


(Spatial) Entity Search and Intelligence

Data Intelligence Lab
Yonsei University, Korea

Seung-won Hwang
<http://dilab.yonsei.ac.kr/~swhwang>

Research background

- PhD(@2005) on left-hand side
- Recent work on right-hand side



Web Images Videos Maps News More


29,500,000 RESULTS Any time ▾




Microsoft Corporation
www.microsoft.com/en-us/default.aspx
 Customer service 800-642-7676
 Get product information, support, and news from Microsoft.

<p>Download Center Welcome to the Microsoft Download Center. Find ...</p> <p>Support Microsoft Help and Support provides support for ...</p> <p>Internet Explorer Official Download—Get the latest IE internet browser ...</p>	<p>Office Try or buy Office 2013, view product information, get ...</p> <p>Products All Microsoft devices and services. Business ...</p> <p>Microsoft Dynamic... Microsoft Dynamics CRM is our customer relationship ...</p>	<p>Downloads</p> <p>Contact us</p>
---	---	--


Windows - Microsoft Windows
windows.microsoft.com
 The official website for the Microsoft Windows operating system. Explore Windows information, get downloads, and find the latest PCs for Windows.
[Downloads](#) · [Windows 8.1 Preview](#) · [Support](#) · [What is Windows 7](#) · [Meet Windows 8](#)

Microsoft - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Microsoft
 Microsoft Corporation / ˈm aɪ k r ə s ɔː f t / is an American multinational corporation headquartered in Redmond, Washington, that develops, manufactures ...
[History](#) · [Businesses](#) · [Culture](#) · [Criticism](#) · [Corporate affairs](#)



0  Sign in  

Microsoft





 Microsoft Corporation is an American multinational corporation headquartered in Redmond, Washington, that develops, manufactures, licenses, suppo... +

Founded: Apr 04, 1975 · [Albuquerque, New Mexico](#)
 Customer service: +1 800-642-7676
 CEO: [Satya Nadella](#)
 Founders: [Bill Gates](#) · [Paul Allen](#)
 Headquarters: [Redmond](#)

Stock price [See more about MSFT](#)

▼ 40.96 -0.01 -0.02%
 After Hours 40.86 -0.10 (-0.24%)
 Mar 30, 4:14 PM EDT · Market Closed

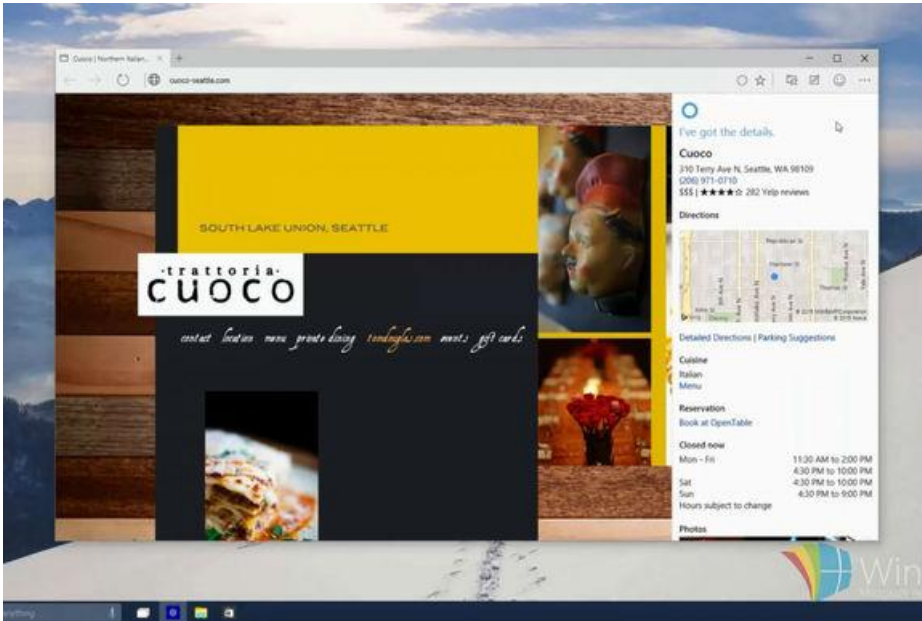
People also search for [See all \(20+\)](#)

