Probase: A Probabilistic Taxonomy for Text Understanding

Wentao Wu [1], Hongsong Li [2], Haixun Wang [2], Kenny Q. Zhu [3]

[1] University of Wisconsin, Madison, WI, USA
[2] Microsoft Research Asia, Beijing, China
[3] Shanghai Jiao Tong University, Shanghai, China

Outline

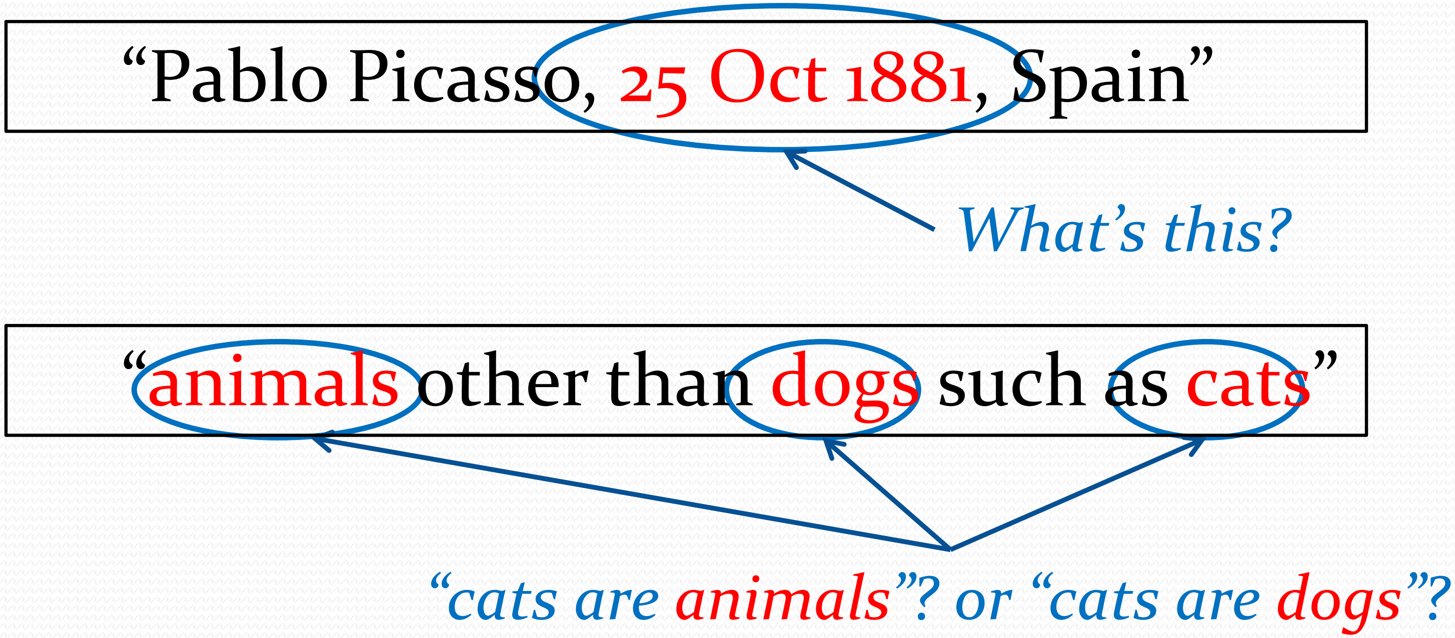
- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion

Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion

Text Understanding

- Machines need to *understand* text to unlock the information confined in Web data.



"Pablo Picasso, 25 Oct 1881, Spain"

What's this?

"animals other than dogs such as cats"

"cats are animals"? or "cats are dogs"?

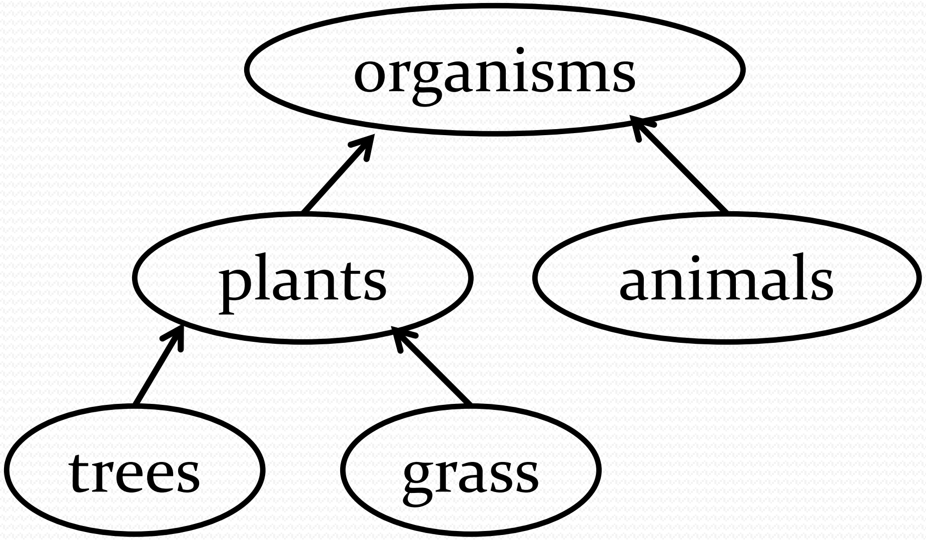
Conceptualization

- A little piece of *knowledge* makes the difference.
  - "Pablo Picasso is a person"
  - "cats are animals"

- Can machines know this?
  - They can't.
  - We need to pass this piece of knowledge to them.

Taxonomies

- A *hierarchical* structure showing the *isA* relationships among concepts.



Limited Size of Concept Space

> “How do we compete with the *largest companies in US*?”

Existing Taxonomies	Number of Concepts
Probase	**2,653,872**
YAGO	352,297
WordNet	25,229
Freebase	1,450
DBPedia	259
NELL	123

Knowledge is Black and White

“How do we compete with the *largest companies in US*?”

- “Vague” concepts
  - *“largest companies in US”* => *Walmart? Microsoft? P&G?*
  - *“beautiful cities”* => *Seattle? Chicago? Shanghai?*

There is inherent uncertainty inside these concepts!

Probase

- Automatically constructed from 1.6 billion web pages (with *92.4%* precision).

- The largest *concept* space so far (*2.6 million*).

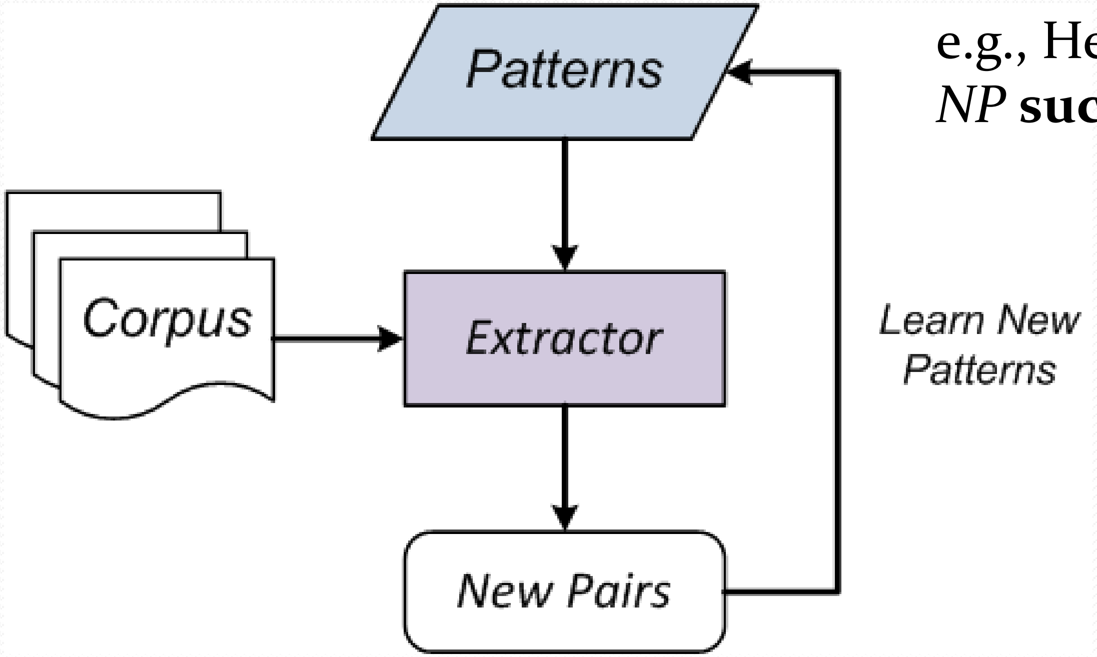
- Use *probabilistic* approach to model the uncertainty inside the concepts.

Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion

Previous Work

- Syntactic Iteration (*KnowItAll*, *TextRunner*, *NELL)*



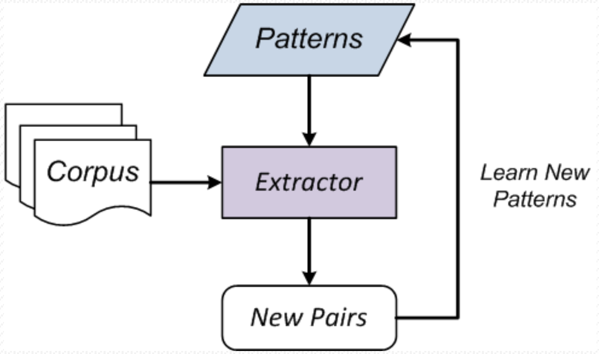
e.g., Hearst Patterns (as seeds):
NP **such as** {*NP*,}*{(**or**|**and**)} *NP*

Problems of Syntactic Iteration

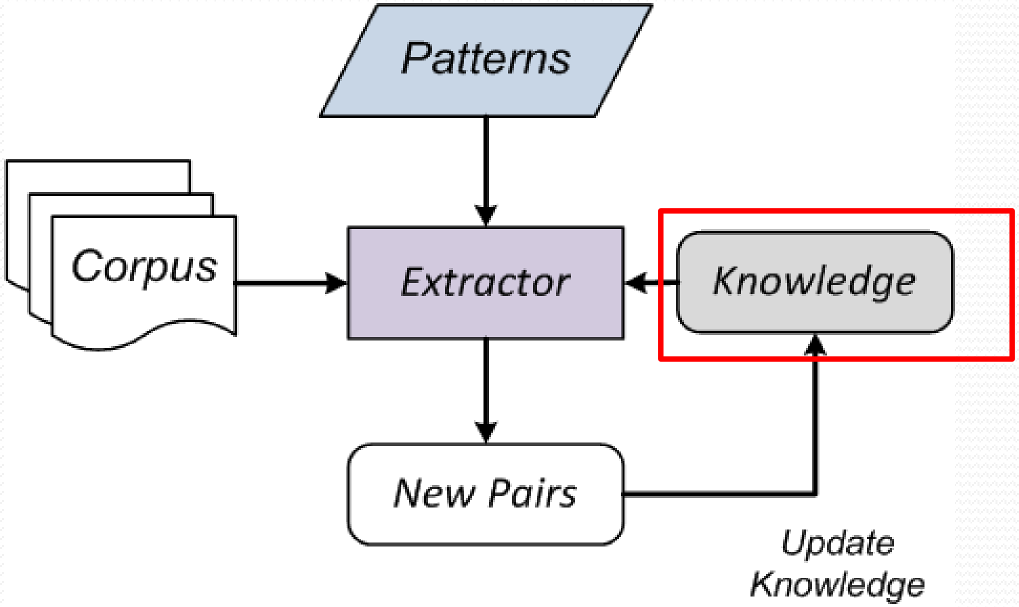
- Syntactic patterns have limited extraction power.
  - "… animals other than dogs such as cats …"
- High quality syntactic patterns are rare.
  - Good patterns: "*x* is a country" => *x* = "China"
  - Bad patterns: "war with *x*" => *x* = "planet Earth"
- Recall is sacrificed for precision.
  - E.g., some methods only focus on extracting *proper nouns*.

