Microsoft

Conceptualization for Short Text Understanding

Zhongyuan Wang (王仲远)

***Joint work with Haixun Wang, Jun Yan, Yanghua Xiao, Ji-Rong Wen, and many interns**

微软亚洲研究院 数据挖掘与企业智能组
Data Mining and Enterprise Intelligence Group
Microsoft Research Asia

Dec 27, 2015

Short Text

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

Short text is *sparse*, *noisy*, *and* *ambiguous*

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 AAAS

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

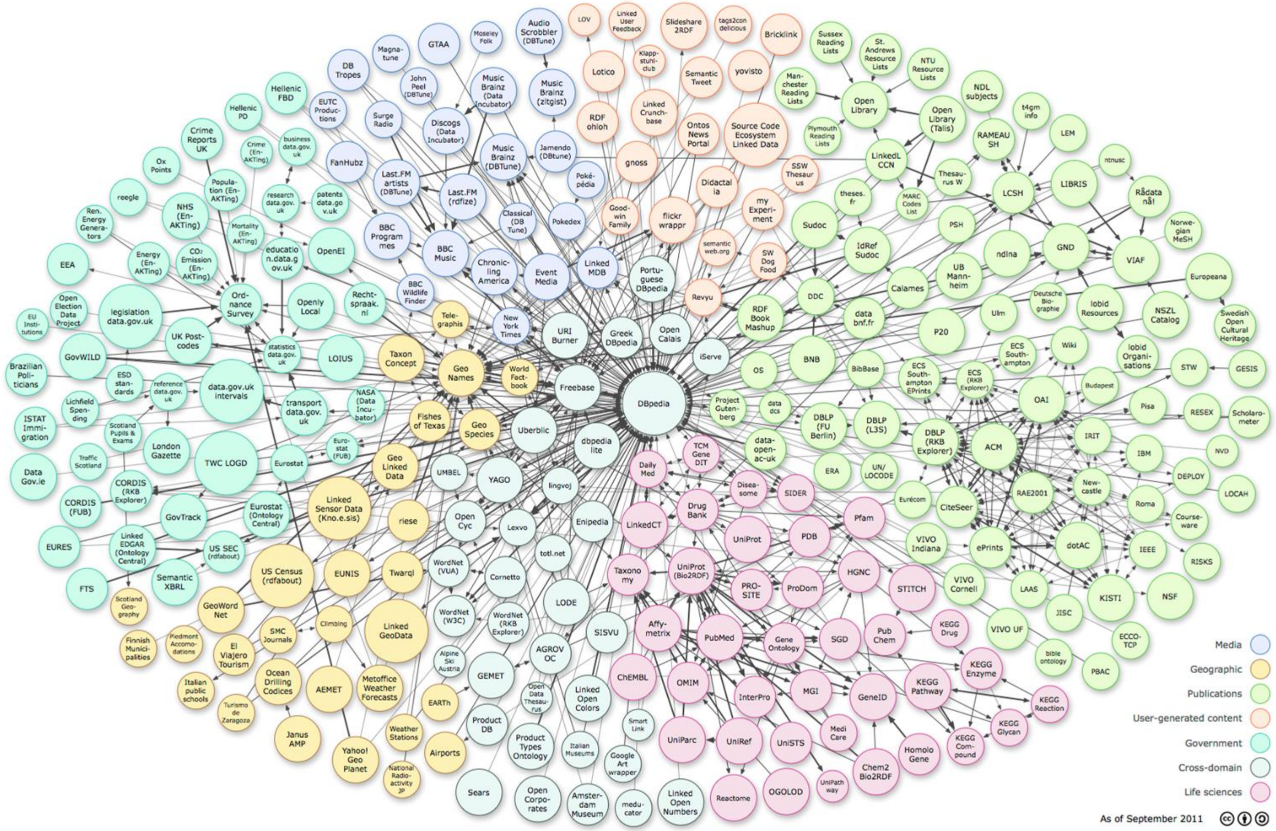
How to Grow a Mind: Statistics, Structure, and Abstraction

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.

Knowledge Base Efforts



http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.png

1. *“Python Tutorial”*
2. *“Who was the U.S. President when the Angels won the World Series?”*

≠

Semantic Network knowledge

knowledge **knowledgebase**

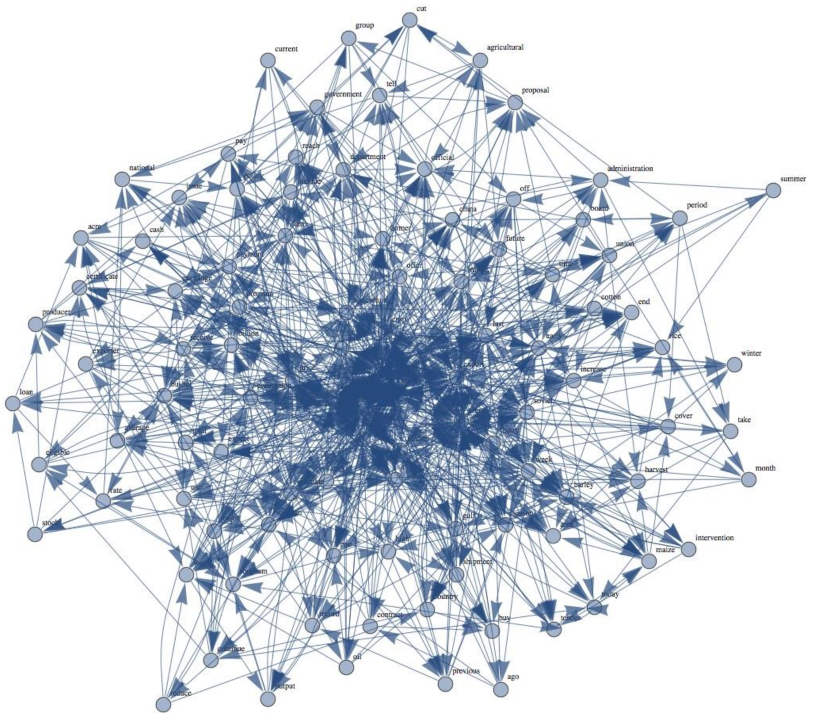
question → understanding (internal representation) → answer

Semantic Network vs. Knowledgebase

Semantic Network	Knowledgebase
Common/linguistic knowledge	Entities Facts
isA isPropertyOf co-occurrence …	DayOfBirth LocatedIn SpouseOf …
Typicality, basic level of categorization	Black or White Precision
KnowItAll, Probase	Freebase, Yago

Probase: A Semantic Network

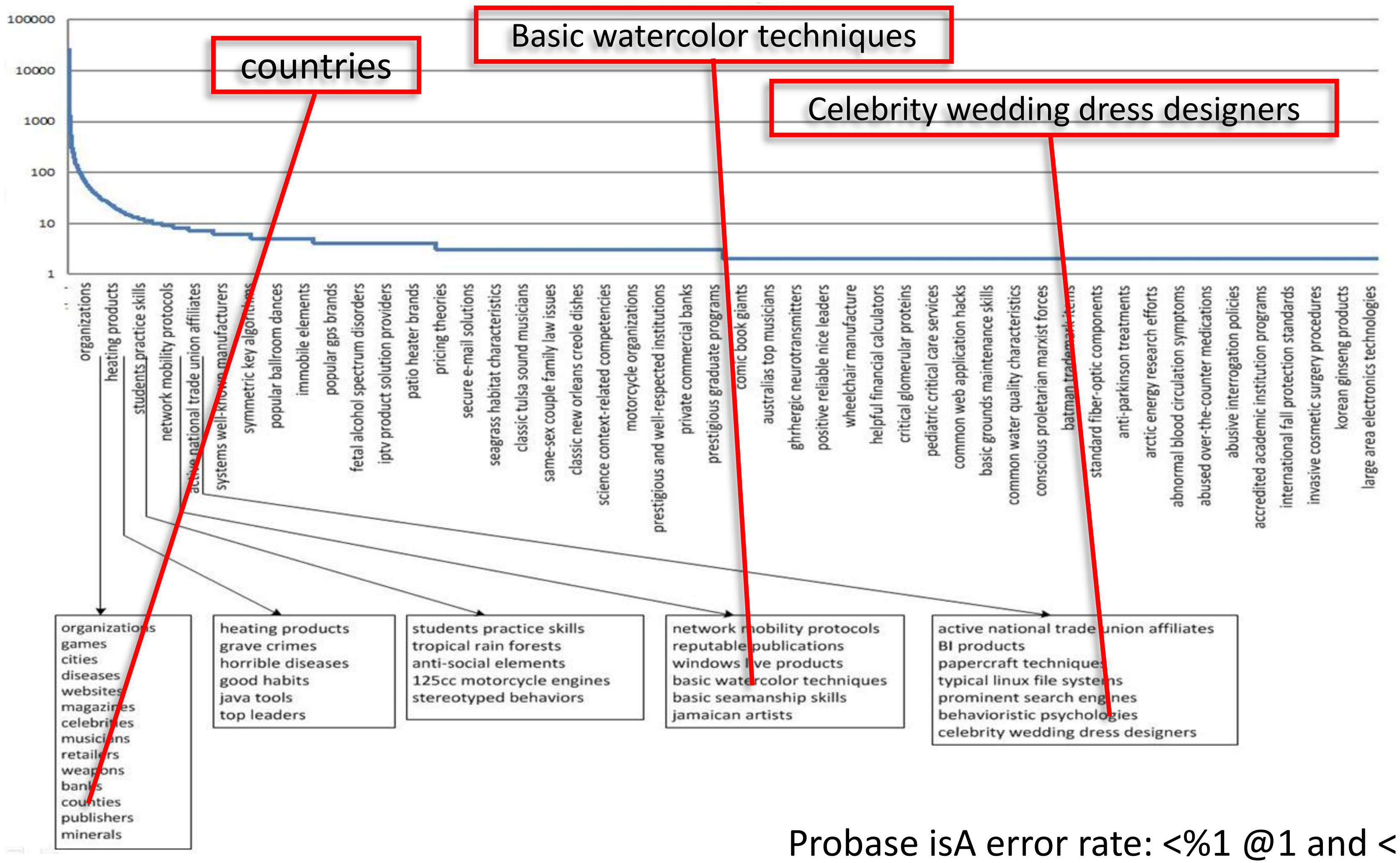
http://research.microsoft.com/probase/



Nodes: Concepts ("Spanish Artists") | Entities ("Pablo Picaso") | Attributes ("Birthday") | Verbs/Adjectives ("Eat", "Sweet")

Edges: isA (concept, entities) | isPropertyOf (attributes) | Co-occurrence (isCEOof, LocatedIn, etc)

Probase Concepts (2.7 million+)



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

Research Roadmap

Application layer

NER

Text Annotation

Ads semantic match

Query Recommendation

Text/table Understanding

SIGKDD’12
WSDM’13
CIKM’14

Conceptualization layer

Term Similarity

Short Text Similarity

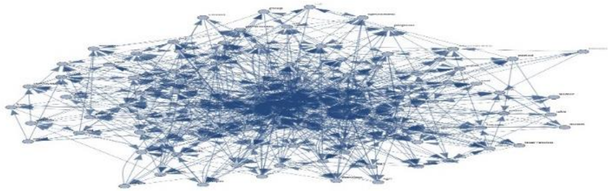
Head-Modifier Detection for Short Texts

Focus on **sense annotation** and **disambiguation**

Conceptualization

IJCAI’11, 13, 15
CIKM’13, 15
ICDE’14, 15
(**Best paper**)

Semantic network layer



The Semantic Network: Probase

VLDB’11
SIGMOD’12, ICDE’13
EDBT’14, TACL’14, TKDE’15

What is short text understanding?

