

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

Steven Aanen, Lennart Nederstigt  
Damir Vadic, Flavius Frasinca



# Terminology

- source taxonomy
- target taxonomy
- category = single node in a taxonomy
- (category) path = list of nodes (starting from root node)

# Product taxonomies

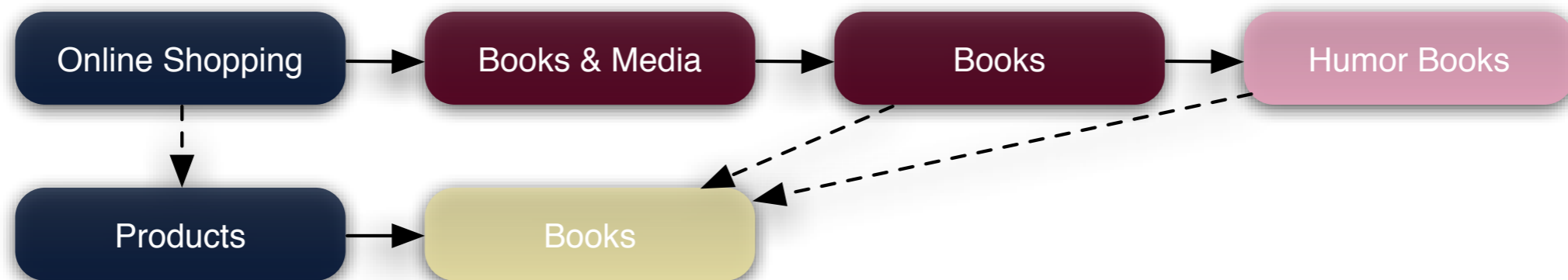
Important aspects of product taxonomies:

- composite categories
- varying degree of granularity
- root category of taxonomies

# Product taxonomies

Important aspects of product taxonomies:

- composite categories
- varying degree of granularity
- root category of taxonomies



# Related work

- The algorithm by Park & Kim  
*“Ontology Mapping between Heterogeneous Product Taxonomies in an Electronic Commerce Environment”*
- PROMPT algorithm in PROMPT Suite  
*“The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping”*

# SCHEMA overview

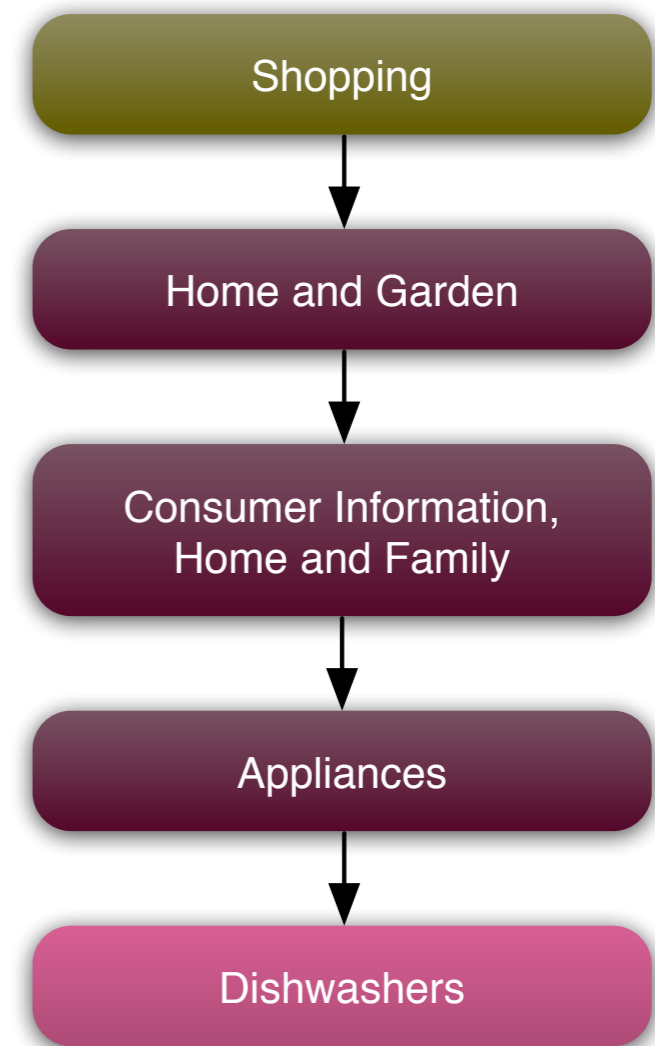
- Input is a source category path
- Output is a target category path (or 'None')
- SCHEMA consists of three steps
  1. source category disambiguation
  2. candidate target category selection
  3. candidate target path key comparison

# SCHEMA overview

- 1. source category disambiguation**
2. candidate target category selection
3. candidate target path key comparison

# Source category disambiguation

- Example category path
  - Dishwashers can have two meanings
  - From the path, the meaning is clear to humans
- Based on the Lesk algorithm





Shopping



Home and Garden



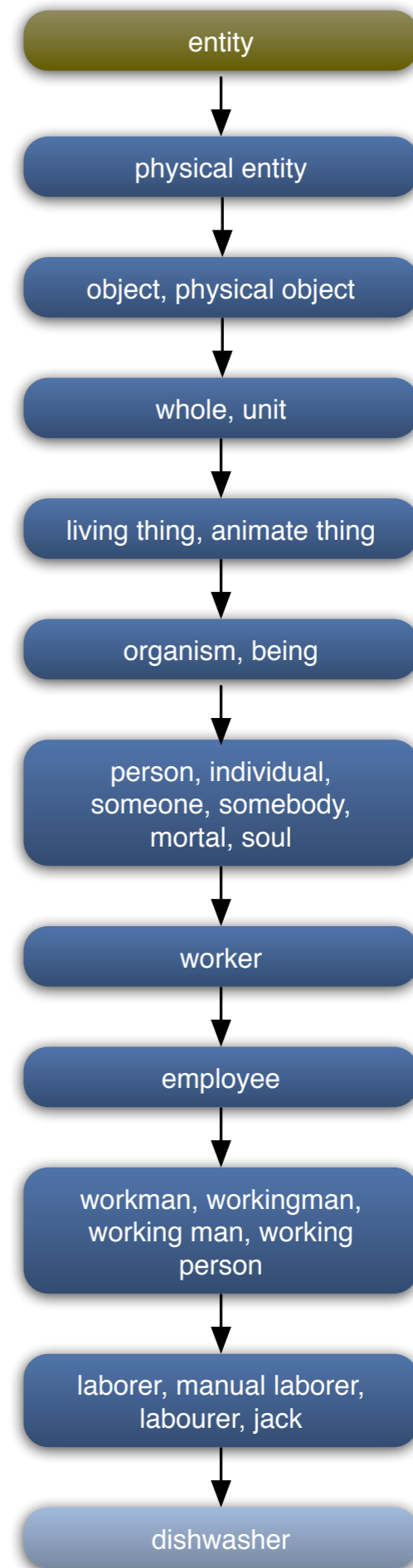
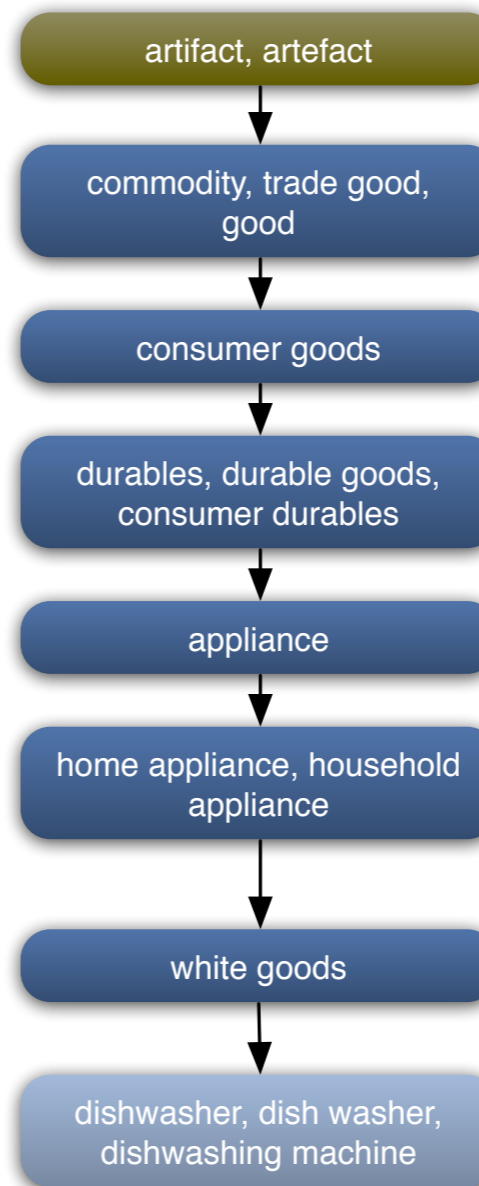
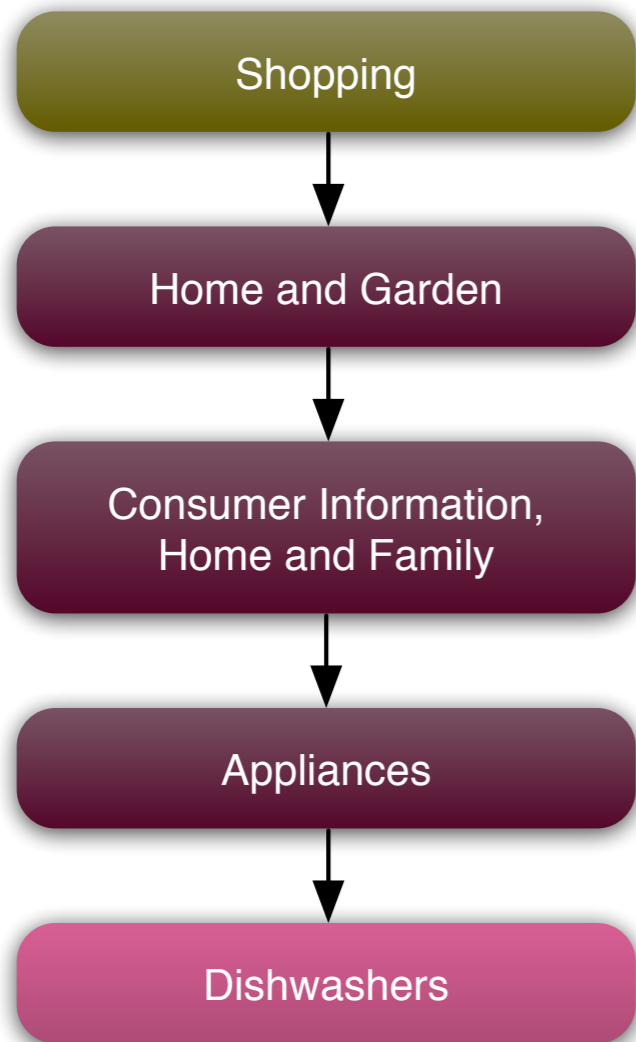
Consumer Information,  
Home and Family

