Learning to Query:
Focused Web Page Harvesting for Entity Aspects

Yuan Fang\textsuperscript{1}, Vincent Zheng\textsuperscript{2}, Kevin Chang\textsuperscript{23}

ICDE 2016 @ Helsinki

\textsuperscript{1} Institute for Infocomm Research, Singapore
\textsuperscript{2} Advanced Digital Sciences Center, Singapore
\textsuperscript{3} University of Illinois at Urbana-Champaign, USA
In this talk

- **Problem: L2Q**
- Challenges and solution
  - Domain-awareness
  - Context-awareness
- Experimental Study
- Conclusion
Entities and their aspects are abundant, but scattered, on the Web

<table>
<thead>
<tr>
<th>Entity type</th>
<th>Common aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>celebrity</td>
<td>spouse, age, net worth, …</td>
</tr>
<tr>
<td>car</td>
<td>safety, cost, interior, …</td>
</tr>
<tr>
<td>business</td>
<td>address, opening hour, phone no., …</td>
</tr>
</tbody>
</table>

People entities alone: 10% of Bing’s search volume
Motivation:
Focused Web Page Harvesting for Entity Aspects

Business Analytics
Vertical portal or search
High level problem: Learning to query (L2Q)

- **Seed query**: Keywords (uniquely) identifying the entity
- **Target aspect**: A pre-trained classifier $Y$, mapping each page to “relevant” or “not relevant” to the target aspect
- **Utility (precision/recall)**: In each iteration, $q^* = \arg \max_q \mathcal{U}^{(Y)}(q)$

![Diagram of L2Q process]

- A pre-trained classifier $Y$, mapping each page to “relevant” or “not relevant” to the target aspect.
- In each iteration, $q^* = \arg \max_q \mathcal{U}^{(Y)}(q)$
In this talk

- Problem: L2Q
- Challenges and solution
  - Domain-awareness
  - Context-awareness
- Experimental Study
- Conclusion
Subproblem #1: Domain-aware L2Q

Target Entity
Marc Snir

 ✓ Gradually harvested in earlier iterations

entity pages $P_E$

Peer Entities
Philips Yu
Andrew Ng

 .....

 ✓ Pre-collected / previously harvested pages
domain pages $P_D$

$$q^* = \arg \max_q \mathcal{U}^{(Y)}(q | P_E, P_D)$$
# Subproblem #1: Vocabulary variations

<table>
<thead>
<tr>
<th>Entity</th>
<th>Example page content</th>
<th>Example query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marc Snir</td>
<td>...many HPC papers in IJHPCA ...</td>
<td>hpc ijhpc</td>
</tr>
<tr>
<td>Philip Yu</td>
<td>...his data mining papers in TKDE ...</td>
<td>data mining tkde</td>
</tr>
<tr>
<td>Andrew Ng</td>
<td>...his recent AI paper in JMLR ...</td>
<td>ai jmlr</td>
</tr>
</tbody>
</table>

**topics**

- hpc
- data mining
- ai

**journals**

- ijhpc
- tkde
- jmlr

(topic) (template)
Subproblem #1: Bridging domain and entity phases

Domain Phase

only once

domain graph

Entity Phase

for each query selection

domain graph

Utility regularization
(i.e. supervision on target aspect)
In this talk

- Problem: L2Q
- **Challenges and solution**
  - Domain-awareness
  - **Context-awareness**
- Experimental Study
- Conclusion
Subproblem #2: Context-aware L2Q

- In iteration-\(i\), a context of already fired queries
  \(\Phi = \{q^{(0)}, q^{(1)}, \ldots, q^{(i-1)}\}\).
- Queries can retrieve redundant pages.
Subproblem #2: Accounting for redundancy

Collective Utilities

Collective precision: \[ \frac{|(\Omega(q) \cup \Omega(\Phi)) \cap \Omega(Y)|}{|\Omega(q) \cup \Omega(\Phi)|} \]

Collective recall: \[ \frac{|(\Omega(q) \cup \Omega(\Phi)) \cap \Omega(Y)|}{|\Omega(Y)|} \]
In this talk

- Problem: L2Q
- Challenges and solution
  - Domain-awareness
  - Context-awareness
- Experimental Study
- Conclusion
Experiment setup

- **Datasets:** two domains
  - 996 researchers & 143 car models
  - Pre-collected pages to simulate the corpus
- **Search engine:** language model
- **Dictionaries for templates**
  - Gathered from existing knowledge base
  - Manually compiled
- **Entity aspects**
  - 7 attributes for each domain
  - Pre-trained aspect classifier with high accuracy
Experiment methodology

- **Utilities of two forms:** precision & recall
- **Evaluation metrics**
  - Precision, recall
  - Combined F-score
- **Metrics reported are normalized**
  - Against ideal precision/recall
  - Ideal metrics computed by “peeking” at unretrieved pages
Finding #1: Effect of domain and context-awareness

- **RND**: select query randomly
- **P/R**: optimize precision/recall without domain and context-awareness
- **P/R+q**: with domain pages, but do not employ templates, and without context
- **P/R+t**: with domain pages and templates, without context
- **L2QP/L2QR**: full approaches optimizing precision/recall
Finding #2(a):
Comparing precision with indep. baselines

- **LM**: language feedback model
- **AQ**: adaptive querying for text databases
- **HR**: harvest rate for hidden structured databases
- **MQ**: manually designed queries
Finding #2(b): Comparing recall with indep. baselines

- **LM**: language feedback model
- **AQ**: adaptive querying for text databases
- **HR**: harvest rate for hidden structured databases
- **MQ**: manually designed queries
Finding #2(c):
Comparing F-score with indep. baselines

- **L2QBAL**: optimize for F-score, balancing L2QP & L2QR
- **LM**: language feedback model
- **AQ**: adaptive querying for text databases
- **HR**: harvest rate for hidden structured databases
- **MQ**: manually designed queries
In this talk

- Problem: L2Q
- Challenges and solution
  - Domain-awareness
  - Context-awareness
- Experimental Study
- Conclusion
Conclusion

- L2Q: a novel paradigm of crawling
- Domain-aware L2Q
  - Templates to handle vocabulary variations
- Context-aware L2Q
  - Collective utilities to account for page redundancy