Add Common Sense to Computing

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

China **Brazil** **India**

emerging market

country

The engineer is eating an apple

IT company

Conceptualization On Knowledge Engineering (COKE)

- **Conceptualization**: An **explicit** representation for the **short text**

P(concept | short text)

a domain millions of concepts
used in day to day communication

search query, anchor text
twitter, ads keywords, …

- **Short text** is *sparse, noisy, and ambiguous*
- **Explicit** means
  - Conceptualization results can be *easily understood* by human beings
  - Conceptualization model can be *easily customized* for different scenarios

Conceptualization On Knowledge Engineering (COKE)

- **Conceptualization**: An **explicit** representation for the **short text**

ShortText: pear apple

Show Parameters

Elapsed Time = 00:00:00.0140014

pear [25/fruit]		apple [25/fruit]	
25/fruit	**0.5724769**	**25/fruit**	**0.5718007**
fruit	0.0562538	fruit	0.1523192
fresh fruit	0.02918897	fresh fruit	0.05945546
tree fruit	0.01260049	tree fruit	0.01657355
dried fruit	0.01165293	dried fruit	0.01593271
seasonal fruit	0.01160144	seasonal fruit	0.01553914
juice	0.01062348	juice	0.01546597
hard fruit	0.01011614	fruit juice	0.01309836
climacteric fruit	0.009254614	hard fruit	0.01297377
fruit juice	0.009048668	climacteric fruit	0.01062575
sweet fruit	0.008924312	sweet fruit	0.01033749
9405/food	**0.1058999**	**9405/food**	**0.1241537**
food	0.03129783	food	0.06844553
high fiber food	0.008663075	high fiber food	0.01028461
ingredient	0.007349567	ingredient	0.009757149
high-fiber food	0.004699597	fresh food	0.004135756
fresh food	0.004038723	hard food	0.004045118
hard food	0.003555713	high-fiber food	0.003733333
fib[illegible]	[illegible]	[illegible]	[illegible]375
fib[illegible]	[illegible]	[illegible]	[illegible]433
[illegible]	[illegible]	[illegible]	[illegible]444
low-g[illegible]	[illegible]	[illegible]	[illegible]483

"pear apple"

ShortText: ipad apple Conceptualize

Show Parameters

Elapsed Time = 00:00:00.6670667

ipad [15/mobile device/device]		apple [1/technology company/company]	
15/mobile device/device	**0.8072805**	**1/technology company/company**	**0.9623328**
mobile device	0.01889746	technology company	0.005182603
apple device	0.01723156	computer manufacturer	0.005060604
tablet device	0.01718674	tech company	0.004826283
ios device	0.01706666	innovative company	0.00475833
portable device	0.01549841	computer company	0.004576935
gadget	0.0121459	tech giant	0.004452952
handheld device	0.0105837	technology giant	0.004435823
digital device	0.01037961	successful company	0.00422191
multimedia device	0.009883645	tech stock	0.004145819
wireless device	0.009499655	software company	0.004118673
3/apple product/product	**0.1443738**	**1053/laptop brand/top brand name/brand**	**0.02506031**
apple product	0.01561871	laptop brand	0.0004202249
apple's product	0.005314387	iconic brand	0.0004185871
electronic product	0.004432585	great brand	0.0004146745
hot new apple product	0.004225718	global brand	0.0004121742
apple's high technology product	0.004222373	big brand	0.0004073361
popular apple product	0.004221935	strong brand	0.0003894416
iconic product	0.004196884	popular brand	0.0003739756
revolutionary product	[illegible]	[illegible]	0.0003603094
digital product	[illegible]	[illegible]	0.0003555858
popular product	[illegible]	[illegible]	0.0003552496

"ipad apple"

Recap: Conceptualization

Production Impacts (Shippings):
- Ads relevance (2012)
- MSN Query Recommendation (2012)
- Bing Image Search (2013)
- Table understanding in Power Query (2013)
- Definition Answer in EQnA (2014, 2015**)**

Conceptualization 1.0 [IJCAI'11, CIKM'13]: *mapping terms to concept space based on Bayesian Inference*

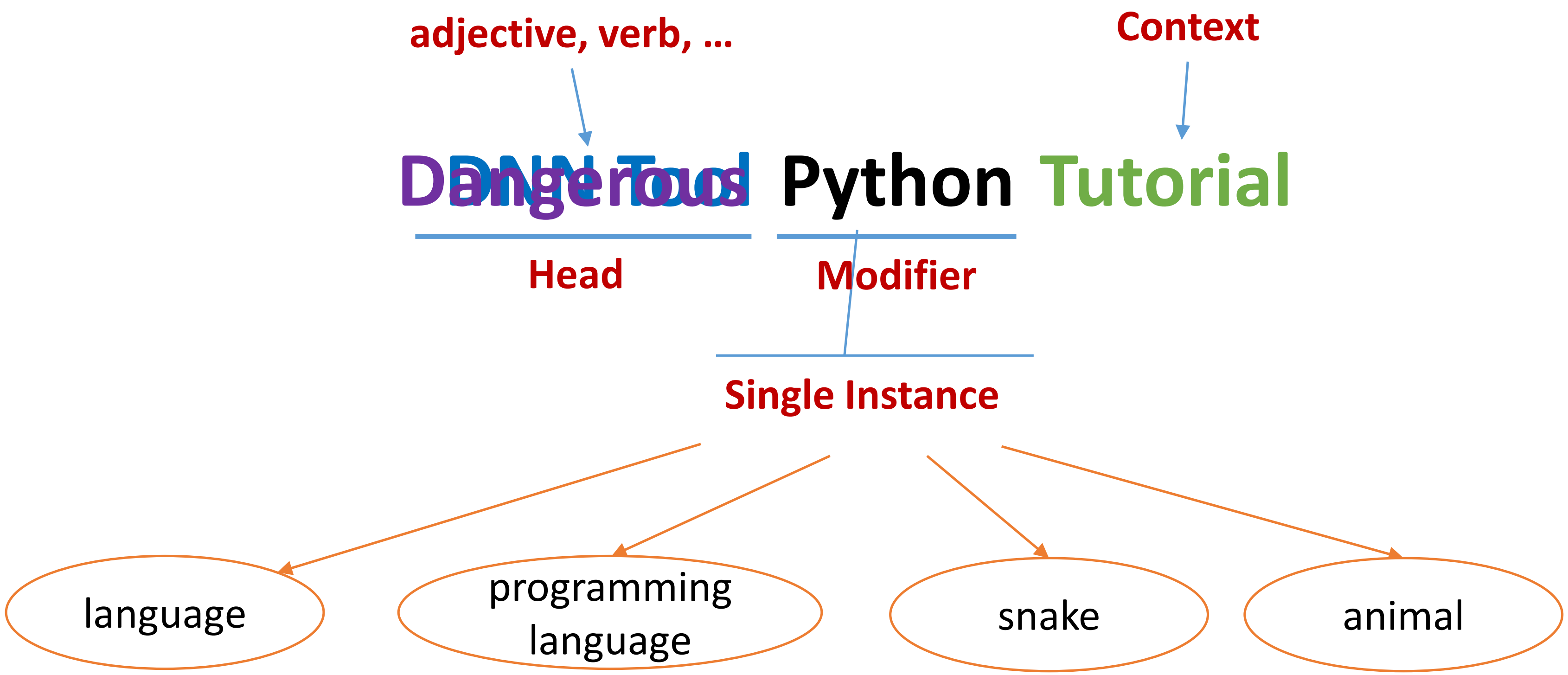
Conceptualization 2.0 [ICDE'14, CIKM'14, **ICDE'15(Best Paper)**]: *incorporating co-occurrence network*

Conceptualization 2.5 [IJCAI'15]: *leveraging verbs, adjective, attribute, etc.*

Conceptualization 3.0 [CIKM'15]: learning-based conceptualization/ leverage embedding

What we resolved?

- Short Text Understanding



Short Text Understanding

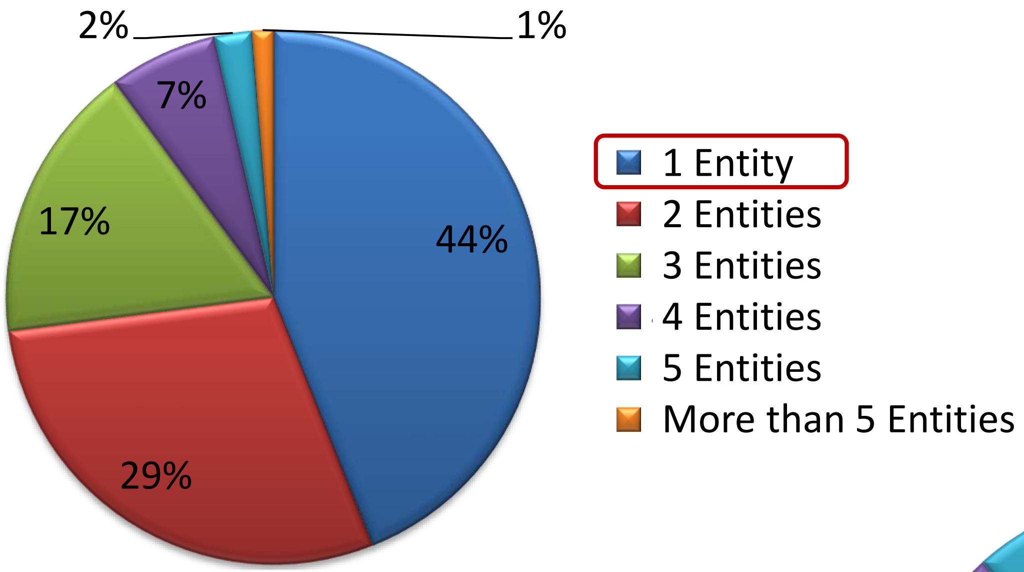
- If the short text is a **single instance**…
  - *SIGMOD 2012, CIKM 2015*
- If the short text has **context** for the instance…
  - *IJCAI 2011/2013, ICDE 2015*
- If the short text contains **verb, adjective**…
  - *IJCAI 2015*
- If the short text contains **multiple instance**…
  - *ICDE 2014*
- **Applications**
  - *WSDM 2013, CIKM 2013/2014*

If the short text is a single instance…

“Python”

- Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Zhu, Probase: A Probabilistic Taxonomy for Text Understanding, in *ACM International Conference on Management of Data* (***SIGMOD***), May 2012.
- Zhongyuan Wang, Haixun Wang, Ji-Rong Wen, and Yanghua Xiao, An Inference Approach to Basic Level of Categorization, in *ACM International Conference on Information and Knowledge Management* (***CIKM***), October 2015.

