

An Automated Approach to Product Taxonomy Mapping in E-commerce

Lennart Nederstigt, Damir Vandić, Flavius Frasinca



Terminology

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

Product taxonomies

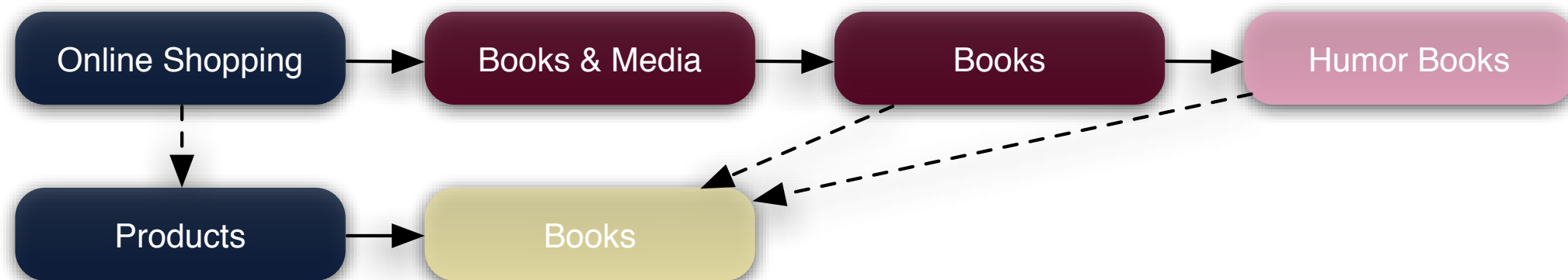
Important aspects of product taxonomies:

- 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



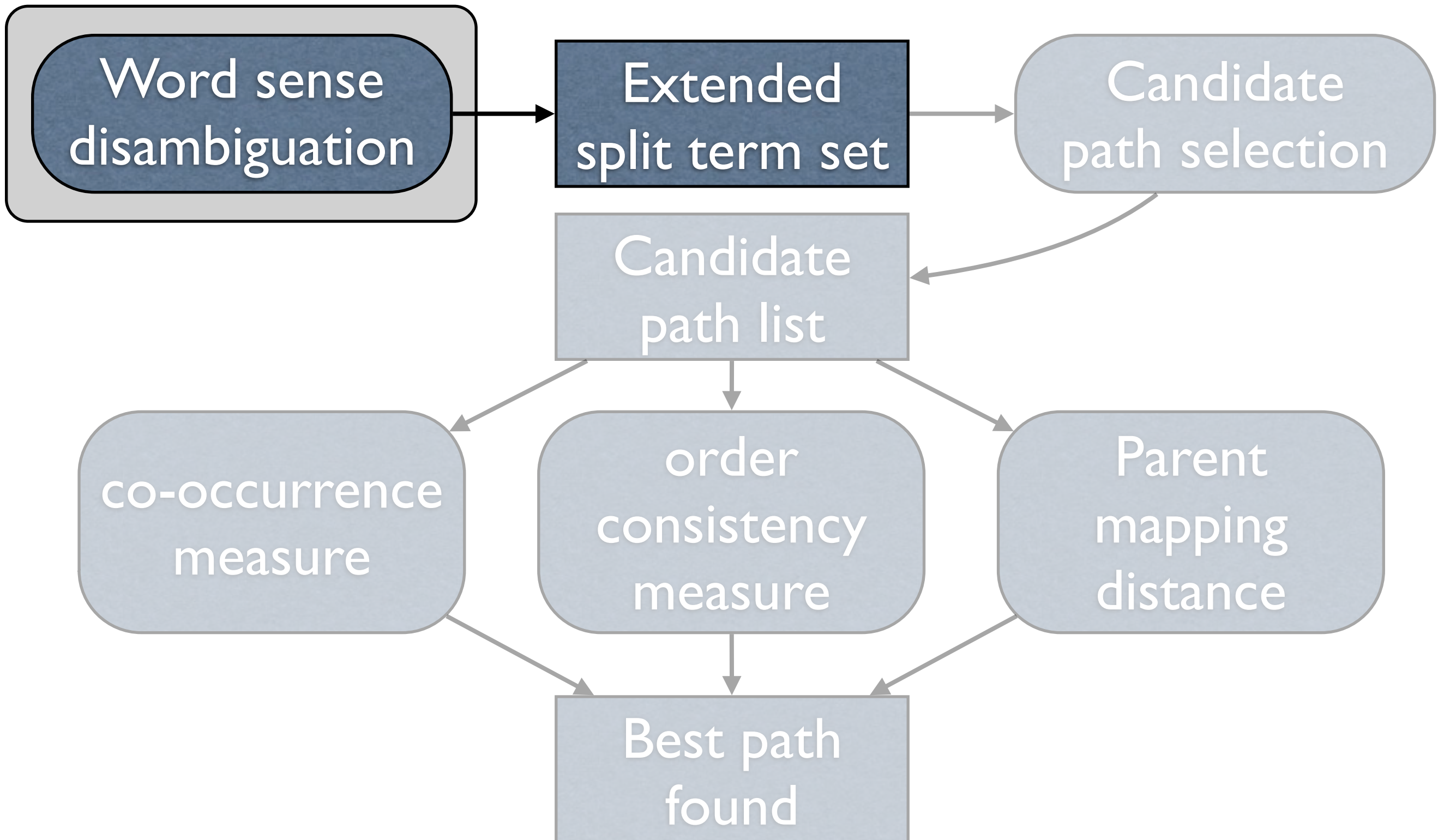
Related work

- 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”

Algorithm overview

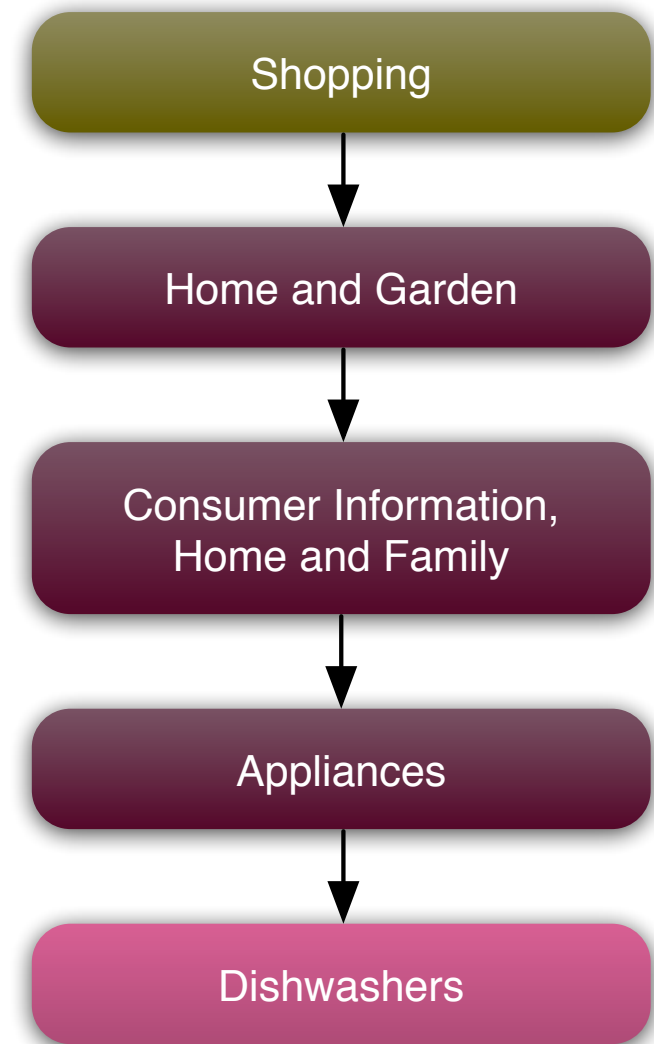
- Input is a source category path
- Output is a target category path (or 'None')
- Three main steps
 1. word sense disambiguation
 2. candidate path search
 3. best path selection

Part I



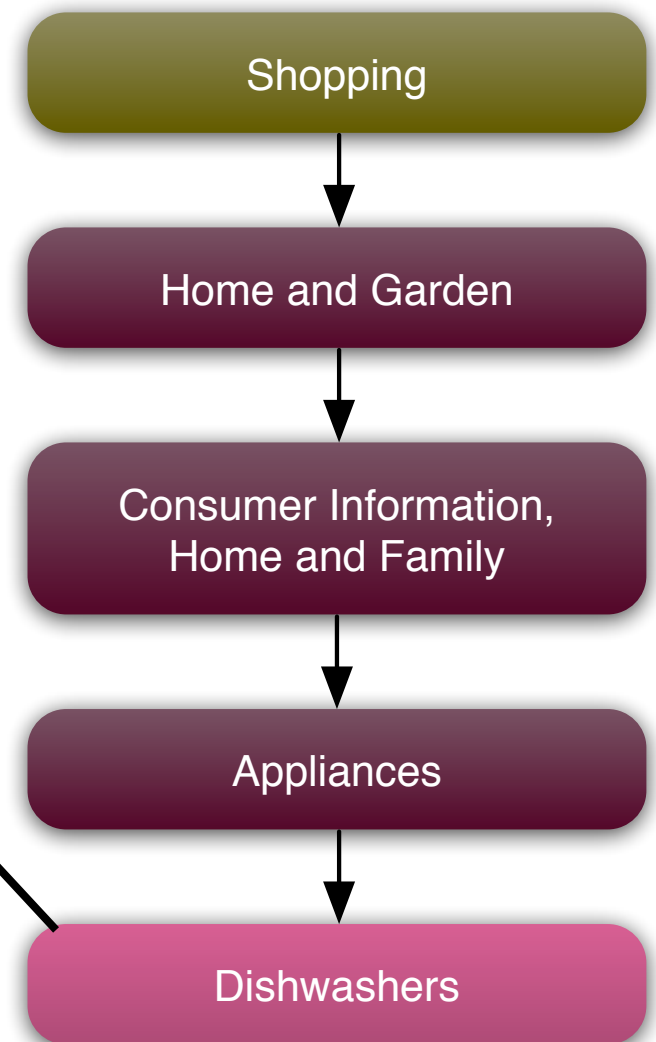
Word sense disambiguation

- Example category path
 - Dishwashers can have two meanings
 - From the path, the meaning is clear to humans
- Word sense disambiguation for source category



