Research on Expert Search at Enterprise Track of TREC 2006

Presenter: Nick Craswell

Shenghua Bao, Huizhong Duan, Qi Zhou, Miao Xiong, Yunbo Cao, Yong Yu

APEX Data and Knowledge Management Lab @ SJTU
Nov.17, 2006
Outline

- **Overview**
- **Our approach**
  - Offline processing
    - Personal name identification
    - Metadata extraction
  - Online processing
    - Evidence extraction
    - Evidence evaluation
    - Evidence merging
- **Submitted runs**
- **Summary**
Overview

- Evidence plays a key role in expert search
- Three issues on using evidence
  - Evidence extraction
    - Q: relationship cardinalities
  - Evidence quality evaluation
    - How’s the differences between queries, experts, relations and documents?
    - And How much?
  - Evidence merging
    - How can each evidence support an expert?
    - Should less evidences with high quality beat more ones with low quality?
Overview

Fig.1 Overview of the Evidence Oriented Framework
Outline

- Overview
- Our approach
  - Offline processing
    - Personal name identification
    - Metadata extraction
  - Online processing
    - Evidence extraction
    - Evidence evaluation
    - Evidence merging
- Submitted runs
- Summary
Offline Processing

- **Person Name Identification**
  - Person Name Identification
  - Person Name Disambiguation
  - Probability of Expert Identification Mask

- **Metadata Extraction**
  - Title
  - Author
  - Reference Block
  - Email & Thread
 Masks of Experts

• Full Name/Email (given)
• Combine Name
• Abbreviated name
• Short Name
• Nick Name
• New Email

<table>
<thead>
<tr>
<th>Mask</th>
<th>Sample</th>
<th>Rate</th>
<th>Ambiguity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Name</td>
<td>Ritu Raj Tiwari Tiwari, Ritu Raj</td>
<td>48.2%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:rtiwari@nuance.com">rtiwari@nuance.com</a></td>
<td>20.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Combine Name</td>
<td>Tiwari, Ritu R R Tiwari</td>
<td>4.2%</td>
<td>39.92%</td>
</tr>
<tr>
<td>Abbreviated Name</td>
<td>Ritu Raj Ritu</td>
<td>21.2%</td>
<td>48.90%</td>
</tr>
<tr>
<td>Short Name</td>
<td>RRT</td>
<td>0.7%</td>
<td>63.96%</td>
</tr>
<tr>
<td>Nick &amp; New Mail</td>
<td>Ritiwari</td>
<td>7%</td>
<td>0.46%</td>
</tr>
</tbody>
</table>

http://apex.sjtu.edu.cn/apex_wiki/FrontPage
Person Name Disambiguation

- **Person Name Disambiguation**
  - Scan document to identify all unambiguous names
  - Find the ambiguous name and measure the ‘distance’ with its possible experts
  - Choose the closest pair of ambiguous name and its unambiguous form by the distance

- **Probability of Different Expert Mask**
  - Uncertainty from expert identifications
  - Different weights of experts identification masks.
Metadata Extraction

- **Title**
  - Brief summary of document
  - Extract from `<title>`, `<h1>` or `<h2>`

- **Author**
  - Strong associate with document
  - Experts within the author scale from document title is considered
Metadata Extraction

- **Reference Block**
  - Publish and Author
  - Very traceable structure from documents

- **Email Handling**
  - Remove the noise by ‘Next’ or ‘Previous’
  - Remove the duplicated Email

- **PageRank**
  - Calculated Based on Link Structure of W3C Corpus
  - Using the standard PageRank algorithm including damping factor \( q=0.85 \)
Outline

- Overview
- Our approach
  - Offline processing
    - Personal name identification
    - Metadata extraction
  - Online processing
    - Evidence extraction
    - Evidence evaluation
    - Evidence merging
- Submitted runs
- Summary
Query preprocessing

- Phrase query
  - Original query

- Bi-gram query
  - “Semantic Web Rule Language”
    - “Semantic Web”, “Web Rule”, “Rule Language”

- Proximity query
  - “Semantic Web Rule Language”~20
    - Matched when each two words appear within 20 words distance
Query preprocessing

