Probase:
A Knowledge Base for Text Understanding

Haixun Wang
Microsoft Corporation
Pablo Picasso | 25 Oct 1881 | Spanish
... animals other than cats such as dogs ...
... household pets other than animals such as reptiles, aquarium fish ...
ProBase

1. More than 2.7 million concepts automatically harnessed from 1.68 billion documents

2. Computation/Reasoning enabled by scoring:
   - Consensus: e.g., is there a company called Apple?
   - Typicality: e.g., how likely you think of Apple when you think about companies?
   - Ambiguity: e.g., does the word Apple, sans any context, represent Apple the company?
   - Similarity: e.g., how likely is an actor also a celebrity?
   - Freshness: e.g., Pluto as a dwarf planet is a claim more fresh than Pluto as a planet.

3. Give machines a new CPU (Commonsense Processing Unit) powered by a distributed graph engine called Trinity.

4. A little knowledge goes a long way after machines acquire a human touch.
Probase Internals

artist

painter

Picasso

Born 1881  Died 1973  Movement Cubism

art

painting

Guernica

Year 1937  Type Oil on Canvas
2.7 million concepts

Basic watercolor techniques

Countries

Celebrity wedding dress designers
Data Sources

Patterns for single statements

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

Examples:
- Good: “rich countries such as USA and Japan …”
- Tough: “animals other than cats such as dogs …”
- Hopeless: “At Berklee, I was playing with cats such as Jeff Berlin, Mike Stern, Bill Frisell, and Neil Stubenhaus.”
Properties

- Given a **class**, find its properties

- Candidate seed properties:
  - “**What** is the [property] of [instance]?”
  - “**Where**, **When**, **Who**” are also considered
Similarity between two concepts

- Weighted linear combinations of
  - Similarity between the set of instances
  - Similarity between the set of attributes
- (nation, country)
- (celebrity, well-known politicians)
Beyond noun phrases

Example: the verb “hit”

- Small object, Hard surface
  - (bullet, concrete), (ball, wall)

- Natural disaster, Area
  - (earthquake, Seattle), (Hurricane Floyd, Florida)

- Emergency, Country
  - (economic crisis, Mexico), (flood, Britain)
Quantify Uncertainty

**Typicality**
- \( P(\text{concept} \mid \text{instance}) \)
- \( P(\text{instance} \mid \text{concept}) \)
- \( P(\text{concept} \mid \text{property}) \)
- \( P(\text{property} \mid \text{concept}) \)

**Similarity**
- \( \text{sim}(\text{concept}_1, \text{concept}_2) \)

the foundation of text understanding and reasoning
Text Mining / IE: State of the Art

- Bag of words based approach: e.g., LDA
  - Based on multiple document statistics
  - Simple bag-of-words, no semantics

- Supervised learning: e.g., CRF
  - Labeled training data required
  - Difficulty for out-of-sample features

- Lack of semantics

- What role can a knowledgebase play?
Five of us bought 5 Kinects and posed in front of an Apple store.

Apple store that sells fruits or apple store that sells iPads?
Step by Step Understanding

- Entity abstraction
- Attribute abstraction
- Short text/query (1-5 words) understanding
- Text block/document understanding
Short Text

Challenge:
- Not enough statistics

Applications
- Twitter
- Query/Search Log
- Anchor Text
- Image/video tag
- Document paraphrasing and annotation
## Comparison of Knowledge Bases

