Probase

Haixun Wang
Microsoft Research Asia

Short Text

- Search
- Ad keywords
- Anchor text
- Document Title
- Caption
- Question

The big question

- How does the mind get so much out of so little?

- Our minds build rich models of the world and make strong generalizations from input data that is *sparse, noisy, and ambiguous* – in many ways far too limited to support the inferences we make.

- How do we do it?

Science **331**, 1279 (2011);

How to Grow a Mind: Statistics, Structure, and Abstraction

Joshua B. Tenenbaum,[1*] Charles Kemp,[2] Thomas L. Griffiths,[3] Noah D. Goodman[4]

MIT

CMU

Berkeley

Stanford

If the mind goes beyond the data given, another source of information must make up the difference.

h1: all and only horses

h2: all horses except Clydesdales

h3: all animals

likelihood prior

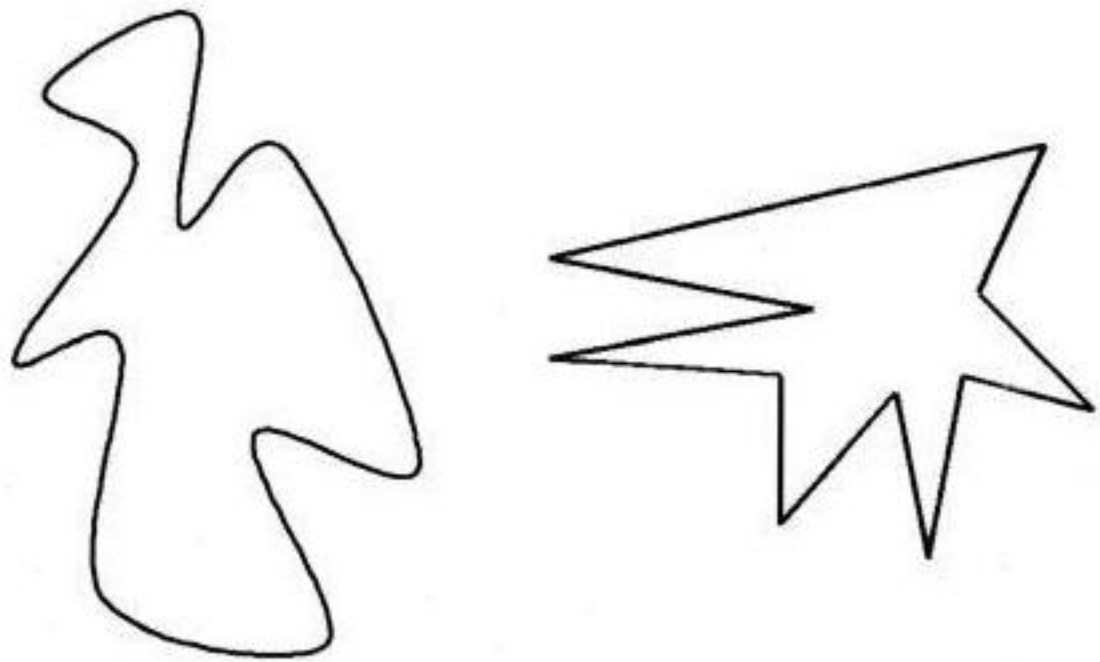
- $P(h|d) \propto P(d|h)P(h)$

h1
h2

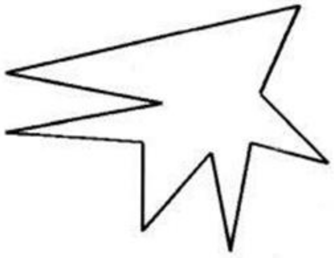
h1
h3

h1: all and only horses
h2: all horses except Clydesdales
h3: all animals

Which is “kiki” and which is “bouba”?



$\backslash' \boldsymbol{k\bar{e}k\bar{e}}$



sound

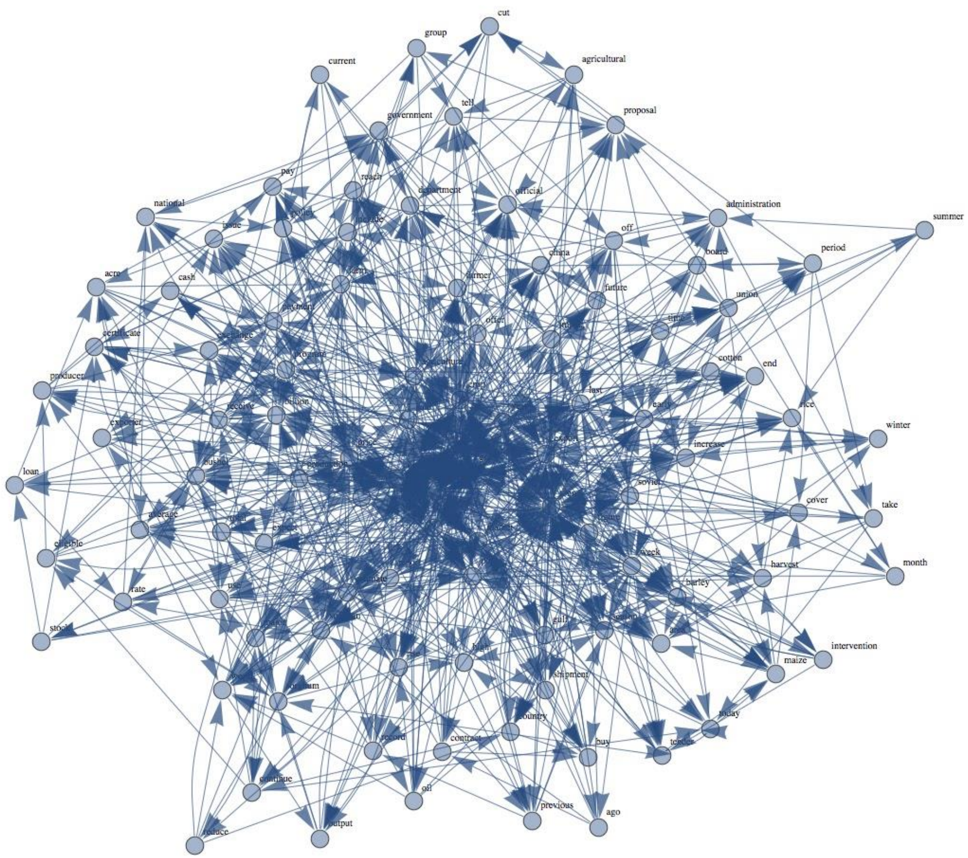
shape

zigzaggedness

Another example

Pablo Picasso	**25 Oct 1881**	**Spanish**

Probase: a semantic network for text understanding



Concepts Entities isA isPropertyOf Co-occurrence

isA Extraction

- Hearst pattern

  NP such as NP, NP, ..., and|or NP

  such NP as NP,* or|and NP

  NP, NP*, or other NP

  NP, NP*, and other NP

  NP, including NP,* or | and NP

  NP, especially NP,* or|and NP

- *domestic animals* such as *cats* and *dogs* …

- animals other than *cats* such as *dogs* …

- … is a … pattern

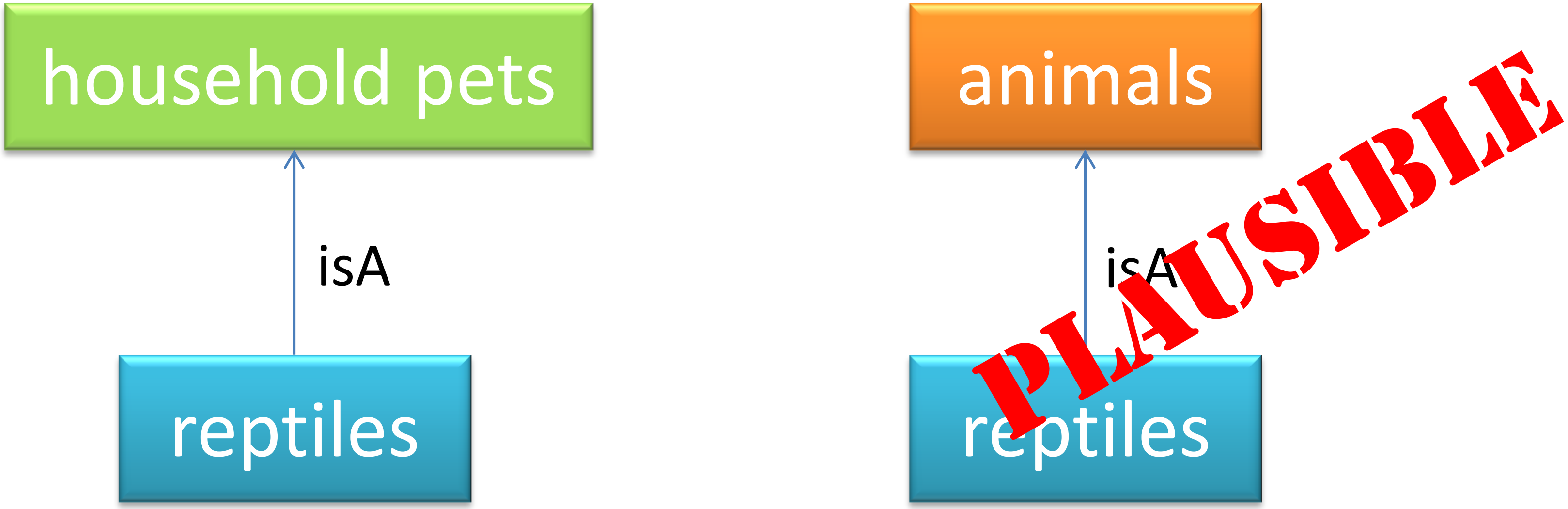
  NP is a/an/the NP

- *China* is a *developing country.*

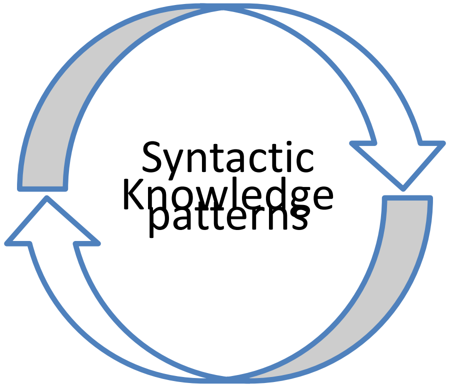
- *Life* is a box of *chocolate*.