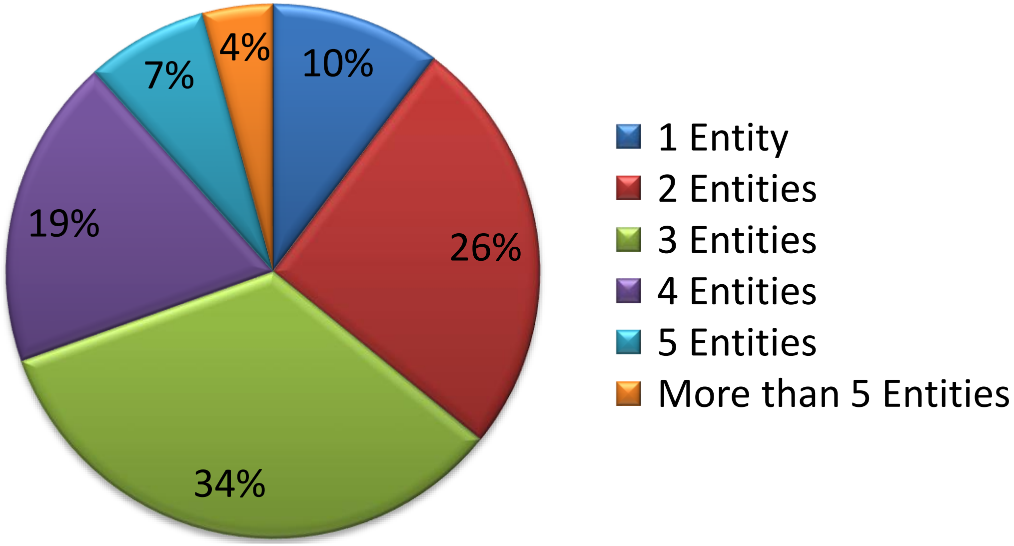
Statistics of Search Queries



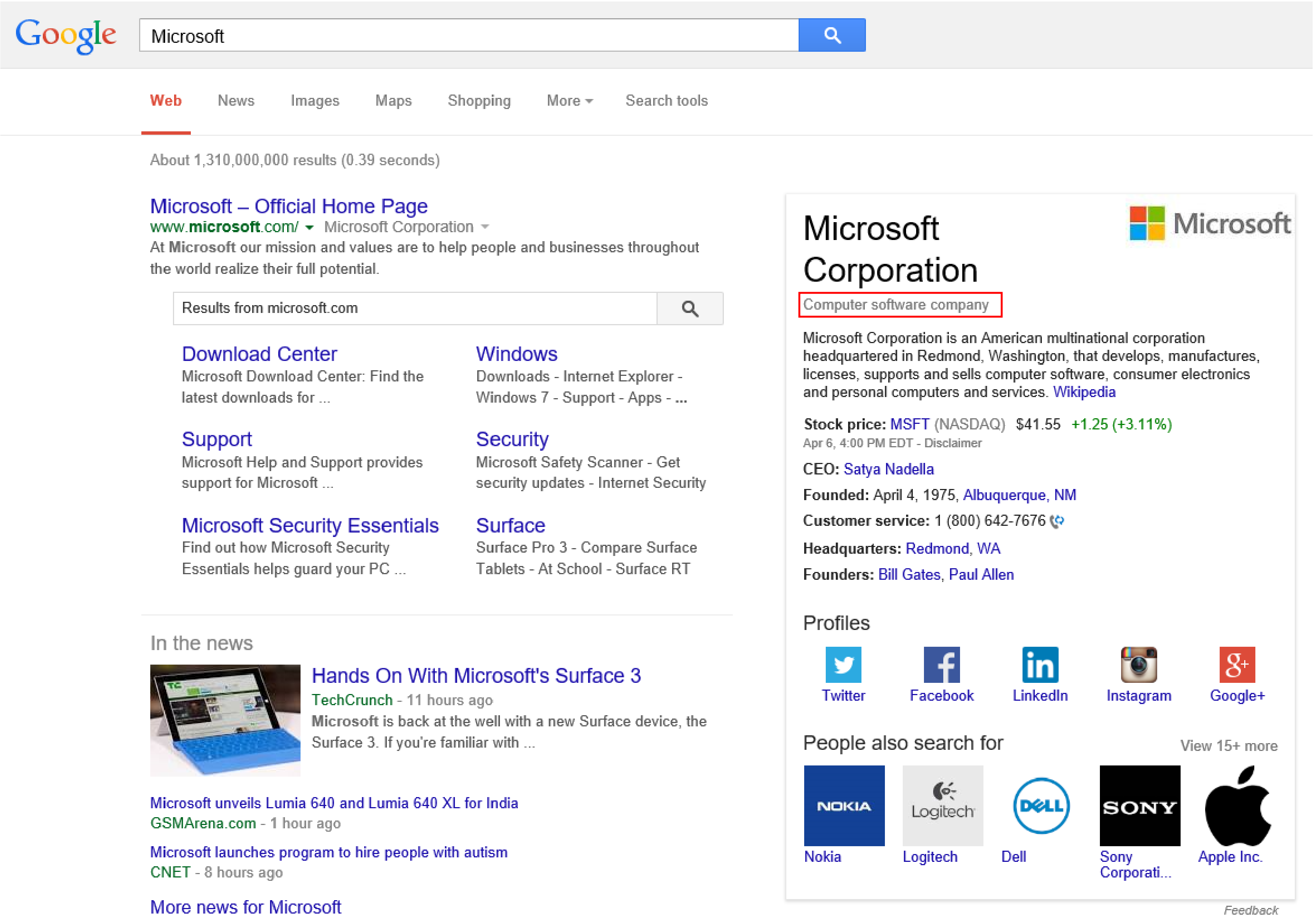
(a) By traffic

angry birds windows phone 8

Entity 1 *Entity 2*

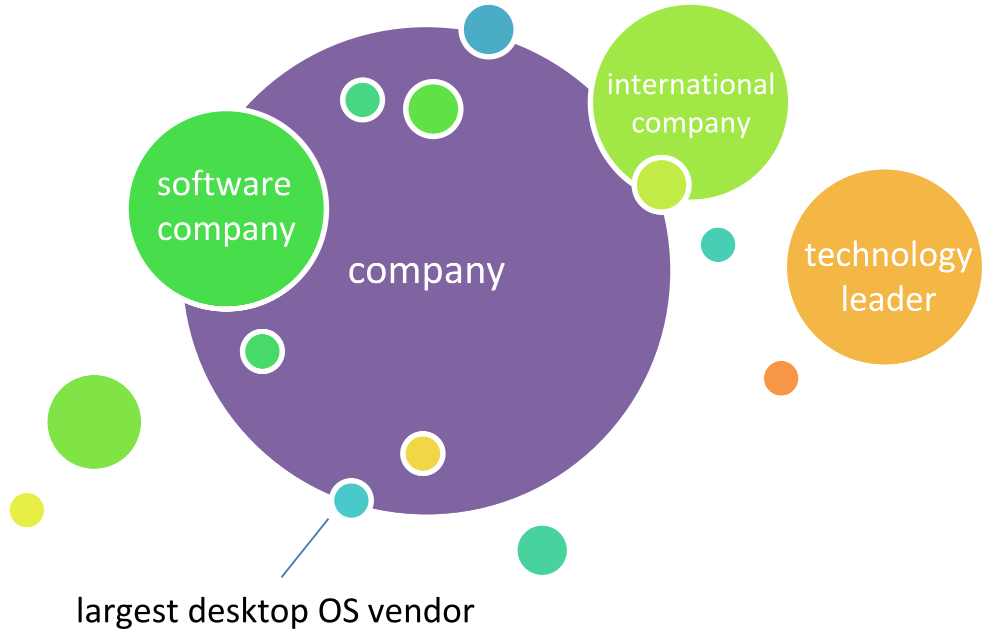


(b) By # of distinct queries



Knowledge Panel

A Concept View of "Microsoft"



Basic-level Conceptualization (BLC)

Category Level	Informative?	Distinctive?
Superordinate	No	Yes
Basic-level	Yes	Yes
Subordinate	Yes	No

Basic-level conceptualization

software company

company ... ... **largest desktop OS vendor**

KFC

BMW

Microsoft

Using *Rep*(*e*, *c*) for BLC

- **Our measure** $Rep(e,c) = P(c|e) * P(e|c)$ means:

Given *e*, the *c* should be its typical concept (shortest distance)

Given *c*, the *e* should be its typical entity (shortest distance)

A process of finding *concept nodes* having shortest expected distance with *e*

- **(With PMI)** If we take the logarithm of our scoring function, we get:

$$\log Rep(e,c) = \log P(c|e) * P(e|c) = \log \frac{P(e,c)}{P(e)} * \frac{P(e,c)}{P(c)} = \log \frac{P(e,c)^2}{P(e)P(c)} = PMI(e,c) + \log P(e,c) = PMI^2$$

- **(With Commute Time)** The commute time between an instance *e* and a concept *c* is:

$$Time(e,c) = \sum_{k=1}^{\infty} (2k) * P_k(e,c) = \sum_{k=1}^{T} (2k) * P_k(e,c) + \sum_{k=T+1}^{\infty} (2k) * P_k(e,c)$$

$$\geq \sum_{k=1}^{T} (2k) * P_k(e,c) + 2(T+1) * \left(1 - \sum_{k=1}^{T} P_k(e,c)\right) = 4 - 2 * Rep(e,c)$$

Precision@K & NDCG@K

- Metrics
  - $Precision@K = \frac{\sum_{i=1}^{K} rel_i}{K}$ (for **correctness** of concepts)
  - $nDCG_K = \frac{rel_1 + \sum_{i=2}^{K} \frac{rel_i}{\log i}}{ideal\_rel_1 + \sum_{i=2}^{K} \frac{ideal\_rel_i}{\log i}}$ (for **ranking** of concepts)
- Results

Precision@K	1	2	3	5	10	15	20
MI(e)	0.769	0.692	0.705	0.685	0.719	0.705	0.690
PMI^3(e)	**0.885**	0.769	0.756	0.800	0.754	**0.733**	**0.721**
NPMI(e)	0.692	0.692	0.667	0.638	0.627	0.610	0.610
Typicality P(c|e)	0.462	0.577	0.603	0.577	0.569	0.564	0.556
Typicality P(e|c)	0.500	0.462	0.526	0.523	0.523	0.510	0.521
Rep(e)	0.846	**0.865**	**0.872**	**0.862**	**0.758**	0.731	0.719

NDCG@K	1	2	3	5	10	15	20
MI(e)	0.516	0.531	0.519	0.531	0.562	0.574	0.594
PMI^3(e)	0.725	0.664	0.652	0.660	0.628	0.631	0.646
NPMI(e)	0.599	0.597	0.579	0.554	0.540	0.539	0.549
Typicality P(c|e)	0.297	0.380	0.409	0.422	0.438	0.446	0.461
Typicality P(e|c)	0.401	0.386	0.396	0.398	0.401	0.410	0.428
Rep(e)	**0.758**	**0.771**	**0.745**	**0.723**	**0.656**	**0.647**	**0.661**

