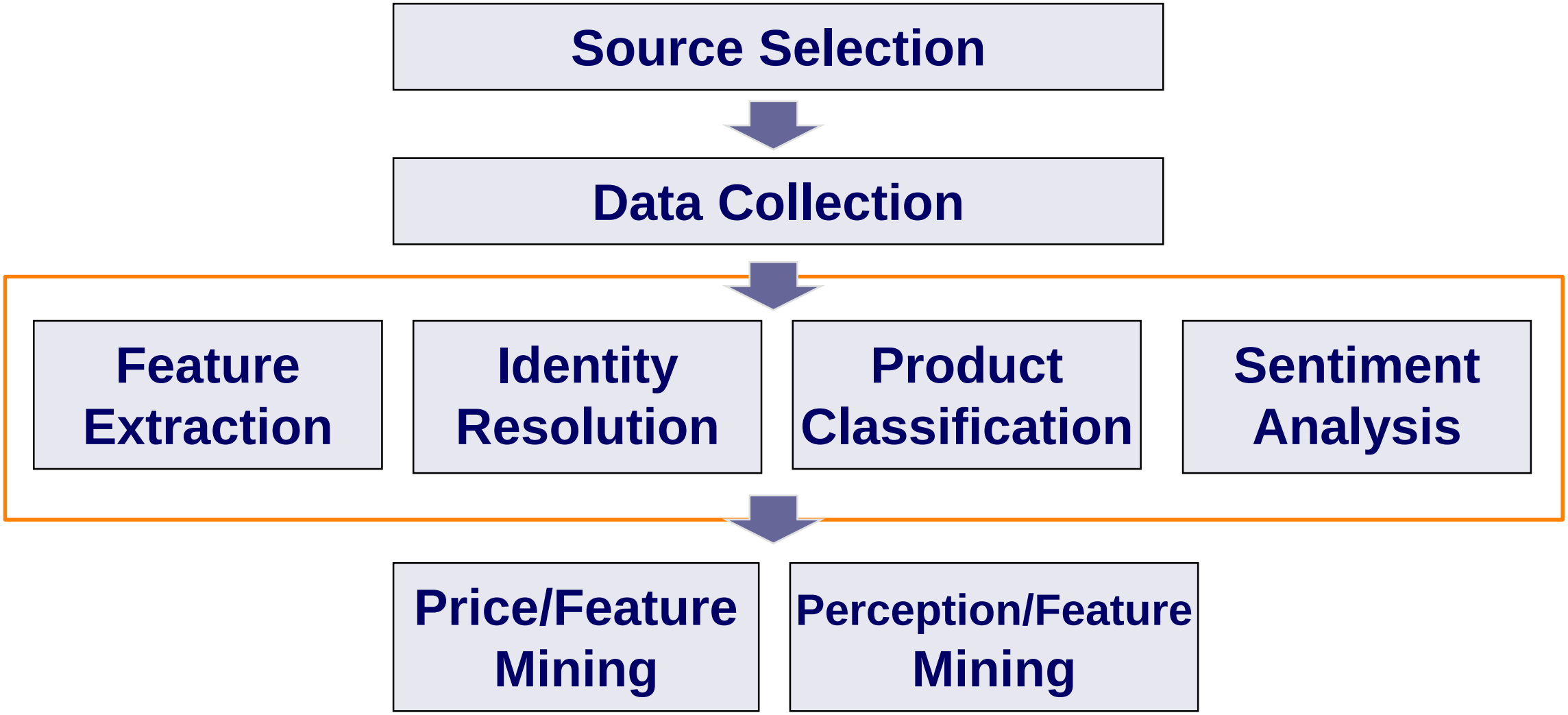
5. How to understand user perception? → **Sentiment Analysis**

Phase 3 is 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 3

Date	Session
Friday, 18.05.2018, 9:15am	Introduction to Phase 3
Tuesday, 29.05.2018, 8:30am	Meet Anna and discuss plans
Sunday, 01.07.2018	**3rd Deliverable:** 8-12 pages report from each subgroup, code & data

Results from Phase 2

What should you have from Phase 2?