Word sense disambiguation

Terms of source node = {Dishwashers}



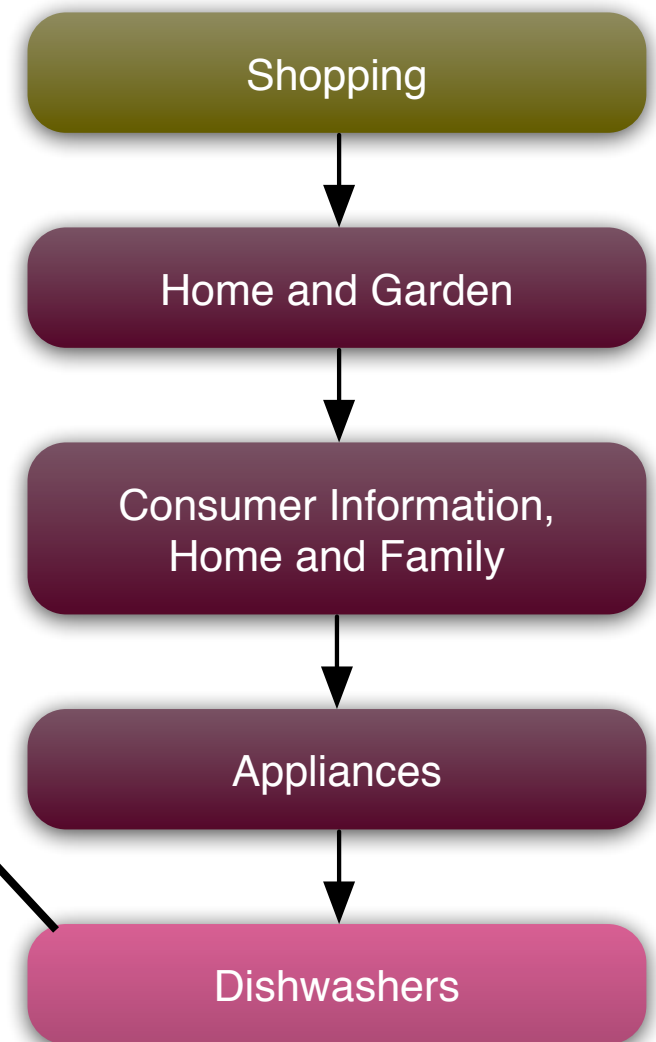
Word sense disambiguation

Terms of source node = {Dishwashers}

for each

{*splitTermSet*, ...}

Synonyms of a split term
that have the correct
meaning in this context



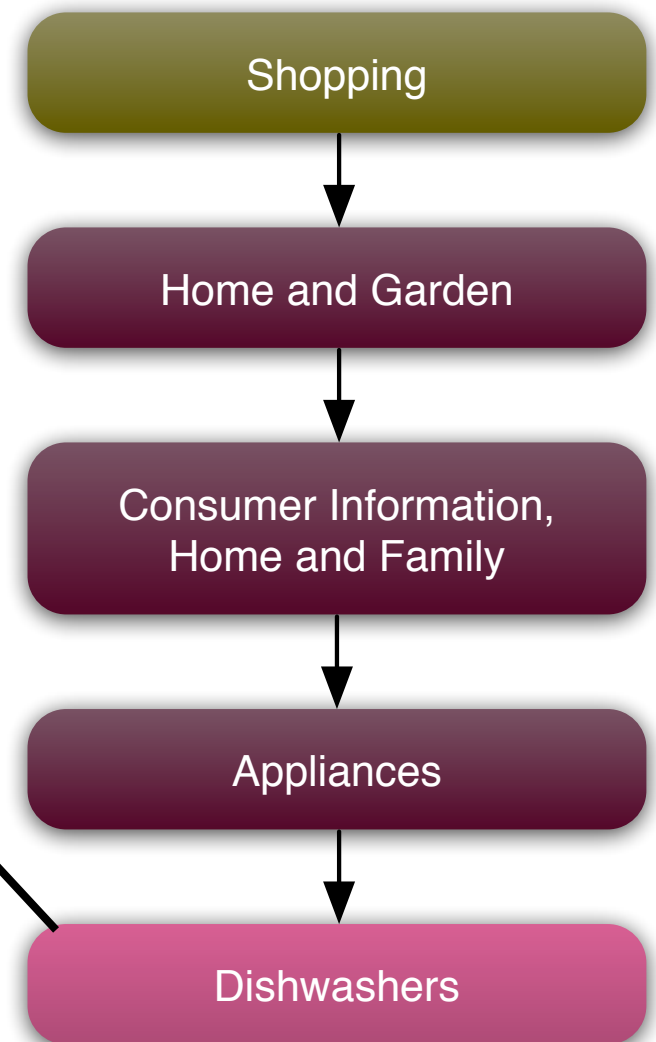
Word sense disambiguation

Terms of source node = {Dishwashers}

for each

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

Synonyms of a split term
that have the correct
meaning in this context



Shopping



Home and Garden



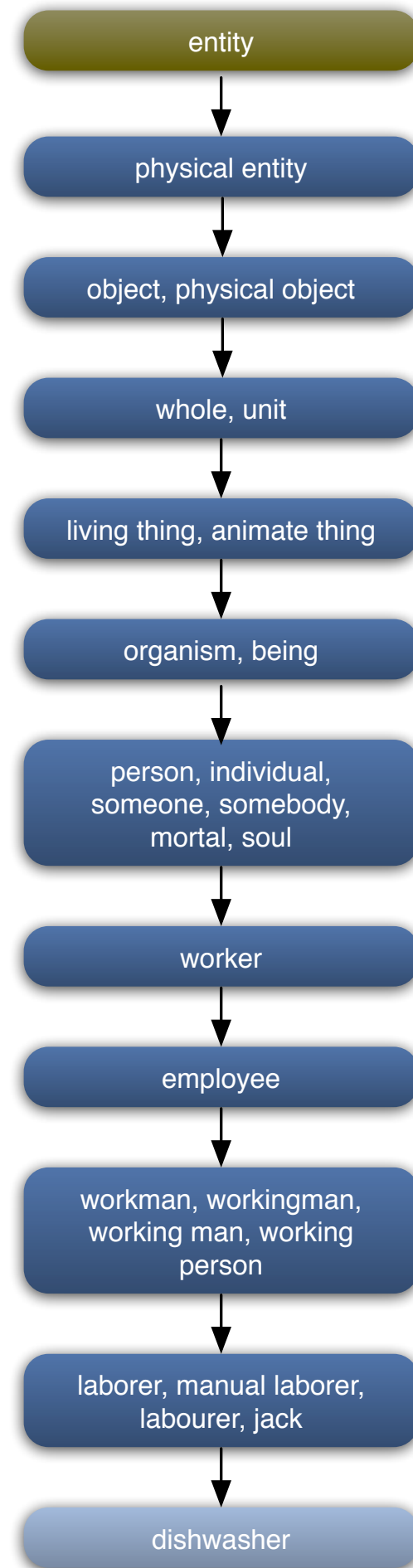
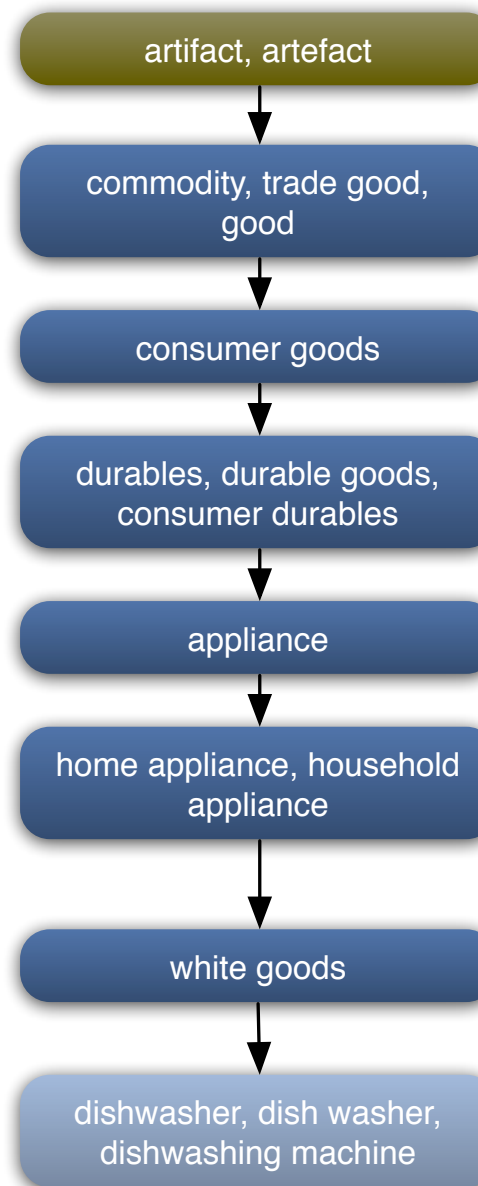
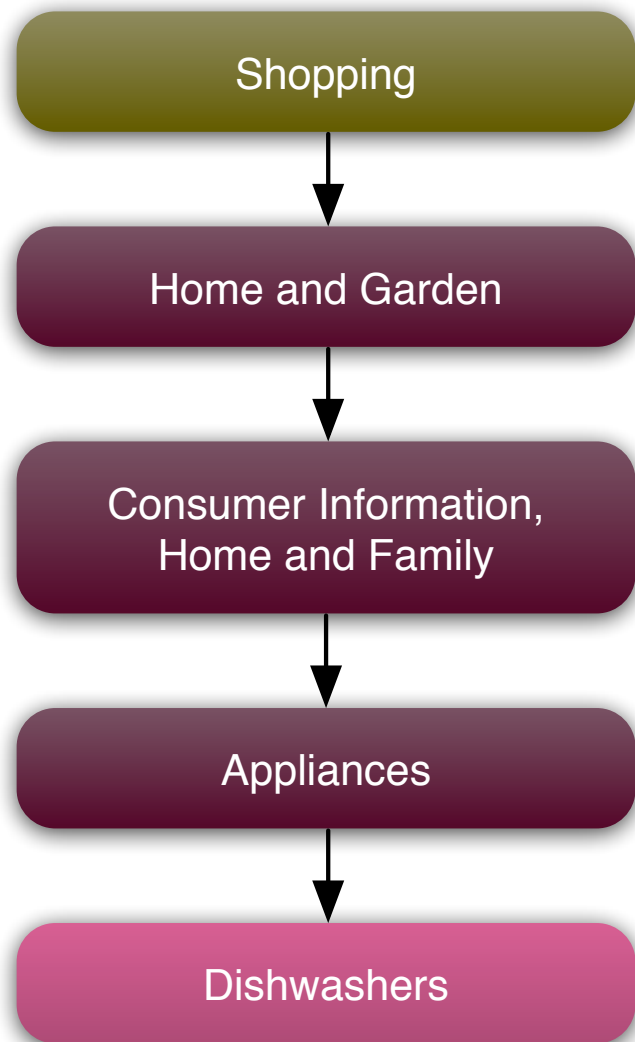
Consumer Information,
Home and Family

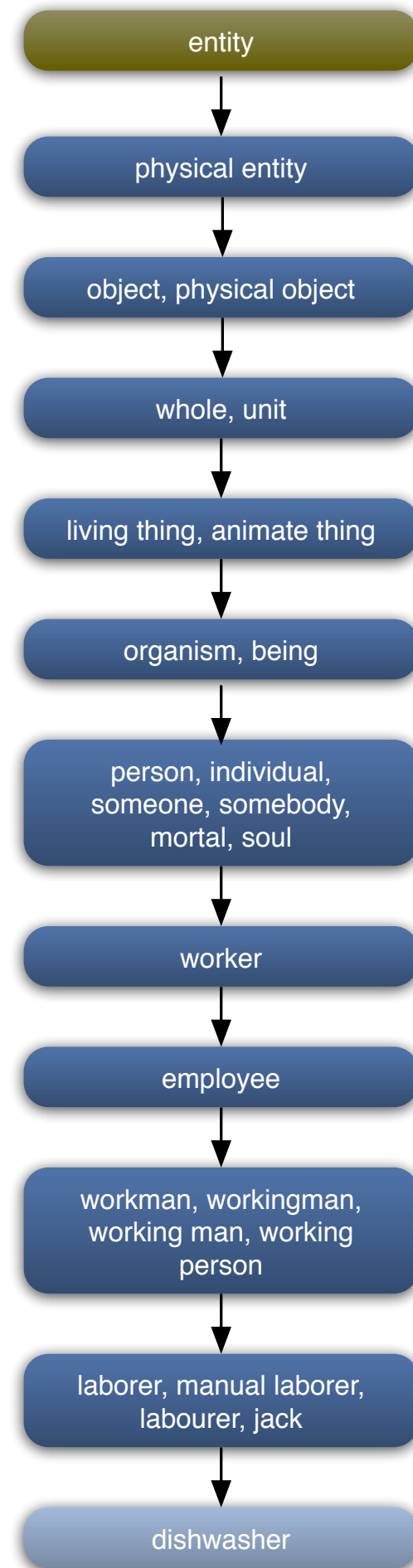
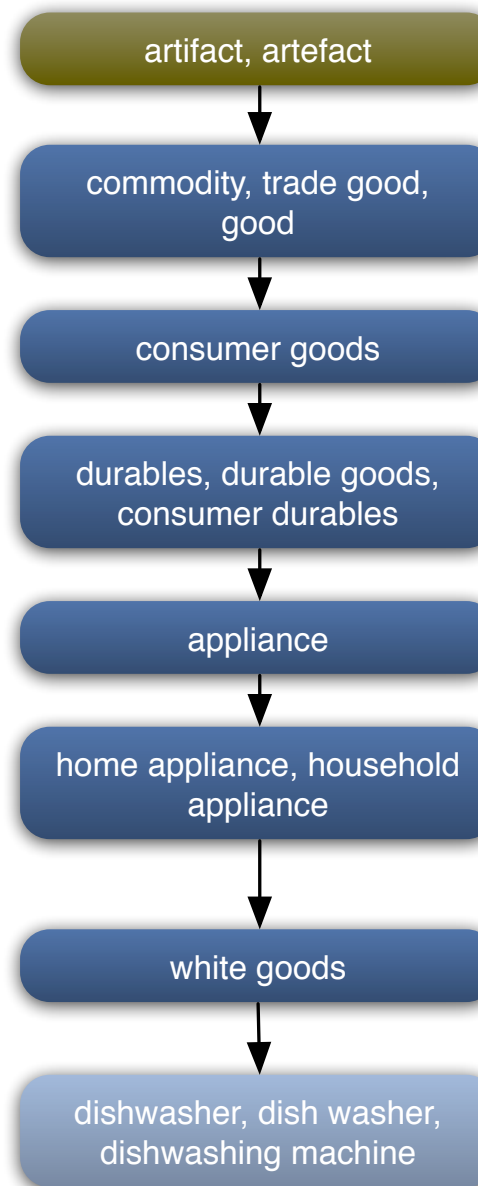
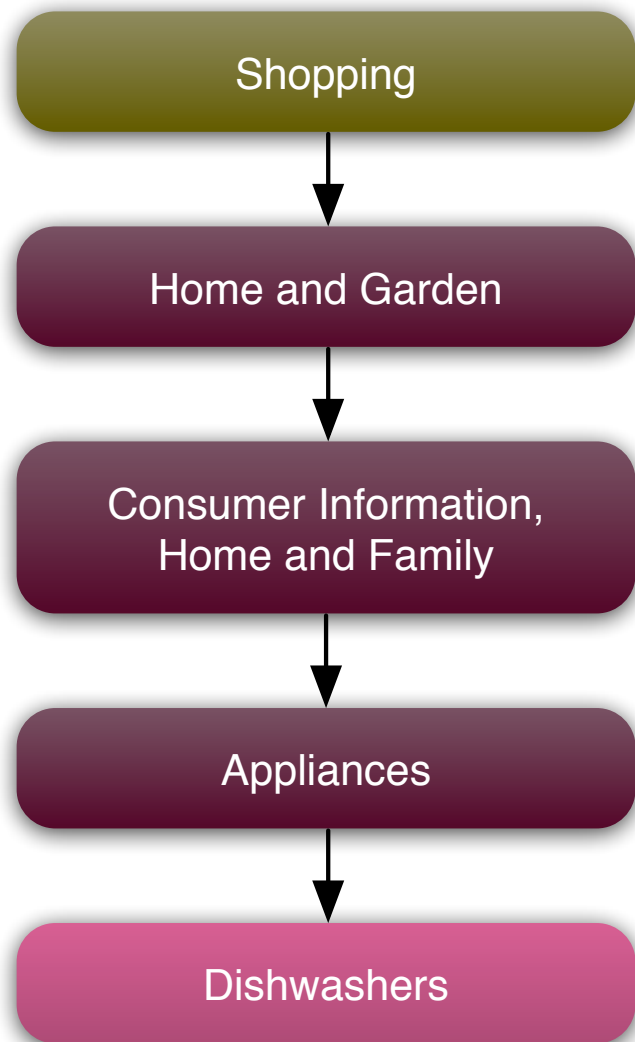