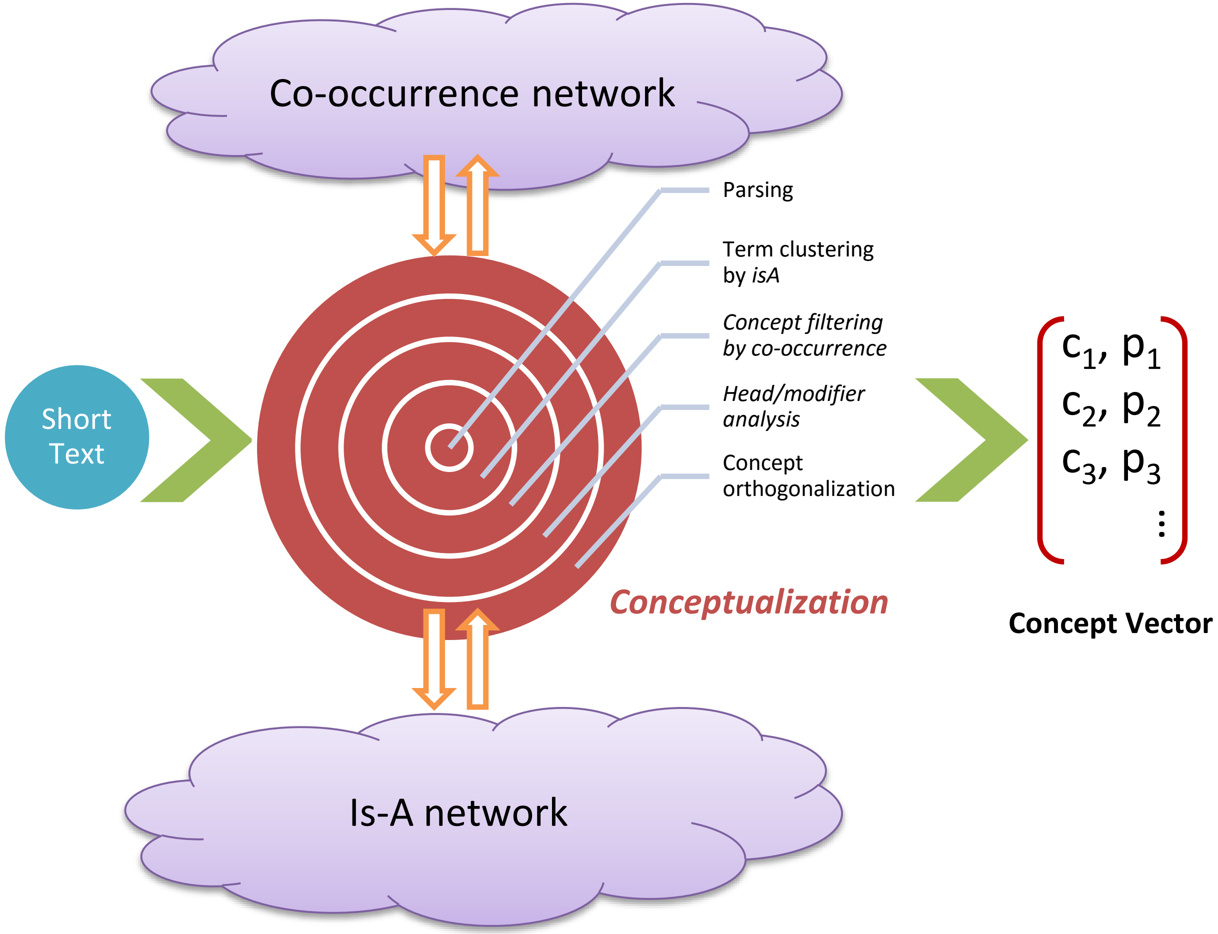
- Overall, our measure *Rep* performs well in both *Precision* and *NDCG*.
- Most important, it's well interpreted in theory

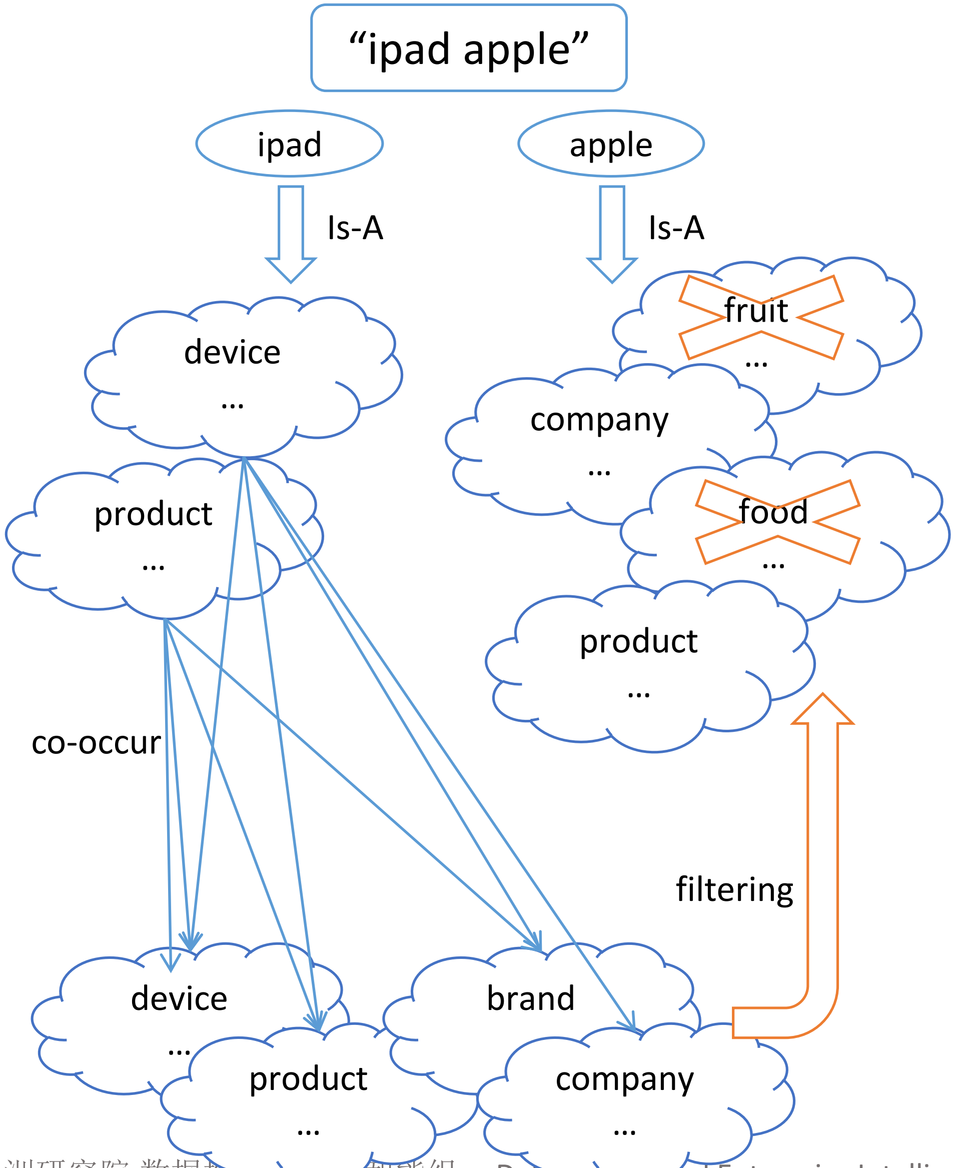
If the short text has context for the instance…

“Python Tutorial”

- Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou, Short Text Understanding Through Lexical-Semantic Analysis, in *International Conference on Data Engineering (**ICDE**)*, April 2015. (**Best Paper Award**)
- Dongwoo Kim, Haixun Wang, and Alice Oh, Context-Dependent Conceptualization, in ***IJCAI***, 2013.
- Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen, Short Text Conceptualization using a Probabilistic Knowledgebase, in ***IJCAI***, 2011.

Conceptualization Framework
(ICDE 2015 Best Paper)



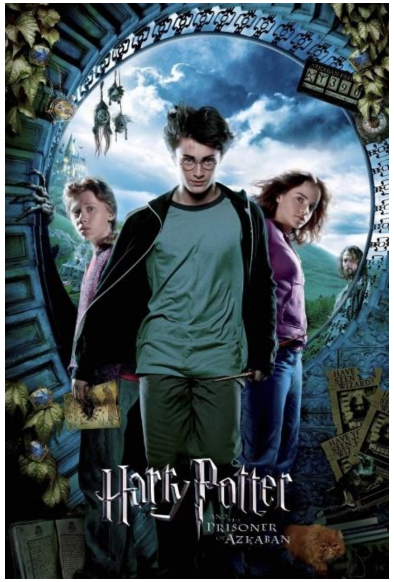


If the short text contains verb, adjective…

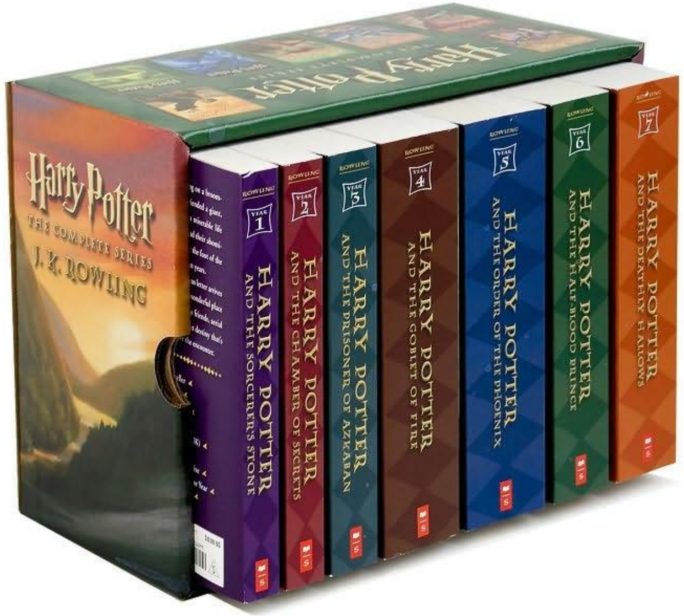
"Dangerous Python"

- Zhongyuan Wang, Kejun Zhao, Haixun Wang, Xiaofeng Meng, and Ji-Rong Wen, Query Understanding through Knowledge-Based Conceptualization, in *IJCAI*, 2015.

