An Automatic Approach for Mapping Product Taxonomies in E-Commerce Systems

Lennart Nederstigt, Steven Aanen,
Damir Vandic, Flavius Frasincar

ERASMUS UNIVERSITEIT ROTTERDAM

Terminology

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

Product taxonomies

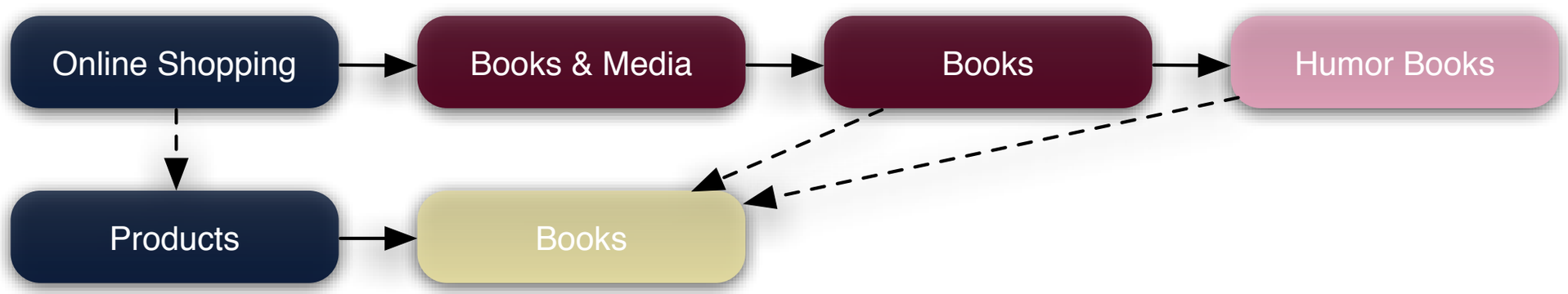
Important aspects of product taxonomies:

- 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



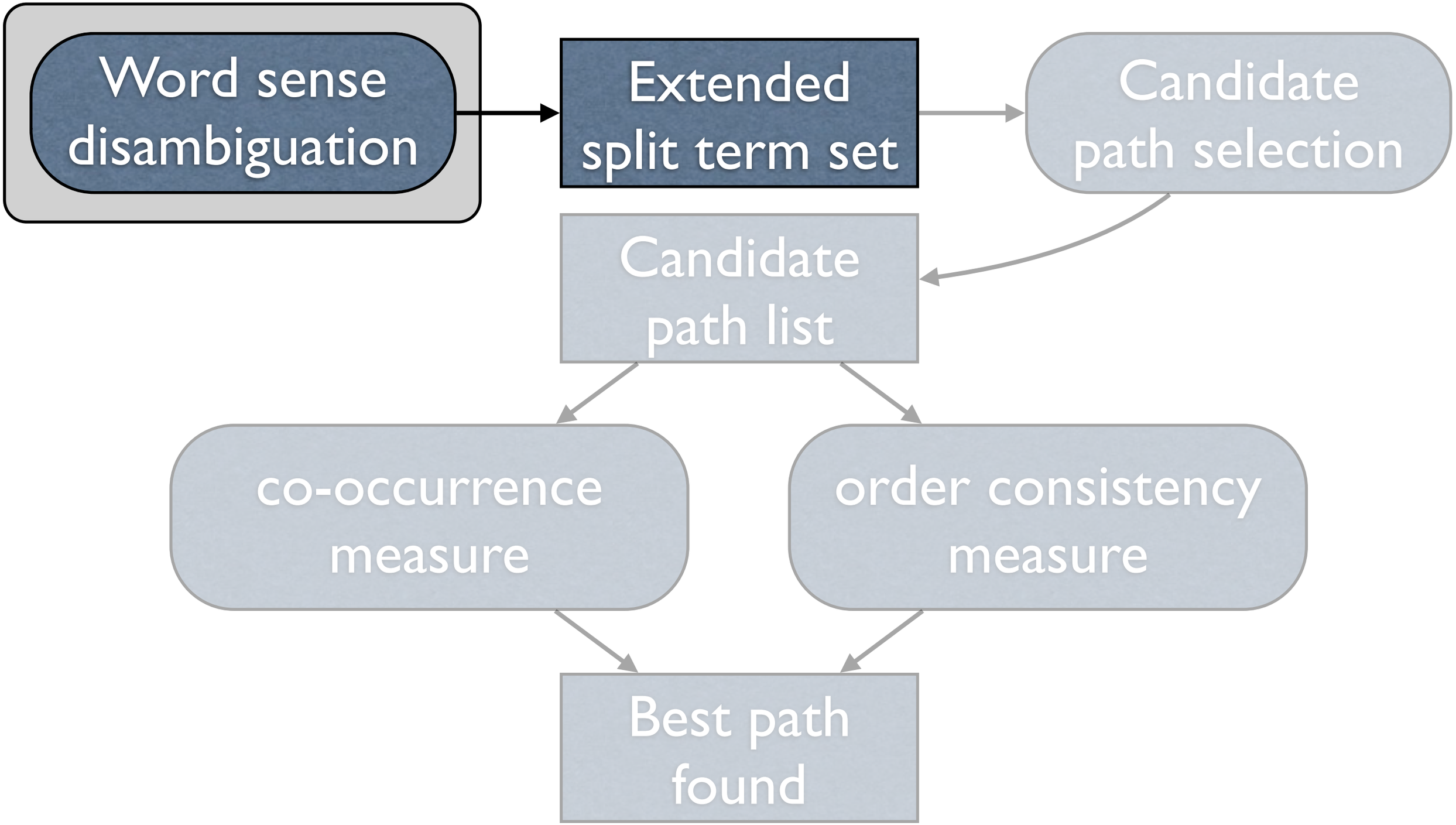
Related work

- 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*"

CMAP overview

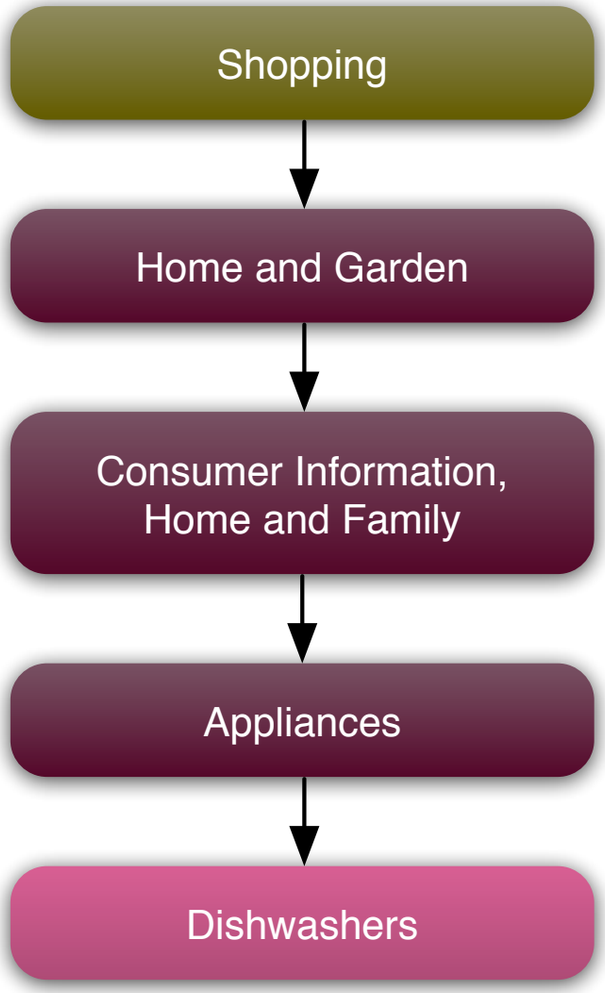
- Input is a source category path
- Output is a target category path (or 'None')
- CMAP consists of three steps
  1. word sense disambiguation
  2. candidate path search
  3. best path selection

CMAP - Part I


Word sense disambiguation
Extended split term set
Candidate path selection
Candidate path list
co-occurrence measure
order consistency measure
Best path found