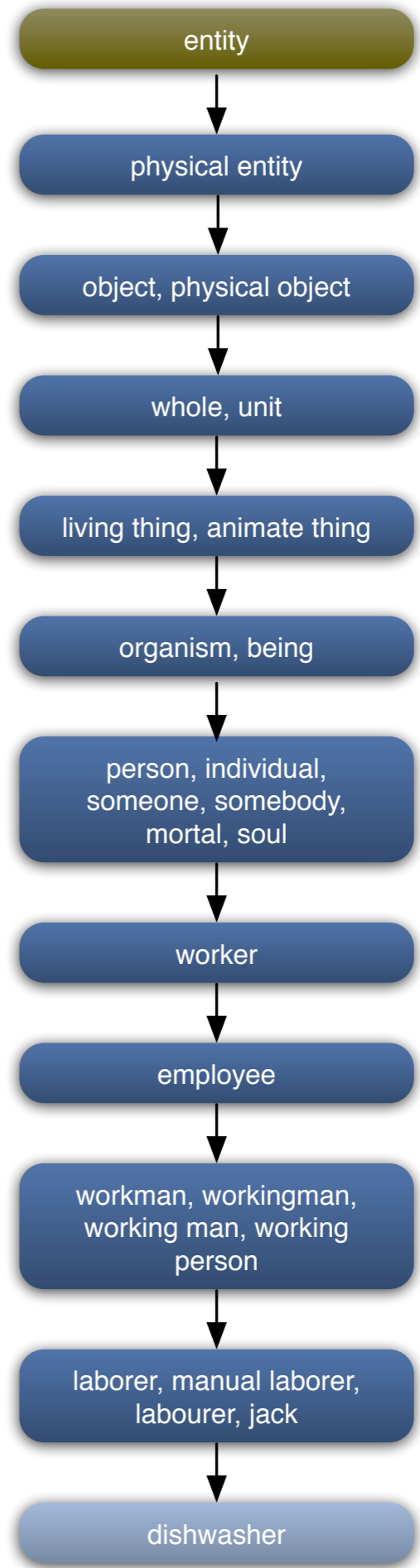
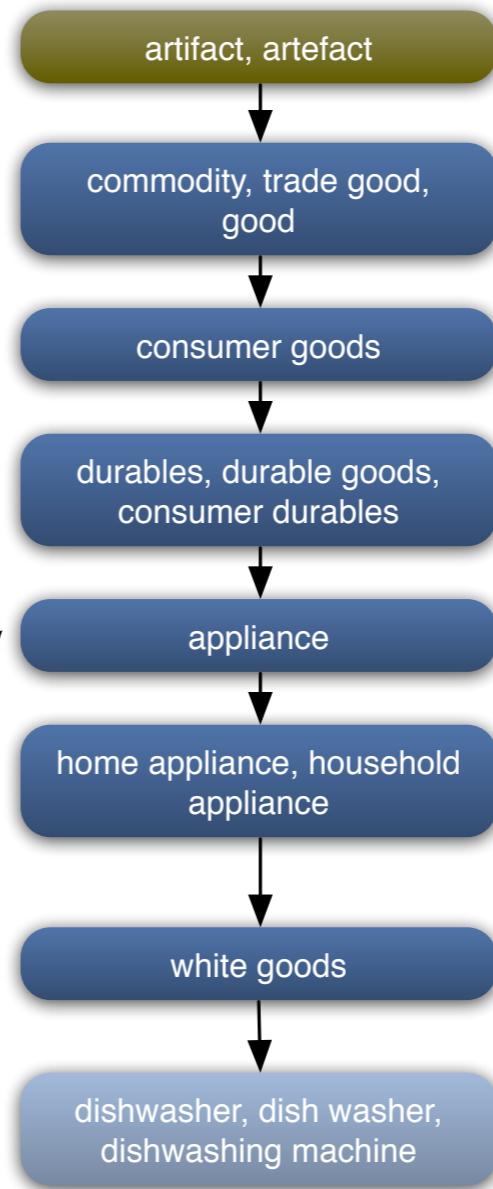
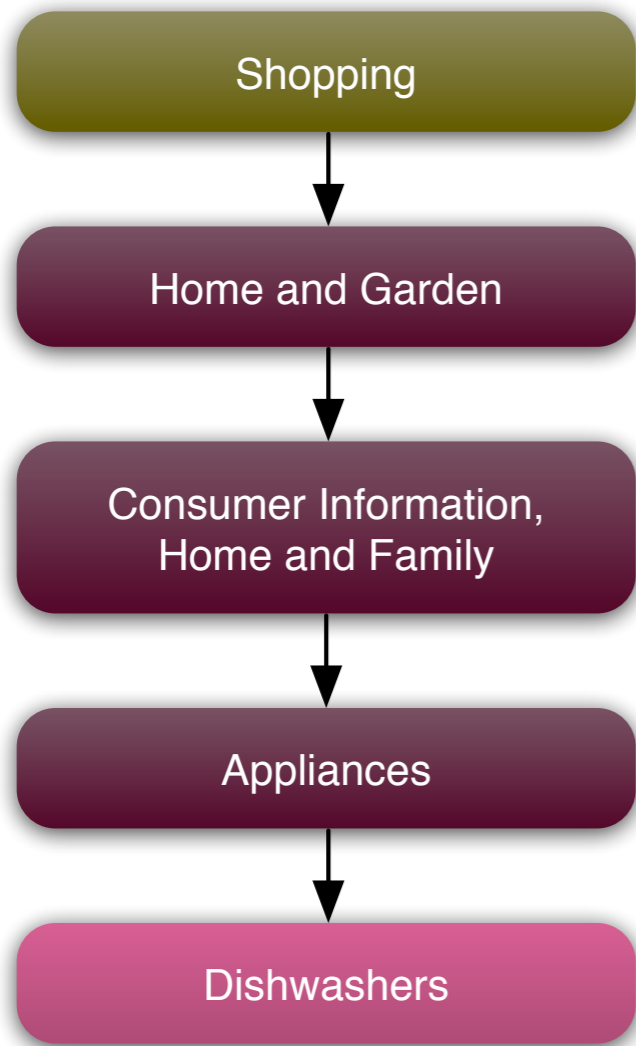


