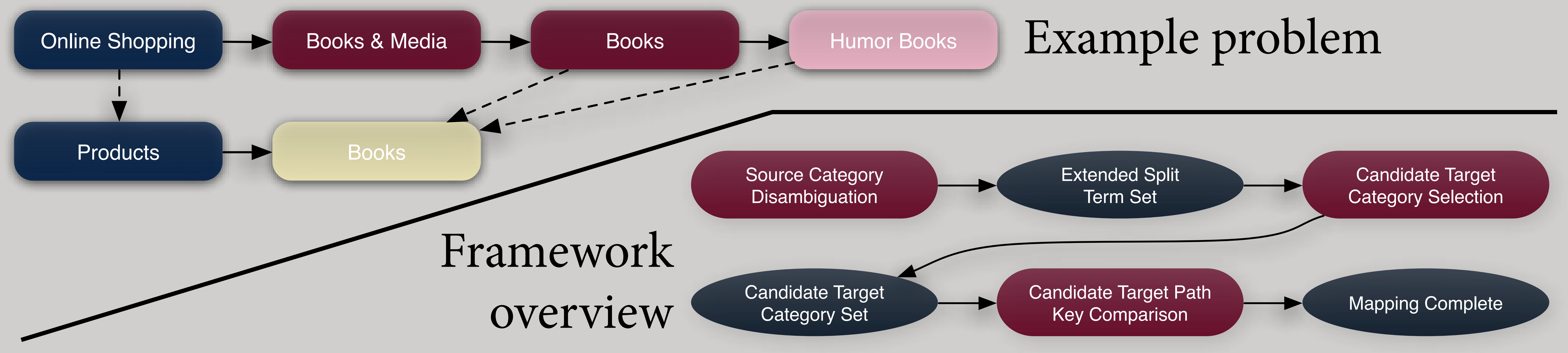
SCHEMA - An Algorithm for Automated Product Taxonomy Mapping in E-commerce

Steven Aanen, Lennart Nederstigt, Damir Vandic, and Flavius Frasincar

In a Nutshell



Online Shopping
Books & Media
Books
Humor Books
Products
Books

Example problem

Framework overview

Source Category Disambiguation
Extended Split Term Set
Candidate Target Category Selection
Candidate Target Category Set
Candidate Target Path Key Comparison
Mapping Complete

(1) Source Category Disambiguation

Generating Extended Split Term Set:

- For parent and each child, create split term set (to address composite categories)
- Disambiguate each split term, which gives the Extended Term Set
- The Extended Split Term Set contains a set of synonyms of the correct sense for each individual split term (set of Extended Term Sets)

Word sense disambiguation procedure:

- Lesk's algorithm with heuristics to reduce computational complexity
- WordNet for finding related synsets based on hypernymy, hyponymy, meronymy, and holonymy

(2) Candidate Target Category Selection

- Use Extended Split Term Set to compare each element from a Extended Term Set to each target category
- Target taxonomy is splitted in the same way into Extended Term Sets
- Comparison done using the normalized Levenshtein Distance metric
- Source category only matches when it is a subset of the target category (e.g., 'Music & Videos' does not match 'Music')

(3) Cand. Target Path Key Comparison

- Procedure to select the 'best' matching candidate target path
- Uses structural and lexical relatedness
- Paths are converted to key sequences
- If two categories have the same Extended Split Term Set, we assign them the same key
- Algorithm uses Damerau-Levenshtein distance to compute a similarity between to key lists

Evaluation

- Three data sets:
  - Amazon.com - ~2,500 categories
  - Overstock.com - ~1,000 categories
  - ODP - 44,000 categories
- Evaluation done on the 6 mapping combinations
- Manually mapped 6x500 categories
- Results:

Algorithm	Precision	Recall	F1-measure
PROMPT	28.93%	16.69%	20.75%
Park & Kim	47.77%	25.19%	32.52%
SCHEMA	42.21%	80.73%	55.10%

Contact

Damir Vandic
Erasmus University Rotterdam
E-mail: vandic@ese.eur.nl
Web: http://damirvandic.com/

This publication is supported by an NWO Mosaic scholarship (project 017.007.142).
Project members: Damir Vandic, Uzay Kaymak (promotor), and Flavius Frasincar (co-promotor).

NWO Netherlands Organisation for Scientific Research
ERASMUS UNIVERSITEIT ROTTERDAM