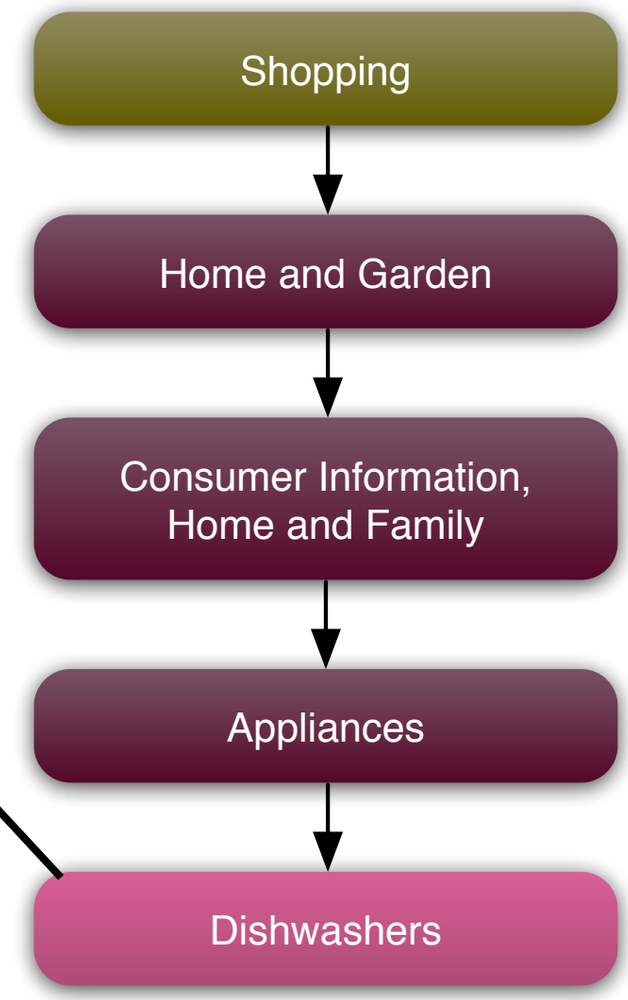
Word sense disambiguation

- 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}



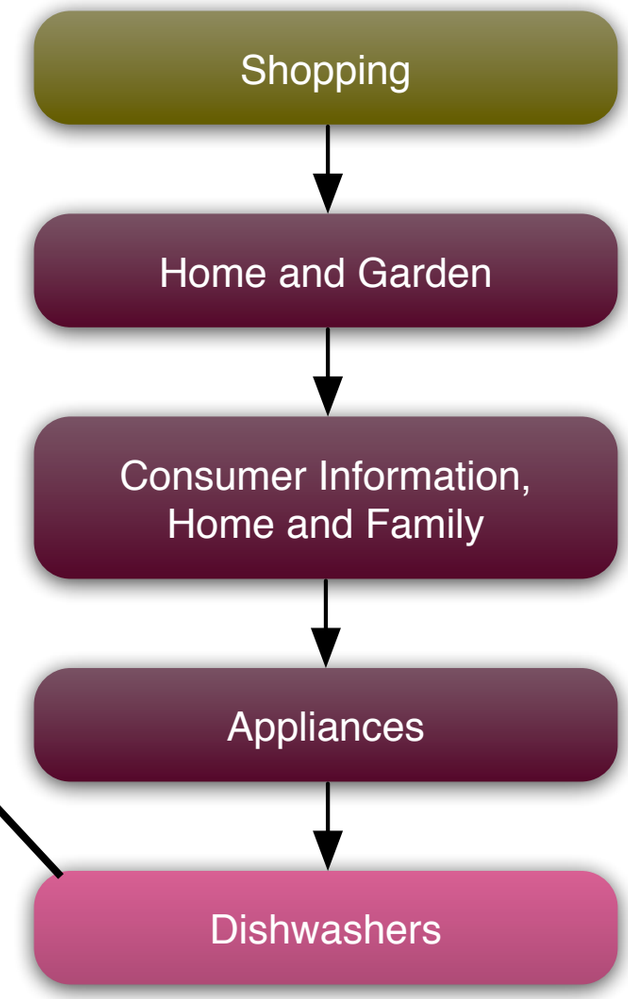
Word sense disambiguation

Terms of source node = {Dishwashers}

for each

{splitTermSet, ...}

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



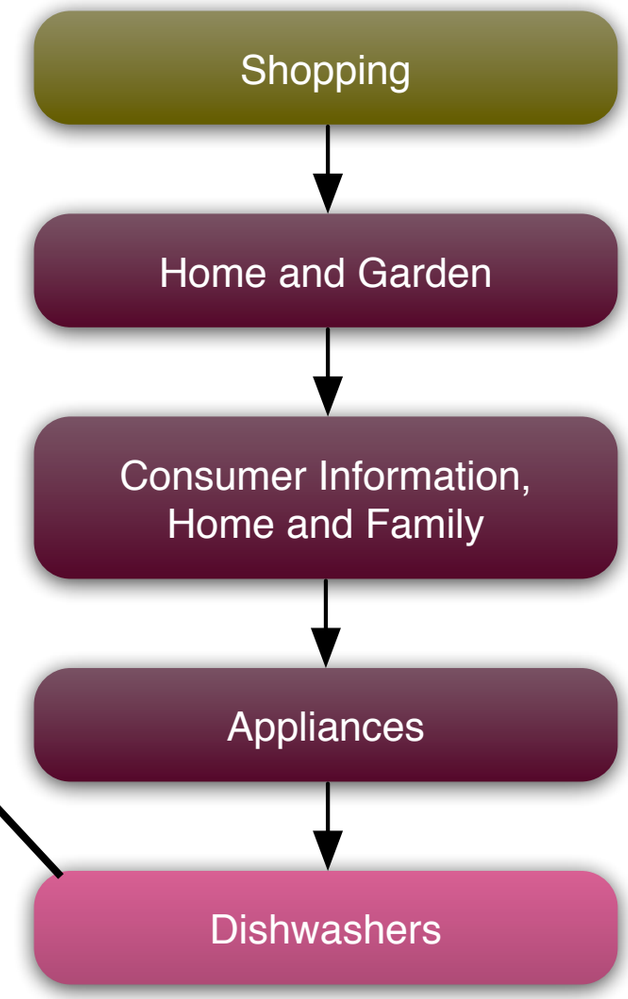
Word sense disambiguation

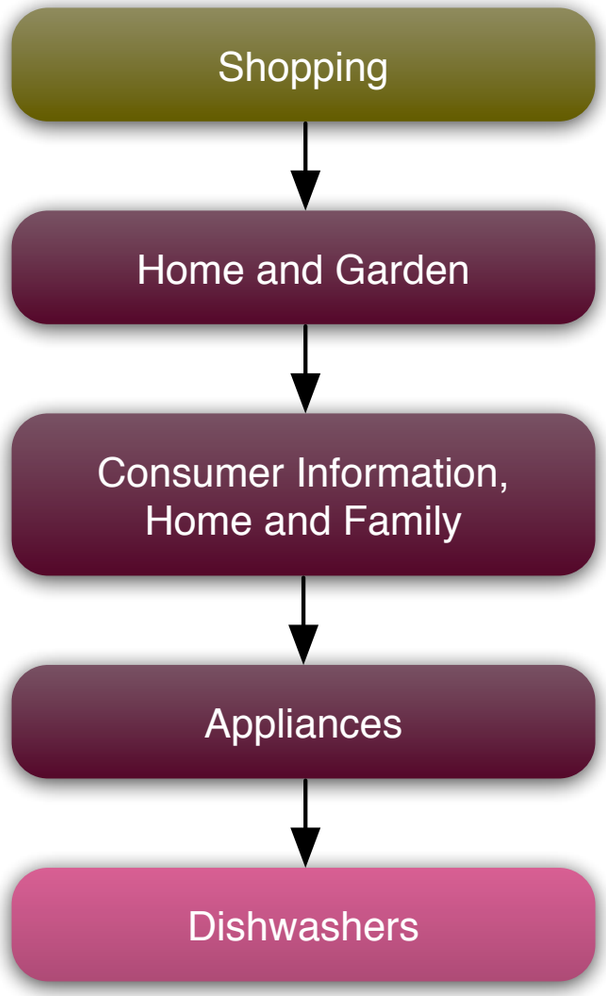
Terms of source node = {Dishwashers}

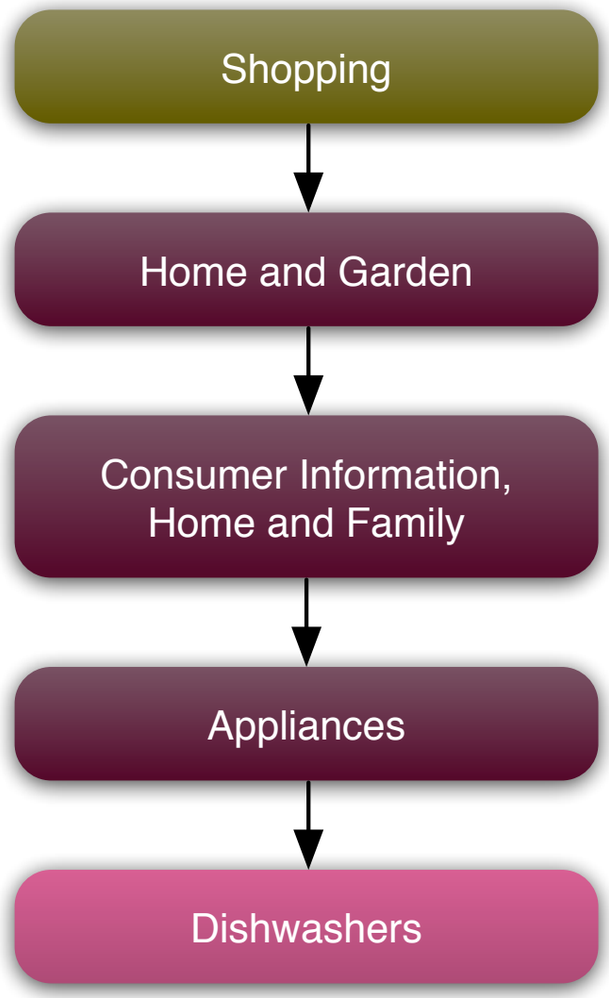
for each

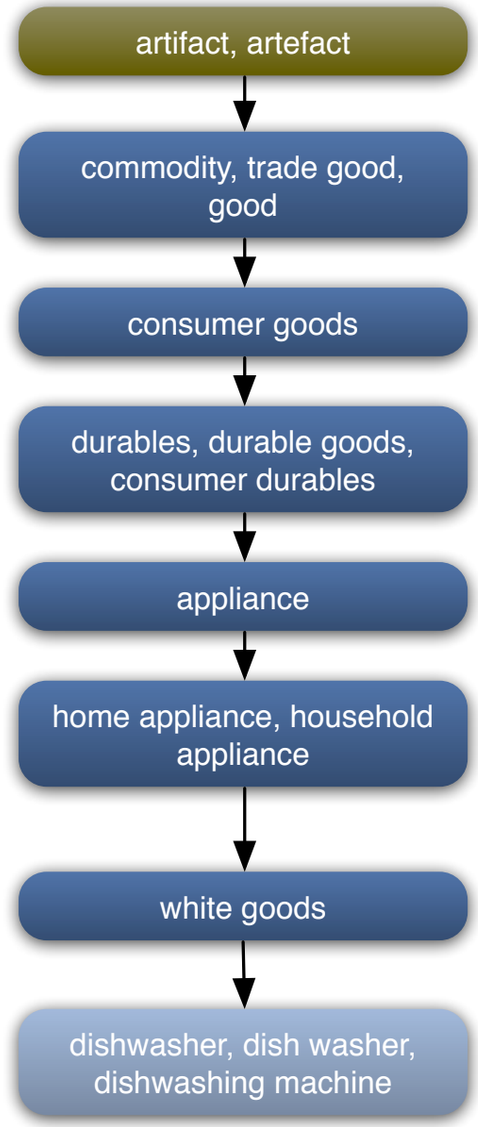
Extended Split Term Set = {*splitTermSet*, ...}

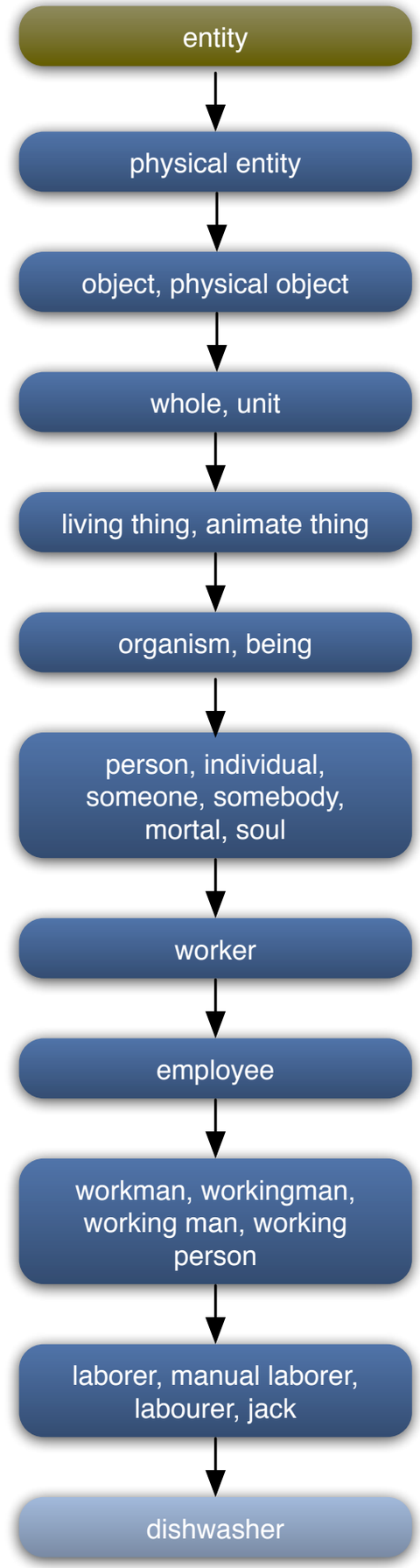
Synonyms of a split term that have the correct meaning in this context

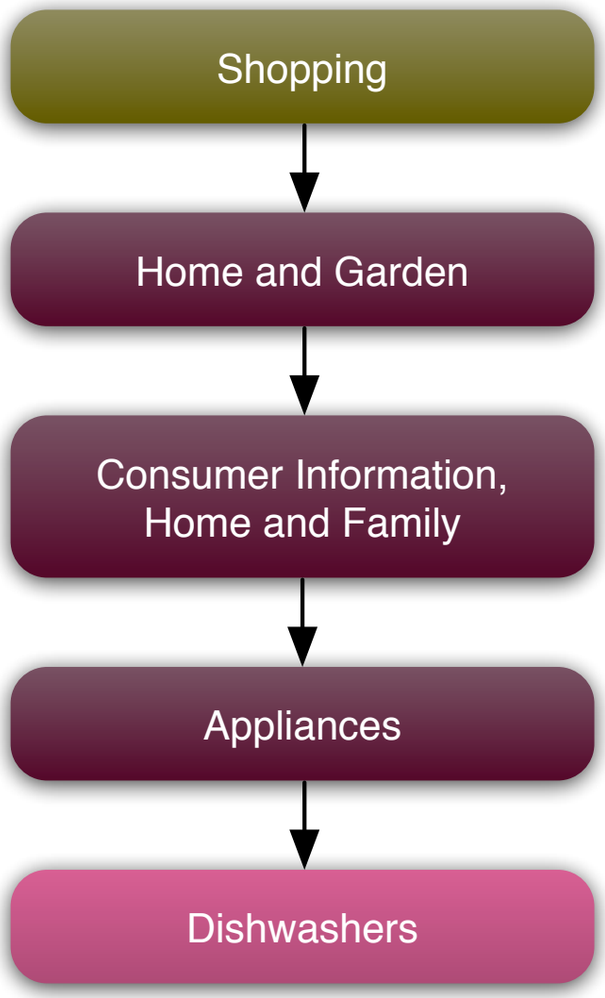


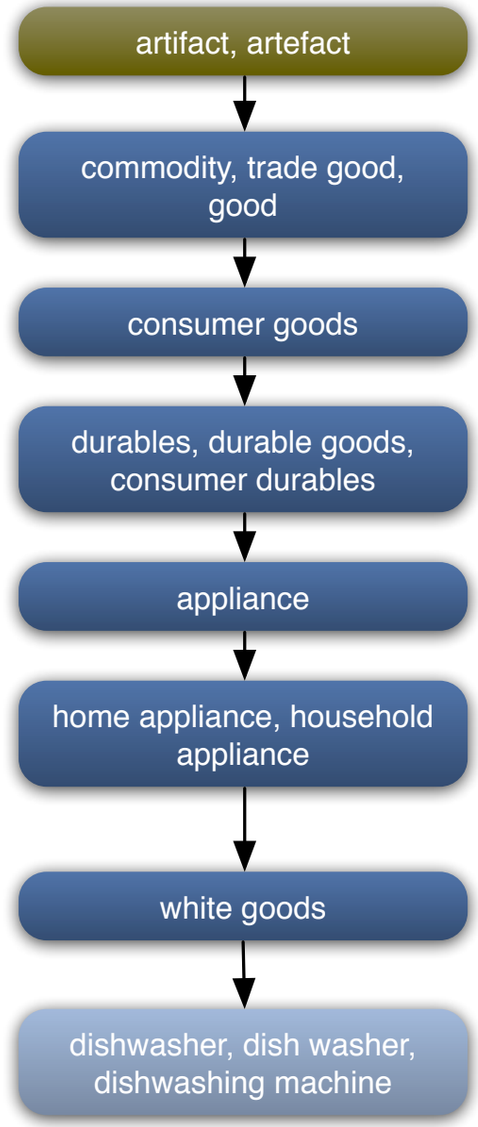

Shopping
Home and Garden
Consumer Information, Home and Family
Appliances
Dishwashers

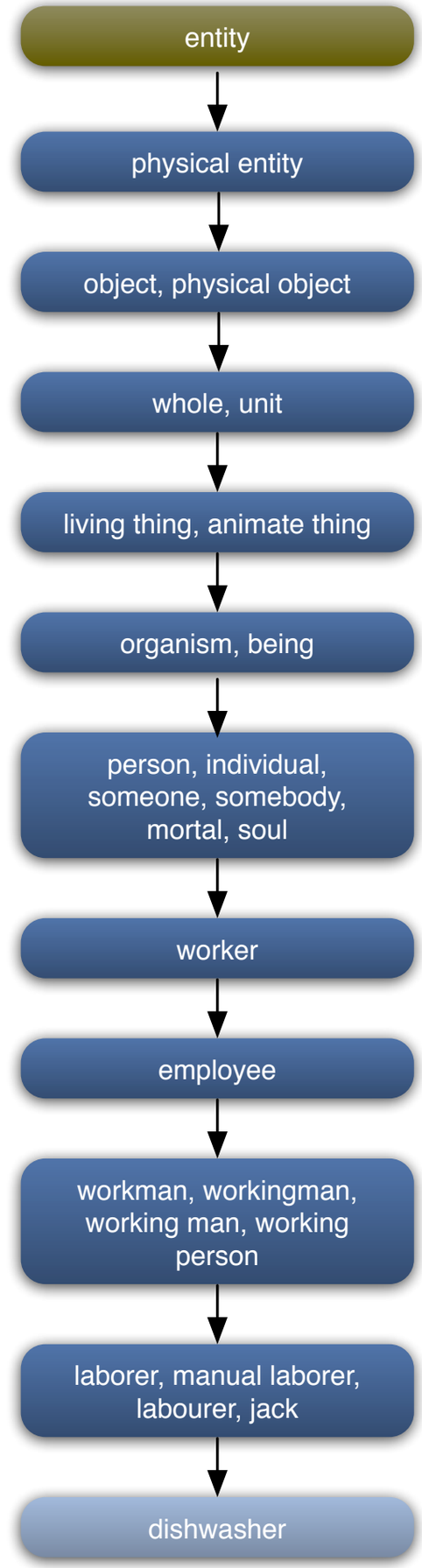

Shopping
Home and Garden
Consumer Information, Home and Family
Appliances
Dishwashers