 Apple Inc.	 Google	 IBM	 Sony
---	---	--	---

Entity search as a platform

□ Browser

□ Spreadsheet



Coffee worldwide 2011

File Edit Tools Help

Country	Population
Austria	8,355,260
Estonia	1,340,415
Belgium	10,754,528
Germany	81,882,342
Bulgaria	7,606,551
Cyprus	793,963
Czech Republic	10,476,543
Dominican Republic	7,142,554

Entity search as a platform


□ Mobile search



how tall is obama

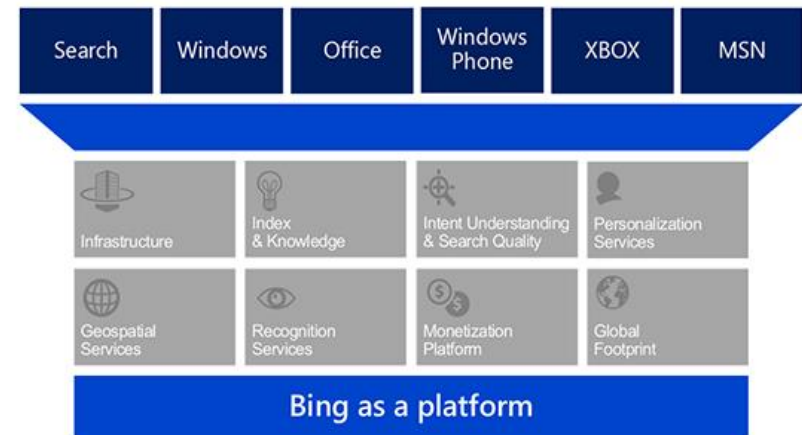
Web Images Videos Maps News More

3,560,000 RESULTS Any time

 **6' 1" (1.85 m)**
Barack Obama · Height

 George W. Bush  Michelle Obama  Bill Clinton

□ And beyond!!



Semantic Web

- ❑ Human readable vs machine readable contents
- ❑ Human defines standard for data formats and models
- ❑ Explicit and *precise* specification of knowledge representation that everyone has to agree upon

Knowledge Web

- ❑ Machine reads human readable contents
- ❑ Machine learns to conflate different formats of the same thing
- ❑ Latent and *fuzzy* representation of knowledge learned by mining big data

Recent Work

□ Harvesting, Completion (#1,#3)

AAAI, ICDE, VLDB, VLDB Journal

□ Linking, Multilingual linking (#2)

ACL, EMNLP, ACM TOIS, IEEE TKDE

□ Performance

SIGIR, WSDM, VLDB



I acquired **Bordeaux** which is 5 years of **age**. It's very pleasing in **texture** and **aroma**.

isA(Bordeaux, wine)=??

isProperty(wine|age,texture,aroma)=0.8

Verb?

bing Seung-won Hwang 황승원 黃升嫻

11,600 RESULTS Any time ▾

Seung-won Hwang | LinkedIn
<https://www.linkedin.com/pub/seung-won-hwang/1/273/458>
 CS faculty@POSTECH,Korea · Higher Education · 225 connections · Korea
 ... helping professionals like **Seung-won Hwang** discover inside connections to recommended job ... Contact **Seung-won** directly; ... **Research Intern at Microsoft**

Speakers and Abstracts - Microsoft Research
<research.microsoft.com/en-us/events/wincore2010/speakers.aspx> ▾
 Microsoft Research Asia is hosting the fourth ... **Dave Probert**, **Microsoft** ... Professor **Seung-won Hwang** is an assistant professor in the Department of ...

Seung-won Hwang
<academic.research.microsoft.com/Author/309089/seung-won-hwang> ▾
Seung-won Hwang (황승원), Pohang University of Science & Technology, Databases, Data Mining, Information Retrieval

Collaborative research improves search engine speed ...
<blogs.msdn.com> ▾ **Microsoft Research** Connections Blog ▾
 Apr 01, 2015 · ... and researchers Sameh Elnikety and Yuxiong He from **Microsoft Research**. Professor **Seung-won Hwang** participated in ... **Seung-won Hwang** and ...