Appliances



Dishwashers





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

artifact, artefact



commodity, trade good,
good



consumer goods



durables, durable goods,
consumer durables



appliance



home appliance, household
appliance



white goods



dishwasher, dish washer,
dishwashing machine

← **synset I**

Shopping



Home and Garden



Consumer Information,
Home and Family



Appliances



Dishwashers

artifact, artefact



commodity, trade good,
good



consumer goods



durables, durable goods,
consumer durables



appliance



home appliance, household
appliance



white goods



dishwasher, dish washer,
dishwashing machine

Shopping



Home and Garden



Consumer Information,
Home and Family

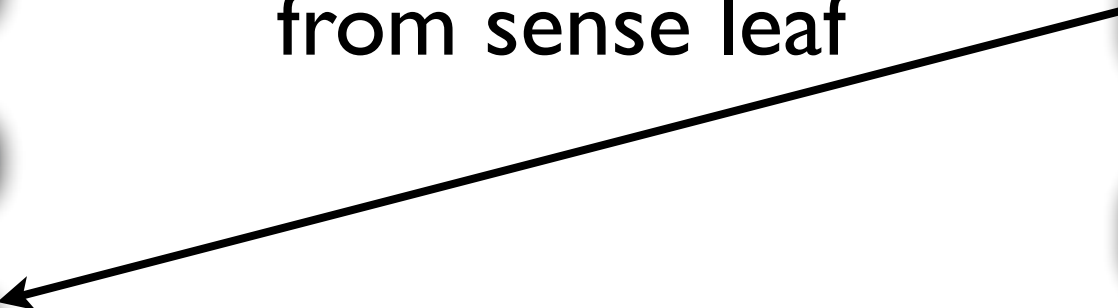


Appliances



Dishwashers

First match
from sense leaf



synset I



artifact, artefact

commodity, trade good,
good

consumer goods

durables, durable goods,
consumer durables

appliance

home appliance, household
appliance

white goods

dishwasher, dish washer,
dishwashing machine

Shopping

Home and Garden

Consumer Information,
Home and Family

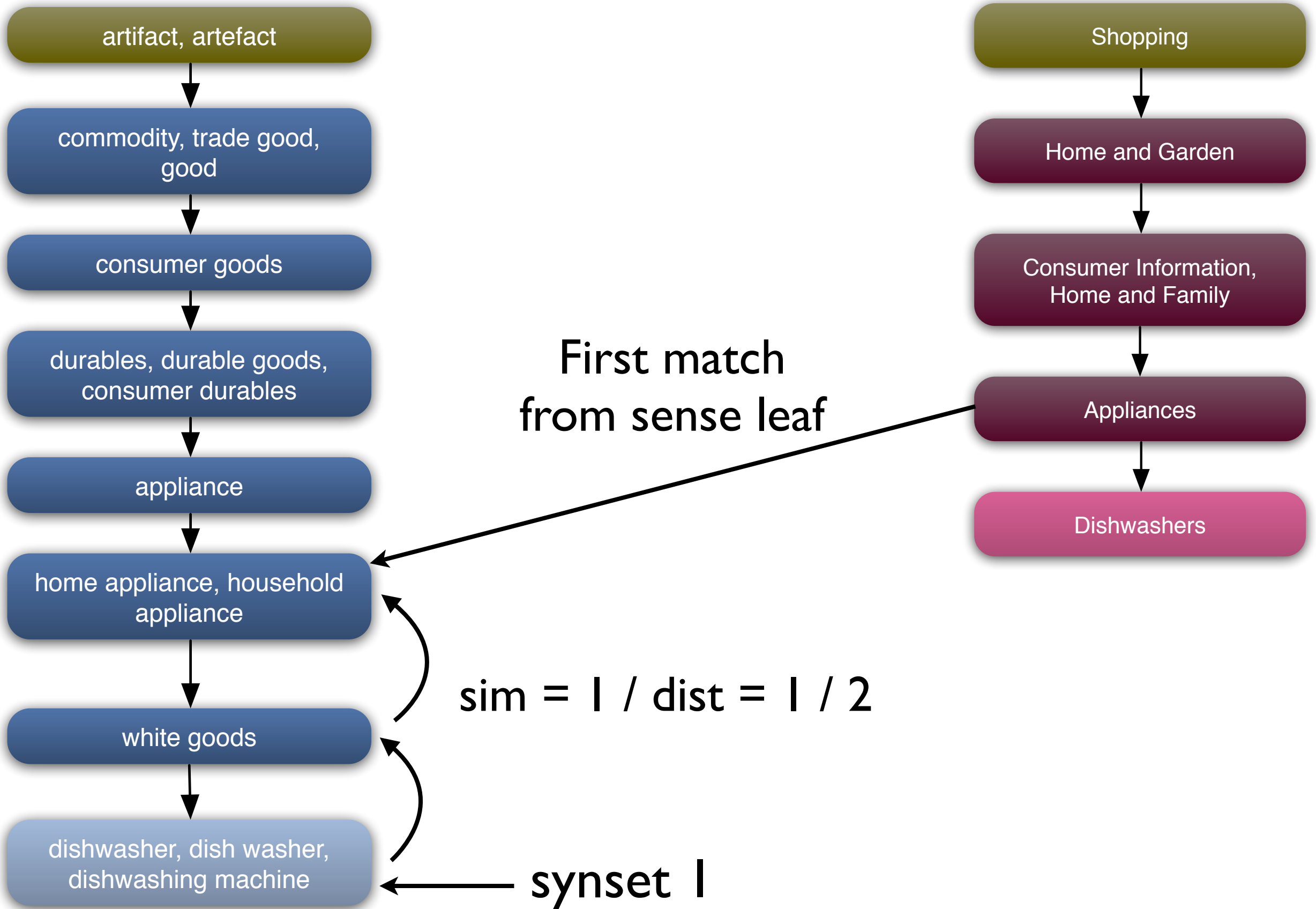
Appliances

Dishwashers

First match
from sense leaf

$$\text{sim} = 1 / \text{dist} = 1 / 2$$

synset 1



artifact, artefact

commodity, trade good, good

consumer goods

durables, durable goods, consumer durables

appliance

home appliance, household appliance

white goods

dishwasher, dish washer, dishwashing machine

Shopping

Home and Garden

Consumer Information, Home and Family

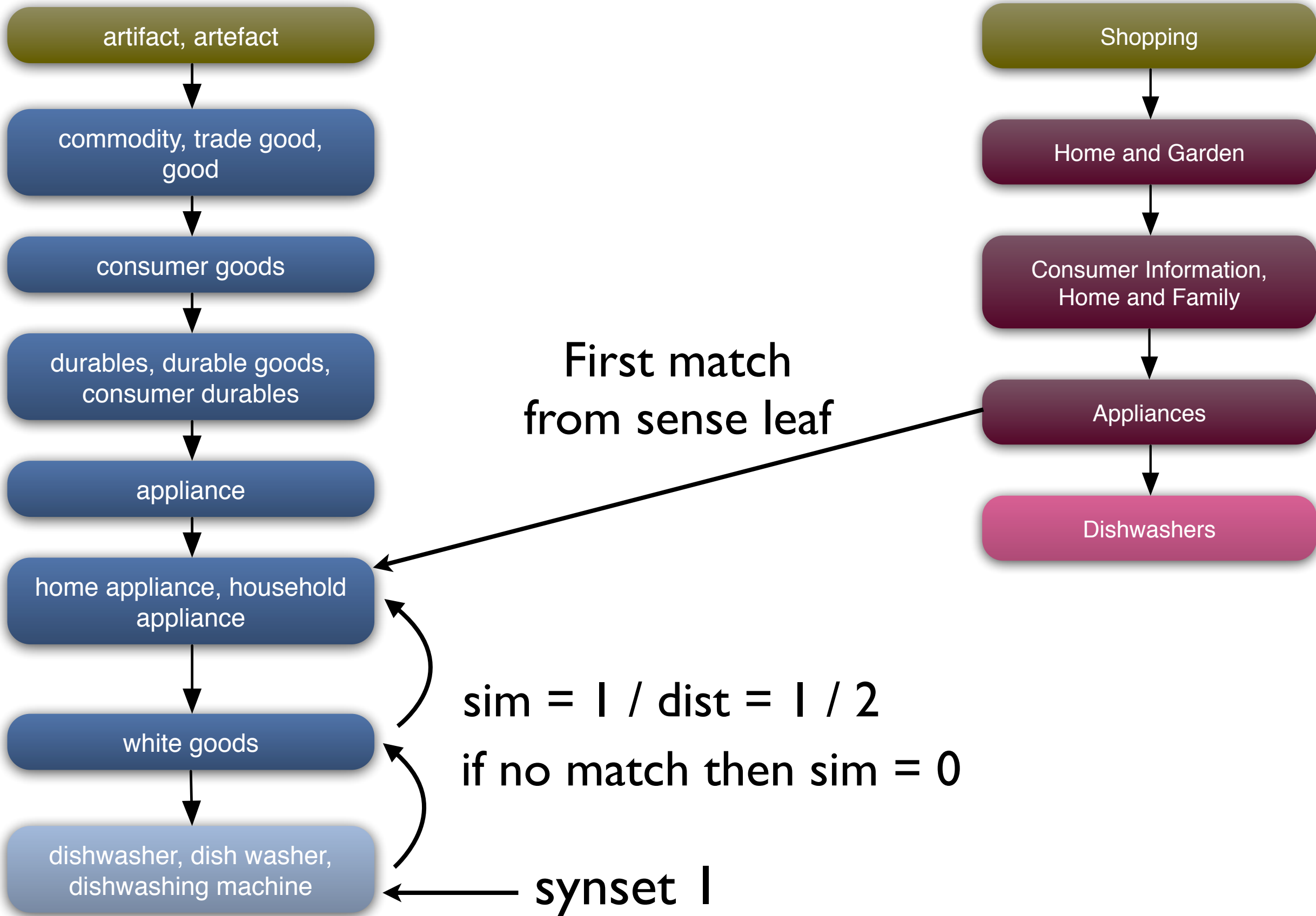
Appliances

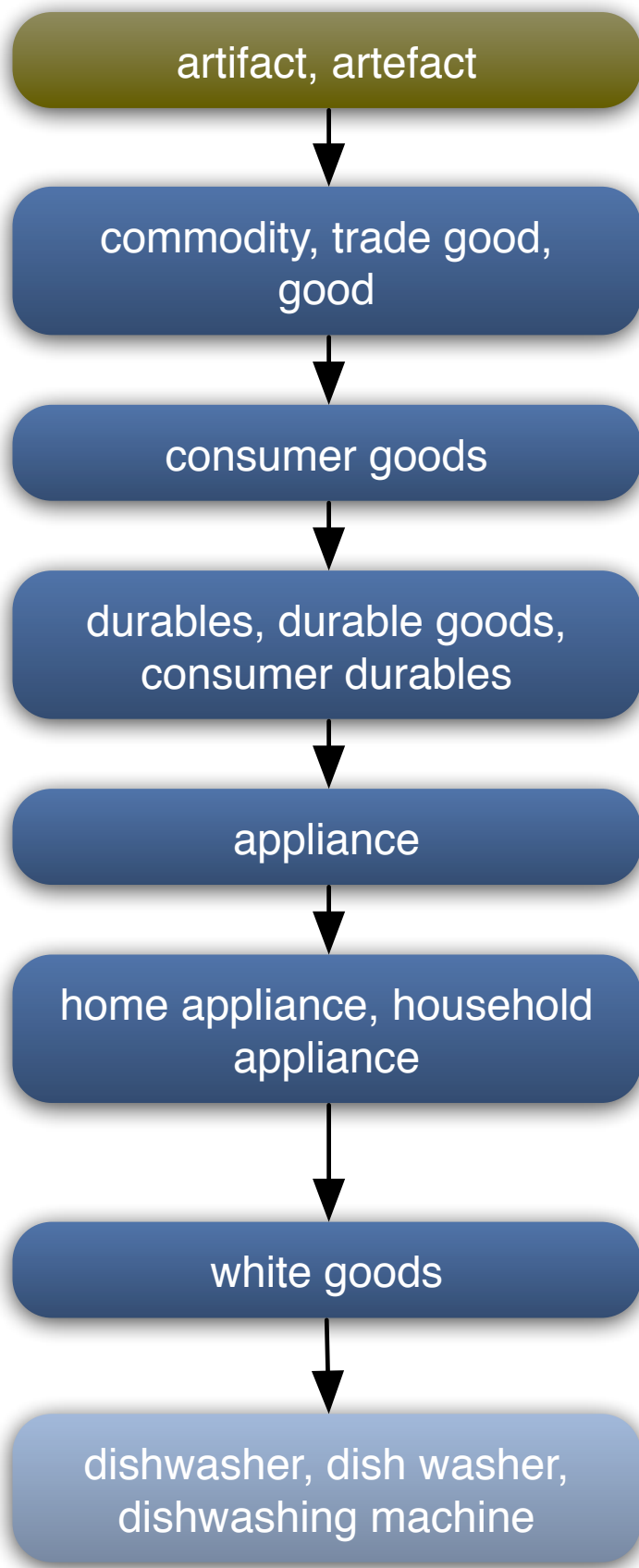
Dishwashers

First match from sense leaf

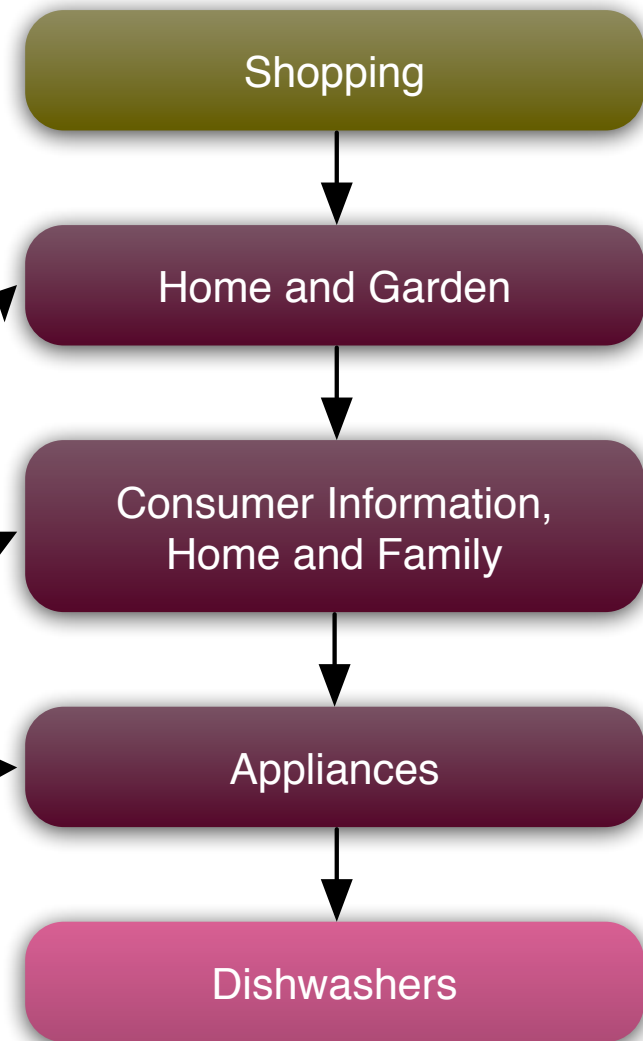
$sim = 1 / dist = 1 / 2$
if no match then $sim = 0$