- Watch Harry Potter
- Read Harry Potter



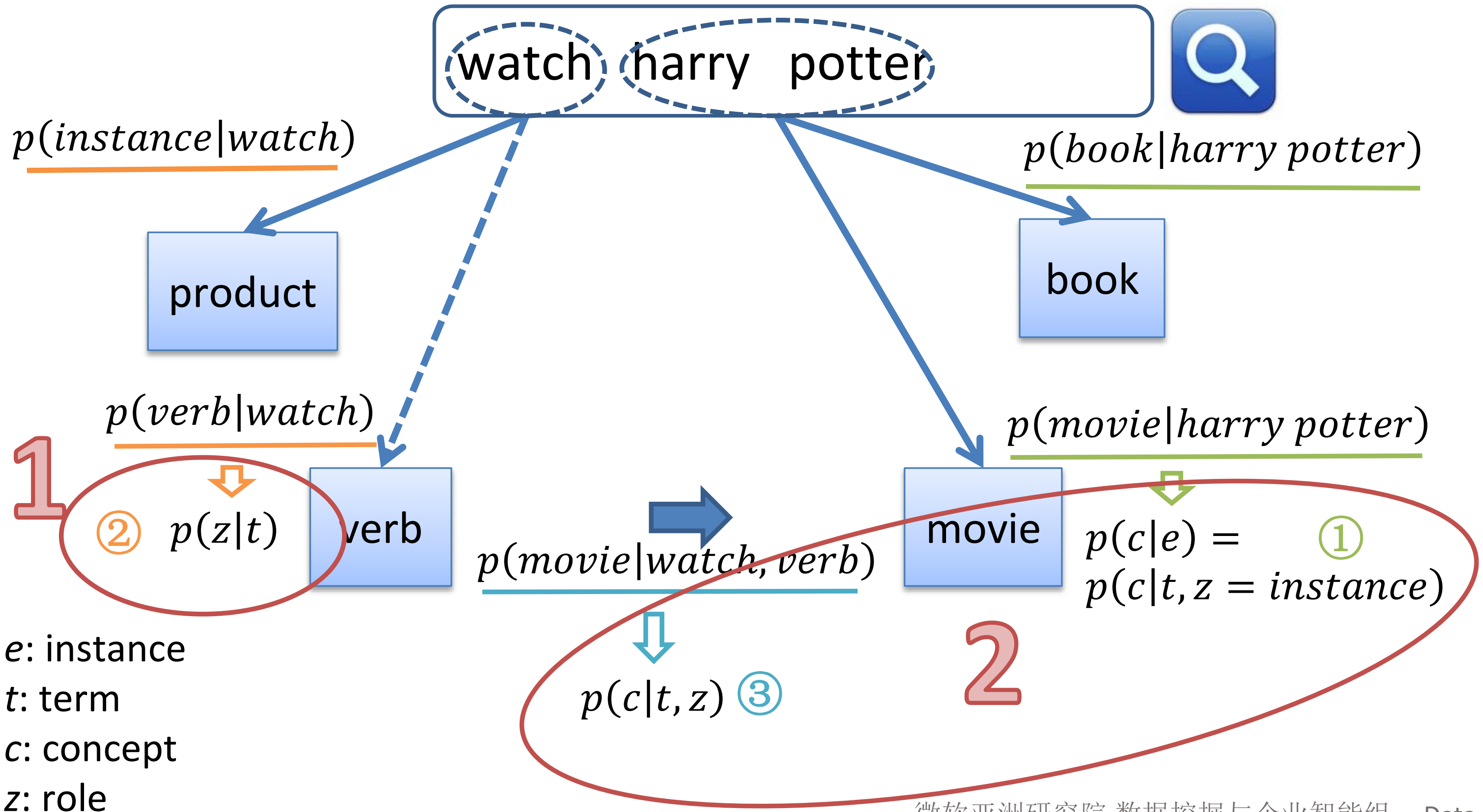
Movie



Book

Mining Lexical Relationships

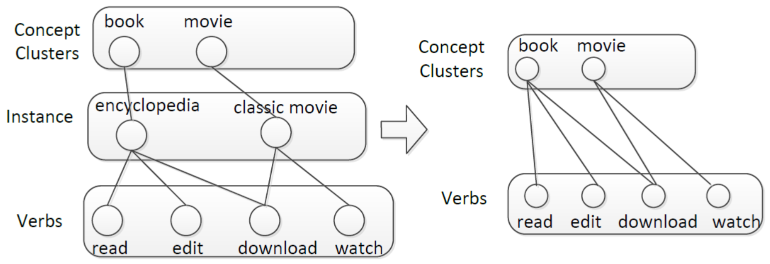
- Lexical knowledge represented by the probabilities



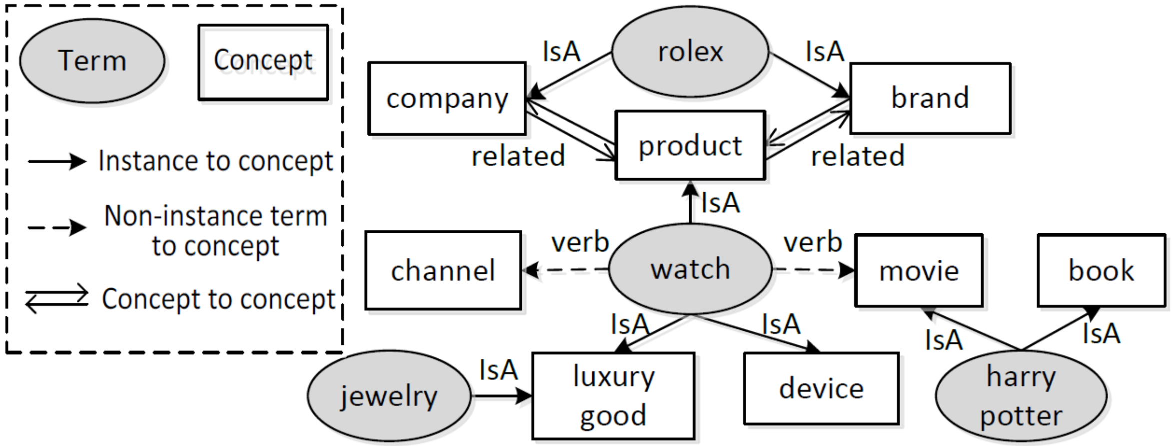
e: instance
t: term
c: concept
z: role

Deriving Probabilities

- **Deriving $\boldsymbol{p(z|t)}$**: $p(z|t) = \frac{n(t,z)}{n(t)}$
- **Deriving $\boldsymbol{P(c|t,z)}$**
  - Case 1: z=instance $P(c|t, z = insctance) = p(c|e)$
  - Case 2: z=attribute $P(c|t, z = attribute) = p(c|a)$
  - Case 3: z=verb
    $$P(c|t, z = verb) = \sum_{e \in c} p(e, c|t, z = verb) = \sum_{e \in c} p(c|e) \times p(e|t, z = verb)$$
  - Case 4: z=adjective
    $$P(c|t, z = adjective) = \sum_{e \in c} p(c|e) \times p(e|t, z = adjective)$$



Constructing an offline semantic network



$$P(c|t) = \sum_z p(c|t,z) \times p(z|t)$$

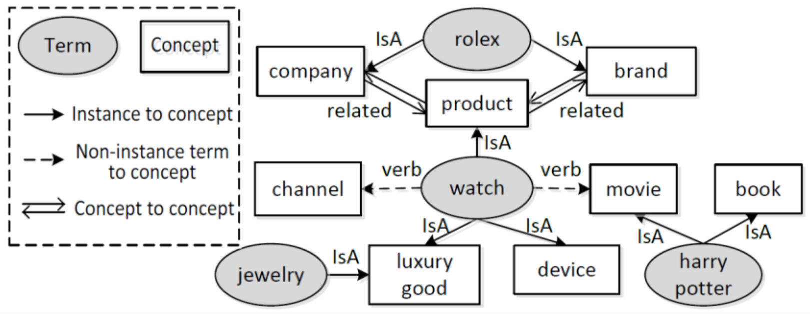
$$P(c_2|c_1) = \frac{\sum_{e_i \in c_1, e_j \in c_2} n(e_i, e_j)}{\sum_c \sum_{e_i \in c_1, e_j \in c} n(e_i, e_j)}$$

Understanding Queries

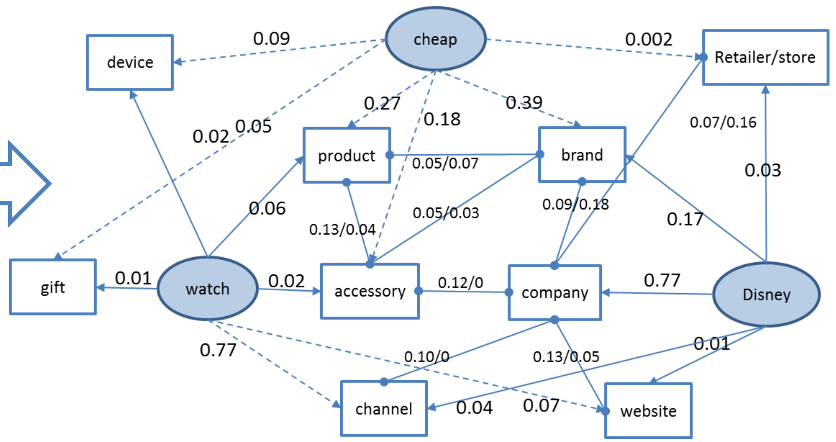
- **Goal**: to rank the concepts and find:

$$\arg\max_{c} p(c|t,q)$$

Query → Segmentation



The offline semantic network



Random walk with restart [Sun et al., 2005] on the online subgraph

If the short text contains multiple instance…

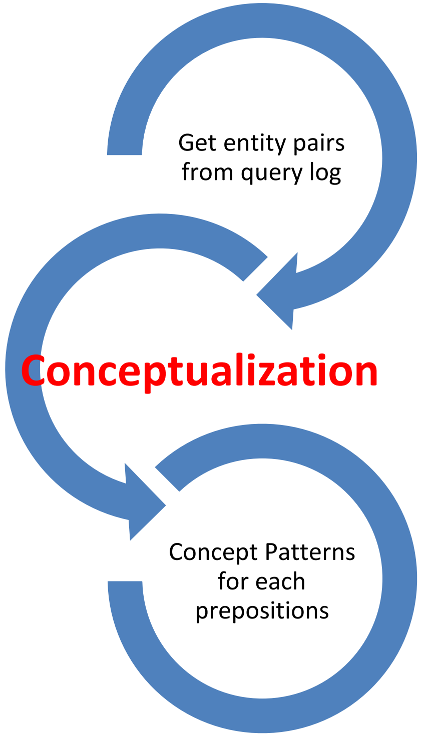
"DNN Tool Python"

Head **Modifier**

- Zhongyuan Wang, Haixun Wang, and Zhirui Hu, Head, Modifier, and Constraint Detection in Short Texts, in *International Conference on Data Engineering (ICDE)*, 2014.