Seung-won Hwang
 large data excite me
[in LinkedIn](#)
 Current: CS faculty@POSTECH,Korea
 Location: Korea

Experience

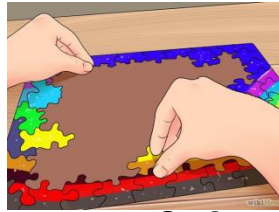
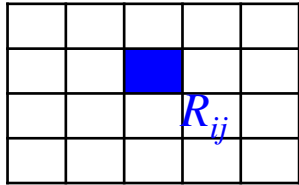
- visiting researcher
Microsoft · 2013 - 2013
- visiting researcher
Microsoft · 2012 - 2012
- research intern
Microsoft · 2004 - 2004

[More on LinkedIn](#)

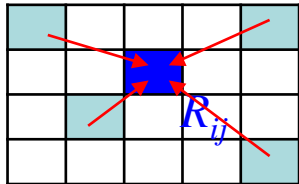
Education ▾

Conflation: Graph

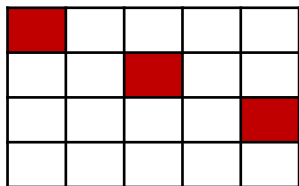
- R_{ij} is confidence of G.i matches G'.j



- Propagate matching confidence of G.i and G'.j neighbors



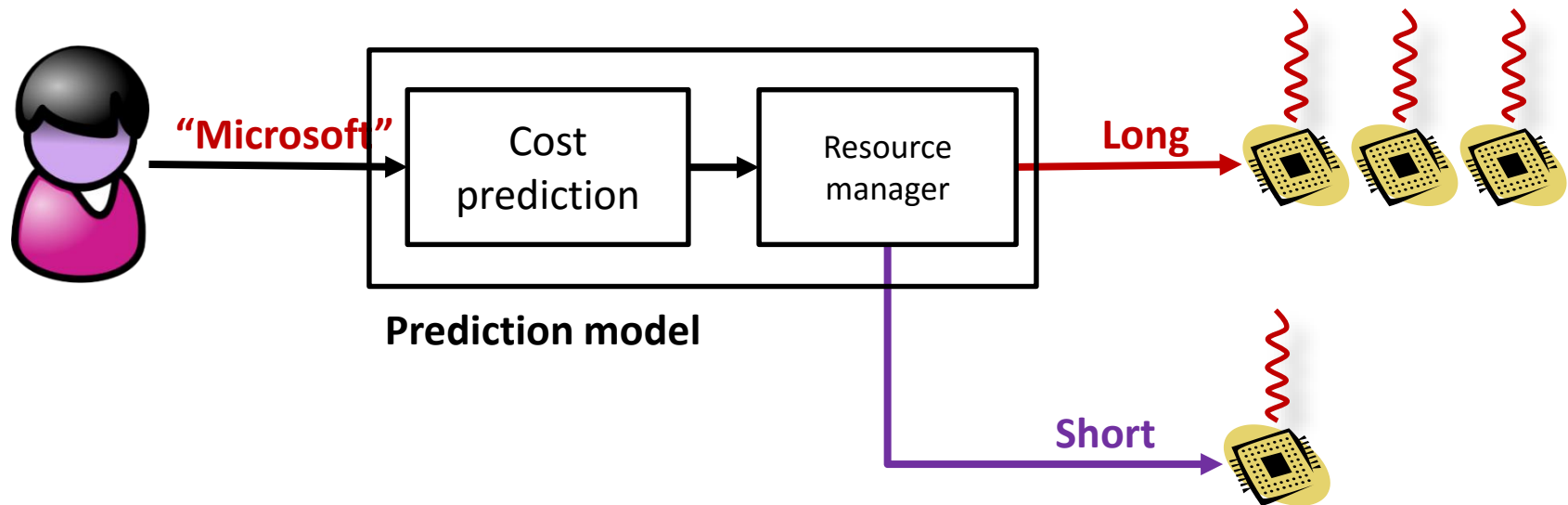
- Repeat #1 and #2 until convergence



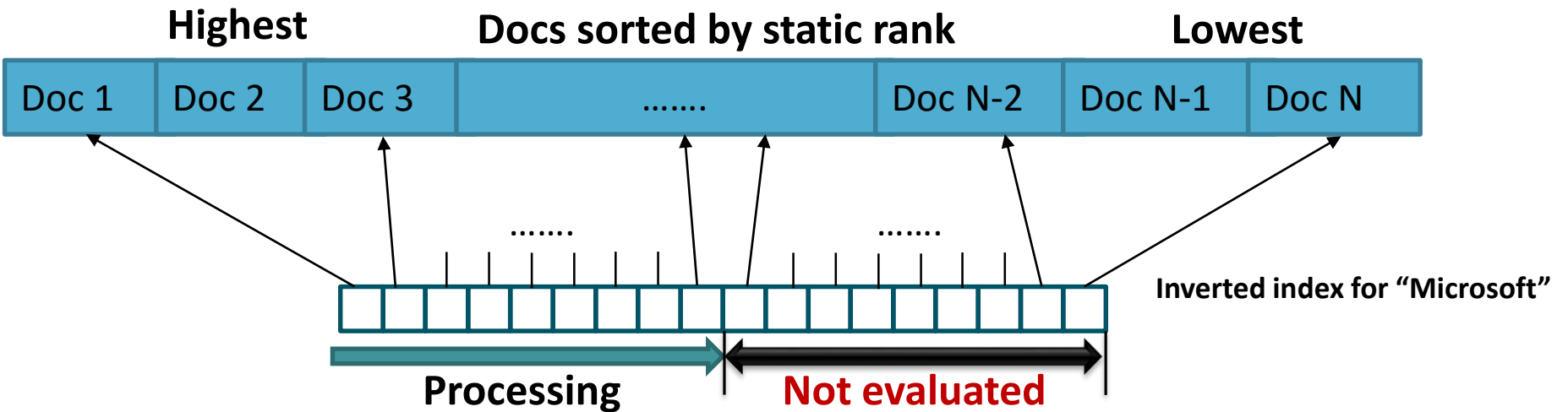
Search performance as a platform

SIGIR14, WSDM15 Best paper runner-up

- Diverse software generate search queries
- Consistent low latency is crucial



Cost prediction features



Score distribution (mean,max,var), #postings, etc

Advanced features for automatic refinement

<Fields related to query execution plan>

rank=BM25F

enablefresh=1 partialmatch=1

language=en location=us

....

<Fields related to search keywords>

Redmond (MS or Microsoft)