synset 1

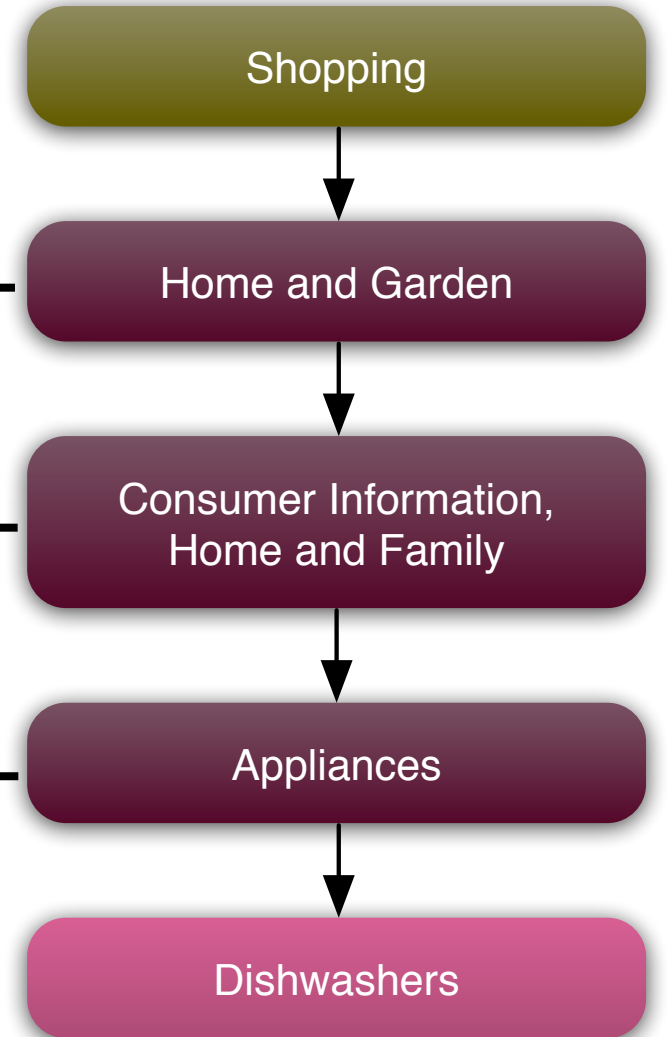
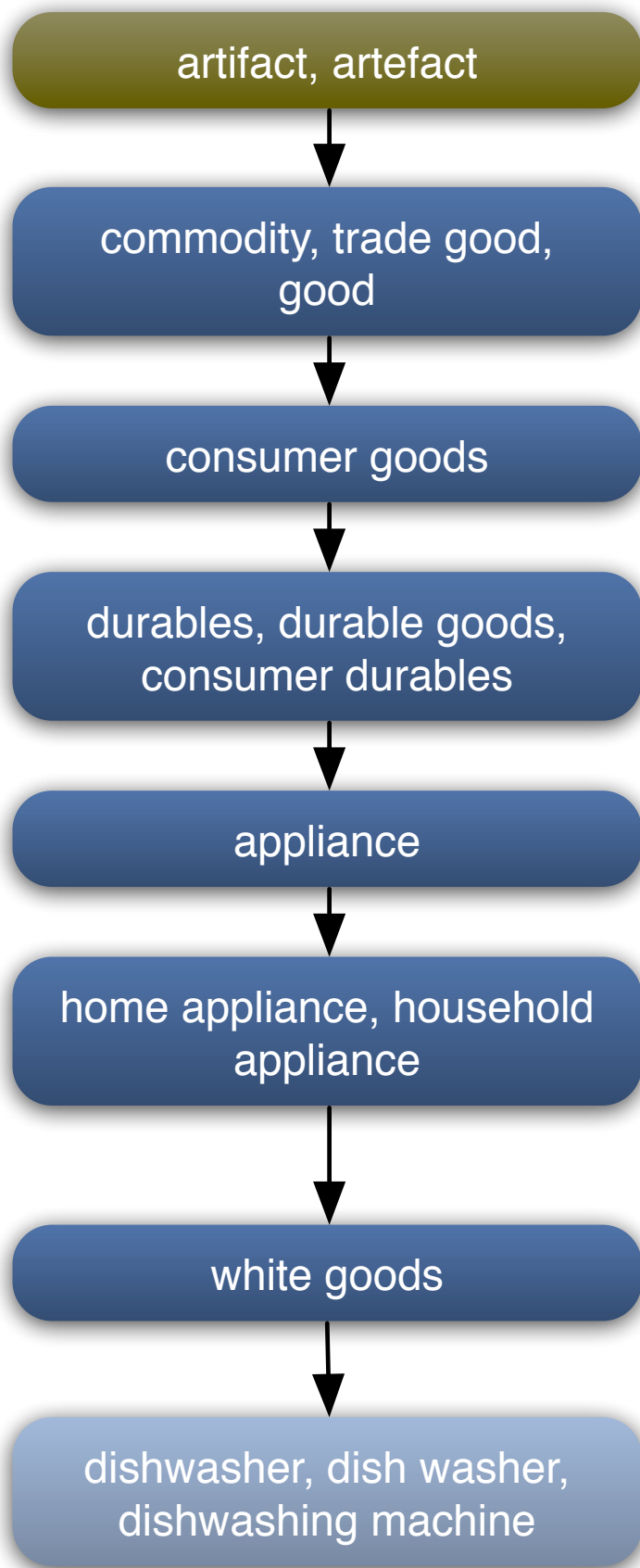




This process is applied for each parent



← synset I

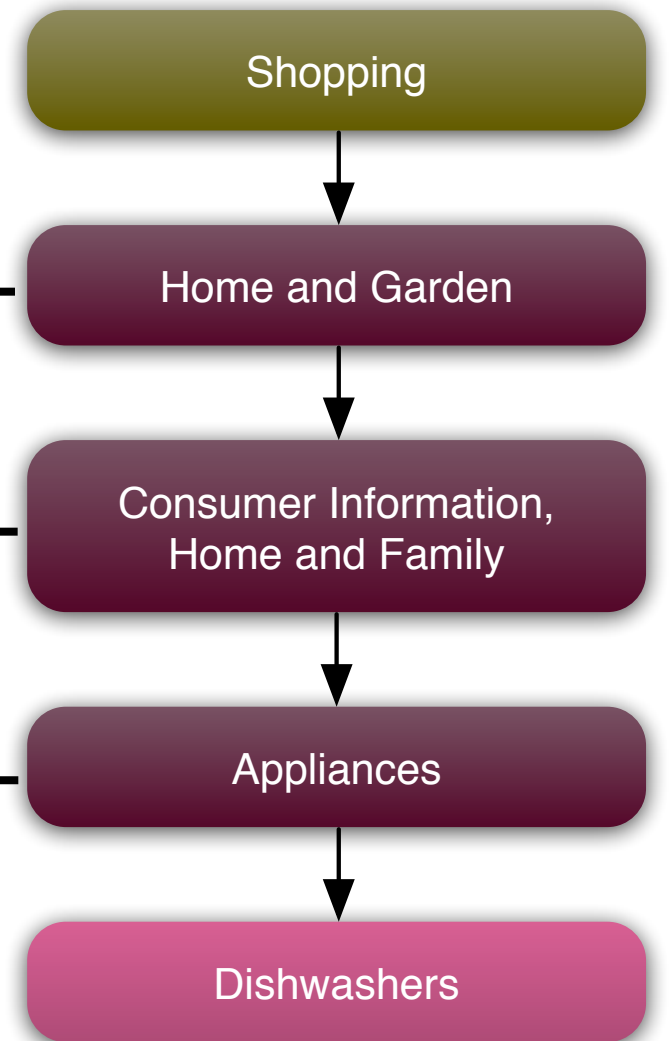
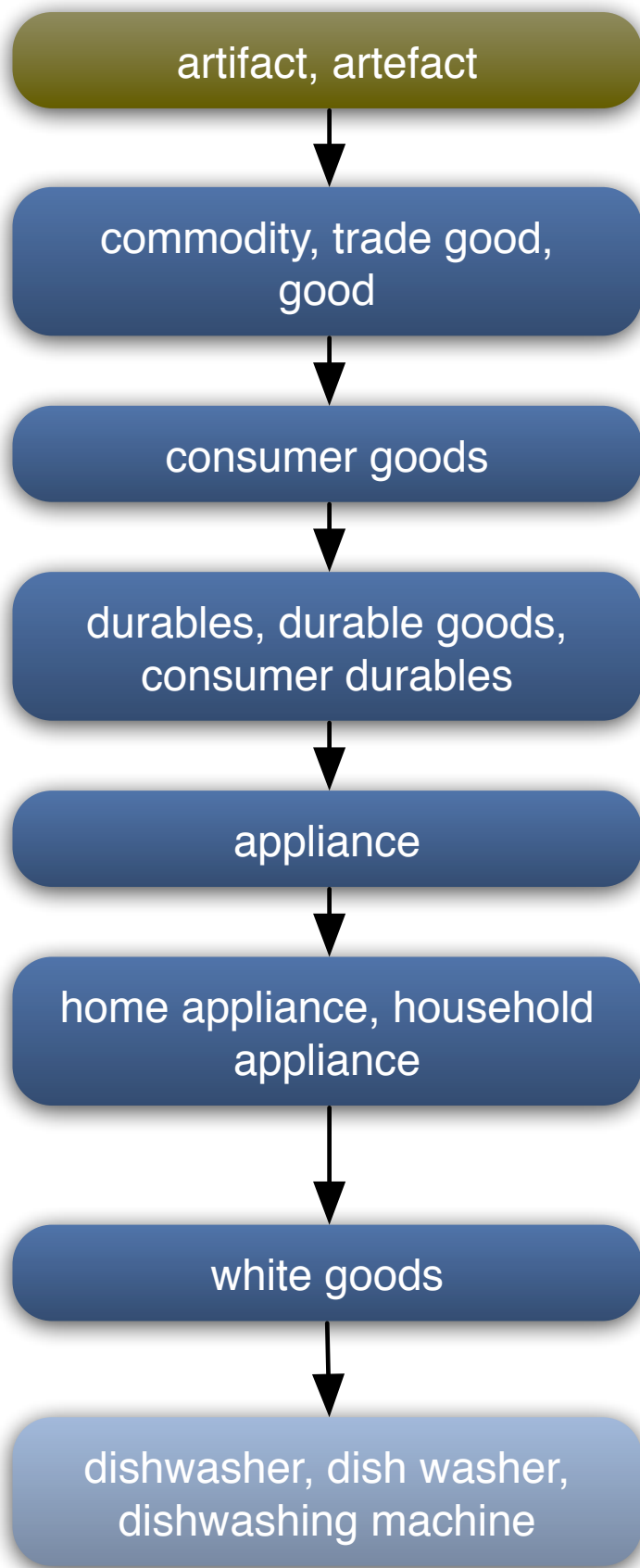


Sim = 0

Sim = 0

Sim = 1/2

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



Sim = 0 ←

Sim = 0 ←

Sim = 1/2 ←

$$\text{Final sim} = \frac{1/2 + 0 + 0}{3} = \frac{1}{6}$$

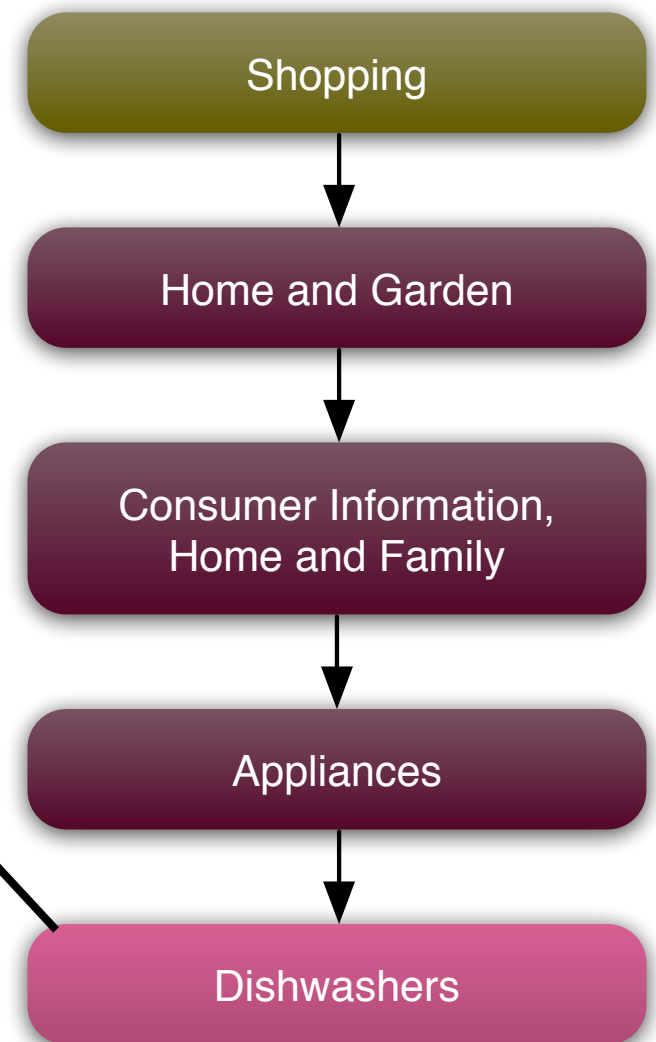
SI = dishwasher, ... (a machine for washing dishes)

