

# On the Investigation of Similarity Measures for Product Resolution

Krisztian Balog  
Norwegian University of Science and Technology

LCIAI'11 Workshop on Discovering Meaning On the Go in Large Heterogeneous Data (LHD-11)  
Barcelona, Spain, July 2011

## Motivation

Google search results for "lézerkard játék". The results list several Star Wars laser sword products, including "STAR WARS Jedi Párbaj Lézerkard" (13990 Ft), "STAR WARS Sith Párbaj Lézerkard" (13990 Ft), "STAR WARS Grievous Tábomok" (14990 Ft), "STAR WARS Elektronikus Lézerkard" (8990 Ft), and "STAR WARS Elektronikus Lézerkard" (8990 Ft). Each result includes a brief description and a price tag.

## Motivation

Google search results for "lézerkard játék". The results are similar to the previous screenshot, but the "STAR WARS DUPLA LEZERKARD" product is highlighted with a red box. The product description mentions it is a Hasbro Star Wars "dupla lézerkard" (double laser sword) with a Klonok (Cloners) theme, featuring a lightsaber and a blaster.

## Motivation

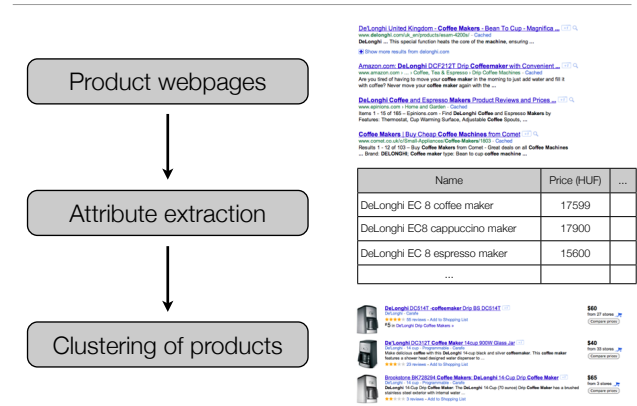
Product page for "STAR WARS DUPLA LEZERKARD". The page shows the product name, a price of 13,990 Ft, and a detailed description. The description states: "A Hasbro Star Wars 'dupla lézerkard' készlet a Klonok háborúja című rajzfilmről megismert különleges kardok tartalmazza. A két kard összeilleszthető, így egyszerre mozgatható harc közben. A két kard világít, a nagy onduló világító és összecsapó haragot is ad. A pengék kihúzhatók, gombnyomással ki és bekapcsolhatók. A két kard csatlakozás elrejtődik a nagy nyelvényben, előzárva a pengékkel harcolhatunk. A készletben két lézerkard található, melyek működéséhez 3 darab AA, valamint 3 darab AAA elem szükséges, melyeket a játék nem tartalmaz. Hat éves kortól ajánlott. - EAN500994432560".

## Motivation

Product listing for various DeLonghi coffee makers. The listing includes several models with their prices and key features:

- DeLonghi DC614T\_coffeemaker Drip BS DC614T: \$60
- DeLonghi DC312T\_Coffee Maker 14cup 800W Glass Jar: \$40
- Brooksstone BK728294\_Coffee Makers DeLonghi 14-Cup Drip Coffee Maker: \$65
- DeLonghi BC0984B Espresso Maker, Caffè Neri 3-in-1: \$161
- DeLonghi DCF212T 12-Cup Glass Carafe Drip Coffee Maker Black: \$32
- DeLonghi DCF212T Programmable Coffee Maker 12 Cup: \$50
- DeLonghi DCF212T Programmable 12-Cup Coffee Maker, Stainless Steel: \$39

## Product resolution pipeline



## Contributions

- Real-world data set for two e-commerce segments: electronics and toys
- Similarity functions for various product attributes
- Metric for measuring the discriminative power of similarity functions

## Data collection

- Product page collection (candidate sets)
  - Product query against a web search engine (Google)
  - Collected to N (~30) pages
  - Removed non-product pages
- Attribute extraction manually
- Clustering into equivalence classes manually (equivalence classes)

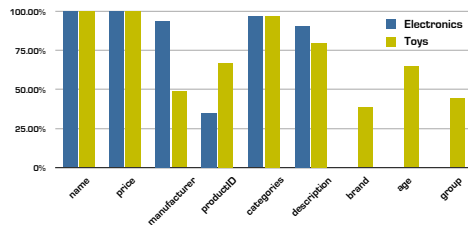
## Example product queries

Electr.	Toys
lenovo thinkpad edge 11	candamir
kingston microsd 8gb	eichhorn railway kit
samsung led tv 32"	eiffel tower puzzle 3d