<table>
<thead>
<tr>
<th>WordNet</th>
<th>Wikipedia</th>
<th>Freebase</th>
<th>Probase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat</td>
<td>Feline; Felid; Adult male; Man; Gossip; Gossiper; Gossipmonger; Rumormonger; Newsmonger; Woman; Adult female; Stimulant; Stimulant drug; Excitant; Tracked vehicle; ...</td>
<td>Domesticated animals; Cats; Felines; Invasive animal species; Cosmopolitan species; Sequenced genomes; Animals described in 1758; ...</td>
<td>Animal; Pet; Species; Mammal; Small animal; Thing; Mammalian species; Small pet; Animal species; Carnivore; Domesticated animal; Companion animal; Exotic pet; Vertebrate; ...</td>
</tr>
<tr>
<td>IBM</td>
<td>N/A</td>
<td>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; ...</td>
<td>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; ...</td>
</tr>
<tr>
<td>Language</td>
<td>Communication; Auditory communication; Word; Higher cognitive process; Faculty; Mental faculty; Module; Text; Textual matter;</td>
<td>Languages; Linguistics; Human communication; Human skills; Wikipedia articles with ASCII art</td>
<td>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; ...</td>
</tr>
</tbody>
</table>
In the mind of the machine: when it sees the word ‘apple’
... when it sees the words ‘apple’ and ‘pear’ together
Entity Abstraction

In the most likely concept which can generalize all the entities. The top entities in the concept are also not "Russia", "India" and "Brazil", then click 'Abstract' and you can find something interesting.

I think you are talking about **country**

Entities in this concept include

1. China
2. Russia
3. India
4. Canada
5. Australia
6. Japan
7. Germany
8. France
9. Russia
10. Brazil
11. Italy
12. Mexico

Entity Abstraction

In the most likely concept which can generalize all the entities. The top entities in the concept are also not "Russia", "India" and "Brazil", then click 'Abstract' and you can find something interesting.

I think you are talking about **emerging market**

Entities in this concept include

1. China
2. India
3. Russia
4. Brazil
5. Asia
6. Latin America
7. Eastern Europe
8. Africa
9. The Middle East
10. Mexico
11. Turkey
12. South Africa
Entity Abstraction

Given a set of entities

\[ E = \{e_i, i \in 1, \ldots, M\}\]

Target Concept (Naïve Bayes Rule)

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

Where \( c_k \) a concept, and

\[ P(e_i|c_k) = \frac{P(e_i,c_k)}{P(c_k)}, \text{ co-occurrence} \]

is computed by
How to Infer Concept from Attribute?

- Given a set of attributes
- The Naïve Bayes Rule gives

\[
    A = \{a_j, j \in 1, \ldots, N\}.
\]

where

\[
    P(c_k | A) = \frac{P(A | c_k) \cdot P(c_k)}{P(A)} \propto P(c_k) \prod_{j=1}^{N} P(a_j | c_k),
\]

\[
    P(a_j | c_k) = \sum_{i: e_i \in E} P(a_j | e_i) P(e_i | c_k),
\]

- (university, florida state university, 75)
- (university, harvard university, 388)
- (university, university of california, 142)
- (country, china, 97346)
- (country, the united states, 91083)
- (country, india, 80351)
- (country, canada, 74481)
- (florida state university, website, 34)
- (harvard university, website, 38)
- (university of california, city, 12)
- (china, capital, 43)
- (the united states, capital, 32)
- (india, population, 35)
- (canada, population, 21)
- (university, website, 4568)
- (university, city, 2343)
- (country, capital, 4345)
- (country, population, 3234)
- ......
When Type of Term is Unknown:

- Given a set of terms with unknown types \( T = \{ t_l \}, l = 1, \ldots, L \)
- **Generative model**

\[
P(t_l | c_k) = P(t_l | z_l = 1, c_k) P(z_l = 1 | c_k) + P(t_l | z_l = 0, c_k) P(z_l = 0 | c_k)
\]

Using Naive Bayes

\[
P(c_k | T) = \frac{P(T | c_k) P(c_k)}{P(T)} \propto P(c_k) \prod_{l} P(t_l | c_k)
\]

- **Discriminative model** (Noisy-OR)

\[
P(c_k | t_l) = 1 - (1 - P(c_k | t_l, z_l = 1))(1 - P(c_k | t_l, z_l = 0))
\]