Word sense disambiguation

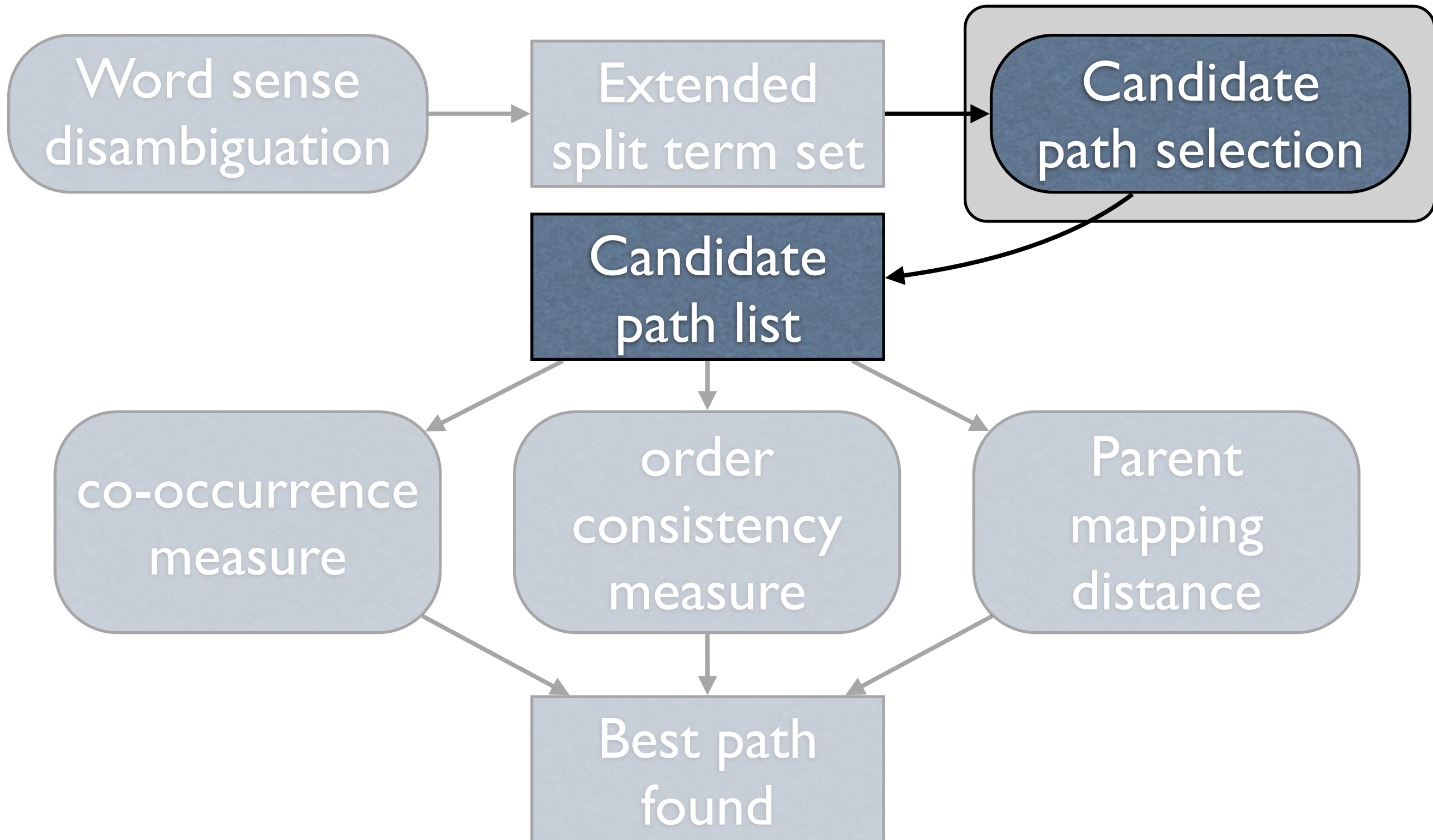
Terms of source node = {Dishwashers}

Extended Split Term Set = {*splitTermSet*}

{Dishwashers, dishwasher, dish washer, dishwashing machine}



Part II

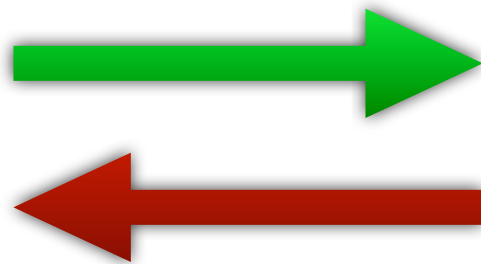


Candidate path search

Candidate path search

Takes into account the
composite categories

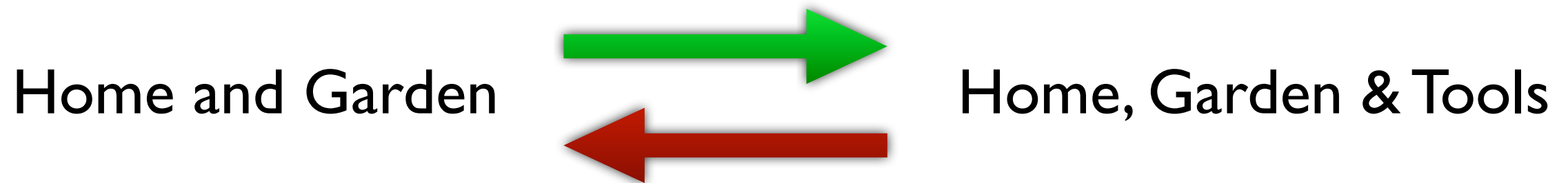
Home and Garden



Home, Garden & Tools

Candidate path search

Takes into account the
composite categories

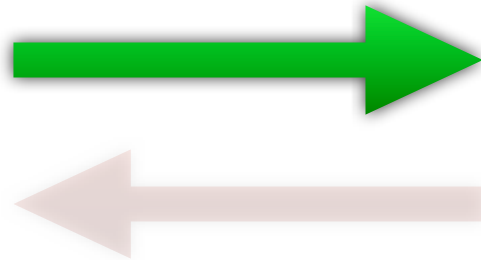


For every target category:
check whether source category is
a 'subset' of the target category

Candidate path search

source

Home and Garden



target

Home, Garden & Tools

Extended Split Term Set =
{{Home, ...}, {Garden, ...}}

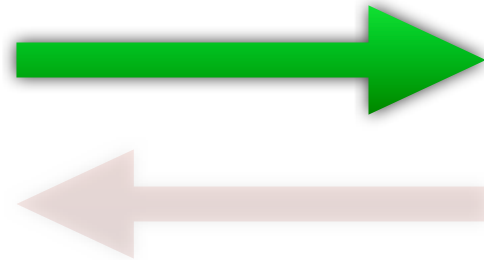
Candidate path search

source

Home and Garden

target

Home, Garden & Tools



Extended Split Term Set =
{{Home, ...}, {Garden, ...}}

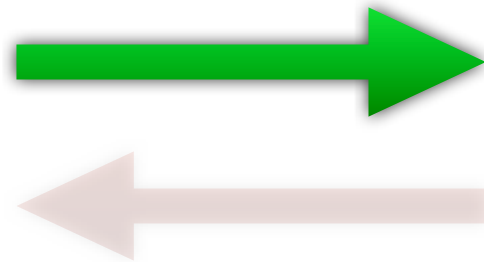


Candidate path search

source

target

Home and Garden



Home, Garden & Tools

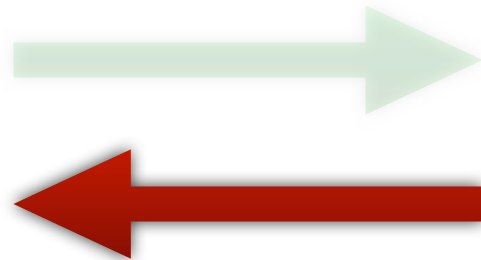
Extended Split Term Set =
{ {Home, ...}, {Garden, ...} }



Candidate path search

target

Home and Garden



source

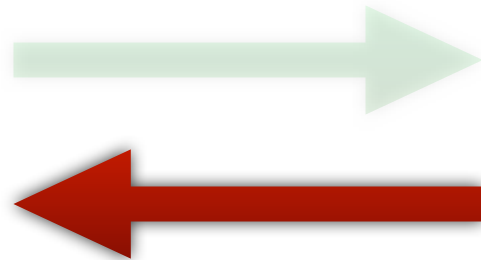
Home, Garden & Tools

Extended Split Term Set =
{ {Home, ...}, {Garden, ...}, {Tools, ...} }

Candidate path search

target

Home and Garden



source

Home, Garden & Tools

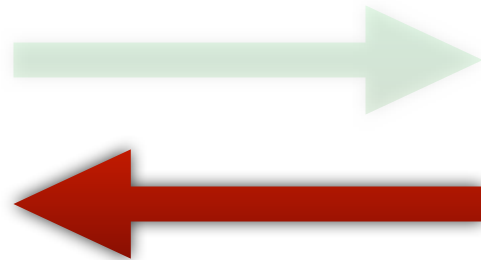
Extended Split Term Set =
{ {Home, ...}, {Garden, ...}, {Tools, ...} }



Candidate path search

target

Home and Garden



source

Home, Garden & Tools

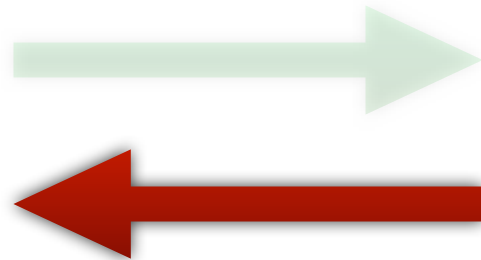
Extended Split Term Set =
{ {Home, ...}, {Garden, ...}, {Tools, ...} }



Candidate path search

target

Home and Garden



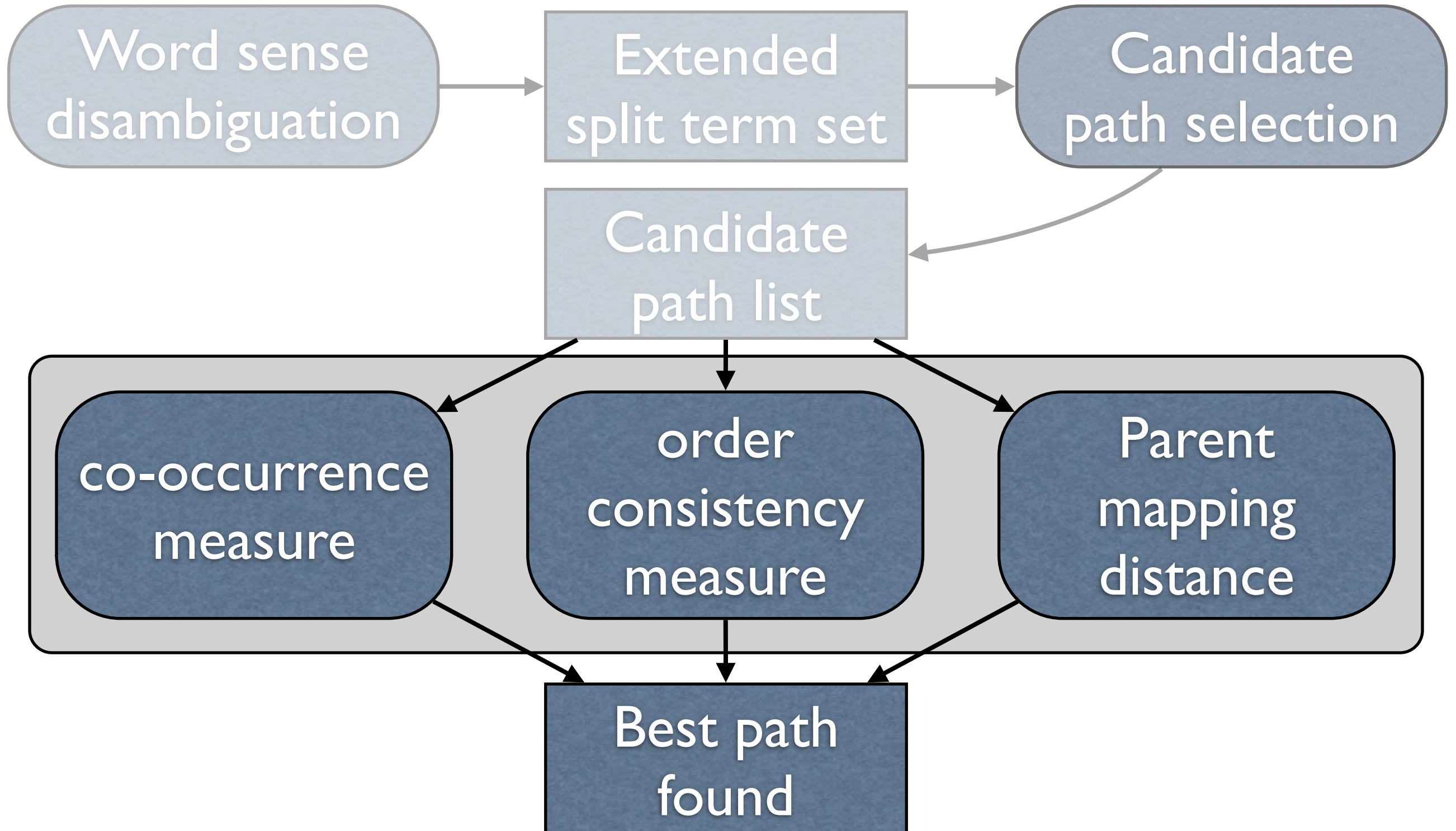
source

Home, Garden & Tools

Extended Split Term Set =
{ {Home, ...}, {Garden, ...}, {Tools, ...} }



Part III



Co-occurrence

$$\text{coOccurrence}(P_{\text{src}}, P_{\text{targ}}) = \left(\sum_{t \in P_{\text{targ}}} \frac{\text{maxSim}(t, P_{\text{src}})}{|P_{\text{targ}}|} \right) \cdot \left(\sum_{t \in P_{\text{src}}} \frac{\text{maxSim}(t, P_{\text{targ}})}{|P_{\text{src}}|} \right)$$

where P_{src} = list of nodes from the current source path

P_{targ} = list of nodes from a candidate target path

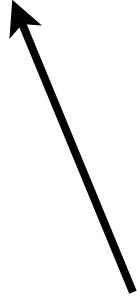
Co-occurrence

$$\text{coOccurrence}(P_{\text{src}}, P_{\text{targ}}) = \left(\sum_{t \in P_{\text{targ}}} \frac{\text{maxSim}(t, P_{\text{src}})}{|P_{\text{targ}}|} \right) \cdot \left(\sum_{t \in P_{\text{src}}} \frac{\text{maxSim}(t, P_{\text{targ}})}{|P_{\text{src}}|} \right)$$

where P_{src} = list of nodes from the current source path
 P_{targ} = list of nodes from a candidate target path