Our Approach

- Semantic Iteration



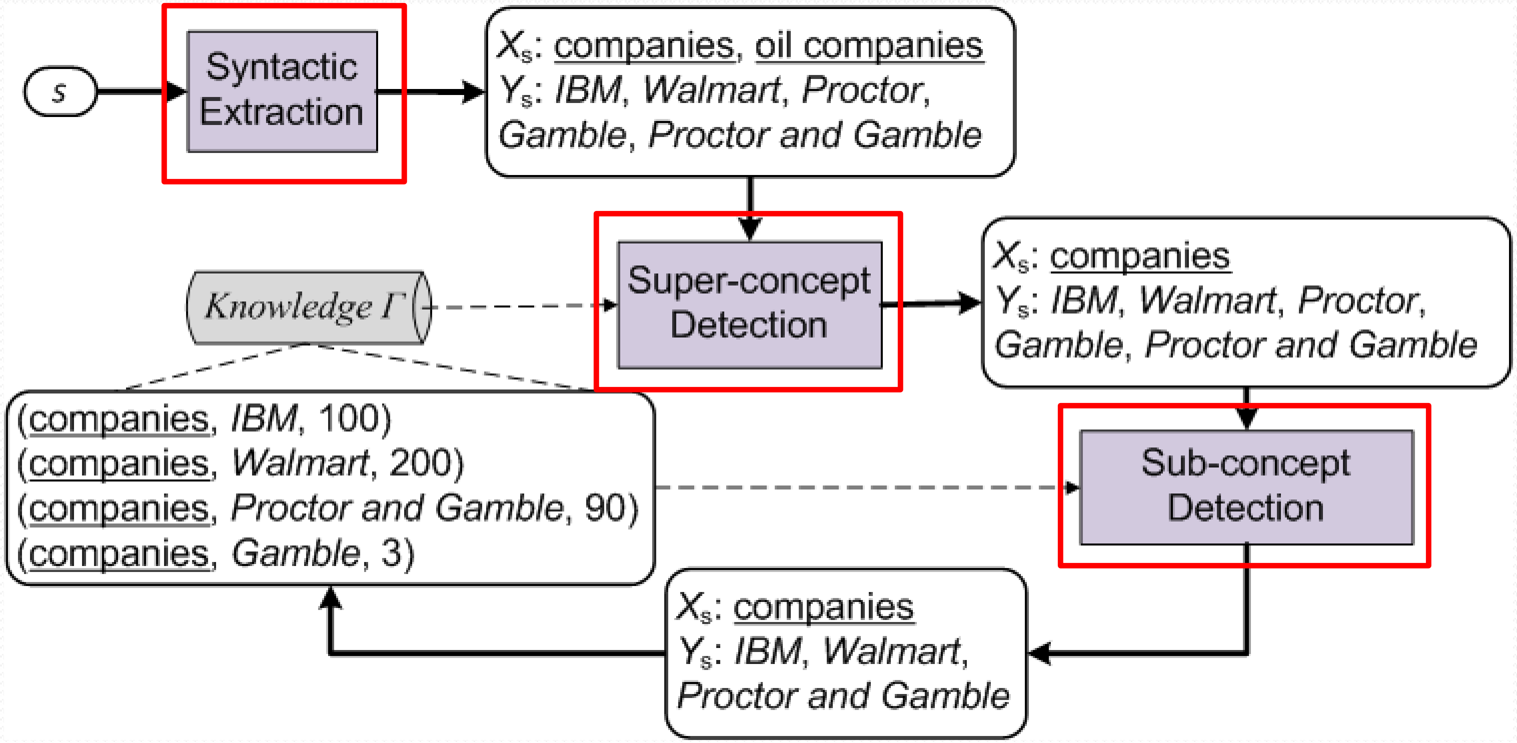
Syntactic Iteration



Semantic Iteration

An Example

s: … companies other than oil companies **such as** *IBM, Walmart, Proctor and Gamble*, …


s
Syntactic Extraction
X_s: companies, oil companies
Y_s: IBM, Walmart, Proctor, Gamble, Proctor and Gamble
Knowledge Γ
Super-concept Detection
X_s: companies
Y_s: IBM, Walmart, Proctor, Gamble, Proctor and Gamble
(companies, IBM, 100)
(companies, Walmart, 200)
(companies, Proctor and Gamble, 90)
(companies, Gamble, 3)
Sub-concept Detection
X_s: companies
Y_s: IBM, Walmart, Proctor and Gamble

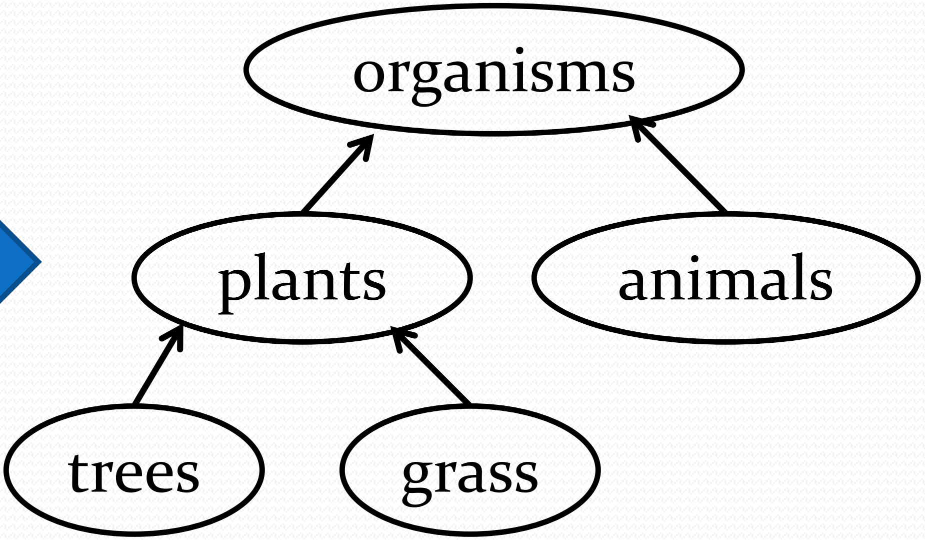
Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion

Goal

- Build a taxonomy *graph* from the *edges* (“*isA*” pairs) from the previous data extraction stage.

(organisms, *animals*)
(organisms, *plants*)
(plants, *trees*)
(plants, *grass*)



Challenges

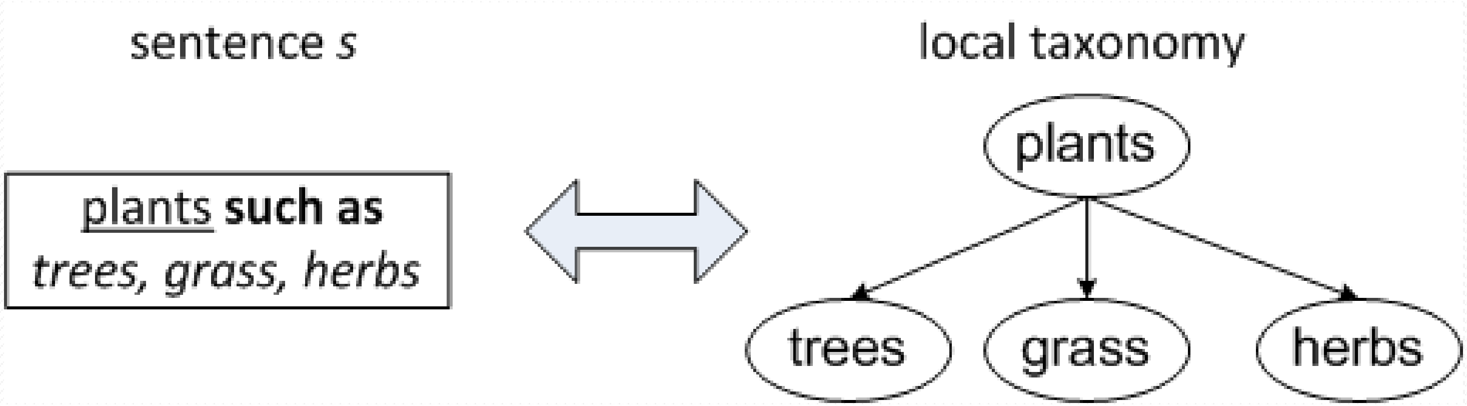
- Should we merge the two "apple" here?
  - e_1 = (fruit, *apple*), e_2 = (companies, *apple*)

- Should we merge the two "plants" here?
  - e_1 = (plants, *tree*), e_2 = (plants, *steam turbines*)

Words such as "apple" and "plants" have multiple meanings (senses).

Properties & Operations(1)

- Example:
  - … plants **such as** *trees*, *grass*, **and** *herbs* …
  - … plants **such as** *steam turbines*, *pumps*, **and** *boilers* …



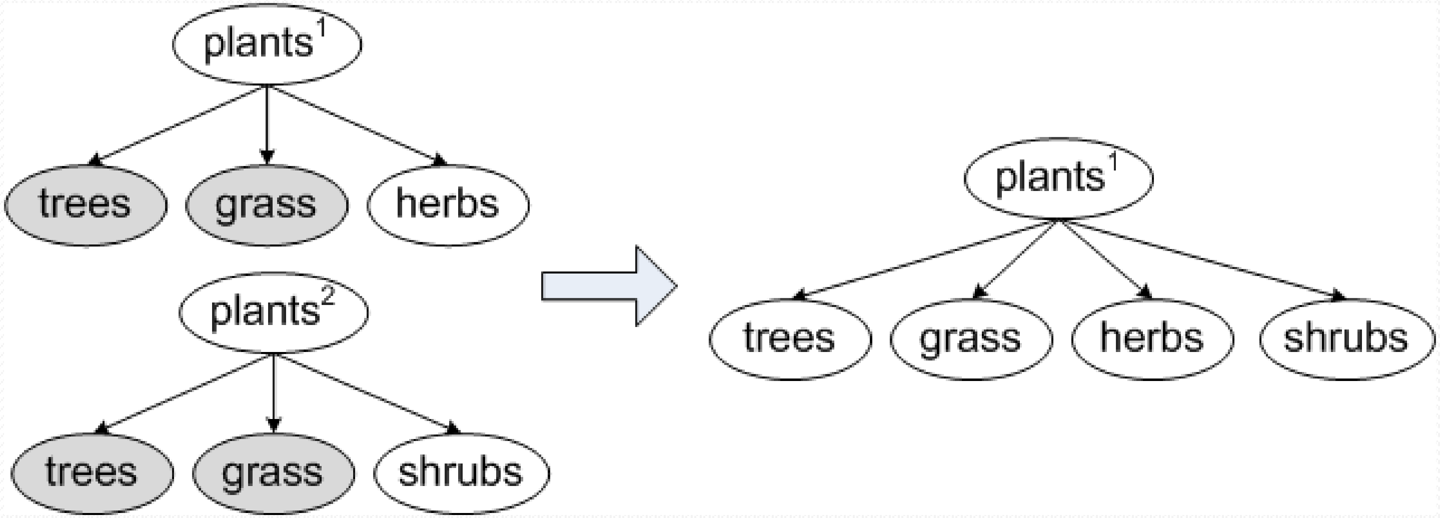
Local Taxonomy Construction

Properties & Operations (2)