Appliances

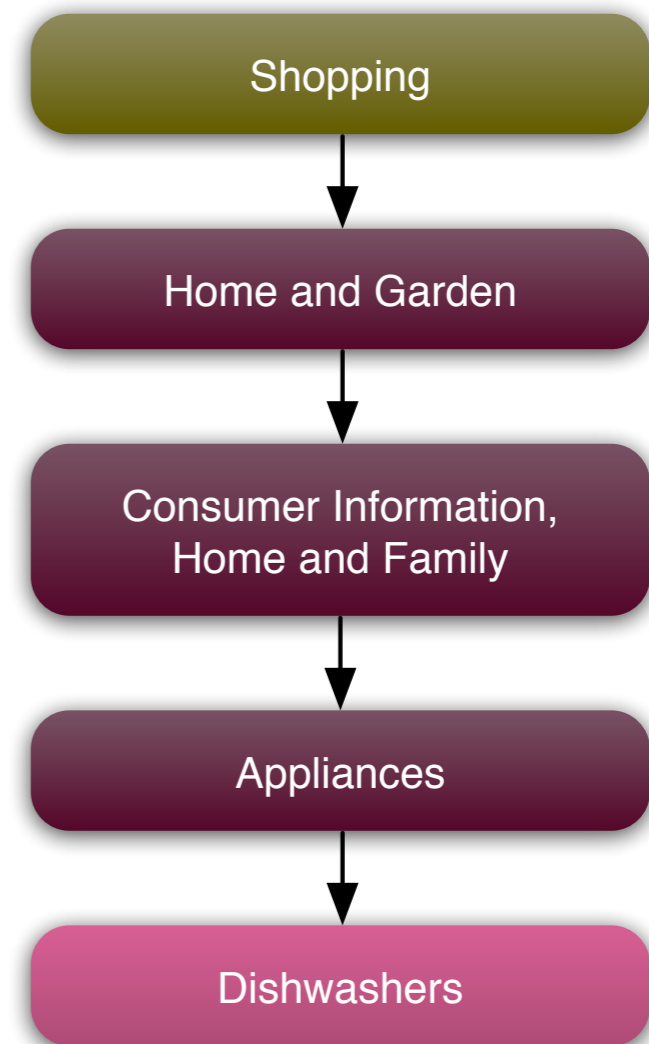


Dishwashers

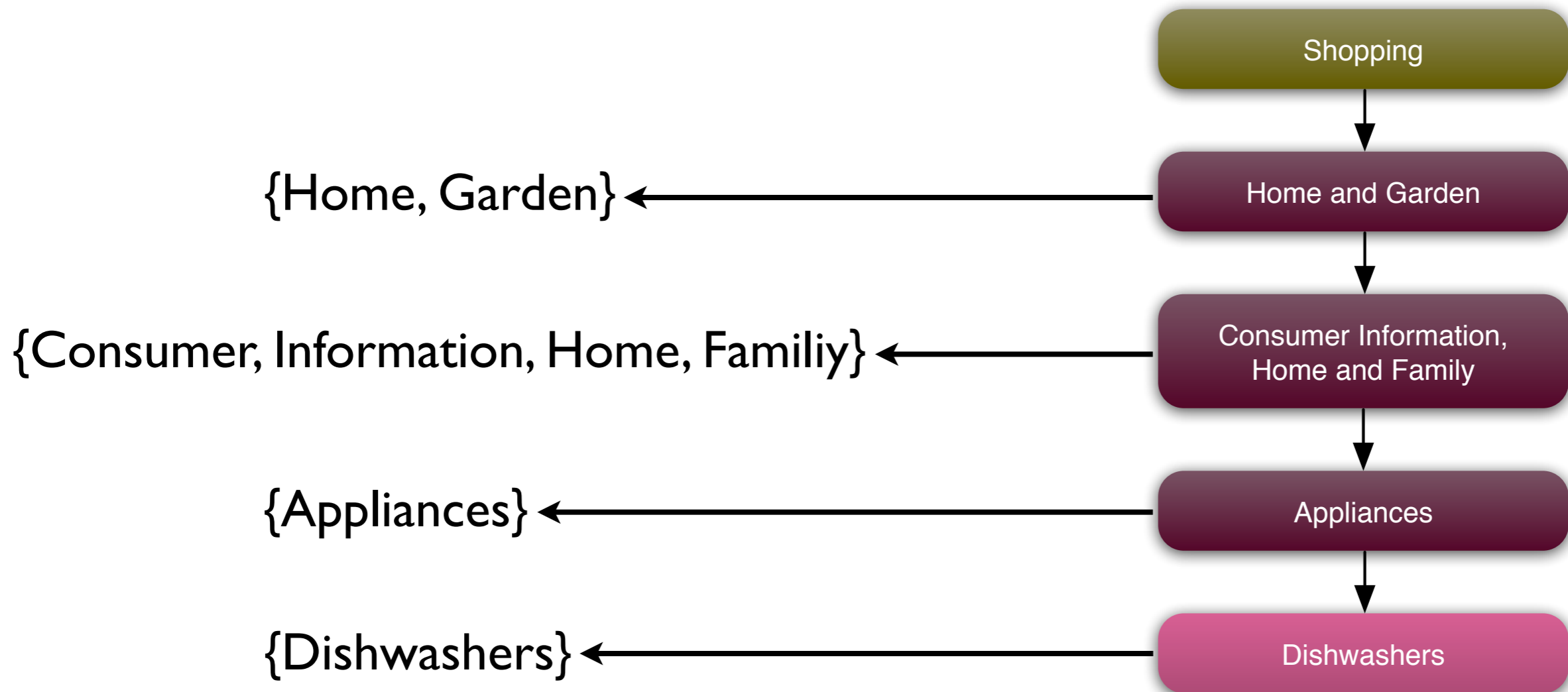




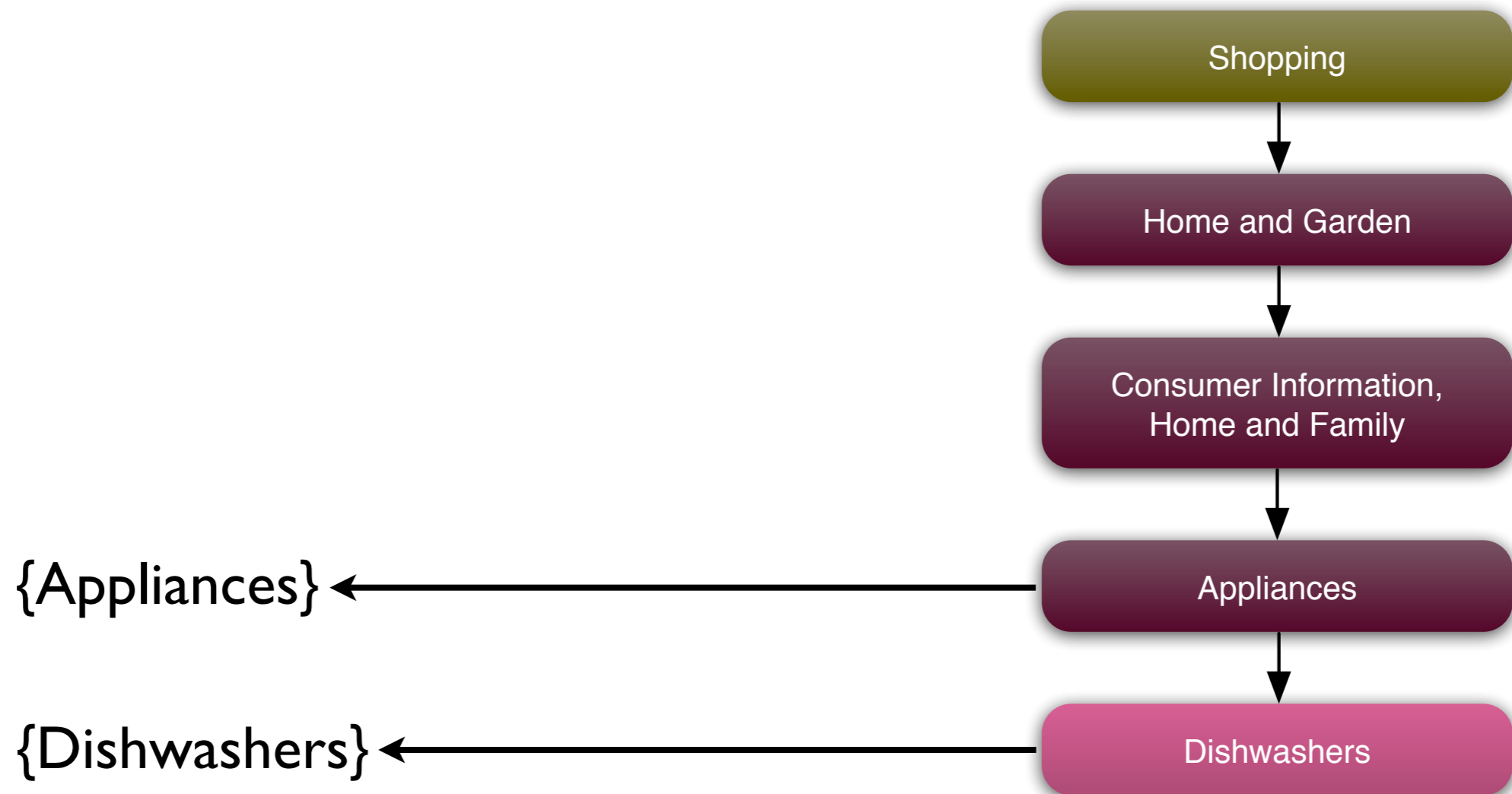
# Source category disambiguation



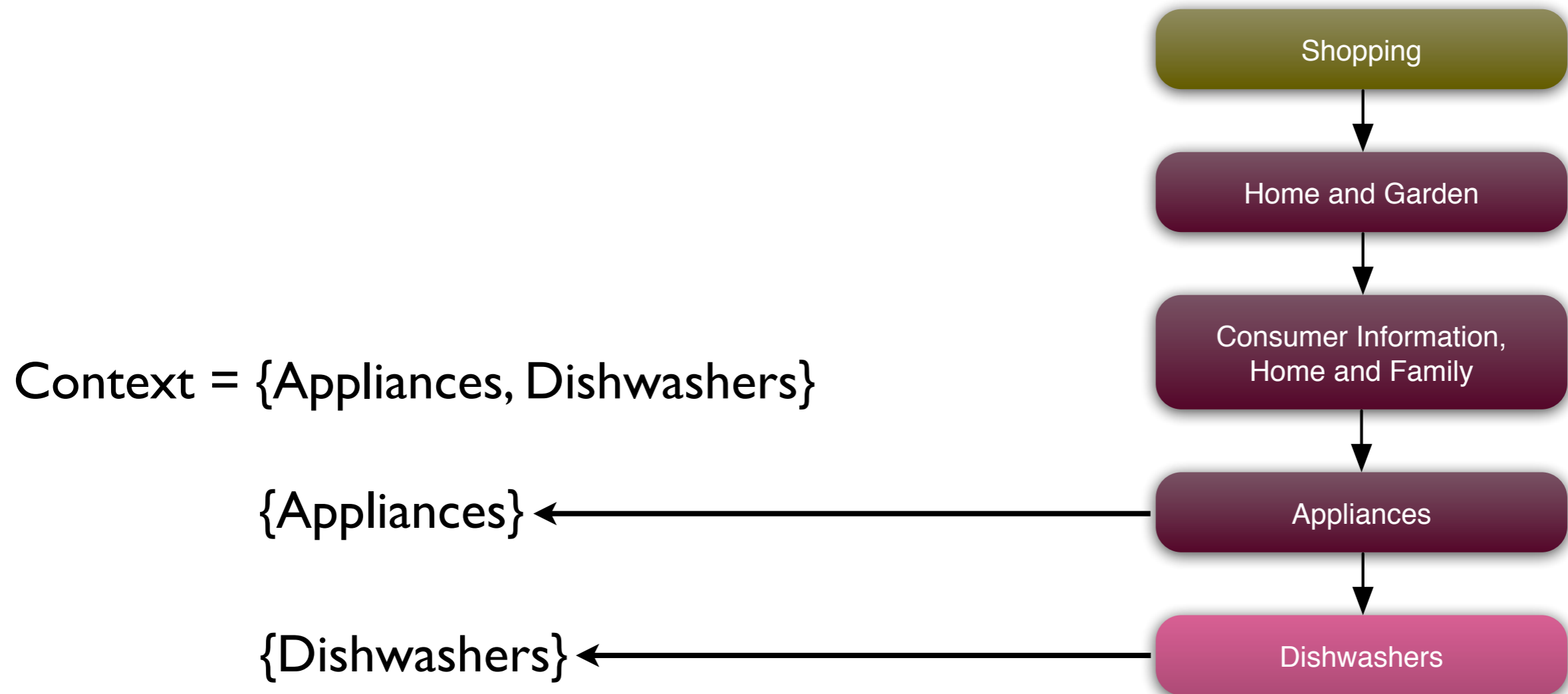
# Source category disambiguation



# Source category disambiguation



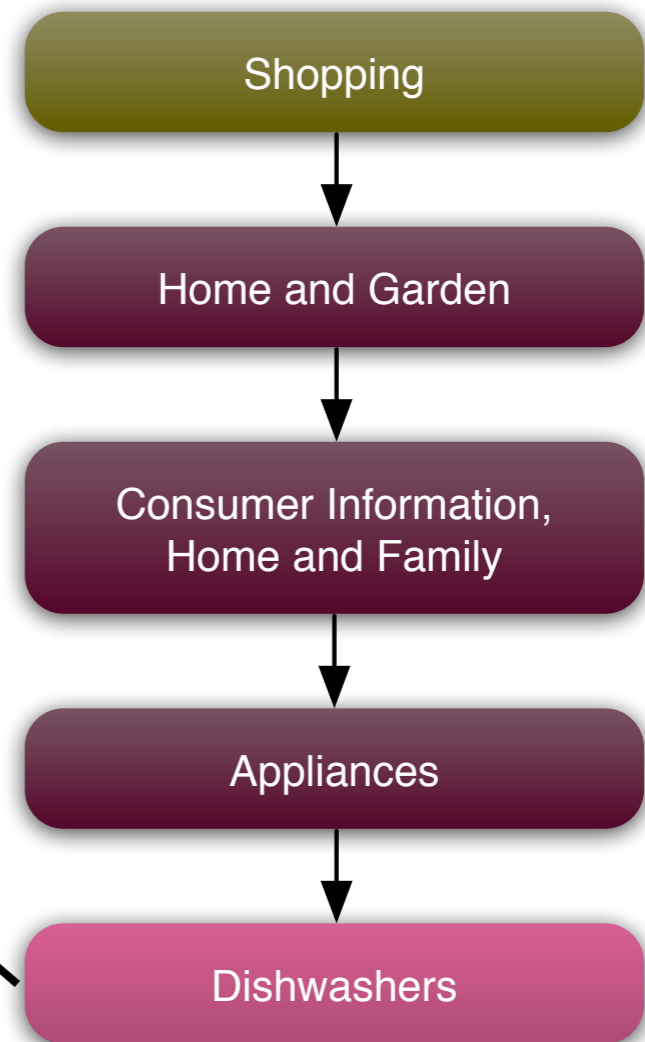
# Source category disambiguation



# Source category disambiguation

Context = {Appliances, Dishwashers}