Co-occurrence

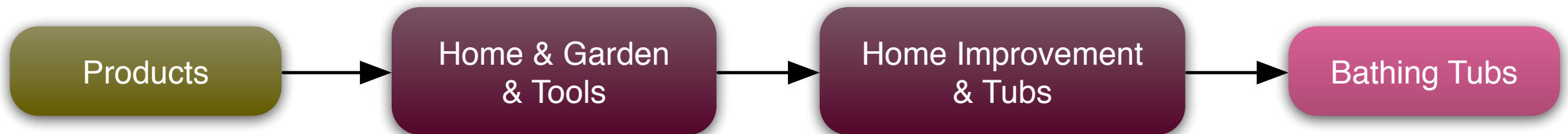
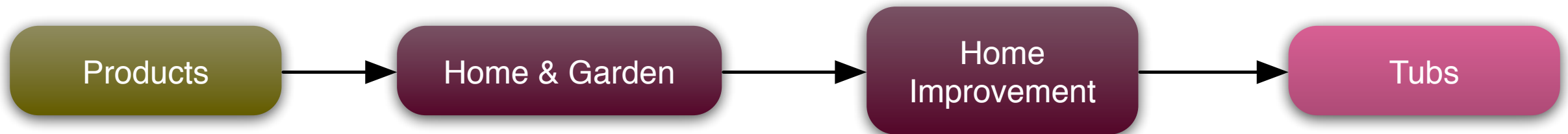
maximum
Jaccard similarity

$$\text{coOccurrence}(P_{\text{src}}, P_{\text{targ}}) = \left(\sum_{t \in P_{\text{targ}}} \frac{\text{maxSim}(t, P_{\text{src}})}{|P_{\text{targ}}|} \right) \cdot \left(\sum_{t \in P_{\text{src}}} \frac{\text{maxSim}(t, P_{\text{targ}})}{|P_{\text{src}}|} \right)$$


where P_{src} = list of nodes from the current source path
 P_{targ} = list of nodes from a candidate target path

Co-occurrence

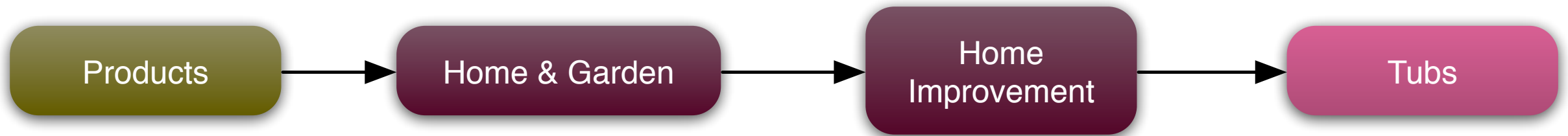
Source category path



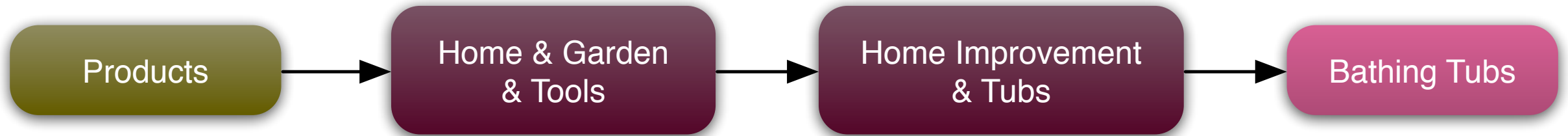
Candidate category path

Co-occurrence

Source category path



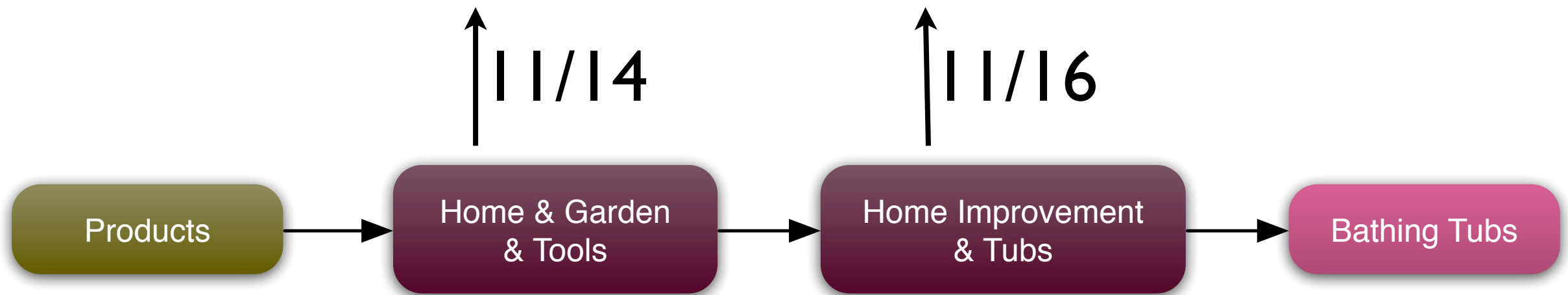
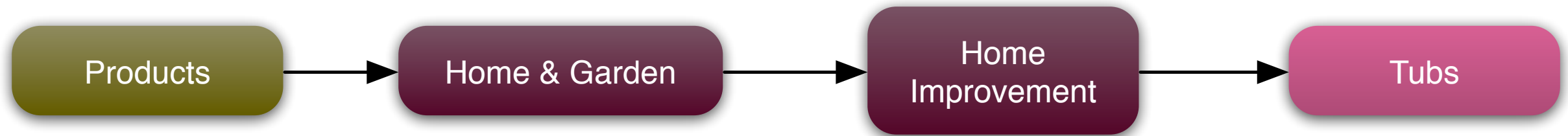
↑ 11/14



Candidate category path

Co-occurrence

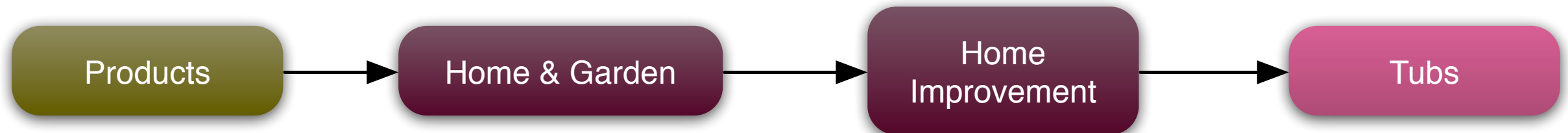
Source category path



Candidate category path

Co-occurrence

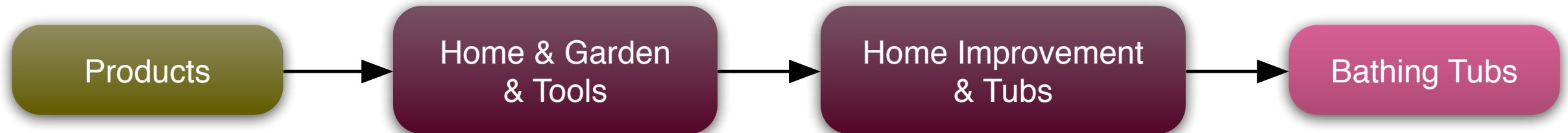
Source category path



↑ 11/14

↑ 11/16

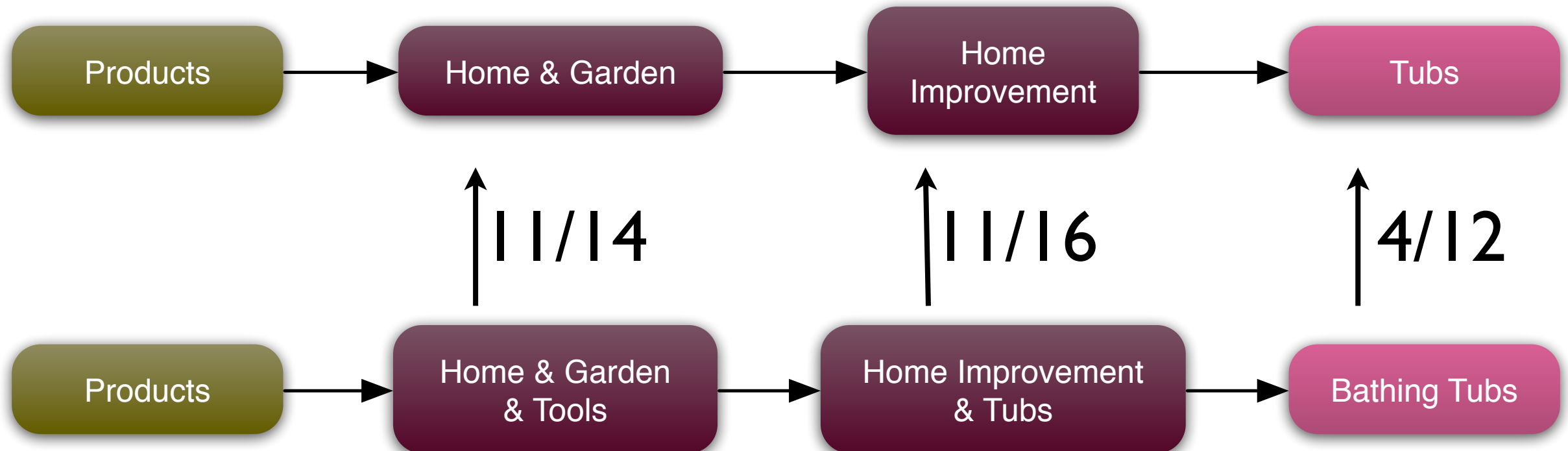
↑ 4/12



Candidate category path

Co-occurrence

Source category path

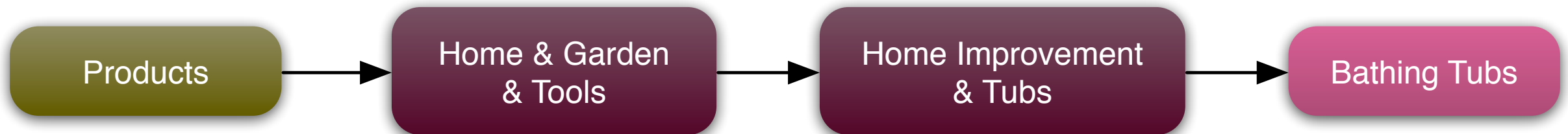
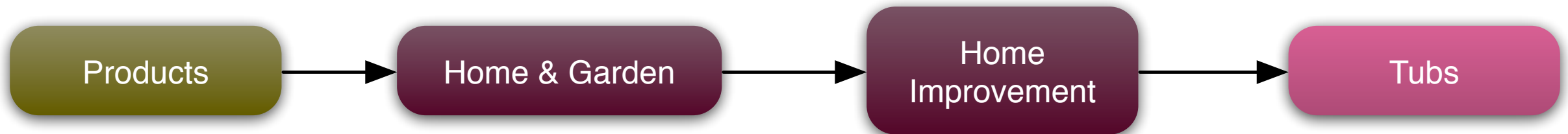


Candidate category path

$$\text{left term of co-occ} = \frac{1}{3} \times \left(\frac{11}{14} + \frac{11}{16} + \frac{4}{12} \right) \approx 0.683$$

Co-occurrence

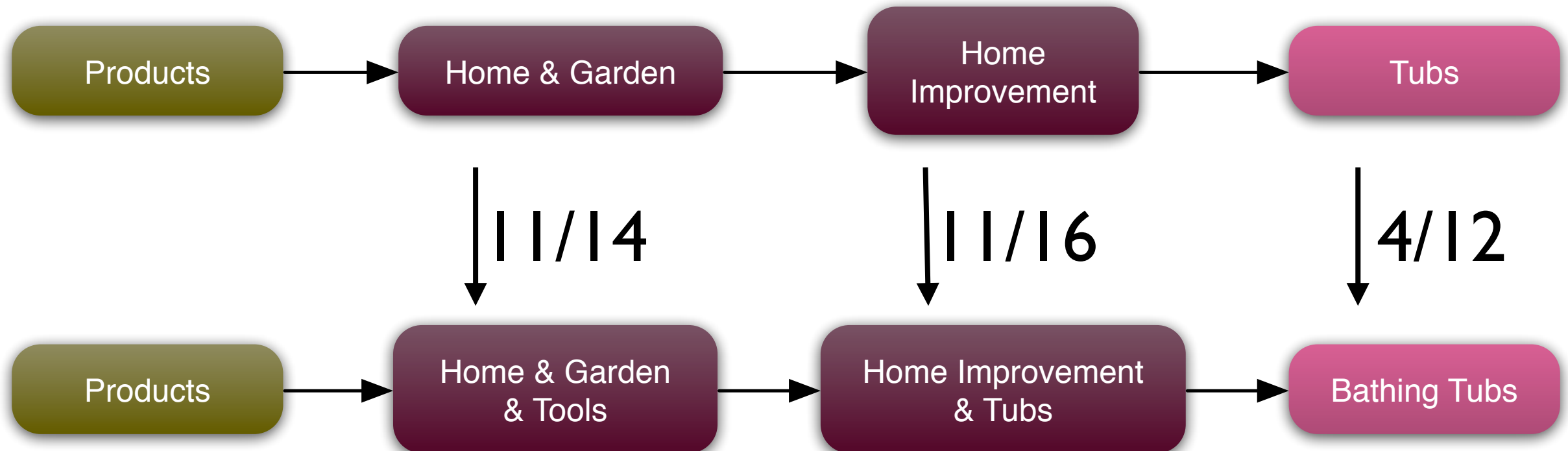
Source category path



Candidate category path

Co-occurrence

Source category path

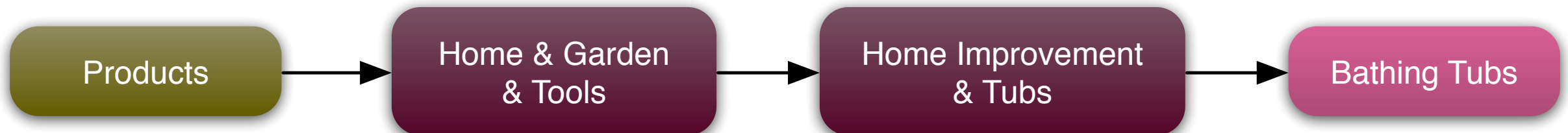
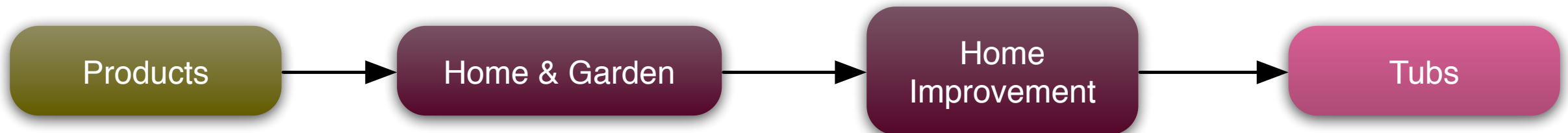