… **animals** other than **cats** such as **dogs** …

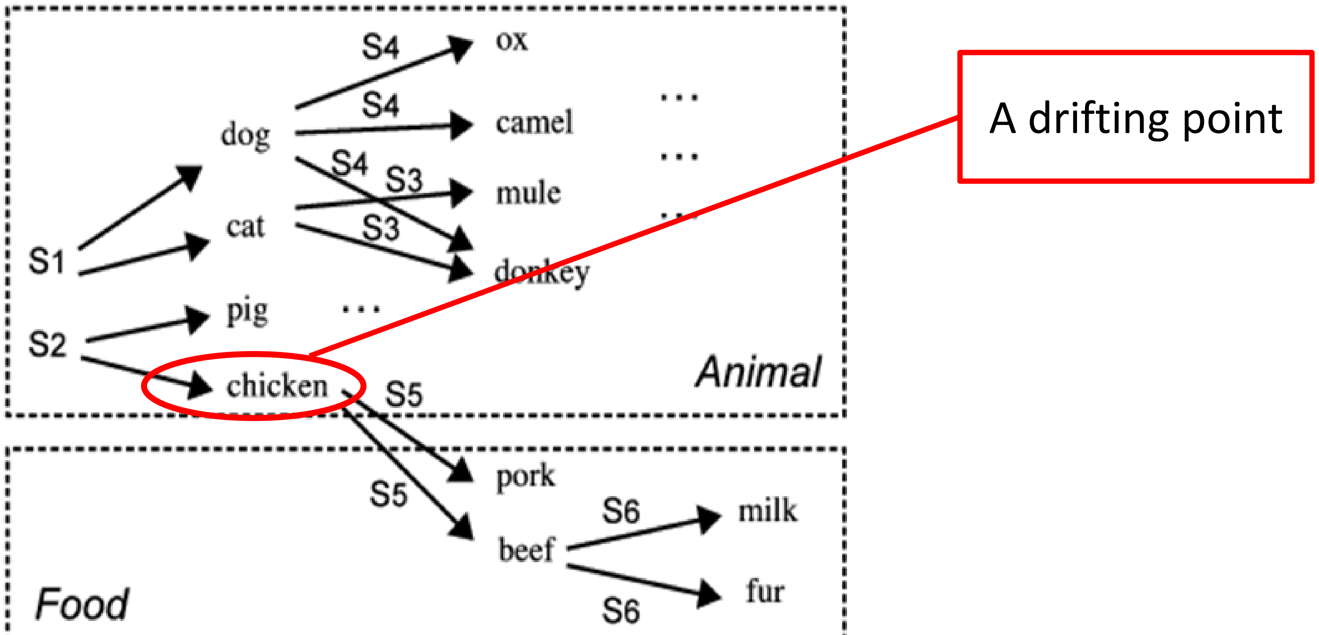
… **household pets** other than **animals** such as **reptiles**, aquarium fish …



Iterative Information Extraction



Semantic Drifts



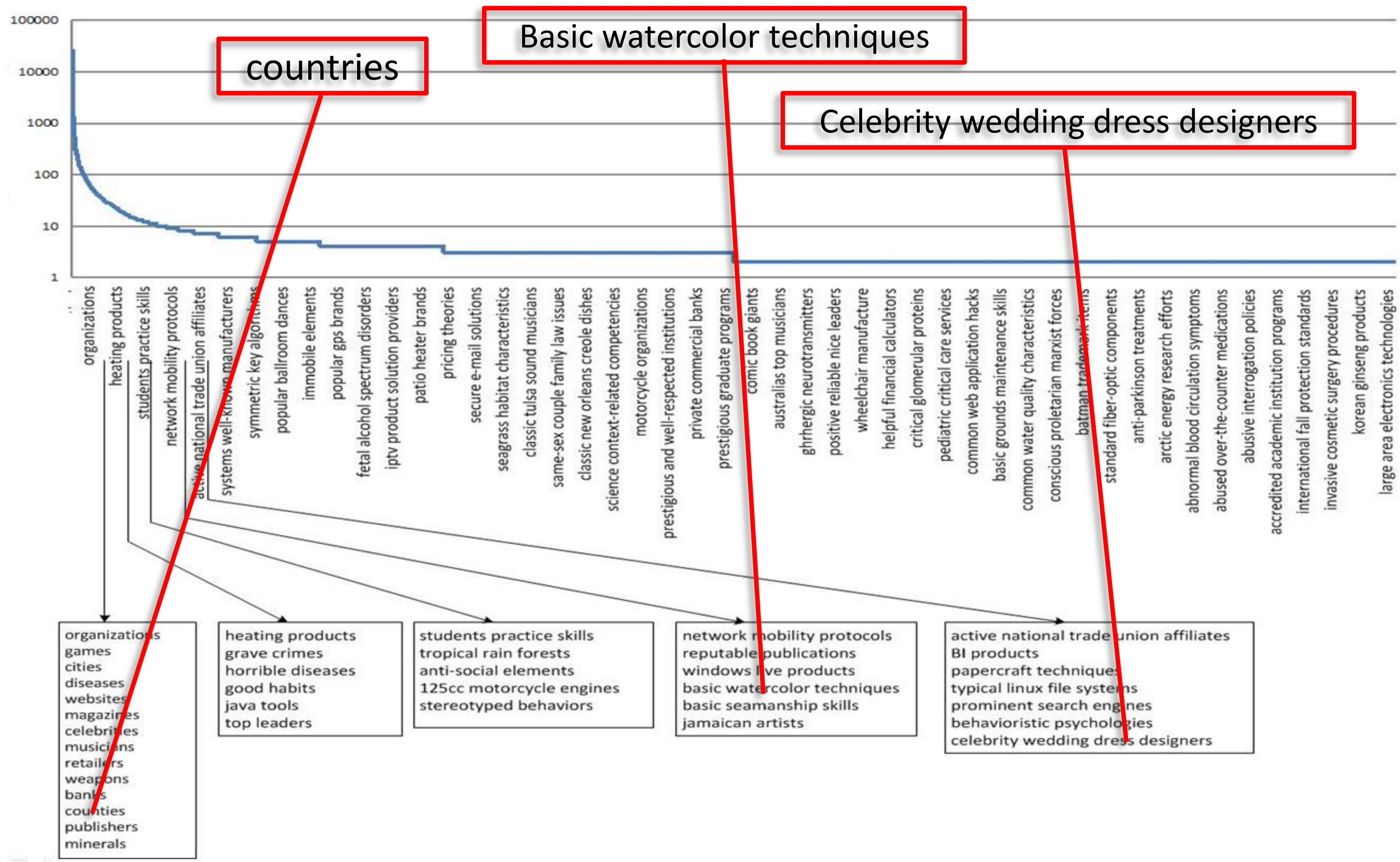
S1="Animals **such as** dogs and cats, grow fast."
S2="Land animals **such as** chicken and pigs – all of which live on land"

S3="Postures are often named after animals, **such as** mule, donkey and cat."
S4="... innkeeper, angels, and animals **such as** ox, camels, donkeys and dog"
S5="Common food from animals such as pork, beef and chicken"

S6="Products from animals **such as** fur, milk and beef are given to families..."

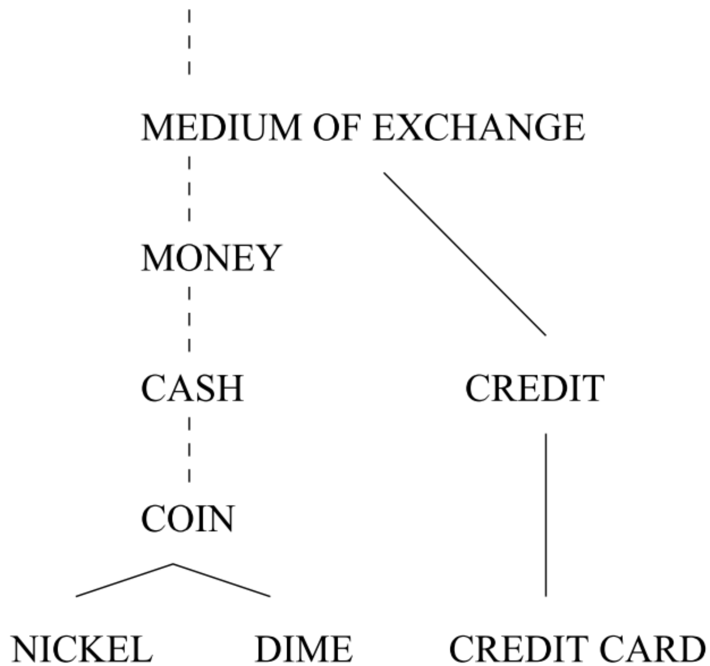
10%-20% Precision Improvement

Probase Concepts (2.7 million+)

