People Search in the Enterprise

Krisztian Balog
Information and Language Processing Systems
University of Amsterdam

Outline

One sentence summary:
Finding and profiling people within an organizational setting

- Setting the scene
- Research questions
- Tasks
- Modeling
- Evaluation
- Wrap-up

Setting the scene

Locating classmates and old friends

Finding partners for dates, romance, and long-term relationships
White and yellow pages
(name, address, and phone number database)

Background check
(address history, property reports, criminal records, etc.)

Setting the scene

My interests

Research Questions

Motivation

- Professional or work-related people search applications
- Example scenarios
  - A personnel officer wants to find information about a person who applied for a specific position.
  - A company requires the state-of-the-art in some field, therefore they want to contact with someone from a knowledge institute.
  - An enterprise needs to set up a task force to accomplish some objective.

- Knowledge Management
  - Getting to an organization’s knowledge
  - Managing the expertise of employees
    - Especially in large or geographically dispersed enterprises

- Information Retrieval, Information Extraction
  - From retrieving documents to retrieving objects/entities
    - Books, CDs, places, answers, people
  - Challenges: representation, modeling, evaluation
Research questions

- **General**
  - What people search tasks should be addressed?
  - How to model these tasks?
  - How to evaluate the proposed solutions?

- **Specific**
  - How to represent topics, documents, candidates?
  - How to represent and make use of structure? (document, topic, people)
  - How to build document-person associations?
  - …

Tasks

- **Two main tasks, in two flavors**
  - **Expert finding**
    - What are experts on the topic "semantic web"?
    - Who are collaborating with X, Y, and Z? Who does person X work with?
  - **Expert profiling**
    - What are the expertise areas of person X?

- **"Other"**
  - Mining contact details
  - Disambiguating person names
  - Finding similar experts

Topical search tasks

- **Probabilistic retrieval framework**
  - **Expert finding**
    - \( p(c|a) \) — What is the probability of a candidate \( c \) being an expert given the query topic \( a \)?
  - **Expert profiling**
    - \( p(a|c) \) — What is the probability of a knowledge area (topic) being part of the candidate’s (expertise) profile?

\[
p(a|c) = \frac{p(q|a)p(a)}{p(q)} = p(c|a) \times [p(q|a)p(a)]
\]
Ingredients

- Problem: how to estimate $p(q|c,a)$?
- Approach: association finding between topics and people
- Candidate-based approaches
  - Create a textual model of candidates’ knowledge according to the documents with which they are associated
  - Model 1: query likelihood
  - Model 3: KL-divergence of candidate and topic models
- Document-based approach
  - Model 2: Find out who is most strongly associated with the documents that best describe the topic

Approach

- Problem: how to estimate $p(q|c,a)$?
- Approach: association finding between topics and people
- Candidate-based approaches
  - Create a textual model of candidates’ knowledge according to the documents with which they are associated
  - Model 1: query likelihood
  - Model 3: KL-divergence of candidate and topic models
- Document-based approach
  - Model 2: Find out who is most strongly associated with the documents that best describe the topic

Candidate-based approaches

- Collect all term information from all documents associated with given candidate
- Smooth it with a background model
- Use this to represent candidate

\[
p(q|\theta_a) = \left(1 - \lambda\right) \cdot p(q|\alpha) + \lambda \cdot p(q)
\]

\[
p(q|\theta_a) = \sum_d p(q|d) \cdot p(d|\alpha)
\]

Document-based approach

- Find documents relevant to the query
- Find out who is most strongly associated with the relevant documents

\[
p(q|\alpha) = \sum_d p(q|d) \cdot p(d|\alpha)
\]

\[
p(q|\theta_a) = \prod_d p(q|d) \cdot p(d|\alpha)\]

\[
p(q|\theta_a) = (1 - \lambda) \cdot p(q|d) + \lambda \cdot p(q)
\]
Evaluation

Test collections

- W3C collection
  - TREC Enterprise Track (2005, 2006)
- UvT Expert Collection
  - Balog et al. SIGIR 2007
- CSIRO collection
  - TREC Enterprise Track (2007)

Test collections

<table>
<thead>
<tr>
<th></th>
<th>W3C</th>
<th>UvT</th>
</tr>
</thead>
<tbody>
<tr>
<td>documents</td>
<td>Web, Bros, wikis ~330,000</td>
<td>Research-, course data, publications, homepages ~26,000</td>
</tr>
<tr>
<td>language</td>
<td>English</td>
<td>English/Dutch</td>
</tr>
<tr>
<td>candidates</td>
<td>102K</td>
<td>11K</td>
</tr>
<tr>
<td># topics</td>
<td>2427</td>
<td>2066</td>
</tr>
<tr>
<td>assessment</td>
<td>59/106/50/05</td>
<td>INL, TREC, ENT, EMN</td>
</tr>
<tr>
<td>organizational structure</td>
<td>www, lists, wikis ~38.000</td>
<td>www, lists, wikis ~330.000</td>
</tr>
</tbody>
</table>

Findings

- Document model performs best for both tasks in nearly all settings
- Quality of document-candidate associations matters a lot
- State-of-the-art performance
  - Best results so far would be in top 3 for TRECEnt’05 and in top 5 for TRECEnt’06 — no heuristics or collection specific features!
- Models generalize well

Advanced models

- Exploiting knowledge area similarity
  - Use similar queries to support the original query
- Contextual information
  - Use organizational units of which the candidate is member of
- Multilingual model
  - Combine results across languages
- Below the document level

Wrap up
My contribution

- Modeling, algorithms
  - Probabilistic framework for two topical expertise finding tasks (finding, profiling)
  - Numerous models implementing different strategies (candidate- and document-based approaches)
  - Incorporating context and structure
- Evaluation
  - State-of-the-art performance on W3C data set
  - Introduction of the UvT Expert Collection

Stepping back

- Tasks are modeled as an association finding problem
  - Models are general; finding associations between topics and entities
- Is it really expertise finding?!
  - Is it “just” association finding? What aspects of real-world expertise finding am I missing (if any)?

Looking forward / Open issues

- Relate work to Knowledge Management
- Document/candidate priors
- Social search
  - What are the suitable workplace/enterprise tasks?
  - Formulate tasks in a probabilistic framework. E.g., $p(c_1|c_2), p(c_1,c_2|q)$
- Dynamic aspects
  - How does expertise change over time?
- Looking at not only individuals but groups of individuals

Questions?
www.science.uva.nl/~kbalog