Candidate category path

$$\text{right term of co-occ} = \frac{1}{3} \times \left(\frac{11}{14} + \frac{11}{16} + \frac{4}{12} \right) \approx 0.683$$

Co-occurrence

Source category path



Candidate category path

$$\text{co-occ} = \left(\frac{1}{3} \times \left(\frac{11}{14} + \frac{11}{16} + \frac{4}{12} \right) \right)^2 \approx 0.466$$

Order consistency

$$\text{orderConsistency}(P_{\text{src}}, P_{\text{targ}}) = \sum_{r \in R} \frac{\text{consistent}(r, P_{\text{targ}})}{\binom{\text{length}(C)}{2}}$$

where P_{src} = list of nodes from the current source path

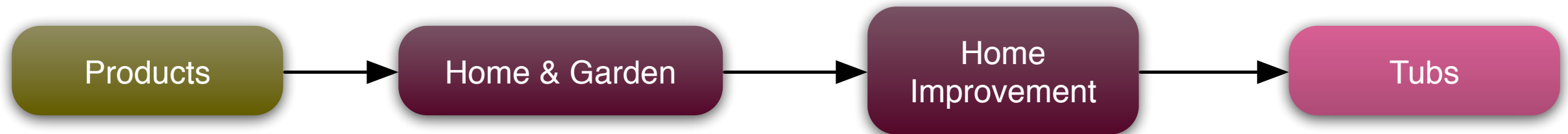
P_{targ} = list of nodes from a candidate target path

C = $\text{common}(P_{\text{src}}, P_{\text{targ}})$

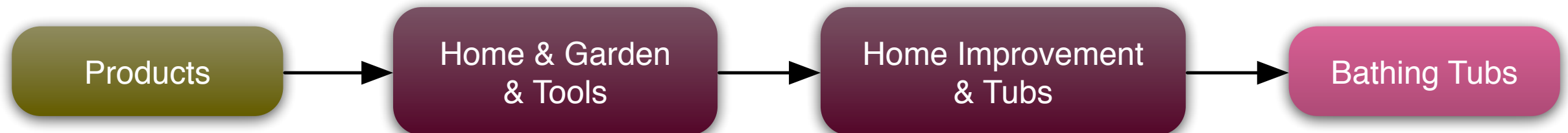
R = $\text{precedenceRelations}(C, P_{\text{src}})$

Order consistency

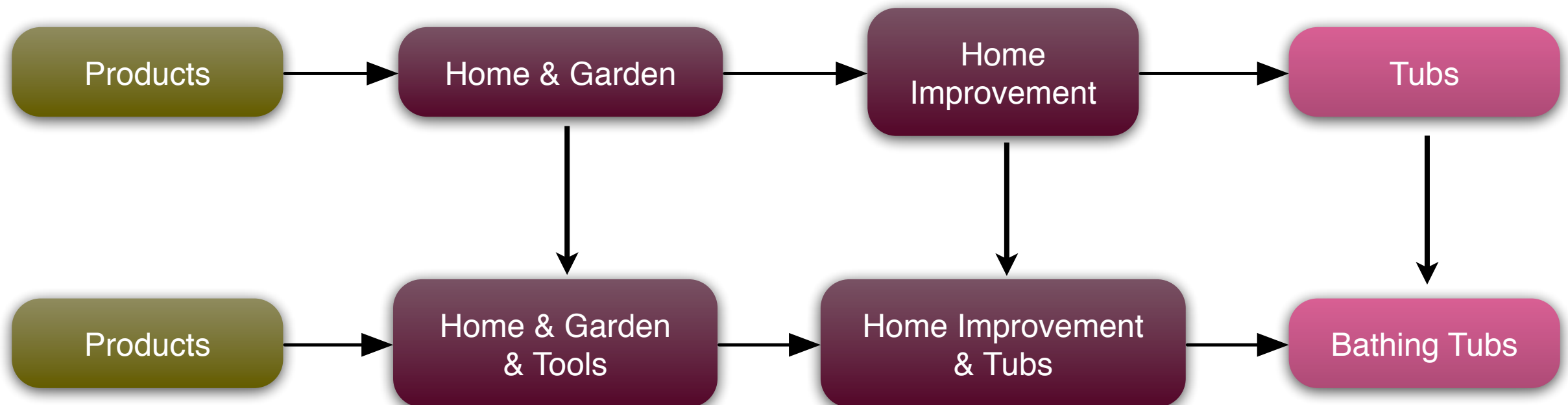
Source category path



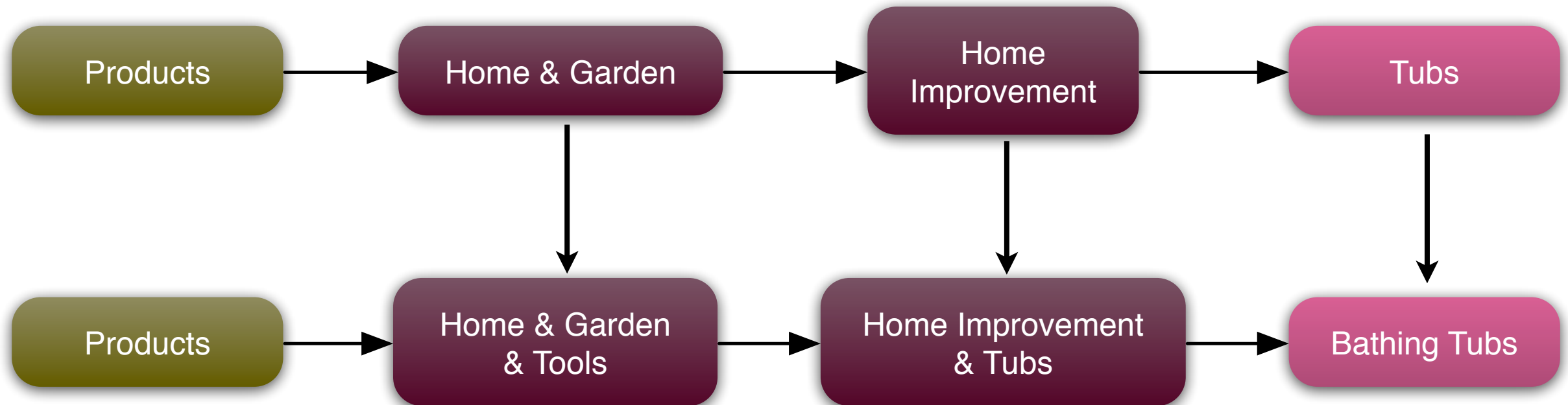
Candidate category path



Order consistency

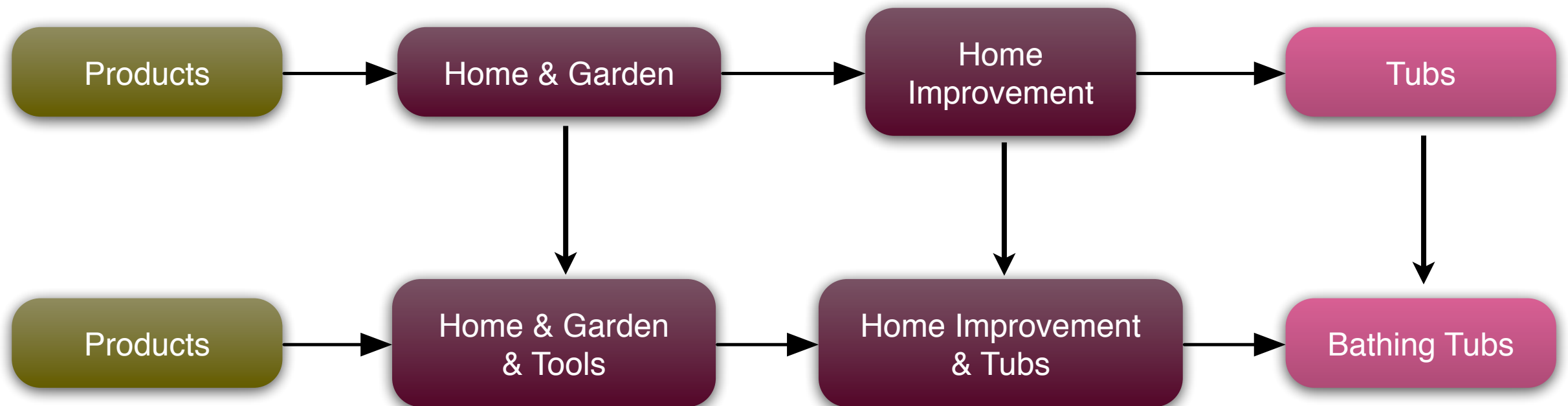


Order consistency

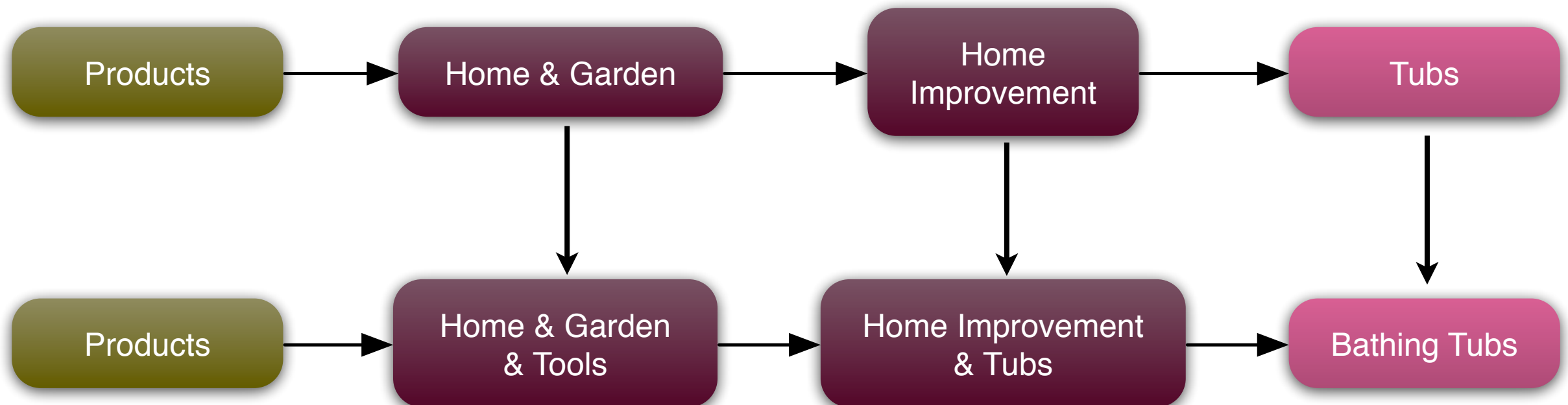


$C = \{$
('Home & Garden', 'Home & Garden & Tools'),
('Home Improvement', 'Home Improvement & Tubs'),
('Tubs', 'Bathing Tubs')
 $\}$

Order consistency

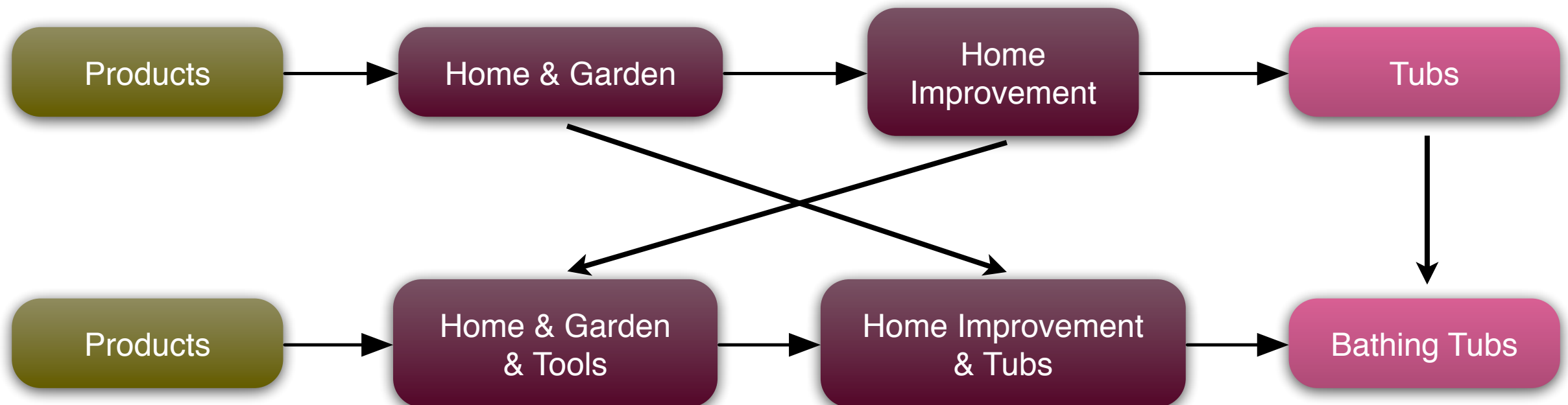


Order consistency



$R = \text{precedenceRelations}(C, P_{\text{src}}) = \{$
 ('Home & Garden & Tools', 'Home Improvement & Tubs'),
 ('Home Improvement & Tubs', 'Bathing Tubs'),
 ('Home & Garden & Tools', 'Bathing Tubs')
}