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

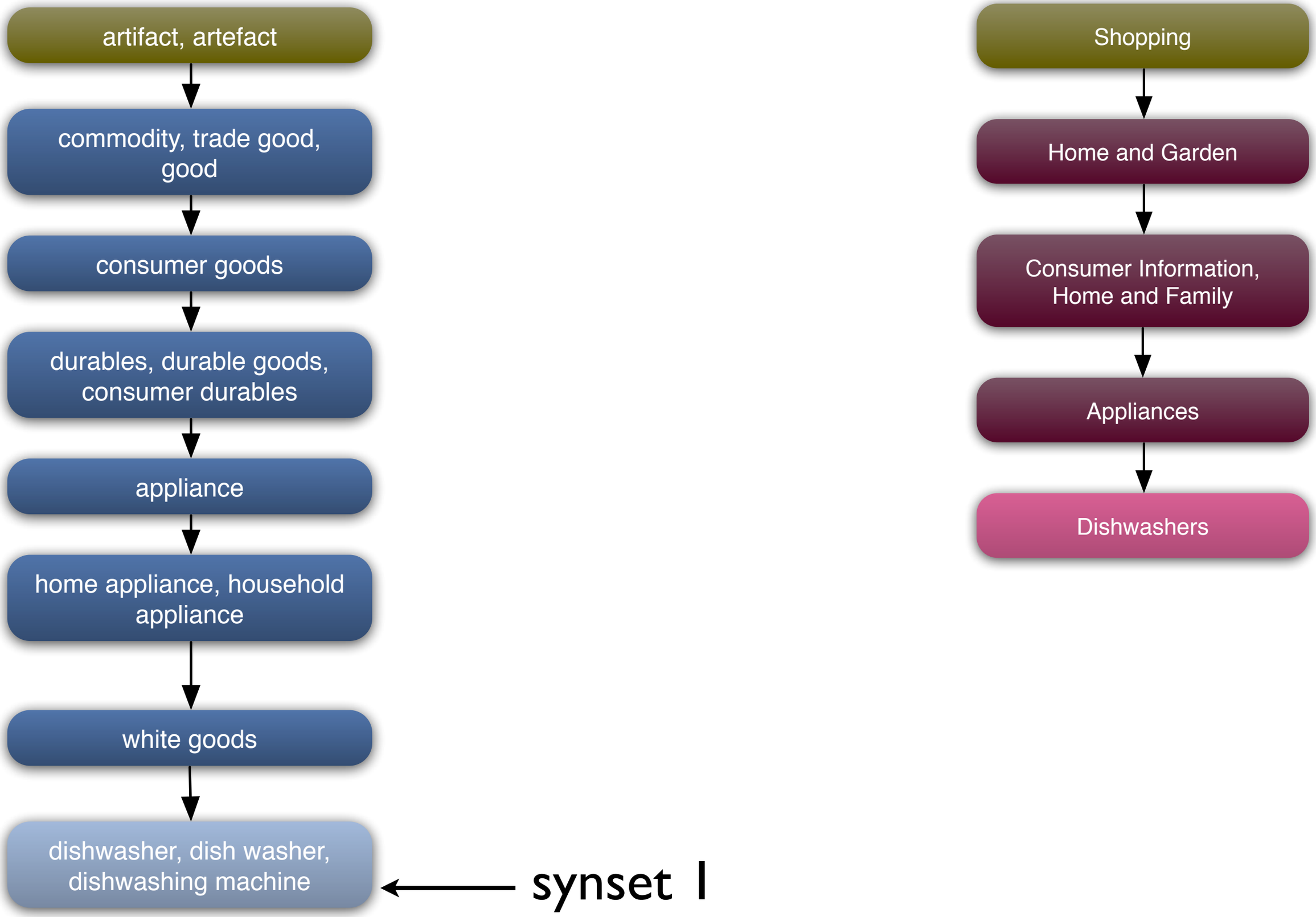

entity
physical entity
object, physical object
whole, unit
living thing, animate thing
organism, being
person, individual, someone, somebody, mortal, soul
worker
employee
workman, workingman, working man, working person
laborer, manual laborer, labourer, jack
dishwasher