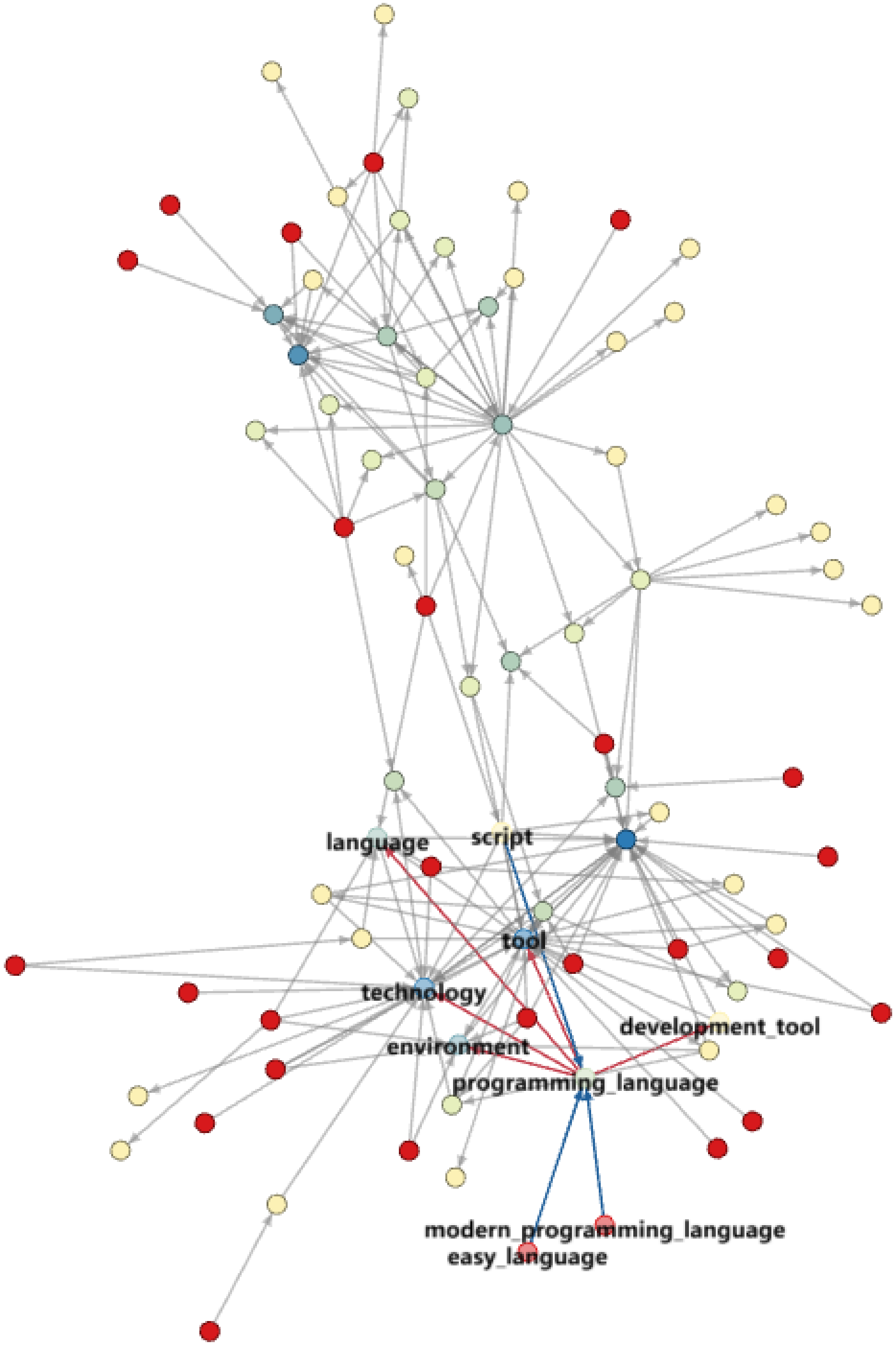


Probase isA error rate: <%1 @1 and <10% for random pair

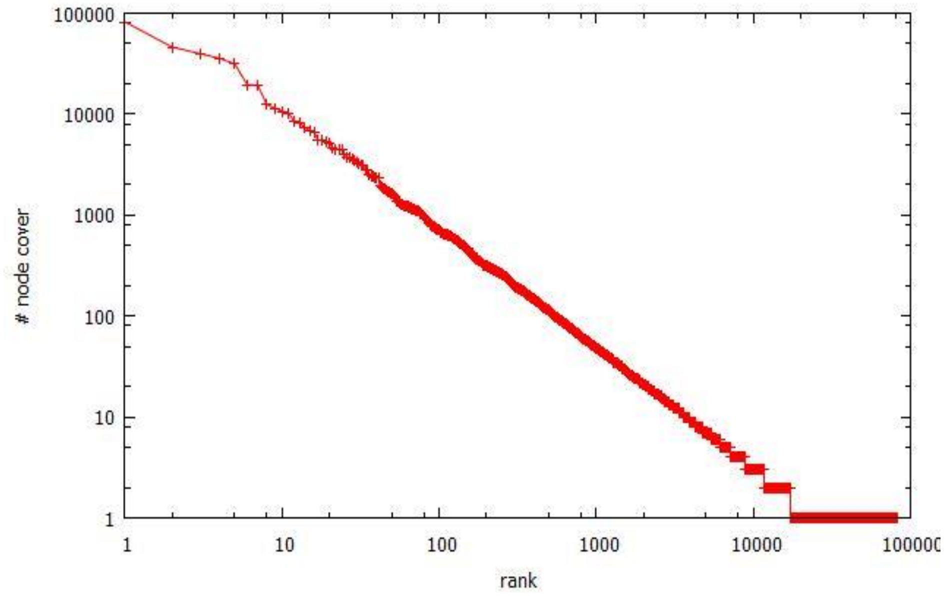
A traditional taxonomy



“python”


language
script
tool
technology
development_tool
environment
programming_language
modern_programming_language
easy_language

of descendants (WordNet)



Transitivity does not always hold

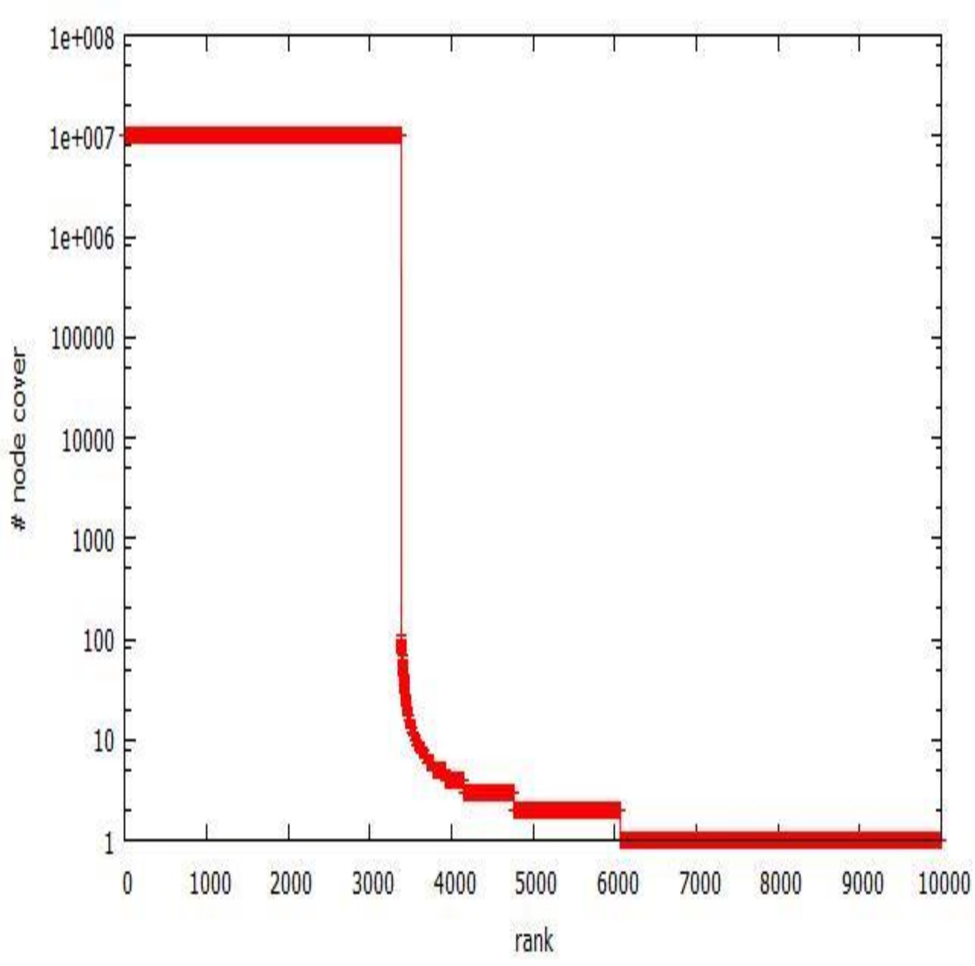
furniture

chair

plastic material

film

of descendants (early version of Probase)



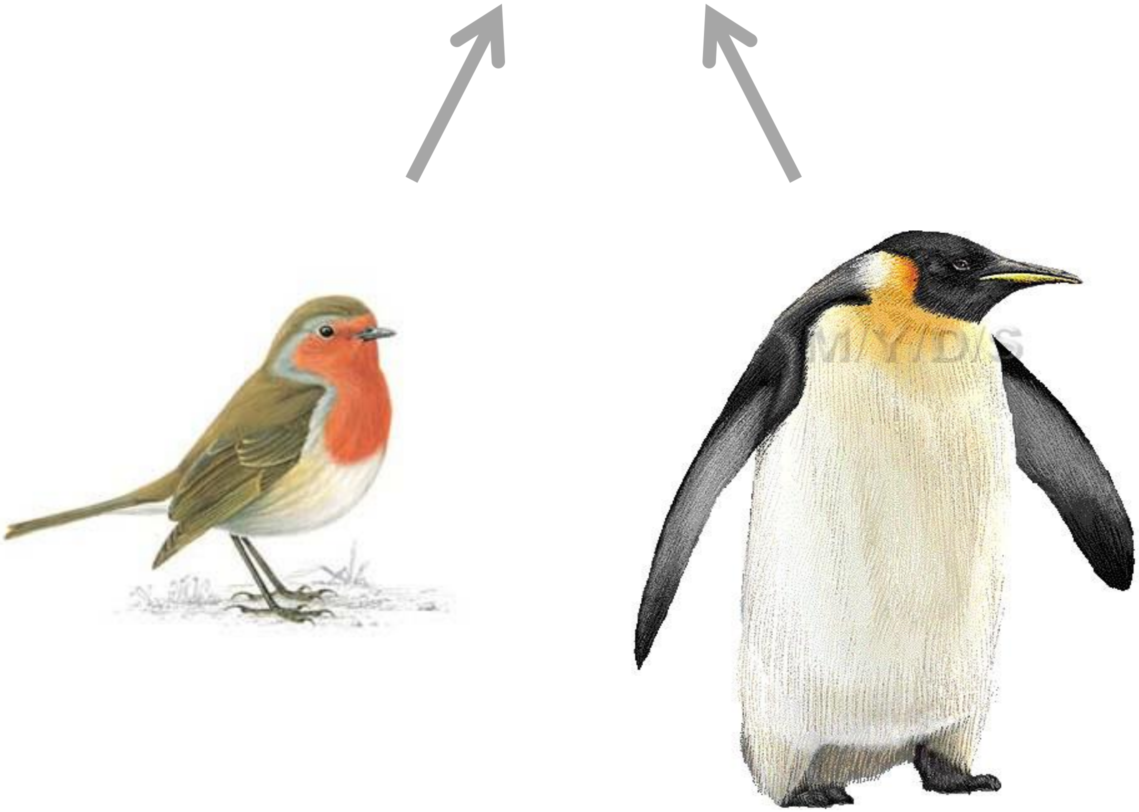
Probase Scores

- Typicality
- Vagueness
- Representativeness
- Ambiguity
- Similarity

foundation for inferencing

Typicality

bird



$$P(e|c) = \frac{n(c,e) + \alpha}{\sum_{e_i \in c} n(c,e_i) + \alpha N}$$

$$P(c|e) = \frac{n(c,e) + \alpha}{\sum_{e \in c_i} n(c_i,e) + \alpha N}$$

"robin" is a more *typical* bird than a "penguin" ➡ $p(robin|bird) > p(penguin|bird)$

Representativeness (basic level of categorization)

software company

$$\max_{c} \; p(c|e) \cdot p(e|c)$$

company ... **?** ... **largest OS vendor**

high typicality p(c|e)

high typicality p(e|c)

Microsoft

Vagueness

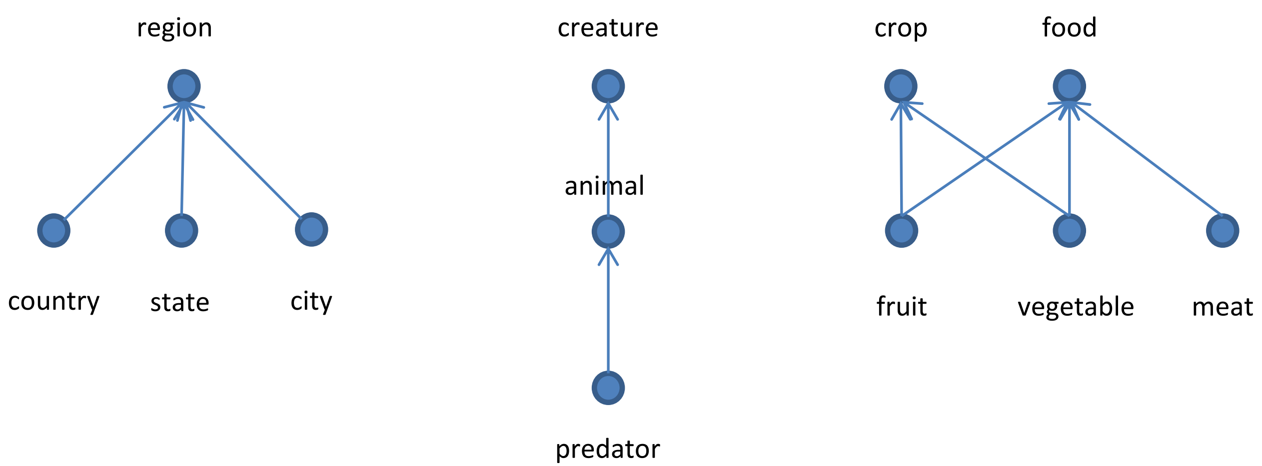
key players
factors
items
things
reasons
…

$$V(C) = \frac{|\{e_i | P(C|e_i) \geq c, \forall e_i \in C\}|}{N(C)}$$

(Do people whom you regard highly regard you highly?)

