(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
Entity search as a platform

- Browser
- Spreadsheet

Coffee worldwide 2011

<table>
<thead>
<tr>
<th>Country</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>8,355,260</td>
</tr>
<tr>
<td>Estonia</td>
<td>1,340,415</td>
</tr>
<tr>
<td>Belgium</td>
<td>10,754,528</td>
</tr>
<tr>
<td>Germany</td>
<td>81,882,342</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>7,606,551</td>
</tr>
<tr>
<td>Cyprus</td>
<td>793,963</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>10,476,543</td>
</tr>
<tr>
<td>Norway</td>
<td>5,389,827</td>
</tr>
</tbody>
</table>
Entity search as a platform

- Mobile search

- And beyond!!

how tall is obama

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

\[
isA(Bordeaux, \text{wine}) = ??
\]

\[
isProperty(\text{wine} | \text{age}, \text{texture}, \text{aroma}) = 0.8
\]

Verb?

황승원  黃升嫄
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)
Performance when deployed

![Graph showing performance metrics](image)

- **Response Time (ms)**
  - Y-axis: 0 to 200
  - X-axis: 50 to 950

- **Query Arrival Rate (QPS)**
  - X-axis values: 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950

- **Throughput Increase**: 50% increase

Legend:
- **Sequential**
- **Degree=3**
- **Predictive**
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”)

- Eiffel Tower
  - Eiffelturm
  - Eiffelgebouw
  - Eiffel Tower
  - Eiffel Tower

- Query: One day tour in Kaohsiung
  - Si Zhi Bay, cuisine

- Recommended Results:
  1. Si Zhi Bay (10:28) ➔ 79 Style ice cream shop (12:20)
     - Average score: PATS:0.96, 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 Zhi Bay (17:00) ➔ New Juyejang Shopping Area (19:00) ➔
       FE21 Mega Kaohsiung (21:00)
     - Average score: PATS:0.97, timeScore:0.99, socialINF:0.16
  4. Guo Ji meat dumpling (11:00) ➔ ShanMinng TeaShop (12:00) ➔
     - The PIER-2 Art Center (15:00) ➔ Si Zhi Bay (17:00) ➔ New
       Juyejang Shopping Area (19:00) ➔ FE21 Mega Kaohsiung (21:00)
     - Average score: PATS:0.96, timeScore:0.99, socialINF:0.16
  5. Guo Ji meat dumpling (11:00) ➔ Si Zhi Bay (17:00) ➔ Love River (20:00)
     ➔ Feng-Shan Night market (04:00)
     - Average score: PATS:0.86, timeScore:1.00, socialINF:0.25
Query Harvest Phase

[ Buca, pepperoni, mozzarella ]

Query Photo

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

Buca di Beppo, pepperoni pizza

Nutrition Facts

Data Integration Phase

Social Media data

rid₁

rid₂

rid₃

rid₄
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 = ((restaurant OR banquet) AND (vegetarian OR halal)
      OR ((hotel OR resort) AND wifi)
      OR ... 
    - S = user location (Seoul)
Additional technical challenges

**Crane:**
Good at narrow-necked vessel

**Fox:**
Good at bowl

<table>
<thead>
<tr>
<th>term</th>
<th>id</th>
<th>df_t</th>
<th>type</th>
<th>ptr</th>
<th>storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>bar</td>
<td>t_1</td>
<td>4</td>
<td>tree</td>
<td>→</td>
<td>aR_{t_1}</td>
</tr>
<tr>
<td>pop</td>
<td>t_2</td>
<td>2</td>
<td>block</td>
<td>→</td>
<td>\langle p_1, p_5 \rangle</td>
</tr>
<tr>
<td>pub</td>
<td>t_3</td>
<td>5</td>
<td>tree</td>
<td>→</td>
<td>aR_{t_3}</td>
</tr>
<tr>
<td>rock</td>
<td>t_4</td>
<td>2</td>
<td>block</td>
<td>→</td>
<td>\langle p_2, p_3 \rangle</td>
</tr>
<tr>
<td>samba</td>
<td>t_5</td>
<td>2</td>
<td>block</td>
<td>→</td>
<td>\langle p_4, p_7 \rangle</td>
</tr>
</tbody>
</table>

**S2I:** Text-first index
Good at selective

**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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Space Reduction</th>
<th>Alg. Cost</th>
<th>Theoretic Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 Single verification pop up</td>
<td>$2^K$</td>
<td>Linear</td>
<td>OPT</td>
</tr>
<tr>
<td>T2 Intersection push down</td>
<td>$2^F$</td>
<td>Linear</td>
<td>$\left(\frac{5}{3}\right)^F \times OPT$</td>
</tr>
<tr>
<td>T3 Least selective intersection first</td>
<td>$\prod_{i=1}^{N} M_i! \cdot C_{M_i}$</td>
<td>Sorting</td>
<td>OPT</td>
</tr>
<tr>
<td>T4 Modified Huffman union tree</td>
<td>$C_N$</td>
<td>Sorting</td>
<td>OPT</td>
</tr>
<tr>
<td>T5 Verification selection</td>
<td>$2^\sum_{i=1}^{N}(M_i-1)$</td>
<td>Exp.</td>
<td>OPT</td>
</tr>
</tbody>
</table>
Additional technical challenges

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

Up to 11 times faster
Thanks!!!

- Understanding Emerging Spatial Entities, AAAI 2016
- Fine-grained Semantic Conceptualization of FrameNet, AAAI 2016
- Verb Pattern: A Probablistic Semantic Representation of Verbs, AAAI 2016
- Processing and Optimizing Main Memory Spatial-Keyowrd 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