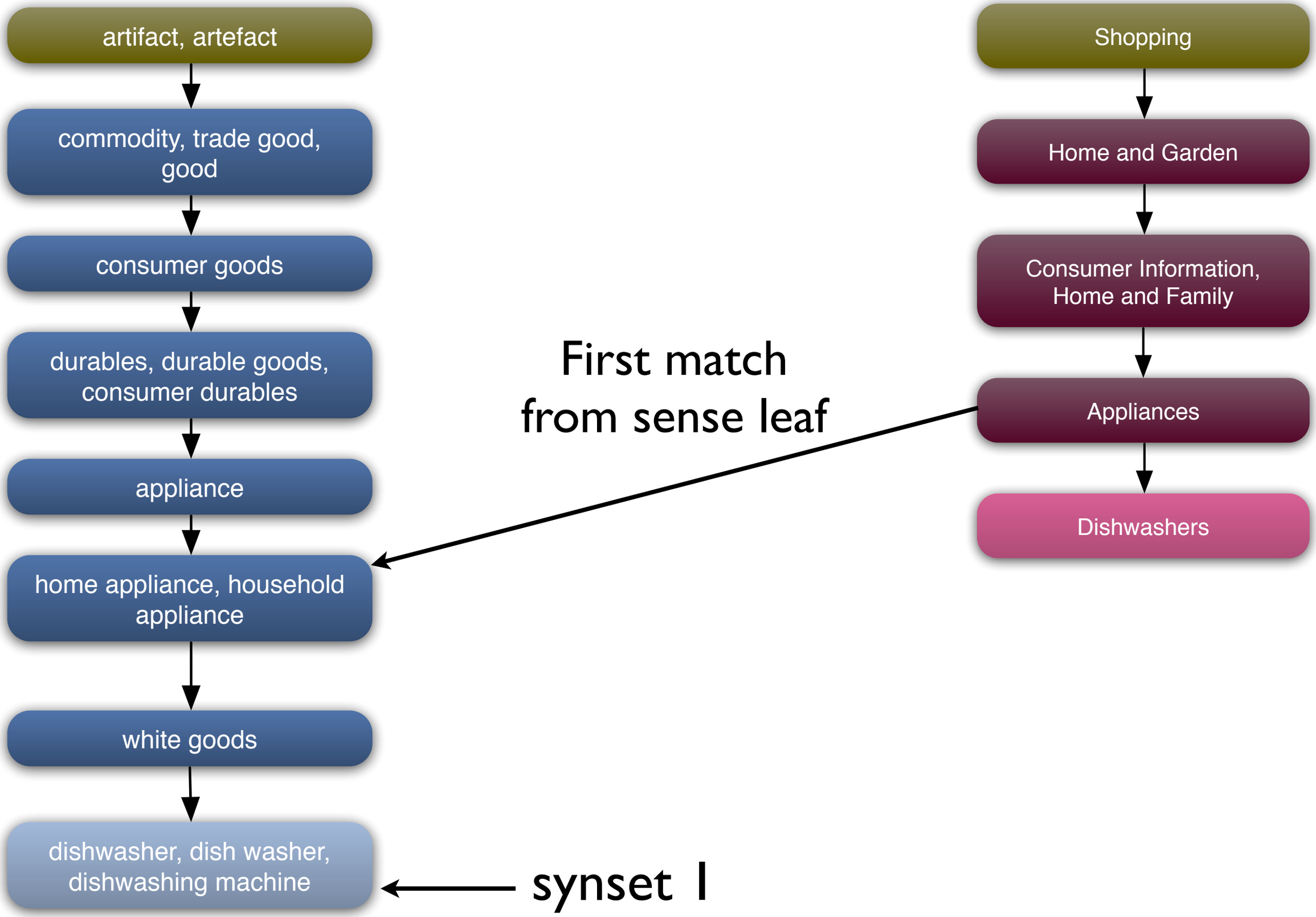


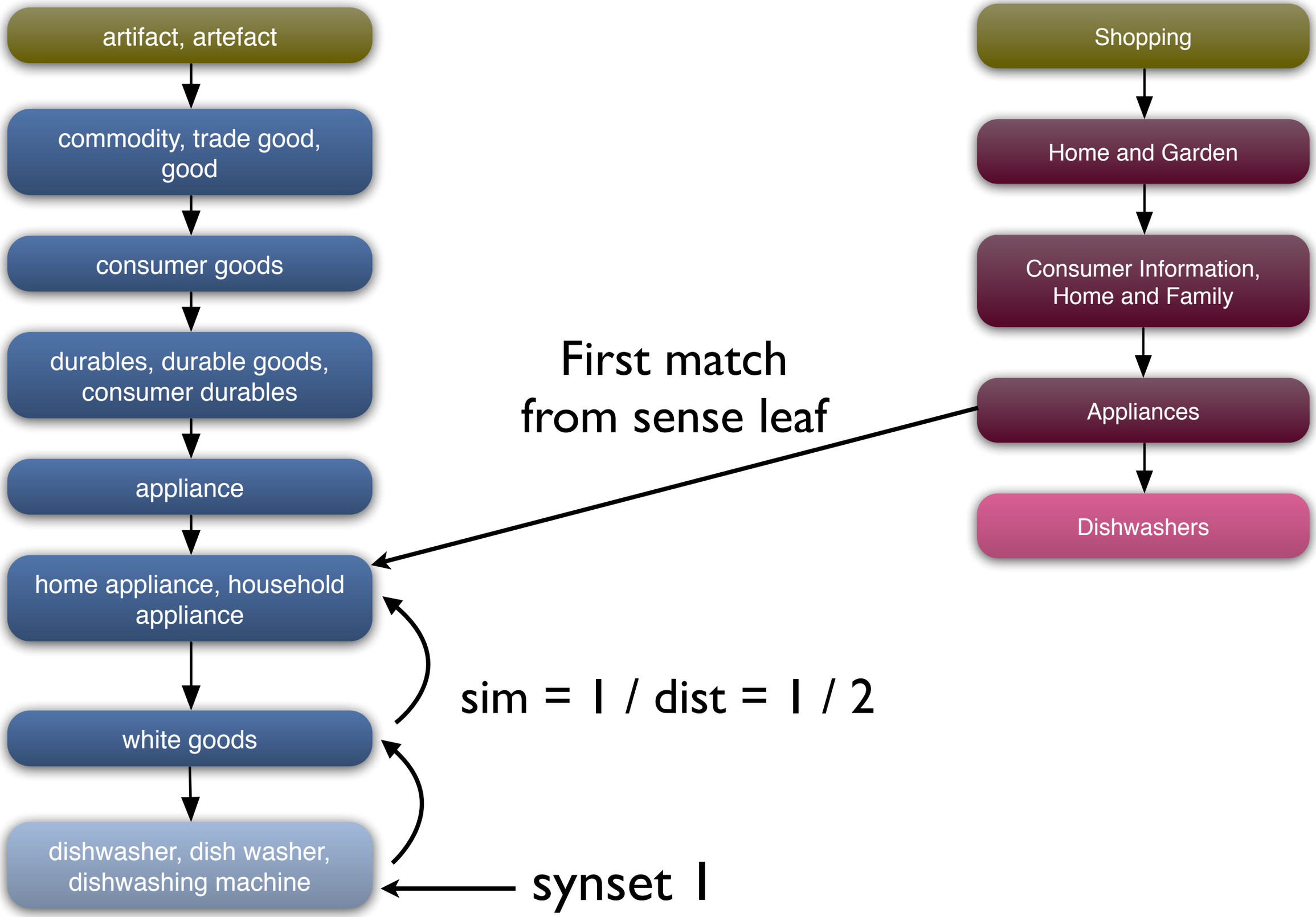


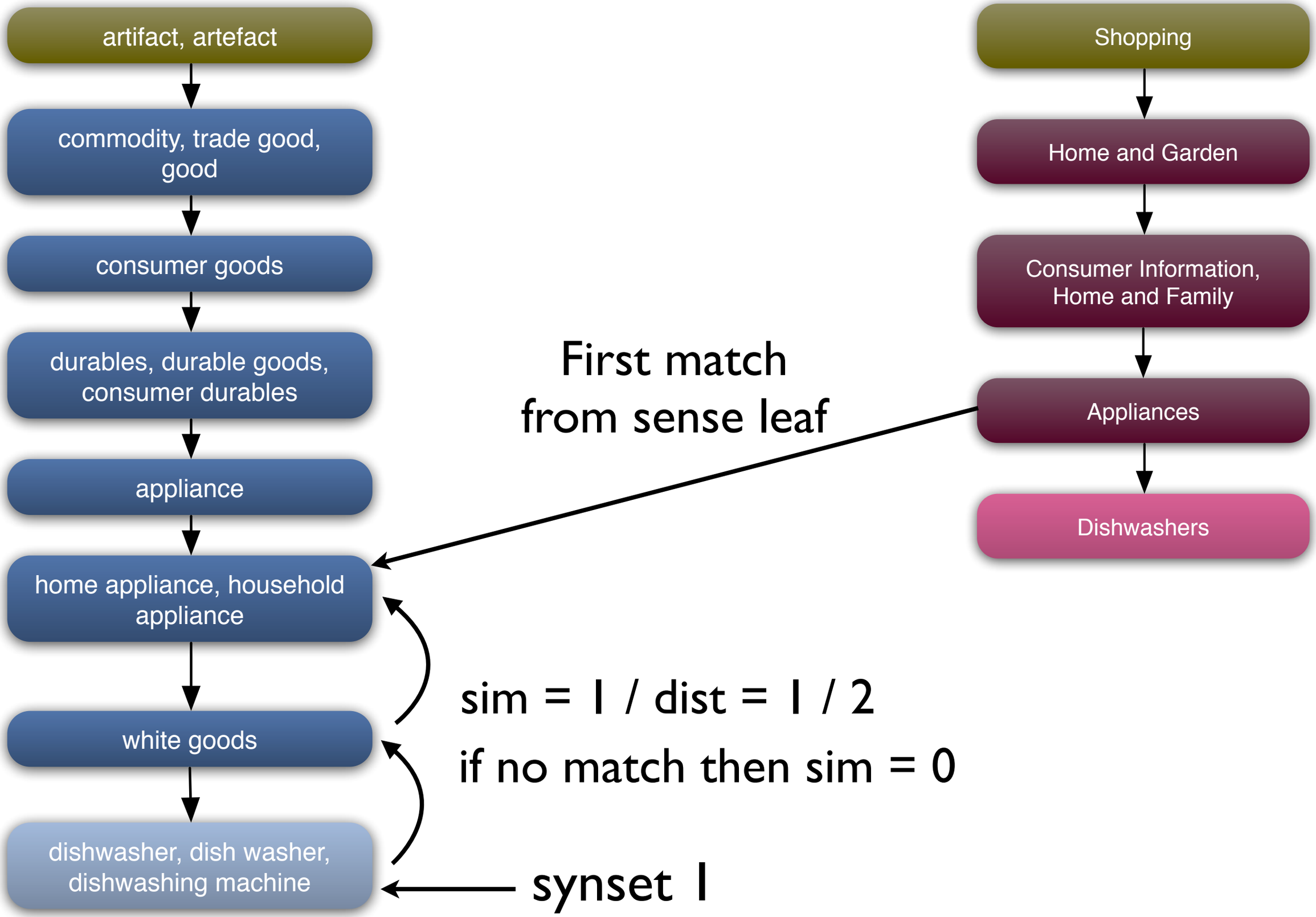


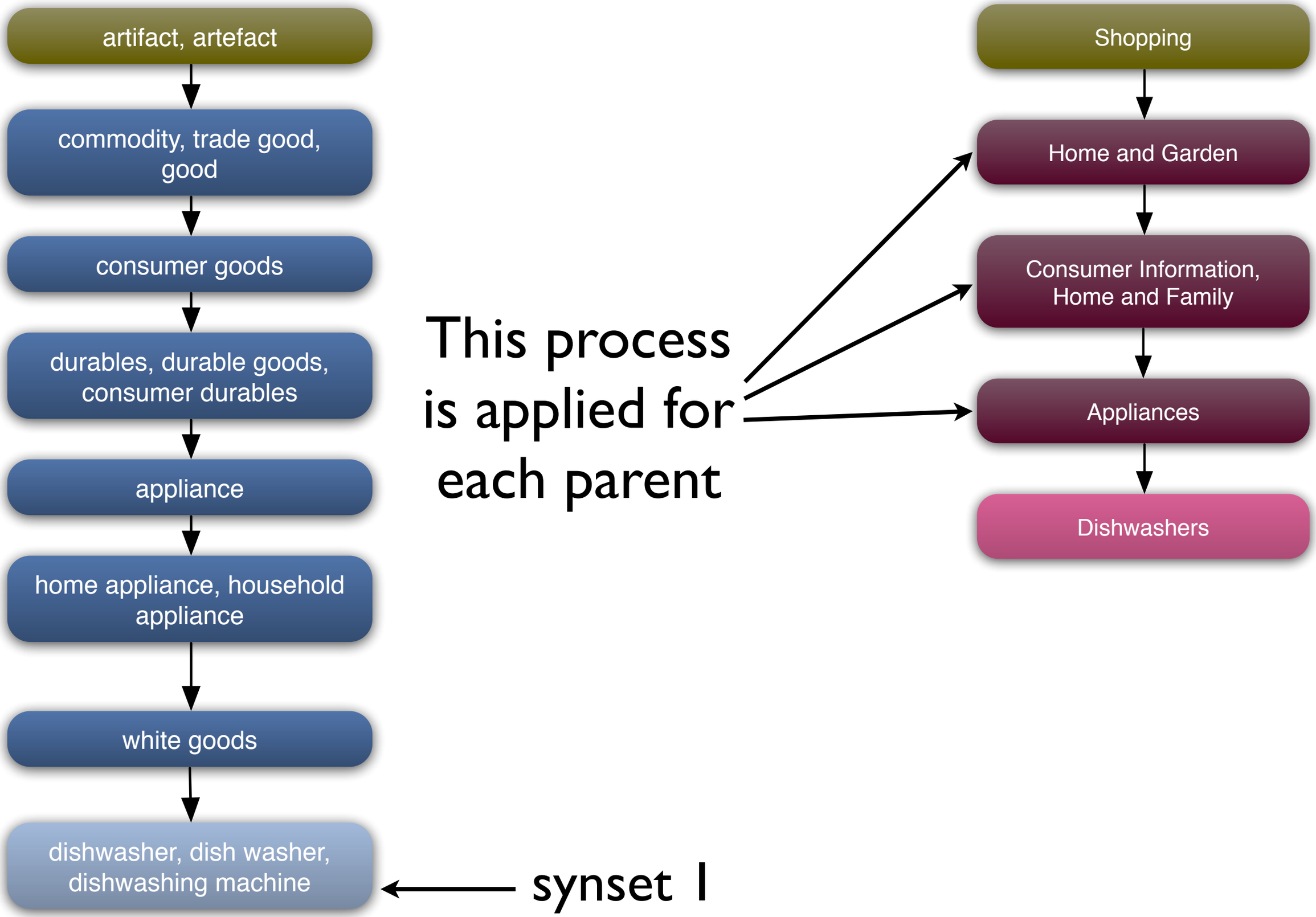
Compute sense score for each sense, highest is selected as correct sense

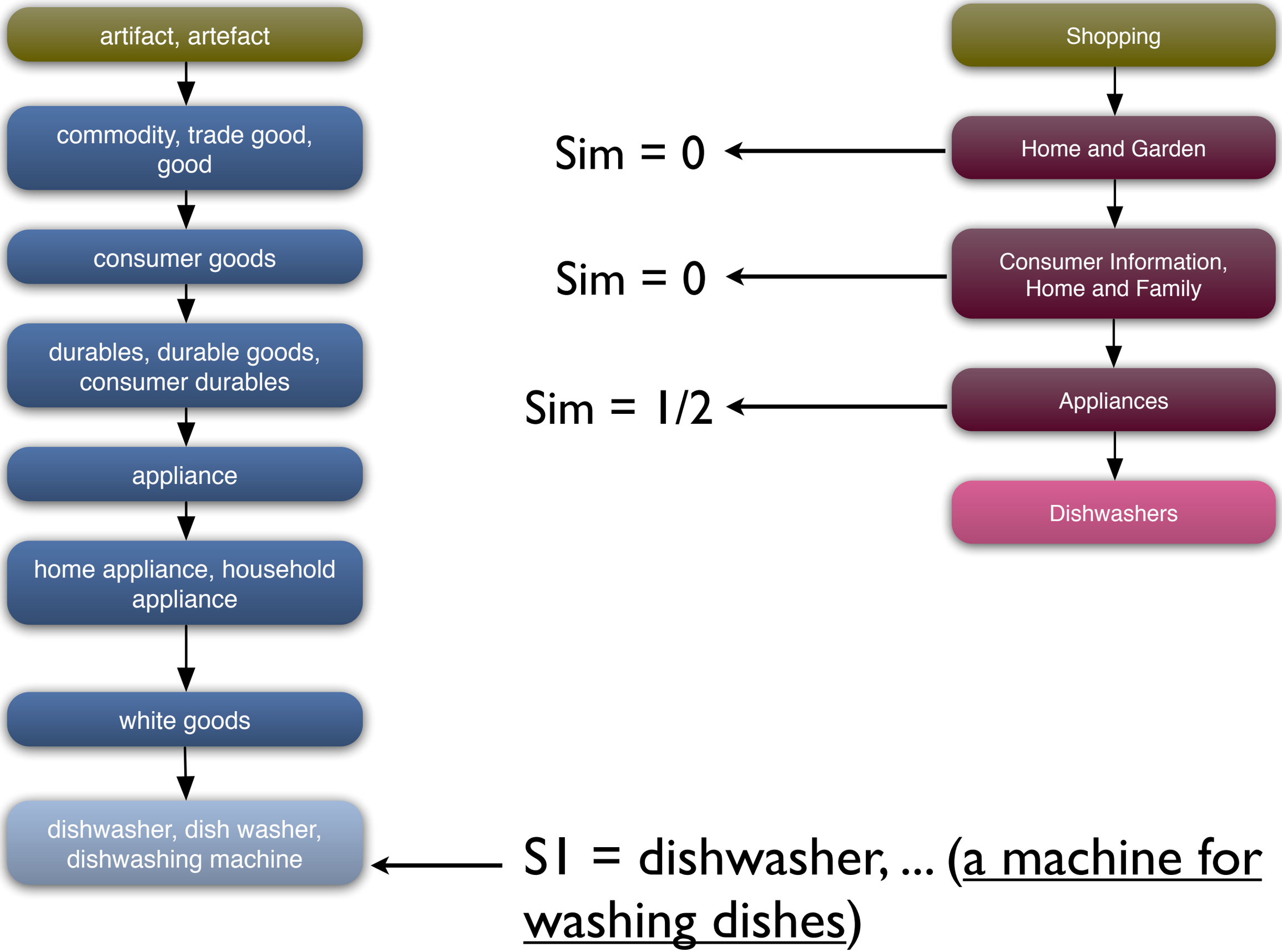

artifact, artefact
commodity, trade good, good
consumer goods
durables, durable goods, consumer durables
appliance
home appliance, household appliance
white goods
dishwasher, dish washer, dishwashing machine
synset 1
Shopping
Home and Garden
Consumer Information, Home and Family
Appliances
Dishwashers

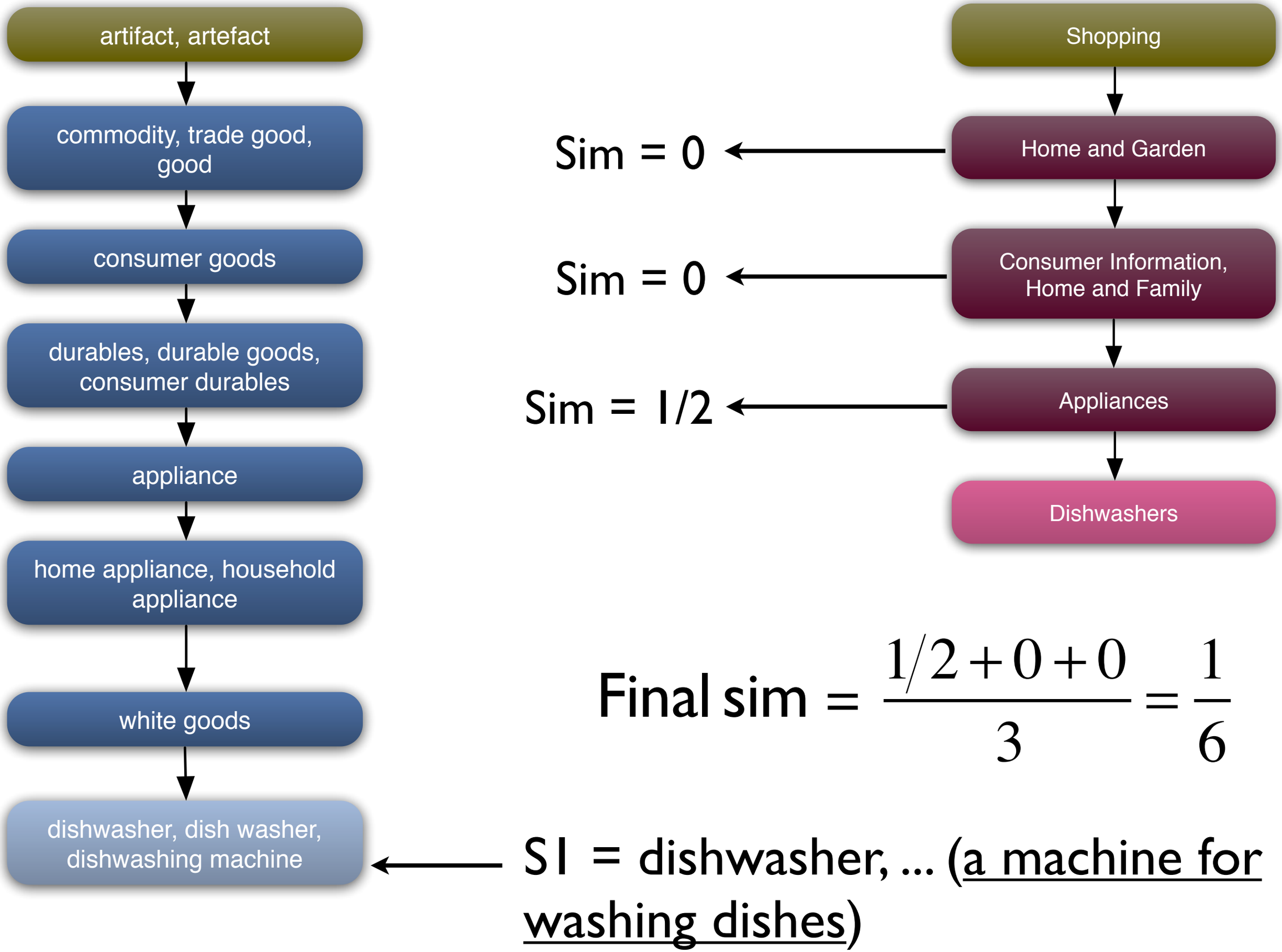

artifact, artefact
commodity, trade good, good
consumer goods
durables, durable goods, consumer durables
appliance
home appliance, household appliance
white goods
dishwasher, dish washer, dishwashing machine
Shopping
Home and Garden
Consumer Information, Home and Family
Appliances
Dishwashers
First match from sense leaf
synset 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
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