Ambiguity

- Probase defines 3 levels of ambiguity
  - Level 0 (1 sense): apple juice
  - Level 1 (2 or more related senses): Google
  - Level 2 (2 or more senses): python
- Concepts form clusters, clusters form senses (through isa relation)



Similarity

- microsoft, ibm ➡ 0.933
- google, apple ➡ 0.378 ??

$$sim(t_1, t_2) = \max_{x,y}\ cosine\ (c_x(t_1), c_y(t_2))$$

Applications

- Query Understanding
  - Head/Modifier/Constraint detection

- …

- SRL (semantic role labeling) with FrameNet
  - e.g. Tom broke the window.

agent

patient

Example: FrameNet

Frame: Apply_heat

FE1 **FE2** **FE3** **FE4**

She was FRYING eggs and bacon and mushrooms on a camp stove in Woolley 's billet .

Concept	P(c|FE)
heat source	0.19
Large metal	0.04
Kitchen appliance	0.02

Instance	P(w|FE)
Stove	0.00019
Radiator*	0.00015
Oven	0.00015
Grill*	0.00014
Heater*	0.00013
Fireplace*	0.00013
Lamp*	0.00013
Hair dryer*	0.00012
Candle*	0.00012

Example: Head and Modifier Detection

- toy kid
- cover iphone → (accessory, smart phone)
- seattle hotel jobs

When concepts are too specific

- Example:

  mobile windows operating system / head
  large and inferential software vendor / modifier

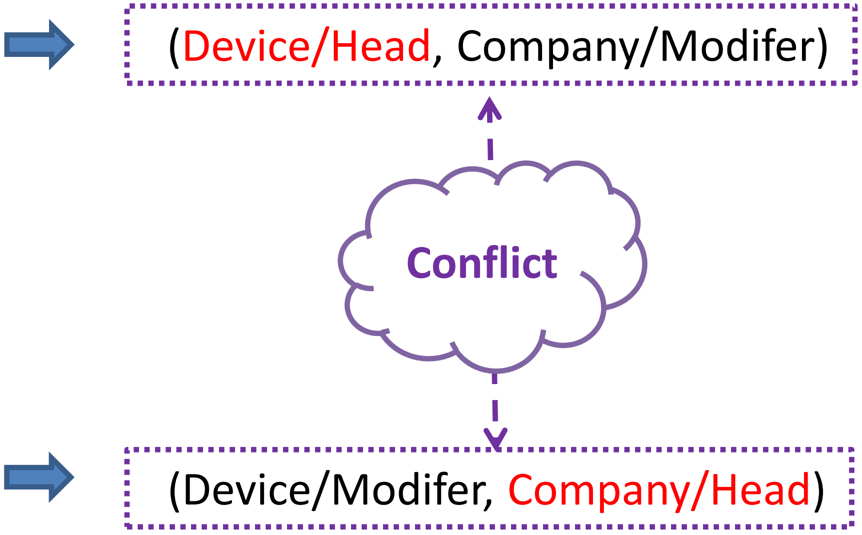
- No generalization power

- $million^2$ patterns

When concepts are too general

Head	**Modifier**
…	…
modem	comcast
wireless router	comcast
…	…

Head	**Modifier**
…	…
netflix	touchpad
skype	windows phone
…	…



Knowledge Bases

	WordNet	Wikipedia	Freebase	Probase
Cat	Feline; Felid; Adult male; Man; Gossip; Gossiper; Gossipmonger; Rumormonger; Rumourmonger; Newsmonger; Woman; Adult female; Stimulant; Stimulant drug; Excitant; Tracked vehicle; ...	Domesticated animals; Cats; Felines; Invasive animal species; Cosmopolitan species; Sequenced genomes; Animals described in 1758;	TV episode; Creative work; Musical recording; Organism classification; Dated location; Musical release; Book; Musical album; Film character; Publication; Character species; Top level domain; Animal; Domesticated animal; ...	Animal; Pet; Species; Mammal; Small animal; Thing; Mammalian species; Small pet; Animal species; Carnivore; Domesticated animal; Companion animal; Exotic pet; Vertebrate; ...
IBM	N/A	Companies listed on the New York Stock Exchange; IBM; Cloud computing providers; Companies based in Westchester County, New York; Multinational companies; Software companies of the United States; Top 100 US Federal Contractors; ...	Business operation; Issuer; Literature subject; Venture investor; Competitor; Software developer; Architectural structure owner; Website owner; Programming language designer; Computer manufacturer/brand; Customer; Operating system developer; Processor manufacturer; ...	Company; Vendor; Client; Corporation; Organization; Manufacturer; Industry leader; Firm; Brand; Partner; Large company; Fortune 500 company; Technology company; Supplier; Software vendor; Global company; Technology company; ...
Language	Communication; Auditory communication; Word; Higher cognitive process; Faculty; Mental faculty; Module; Text; Textual matter;	Languages; Linguistics; Human communication; Human skills; Wikipedia articles with ASCII art	Employer; Written work; Musical recording; Musical artist; Musical album; Literature subject; Query; Periodical; Type profile; Journal; Quotation subject; Type/domain equivalent topic; Broadcast genre; Periodical subject; Video game content descriptor; ...	**Instance of**: Cognitive function; Knowledge; Cultural factor; Cultural barrier; Cognitive process; Cognitive ability; Cultural difference; Ability; Characteristic; **Attribute of:** Film; Area; Book; Publication; Magazine; Country; Work; Program; Media; City; ...

What can Probase do?

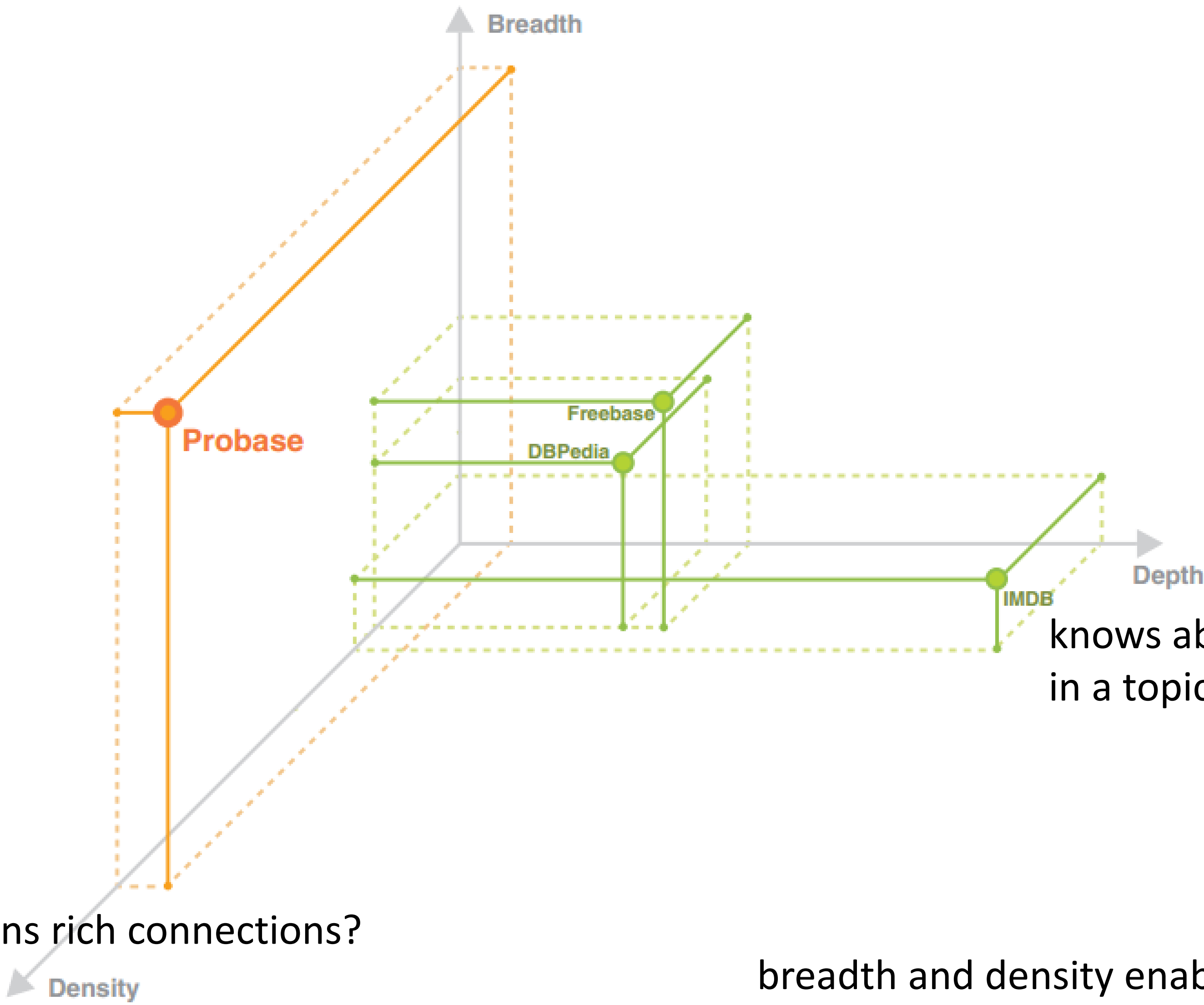
enable *understanding*

and

make up for the lack of depth

Knowledgebases

covers every topic?



knows about everything in a topic?

contains rich connections?