Ads related to microsoft office

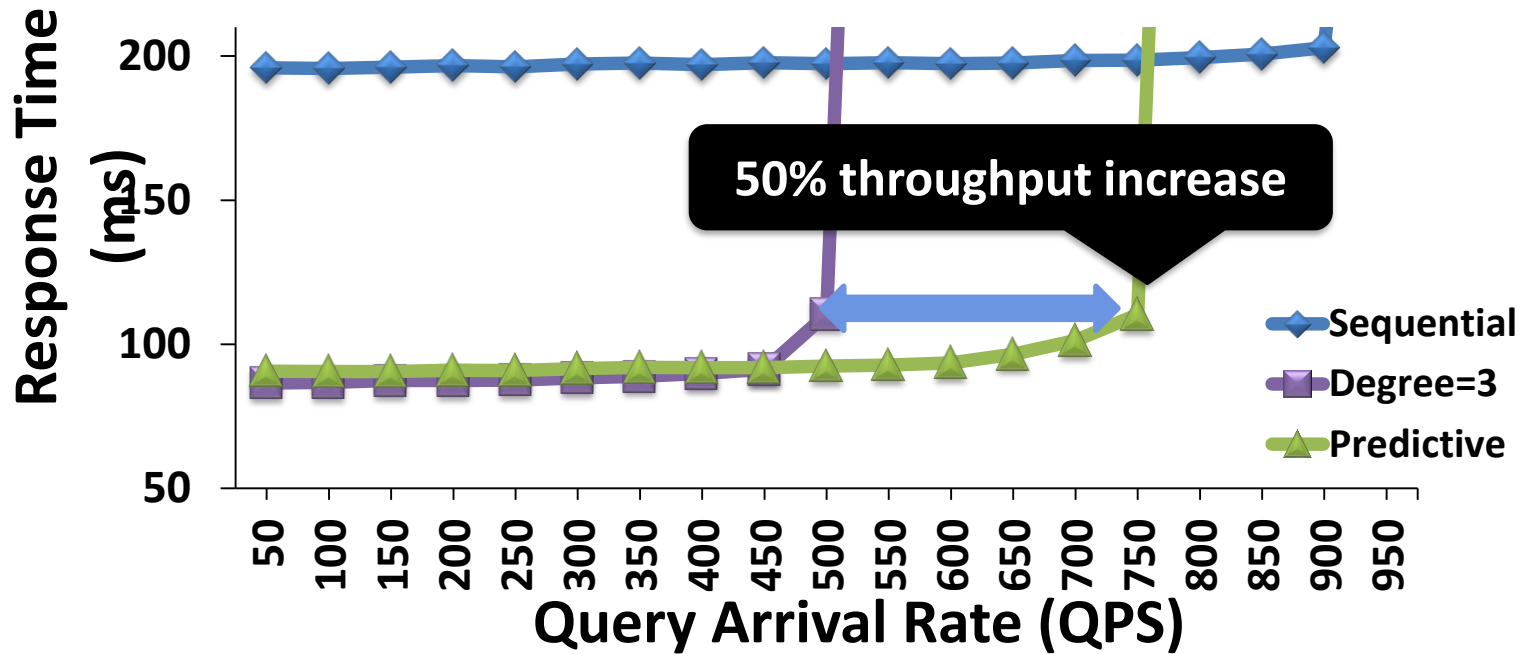
Be an **MS Office** Expert

[Simplilearn.com/MS-Office](https://simplilearn.com/MS-Office)

3,100+ followers on Twitter

5 Courses on **MS Office** - Word, Excel, PowerPoint, Access & Outlook.

Performance when deployed



Spatial KB and search as a platform

- Devices as a producer/consumer of information
- Location as a first-class citizen context



Piazza San Marco ★

Plaza

Piazza San Marco, is the principal public square of Venice, Italy, where it is generally known just as "the Piazza". All other urban spaces in the city are called "campi". [Wikipedia](#)

Address: Piazza San Marco, Venezia, Italy

Province: [Province of Venice](#)

Construction started: 1100

[Claim this business](#)

Reviews

4.6 ★★★★★ 694 Google reviews

Iconic square & gathering place dating back to the 12th century & dominated by St. Mark's Basilica. - Google

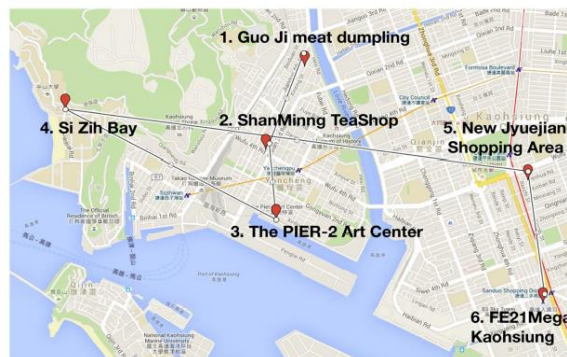
Conflation for spatial entity [AAAI16, ICDM15]

□ KB harvesting

- Map translation
- Intelligent query expansion (“seattle center” → “seattle center” or “space needle” or “Chihuly museum”)



