Spatio-Textual Similarity Joins

Panagiotis Bouros\textsuperscript{1,2}, Shen Ge\textsuperscript{1}, Nikos Mamoulis\textsuperscript{1}

\textsuperscript{1} University of Hong Kong
\textsuperscript{2} Humboldt-Universität zu Berlin

39\textsuperscript{th} International Conference on Very Large Data Bases
August 29, 2013
Complex data

• Data are becoming more complex
  – FLICKR, Foursquare, Twitter, Facebook...
    • Spatial locations
    • Textual description
    • Timestamps
    • Connectivity information (social)
  – Emerging geo-scientific fields, oceanography, seismology
    • Numerical attributes (measurements)
• Challenges for new complex queries
  – Research and industry, space as another dimension for set-value data
Motivation examples

• Social recommendation
  • Match men and women
  • Spatial locations
  • Interests

• Data de-duplication
  • Find FLICKR duplicates
  • Spatial locations
  • Tags description

39th International Conference on Very Large Data Bases
August 29, 2013
Problem definition

- **Spatio-textual objects** $o(id, loc, text)$
- **ST-SJOIN(R, S, ε, θ)**
  - Pair of objects close in space with similar textual description
  - Euclidean spatial distance
    \[ dist_l(r, s) = dist(r.loc, s.loc) \]
  - Jaccard textual similarity
    \[ sim_t(r, s) = \frac{|r.text \cap s.text|}{|r.text \cup s.text|} \]
  - Subset of R x S with $dist_l(r, s) \leq \varepsilon$ and $sim_t(r, s) \geq \theta$
Problem definition (cont’d)

\[
\text{ST-SJOIN}(R, R, \varepsilon = 0.2, \theta = 0.7)
\]

\[
\begin{align*}
  x_1 & \quad \{B,C\} & x_6 & \quad \{C,D,E,F\} \\
  x_2 & \quad \{E,F\} & x_7 & \quad \{A,B,C,D,F\} \\
  x_3 & \quad \{D,E,F\} & x_8 & \quad \{A,B,D,E,F\} \\
  x_4 & \quad \{A,B,E,F\} & x_9 & \quad \{A,B,C,D,E\} \\
  x_5 & \quad \{C,D,E,F\} \\
\end{align*}
\]
Problem definition (cont’d)

ST-SJOIN(R, R, \( \epsilon = 0.2 \), \( \theta = 0.7 \))

- \( x_1 \) \{B,C\}
- \( x_2 \) \{E,F\}
- \( x_3 \) \{D,E,F\}
- \( x_4 \) \{A,B,E,F\}
- \( x_5 \) \{C,D,E,F\}
- \( x_6 \) \{C,D,E,F\}
- \( x_7 \) \{A,B,C,D,F\}
- \( x_8 \) \{A,B,D,E,F\}
- \( x_9 \) \{A,B,C,D,E\}
Problem definition (cont’d)

ST-SJOIN(R, R, $\varepsilon = 0.2$, $\theta = 0.7$)

- $x_1 \{B, C\}$
- $x_2 \{E, F\}$
- $x_3 \{D, E, F\}$
- $x_4 \{A, B, E, F\}$
- $x_5 \{C, D, E, F\}$
- $x_6 \{C, D, E, F\}$
- $x_7 \{A, B, C, D, F\}$
- $x_8 \{A, B, D, E, F\}$
- $x_9 \{A, B, C, D, E\}$
Outline

• Introduction
• Background on set similarity joins
• Computing spatio-textual similarity joins
• Experimental analysis
• Conclusions and future work
Set similarity joins and PPJOIN

[Xiao et al @ WWW’08]

- Inverted index to compute overlaps [Sarawagi et al @ SIGMOD’04]
- Prefix filtering [Chaudhuri et al @ ICDE’06]
- Two-phase method [Bayardo et al @ WWW’07]
  - Objects by length
  - Read-Probe-Index
- Positional filter
- Suffix filter

Hamming distance lower bound
Overlap upper bound
Computing ST-SJOIN

• **Textual similarity** join
  – Build upon **PPJOIN**

• **Spatial distance** join
  – Filtering, dynamic grid partitioning, R-tree

• **Methods**
  – PPJ
  – PPJ-I
  – PPJ-C
  – PPJ-R

• **Grouping**
Spatial filtering and PPJ

- **Straightforward approach**
  - Extend PPJOIN
  - Add another filter before positional and suffix

\[ \text{dist}_I(r,s) \leq \varepsilon \]