breadth and density enable understanding

Concept Learning

China **Brazil** **India**

emerging market

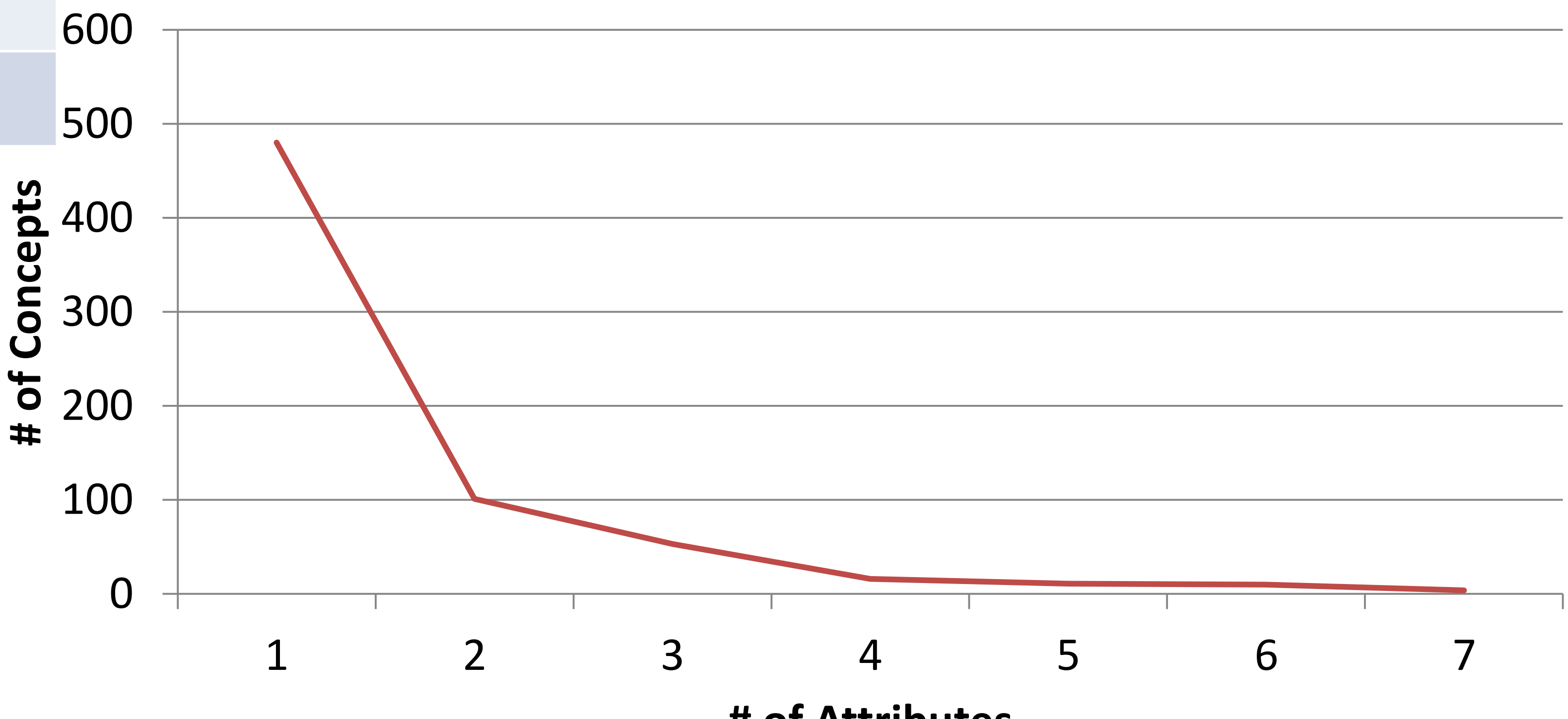
country

body smell taste

$$\underbrace{\hspace{12em}}_{\textit{wine}}$$

Understanding Web Tables

	website	president	city	motto	state	type	director



$\underbrace{\text{china}}_{\textit{country}}$ population collector of fine china

earthenware

Bayesian

$$P(c_k|E) = \frac{P(E|c_k)P(c_k)}{P(E)} \quad \propto P(c_k)\prod_{i=1}^{M} P(e_i|c_k).$$

- For a mixture of instances and properties: Noisy-Or model

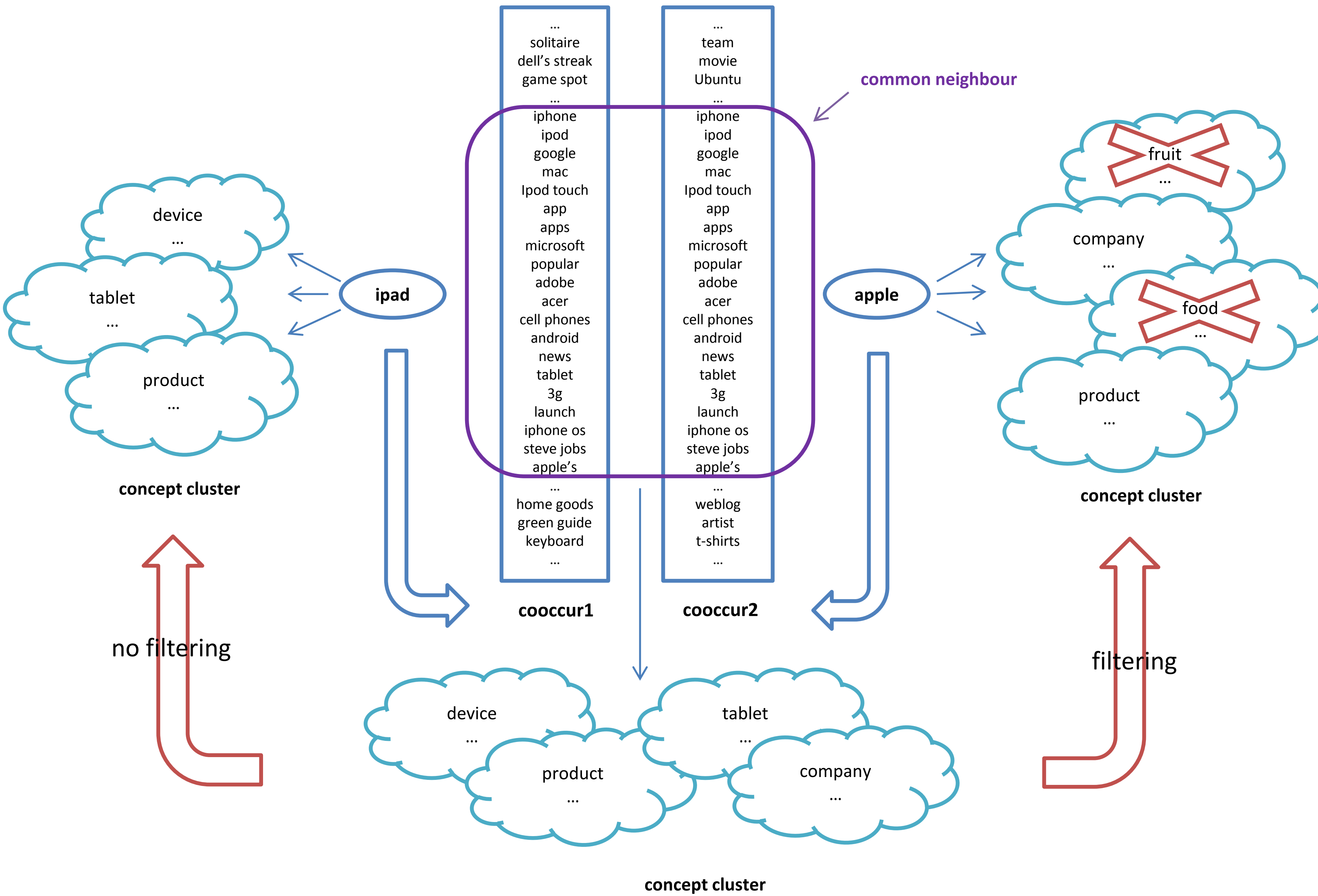
$$P(c|t_l) = 1 - \big(1 - P(c|t_l, z_l = 1)\big)\big(1 - P(c|t_l, z_l = 0)\big)$$

where $z_l = 1$ indicates t_l is an entity, $z_l = 0$ indicates t_l is a property

- Bayesian rule gives:

$$P(c|T) \ \propto P(c)\prod_l^L P(t_l|c) \propto \frac{\prod_l P(c|t_l)}{P(c)^{L-1}}$$

$\underbrace{\text{apple}}_{\textit{company}}$ $\qquad$ $\underbrace{\text{iPad}}_{\textit{device}}$



device
...
tablet
...
product
...
concept cluster
no filtering
ipad
...
solitaire
dell's streak
game spot
...
iphone
ipod
google
mac
Ipod touch
app
apps
microsoft
popular
adobe
acer
cell phones
android
news
tablet
3g
launch
iphone os
steve jobs
apple's
...
home goods
green guide
keyboard
...
cooccur1
...
team
movie
Ubuntu
...
iphone
ipod
google
mac
Ipod touch
app
apps
microsoft
popular
adobe
acer
cell phones
android
news
tablet
3g
launch
iphone os
steve jobs
apple's
...
weblog
artist
t-shirts
...
cooccur2
common neighbour
apple
fruit
...
company
...
food
...
product
...
concept cluster
filtering
device
...
product
...
tablet
...
company
...
concept cluster