Eiffeltoring 에펠탑 torre Eiffel আইফেল টাওয়ার eiffel torre ку
Eifel Айфеловата кула Eiffelov toranj Эйфелева вежа 艾菲尔
eiffelova věž エッフェル塔 eiffel tower Eiffel Dorrea πύρ
του Άιφελ tour eiffel Eyfel qülləsi Эйфелева башня راج ايفل



Query: One day tour in Kaohsiung
Si Zih Bay, cuisine

Recommended Results:

1. Si Zih Bay (10:28) -> 79 Style ice cream shop (12:20)
Average score: PATS:0.86, timeScore:0.97
2. Formosa Boulevard station (12:24) -> Kenting Hostel (06:14)
Average score: PATS:0.94, timeScore:0.93, socialINF:0.33
3. Formosa Boulevard station (10:00) -> seafood restaurant(14:00) ->
Si Zih Bay (17:00) -> New Jyuejiang Shopping Area (19:00) ->
FE21Mega Kaohsiung (21:00)
Average score: PATS: 0.87, timeScore:0.90, socialINF:0.4
4. Guo Ji meat dumpling (11:00) -> ShanMinng TeaShop (12:00) ->
The PIER-2 Art Center (15:00) -> Si Zih Bay (17:00) -> New
Jyuejiang Shopping Area (19:00) -> FE21Mega Kaohsiung (21:00)
Average score: PATS:0.87, timeScore:0.99, socialINF:0.16
5. Guo Ji meat dumpling (11:00) -> Si Zih Bay (17:00) -> Love River (20:00)
-> Feng-Shan Night market (04:00)
Average score: PATS:0.86, timeScore:1.0, socialINF:0.25

Query Harvest Phase



[Buca , pepperoni,
mozzarella]