**ST-SJOIN(R, R, \varepsilon = 0.2, \theta = 0.7)**

\[
\begin{align*}
    x_1 & \quad \{B,C\} & x_6 & \quad \{C,D,E,F\} \\
    x_2 & \quad \{E,F\} & x_7 & \quad \{A,B,C,D,F\} \\
    x_3 & \quad \{D,E,F\} & x_8 & \quad \{A,B,D,E,F\} \\
    x_4 & \quad \{A,B,E,F\} & x_9 & \quad \{A,B,C,D,E\} \\
    x_5 & \quad \{C,D,E,F\}
\end{align*}
\]
Spatial filtering and PPJ

• **Straightforward approach**
  – Extend PPJOIN
  – Add another filter before positional and suffix
  
  $\text{dist}_1(r,s) \leq \varepsilon$

• **Problem**
  – Lack of spatial indexing
  – Examines objects no matter how far from $x_3$

$\text{ST-SJOIN}(R, R, \varepsilon = 0.2, \theta = 0.7)$

---

$x_1 \{B, C\}$ $x_6 \{C, D, E, F\}$
$x_2 \{E, F\}$ $x_7 \{A, B, C, D, F\}$
$x_3 \{D, E, F\}$ $x_8 \{A, B, D, E, F\}$
$x_4 \{A, B, E, F\}$ $x_9 \{A, B, C, D, E\}$
$x_5 \{C, D, E, F\}$
Dynamic grid partitioning

- Grid partitioning
  - On the fly
  - Extend of a grid cell equals $\varepsilon$
  - Numbering from left to right from bottom to top

39th International Conference on Very Large Data Bases  August 29, 2013
Dynamic grid partitioning

- Grid partitioning
  - On the fly
  - Extend of a grid cell equals $\epsilon$
  - Numbering from left to right from bottom to top

- Property
  - Objects spatially joinable inside at most 9 cells
  - Still need to verify w.r.t. $\epsilon$
Dynamic grid partitioning

- Grid partitioning
  - On the fly
  - Extend of a grid cell equals $\epsilon$
  - Numbering from left to right from bottom to top

- Property
  - Objects spatially joinable inside at most 9 cells
  - Still need to verify w.r.t. $\epsilon$
Dynamic grid partitioning and PPJ-I

• Spatial information inside inverted index
  – Sort postings by cell id
  – Lightweight index on top of postings

---

Spatial distance join with space filling curve
Dynamic grid partitioning and PPJ-I

- **Spatial information inside inverted index**
  - Sort postings by cell id
  - Lightweight index on top of postings

- **Joinable neighborhood**
  - At most three cell intervals

- **Spatial distance join with space filling curve**

<table>
<thead>
<tr>
<th>c1</th>
<th>c2</th>
<th>c3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- c\(_{37}\): [28,30], [36,38], [44,46]
- c\(_{2}\): [1,3], [9,11]

---

39\(^{th}\) International Conference on Very Large Data Bases  
August 29, 2013
Dynamic grid partitioning and PPJ-C

• Working at the cell-level
• For each cell
  – Build inverted index
  – Define set $A[c]$, cells among 9 adjacent with smaller or equal id

$A[c_{37}] = \{c_{28}, c_{29}, c_{30}, c_{36}, c_{37}\}$

$A[c_{2}] = \{c_{1}, c_{2}\}$
Dynamic grid partitioning and PPJ-C

- Working at the cell-level
- For each cell
  - Build inverted index
  - Define set $A[c]$, cells among 9 adjacent with smaller or equal id
  - $ST$-$SJOIN(c, c, \varepsilon, \theta)$
  - $ST$-$SJOIN(c, c', \varepsilon, \theta)$ for each cell $c'$ in $A[c]$
  - Discard $c$ after finish with all cell in $A[c]$
R-tree and PPJ-R

- **Similar** to PPJ-C but:
  - *Static* partitioning, objects **indexed offline** by R-tree
  - **No connection** to $\varepsilon$
- **ST-SJOIN** based on **$\varepsilon$-distance join using R-trees**
  - Traversing R-tree determines which partitions to join
Grouping

<table>
<thead>
<tr>
<th>object</th>
<th>x.text</th>
<th>( \text{ppref}(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>{B,C}</td>
<td>{B}</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>{E,F}</td>
<td>{E}</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>{D,E,F}</td>
<td>{D}</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>{A,B,E,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td>( x_5 )</td>
<td>{C,D,E,F}</td>
<td>{C,D}</td>
</tr>
<tr>
<td>( x_6 )</td>
<td>{C,D,E,F}</td>
<td>{C,D}</td>
</tr>
<tr>
<td>( x_7 )</td>
<td>{A,B,C,D,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td>( x_8 )</td>
<td>{A,B,D,E,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td>( x_9 )</td>
<td>{A,B,C,D,E}</td>
<td>{A,B}</td>
</tr>
</tbody>
</table>