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 1
This process is applied for each parent
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
Sim = 0
Sim = 0
Sim = 1/2
S1 = dishwasher, ... (a machine for washing dishes)

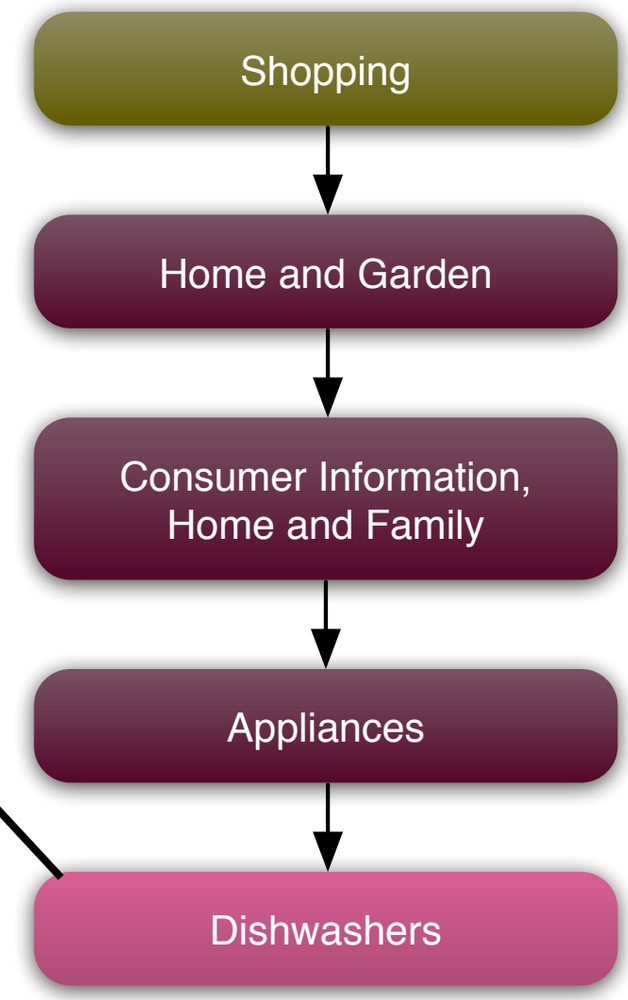

artifact, artefact
commodity, trade good, good
consumer goods
durables, durable goods, consumer durables
appliance
home appliance, household appliance
white goods
dishwasher, dish washer, dishwashing machine
Shopping
Home and Garden
Consumer Information, Home and Family
Appliances
Dishwashers
Sim = 0
Sim = 0
Sim = 1/2
Final sim = $\frac{1/2+0+0}{3} = \frac{1}{6}$
S1 = dishwasher, ... (a machine for washing dishes)

Word sense disambiguation

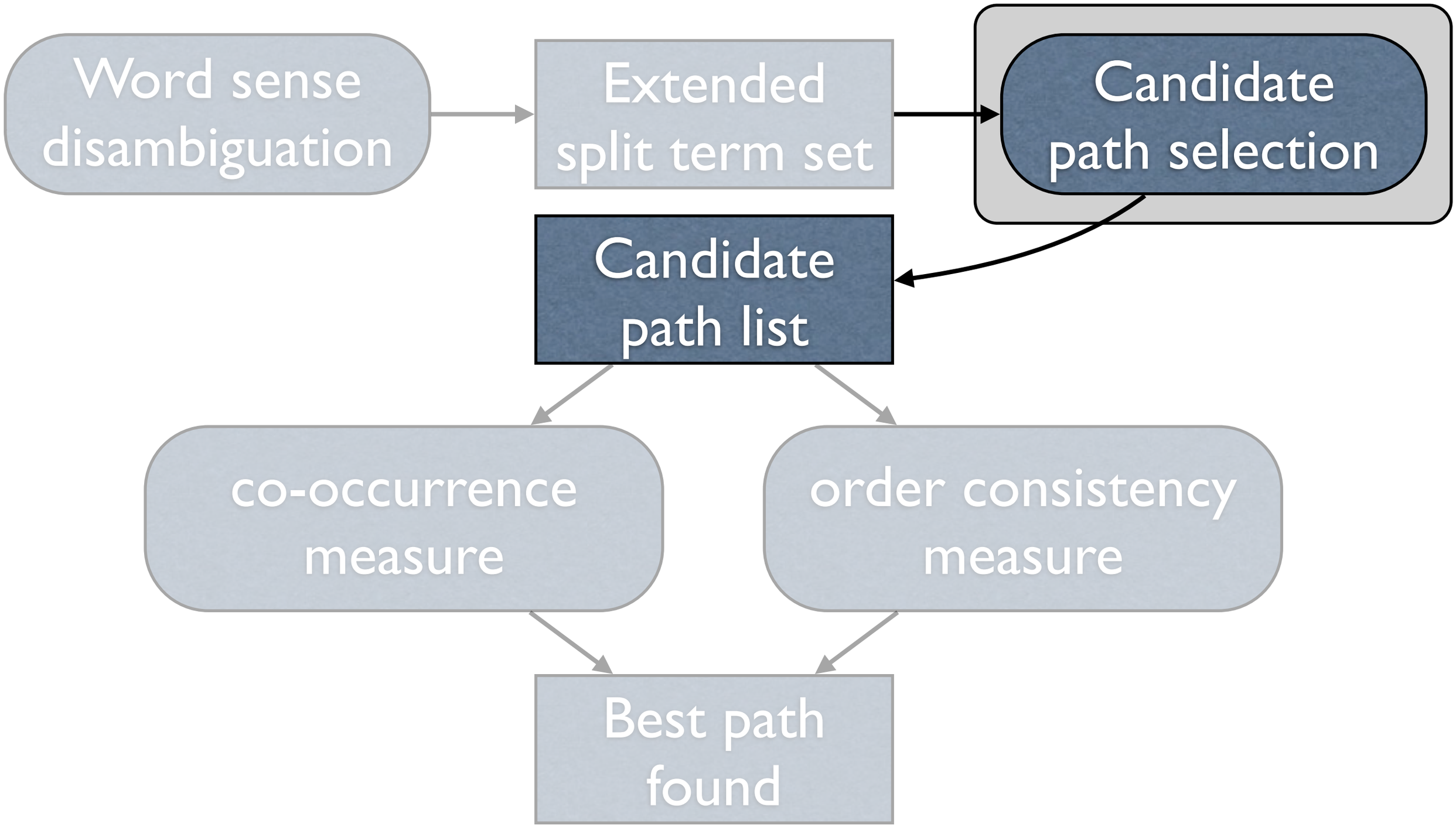
Terms of source node = {Dishwashers}

Extended Split Term Set = {*splitTermSet*}

{Dishwashers, dishwasher, dish washer, dishwashing machine}



CMAP - Part II

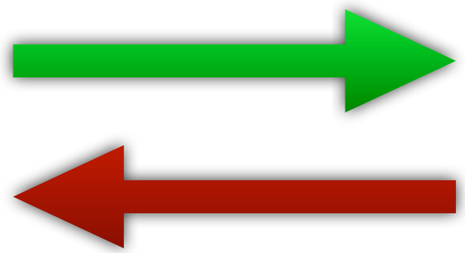

Word sense disambiguation
Extended split term set
Candidate path selection
Candidate path list
co-occurrence measure
order consistency measure
Best path found

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

Home and Garden ⇄ Home, Garden & Tools

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

Home and Garden

target

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

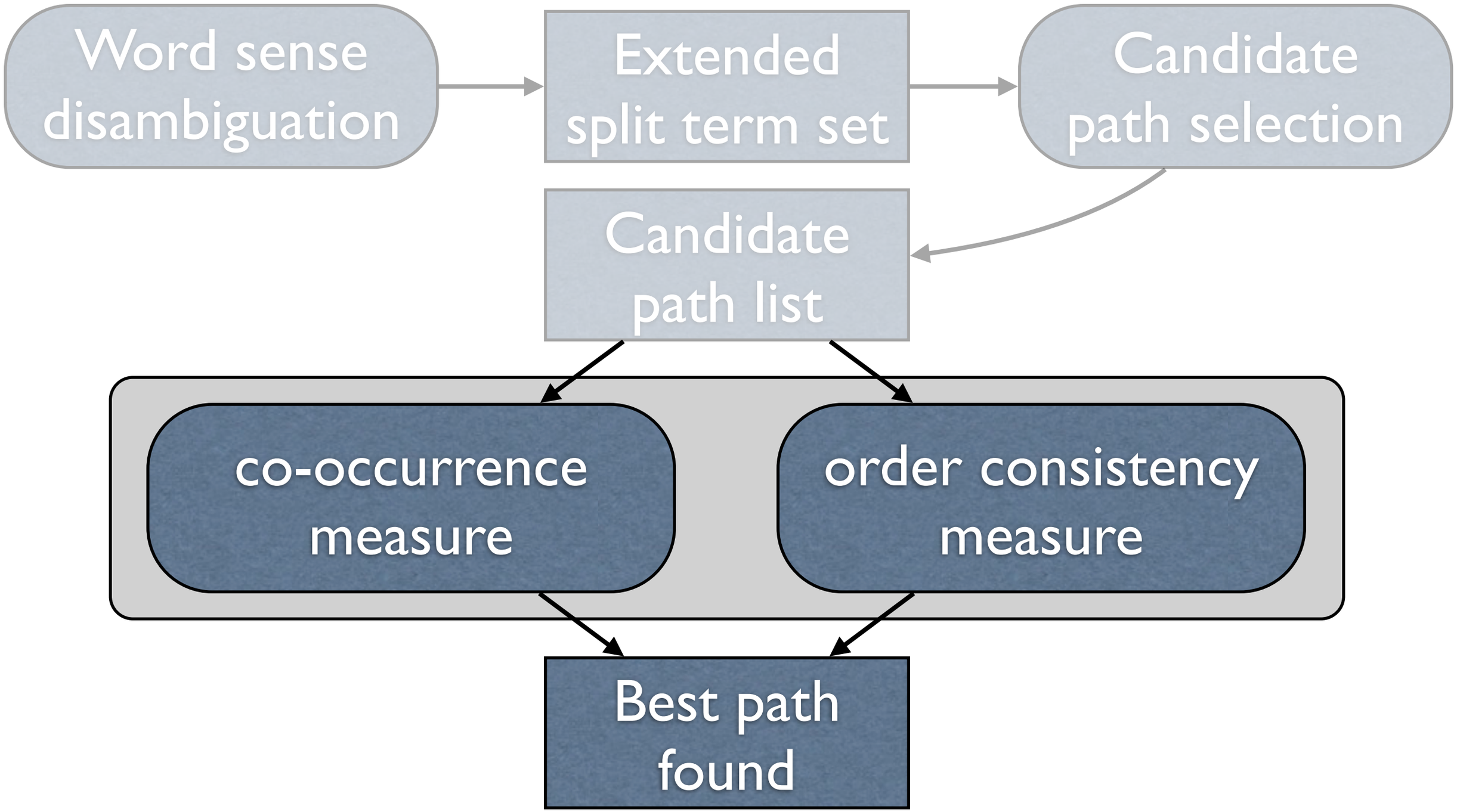
Home and Garden

source

Home, Garden & Tools

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

CMAP - 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

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

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

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

Co-occurrence

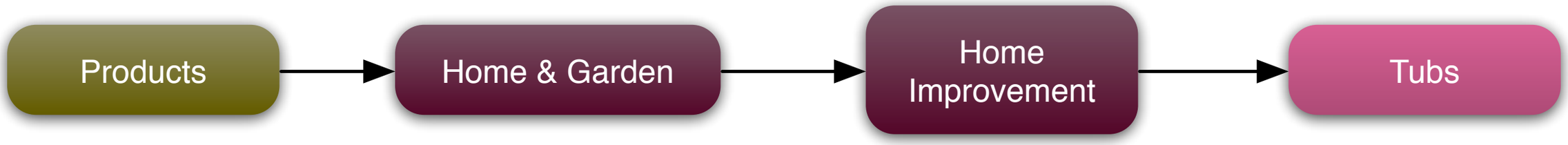
maximum
Jaccard similarity

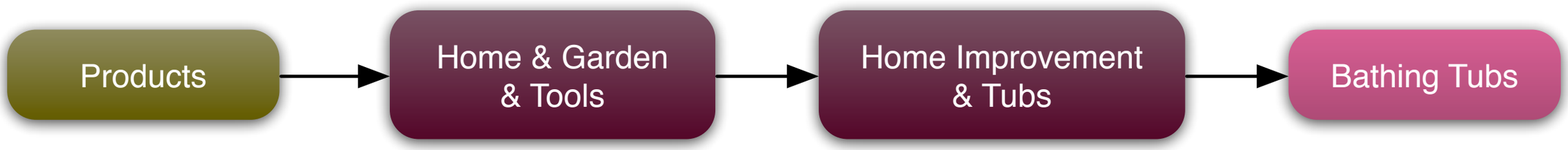
$$\text{coOccurrence}(P_{\text{src}}, P_{\text{targ}}) = \left(\sum_{t \in P_{\text{targ}}} \frac{\text{maxSim}(t, P_{\text{src}})}{|P_{\text{targ}}|} \right) \cdot \left(\sum_{t \in P_{\text{src}}} \frac{\text{maxSim}(t, P_{\text{targ}})}{|P_{\text{src}}|} \right)$$