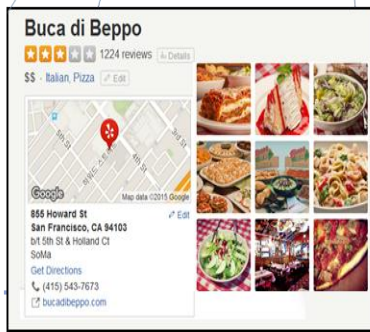
Query Photo



[pepperoni , cheese , tomato, mozzarella ,
Buca di Beppo , pepperoni pizza]



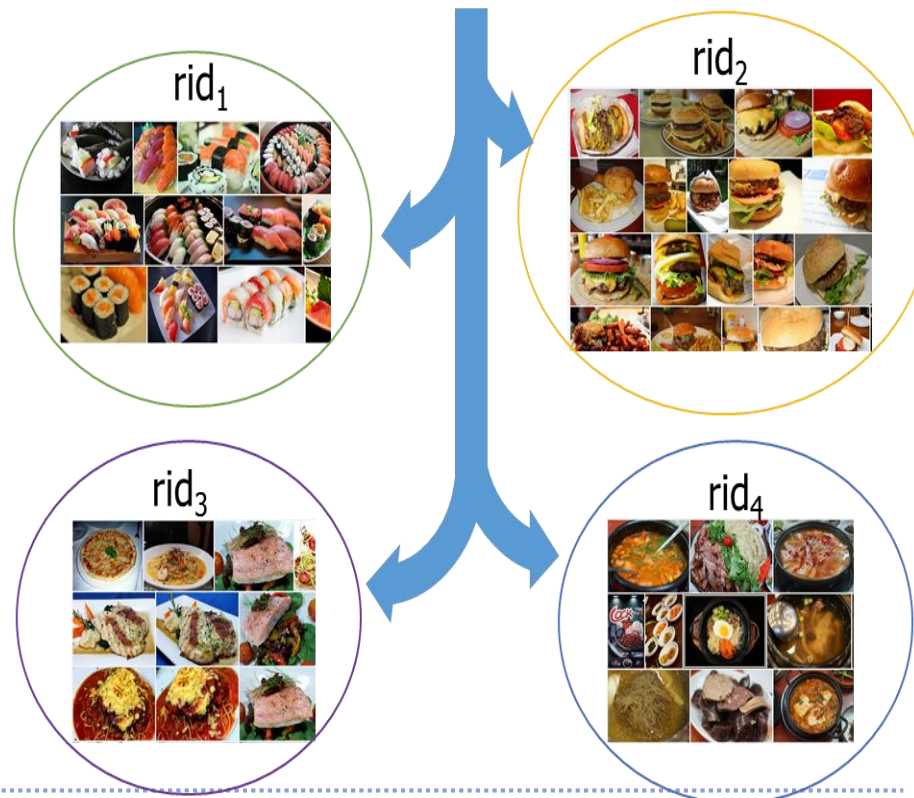
[pepperoni , cheese , tomato, mozzarella ,
Buca di Beppo , pepperoni pizza]



Nutrition Facts	
Serving Size 105.2g	
Amount Per Serving	Calories from Fat 51
Calories 234	% Daily Value*
Total Fat 5.7g	9%
Saturated Fat 2.6g	13%
Trans Fat 0.0g	
Cholesterol 13mg	4%
Sodium 13mg	19%
Potassium 100mg	3%
Total Carbohydrates 35.0g	12%
Dietary Fiber 1.6g	6%
Sugars 2.4g	
Protein 9.0g	
Vitamin A 2%	Vitamin C 2%
Calcium 12%	Iron 13%

Data Integration Phase

Social Media data



Performance [VLDB16]

□ Automatic query expansion

- restaurant → restaurant OR banquet
- “seattle center” → “seattle center” or “space needle”

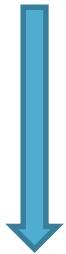
□ Multiple keywords

- Complex AND/OR with location
- Example
 - $T = ((\text{restaurant OR banquet}) \text{ AND } (\text{vegetarian OR halal}) \text{ OR } ((\text{hotel OR resort}) \text{ AND } \text{wifi}) \text{ OR } \dots$
 - S = user location (Seoul)

Additional technical challenges

Crane:

Good at narrow-necked vessel



term	id	df _t	type	ptr	storage
bar	t_1	4	tree	\hookrightarrow	aR^{t_1}
pop	t_2	2	block	\hookrightarrow	$\langle p_1, p_5 \rangle$
pub	t_3	5	tree	\hookrightarrow	aR^{t_3}
rock	t_4	2	block	\hookrightarrow	$\langle p_2, p_3 \rangle$
samba	t_5	2	block	\hookrightarrow	$\langle p_4, p_7 \rangle$