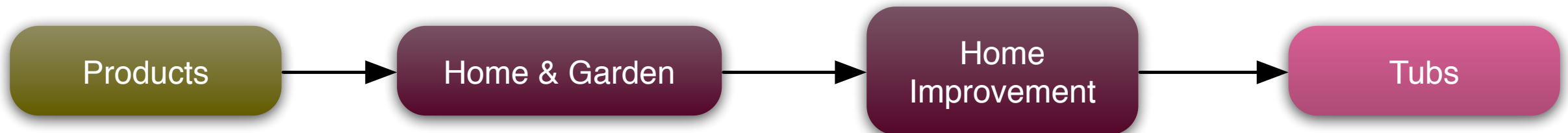
Order consistency



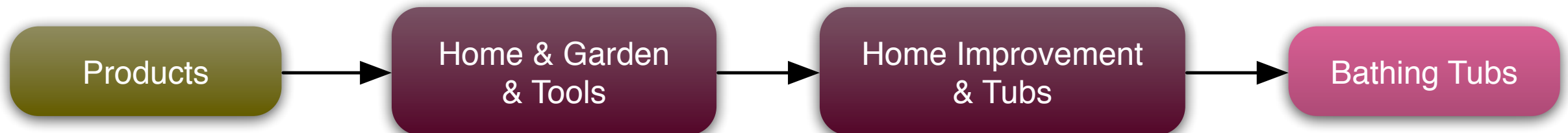
$R = \text{precedenceRelations}(C, P_{\text{src}}) = \{$
 ('Home Improvement & Tubs', 'Home & Garden & Tools'),
 ('Home & Garden & Tools', 'Bathing Tubs'),
 ('Home Improvement & Tubs', 'Bathing Tubs')
}

Order consistency

Source category path



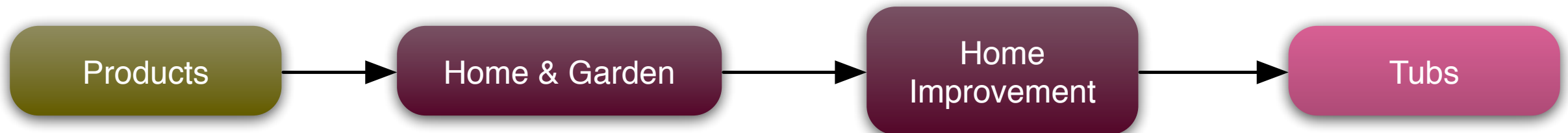
Candidate category path



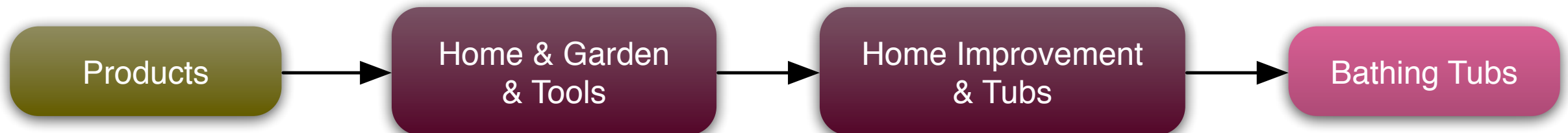
$R = \text{precedenceRelations}(C, P_{\text{src}}) = \{$
 ('Home & Garden & Tools', 'Home Improvement & Tubs'),
 ('Home Improvement & Tubs', 'Bathing Tubs'),
 ('Home & Garden & Tools', 'Bathing Tubs')
}

Order consistency

Source category path



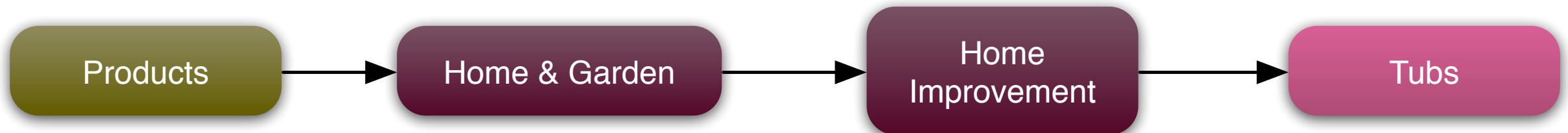
Candidate category path



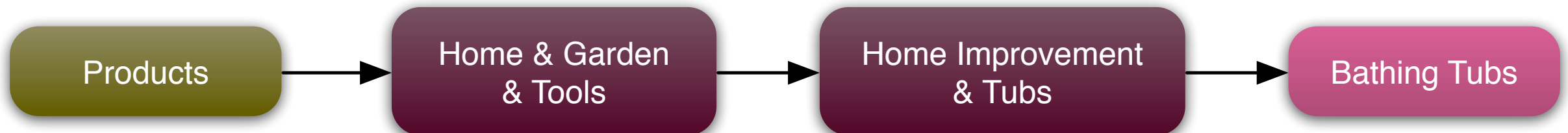
$R = \text{precedenceRelations}(C, P_{\text{src}}) = \{$
✓ ('Home & Garden & Tools', 'Home Improvement & Tubs'),
('Home Improvement & Tubs', 'Bathing Tubs'),
('Home & Garden & Tools', 'Bathing Tubs')
 $\}$

Order consistency

Source category path



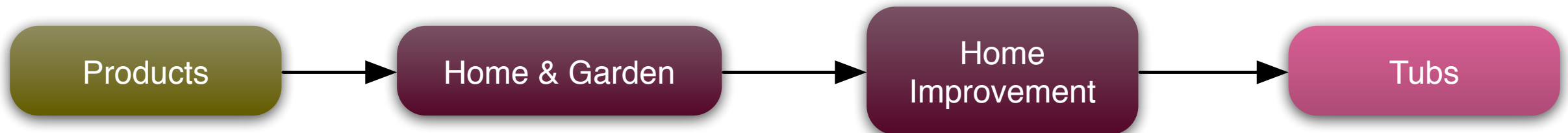
Candidate category path



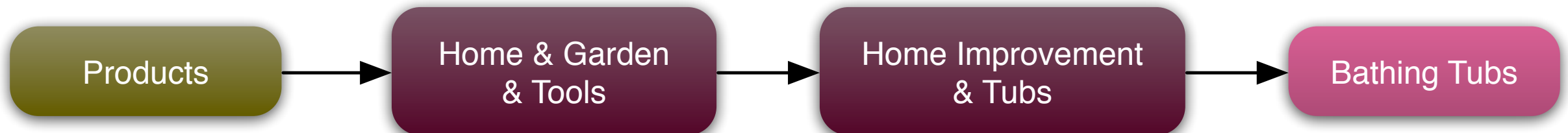
$R = \text{precedenceRelations}(C, P_{\text{src}}) = \{$
✓ ('Home & Garden & Tools', 'Home Improvement & Tubs'),
✓ ('Home Improvement & Tubs', 'Bathing Tubs'),
('Home & Garden & Tools', 'Bathing Tubs')
 $\}$

Order consistency

Source category path



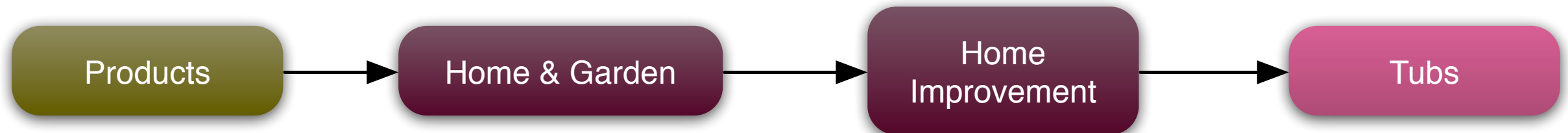
Candidate category path



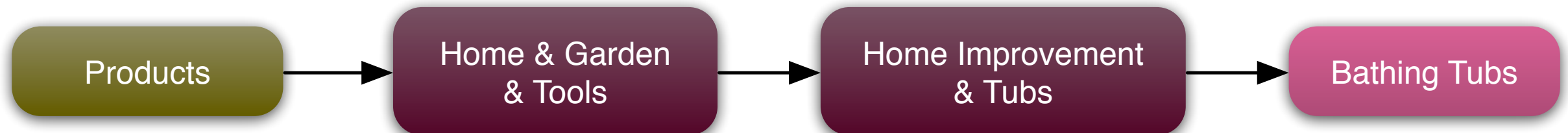
R = precedenceRelations(C, Psrc) = {
✓ ('Home & Garden & Tools', 'Home Improvement & Tubs'),
✓ ('Home Improvement & Tubs', 'Bathing Tubs'),
✓ ('Home & Garden & Tools', 'Bathing Tubs')
}

Order consistency

Source category path

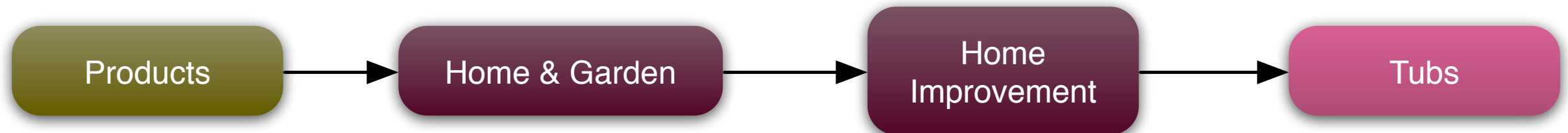


Candidate category path

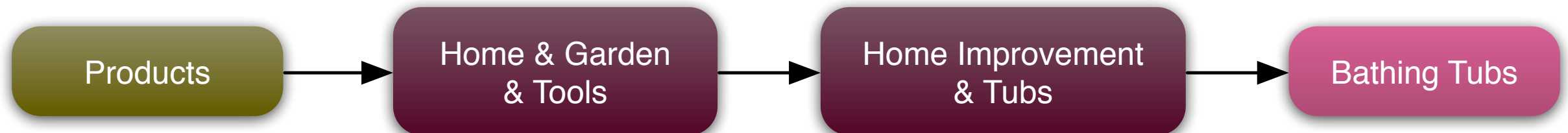


Order consistency

Source category path



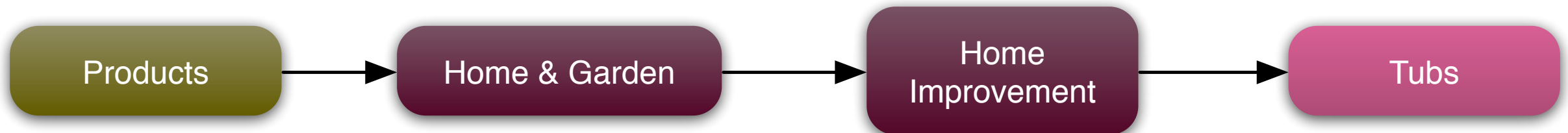
Candidate category path



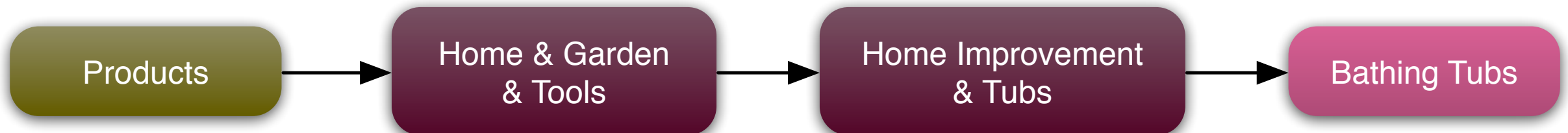
$$\text{Order consistency} = 3 / 3 = 1$$

Order consistency

Source category path



Candidate category path

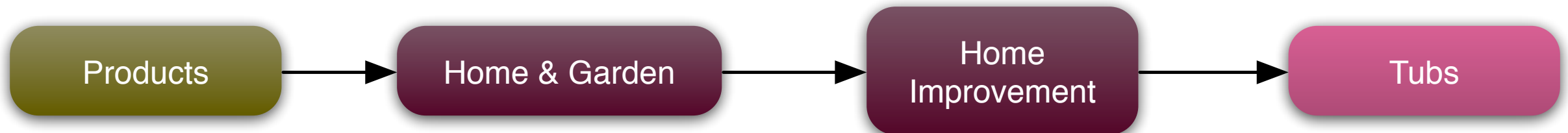


Total number of precedence relations
between common nodes

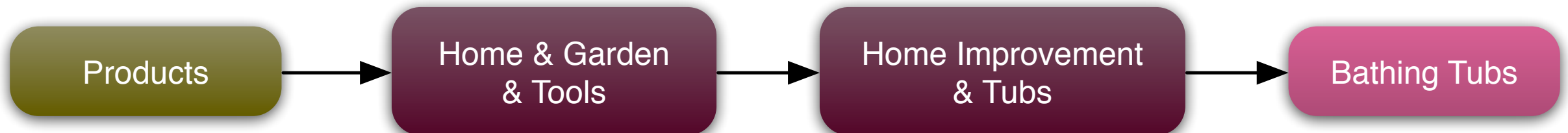
$$\text{Order consistency} = 3 / 3 = 1$$

Order consistency

Source category path



Candidate category path



Total number of precedence relations
between common nodes

$$\text{Order consistency} = 3 / 3 = 1$$

Total number of

consistent precedence relations between common nodes

Parent mapping distance

- Exploit our knowledge of how the parent of the source node was mapped
- Distance is computed using edge counting

Finding optimal path

- Aggregated similarity is the harmonic mean of the three measures:
 - co-occurrence
 - order-consistency
 - parent mapping distance
- Path with highest aggregated similarity is chosen (if above a certain threshold)

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
 - 6 data set combinations, sample size of 500
 - 3 individuals performed the evaluation

Evaluation

Overall results

Algorithm	Precision	Recall	F ₁	Computation Time
PROMPT	28.93%	16.69%	20.75%	0.47 sec
Park & Kim	47.77%	25.19%	32.52%	4.99 sec
Our algorithm	38.28%	83.66%	52.31%	20.71 sec

Questions?