{Dishwashers}





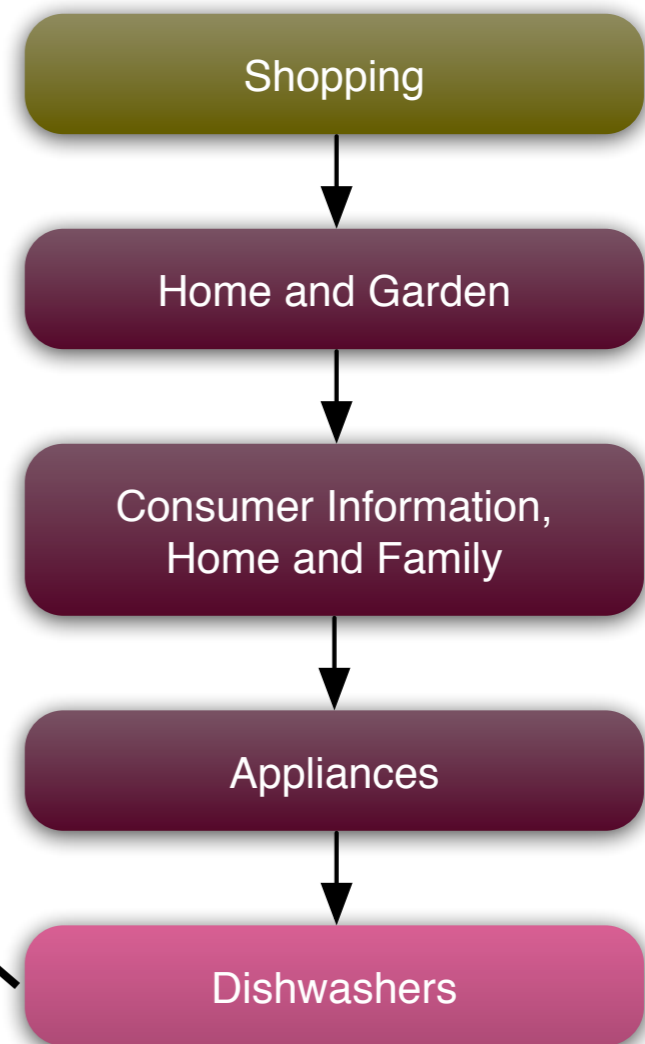
# Source category disambiguation

Context = {Appliances, Dishwashers}

{Dishwashers}

Extended Split Term Set = {*extendedTermSet*, ...}

Synonyms of 'Dishwashers'  
with the correct meaning



# Source category disambiguation

S1 = dishwasher, dish washer, dishwashing  
machine (a machine for washing dishes)