- Example:

  a) … plants **such as** *trees, grass,* **and** *herbs* …

  b) … plants **such as** *trees, grass,* **and** *shrubs* …



Horizontal Merge

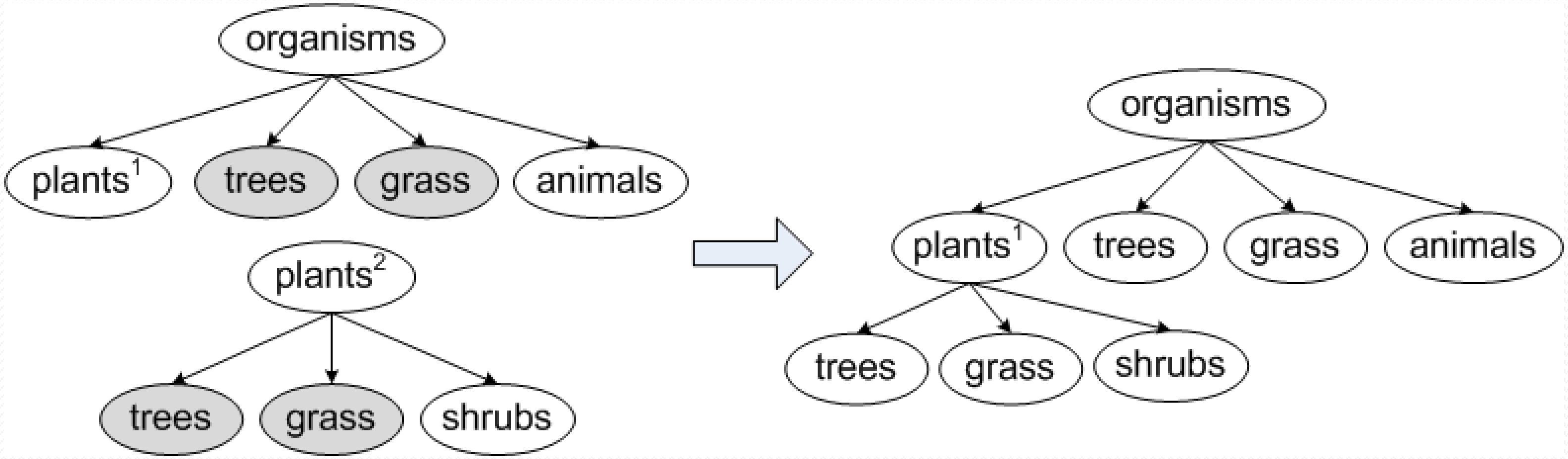
Properties & Operations (3)

- Example:

  a) … organisms **such as** *plants*, *trees*, *grass* **and** *animals* …

  b) … plants **such as** *trees*, *grass*, **and** *shrubs* …

  c) … plants **such as** *steam turbines*, *pumps*, **and** *boilers* …



Vertical Merge

Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion

Plausibility

How likely is that the claim “*y* is an *x*” is true?

$$P(x, y) = 1 - p(\overline{E}) = 1 - p(\bigcap_{i=1}^{n} \overline{s_i}) = 1 - \prod_{i=1}^{n} (1 - p_i)$$

s_i: evidence (or sentence) that supports (x, y)
p_i: the probability that the evidence s_i is true

Typicality

- Which one is more *typical* for the concept “bird”? a *robin* or *ostrich*?

$$T(i \mid x) = \frac{n(x,i) \cdot P(x,i)}{\sum_{i' \in I_x} n(x,i') \cdot P(x,i')}$$

⇓

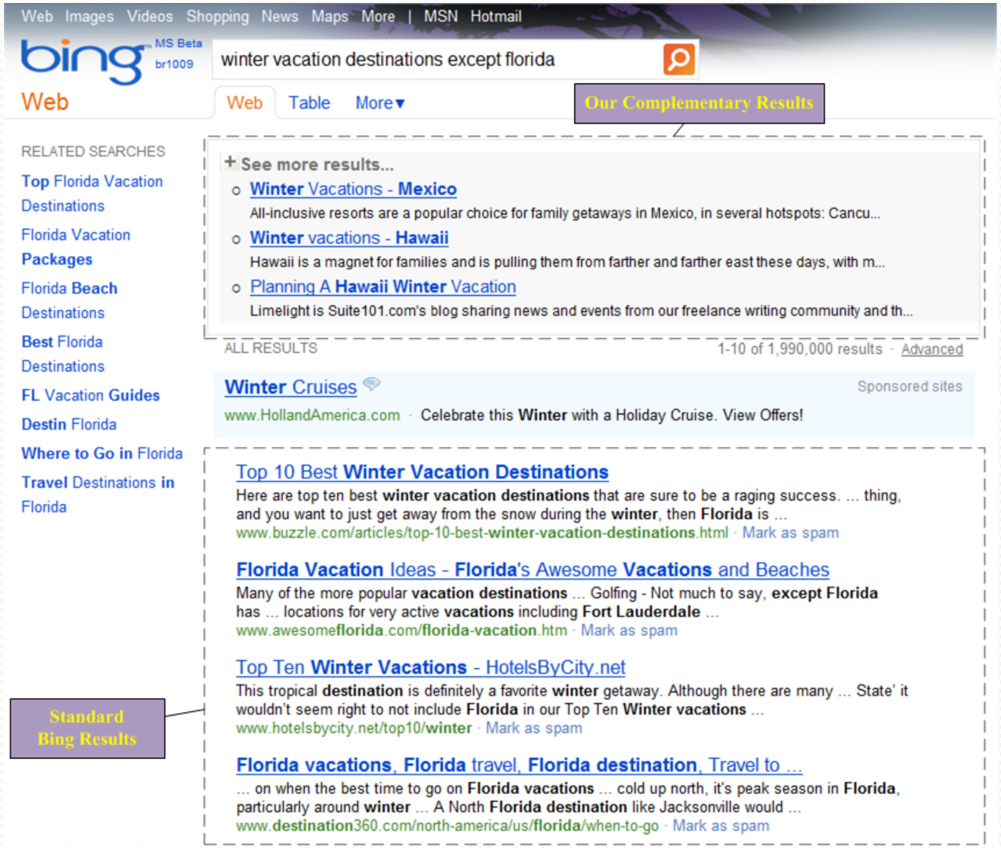
An instance of “big company” is also an instance of “company”.

$$T(i \mid x) = \frac{\sum_{y \in D(x)} \tilde{P}(x,y) \cdot n(y,i) \cdot P(y,i)}{\sum_{i' \in I_x} \sum_{y \in D(x)} \tilde{P}(x,y) \cdot n(y,i') \cdot P(y,i')}$$

$\tilde{P}(x,y)$ is the *plausibility* that y is a *descendant* concept of x.

Application of Typicality (1)

- Semantic Web Search (ER'12)



winter vacation destinations except florida

Our Complementary Results

RELATED SEARCHES

Top Florida Vacation Destinations

Florida Vacation Packages

Florida Beach Destinations

Best Florida Destinations

FL Vacation Guides

Destin Florida

Where to Go in Florida

Travel Destinations in Florida

+ See more results...

- Winter Vacations - Mexico
  All-inclusive resorts are a popular choice for family getaways in Mexico, in several hotspots: Cancu...
- Winter vacations - Hawaii
  Hawaii is a magnet for families and is pulling them from farther and farther east these days, with m...
- Planning A Hawaii Winter Vacation
  Limelight is Suite101.com's blog sharing news and events from our freelance writing community and th...

ALL RESULTS 1-10 of 1,990,000 results · Advanced

Winter Cruises Sponsored sites
www.HollandAmerica.com · Celebrate this Winter with a Holiday Cruise. View Offers!

Top 10 Best Winter Vacation Destinations
Here are top ten best winter vacation destinations that are sure to be a raging success. ... thing, and you want to just get away from the snow during the winter, then Florida is ...
www.buzzle.com/articles/top-10-best-winter-vacation-destinations.html · Mark as spam

Florida Vacation Ideas - Florida's Awesome Vacations and Beaches
Many of the more popular vacation destinations ... Golfing - Not much to say, except Florida has ... locations for very active vacations including Fort Lauderdale ...
www.awesomeflorida.com/florida-vacation.htm · Mark as spam

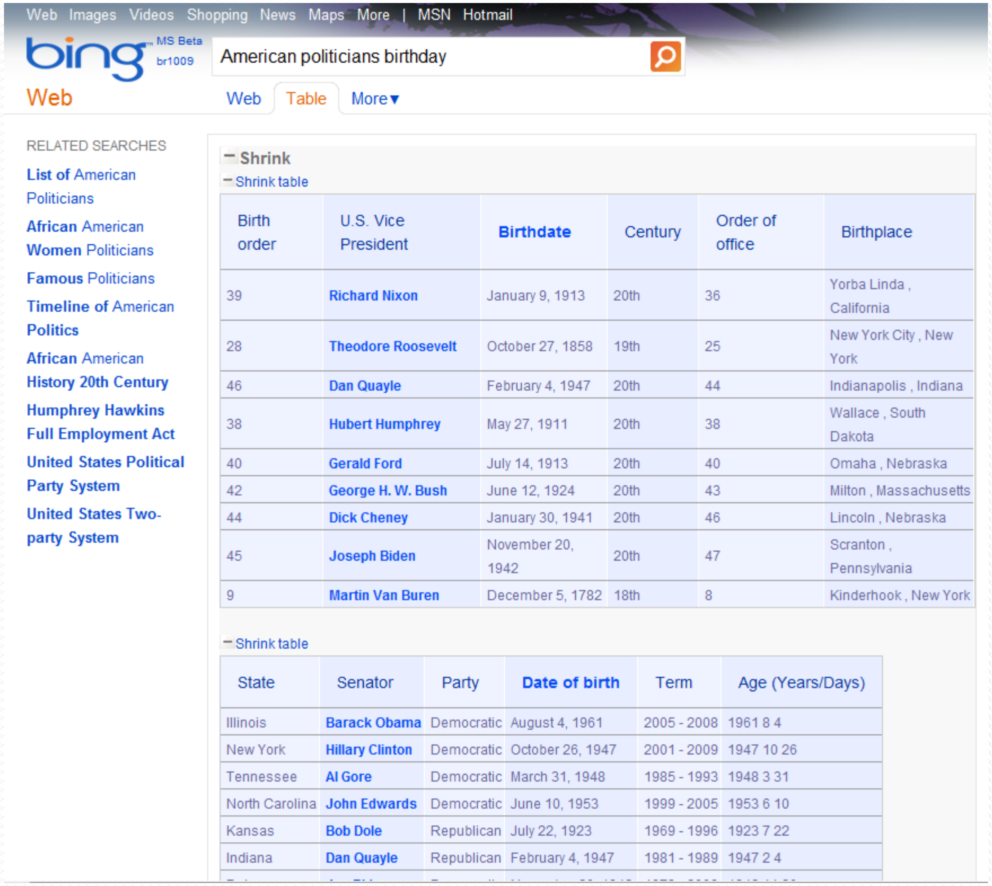
Top Ten Winter Vacations - HotelsByCity.net
This tropical destination is definitely a favorite winter getaway. Although there are many ... State' it wouldn't seem right to not include Florida in our Top Ten Winter vacations ...
www.hotelsbycity.net/top10/winter · Mark as spam