## Data collection

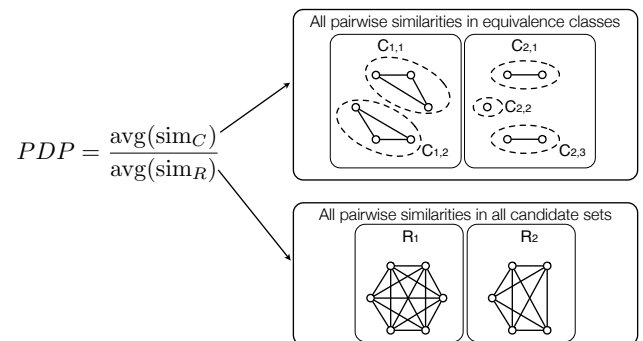
	Electr.	Toys
#queries	25	30
#queries with >1 clusters	6	17
avg query length in characters	19.56	21.6
avg query length in terms	3.4	3.0
avg #product pages per query	4.6	5.7
min/max #product pages per query	1/8	1/11
avg #equivalence classes per query	1.24	2.66
min/max #equivalence classes per query	1/2	1/7
#different web-shops	46	38

## Product attributes



## Pairwise Discriminative Power (PDP)

- Goal: to be able to measure how well similarity metrics can set apart entities



## Similarity measures for product attributes

## Product name

- Various string matching methods

Distance function	Electr.	Toys	
Levenshtein	1.0343	1.2556	} Character-based
MongeElkan	1.0248	1.1456	
Jaro	1.0184	1.1508	
JaroWinkler	1.0167	1.1309	
QGramsDistance	1.0470	1.2961	} Term-based
MatchingCoefficient	1.0511	1.3387	
DiceSimilarity	1.0501	1.3181	
OverlapCoefficient	1.0491	1.2804	
JaccardSimilarity	1.0602	1.4190	
CosineSimilarity	1.0501	1.3123	

## Product name - examples



- Star-wars laser sword
- Star-Wars Jedi Dual Sword
- Star Wars laser sword (TV advertised product)
- Laser sword with lights and sound
- Space laser sword
- Hasbro Star Wars Clone Wars electronic laser sword / lightsaber



- DeLonghi EC 8 coffee maker
- DeLonghi EC8 cappuccino maker
- DeLonghi EC 8 espresso maker
- DeLonghi EC-8 espresso maker
- DeLonghi EC 8 presso and cappuccino maker

## Price

$$\text{sim}_{\text{price}}(p_i, p_j) = \frac{\min(p_i, \text{price}, p_j, \text{price})}{\max(p_i, \text{price}, p_j, \text{price})}$$

- Value between 0 and 1
- 1 iff the two products have the exact same price

	Electr.	Toys
Price	1.0306	1.1882

## Manufacturer

---

- Same string distance metrics as for product name

Distance function	Electr.	Toys
QGramsDistance	1.0276	1.0979
JaccardSimilarity	1.0275	1.1056

## Product ID

---

- Strict string matching (i.e., binary function)

## Comparison

---

Attribute	Electronics			Toys		
	%R	%C	PDP	%R	%C	PDP
Product name	100.0	100.0	1.0602	100.0	100.0	1.4190
Price	96.9	98.2	1.0306	89.6	99.1	1.1882
Manufacturer	95.3	94.6	1.0276	30.8	36.4	1.1056
Product ID	10.9	9.4	1.3333	45.0	44.2	1.9212

- %R, %C: fraction of all product pairs for which similarity can be established in the candidate sets and equivalence classes, respectively

## Summary

---

- First steps towards the **task of automatic product webpage resolution**
- Defined **similarity functions** for various **product attributes**
- Introduced a **metric for comparing similarity functions**
- Performed an **experimental evaluation** on two e-commerce domains

## Future directions

## Additional product attributes

---

- Categories
- Description
- Toys-specific
  - Age group (e.g., 3-6 years)
  - Target group (for boys, girls, or both)

## PDP measure

---

- Simple and intuitive, but yet to be validated
  - Do higher PDP values indeed lead to better clustering performance?

## Data collection

---

- Current data set is too small for statistically robust comparisons
  - Repeat experiments with a larger collection
- Finding product pages using web search engines is a viable method, but only 1 out of 6 web search results is a product page
  - Consider product comparison sites too
  - Issue more targeted queries, e.g., append currency to the product name
  - Query reformulation techniques, e.g., blind relevance feedback

## Questions?

---



@krisztianbalog



<http://krisztianbalog.com>



<http://www.linkedin.com/in/krisztianbalog>