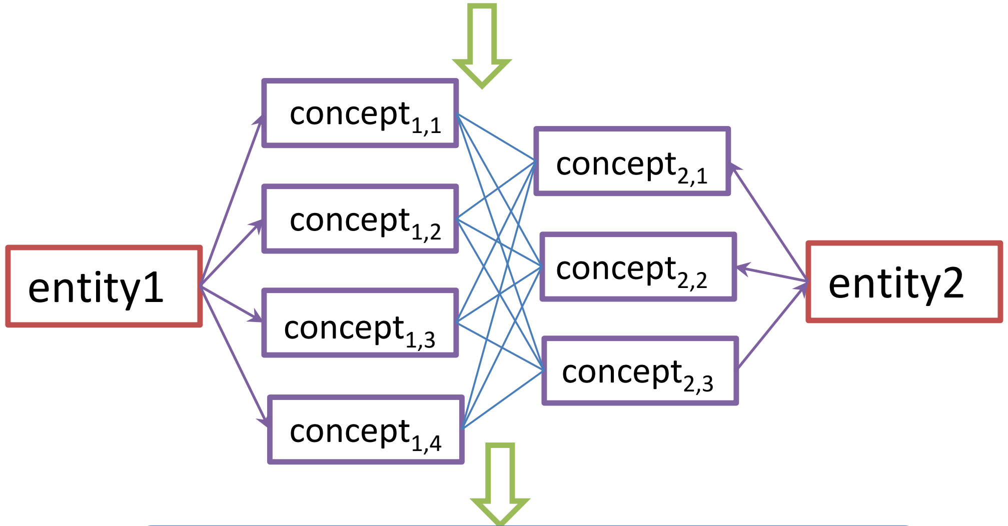
Mining Concept Patterns

Building concept pattern dictionary:



Query Logs	Extract Patterns
cover for iphone 5s, battery for sony a7r wicked on broadway	A for B, A of B, A with B, A in B, A on B, A at B …

entity 1/head   entity 2/constraint



Concept Pattern Dictionary

(concept$_{1,1}$, concept$_{2,1}$), (concept$_{1,1}$, concept$_{2,2}$)
(concept$_{1,1}$, concept$_{2,3}$)…

Why Concepts Can't Be Too General

- It may cause too many concept pattern conflicts: can't distinguish head and modifier for general concept pairs

	Head	Modifier
Derived Concept Pattern	device	company
Supporting Entity Pairs	iphone 4	verizon
	modem	comcast
	wireless router	comcast
	iphone 4	tmobile

	Head	Modifier
Derived Concept Pattern	company	device
Supporting Entity Pairs	amazon books	kindle
	netflix	touchpad
	skype	windows phone
	netflix	ps3

Why Concepts Can't Be Too Specific

- It may generate concepts with less representation

…	…
device	largest desktop OS vendor
device	largest software development company
device	largest global corporation
device	latest windows and office provider
…	…

- Concept level may regress to entity level
  - Large storage space: up to (million * million) patterns

We should use Basic-level Conceptualization (BLC)

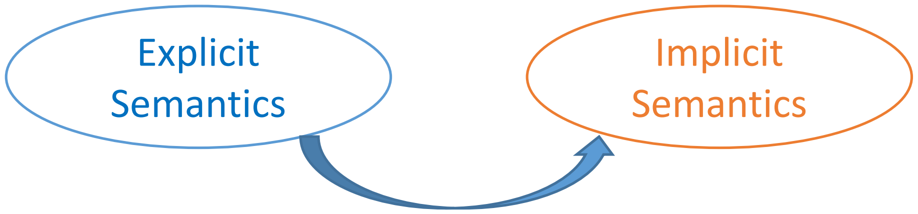
Top Concept Patterns

Cluster size	Sum of Cluster Score	head;modifier;score
615	21146.91	breed;state;3572.98460224501
296	7752.357	game;platform;627.403476771856
153	3466.804	accessory;vehicle;533.93705094809
70	1182.59	browser;platform;132.612807637391
22	1010.993	requirement;school;271.407526294823
34	948.9159	drug;disease;154.602405333541
42	899.2995	cosmetic;skin condition;81.4659415003929
16	742.1599	job;city;279.03732555528
32	710.403	accessory;phone;246.513830851194
18	669.2376	software;platform;210.126322725878
20	644.4603	test;disease;239.774028397537
27	599.4205	clothes;breed;98.773996282851
19	591.3545	penalty;crime;200.544192793488
25	584.8804	tax;state;240.081818612579
16	546.5424	sauce;meat;183.592863621553
18	480.9389	credit card;country;142.919087972152
14	473.0792	food;holiday;145.54140330924
11	453.6199	mod;game;257.163856882439
29	435.0954	garment;sport;47.1533326845442
23	399.4886	career information;professional;73.2726483731257
15	386.065	song;instrument;128.189481818135
18	378.213	bait;fish;78.0426514113169
22	372.2948	study guide;book;50.8339765053921
19	340.8953	plugins;browser;55.0326072627126
14	330.5753	recipe;meat;88.2779863422951
18	321.4226	currency;country;110.825444188352
13	318.0272	lens;camera;186.081673263957
9	316.973	decoration;holiday;130.055844126533
16	314.875	food;animal;73.38544366514

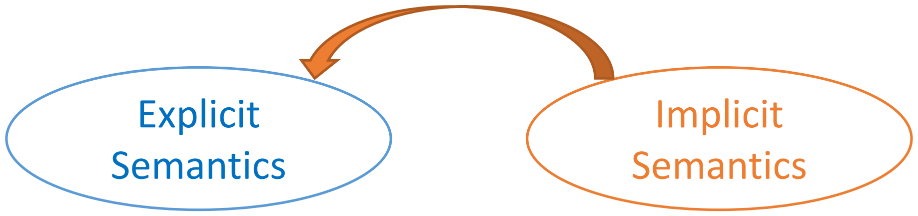
game	platform
game	device
video game	platform
game	console game pad
game	gaming platform

Detection

Game (Head)	Platform (Modifier)
angry birds	android
angry birds	ios
angry birds	windows 10
…	…



Existing Work

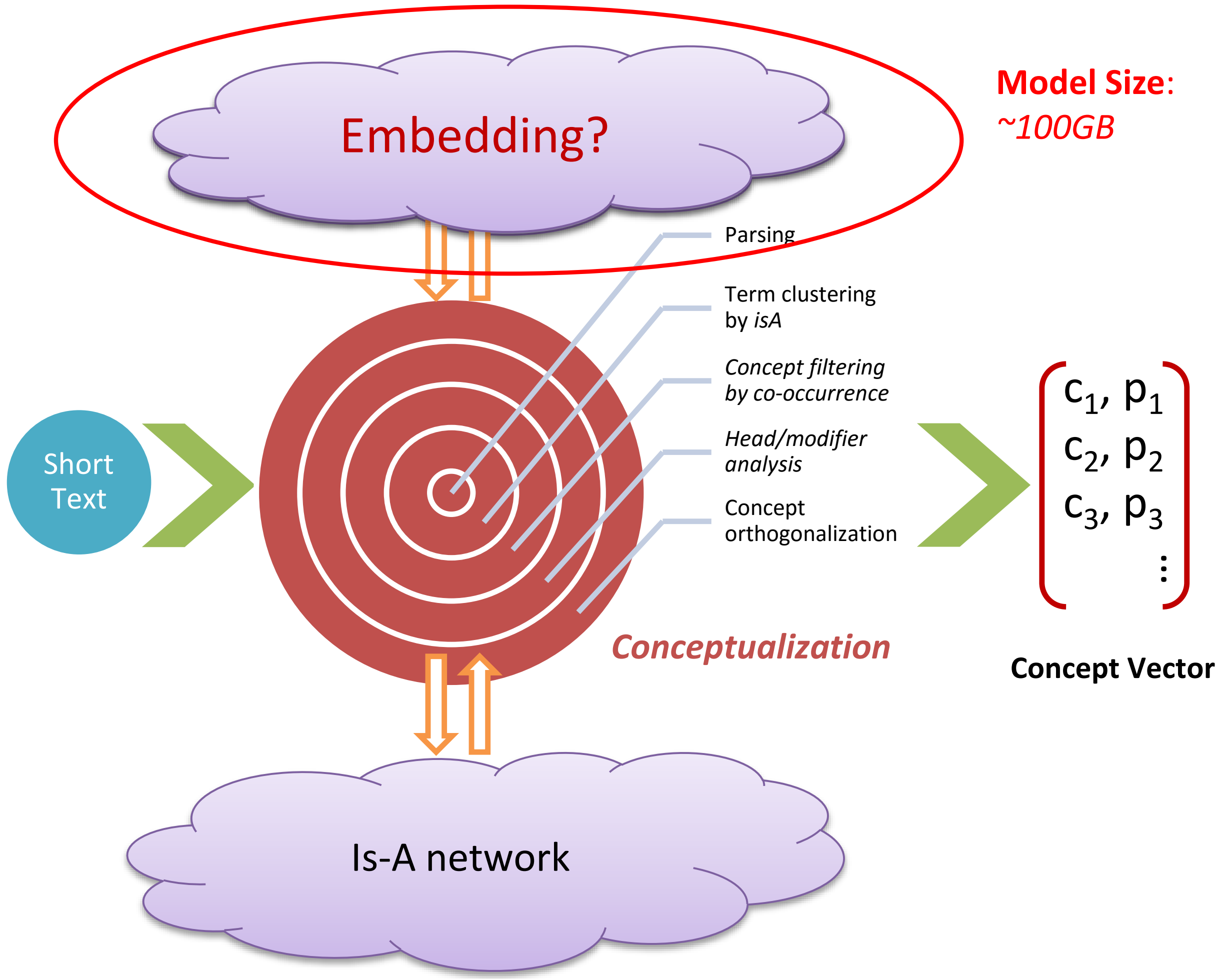


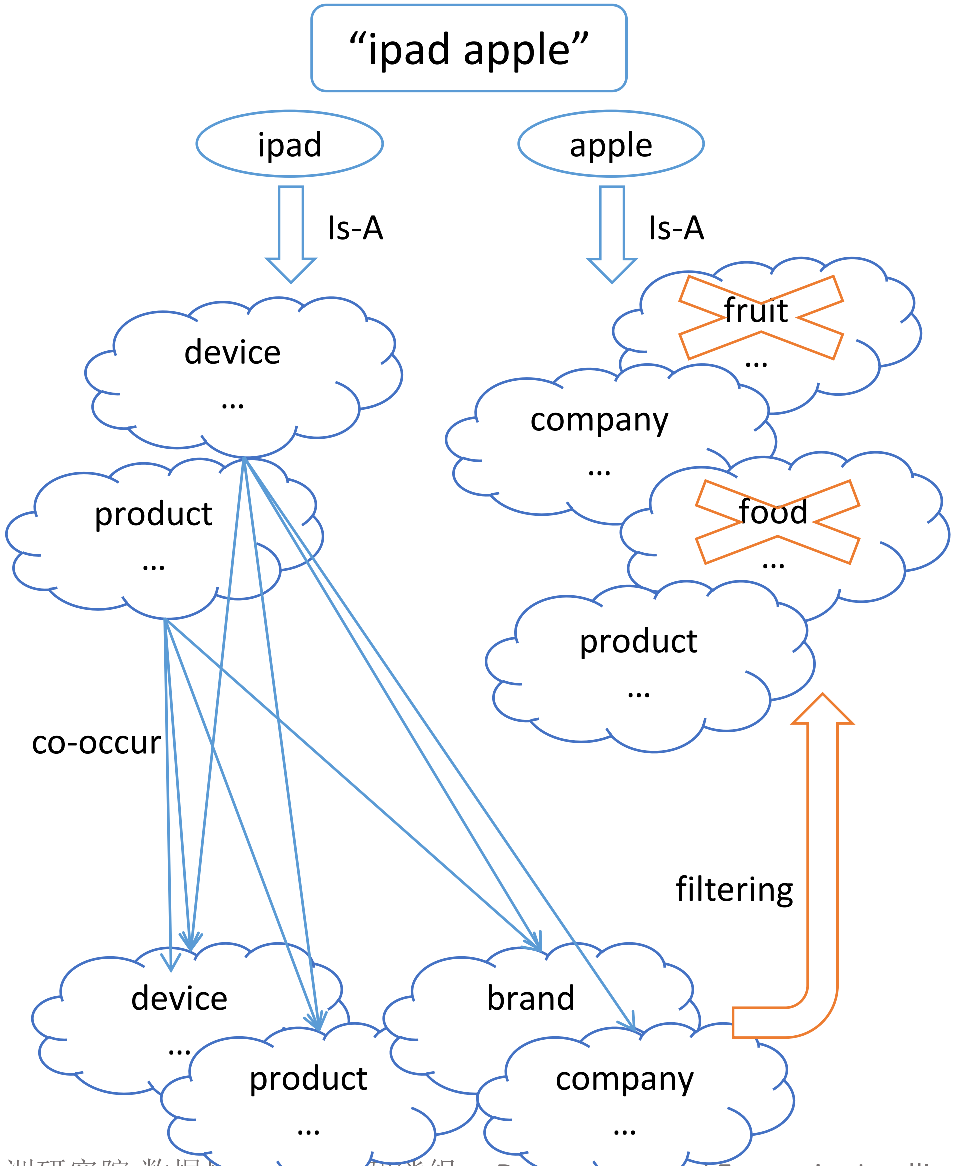
Learning-based Conceptualization

Combine Explicit and Implicit Semantics (First Attempt): **Learning-based Conceptualization**