And using twice

\[
P(c_k | T) \propto P(c_k) \prod_{l} P(t_l | c_k) \propto \frac{\prod_{l} P(c_k | t_l)}{P(c_k)^{L-1}}
\]

- where \( z_l = 1 \) indicate “entity” and \( z_l = 0 \) indicate “attribute”
When you see attributes …

<table>
<thead>
<tr>
<th>website</th>
<th>president</th>
<th>city</th>
<th>motto</th>
<th>state</th>
<th>type</th>
<th>director</th>
</tr>
</thead>
</table>

![Graph showing the relationship between the number of concepts and the number of attributes.](image-url)
Understanding = Concept Forming

<table>
<thead>
<tr>
<th>apple</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pear</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Short Text Conceptualization

Please input your query/text

What happens to lakes in an area hit by forest fires and floods? Some will glow in the dark.

Parsed text: **area**; **hit**; **forest**; **will**; **glow**; **dark**;

Recommend Attribute:
Recommend Concept:
Recommend Entity:
Clustering Twitter Messages

Problem 1 (unique concepts): use keywords to retrieve tweets in 3 categories:
1. Microsoft, Yahoo, Google, IBM, Facebook
2. cat, dog, fish, pet, bird
3. Brazil, China, Russia, India

Problem 2 (concepts with subtle differences): use keywords to retrieve tweets in 4 categories:
1. United states, American, Canada
2. Malaysia, China, Singapore, India, Thailand, Korea
3. Angola, Egypt, Sudan, Zambia, Chad, Gambia, Congo
4. Belgium, Finland, France, Germany, Greece, Spain, Switzerland
## Comparison Results

Clustering NMI scores on Twitter data.

<table>
<thead>
<tr>
<th>Method</th>
<th>@Problem1</th>
<th>@Problem2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Data</td>
<td>0.215±0.010</td>
<td>0.452±0.076</td>
</tr>
<tr>
<td>LDA (1×Cluster Num)</td>
<td>0.161±0.065</td>
<td>0.114±0.037</td>
</tr>
<tr>
<td>LDA (2×Cluster Num)</td>
<td>0.067±0.022</td>
<td>0.069±0.024</td>
</tr>
<tr>
<td>WordNet</td>
<td>0.195±0.070</td>
<td>0.074±0.074</td>
</tr>
<tr>
<td>Freebase</td>
<td>0.531±0.164</td>
<td>0.204±0.037</td>
</tr>
<tr>
<td>Wikipedia (Category-Link)</td>
<td>0.540±0.077</td>
<td>0.336±0.089</td>
</tr>
<tr>
<td>Wikipedia (ESA)</td>
<td>0.351±0.132</td>
<td>0.340±0.800</td>
</tr>
<tr>
<td>Probase (Top 10)</td>
<td>0.318±0.110</td>
<td>0.490±0.029</td>
</tr>
<tr>
<td>Probase (Top 20)</td>
<td>0.479±0.111</td>
<td>0.555±0.019</td>
</tr>
<tr>
<td>Probase (Top 50)</td>
<td>0.559±0.123</td>
<td><strong>0.632±0.066</strong></td>
</tr>
<tr>
<td>Probase (Top 500)</td>
<td><strong>0.826±0.062</strong></td>
<td>0.301±0.189</td>
</tr>
<tr>
<td>Probase (Top 5000)</td>
<td>0.690±0.176</td>
<td>0.095±0.084</td>
</tr>
</tbody>
</table>
Many Applications …

- **Mapping questions to knowledge**
  - How many people are in China? → entity: China, Attribute: population
  - Where is MSR? → entity: MSR, Attribute: location
  - How long does it take for Asclepius to take effect? → entity: Asclepius, Attribute: pharmaceutical effect

- **Synonym**
  - China national song → entity: China, Attribute: national anthem
  - USA headline → entity: USA, Attribute: news
  - India demographic → entity: India, Attribute: population

- **Misspelling**
  - Japan poulation → entity: Japan, Attribute: population

- **Correlated indirectly**
  - google earth China → entity: China, Attribute: map
  - China dishes → entity: China, Attribute: food
  - what is the exchange rate for UK → entity: UK, Attribute: currency
Summary

- A little knowledge goes a long way

- A concept space large enough to model the concepts in a human mind

- Scores and weights that enable Bayesian reasoning.