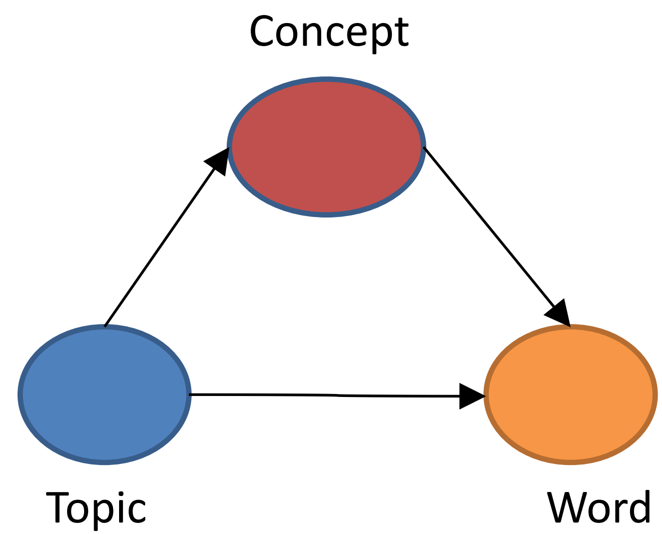
Modeling Co-occurrence

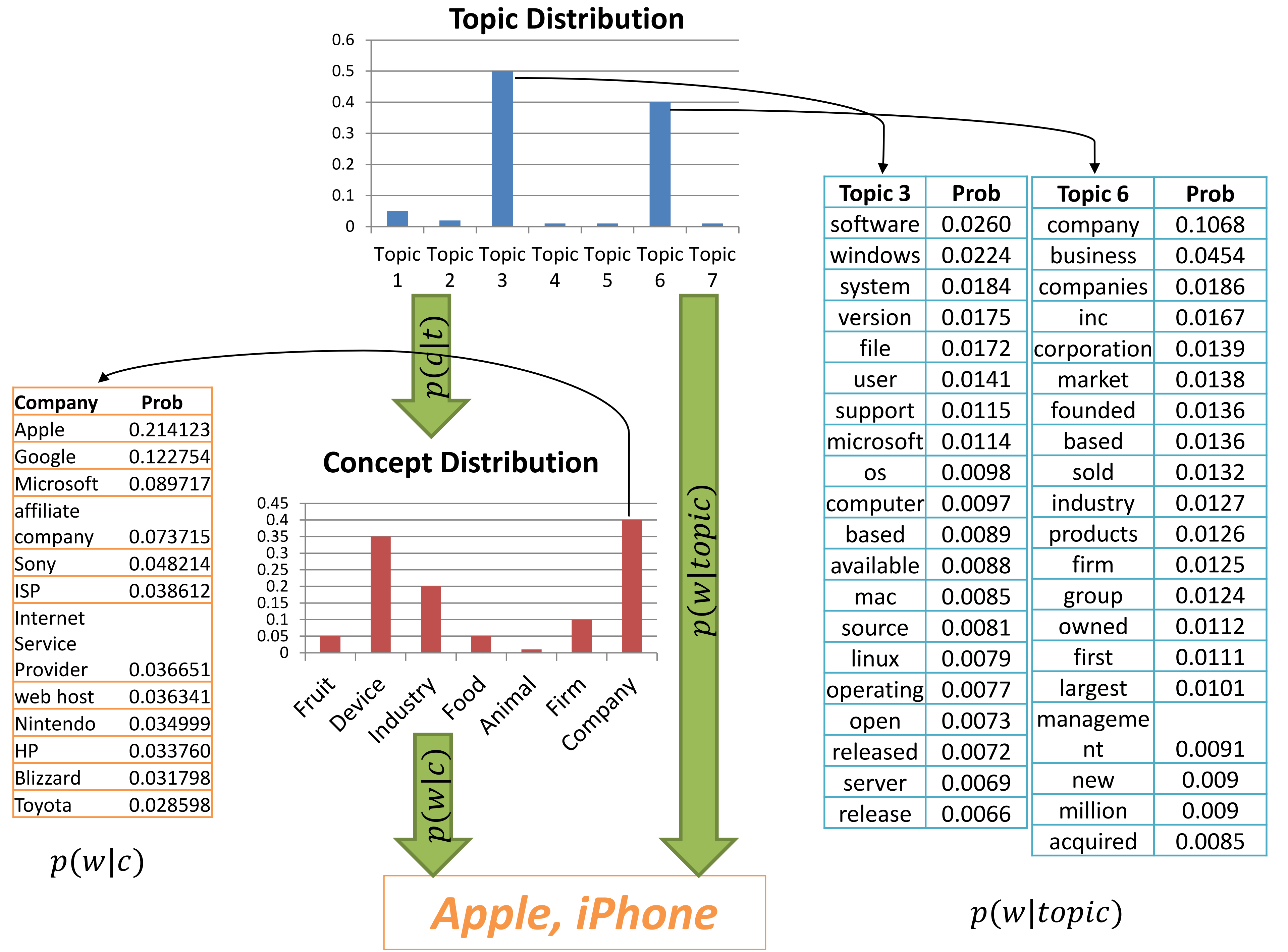
Probase

+

Wikipedia

⇨ LDA model





Company	Prob
Apple	0.214123
Google	0.122754
Microsoft	0.089717
affiliate company	0.073715
Sony	0.048214
ISP	0.038612
Internet Service Provider	0.036651
web host	0.036341
Nintendo	0.034999
HP	0.033760
Blizzard	0.031798
Toyota	0.028598

$p(w|c)$

Topic 3	Prob
software	0.0260
windows	0.0224
system	0.0184
version	0.0175
file	0.0172
user	0.0141
support	0.0115
microsoft	0.0114
os	0.0098
computer	0.0097
based	0.0089
available	0.0088
mac	0.0085
source	0.0081
linux	0.0079
operating	0.0077
open	0.0073
released	0.0072
server	0.0069
release	0.0066

Topic 6	Prob
company	0.1068
business	0.0454
companies	0.0186
inc	0.0167
corporation	0.0139
market	0.0138
founded	0.0136
based	0.0136
sold	0.0132
industry	0.0127
products	0.0126
firm	0.0125
group	0.0124
owned	0.0112
first	0.0111
largest	0.0101
manageme nt	0.0091
new	0.009
million	0.009
acquired	0.0085

$p(w|topic)$

- Infer topics z from text s using collapsed Gibbs sampling:

$$p(z_i = k|\vec{s}, z_{-i}, C) \propto (n_{\cdot k} + \alpha) \times \frac{C_{s_i k} + n_{s_i k} + \beta}{\sum_w C_{wk} + n_{wk} + |W|\beta},$$

- Estimate the concept distribution for each term w in s:

$$p(c|w, z) \propto p(c|w) \sum_k \pi_{wk} \phi_{ck},$$

$$\phi_{ck} = \frac{C_{ck} + \beta}{\sum_w C_{wk} + |W|\beta},$$

Examples

ShortText: fox fur

Conceptualize

Show Parameters

Elapsed Time = 00:00:00.2360236

fox

[159/wild animal/pet/animal][v]

159/wild animal/pet/animal	**0.5956765**
wild animal	0.0169223
feral animal	0.01490341
introduced animal	0.01263432
pest animal	0.01216037
small animal	0.01138677
nocturnal animal	0.01060585
native animal	0.01022427
predatory animal	0.009197926
animal	0.008580011
large animal	0.007967799

wild animal	0.004003763
feral animal	0.003526101
introduced animal	0.00298924
pest animal	0.002877105
small animal	0.002694075
nocturnal animal	0.002509312
native animal	0.002419031
predatory animal	0.002176201

fur

[4/texture/material][v]

4/texture/material	**0.2107609**
texture	0.01112421
organic material	0.007871442
soft material	0.007446955
luxury material	0.007329956
luxurious material	0.007232076
raw material	0.006870993
natural material	0.006293016
real world surface	0.00589916
locally available raw material	0.005892543
dead material	0.005889004

electronic product	0.01436949
electronic good	0.01051342
high-tech product	0.009497663
electrical good	0.006462679
consumer electronic product	0.006424694
electrical product	0.006299805
consumer product	0.005079063
range electrical product	0.004229661

/channel/network]

nel/network	**0.6562241**
	0.1072035
	0.0970483
	0.06378444
	0.05830856
	0.0403064
	0.0391444
	0.03133982
	0.0295761
	0.02876115
	0.02717765

Examples

ShortText: read harry potter [Conceptualize]

Good HalfGood NotGood

[Report]

Show Parameters

Elapsed Time = 00:00:00.0156005

read[v] **harry potter**

[67/book]

67/book	**0.543426**
book	0.07531892
fantasy book	0.04780534
popular book	0.03634102
children's book	0.02661931
fiction book	0.02292863
chapter book	0.02292863
modern book	0.01817051
long book	0.01817051
series book	0.01146431
interesting book	0.01146431

254/novel	**0.2113914**
novel	0.03902724
fantasy novel	0.03693517
popular novel	0.01231172
great novel	0.01231172

Examples

SHORT TEXT CONCEPTUALIZATION [Log In]

Conceptualization | ConceptualizationGraphic | AmbiguityView | CooccurView | BenchmarkView

SHORT TEXT CONCEPTUALIZATION

(*This is only for demo. Please note that this is not necessarily the up-to-date version*)

ShortText: apple engineer is eating the apple [Conceptualize]

○ Good ○ HalfGood ○ NotGood

[Report]

Show Parameters

Elapsed Time = 00:00:01.1051105

apple [1/company]		engineer [805/professional][v]		eating [2/activity]		apple [9405/food]	
1/company	**0.9556221**	**805/professional**	**0.5360888**	**2/activity**	**0.9672647**	**9405/food**	**0.7455807**
company	0.01050991	professional	0.02111498	activity	0.04600645	food	0.01541176
corporation	0.006285705	expert	0.0127867	everyday activity	0.03235053	ingredient	0.009355835
firm	0.006132113	occupation	0.0127867	simple activity	0.02173292	high fiber food	0.008366366
large company	0.005865776	design professional	0.01129569	daily living activity	0.02010534	hard food	0.008017257
client	0.005627672	licensed professional	0.009778754	hobby	0.0180488	crunchy food	0.00769472
player	0.005538661	technical professional	0.009208023	basic activity	0.0180488	fiber-rich food	0.007606609
stock	0.005443777	professional group	0.008764553	normal daily activity	0.0180488	healthy food	0.00751504
technology company	0.005443777	skilled professional	0.008764553	hand-to-mouth activity	0.015253	fresh food	0.007216591
big company	0.005155101	construction professional	0.008252603	life-sustaining activity	0.015253	fiber rich food	0.006322427
giant	0.004985663	industry professional	0.006906016	day activity	0.01404469	wholesome ingredient	0.006161352

Similarity between Two Short Texts

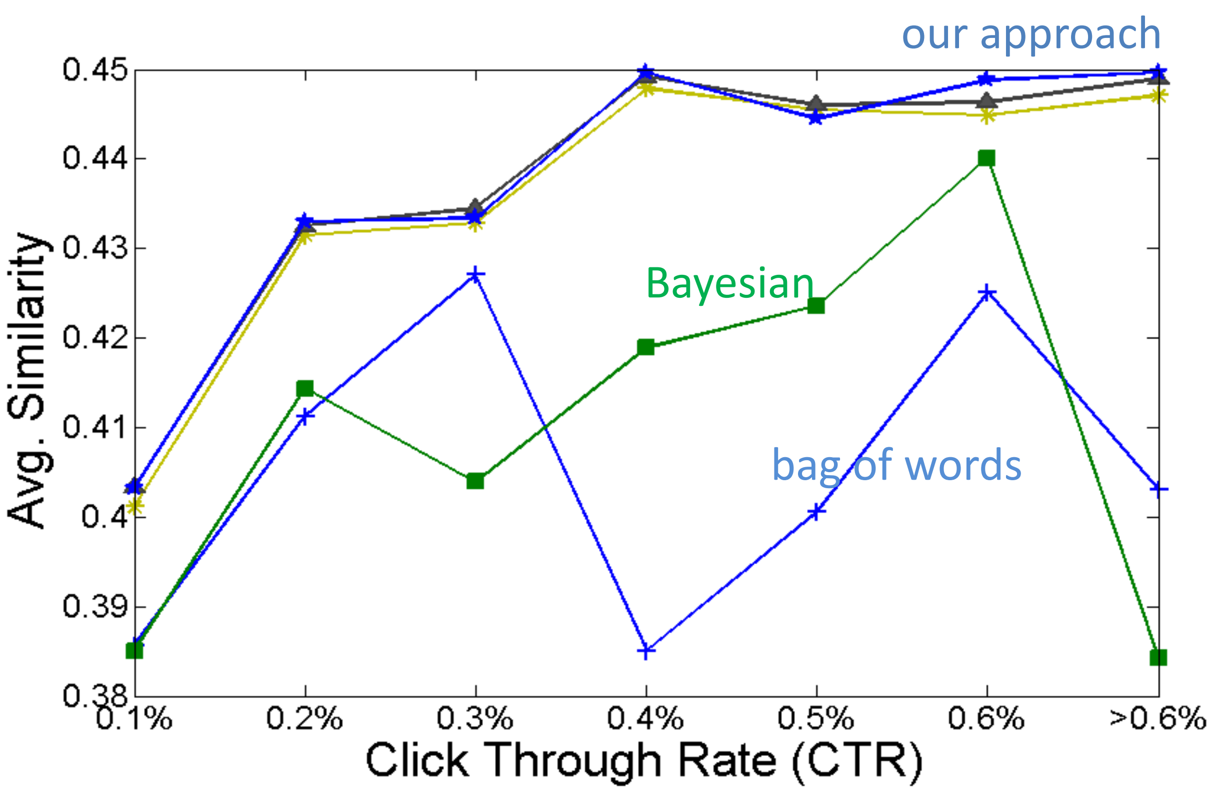
Search and URL title:

	Bayesian	LDA	LDA+Probase
T100	0.31 (0.29↑)	0.55 (0.31↑)	0.42 (0.39↑)
T200		0.52 (0.31↑)	0.42 (0.39↑)
T300		0.50 (0.31↑)	0.43 (0.40↑)

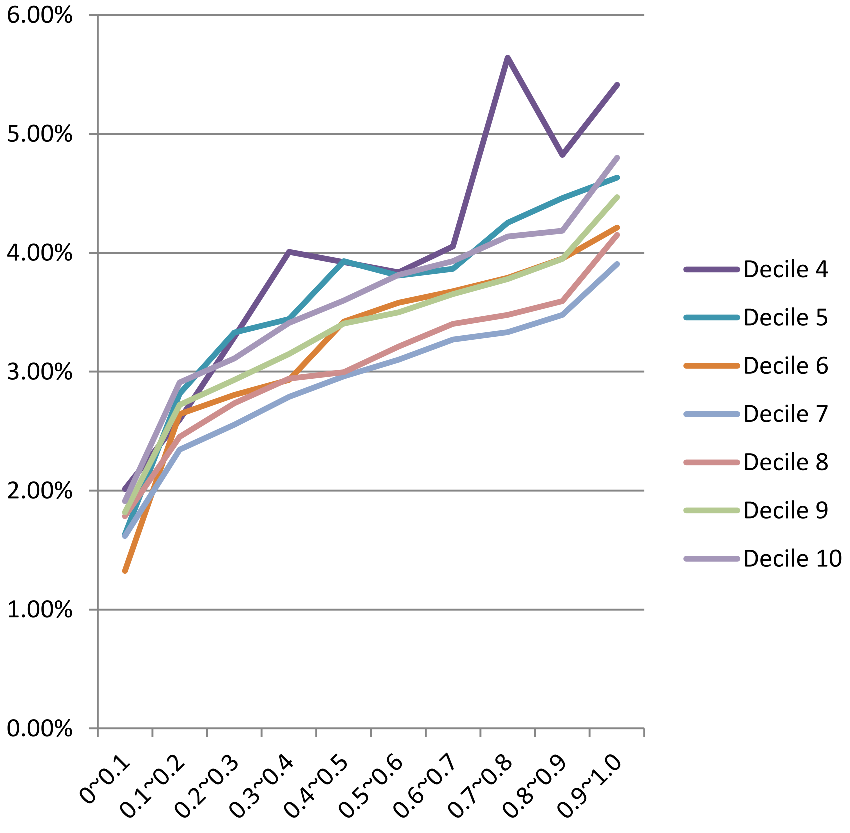
Two random searches:

	Bayesian	LDA	LDA+Probase
T100	0.02	0.24	0.03
T200		0.21	0.03
T300		0.19	0.03

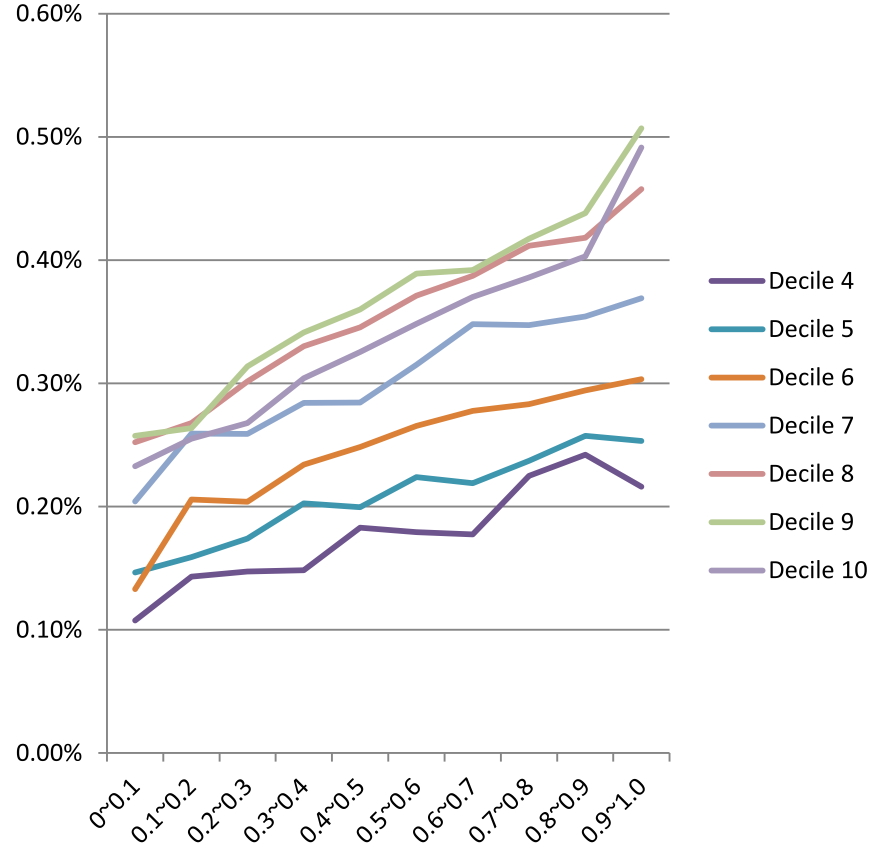
CTR and search/ads similarity



CTR and search/ads similarity (torso and tail queries)



Mainline Ads



Sidebar Ads

FrameNet Sentences

	Basic	Context Sensitive		
		T100	T200	T300
Fold 1	-4.716	-3.401	-3.385	-3.378
Fold 2	-4.728	-3.409	-3.393	-3.389
Fold 3	-4.741	-3.432	-3.417	-3.410
Fold 4	-4.727	-3.413	-3.399	-3.392
Fold 5	-4.740	-3.433	-3.417	-3.413

Log-likelihood of frame elements with five-fold validation.

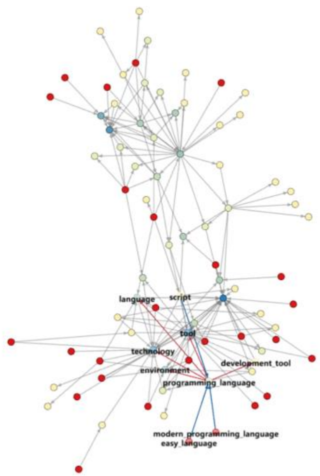
Many applications

We mainly worked in the Search/Ads domain

- Related search
- Ads selection
- Bid keyword suggestion
- Search suggestion
- …