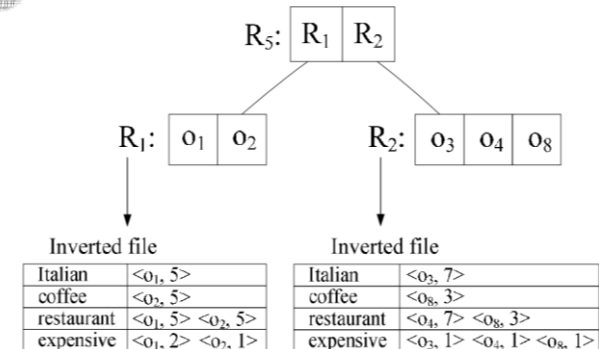
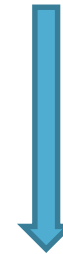
S2I: Text-first index

Good at selective



Fox:

Good at bowl



IR-tree: Augmented R-tree

Good at selective spatial



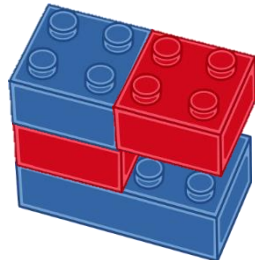
Additional technical challenges

□ Our approach

Measuring
the problem
(Cost model)

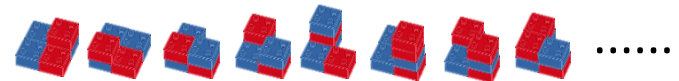


Proposing
the solution
(Optimization)



□ Challenges

- Cost model design
- Exponential possible ways (solution space)



- Efficient optimization
- Theoretic guarantee

Additional technical challenges

□ Base mapping

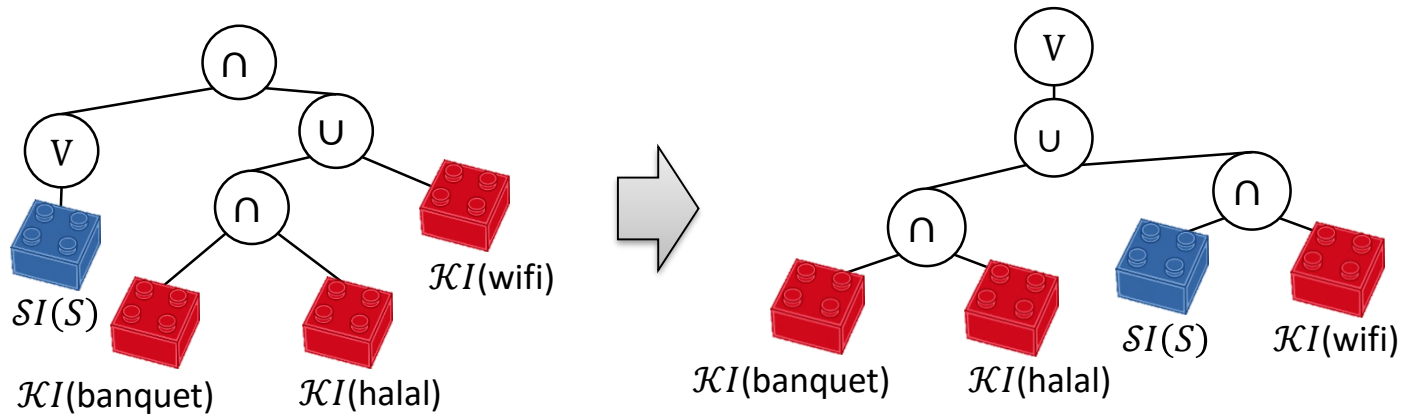
- (spatial keyword processing part) Intersection (keyword predicate processing part)