S2 = dishwasher (someone who washes dishes)

# Source category disambiguation

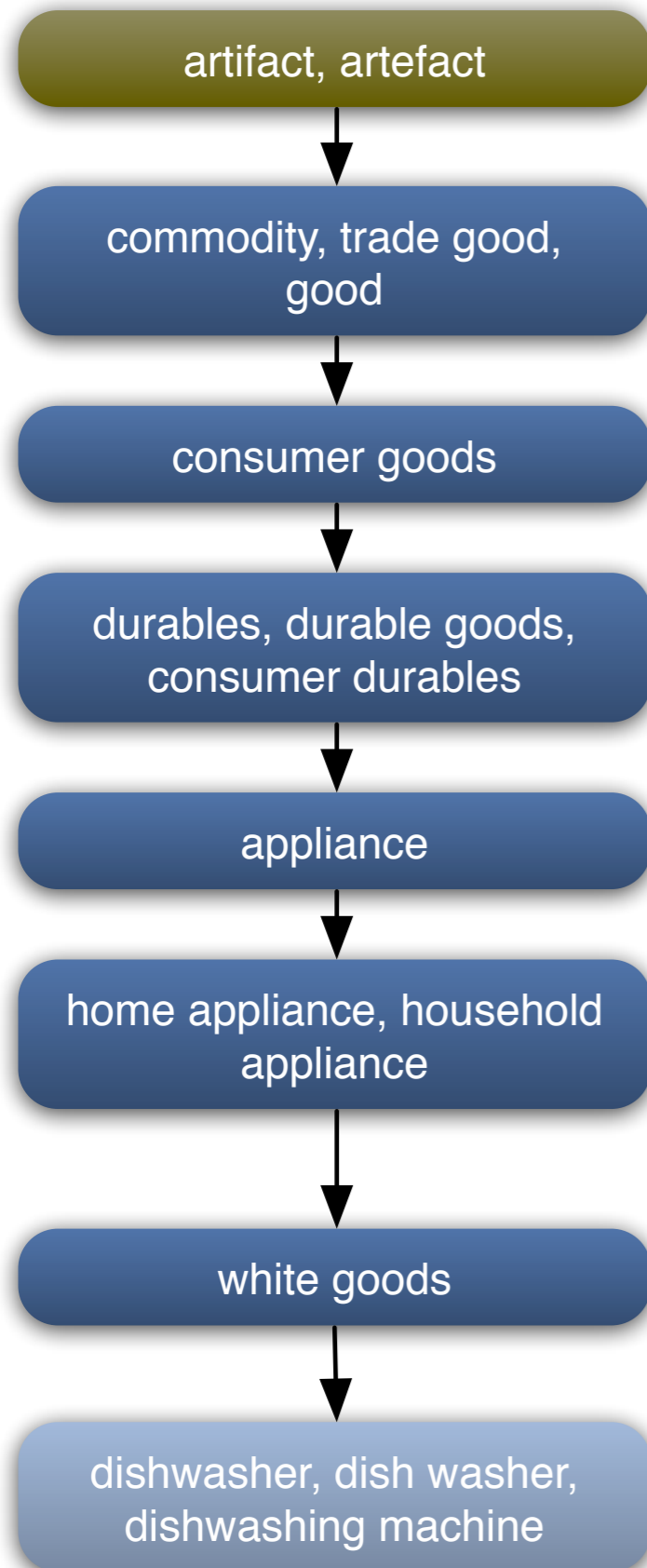
S1 = dishwasher, dish washer, dishwashing  
machine (a machine for washing dishes)

S2 = dishwasher (someone who washes dishes)

Compute sense score for each sense, highest is  
selected as correct sense

SI = dishwasher, ... (a machine for washing dishes)

# Related synsets based on hypernymy, hyponymy, meronymy and holonymy



SI = dishwasher, ... (a machine for washing dishes)

# Related synsets based on hypernymy, hyponymy, meronymy and holonymy

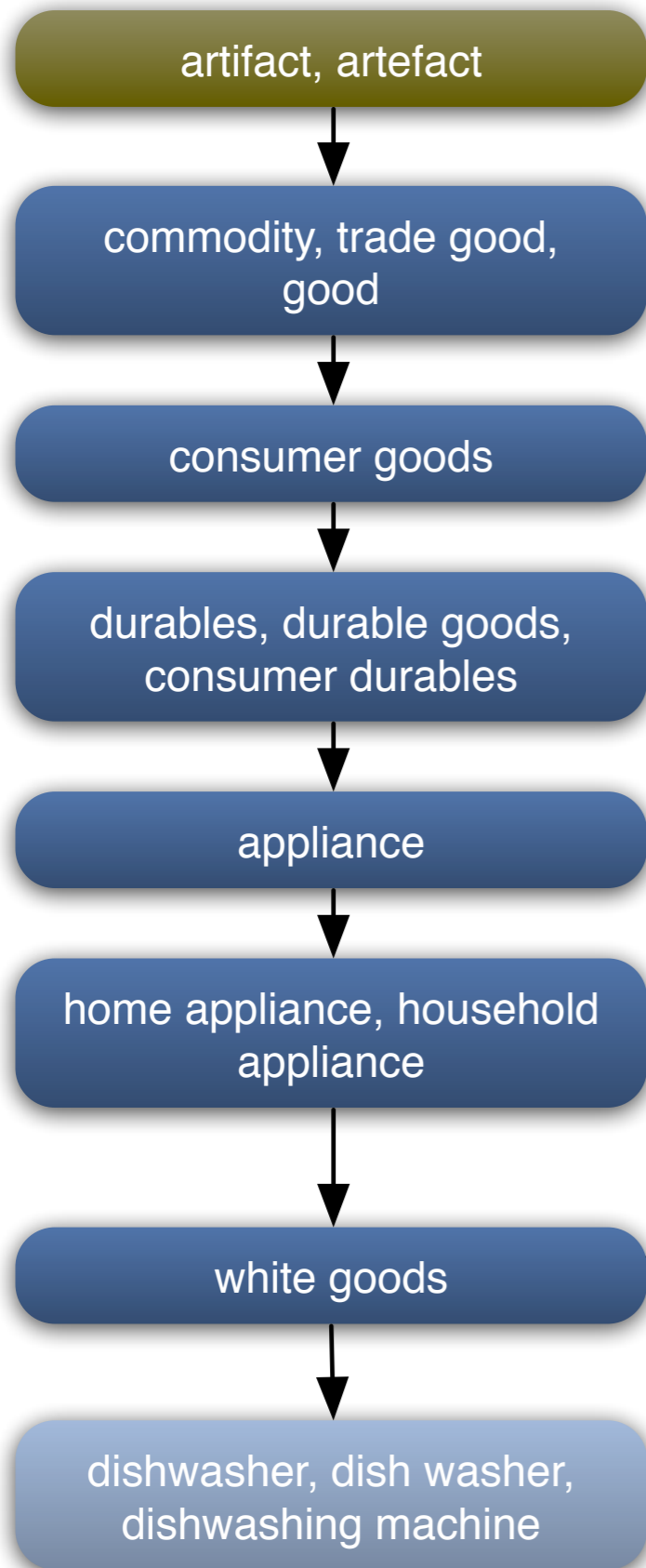
Context = {

Appliances,

...