Feature Extraction team

- Dictionary-based feature extraction method
- A baseline generic approach
- Gold standard
- Evaluation of your methods
- Set of extracted features per page and profiling statistics about it

Identity Resolution team

- Baseline BoW identity resolution approach
- Machine learning based BoW identity resolution approach
- Gold standard
- Experiments with different sets of features (e.g. schema.org annotations, all html)
- Evaluation of your methods
- Set of corrrespondences and profiling statistics about it

Results from Phase 2

What should you have from Phase 2?

Product Classification team

- Taxonomy of product categories for bags and cameras
- Baseline hierarchical classifier
- Experiments with different sets of features
- Gold standard
- Evaluation of your method
- Profiling statistics about the categories of the pages of your corpus

Sentiment Analysis team

- Gather bigger review corpus and profiling statistics about it
- Define your aspects for sentiment analysis (think about future mining e.g. include price)
- Baseline method for extracting polarity of aspects in reviews
- Evaluate baseline method

Phase 3 – Subgroup 1: Feature Extraction

Goal: Extract clean feature - value pairs from the product pages and perform schema maching (refinement and wider set of attributes)

How?

1. Perform error analysis on your current results
   - What went wrong in the feature extraction?
   - What went wrong in the schema matching?
2. Enhance your feature extraction techniques
   - Extract tables and classify them into specification/ non-specification ones. Consider [1,2,3]
3. Enhance your schema matching techniques
   - Use the results of the identity resolution team (or their gold standard) for duplicate-based schema matching

[1] 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
[2] 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.
[3] 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

Phase 3 – Subgroup 2: Identity Resolution

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

(using a wider set of attributes and additional webpages to learn product recognizers)

How?

1. Use the results of the feature extraction subteam for phase 2 and test more expressive learning algorithms
   1. linear regression on attribute similarity scores
   2. tree-based models on attribute similarity scores
2. Perform error analysis on best BoW- and Attribute-based models
3. Use product IDs (GTIN) to learn better product recognizers
   - Learn product recognizers (one binary classifier per product) based on the data in the catalog
   - Use products identifiers to find additional web pages describing the product (google the IDs).
   - Learn improved product recognizers using the catalog and the new pages as training data
   - Compare performance of base recognizers and improved recognizers: Find out how much the results can be improved using additional webpages that were found using product IDs (like GTIN).

Phase 3 – Subgroup 3: Categorization

Goal: Profile/Cluster categorization taxonomies from different websites in order to identify alternative categorization approaches.

How?

1. Improve hierarchical classification
    1. Use the results of the feature extraction subteam for phase 2 and update your features
    2. Apply more sophisticated hierarchical classification methods
2. Profile Categorization Taxonomies of different websites
    1. Extract alternative categorization taxonomies/paths from the shops using schema.org breadcrumb, schema.org product or offer category, or regular expressions.
    2. Profile the categorization taxonomies/paths
    3. Manually group the taxonomies/paths by categorization „idea“ / level of detail.
3. Automatically Cluster Categorization Taxonomies
    1. Apply taxonomy clustering method to automatically find grouping (using results from 3. as ground truth).

Phase 3 – Subgroup 4: Sentiment Analysis

Goal: Perform aspect based sentiment analysis on product reviews (refinement)

How?

1. Perform error analysis
   - Identify the most difficult factors for extracting sentiments, i.e. degree words, sentence structure, negation
2. Use feature extraction methods or feature names/values from the Feature Extraction subteam to locate more easily the aspects in your texual reviews
3. Use supervised models to learn important polarity words for every aspect
4. Create a gold standard of the form <Review_id, Aspect_id, Polarity>

Phase 3 Results & Deliverable

Duration: 18.05.2018 – 01.07.2018

Deliverables:

1. 8 – 12 from each subgroup

Reports should be 8-12 pages single column

- including appendixes
- not including the bibliography
- every additional page reduces your grade by 0.3
- Created with Latex template of the Data and Web Science group (http://dws.informatik.uni-mannheim.de/en/thesis/masterthesis/)

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?