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

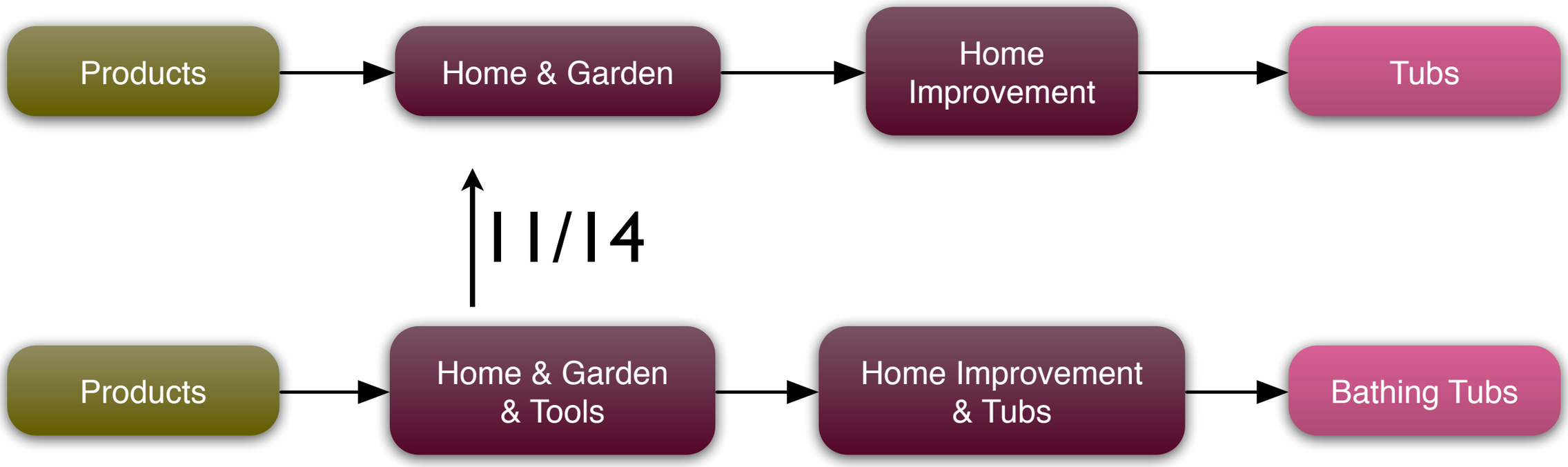




Candidate category path

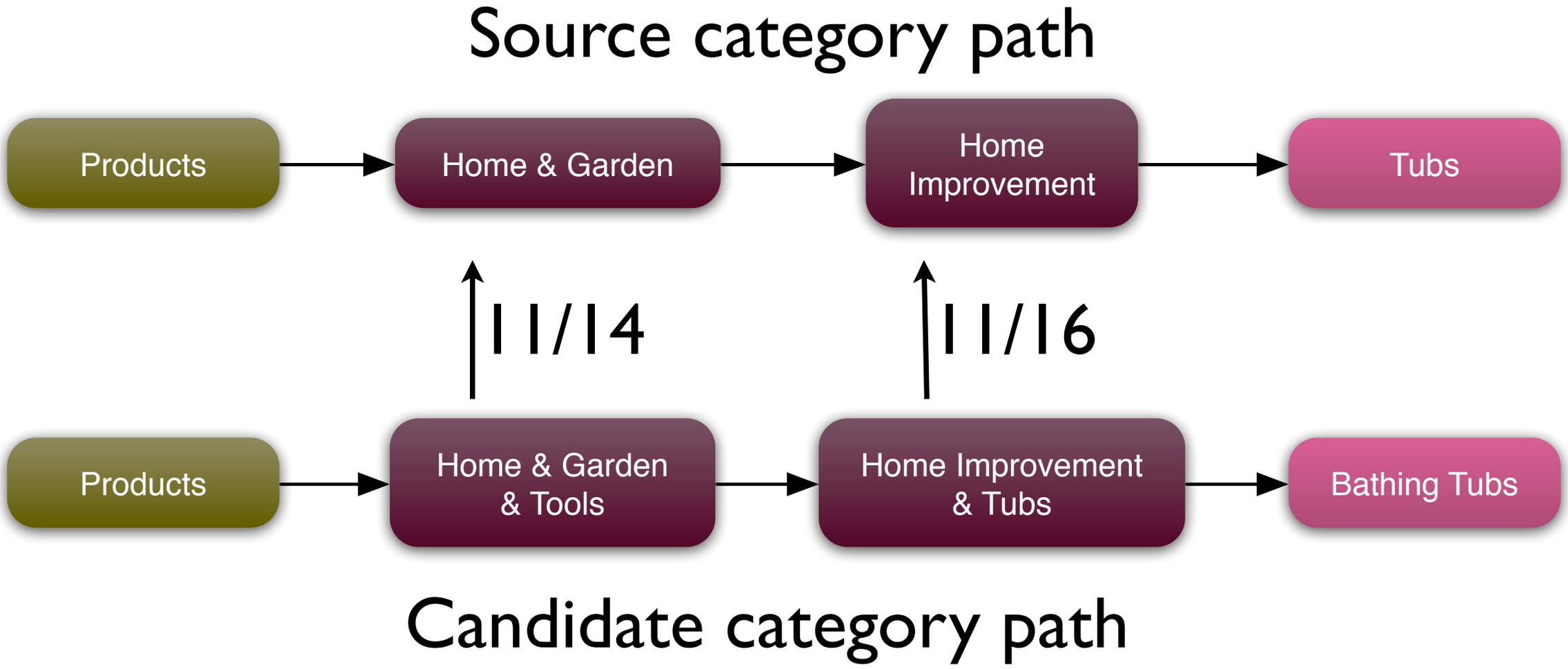
Co-occurrence

Source category path

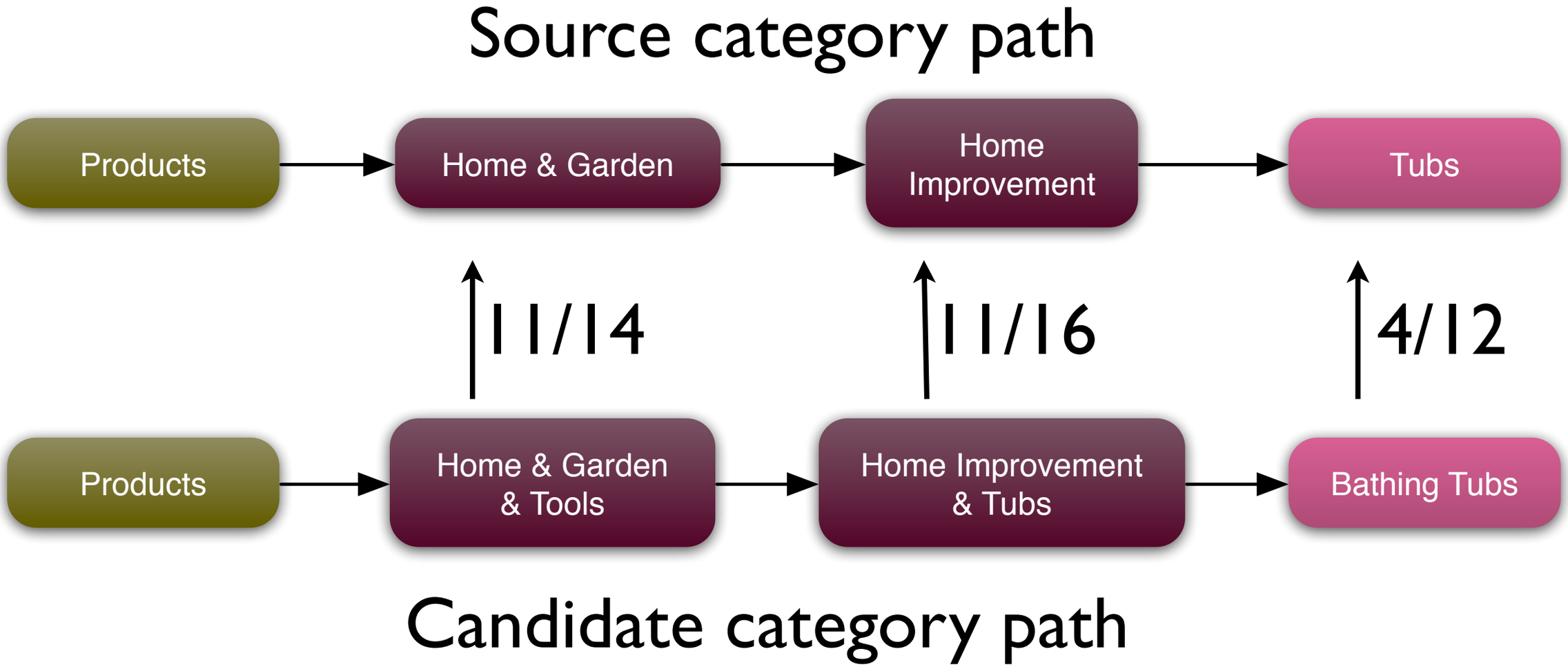


Candidate category path

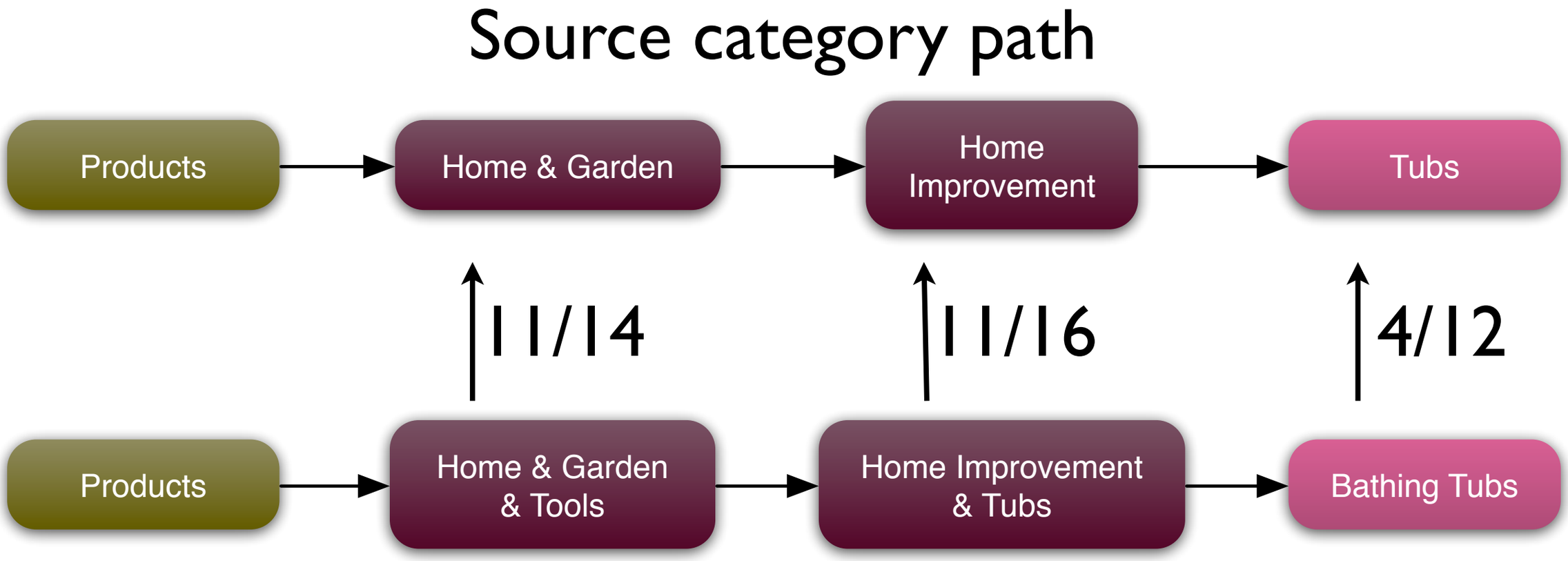
Co-occurrence



Co-occurrence



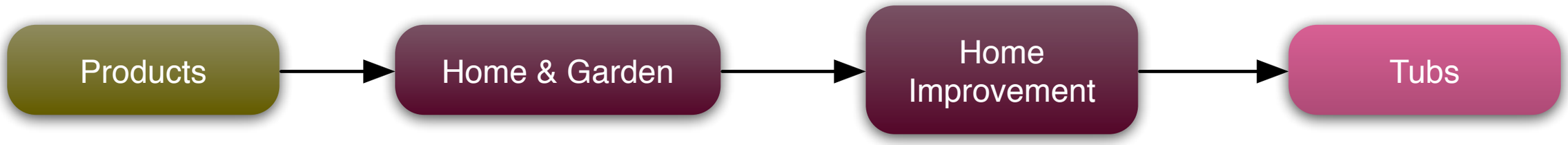
Co-occurrence

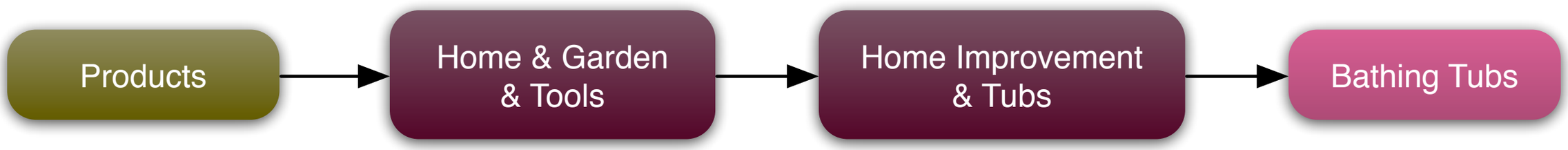


$$\text{left term of co-occ} = \frac{1}{3} \times \left(\frac{11}{14} + \frac{11}{16} + \frac{4}{12} \right) \approx 0.683$$