Dishwashers

}



SI = dishwasher, ... (a machine for washing dishes)

# Related synsets based on hypernymy, hyponymy, meronymy and holonymy

Context = {

Appliances,

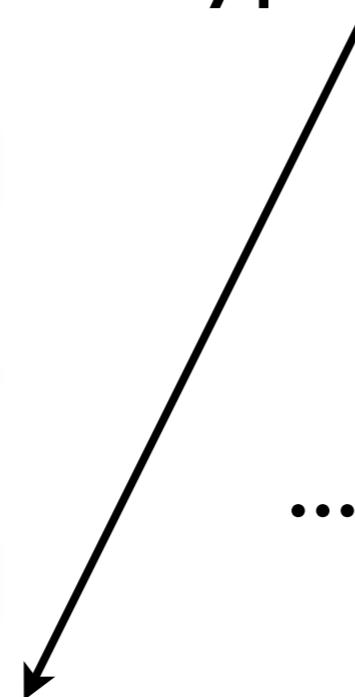
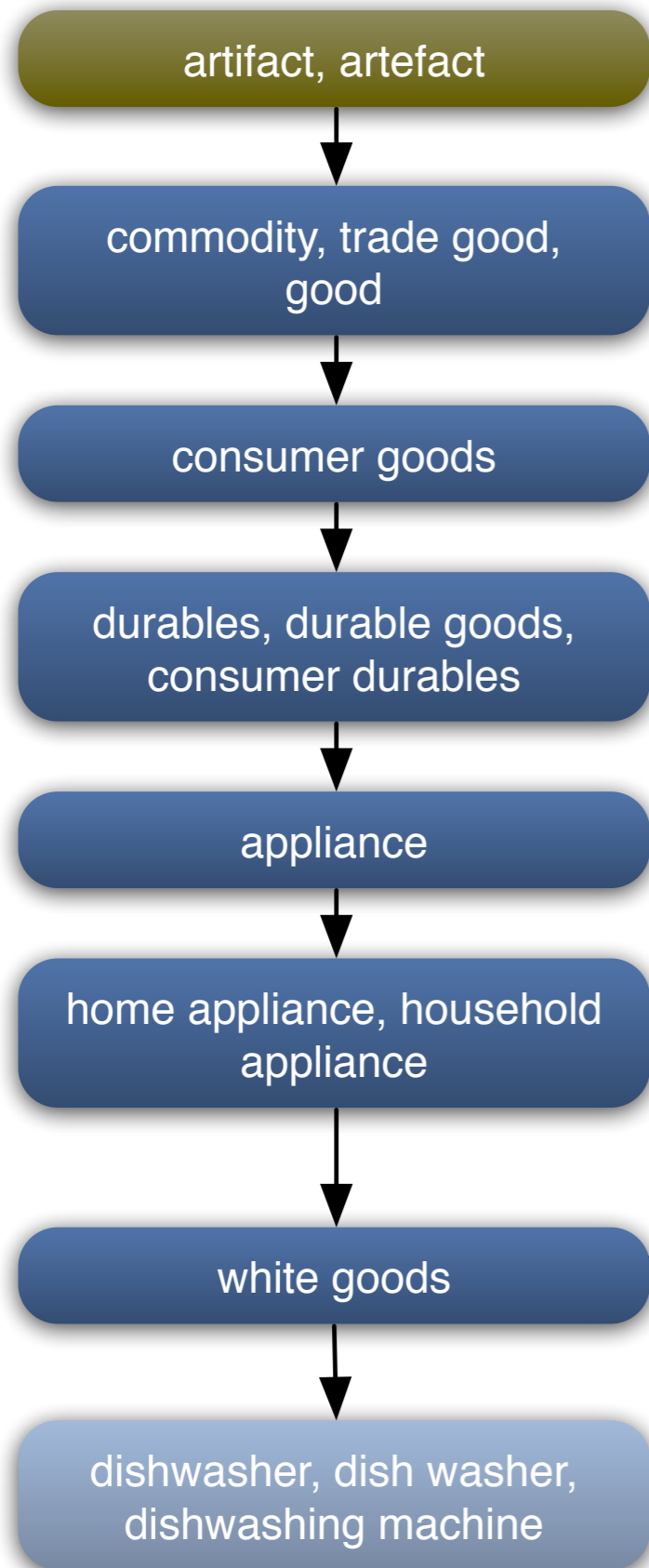
...

Dishwashers

}

**longest common substring** is used to compute the score

SI = dishwasher, ... (a machine for washing dishes)



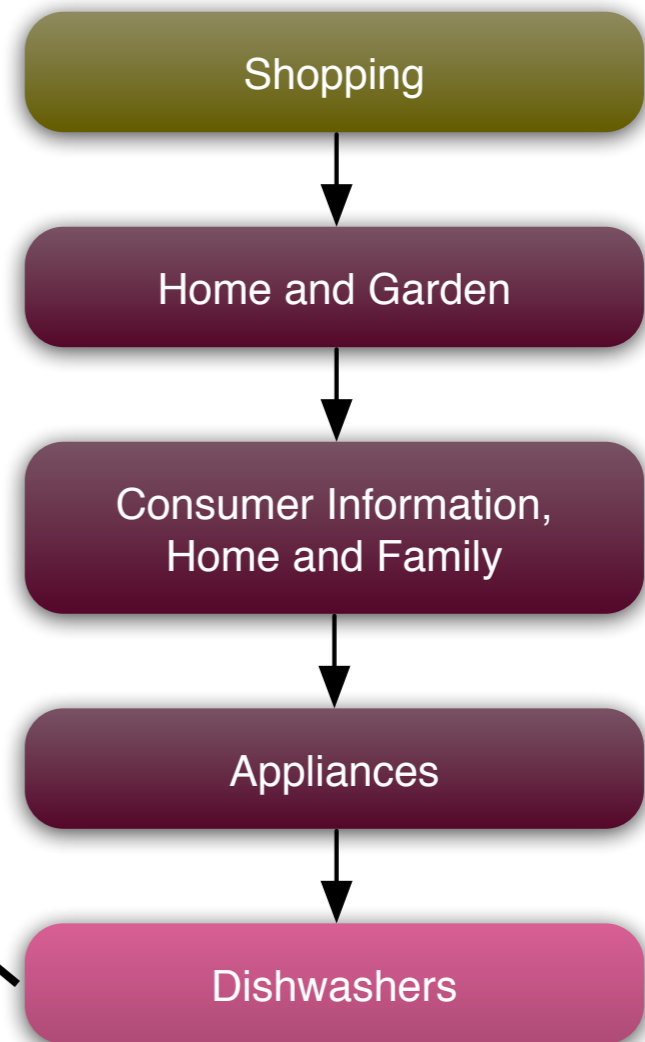
# Source category disambiguation

Context = {Appliances, Dishwashers}

{Dishwashers}

Extended Split Term Set = {*extendedTermSet*, ...}

{Dishwashers, dishwasher, dish  
washer, dishwashing machine}

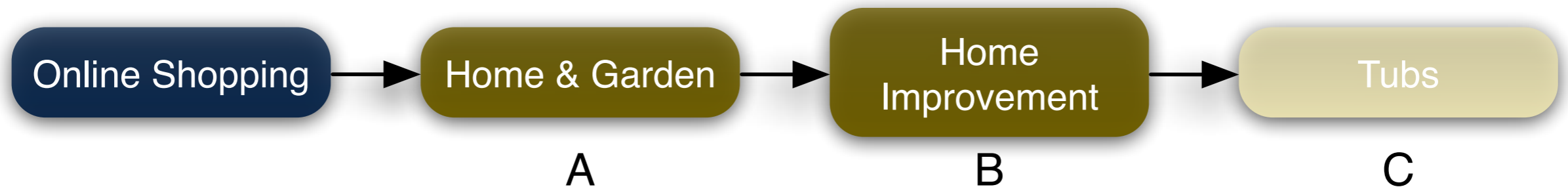




# Candidate target category selection

- Algorithm 'Semantic Search'
- Input:
  - a source category name with the disambiguation results
  - a target category name
- Output: true if source category matches and is a subset of target category

# Candidate target category selection



Disambiguation result for 'Tubs':  
{{Tubs, bathtub, bathing tub, bath, tub}}

# Candidate target category selection

Disambiguation result for 'Tubs':  
{{Tubs, bathtub, bathing tub, bath, tub}}

# Candidate target category selection

Disambiguation result for 'Tubs':  
{Tubs, bathtub, bathing tub, bath, tub}

Target path: Kitchen & Bath Fixtures

# Candidate target category selection

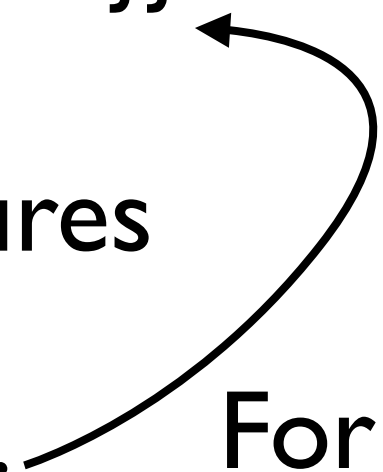
Disambiguation result for 'Tubs':  
{Tubs, bathtub, bathing tub, bath, tub}

