Model Words-Driven Approaches for Duplicate Detection on the Web

Marnix de Bakker, Flavius Frasincar,
Damir Vandic, and Uzay Kaymak

vandic@ese.eur.nl
Erasmus University Rotterdam

ERASMUS UNIVERSITEIT ROTTERDAM

Introduction

- Duplicate detection of products
- Aggregation of Web product offerings
- Example:
  - Samsung - 40” Class / LCD / 1080p / 60Hz / HDTV
  - Samsung 40” 1080p 60Hz LCD HDTV LN40D503

Algorithms

We investigate three algorithms:

- title model words method
  D. Vandic et. al. *Faceted Product Search Powered by the Semantic Web* Decision Support Systems, 53(3):425–437, 2012.
- attribute distance method (new)
- extended model words method (new)

Title model words method

The main steps (high-level):

1. First, perform a word-based cosine similarity check
2. Search for a model word pair where the non-numeric parts are *approximately* the same, but the numeric parts are different.
3. Otherwise, compute average similarity between model words

Title model words method

Example 1

- ‘Samsung - 46” Class/ LED / 1080p / **120Hz** / HDTV’

vs.

- ‘Samsung - 46” Class/ LED / 1080p / **200Hz** / HDTV’

Title model words method

Example 2

- ‘Samsung - **55”** Class/ LED / 1080p / 120Hz / HDTV’

vs.

- ‘Samsung - **46”** Class/ LED / 1080p / 120Hz / HDTV’

Attribute Distance Method

- Uses key/value pairs (KVP’s) in the process
- Starts with each product in separate cluster
- Matches products using previous method
- In case of no match, KVP’s are employed:
    - all matching keys are found and similarity is updated by the KVP value distances for these keys

Extended Model Words Method

- Same as previous algorithm, only in case of no match a different approach is taken:
  - instead of computing similarity between values for matching keys, we compute the similarity for all pair of words (not only model words)
  - reason: data often differently structured

Extended Model Words Method

Example differently structured data

- TV from Bestbuy.com has the KVP:
  [‘Product Weight’,
  ‘19.1 lbs. with stand (16.9 lbs. without)’
  ]
- Same TV on NewEgg.com:
  [‘Weight Without Stand’, ‘16.9 lbs.’]
  [‘Weight With Stand’, ‘19.1 lbs.’]

Evaluation setup

- Data set of 282 TV's from two Web shops
  - BestBuy.com and NewEgg.com
- There are 82 pairs (164 products) that are duplicates
- 20 random test sets (10% of total size)
- Wilcoxon signed rank test

Evaluation results

Method	Average F1-measure	Average precision	Average recall
Title model words	0.357	0.556	0.279
Attribute distance	0.529	0.531	0.556
Extended model words	0.607	0.637	0.597

Evaluation results

H0: row < col

p-values	Title model words	Attribute distance	Extended model words
Title model words	X	0.082	0.002
Attribute distance	0.923	X	0.285
Extended model words	0.999	0.727	X

Conclusions and future work

- Two new methods proposed for product duplicate detection
- Benchmarked against an existing approach
- Extended model words method is best performing on F1
- Recall is boosted for the new methods because KVP's are taken into account

Conclusions and future work

Future work

- Experiment with more distance measures
- Use semantics of product attributes/values
- Investigate a hybrid method that combines the good aspects of the ‘attribute distance’ and ‘extended model words’ methods

Questions?