Co-occurrence

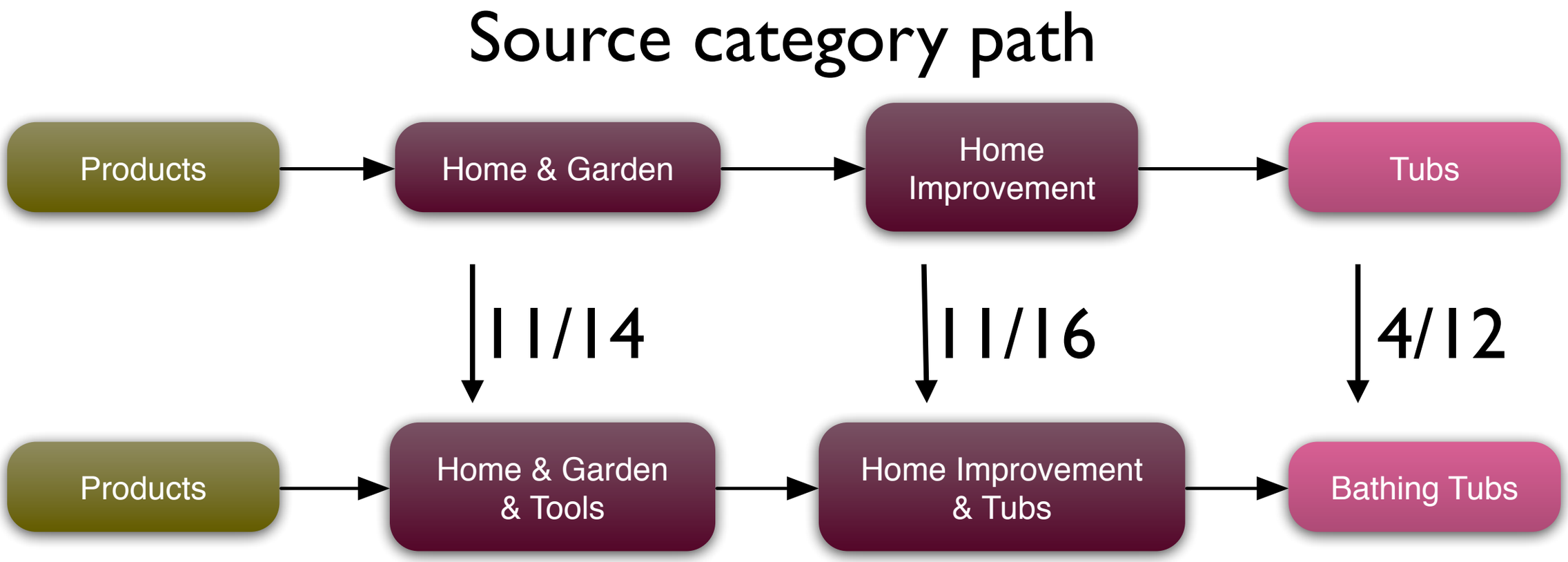
Source category path





Candidate category path

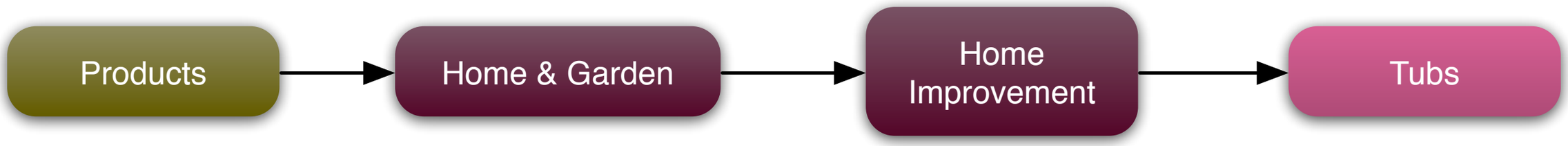
Co-occurrence

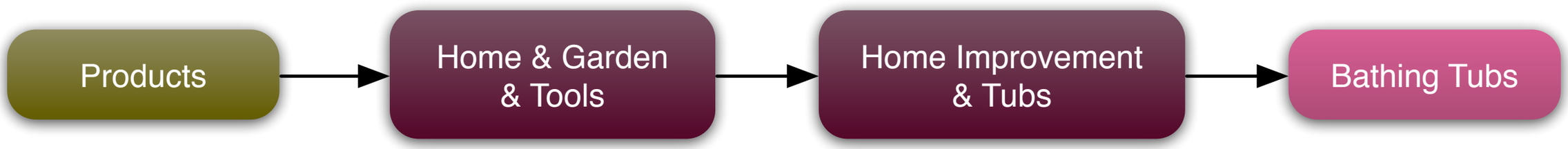


$$\text{right term of co-occ} = \frac{1}{3} \times \left(\frac{11}{14} + \frac{11}{16} + \frac{4}{12} \right) \approx 0.683$$

Co-occurrence

Source category path





Candidate category path

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

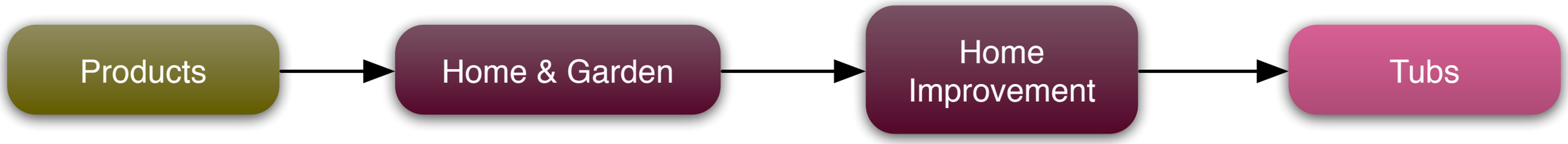
Order consistency

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

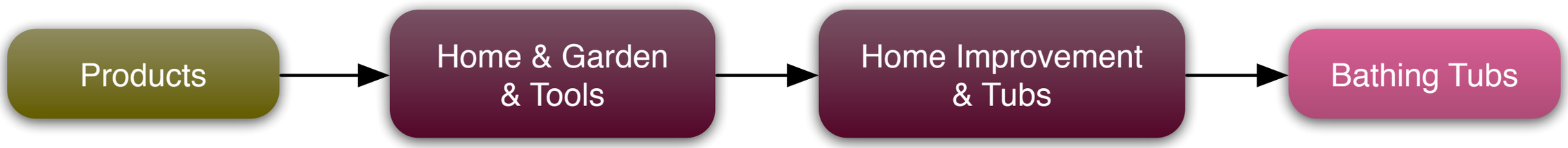
$$\begin{aligned}
\text{where } P_{\text{src}} &= \text{list of nodes from the current source path} \\
P_{\text{targ}} &= \text{list of nodes from a candidate target path} \\
C &= \text{common}(P_{\text{src}}, P_{\text{targ}}) \\
R &= \text{precedenceRelations}(C, P_{\text{src}})
\end{aligned}$$

Order consistency

Source category path

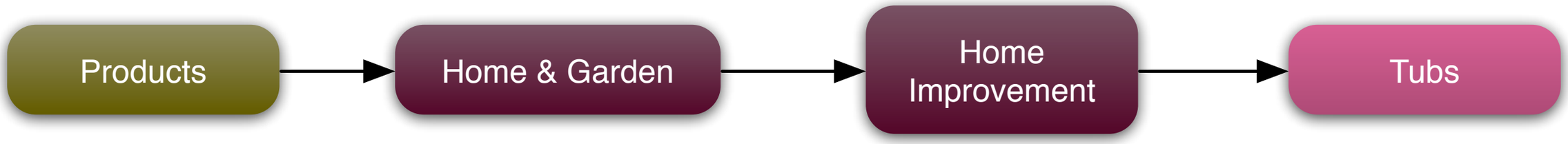


Candidate category path

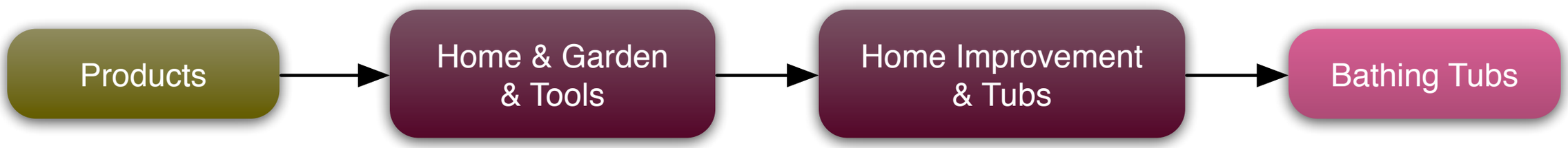


Order consistency

Source category path



Candidate category path



C = {
(‘Home & Garden’, ‘Home & Garden & Tools’),
(‘Home Improvement’, ‘Home Improvement & Tubs’),
(‘Tubs’, ‘Bathing Tubs’)
}

Order consistency

C = {
(‘Home & Garden’, ‘Home & Garden & Tools’),
(‘Home Improvement’, ‘Home Improvement & Tubs’),
(‘Tubs’, ‘Bathing Tubs’)
}

Order consistency

C = {
(‘Home & Garden’, ‘Home & Garden & Tools’),
(‘Home Improvement’, ‘Home Improvement & Tubs’),
(‘Tubs’, ‘Bathing Tubs’)
}

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

Order consistency

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

$$\begin{aligned}
\text{where } P_{\text{src}} &= \text{list of nodes from the current source path} \\
P_{\text{targ}} &= \text{list of nodes from a candidate target path} \\
C &= \text{common}(P_{\text{src}}, P_{\text{targ}}) \\
R &= \text{precedenceRelations}(C, P_{\text{src}})
\end{aligned}$$

Order consistency

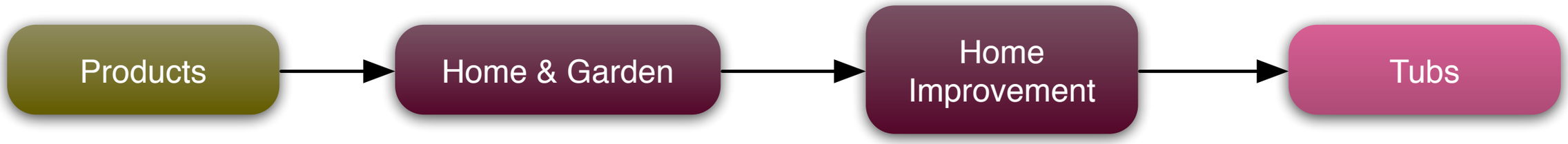
$$\text{orderConsistency}(P_{\text{src}}, P_{\text{targ}}) = \sum_{r \in R} \frac{\text{consistent}(r, P_{\text{targ}})}{\binom{\text{length}(C)}{2}}$$

$$\begin{aligned} \text{where } P_{\text{src}} &= \text{list of nodes from the current source path} \\ P_{\text{targ}} &= \text{list of nodes from a candidate target path} \\ C &= \text{common}(P_{\text{src}}, P_{\text{targ}}) \\ R &= \text{precedenceRelations}(C, P_{\text{src}}) \end{aligned}$$

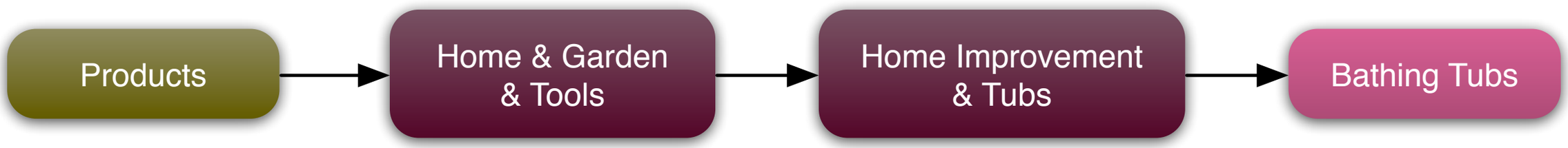
$$\mathit{consistent}\big((a,b), P_{\text{targ}}\big) = \begin{cases} 1, \text{ if } a \text{ precedes } b \text{ in } P_{\text{targ}} \\ 0, \text{ otherwise} \end{cases}$$