- **Problems**
  - Same prefix index more than once
## Grouping

<table>
<thead>
<tr>
<th>object</th>
<th>x.text</th>
<th>ppref(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>{B,C}</td>
<td>{B}</td>
</tr>
<tr>
<td>$x_2$</td>
<td>{E,F}</td>
<td>{E}</td>
</tr>
<tr>
<td>$x_3$</td>
<td>{D,E,F}</td>
<td>{D}</td>
</tr>
<tr>
<td>$x_4$</td>
<td>{A,B,E,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td>$x_5$</td>
<td>{C,D,E,F}</td>
<td>{C,D}</td>
</tr>
<tr>
<td>$x_6$</td>
<td>{C,D,E,F}</td>
<td>{C,D}</td>
</tr>
<tr>
<td>$x_7$</td>
<td>{A,B,C,D,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td>$x_8$</td>
<td>{A,B,D,E,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td>$x_9$</td>
<td>{A,B,C,D,E}</td>
<td>{A,B}</td>
</tr>
</tbody>
</table>

### Problems

- Same prefix index more than once
- Some overlap calculated more than once
## Grouping

<table>
<thead>
<tr>
<th>group</th>
<th>object</th>
<th>x.text</th>
<th>ppref(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>$x_1$</td>
<td>{B,C}</td>
<td>{B}</td>
</tr>
<tr>
<td>$g_2$</td>
<td>$x_2$</td>
<td>{E,F}</td>
<td>{E}</td>
</tr>
<tr>
<td>$g_3$</td>
<td>$x_3$</td>
<td>{D,E,F}</td>
<td>{D}</td>
</tr>
<tr>
<td></td>
<td>$x_5$</td>
<td>{C,D,E,F}</td>
<td>{C,D}</td>
</tr>
<tr>
<td>$g_4$</td>
<td>$x_4$</td>
<td>{A,B,E,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td></td>
<td>$x_6$</td>
<td>{C,D,E,F}</td>
<td>{C,D}</td>
</tr>
<tr>
<td></td>
<td>$x_7$</td>
<td>{A,B,C,D,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td></td>
<td>$x_8$</td>
<td>{A,B,D,E,F}</td>
<td>{A,B}</td>
</tr>
<tr>
<td></td>
<td>$x_9$</td>
<td>{A,B,C,D,E}</td>
<td>{A,B}</td>
</tr>
</tbody>
</table>

### Problems
- Same prefix index more than once
- Some overlap calculated more than once

### Grouping objects by prefix
- Massive pruning
Grouping for ST-SJOIN

• **Textually**
  – Group objects by *length and prefix*
  – Examination order retained, PPJOIN fully applicable
  – If $|g_x| \geq |g_y|$ then $|x| \geq |y|$ for $x$ in $g_x$ and $y$ in $g_y$

• **Spatially**
  – PPJ: group objects *no matter how far*
  – PPJ-I,PPJ-C: group objects inside *grid cells*

• **Join process**
  – Probing and indexing *over groups*
  – Suffix filter *not useful*
  – *Unfold groups during verification*
Experimental analysis

- **Real** collections
  - FLICKR, NY, |R| = 1.5M, |T| = 730K, avg size 10.5
  - POI-USCA, California state, |R| = 1.5M, |T| = 16K, avg size 4.4
  - POI-AU, Australia, |R| = 700K, |T| = 2.6K, avg size 4.7

- **Synthetic** collections
  - |R| = {30K, 100K, 500K, 1M, 3M}
  - |T| = {5K, 10K, 50K, 100K, 300K}
  - Spatial distribution, uniform or clustered
  - Correlated

- **Experiments**
  - Measure response time
  - Vary $\varepsilon = \{0.001, 0.005, 0.01, 0.05, 0.1\}$ synthetic $\{0.001, 0.005, 0.01, 0.05\}$ real
  - Vary $\theta = \{0.5, 0.6, 0.7, 0.8, 0.9\}$ synthetic, $\{0.6, 0.7, 0.8, 0.9\}$ real
To group or not to group