Target path: Kitchen & Bath Fixtures

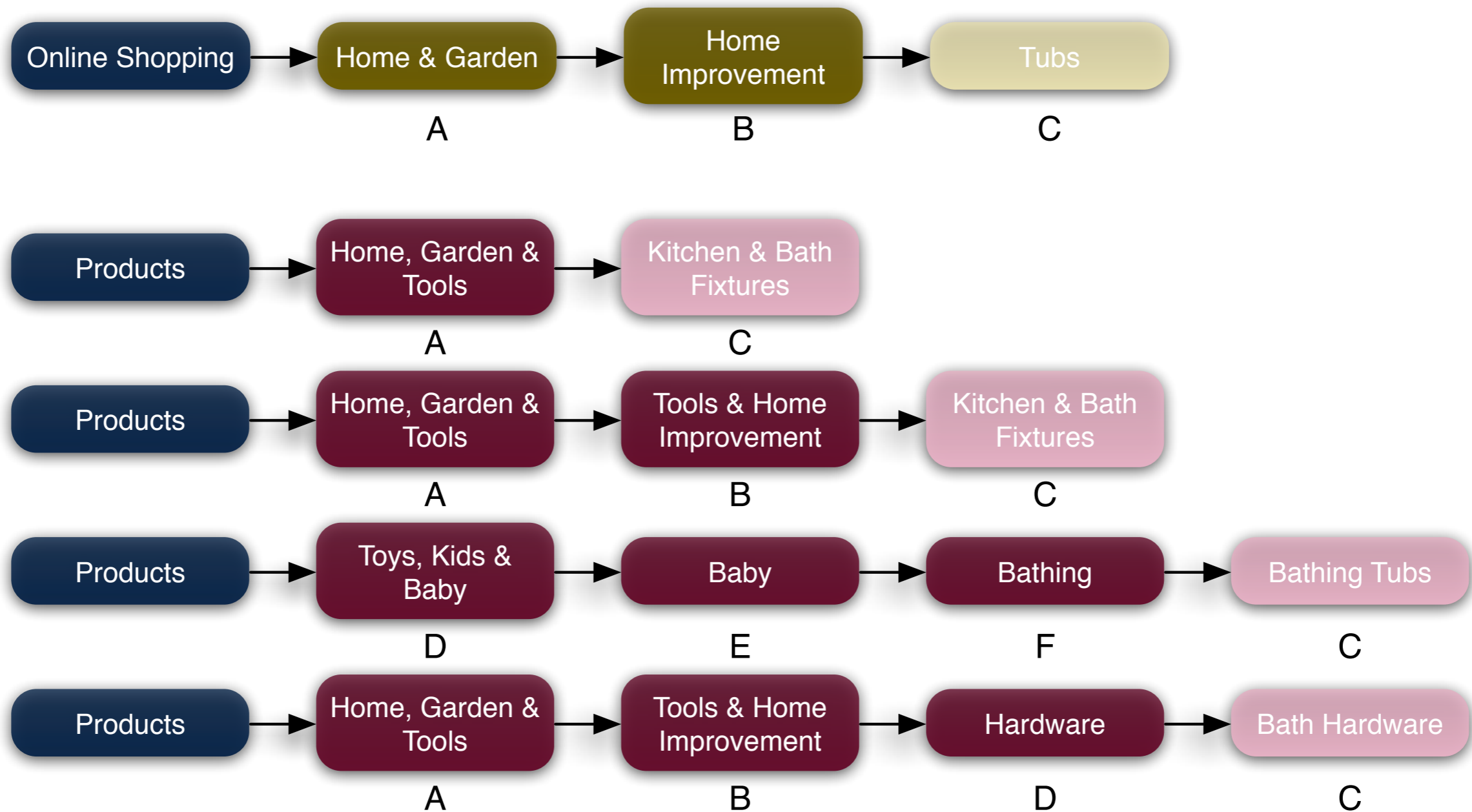
Match for at least one split term:

- source term is part of target category as separate term, or
- normalized Levenshtein similarity is above a certain threshold

For each extended term set



# Candidate target category selection



# SCHEMA overview

1. source category disambiguation
2. candidate target category selection
- 3. candidate target path key comparison**

# Candidate target path key comparison

- Damerau-Levenshtein applied on paths
- Category paths are converted to list of generated ID's
- Equal nodes get the same ID
- Equality determined by 'Semantic Search' algorithm (candidate target selection)



# Candidate target path key comparison

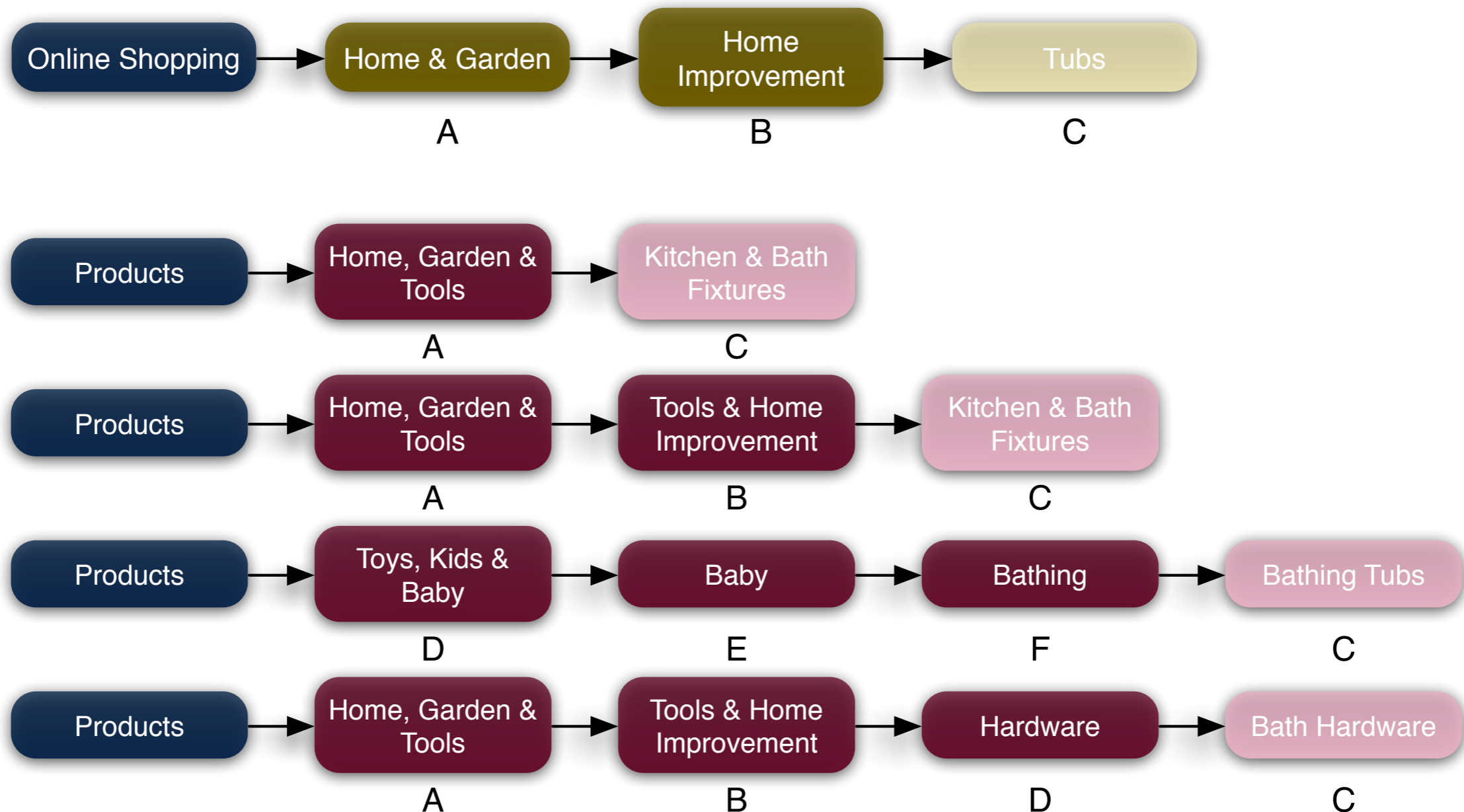
Final score:

$$score(K_{src}, K_{cand}) = 1 - \frac{damLev(K_{src}, K_{cand}) + p}{\max(\text{len}(K_{src}), \text{len}(K_{cand})) + p}$$

