Finding People and Documents, Using Web 2.0 Data

Nadav Har'El
Einat Amitay
David Carmel
Nadav Golbandi
Shila Ofek-Koifman
Sivan Yogev

IBM Haifa Research Lab
Web 2.0 data

• The traditional Web:
  – Considerable effort to publish content.
  – Most users are information consumers only.

• Web 2.0:
  – Ordinary users easily produce information.
  – Services such as forums, wikis, blogs, collaborative, bookmarking, etc.
Web 2.0 data

- Web 2.0 data gives us
  - New wealth of information (produced by ordinary users)
  - New types of information – **social information**:
    - User-supplied metadata for documents (bookmarks, tags, ratings, comments)
    - **Relationships between people and documents** (who wrote a document, who tagged it, etc.)
    - Relationships between people and people.
Social search

• Our goal: use social information to improve search in an enterprise intranet (IBM).
  - Improve the relevance of document results:
    • Tags and comments supply more text to be searched.
    • Important documents can be recognized by user activity around them (bookmarking, comments, etc.)
    • Our research shows precision is vastly improved over standard full-text search (P@10 between 0.7-0.8).
  - How use person-document relationships?
Outline of this talk

• Unified search: document & person.
• How the document-person relationships enable person search.
• Implementation of the unified search using faceted search.
• The system and its evaluation.
Unified search

• When in need of information,
  – Some people like to find a written document.
  – Some people like to find a person to ask.
  – Most people are between these extremes.
  – And each source is better in different situations.
Unified search

• So given a query, we want the search engine to return:
  – A ranked list of documents relevant to the query
  – A ranked list of people interested in the query topic

• We also want to use people in queries:
  – “John Smith”
  – information retrieval “John Smith”
Person search

- Using person-documents relationship:
- A person is relevant to a query if he or she are related to documents relevant to the query.
  - Given a query
  - Find all documents relevant to this query
  - Find people relevant to these documents
- [McDonald & Ounis, Balog & de Rijke, 2006]
- But how to score?
Person search

- Returning to the Vector Space Model:
  - In VSM, documents define relevance matrix $D$, between documents and terms.
  - A query is also a vector $q$. Search results: $Dq$.
  - Document-person relationships define relevance matrix $P$ between documents and people.
  - $P^TD$ is a relevance matrix between terms and people. $P^TDq$ are (scored) people search results.
Person search

- But using $P^T D q$ directly is inconvenient:
  - Keeping $P^T D$ up-to-date is hard
  - Document and person search done using two different matrices ($D$ and $P^T D$)
  - Lose non-VSM search engine features (phrase, etc)
- We prove that the following more-useful formula is equivalent:
Person search

- Score for person i, \((P^T D q)_i\) =

\[
\sum_{\text{matching documents } d} \text{relation}(d, \text{person } i) \cdot \text{score}_q(d)
\]

- Already proposed in Balog & de Rijke, with different (probabilistic) justification.

- Can be calculated using \textbf{faceted search}: 
Faceted search

- Commonly used technique for adding navigation to a search engine.
- A **facet** is a single attribute of the document.
- In a camera search application, documents might have a “Brand” and “Price” facets.
- To each document, several **categories** are added. For example “Brand/索尼” or “Price Range/$90-$40”.

Faceted search

- Simplest faceted search goes over matching documents, counting for each category the number of documents:

<table>
<thead>
<tr>
<th>Price Range</th>
<th>Brand</th>
<th>LCD Display Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below $90 (116)</td>
<td>Canon (170)</td>
<td>Less than 1.5 in. (62)</td>
</tr>
<tr>
<td>$90 - $140 (106)</td>
<td>Olympus (214)</td>
<td>1.5 - 2.0 in. (1,262)</td>
</tr>
<tr>
<td>$140 - $170 (96)</td>
<td>Nikon (158)</td>
<td>2.0 - 2.4 in. (390)</td>
</tr>
<tr>
<td>$170 - $210 (105)</td>
<td>Sony (169)</td>
<td>More than 2.4 in. (754)</td>
</tr>
<tr>
<td>$210 - $260 (112)</td>
<td>Panasonic (104)</td>
<td>Select more than one</td>
</tr>
<tr>
<td>$260 - $350 (117)</td>
<td>Kodak (164)</td>
<td></td>
</tr>
<tr>
<td>$350 - $650 (112)</td>
<td>Fuji (161)</td>
<td></td>
</tr>
<tr>
<td>Above $650 (112)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Faceted search