FLICKR $\varepsilon = 0.005$, $\theta = 0.9$

POI-AU $\varepsilon = 0.05$, $\theta = 0.7$
Comparison with baseline methods

**FLICKR**

<table>
<thead>
<tr>
<th>Number of results</th>
<th>Response time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>53M</td>
<td>10^5</td>
</tr>
<tr>
<td>44M</td>
<td>10^4</td>
</tr>
<tr>
<td>22M</td>
<td>10^3</td>
</tr>
<tr>
<td>17M</td>
<td>10^2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\epsilon/\theta$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001/0.6</td>
<td>RT</td>
</tr>
<tr>
<td>0.005/0.7</td>
<td>PPJOIN</td>
</tr>
<tr>
<td>0.01/0.8</td>
<td>PPJ</td>
</tr>
<tr>
<td>0.05/0.9</td>
<td>PPJ-R</td>
</tr>
</tbody>
</table>

**POI-USCA**

<table>
<thead>
<tr>
<th>Number of results</th>
<th>Response time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>44M</td>
<td>10^6</td>
</tr>
<tr>
<td>100M</td>
<td>10^5</td>
</tr>
<tr>
<td>130M</td>
<td>10^4</td>
</tr>
<tr>
<td>700M</td>
<td>10^3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\epsilon/\theta$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001/0.6</td>
<td>RT</td>
</tr>
<tr>
<td>0.005/0.7</td>
<td>PPJOIN</td>
</tr>
<tr>
<td>0.01/0.8</td>
<td>PPJ</td>
</tr>
<tr>
<td>0.05/0.9</td>
<td>PPJ-R</td>
</tr>
</tbody>
</table>

39th International Conference on Very Large Data Bases
August 29, 2013
IR-tree and PPJ-IR

FLICKR

POI-USCA

Response time (sec)

Number of results

247M  218M  127M  126M

401M  773M  353M  516M

$\frac{\epsilon}{\theta}$

$0.001/3$  $0.005/5$  $0.01/7$  $0.05/9$

$0.001/1$  $0.005/2$  $0.01/3$  $0.05/4$

PPJ-IR  PPJ-R  PPJ-C
Conclusions and future work

• Conclusions
  – New join query, ST-SJOIN
  – Evaluation algorithms
    • State-of-the-art on set similarity joins
    • Spatial indexing
    • PPJ-C in general most efficient method

• Future work
  – Study PPJ-C’s advantage on distributed environments
  – Consider other dimensions, e.g., time or graph
Questions?
Backup slides
Set similarity joins

- For every term $t$ in object $S$: [Sarawagi et al @ SIGMOD’04]
  - Probe inverted index, traverse postings list $L_t$
  - Compute overlap $O[\text{ }, \text{ }]$ with every object

- Optimization
  - Build inverted index on the fly, incrementally
  - Compute overlap between two object only once
Set similarity joins

- Prefix filtering [Chaudhuri et al @ ICDE’06]
  - Global ordering of terms, canonicalized objects
  - Prefixes w.r.t. \( \theta \) should share at least one term
Set similarity joins

- AllPairs [Bayardo et al @ WWW’07]
  - Builds upon prefix-filtering
  - Examine objects by length, ascending
  - Reduce indexing cost
    - Index prefix of an object
  - Length filter
Set similarity joins

- **PPJOIN** [Xiao et al @ WWW’08]
  - Builds upon AllPairs
  - Positional filter
  - Suffix filter

[Diagram showing set similarity join process with objects and inverted index filtering.]
Dynamic grid partitioning and PPJ-I

- When examining $x_4$ in $c_7$

$c_7$: [1,3], [6,8], [11,13]
- $c_{15}$ is not inside the joinable neighborhood of $c_7$
Dynamic grid partitioning and PPJ-I

- When examining $x_5$ in $c_{19}$
  - $c_{25}$ is inside the joinable neighborhood of $c_{19}$
  - Need to check Euclidean distance

```
\begin{align*}
&c_{19}: [13,15], [18,20], [23,25] \\
&- c_{25} \text{ is inside the joinable neighborhood of } c_{19}
\end{align*}
```

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>${B,C}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2$</td>
<td>${E,F}$</td>
</tr>
<tr>
<td>$x_3$</td>
<td>${D,E,F}$</td>
</tr>
<tr>
<td>$x_4$</td>
<td>${A,B,E,F}$</td>
</tr>
<tr>
<td>$x_5$</td>
<td>${C,D,E,F}$</td>
</tr>
<tr>
<td>$x_6$</td>
<td>${C,D,E,F}$</td>
</tr>
<tr>
<td>$x_7$</td>
<td>${A,B,C,D,F}$</td>
</tr>
<tr>
<td>$x_8$</td>
<td>${A,B,D,E,F}$</td>
</tr>
<tr>
<td>$x_9$</td>
<td>${A,B,C,D,E}$</td>
</tr>
</tbody>
</table>

ST-SJOIN($R$, $R$, $\varepsilon = 0.2$, $\theta = 0.7$)