- Fuzzy query
  - “mereology”
    - “methology”
    - “ontology”
    - And so on…

- Stemming query
  - “SOAP security considerations”
    - “soap secur consider”
Evidence extraction

- window-based co-occurrence relation
- section-constrained co-occurrence relation
- block-based co-occurrence relation
- metadata-based co-occurrence relation
- semantic-block-based co-occurrence relation
Examples of Improved block-based co-occurrence relation

- Len.Greski@granger.com
  - Re: quote use case (Monday, 16 June)
  - Re: quote use case (Wednesday, 11 June)
  - quote use case (Wednesday, 11 June)
- Lipton, Paul C
  - RE: Requirements: Decision Points Requirement Proposals (Tuesday, 10 June)
- Martin Chapman
  - RE: Requirements: Decision Points Requirement Proposals (Monday, 30 June)
  - FW BPSS.tfj.june03.ppt (Thursday, 19 June)
  - FW BPSS.tfj.june03.ppt (Thursday, 10 June)
  - RE: relationship cardinalities (was... RE: Requirements: Decision Points Requirement Proposals)
  - RE: Requirements: Decision Points Requirement Proposals (Tuesday, 10 June)
  - RE: relationship cardinalities (was... RE: Requirements: Decision Points Requirement Proposals)
  - RE: relationship cardinalities (was... RE: Requirements: Decision Points Requirement Proposals)
  - RE: relationship cardinalities (was... RE: Requirements: Decision Points Requirement Proposals)
  - RE: relationship cardinalities (was... RE: Requirements: Decision Points Requirement Proposals)
  - Agenda for tomorrow's call (Monday, 2 June)
  - updated minutes con call 26 May 2003 (Monday, 2 June)
  - RE: Partial executability determination of a Chore description language (Monday, 2 June)
- Monica J. Martin
  - Re: travel agency use cases (Monday, 30 June)
  - Re: use case submission deadline reminder (Monday, 30 June)
  - Re: Revised Mission Statement (Monday, 30 June)
  - Re: Choreography State Definition (was: RE: More requirement) (Monday, 30 June)
Online Processing

Examples of semantic-block-based co-occurrence relation

- Friday, 20 June
- Why event calculus might be the right model
- Jean-Jacques Dubray
- RE: BPSS_TF_june03.ppt Jean-Jacques Dubray

- Thursday, 19 June
- RE: BPSS_TF_june03.ppt Joo Cordes
- Re: FW: BPSS_TF_june03.ppt Anders W. Telg
- Re: FW: BPSS_TF_june03.ppt Monica J. Martin
- RE: FW: BPSS_TF_june03.ppt Martin Chapman
- Re: FW: BPSS_TF_june03.ppt Anders W. Telg
- Choreography Specification Draft - overview presentation Burdett, David
- Signing off for today Asset Arkin
- FW: BPSS_TF_june03.ppt Martin Chapman
- RE: Revised Mission Statement Jean-Jacques Dubray
- RE: Draft Choreography Spec Burdett, David

- Monday, 16 June
- Re: quote use case Len_Greski@grainger.com

- Wednesday, 16 June
- RE: Revised Mission Statement Thomas Pimpe
- RE: Revised Mission Statement Yaron Y. Golano
- Presentation on Event Calculus Frank McCabe
- RE: Draft Choreography Spec Yaron Y. Golano
- Presentation on Event Calculus Frank McCabe
- RE: More requirement Jean-Jacques Dubray

- Tuesday, 17 June
- RE: Draft Choreography Spec Burdett, David
- RE: Revised Mission Statement Bruno, Richard
- Re: Revised Mission Statement Asset Arkin
- RE: More requirement Yaron Y. Golano
- RE: Draft Choreography Spec Yaron Y. Golano
- RE: Revised Mission Statement Yaron Y. Golano
Basic Model

\[
S_{base}(q,e,r,d) = \mu \frac{freq(e,d,r)}{L(d,r)} + (1 - \mu) \sum_{d' \in d} \frac{freq(e,d',r)}{L(d',r)} / df_e
\]

\[
\mu = \frac{L(d,r)}{L(d,r) + K}
\]
Evidence evaluation

- Relation-type quality
  \[ Q_r = W_r \cdot D(r, q, e) \]
  \[ D(r, q, e) = \begin{cases} 
  \frac{100}{\text{dis}(e, q) + 1} & \text{if } r = \text{window constrained relation} \\
  1 & \text{else}
\end{cases} \]