- Contextual Text Understanding in Distributional Semantic Space (CIKM2015)

Previous Conceptualization Framework
(ICDE 2015 Best Paper)





Basic Idea of Learning-based Conceptualization

- **Target**: best representation for the short text:

Conceptualization + Embedding = 

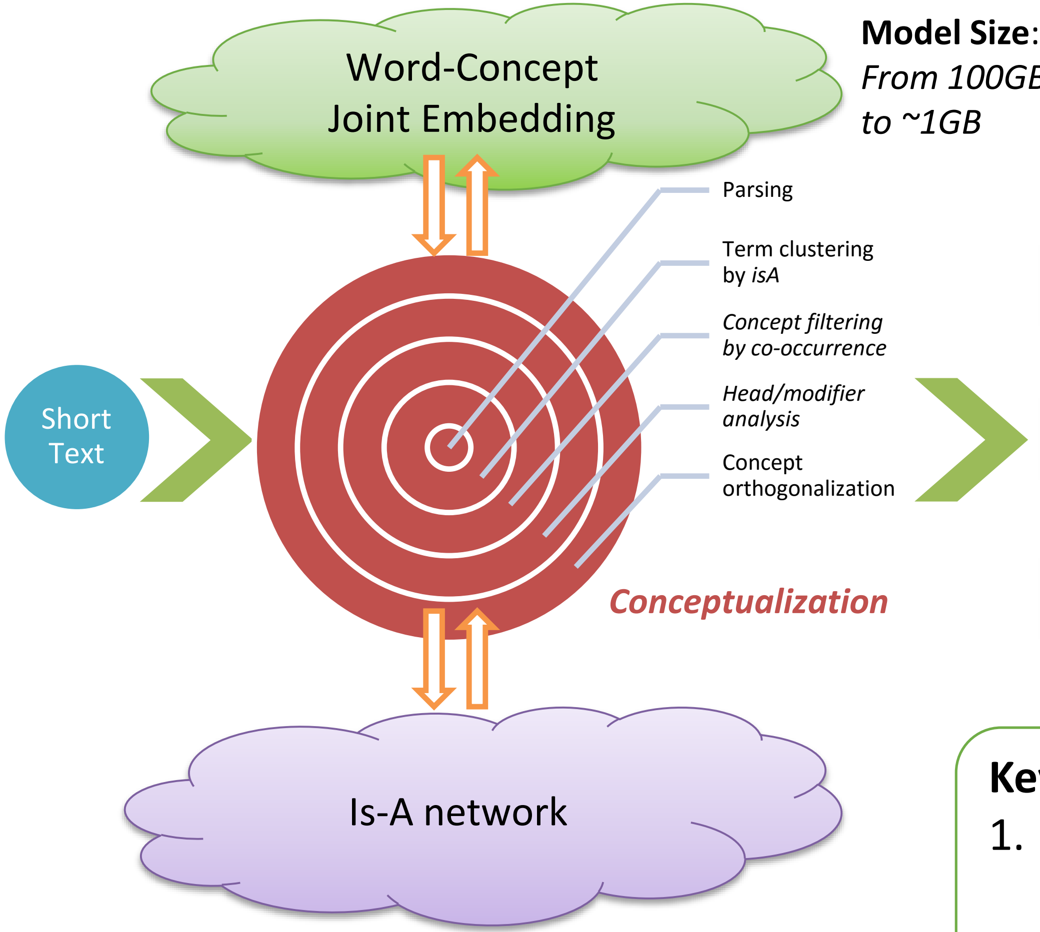
- **Learning-based Conceptualization**: given a term *t*, with its context *q*, we want to find the probability of concept *c*

$$f_{conceptualization}(c,t,q) = \underbrace{p(c|t)}_{\text{Traditional conceptualization based on Is-A network}} * \underbrace{p(c|q)}_{\text{Contextual knowledge based on Embedding}} = p(c|t) * \frac{\cos(\vec{q},\vec{c})}{\sum_{c_i \in c(t)} \cos(\vec{q},\vec{c_i})}$$

Traditional conceptualization
based on **Is-A network**

Contextual knowledge
based on **Embedding**

Learning-based Conceptualization Framework



Model Size:
From 100GB
to ~1GB

Concept Vector : $[(c_1, p_1), (c_2, p_2), (c_3, p_3) \ldots]$	**Concepts with scores in context** **(Explicit Representation)**
Contextual Word Vector	**Word vectors in context** **(Implicit Representation)**
Contextual Text Vector	**Text vectors in context** **(Implicit Representation)**

Key Problems:

1. Joint word-concept embedding: project **words** and **concepts** to the **same semantic space**
2. Refined **concept inference process**

Joint Word-Concept Embedding

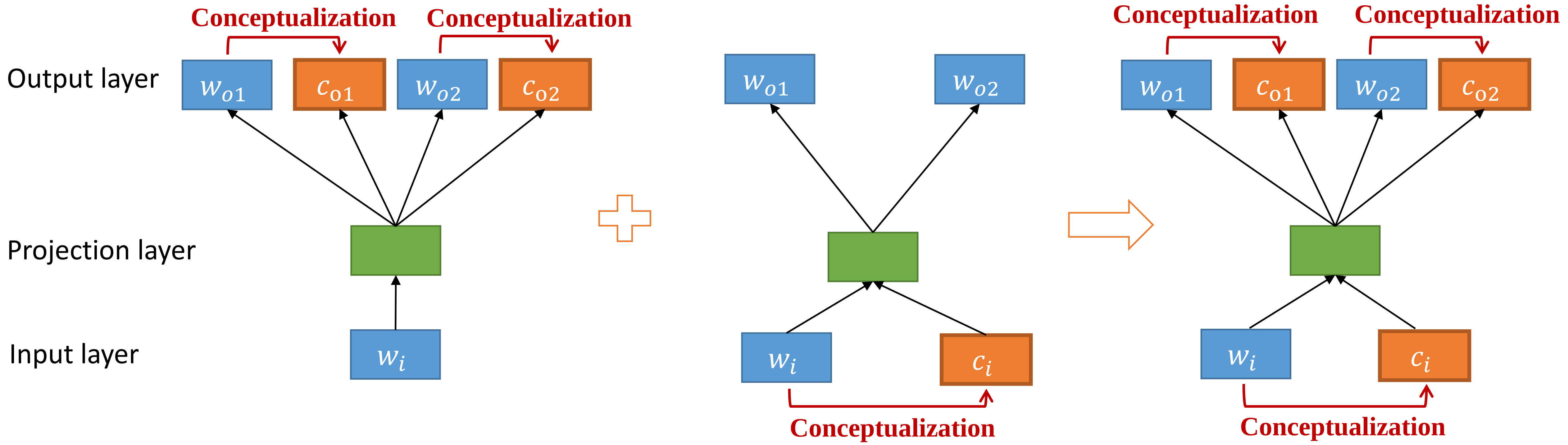
- Word embedding – Skip gram model

$$\sum_{i=1}^{N} \sum_{o=i-c}^{i+c} logP(w_o|w_i)$$

- How to incorporate concept
  - Concept as input $\sum_{i=1}^{N} \sum_{o=i-c}^{i+c} logP(w_o|c_i)$
  - Concept as output $\sum_{i=1}^{N} \sum_{o=i-c}^{i+c} logP(c_o|w_i)$
  - Concept as both input and output $\sum_{i=1}^{N} \sum_{o=i-c}^{i+c} logP(c_o|c_i)$

Approach 1: Parallel Joint-Embedding Models

- Assume conditionally independent between the word and concept



$$\sum_{i=1}^{N}\sum_{o=i-c}^{i+c} log\, P(w_o|w_i)P(c_o|w_i)$$

PM-1

$$\sum_{i=1}^{N}\sum_{o=i-c}^{i+c} log\, P(w_o|w_i)P(w_o|c_i)$$

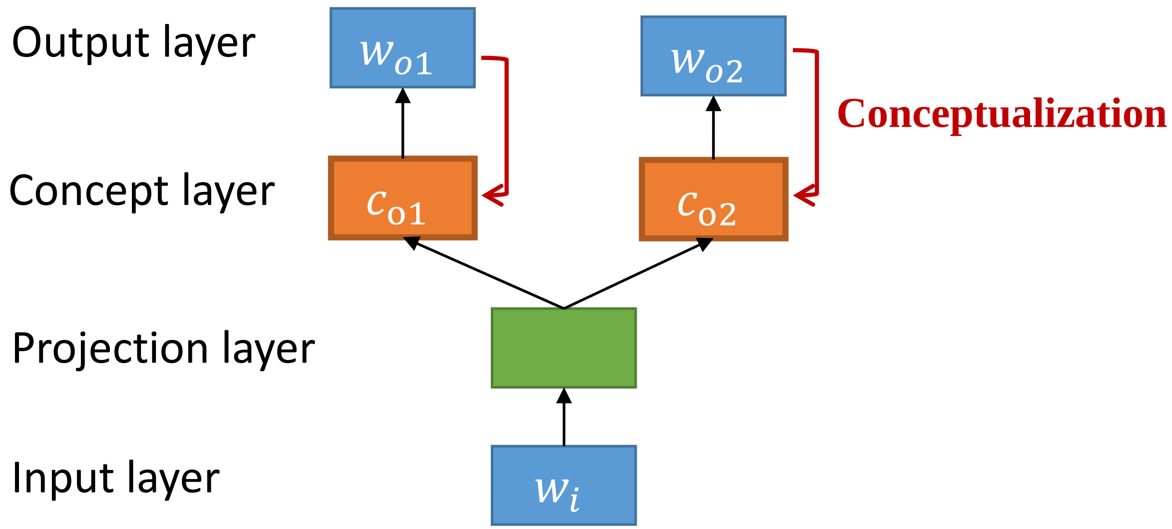
PM-2

$$\sum_{i=1}^{N}\sum_{o=i-c}^{i+c} log\, P(w_o|w_i)P(w_o|c_i)P(c_o|w_i)P(c_o|c_i)$$