□ Optimized solution

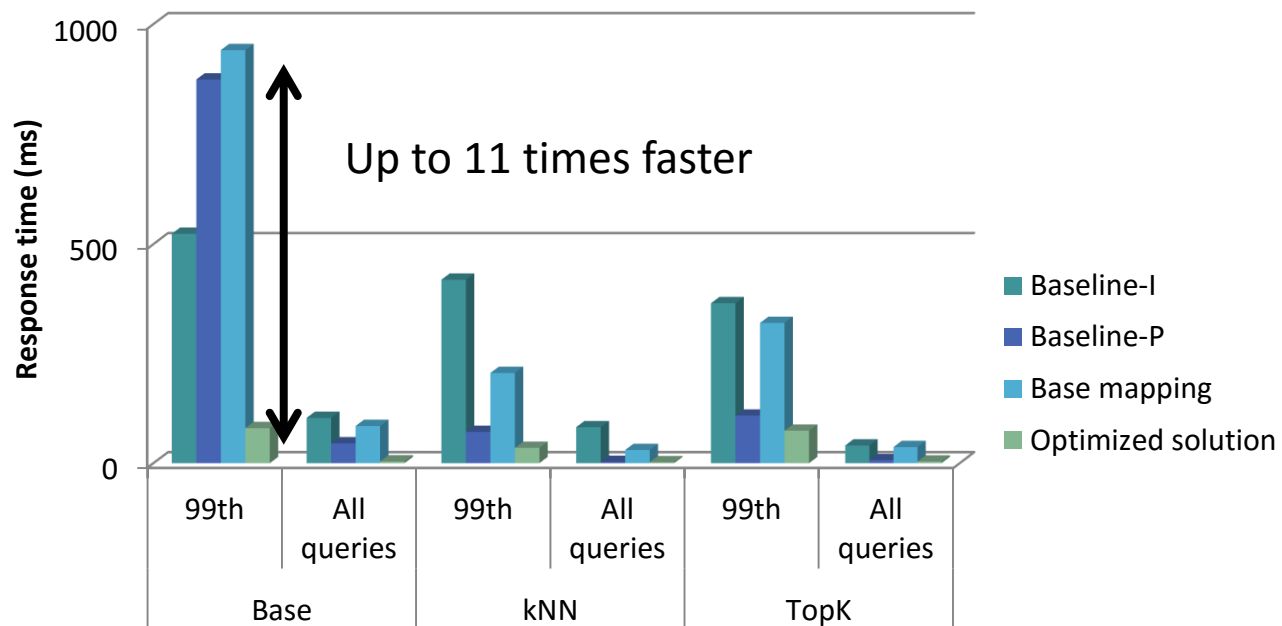
- Base mapping is optimized with the following five techniques.

	Name	Space Reduction (↑ better)	Alg. Cost (↓ better)	TheoreticBound (↓ better)
T1	Single verification pop up	2^K	Linear	OPT
T2	Intersection push down	2^F	Linear	$\left(\frac{5}{3}\right)^F \times \text{OPT}$
T3	Least selective intersection first	$\prod_{i=1}^N M_i! \cdot C_{M_i}$	Sorting	OPT
T4	Modified Huffman union tree	C_N	Sorting	OPT
T5	Verification selection	$2^{\sum_{i=1}^N (M_i - 1)}$	Exp.	OPT
			Linear	2X OPT

Additional technical challenges



Base mapping [134.7 ms] vs. Optimized solution [1.8 ms]



Thanks!!!

- Understanding Emerging Spatial Entities, AAAI 2016
- Fine-grained Semantic Conceptualization of FrameNet, AAAI 2016
- Verb Pattern: A Probabilistic Semantic Representation of Verbs, AAAI 2016
- Processing and Optimizing Main Memory Spatial-Keyword Queries, VLDB 2016
- KSTR: Keyword-aware Skyline Travel Route Recommendation, ICDM 2015
- Delayed-Dynamic-Selective (DDS) Prediction for Reducing Extreme Tail Latencies in Web Search, WSDM 2015 (Best Paper Runner-up)
- Predictive Parallelization: Taming Tail Latencies in Web Search, SIGIR 2014
- Overcoming Asymmetry in Entity Graphs, IEEE TKDE 14
- ARIA: Asymmetry-Resistant Instance Alignment, AAAI 14
- Bootstrapping Entity Translation on Weakly Comparable Corpora, ACL 13
- Entity Translation Mining from Comparable Corpora: Combining Graph Mapping with Corpus Latent Features, IEEE TKDE 13
- Efficient Entity Translation Mining: A Parallelized Graph Alignment Approach, ACM TOIS 12
- Web Scale Taxonomy Cleansing, VLDB 2011
- Mining Entity Translations from Comparable Corpora: A Holistic Graph Mapping Approach, CIKM 2011
- SocialSearch: Enhancing Entity Search with Social Network Matching, EDBT 2011

Any Questions?

Visit dilab.yonsei.ac.kr/~swhwang