- Expert-matching quality
  - An expert candidate can occur in the documents in various ways. The most confident occurrence should be the ones in full name or email address.
Evidence evaluation

- Query-matching quality

\[ Q_q = \begin{cases} 
W(t_q) \frac{\text{MAX}_{PQ(q) = PQ(q)}(df_q)}{df_q} & \text{if } t_q = \text{bi-gram matching} \\
W(t_q) & \text{else}
\end{cases} \]

- Context quality

\[ Q_d = Q_{PR}(d)Q_{CP}(d)Q_{DY}(q,d) \]
Quality Based Model for Expert Search

\[ S_{\text{quality}} (q, e, r, d) = Q_q Q_e Q_r Q_d S_{\text{base}} (q, e, r, d) \]

- Query-matching quality
- Expert-matching quality
- Relation extraction quality
- Document context quality
Evidence Merging

- Simple Merging method

\[ S_{\text{simple}}(e) = \sum_{q} \sum_{r} \sum_{d} S_{\text{quality}}(q, e, r, d) \]

- Smoothed Merging method

\[ S_{\text{smoothed}}(e) = \lambda S_{\ln}(e) + (1 - \lambda) \sum_{e' \in C} S_{\ln}(e')/|C| \]

\[ S_{\ln}(e) = \ln(S_{\text{simple}}(e)) \]
Outline

- Overview
- Our approach
  - Offline processing
    - Personal name identification
    - Metadata extraction
  - Online processing
    - Evidence extraction
    - Evidence evaluation
    - Evidence merging
- Submitted runs
- Summary
## Submitted runs

<table>
<thead>
<tr>
<th>Runs</th>
<th>Query Processing</th>
<th>Evidence Extraction</th>
<th>Quality Evaluation</th>
<th>Merging Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>+ Phrase query + Bi-gram query + Proximity query + Fuzzy query + Stemming query</td>
<td>+ Window-based + Section-based + Title-Author + Semantic-block + Improved-block</td>
<td>+ Relation-type quality + Query-matching quality + Expert-matching quality + Context quality (Similarity)</td>
<td></td>
</tr>
<tr>
<td>SJTU01</td>
<td>Same as Base</td>
<td>Same as Base</td>
<td>+ Context quality (PageRank)</td>
<td>Simple Merging</td>
</tr>
<tr>
<td>SJTU02</td>
<td>Same as Base</td>
<td>Same as Base</td>
<td>+ Context quality (PageRank) + Context quality (Corpus weighting)</td>
<td>Simple Merging</td>
</tr>
<tr>
<td>SJTU03</td>
<td>Same as Base</td>
<td>Same as Base</td>
<td>+ Context quality (Corpus weighting)</td>
<td>Simple Merging</td>
</tr>
<tr>
<td>SJTU04</td>
<td>Same as Base</td>
<td>Same as Base</td>
<td>+ Context quality (PageRank) + Context quality (Corpus weighting)</td>
<td>Smoothed Merging</td>
</tr>
<tr>
<td></td>
<td>MAP</td>
<td>Bpref</td>
<td>Prec@10</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>SJTU01</td>
<td>0.5829</td>
<td>0.5835</td>
<td>0.6878</td>
<td></td>
</tr>
<tr>
<td>SJTU02</td>
<td>0.5860</td>
<td>0.5863</td>
<td>0.6918</td>
<td></td>
</tr>
<tr>
<td>SJTU03</td>
<td>0.5851</td>
<td>0.5849</td>
<td>0.6776</td>
<td></td>
</tr>
<tr>
<td>SJTU04</td>
<td>0.5947</td>
<td>0.5913</td>
<td>0.7041</td>
<td></td>
</tr>
</tbody>
</table>
Summary

- Proposed an evidence oriented framework for expert search
- Studied different techniques for evidence extraction
- Evaluated various evidence qualities
- Compared two evidence merging methods
Thanks!