Florida vacations, Florida travel, Florida destination, Travel to ...
... on when the best time to go on Florida vacations ... cold up north, it's peak season in Florida, particularly around winter ... A North Florida destination like Jacksonville would ...
www.destination360.com/north-america/us/florida/when-to-go · Mark as spam

Standard Bing Results

Application of Typicality (2)

- Understanding Web Tables (ER'12)


Web Images Videos Shopping News Maps More | MSN Hotmail
bing MS Beta br1009
American politicians birthday
Web
Web Table More▾

RELATED SEARCHES

List of American Politicians
African American Women Politicians
Famous Politicians
Timeline of American Politics
African American History 20th Century
Humphrey Hawkins Full Employment Act
United States Political Party System
United States Two-party System

− Shrink

− Shrink table

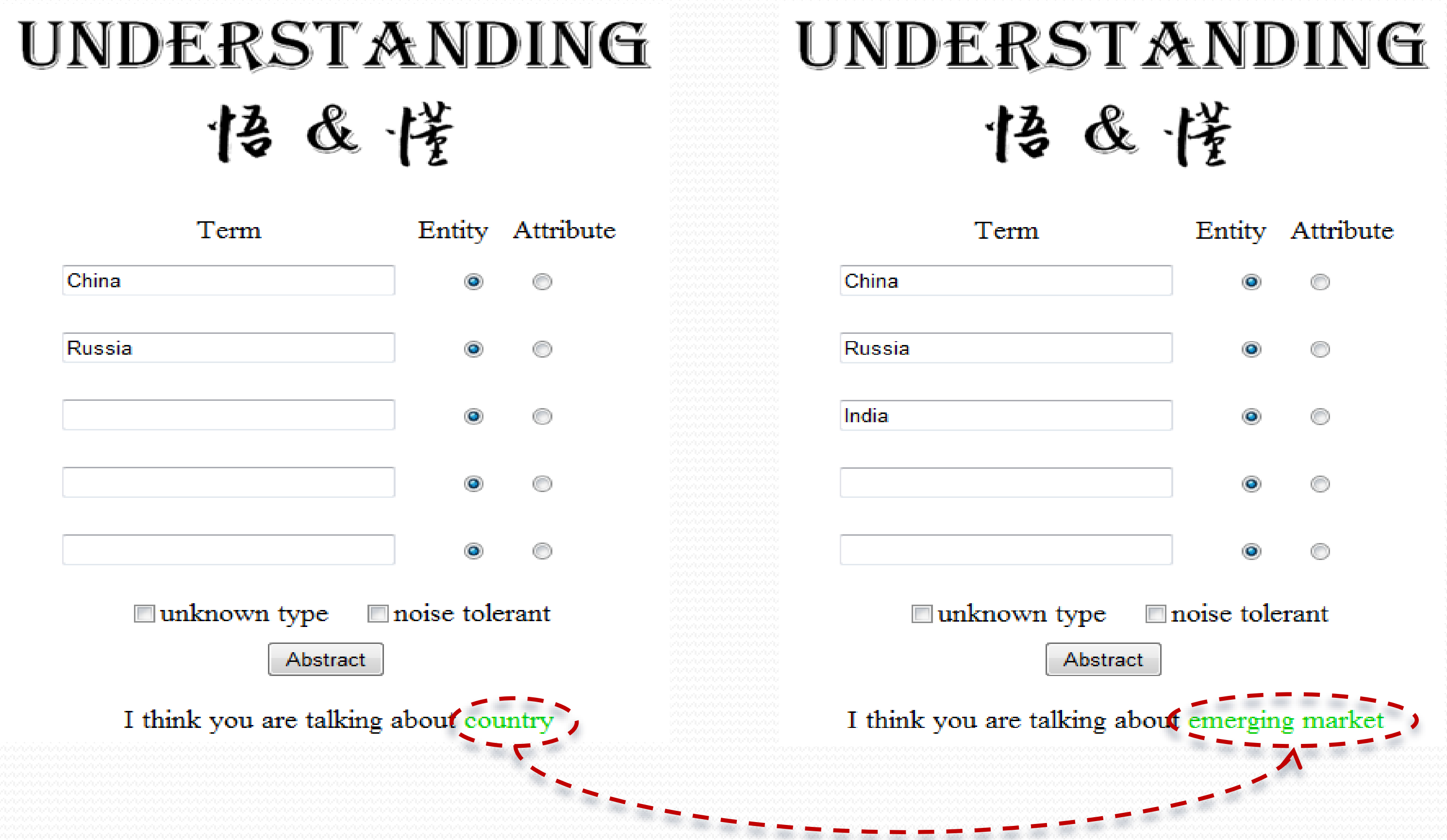
Birth order	U.S. Vice President	Birthdate	Century	Order of office	Birthplace
39	Richard Nixon	January 9, 1913	20th	36	Yorba Linda , California
28	Theodore Roosevelt	October 27, 1858	19th	25	New York City , New York
46	Dan Quayle	February 4, 1947	20th	44	Indianapolis , Indiana
38	Hubert Humphrey	May 27, 1911	20th	38	Wallace , South Dakota
40	Gerald Ford	July 14, 1913	20th	40	Omaha , Nebraska
42	George H. W. Bush	June 12, 1924	20th	43	Milton , Massachusetts
44	Dick Cheney	January 30, 1941	20th	46	Lincoln , Nebraska
45	Joseph Biden	November 20, 1942	20th	47	Scranton , Pennsylvania
9	Martin Van Buren	December 5, 1782	18th	8	Kinderhook , New York

− Shrink table

State	Senator	Party	Date of birth	Term	Age (Years/Days)
Illinois	Barack Obama	Democratic	August 4, 1961	2005 - 2008	1961 8 4
New York	Hillary Clinton	Democratic	October 26, 1947	2001 - 2009	1947 10 26
Tennessee	Al Gore	Democratic	March 31, 1948	1985 - 1993	1948 3 31
North Carolina	John Edwards	Democratic	June 10, 1953	1999 - 2005	1953 6 10
Kansas	Bob Dole	Republican	July 22, 1923	1969 - 1996	1923 7 22
Indiana	Dan Quayle	Republican	February 4, 1947	1981 - 1989	1947 2 4

Application of Typicality (3)

- Short Text Understanding (IJCAI’11)



UNDERSTANDING

悟 & 懂

Term	Entity	Attribute
China	●	○
Russia	●	○
	●	○
	●	○
	●	○

□ unknown type □ noise tolerant

I think you are talking about country

UNDERSTANDING

悟 & 懂

Term	Entity	Attribute
China	●	○
Russia	●	○
India	●	○
	●	○
	●	○

□ unknown type □ noise tolerant

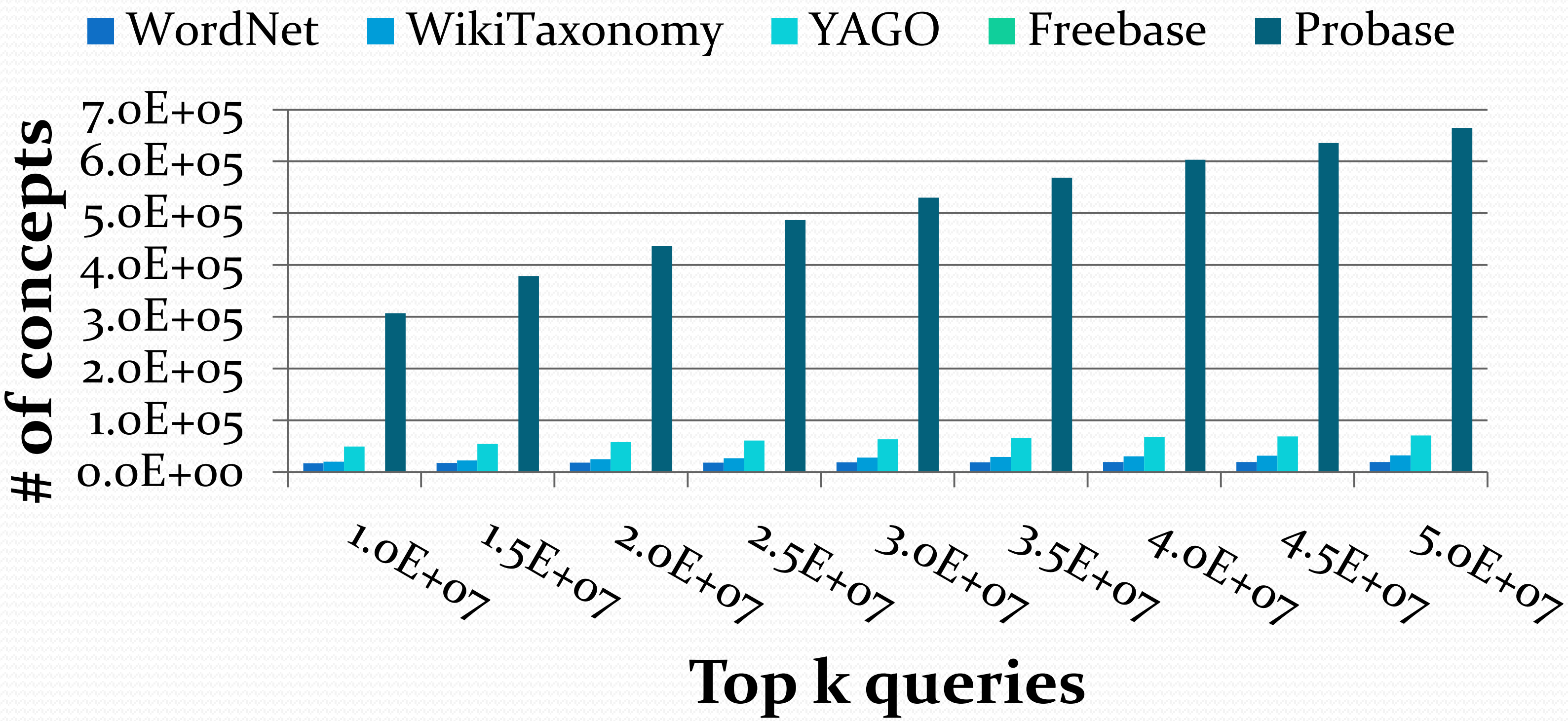
I think you are talking about emerging market

Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion

Concept Space

- A concept is *relevant* if it appears at least once in the top 50 million popular queries in Bing's query log.