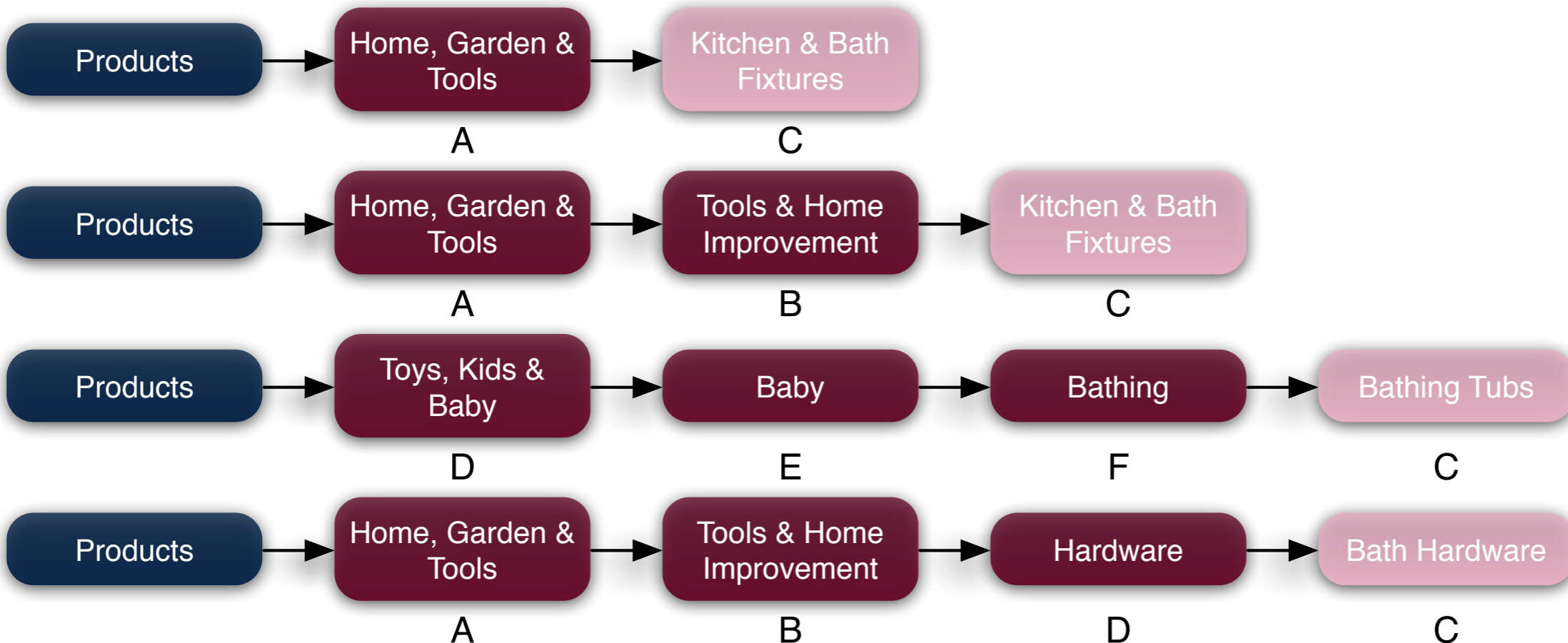
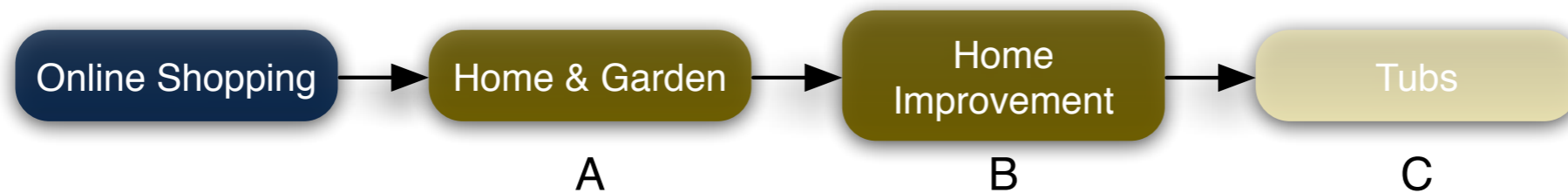
where:

- $K$  is a key list
- $p$  is the penalty (# absent nodes in candidate path)
- $damLev()$  computes the Damerau-Levenshtein distance between two key lists

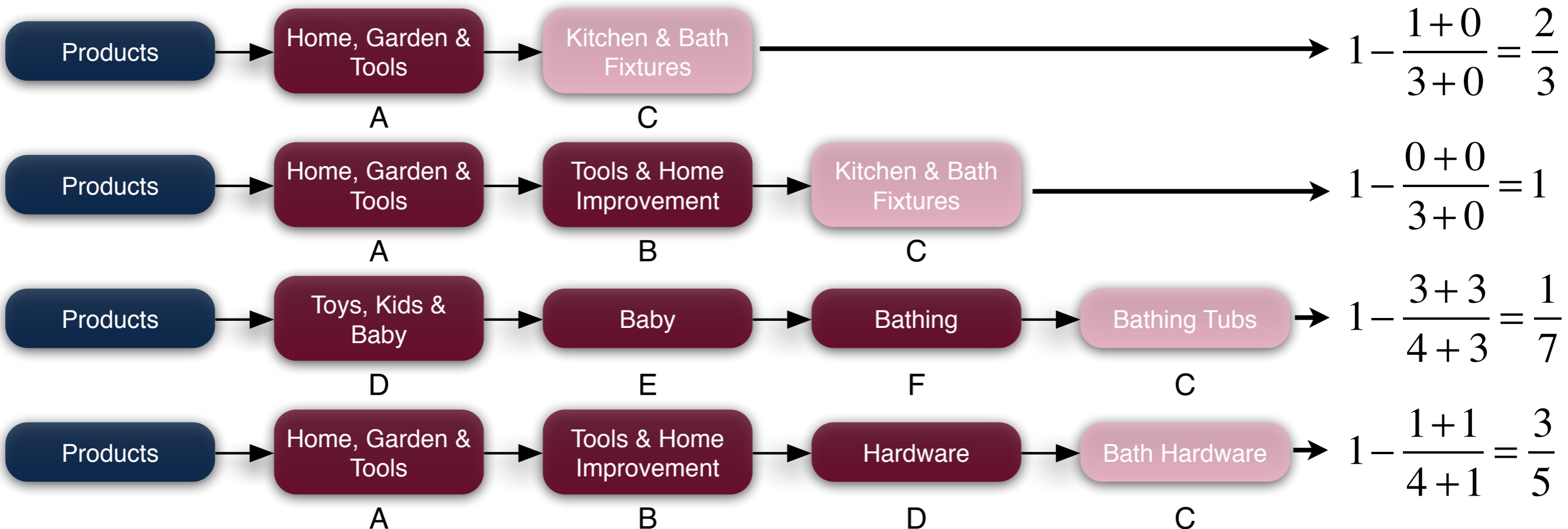
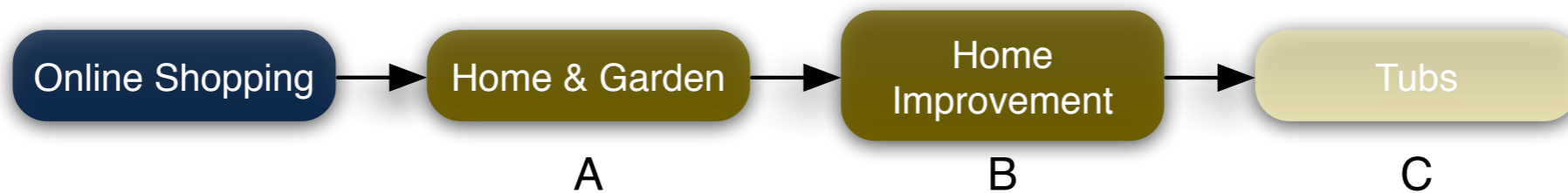
# Candidate target category selection



# Candidate target category selection



# Candidate target category selection



# Evaluation

# Evaluation

- Datasets
  - Amazon.com, ~2,500 categories
  - Overstock.com, ~1,000 categories
  - Dmoz.org, ~44,000 categories

# Evaluation

- Datasets
  - Amazon.com, ~2,500 categories
  - Overstock.com, ~1,000 categories
  - Dmoz.org, ~44,000 categories
- Manual mapping of 3000 categories
  - the 6 data set combinations, sample size 500
  - 3 individuals performed the evaluation

# Evaluation

## Overall results

Algorithm	Precision	Recall	F <sub>1</sub>	# Senses found	WSD accuracy
PROMPT	28.93%	16.69%	20.75%	n/a	n/a
Park & Kim	47.77%	25.19%	32.52%	5.70%	83.72%
SCHEMA	42.21%	80.73%	55.10%	82.03%	84.01%



# Questions?