- Many applications
Thanks!
## Examples

### Concept and Entity Co-occurrence

| Concept          | Entity  | Co-occurrence | Concept Number | Entity Number | P(e|c)  | P(c|e)  |
|------------------|---------|---------------|----------------|---------------|--------|--------|
| country          | india   | 80905         | 2262485        | 197915        | 0.03576| 0.40879|
| country          | china   | 98517         | 2262485        | 269127        | 0.04354| 0.36606|
| emerging market  | china   | 6556          | 29298          | 269127        | 0.22377| 0.02436|
| emerging market  | india   | 5702          | 29298          | 197915        | 0.19462| 0.02881|
| area             | china   | 2231          | 2525020        | 269127        | 0.00088| 0.00829|
| area             | india   | 1797          | 2525020        | 197915        | 0.00071| 0.00908|

### Concept and Attribute

| Concept          | Attribute | P(c, a) | P(c)         | P(a)         | P(a|c)  | P(c|a)  |
|------------------|-----------|---------|--------------|--------------|--------|--------|
| country          | population| 4.08183 | 173.44931    | 41736.78060  | 0.02353| 0.00010|
| country          | language  | 1.48795 | 173.44931    | 58584.50905  | 0.00858| 0.00003|
| emerging market  | language  | 4.52949 | 402.13772    | 58584.50905  | 0.01126| 0.00008|
| emerging market  | population| 16.54701| 402.13772    | 41736.78060  | 0.04115| 0.00040|
Examples

Given “china”, “india”, “language” and “population”, “emerging market” will be ranked as 1st

| Concept          | Entity    | Co-occurrence | Concept Number | Entity Number | P(e|c)  | P(c|e)  |
|------------------|-----------|---------------|----------------|---------------|--------|--------|
| country          | india     | 80905         | 2262485        | 197915        | 0.03576| 0.40879|
| country          | china     | 98517         | 2262485        | 269127        | 0.04354| 0.36606|
| emerging market  | china     | 6556          | 29298          | 269127        | 0.22377| 0.02436|
| emerging market  | india     | 5702          | 29298          | 197915        | 0.19462| 0.02881|
| area             | china     | 2231          | 2525020        | 269127        | 0.00088| 0.00829|
| area             | india     | 1797          | 2525020        | 197915        | 0.00071| 0.00908|

| Concept          | Attribute | P(c, a) | P(c)  | P(a)  | P(a|c)  | P(c|a)  |
|------------------|-----------|---------|-------|-------|--------|--------|
| factor           | population| 75.74704| 71073.46656| 41736.78060| 0.00107| 0.00181|
| factor           | language  | 113.32628| 71073.46656| 58584.50905| 0.00159| 0.00193|
| countries        | population| 4.08183 | 173.44931| 41736.78060| 0.02353| 0.00010|
| countries        | language  | 1.48795  | 173.44931| 58584.50905| 0.00858| 0.00003|
| emerging market  | language  | 4.52949  | 402.13772| 58584.50905| 0.01126| 0.00008|
| emerging market  | population| 16.54701| 402.13772| 41736.78060| 0.04115| 0.00040|
Example (Cont’d)

(a) China (I), Russia (I), India (I), Brazil (I)
(b) China (I), India (I), Japan (I), Singapore (I)
(c) population (A), location (A), president (A)
(d) California (U), Florida (U), population (U)
(e) China (U), Brazil (U), Russia (U), apple (U), banana (U), BBC (U), New York Time (U)