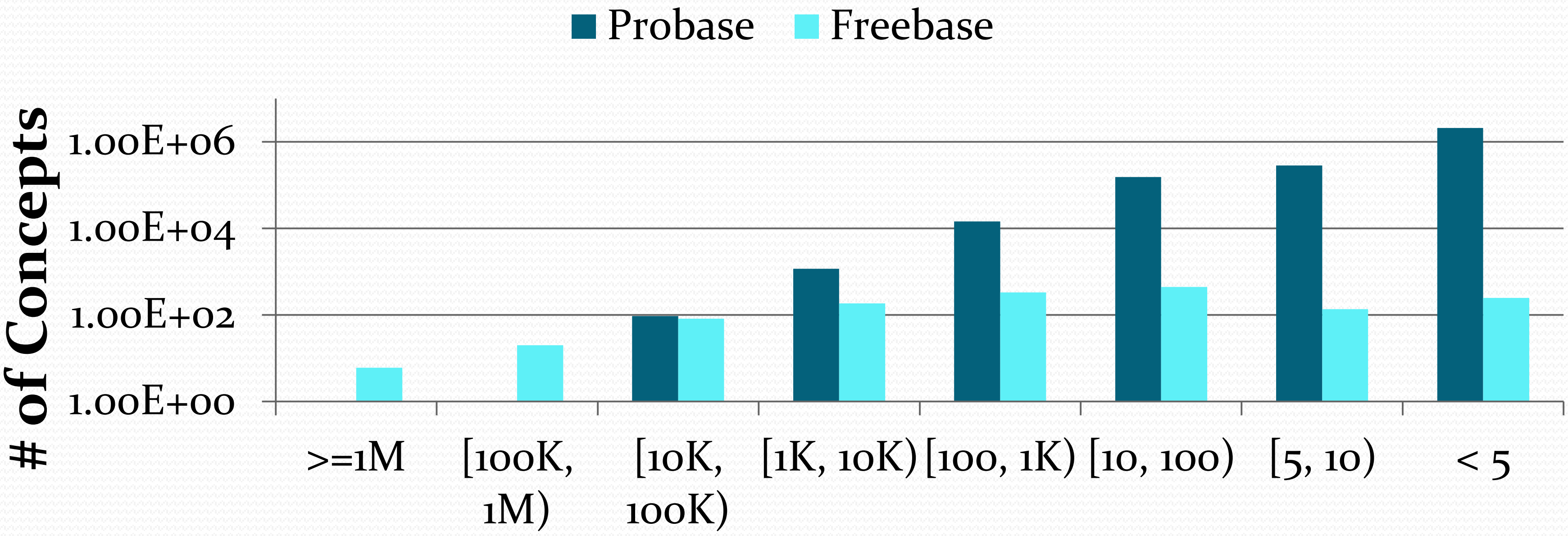
IsA Relationship Space (1)

- The Concept-Subconcept Relationship Space

	# of *isA* pairs	Avg # of children	Avg # of parents	Avg level	Max level
Probase	**4,539,176**	**7.53**	**2.33**	**1.086**	**7**
WordNet	283,070	11.0	2.4	1.265	14
WikiTaxonomy	90,739	3.7	1.4	1.483	15
YAGO	366,450	23.8	1.04	1.063	18
Freebase	0	0	0	1	1

IsA Relationship Space (2)

- The Concept-Instance Relationship Space

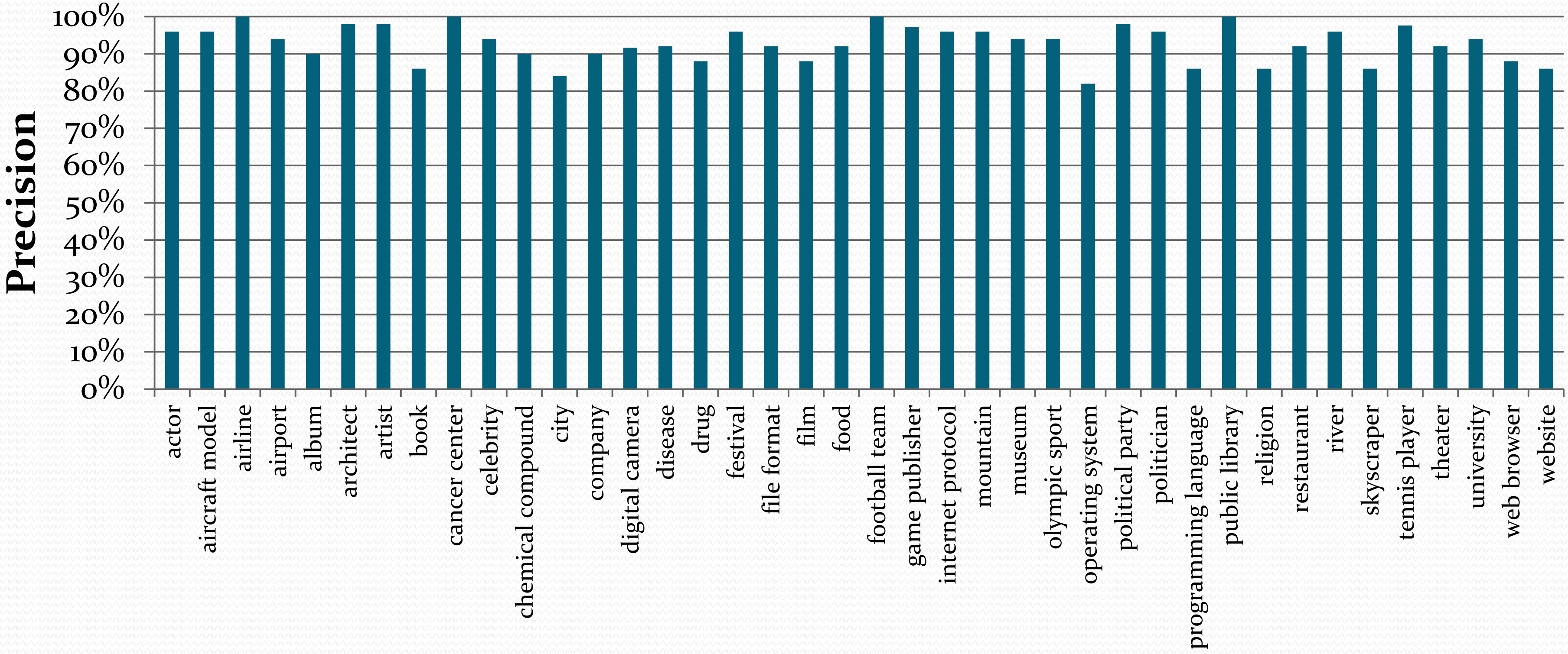


Interval of Concept Size

Concept Size Distribution in Probase v.s. Freebase

Precision of the Extracted Pairs

- 92.4% precision in average over the 40 benchmark concepts.



Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion

Conclusion

- We present a novel iterative extraction framework to extract the isA relationships from text.

- We present a novel taxonomy construction framework based on merging concepts by their senses.

- We use the above techniques to build Probase, which is currently the largest taxonomy in terms of concepts.

- We present a novel probabilistic approach to model the plausibility and typicality of the facts in Probase, and demonstrate its effectiveness in important text understanding applications.

Q & A

Thank you ☺

Please visit our website:
http://research.microsoft.com/probase/
for more information about Probase!

Backup Slides

Algorithm Outline (Extraction)

- **Input**: *S*, the set of sentences matching Hearst Patterns
- **Output**: *Γ*, the set of *isA* pairs

```
Repeat
    foreach $s$ in $S$ do
        $X_s, Y_s \leftarrow SyntacticExtraction(s)$;
        if $|X_s|>1$: $X_s \leftarrow SuperConceptDetection(X_s, Y_s, \Gamma)$;
        if $|X_s|=1$: $Y_s \leftarrow SubConceptDetection(X_s, Y_s, \Gamma)$;
                add valid isA pairs to $\Gamma$;
    end
Until no new pairs added into $\Gamma$;
Return $\Gamma$;
```

Syntactic Extraction

- Challenges
  - … animals other than dogs **such as** *cats* …
  - … classic movies **such as** *Gone with the Wind* …
  - … companies **such as** *IBM*, *Nokia*, *Proctor* **and** *Gamble* …

- Strategy
  - Use "," as the delimiter to obtain the candidates.
  - For the *last* element, also use "and" and "or" to break it down.

Super-Concept Detection

- Find the most likely super-concept among the candidates.

$$r(x_1,x_2)=\frac{p(x_1\,|\,Y_s)}{p(x_2\,|\,Y_s)}=\frac{p(Y_s\,|\,x_1)\,p(x_1)}{p(Y_s\,|\,x_2)\,p(x_2)}$$

Pick x_1 if $r\ (x_1, x_2) > \varepsilon$

$\Downarrow$ Assuming independence of y_i's

$$r(x_1,x_2)=\frac{p(x_1)\prod_{i=1}^{n}p(y_i\,|\,x_1)}{p(x_2)\prod_{i=1}^{n}p(y_i\,|\,x_2)}$$

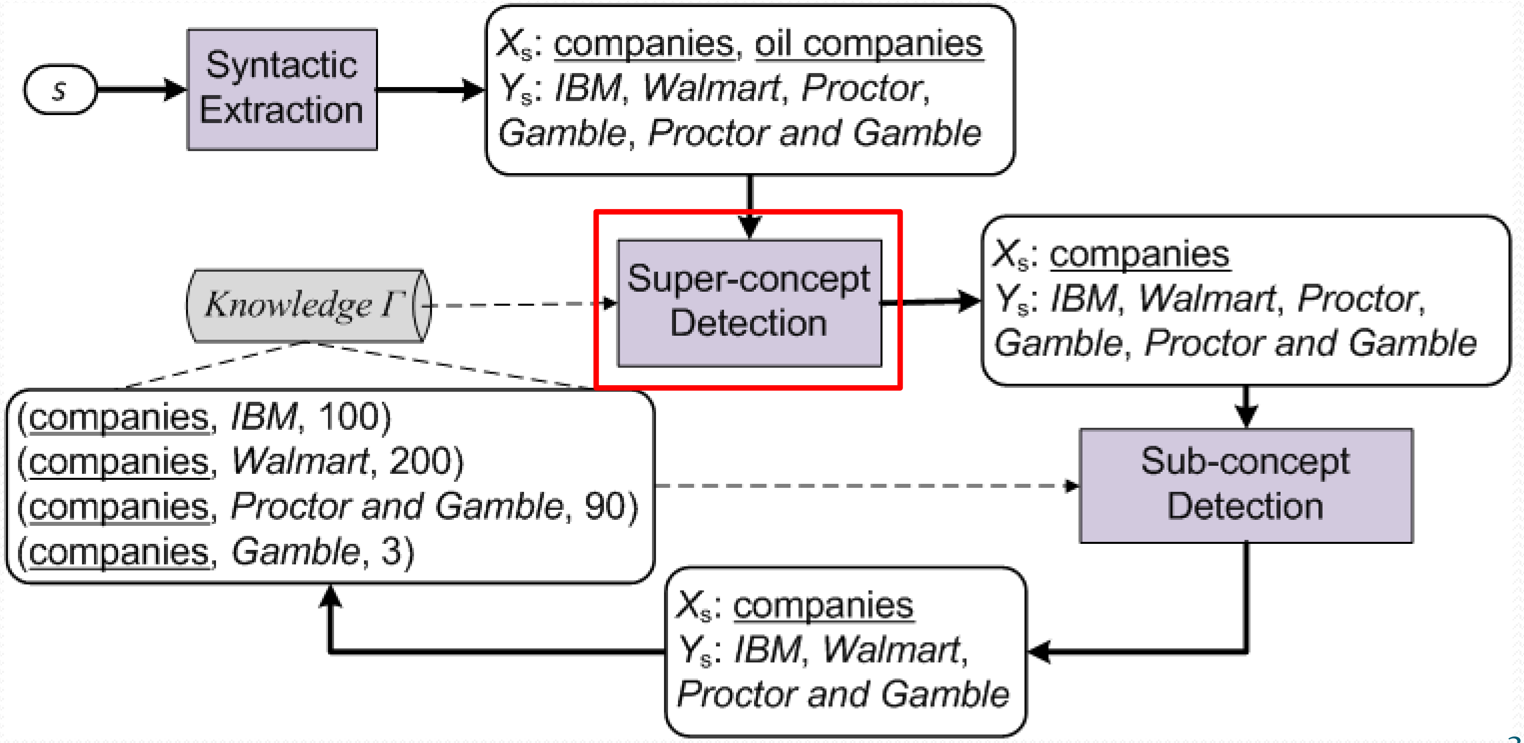
1) Y_s is the set of sub-concepts of the sentence s.
2) $p\ (y_i\,|\,x_1) = p(x_1, y_i)\ /\ p(x_1) = n(x_1, y_i)\ /\ n(x_1)$.