PM-3

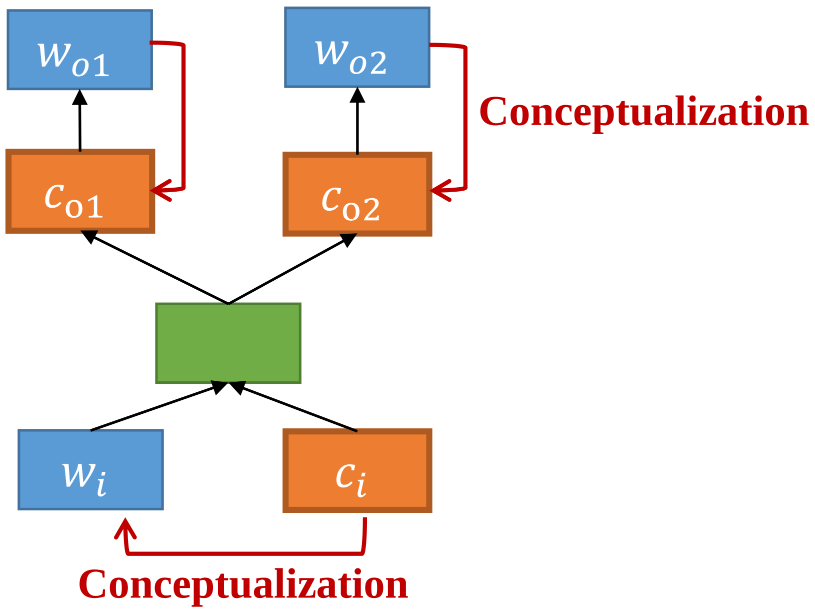
Approach 2: Generative Joint-Embedding Models

- Assume conditionally dependent between output word and output concept: A word is selected by firstly select the class it belongs to.



$$\sum_{i=1}^{N} \sum_{t=i-c}^{i+c} logP(w_o|w_i) = \sum_{i=1}^{N} \sum_{t=i-c}^{i+c} logP(c_o|w_i)P(w_o|c_o)$$

GM-1



$$\sum_{i=1}^{N} \sum_{t=i-c}^{i+c} logP(w_o|w_i, c_i) = \sum_{i=1}^{N} \sum_{t=i-c}^{i+c} logP(c_o|w_i)P(c_o|c_i)P(w_o|c_o)$$

GM-2

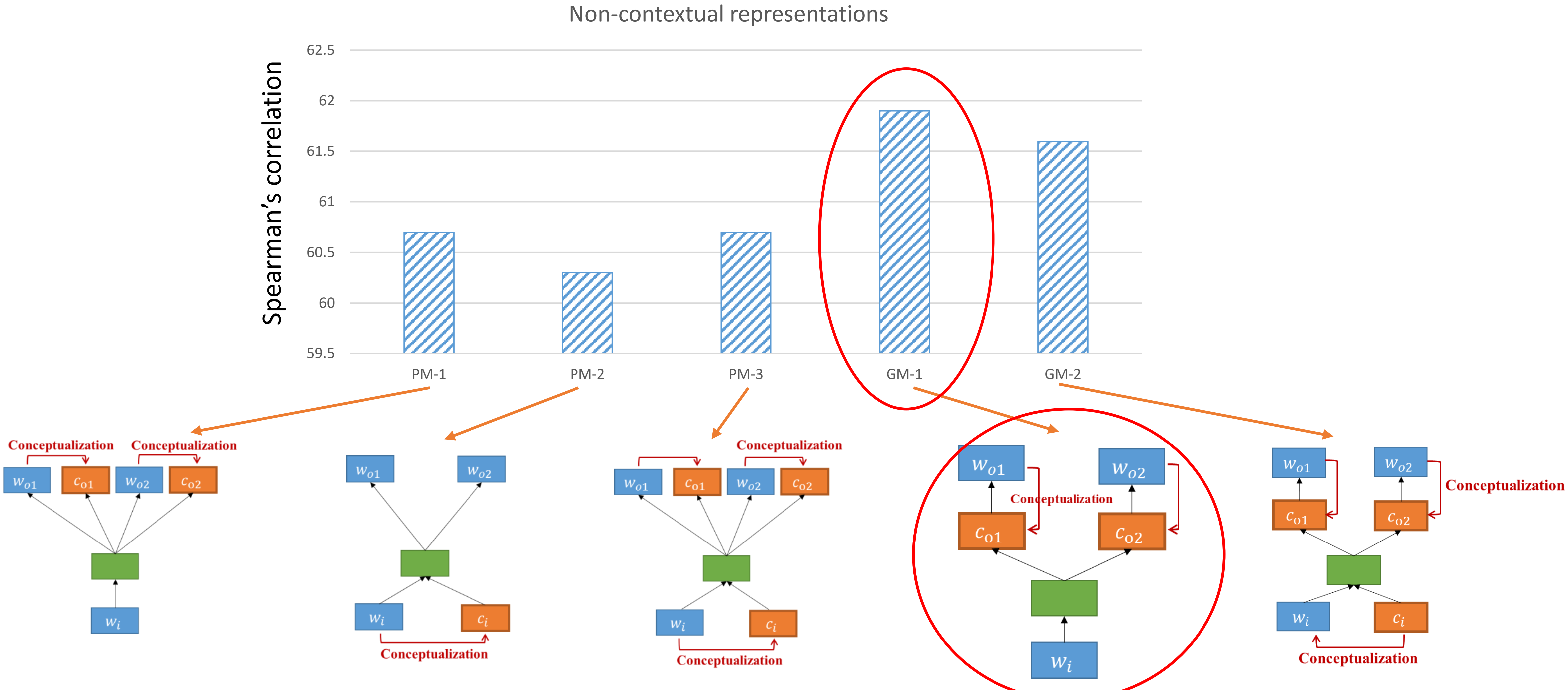
Evaluation: Word Similarity in Context

- Evaluation Setting
  - **Public dataset**: Eric Huang et al. (2012) Improving Word Representations via Global Context and Multiple Word Prototypes.
  - **Task**: Given *word1*, *word2* and their contexts, compute *similarity(word1, word2)*

- Metric: the **Spearman's correlation** between human judgement and embedding similarity score.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Word-Concept Embedding Evaluation Results



- **PM**: Parallel Joint-Embedding Model
- **GM**: Generative Joint-Embedding Model

Conceptualization with Word-Concept Embedding

- **Concept Inference**: given a term *t*, with its context *q*, we want to find the probability of concept *c*

$$f_{conceptualization}(c,t,q) = \underbrace{p(c|t)}_{\text{Traditional conceptualization based on Is-A network}} * \underbrace{p(c|q)}_{\text{Contextual knowledge based on Embedding}} = p(c|t) * \frac{\cos(\vec{q},\vec{c})}{\sum_{c_i \in c(t)} \cos(\vec{q},\vec{c_i})}$$

- **Conceptualization**: $p(c|t,q) = \dfrac{f_{conceptualization}(c,t,q)}{\sum_{c_i \in c(t)} f_{conceptualization}(c_i,t,q)}$

- **Example**: "apple, microsoft and google are world's most valuable brands."

$\underbrace{\text{microsoft and google are world's most valuable brands.}}_{\text{context}}$

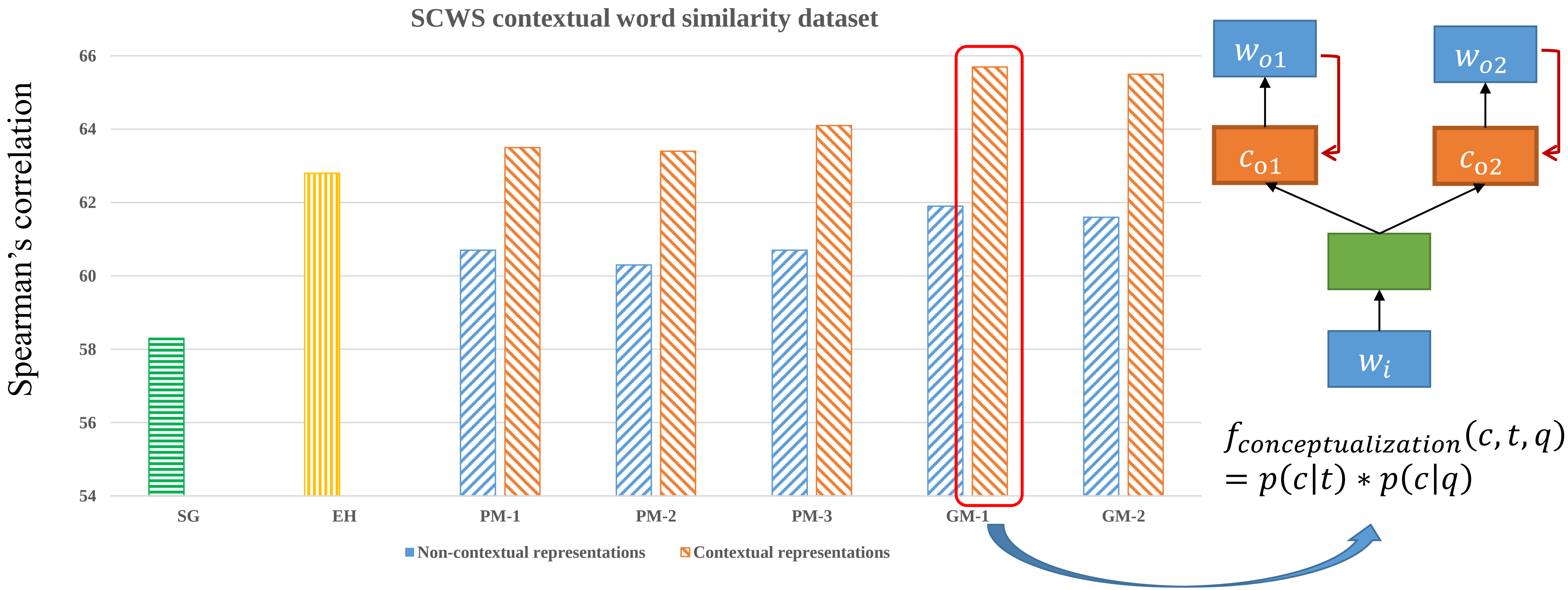
$p(fruit|apple)$ $p(company|apple)$

Concept "fruit" Concept "company"

$p(company|context)$

$p(fruit|context)$

Learning-based Conceptualization Evaluation Results



$$f_{conceptualization}(c,t,q) = p(c|t) * p(c|q)$$

- **SG**: Skip-Gram (Word2Vec)
- **EH**: Eric Huang's Sense Embedding
- **PM**: Parallel Joint-Embedding Model
- **GM**: Generative Joint-Embedding Model

Applications

- Short text understanding
- Short text similarity
- Ads/search semantic match
- Q/A system
- Query recommendation based on channels and articles
- Web table understanding
- …

~Thank You~

http://research.microsoft.com/probase/

Contact: **Zhongyuan Wang**
(email: zhy.wang # microsoft.com)