- In our application, a “Related Person” facet.
- Categories like “Related Person/John Smith” attached to document, with a weight.
- Instead of just counting, can aggregate expressions. For person i category:

\[ \sum_{\text{matching documents } d} \text{relation}(d, \text{person } i) \cdot \text{score}_q(d) \]
Faceted search

• More faceted search features we use:
  – Query-independent static score for categories (category boost).

\[
\text{ief}(person) = \log\left(\frac{N}{N_{person}}\right)
\]

– Special query for “Person P” returns all documents in this category, sorted by the category weight.
The Social Search Application

- Data from some of IBM's internal Web 2.0 sites:
  - 67,564 blog threads (thread = entry + comments)
    - Content: Blog entry, comments, tags
    - Person facet: author, commenter, bookmarker
  - 337,345 bookmarks to 214,633 Web-pages
    - Content: Titles, user descriptions, tags
    - Person facet: bookmarker
  - 15,779 people who created that content
The Social Search Application

Search took 0.05 seconds.
Found 108 results. Showing results 1-10:

- **Welcome to OpenID Enabled! — OpenID Enabled**
  - Resource for OpenID community
  - Bookmarked 2 times

- **Sam Ruby: OpenID for non-SuperUsers**
  - A well written article for novice users to get started with OpenID
  - Bookmarked 4 times
  - [http://intertwingly.net/blog/2007/01/03/OpenID-for-non-SuperUsers](http://intertwingly.net/blog/2007/01/03/OpenID-for-non-SuperUsers)

- **OpenID: an actually distributed identity system**
  - The first piece of the OpenID framework is authentication -- how you prove ownership of a URI. Today, websites require usernames and passwords to login, which means that many people use the same password everywhere.
  - Bookmarked 24 times
  - [http://openid.net/](http://openid.net/)

- **OpenID Authentication 1.1**
  - OpenID Authentication 1.1
  - Bookmarked 1 time
  - [http://openid.net/specs/openid-authentication-1_1.html](http://openid.net/specs/openid-authentication-1_1.html)

- **Microsoft and Google want to support OpenID - The Good, The Bad and The Ugly - BlogCentral**
  - Blog entry by [**@de.ibm.com**](mailto:your@email.ibm.com), with 1 comments
  - Bookmarked 1 time
  - [http://blogs.tap.ibm.com/webblogs/your@email.ibm.com/entry/microsoft_and_google_want_to](http://blogs.tap.ibm.com/webblogs/your@email.ibm.com/entry/microsoft_and_google_want_to)
Evaluation

- We return both documents and people for every query – need to evaluate precision of both.

- Document results evaluated as usual:
  - 50 real queries chosen from query logs
  - The top results judged by humans as being “relevant”, “very relevant” or “irrelevant”.
  - Very high precision demonstrated (P@10 ~ 0.8).
  - Much better than full-text enterprise search.
Evaluation

- “Related people” evaluation – large user study
  - 60 real queries chosen from query logs.
  - 100 related people retrieved for each query.
  - Each person is mailed listing 6-15 queries (some believed to be relevant and some irrelevant): Rate 1-5 whether the topic is relevant to you.
  - 612 people responded, from 116 IBM locations in 38 countries.
  - The ranked list of related people we generate are compared to these self-ratings using NDCG metric.
  - Compare full scoring formula to simpler ones.
## Evaluation

### Evaluation results:

<table>
<thead>
<tr>
<th>Aggregation expression</th>
<th>NDC G 10</th>
<th>NDC G 20</th>
<th>NDC G 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count only “votes”</td>
<td>0.71</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td>Sum of scores “CombSUM”</td>
<td>0.75</td>
<td>0.73</td>
<td>0.72</td>
</tr>
<tr>
<td>+relationship weights</td>
<td>0.76</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>+person static score: ief</td>
<td>0.77</td>
<td>0.76</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Conclusions

- Web 2.0 data provides an excellent source for document and people search in an enterprise.
- Unified (document/person) search can be easily realized using faceted search.
- VSM justification for the scoring formula.
- In a 612-respondent study, the full scoring formula was shown better than simpler versions.
- Also strengthens previously published results by using with a new data set and evaluation.