Order consistency

Source category path



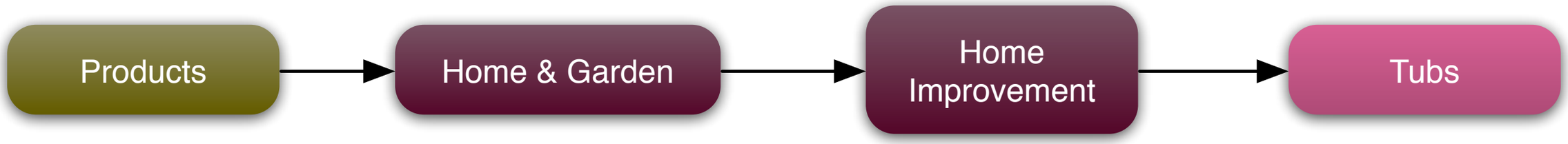
Candidate category path



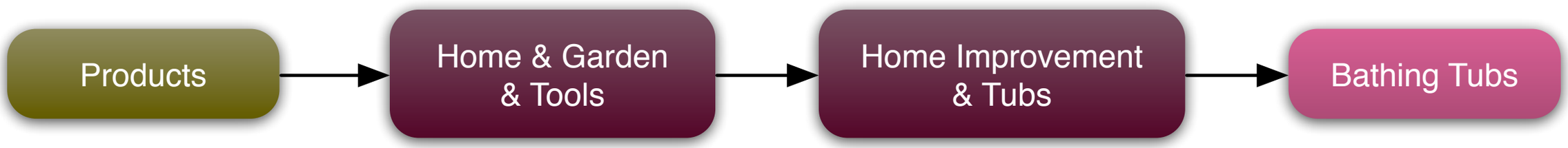
R = precedenceRelations(C, Psrc)= {
(‘Home & Garden & Tools’, ‘Home Improvement & Tubs’),
(‘Home Improvement & Tubs’, ‘Bathing Tubs’),
(‘Home & Garden & Tools’, ‘Bathing Tubs’)
}

Order consistency

Source category path



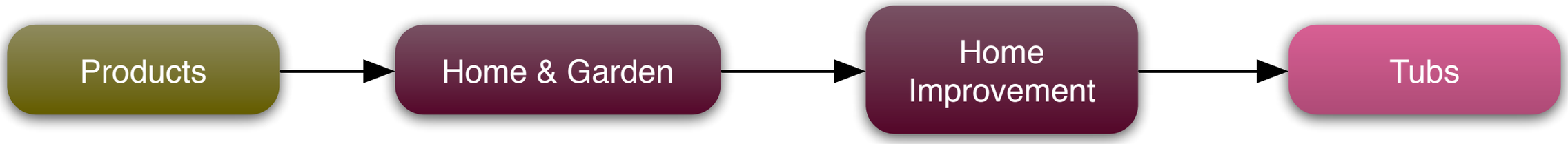
Candidate category path



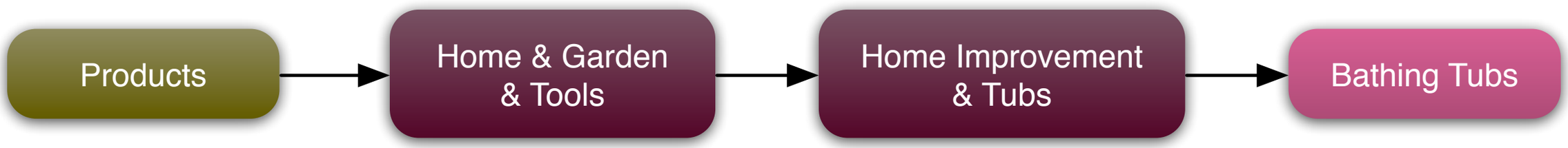
R = precedenceRelations(C, Psrc)= {
✔ (‘Home & Garden & Tools’, ‘Home Improvement & Tubs’),
(‘Home Improvement & Tubs’, ‘Bathing Tubs’),
(‘Home & Garden & Tools’, ‘Bathing Tubs’)
}

Order consistency

Source category path



Candidate category path



R = precedenceRelations(C, Psrc)= {

✔ (‘Home & Garden & Tools’, ‘Home Improvement & Tubs’),

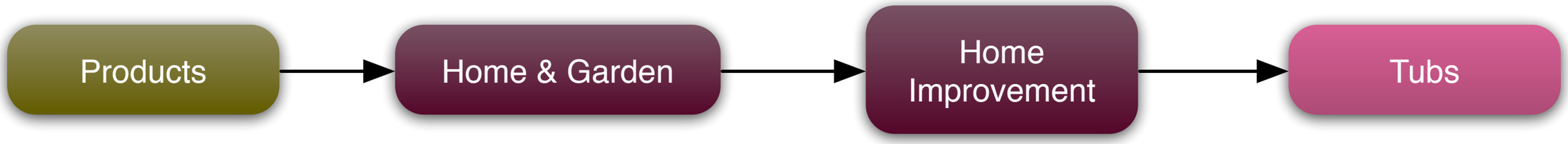
✔ (‘Home Improvement & Tubs’, ‘Bathing Tubs’),

(‘Home & Garden & Tools’, ‘Bathing Tubs’)

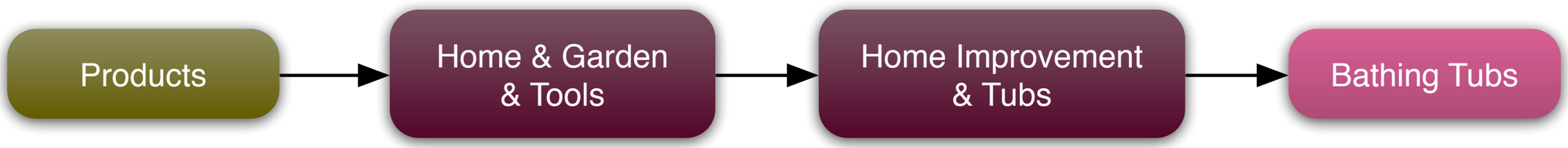
}

Order consistency

Source category path



Candidate category path



R = precedenceRelations(C, Psrc)= {

✔ ('Home & Garden & Tools', 'Home Improvement & Tubs'),

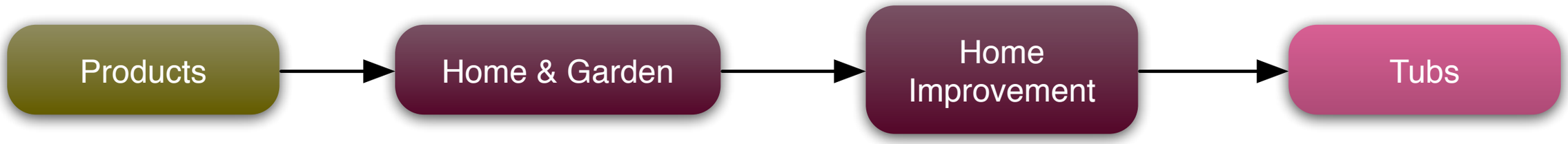
✔ ('Home Improvement & Tubs', 'Bathing Tubs'),

✔ ('Home & Garden & Tools', 'Bathing Tubs')

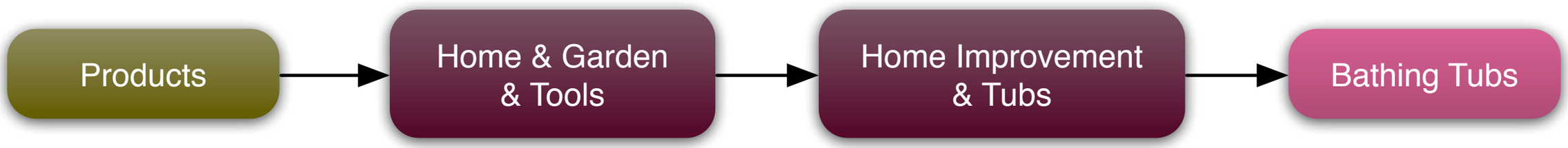
}

Order consistency

Source category path

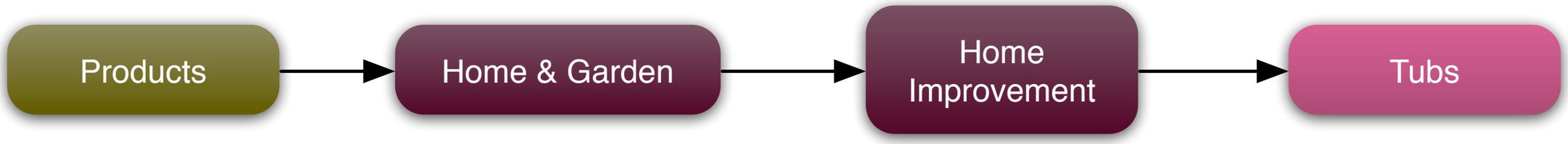


Candidate category path

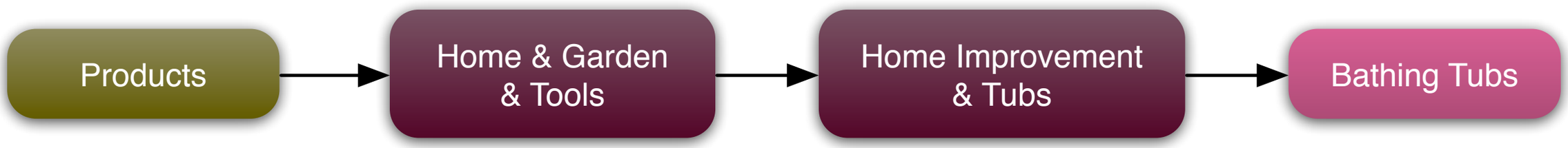


Order consistency

Source category path



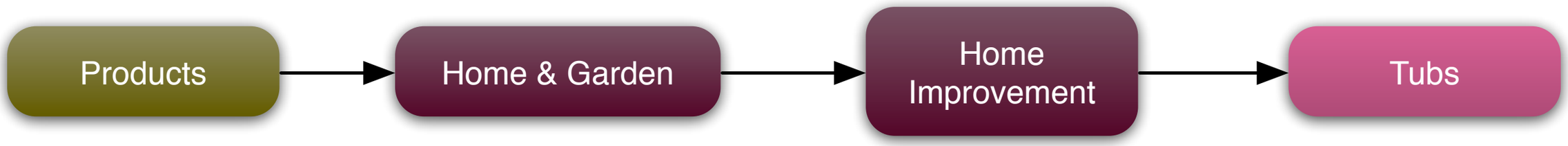
Candidate category path



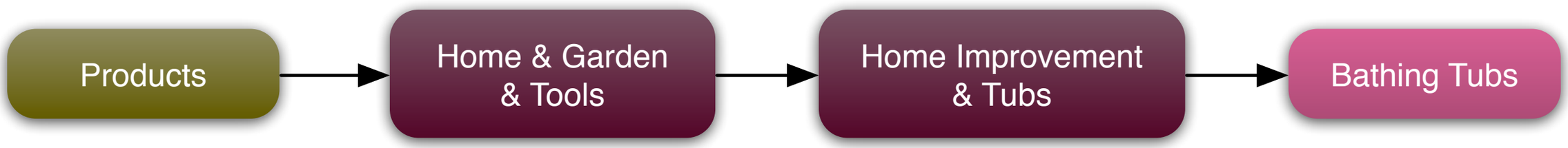
Order consistency = 3 / 3 = 1

Order consistency

Source category path



Candidate category path



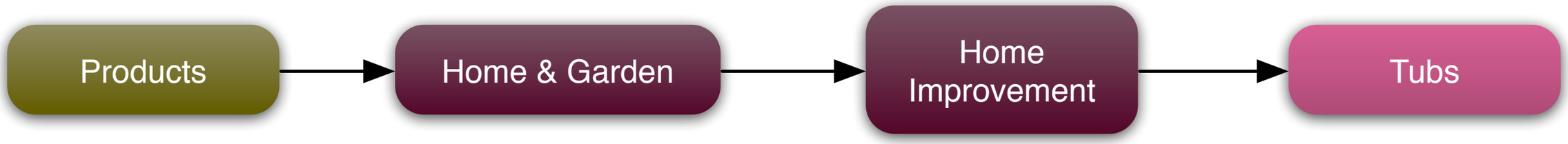
Order consistency = 3 / 3 = 1

Total number of

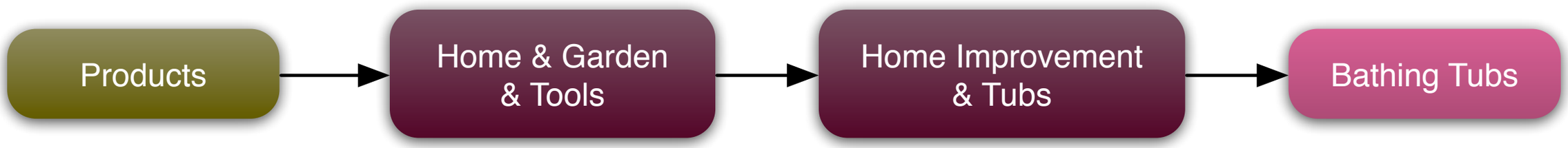
consistent precedence relations between common nodes

Order consistency

Source category path



Candidate category path



Total number of precedence relations between common nodes

Order consistency = 3 / 3 = 1

Total number of **consistent** precedence relations between common nodes

Finding optimal path

$$\begin{aligned}\text{overallSimilarity}(P_{\text{src}}, P_{\text{targ}}) &= (\text{orderConsistency}(P_{\text{src}}, P_{\text{targ}}) + t) \\ &\quad \cdot \text{coOccurrence}(P_{\text{src}}, P_{\text{targ}}) \\ \text{where } P_{\text{src}} &= \text{list of nodes from the current source path} \\ P_{\text{targ}} &= \text{list of nodes from a candidate target path} \\ t &= \text{the similarity threshold}\end{aligned}$$

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	19.82%	10.62%	13.50%	0.47 sec
Park & Kim	37.89%	17.93%	24.15%	4.99 sec
SCHEMA	41.82%	26.03%	31.80%	5.82 sec

Questions?