We maintain a count $n(x, y)$ for each (x, y) in Γ.

Super-Concept Detection (Ex)

$$\left.\begin{array}{l} r(x_1,x_2) = \dfrac{p(x_1 \mid Y_s)}{p(x_2 \mid Y_s)} = \dfrac{p(Y_s \mid x_1)\, p(x_1)}{p(Y_s \mid x_2)\, p(x_2)} \\ p\,(y_i \mid x_1) = p(x_1, y_i) \,/\, p(x_1) = n(x_1, y_i) \,/\, n(x_1) \end{array}\right\}$$

r (companies, oil companies)



s

Syntactic Extraction

X_s: companies, oil companies
Y_s: *IBM*, *Walmart*, *Proctor*, *Gamble*, *Proctor and Gamble*

Knowledge Γ

Super-concept Detection

X_s: companies
Y_s: *IBM*, *Walmart*, *Proctor*, *Gamble*, *Proctor and Gamble*

(companies, *IBM*, 100)
(companies, *Walmart*, 200)
(companies, *Proctor and Gamble*, 90)
(companies, *Gamble*, 3)

Sub-concept Detection

X_s: companies
Y_s: *IBM*, *Walmart*, *Proctor and Gamble*

Sub-Concept Detection (1)

- Find the valid sub-concepts among the candidates.

Observation 1. The *closer* a candidate sub-concept is to the **pattern keywords**, the more likely it is a valid sub-concept.

Observation 2. If we are certain a candidate sub-concept at the *k*-th position from the **pattern keywords** is valid, then most likely candidate sub-concepts from position 1 to position *k*-1 are also valid.

E.g., … representatives in North America, Europe, the Middle East, *Australia*, *Mexico*, *Brazil*, *Japan*, *China*, **and other** countries.

Sub-Concept Detection (2)

- Strategy
  - Find the largest scope wherein sub-concepts are all valid:

    find the maximum k s.t. $p\,(y_k \,|\, x) > \varepsilon'$
  - Address the ambiguity issues inside the scope $y_1, \ldots, y_k$:

$$r(c_1, c_2) = \frac{p(c_1 \,|\, x, y_1, \Lambda\ , y_{j-1})}{p(c_2 \,|\, x, y_1, \Lambda\ , y_{j-1})}$$

Suppose that y_j is ambiguous with two candidates c_1 and c_2.

$$\Big\Downarrow$$

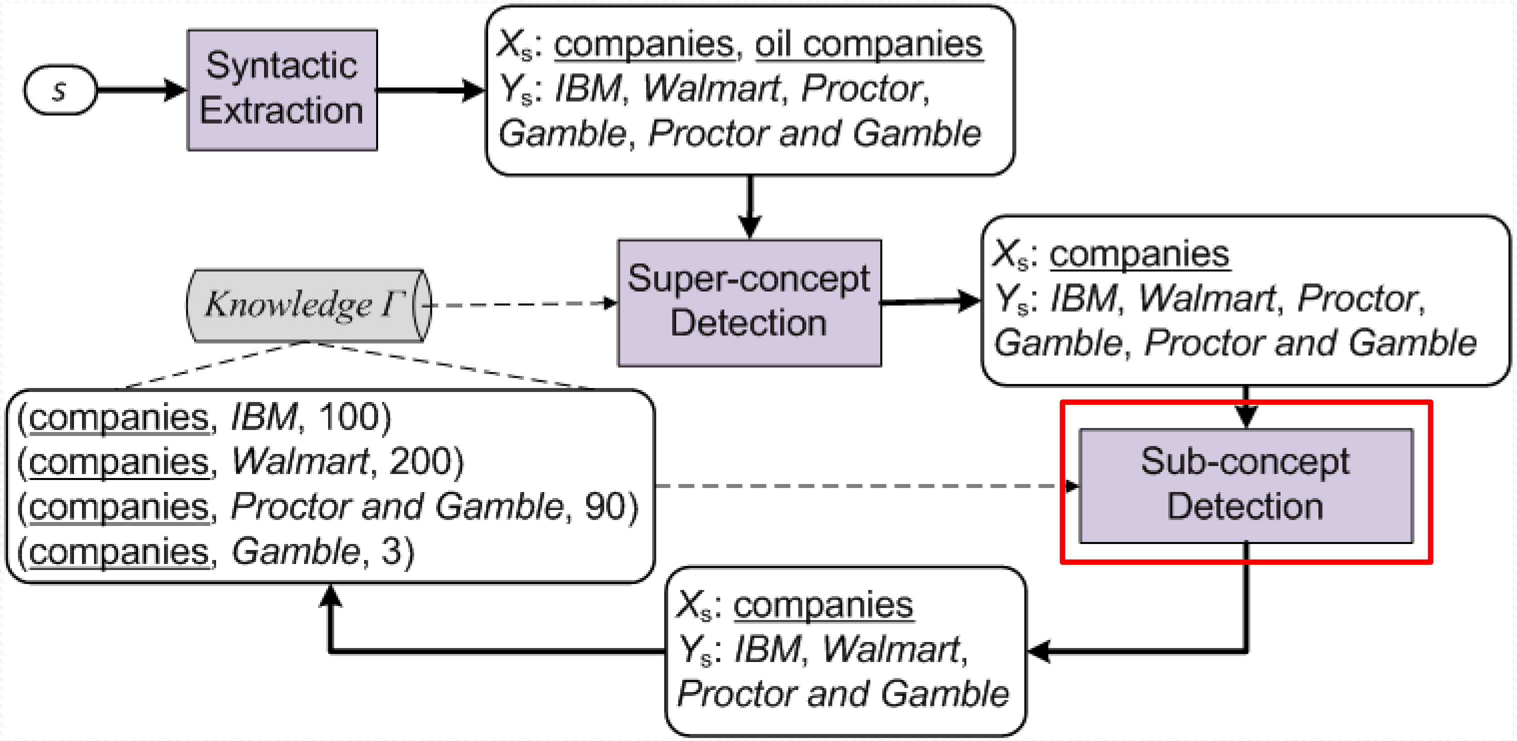
Assuming independence of y_i's

$$r(c_1, c_2) = \frac{p(c_1 \,|\, x) \prod_{i=1}^{j-1} p(y_i \,|\, c_1, x)}{p(c_2 \,|\, x) \prod_{i=1}^{j-1} p(y_i \,|\, c_2, x)}$$

Pick c_1 if $r\,(c_1, c_2) > \varepsilon''$

Sub-Concept Detection (Ex)

$$r(c_1, c_2) = \frac{p(c_1 \mid x, y_1, \Lambda\ , y_{j-1})}{p(c_2 \mid x, y_1, \Lambda\ , y_{j-1})}$$ → r (Proctor and Gamble, Proctor)



s

Syntactic Extraction

X_s: companies, oil companies
Y_s: *IBM*, *Walmart*, *Proctor*, *Gamble*, *Proctor and Gamble*

Knowledge Γ

Super-concept Detection

X_s: companies
Y_s: *IBM*, *Walmart*, *Proctor*, *Gamble*, *Proctor and Gamble*

(companies, *IBM*, 100)
(companies, *Walmart*, 200)
(companies, *Proctor and Gamble*, 90)
(companies, *Gamble*, 3)

Sub-concept Detection

X_s: companies
Y_s: *IBM*, *Walmart*, *Proctor and Gamble*

Properties of "Such As" (1)

Property 1. Let s = $\{(x, y_1), \ldots, (x, y_n)\}$ be the *isA* pairs derived from a sentence . Then, all the *x*'s in *s* have a unique sense, that is, there exists a unique *i* such that $(x, y_j) \mid= (x^i, y_j)$ holds for all $1 \leq j \leq n$.

- Example:
  - … plants **such as** *trees* **and** *grass* …
  - … plants **such as** *steam turbines*, *pumps*, **and** *boilers* …

But sentences like "… plants **such as** *trees* **and** *boilers* …" are extremely rare.

Properties of "Such As" (2)

Property 2. Let $\{(x^i, y_1), \ldots, (x^i, y_m)\}$ denote pairs from one sentence, and $\{(x^j, z_1), \ldots, (x^j, z_n)\}$ from another sentence. If $\{y_1, \ldots, y_m\}$ and $\{z_1, \ldots, z_n\}$ are similar, then it is highly likely that x^i and x^j are equivalent, that is, i = j.

- Example:

  a) … plants **such as** *trees* **and** *grass* …

  b) … plants **such as** *trees*, *grass* **and** *herbs* …

The "plants" in a) and b) are highly likely to have the same sense.

Properties of "Such As" (3)

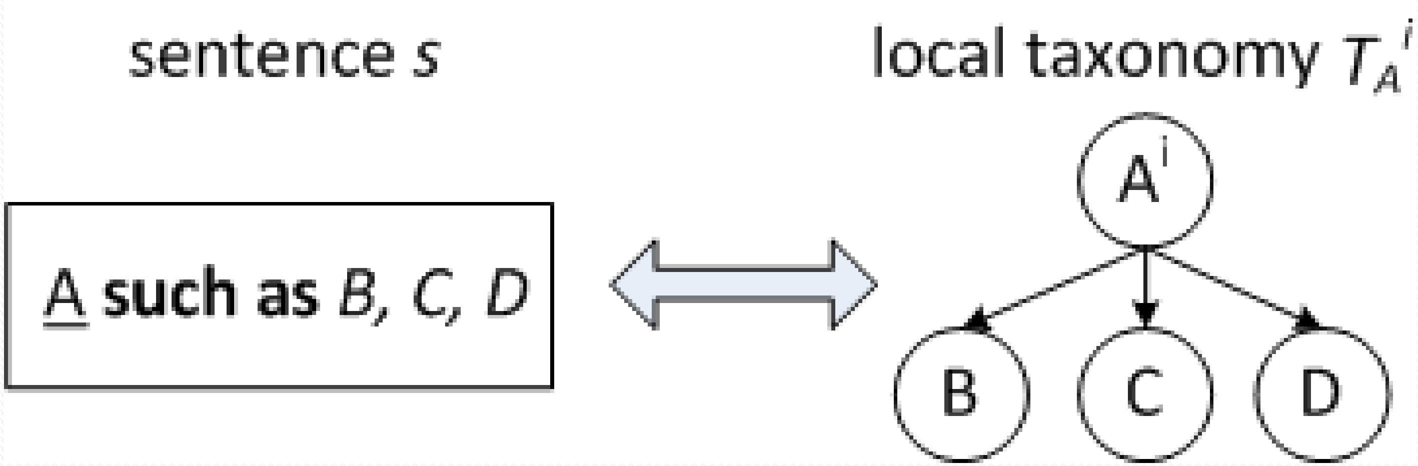
Property 3. Let $\{(x^i, y), (x^i, u_1), \ldots, (x^i, u_m)\}$ denote pairs obtained from one sentence, and $\{(y^k, v_1), \ldots, (y^k, v_n)\}$ from another sentence. If $\{u_1, u_2, \ldots, u_\mathrm{m}\}$ and $\{v_1, v_2, \ldots, v_n\}$ are similar, then it is highly likely that $(x^i, y) \models (x^i, y^k)$.