knowledge

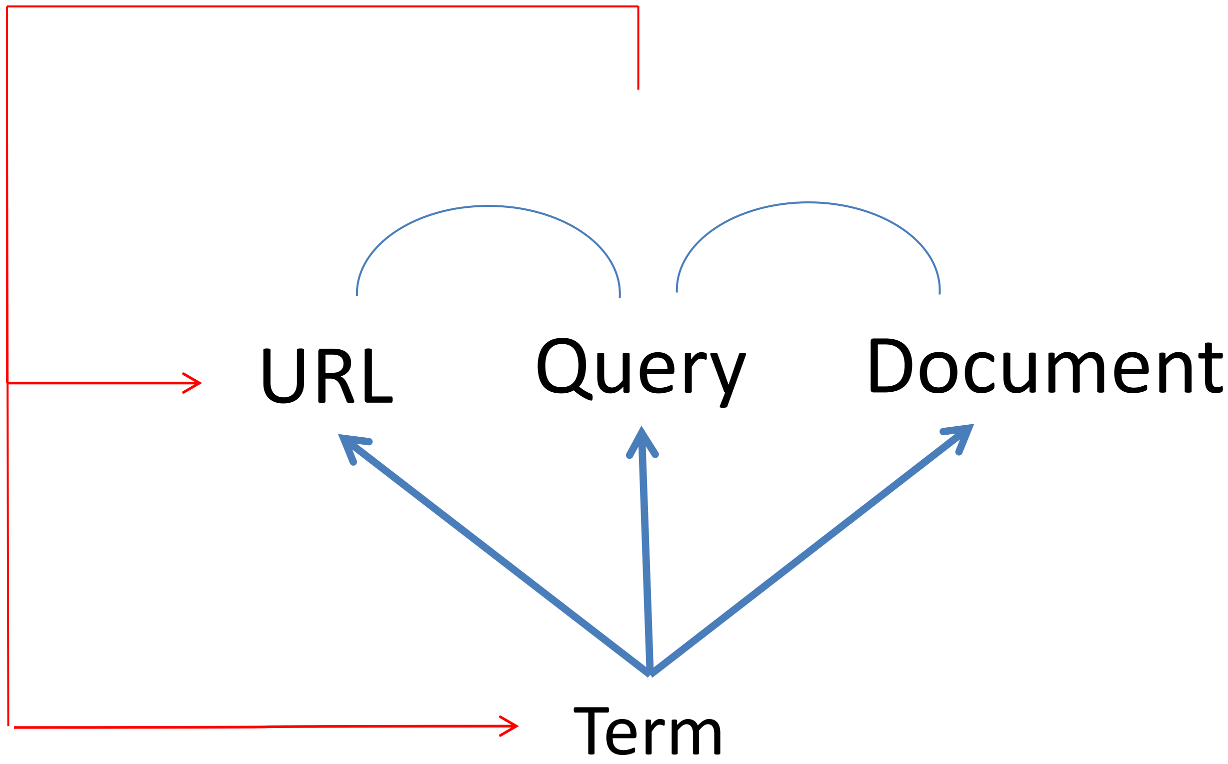
representation



tasks

86%

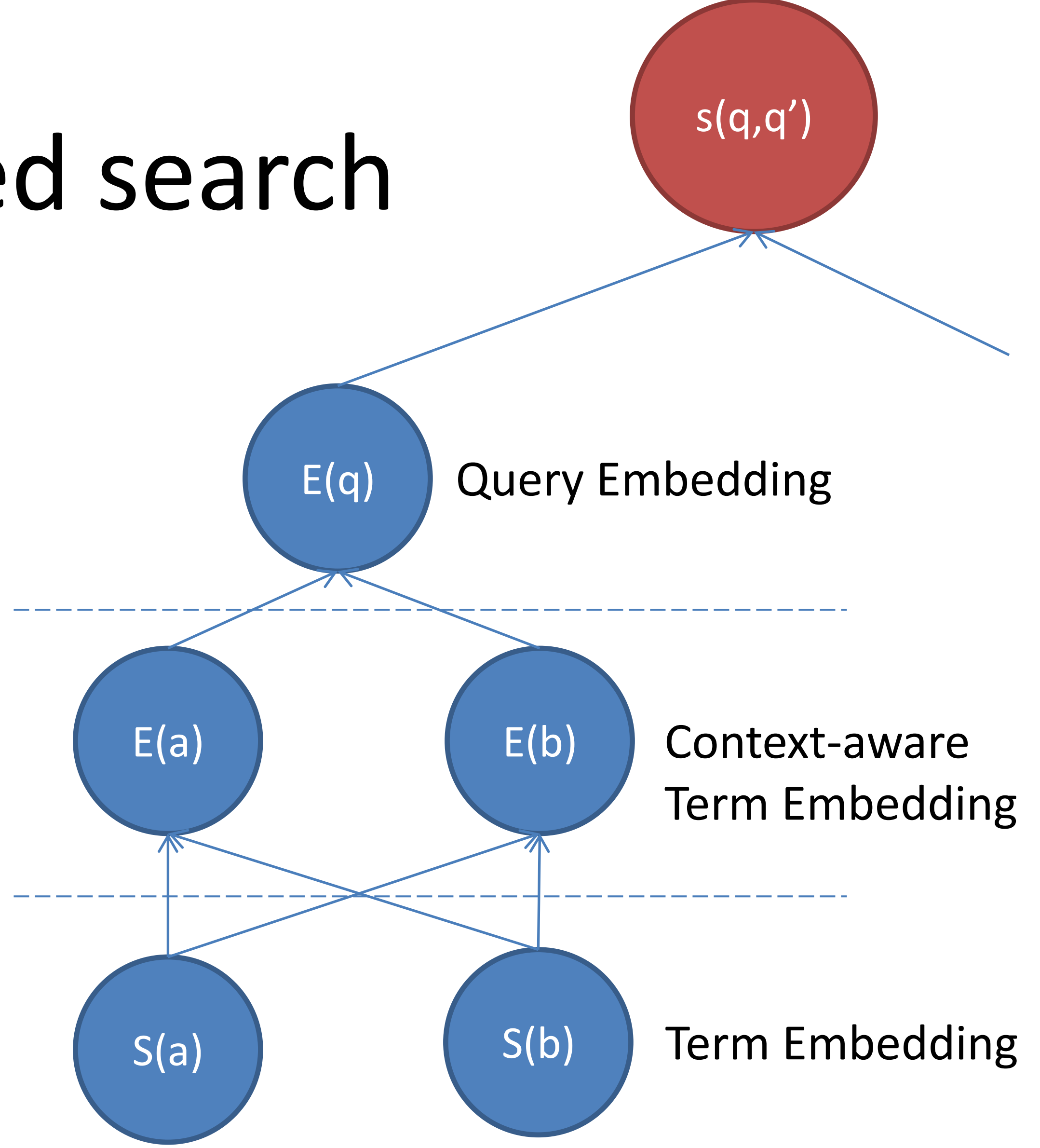
Representation



Example: related search

- Given a query q
- Positive case q^{+}
  - Queries in the same session
- Negative case q^{-}
  - Generated randomly
- Intuition: $s(q, q^{+}) > s(q, q^{-})$
- Objective function:

$$\sum_{q} \sum_{q^{+}, q^{-}} \max\ (0, 1 - S(q, q^{+}) + S(q, q^{-}))$$

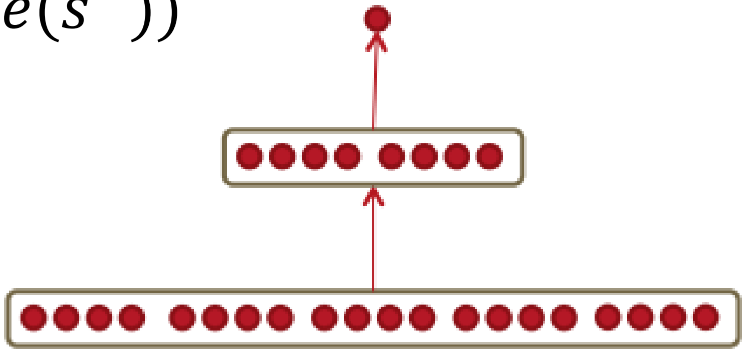


Word embedding [Collobert and Weston 2008]

- Positive:
  - s^+: "… UN assists **China** in developing …"

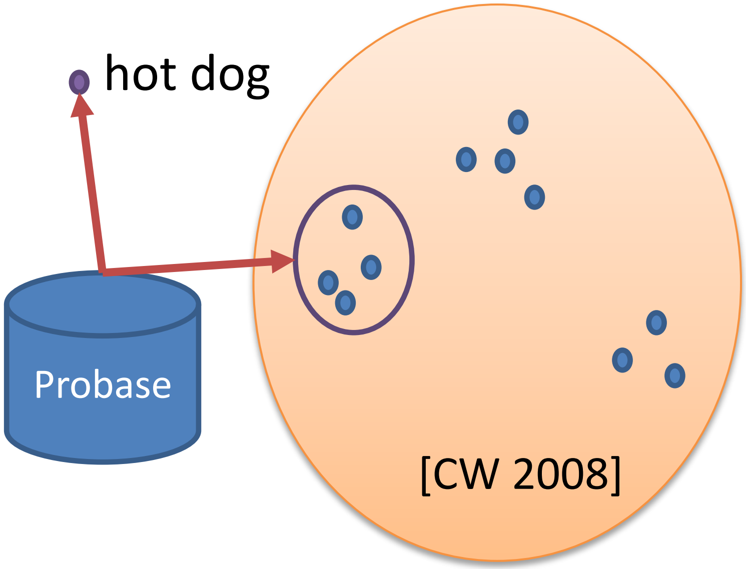
- Negative:
  - s^-: "… UN assists **banana** in developing …"

- $J = \max(0, 1 - Score(s^+) + Score(s^-))$



Extending Embedding using Probase

- For any term not covered by the embedding
  - "hot dog"
- Find its neighbors in Probase conceptual space
  - "bagel", "sandwich", etc.
- Use the average embedding of its top-k neighbors
  - Special case: k = 1
- Handle multi-sense



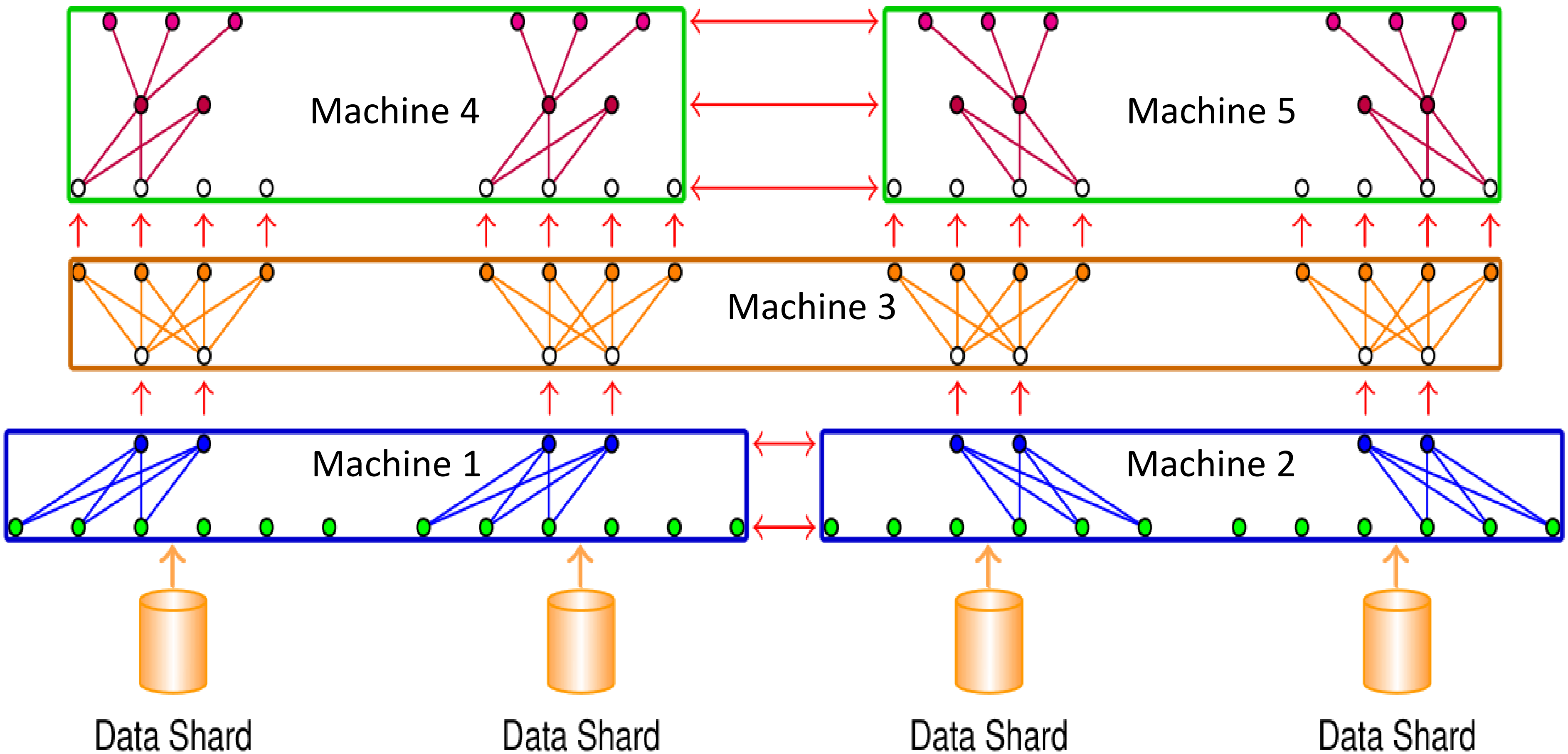
Probase graph embedding

- Probase concept graph G (m concepts)
- w_{ij} : weight (similarity between concept i and j)
- Let $y = (y_1, \dots, y_m)^T$ be the embedding of G
- The optimal y is given by minimizing

$$\sum_{i,j} \left\| y_i - y_j \right\|^2 f(w_{ij})$$

To be updated.

Implementation on Trinity



Data Partitioning → Model Partitioning / Replication → Training Pipeline

Probase Publications

1. Context dependent conceptualization, IJCAI 2013
2. Automatic extraction of top-k lists from web data, ICDE 2013
3. Attribute Extraction and Scoring: A Probabilistic Approach, ICDE 2013
4. Identifying Users' Topical Tasks in Web Search, WSDM 2013
5. Probase: A Probabilistic Taxonomy for Text Understanding, *SIGMOD* 2012
6. Optimizing Index for Taxonomy Keyword Search, *SIGMOD* 2012
7. Automatic Taxonomy Construction from Keywords, *KDD* 2012
8. A System for Extracting Top-K Lists from the Web (demo)*, KDD* 2012
9. Understanding Tables on the Web, *ER* 2012
10. Toward Topic Search on the Web, *ER* 2012
11. Isanette: A Common and Common Sense Knowledge Base for Opinion Mining, *ICDM Workshops* 2011
12. Web Scale Taxonomy Cleansing, *VLDB* 2011
13. Short Text Conceptualization using a Probabilistic Knowledgebase, *IJCAI* 2011

Thanks