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**.
  - $\mathbf{P^T D}$ is a relevance matrix between **terms** and **people**. $\mathbf{P^T Dq}$ are (scored) people search results.

Person search

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

Person search

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

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

- Already proposed in Balog & de Rijke, with different (probabilistic) justification.
- Can be calculated using **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/Sony” or “Price Range/$90-$40”.

Faceted search

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

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:

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

Faceted search

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

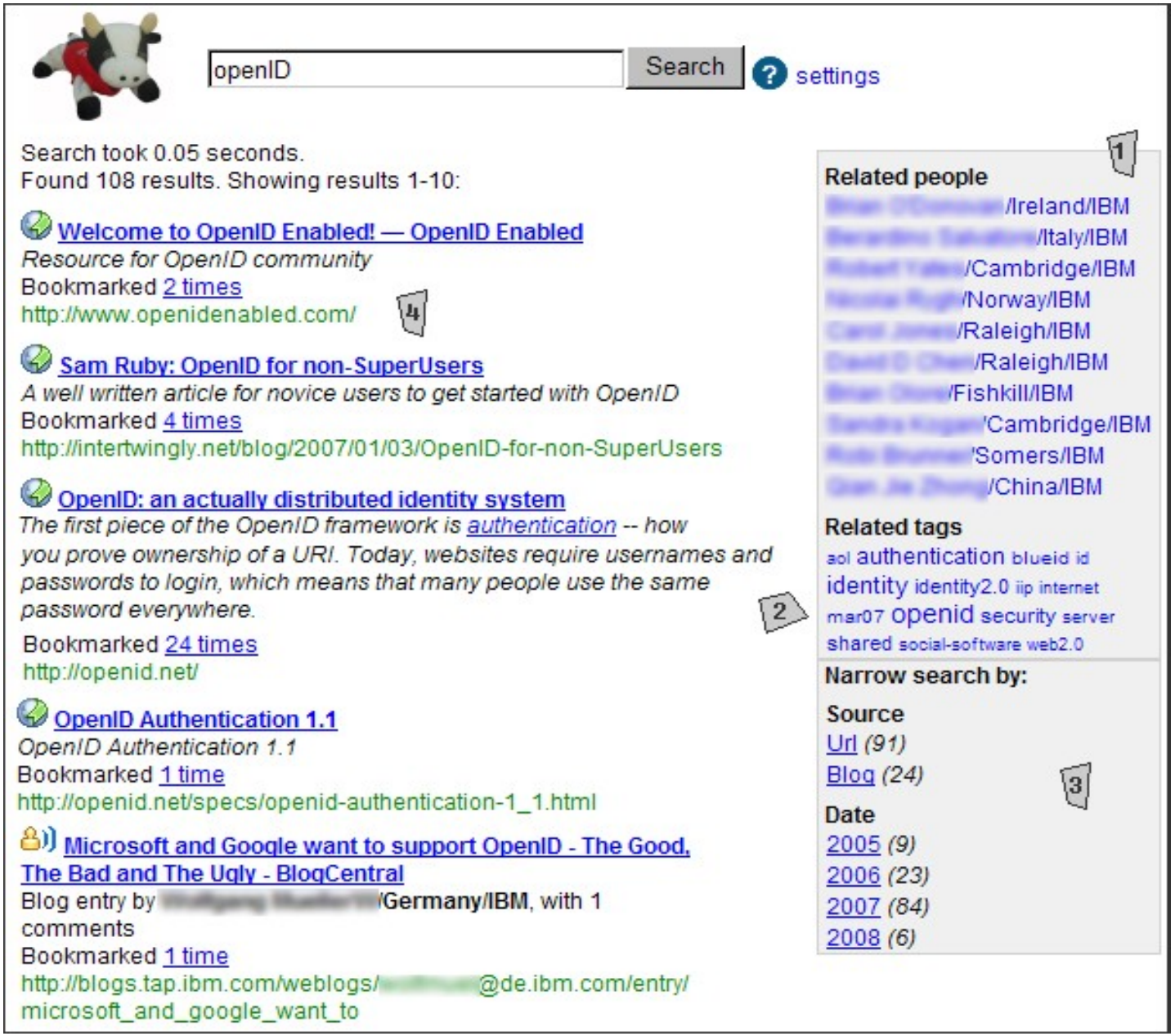
$$\text{ief}(person) = \log(\frac{N}{N_{person}})$$

  - 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



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:

Aggregation expression	**NDCG 10**	**NDCG 20**	**NDCG 30**
Count only "votes"	0.71	0.69	0.68
Sum of scores "CombSUM"	0.75	0.73	0.72
+relationship weights	0.76	0.74	0.73
+person static score: ief	0.77	0.76	0.74

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.