- Example:
  - a) … organisms **such as** *plants*, *trees*, *grass* **and** *animals* …
  - b) … plants **such as** *trees*, *grass*, **and** *shrubs* …
  - c) … plants **such as** *steam turbines*, *pumps*, **and** *boilers* …

The "plants" in a) and b) are highly likely to have the same sense, but not the "plants" in a) and c).

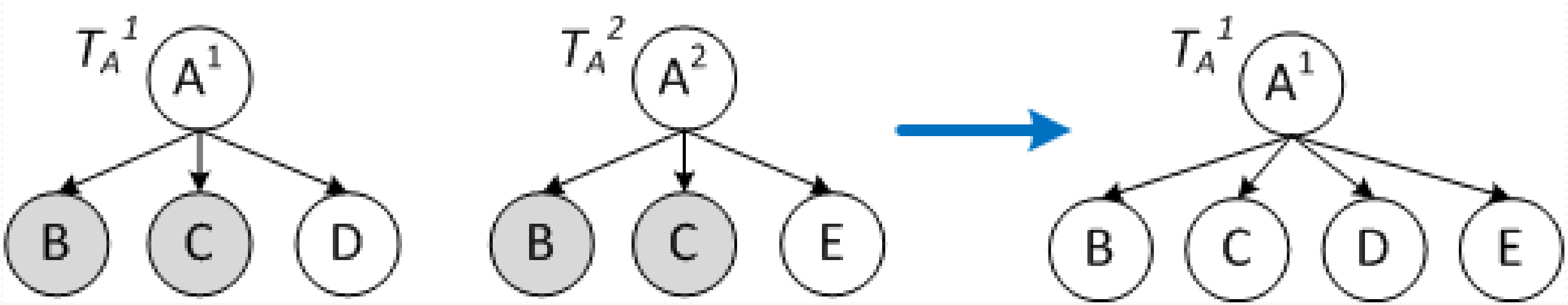
Local Taxonomy

- Based on Property 1



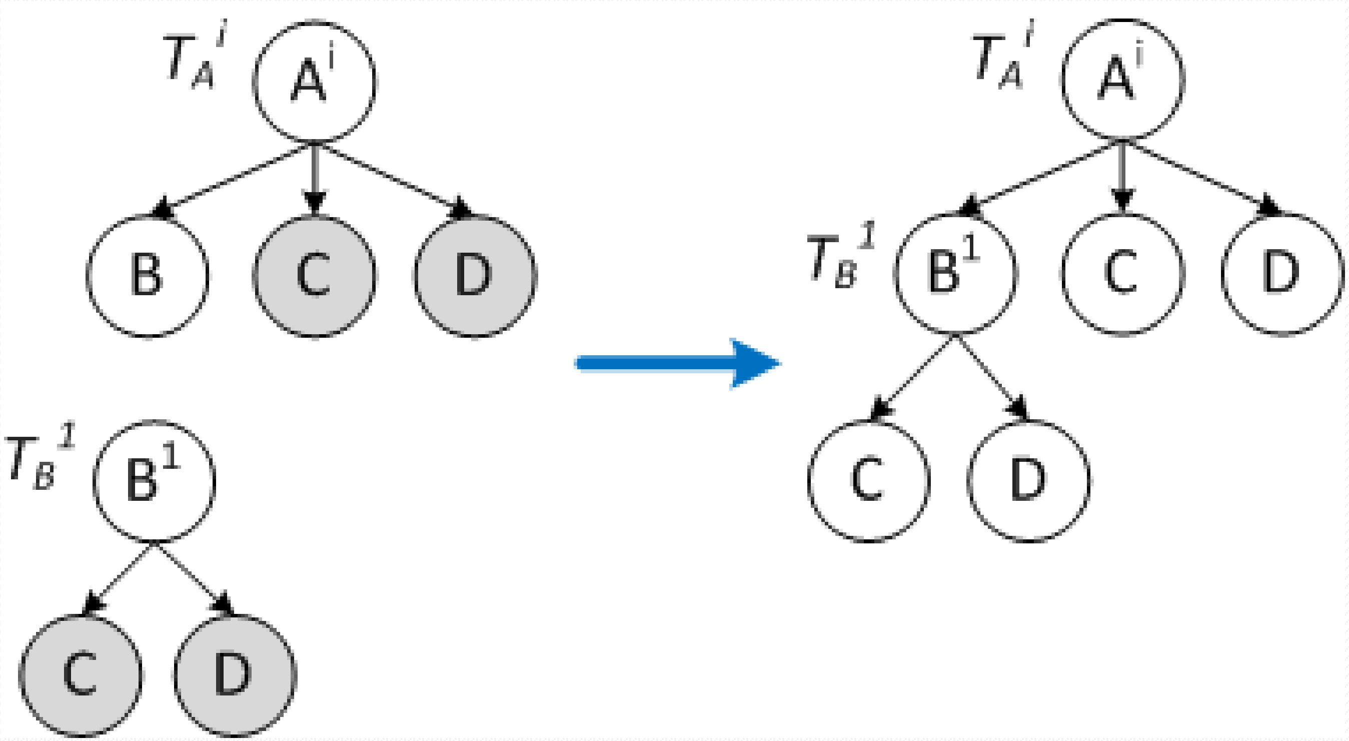
Horizontal Merge

- Based on Property 2



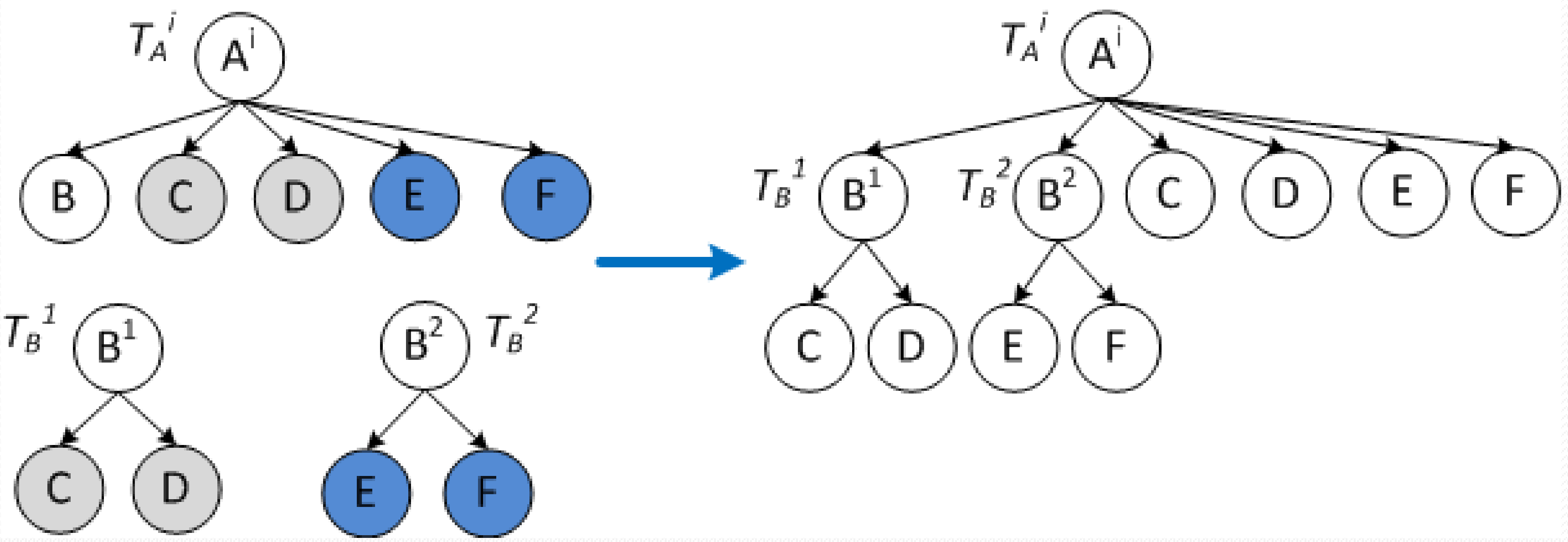
Vertical Merge (1)

- Single Sense Alignment (Based on Property 3)



Vertical Merge (2)

- Multiple Sense Alignment (Based on Property 3)



Similarity Function

- We favor the similarity $f(A, B)$ to be measured by the *absolute* overlap of the two sets A and B.
  - Similarity based on *relative* overlap such as Jaccard similarity will raise weird results (see the paper for an example).

- More generally, the similarity function is desired to have the following *closure* property:

Property 4. If A, A', B, and B' are any sets s. t. $A \subseteq A'$ and $B \subseteq B'$, then $Sim(A, B)$ => $Sim(A', B')$.

Algorithm Outline (Construction)

- **Input**: *S*, the set of sentences with extracted *isA* pairs
- **Output**: *T*, the taxonomy graph

Stage 1: For each *s* in *S*, construct a *local taxonomy*.

Stage 2: Perform all possible *horizontal* merges.

Stage 3: Perform all possible *vertical* merges.

Return the graph *T* after the 3 stages

Theoretical Results

Theorem 1. Let T be a set of local taxonomies. Let $\mathbf{O}^{\alpha}$ and $\mathbf{O}^{\beta}$ be any two sequences of horizontal and vertical merge operations on T. Assume no further operations can be performed on T after $\mathbf{O}^{\alpha}$ or $\mathbf{O}^{\beta}$. Then, the final graph after performing $\mathbf{O}^{\alpha}$ and the final graph after performing $\mathbf{O}^{\beta}$ are identical.

Theorem 2. Let O be the set of all possible sequences of operations, and let $M = \min\{|\mathbf{O}| : \mathbf{O} \in O\}$. Suppose $\mathbf{O}^{\sigma}$ is the sequence that performs all possible horizontal merges first and all possible vertical merges next, then $|\mathbf{O}^{\sigma}| = M$.

Applications of Typicality (1)

- Semantic Web Search

ACM fellows *working on semantic web*

database conferences *in* asian cities

*Are you interested in the **text** or instances of* “ACM fellows”, “database conferences” *and* “asian cities”?

Applications of Typicality (2)

- Short Text Understanding (Y. Song et al. *IJCAI'11*)
  - Conceptualize from a set of words by performing Bayesian analysis based on the (inverse) typicality $T(x|i)$.

Example: India => country / region
India, China => Asian country / developing country
India, China, Brazil => BRIC / emerging market

  - Cluster Twitter messages based on conceptualization signals of words.

Concept Space (1)

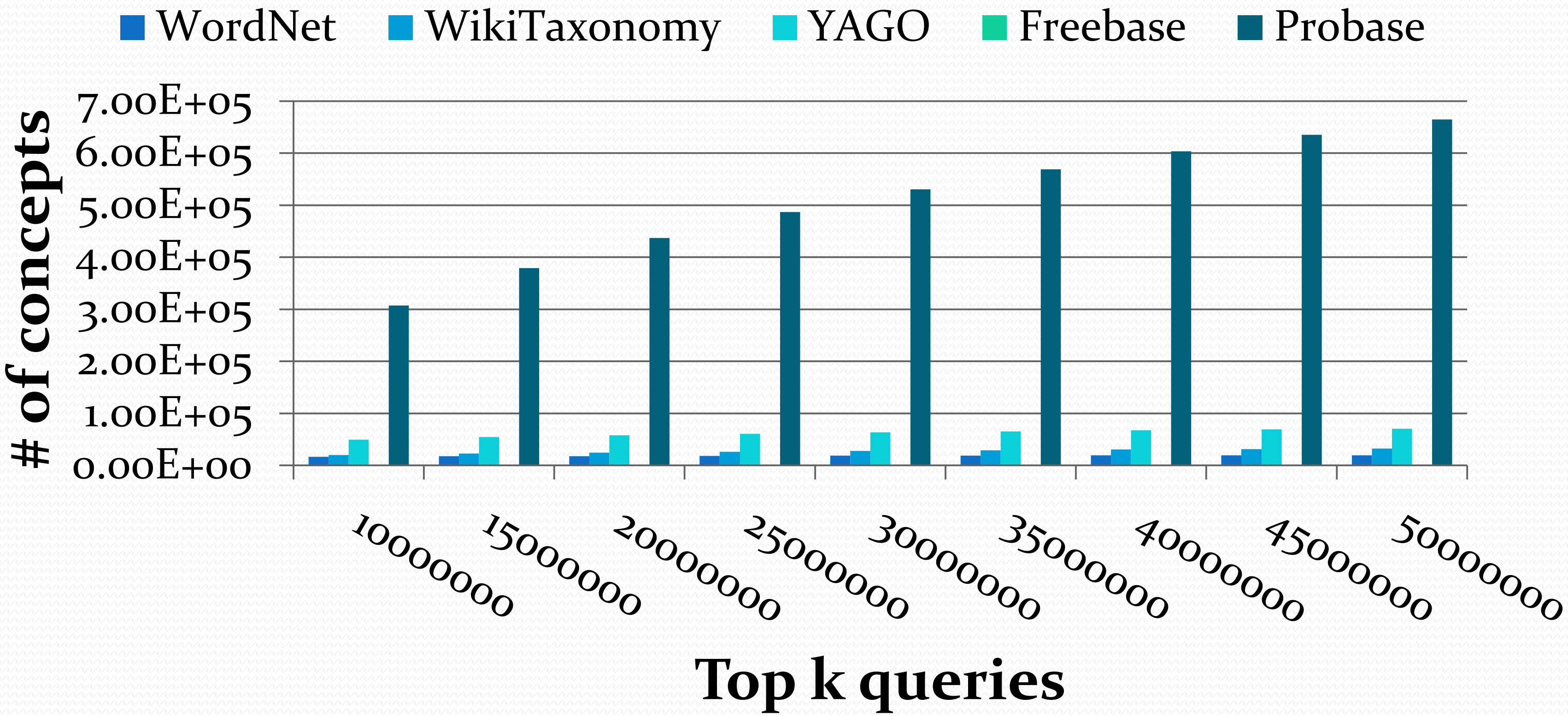
- Probase contains more then 2.6 million concepts. Are they useful?

- Evaluate this using the top 50 million popular queries in Bing's query log from a 2-year period.

- Metrics in the evaluation
  - *Relevance*
  - *Taxonomy Coverage*
  - *Concept Coverage*

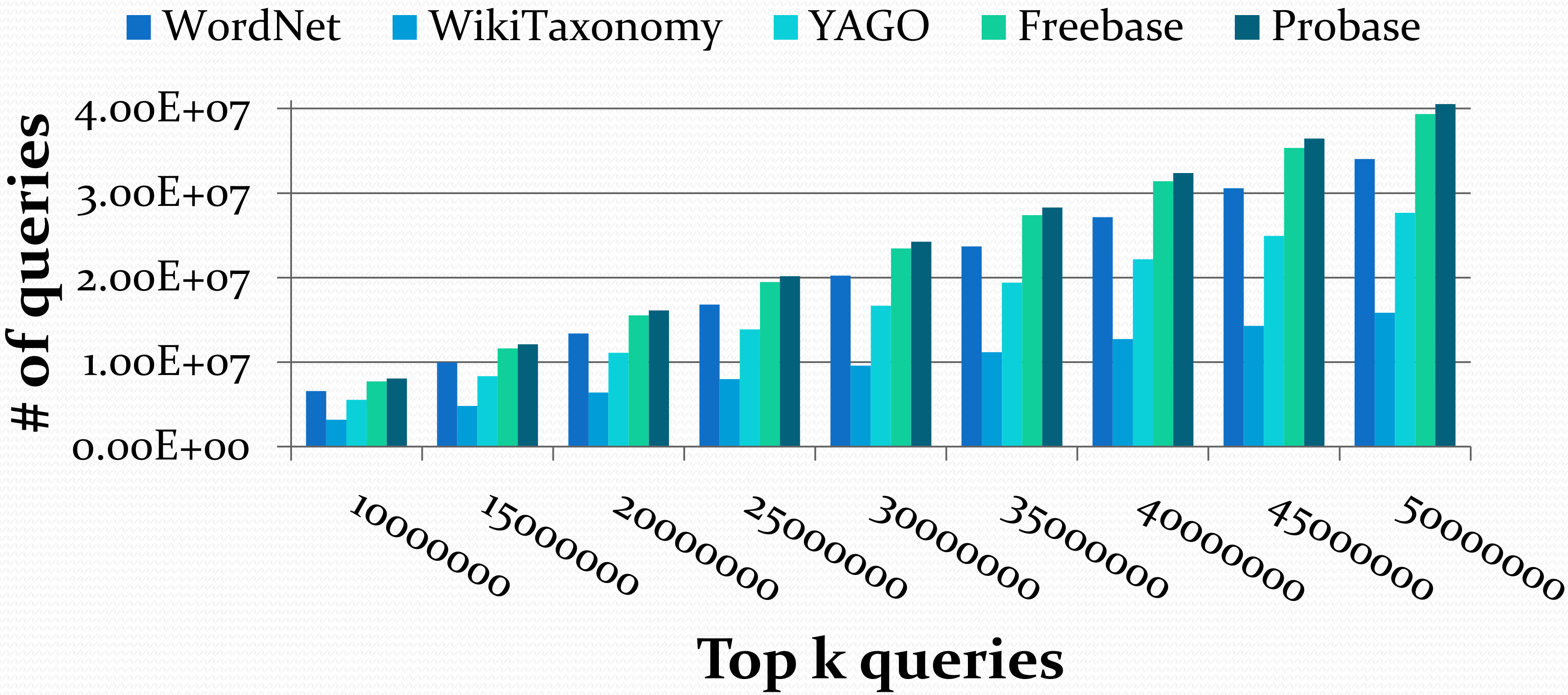
Concept Space (2)

- Relevance: A concept is relevant if it appears at least once.



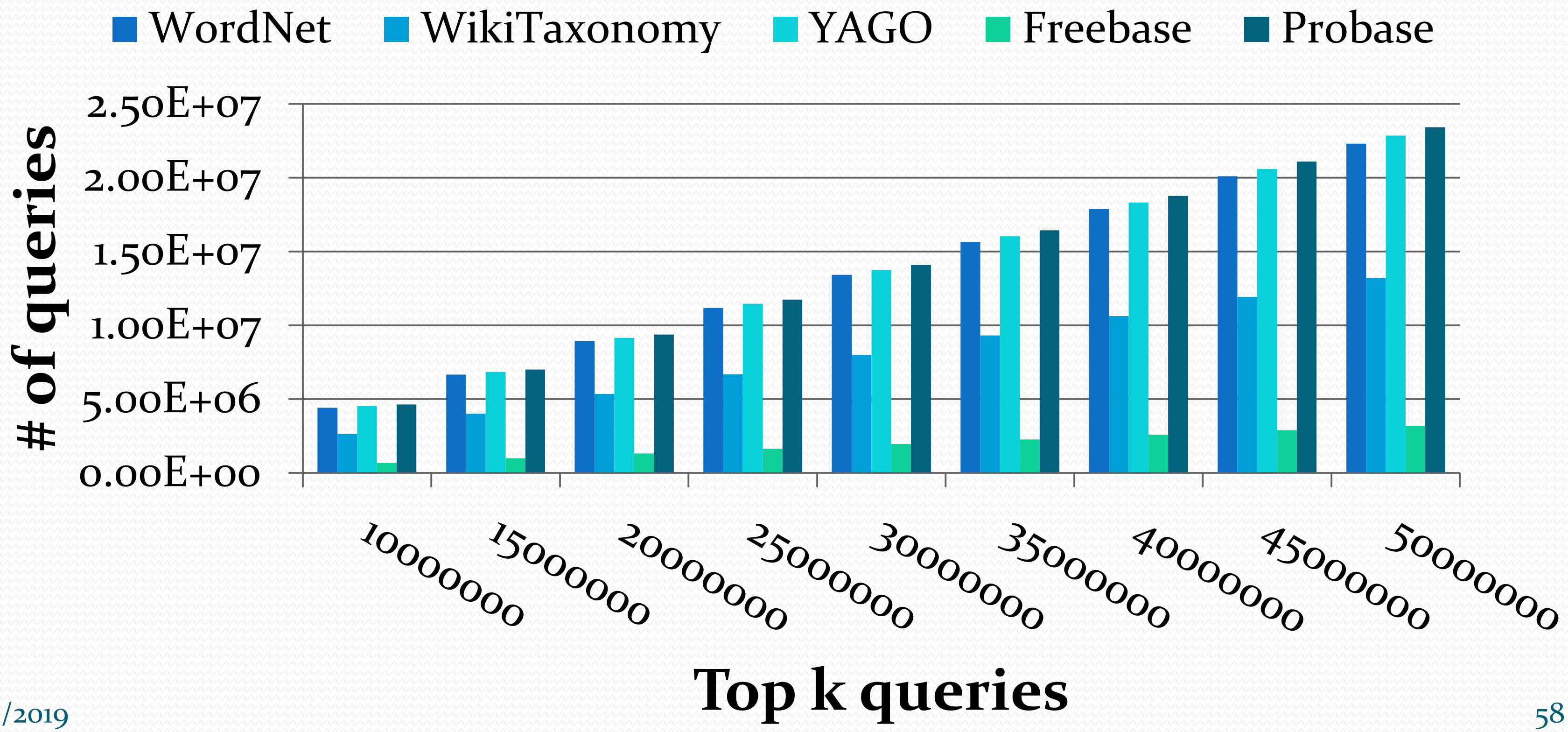
Concept Space (3)

- Taxonomy Coverage: A query is covered if it contains at least one concept *or* instance in the taxonomy.



Concept Space (4)

- Concept Coverage: A query is covered if it contains at least one concept in the taxonomy.


■ WordNet
■ WikiTaxonomy
■ YAGO
■ Freebase
■ Probase
# of queries
2.50E+07
2.00E+07
1.50E+07
1.00E+07
5.00E+06
0.00E+00
Top k queries