Social Search and Discovery Using a Unified Approach

Einat Amitay, David Carmel, Nadav Har'El, Shila Ofek-Koifman,
Aya Soffer, Sivan Yogev, Nadav Golbandi
IBM Research Lab in Haifa
Haifa 31905, Israel
{einat,carmel,nyh,shila,ayas,sivany}@il.ibm.com, nadav.golbandi@gmail.com

ABSTRACT

We explore new ways of improving a search engine using data from Web 2.0 applications such as blogs and social bookmarks. This data contains entities such as documents, people and tags, and relationships between them. We propose a simple yet effective method, based on *faceted search*, that treats all entities in a *unified* manner: returning all of them (documents, people and tags) on every search, and allowing all of them to be used as search terms. We describe an implementation of such a *social search engine* on the intranet of a large enterprise, and present large-scale experiments which verify the validity of our approach.

Categories and Subject Descriptors: H.3.3 [Information storage and retrieval]: Information search and retrieval

General Terms: Algorithms

Keywords: social search, faceted search, enterprise search

1. INTRODUCTION

Traditionally, building a Web site took considerable effort and expertise, so most Web users only consumed information, not produced it. In recent years, however, the tide has turned to letting ordinary users produce information and publish it to their peers. Services such as blogs, forums, wikis, collaborative bookmarking, and others have become common, and collectively called *Web 2.0* applications.

The *social information* in these sources goes beyond just text documents. There are two additional important *entities*, people and tags, and relationships between the three types of entities. For example, a person might be related to a document as a writer, commenter, or a tagger. Additionally, users add metadata to documents, by bookmarking them, commenting on them, rating them, and so on.

The goal of a *social search engine*, is to use the social information to improve the user's search experience over regular full-text search. One way of improving search is to improve the relevance of document results [1, 5, 4]: tags and comments supply more text that can be considered during search, and important documents can be recognized by the amount of user activity around them (such as the number of people who bookmarked them or commented on them).

But using the social information, we want to do more than just return better documents. We would like to treat all three entity types, *documents*, *people* and *tags*, in a *unified* manner.Related people and tags, not just documents, will be returned for every query. People and tags can also be used as query terms.

This *unified search* enables the searcher to get a wider view on the topic, to look deeper into interesting people that are related to the topic and to understand the topic better as described by the tag cloud. The initial search results can be viewed as an entry point to a topic, enabling the user to go deeper according to his interests, and explore a web of entities and relationships between those entities, around his topic of interest.

Our unified search solution uses the enhanced faceted search engine [2]. Basically, a person is highly related to the query if he is highly related to many documents which are each highly relevant to the query. Faceted search first scores the relevant documents, and then sums each document's contribution to its related entities (this contribution is the multiplication of the strength of the document-entity relationship, with the relevance of the document to the query).

In this work, we present a social search engine we developed for *enterprise search* (search in the intranet of an organization), and deployed in IBM. Social search is particularly suited to enterprise search: It aims to increase the notoriously low precision and recall of traditional IR techniques on enterprise data; The user-supplied social information can be trusted; Person identities in the enterprise are more meaningful and easier to correlate across different applications.

We will also present extensive evaluations to show that the social information significantly improves search precision and that the related people and tags found on each search (in addition to the relevant documents) are indeed relevant.

2. THE SOCIAL SEARCH APPLICATION

We implemented a social search engine based on social information from IBM's intranet. From the internal Web 2.0 services in IBM, we chose the currently most used ones: Dogear, a collaborative bookmarking service used to bookmark and tag pages both within and outside the intranet; and BlogCentral, a blog service allowing all IBM employees to manage blogs in the intranet. We collected data from 15,779 employees assigning 337,345 bookmarks to 214,633 pages and writing 67,564 blog threads.

The content we indexed for a bookmarked page contained its title and the users' descriptions and tags as provided by Dogear (the actual content of the page was not available). For blogs, the indexed document was a blog thread, containing the blog entry, comments, and tags. For each person, we had a document containing the person's directory informa-

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

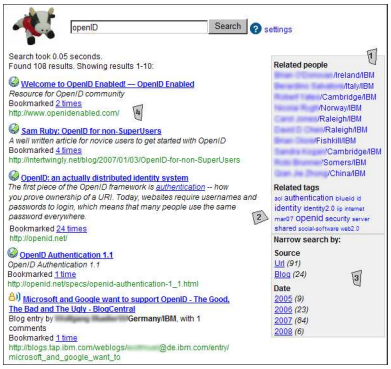

Figure 1: The social search application.

tion (such as name, title, and group). Finally, people were connected, as facets, to the pages they bookmarked, and to their blog entries (as an author or as a commenter). Tags were connected as facets to their related documents.

A static-score (boost) was given to each document based on the amount of activity around it. In essence, a page which was bookmarked by many people, or a blog entry that was heavily commented or rated, is more likely to be a good search result than a document in which hardly anyone expressed interest. Our evaluation showed that this boosting significantly increased the document search precision.

The social search Web application, codenamed Cow Search, was made available to all users of IBM's intranet. As explained above, the application enables searching with any of the supported entity types (terms, people, tags), or a mixture thereof, as queries. Using the faceted search library, three ranked result lists are generated: documents (bookmarked pages, blog threads, and directory entries), related people, and related tags. Tags are shown in a tag cloud (a tag's size is determined by its score) and people are shown as a ranked list.

Figure 1 shows a screenshot of the application, given the query "openID". On the left (marked by 4) we see the most relevant documents — a mix of blogs, Web pages and personal profiles. On the right (1) is the list of related people. The related-tags cloud is shown in (2). (3) shows some additional facets, which aid navigation within the search results.

3. EVALUATION

A significant part of this research was evaluating the effectiveness of social search in the enterprise. Our unified search system returns three result lists for each query: documents, people and tags. Therefore, to evaluate the quality of our system we needed to evaluate the quality of all three lists.

The quality of document results was evaluated using the standard IR evaluation methodology: We picked 50 real users' queries, and ran them on our search engine and IBM's default one. We took the top 30 documents returned for each search, and asked human evaluators to examine each result and judge its relevance to the query with three relevance levels (not relevant, marginally relevant or highly relevant), without knowing where the result came from. These judgments were then compared against our system's result set using the NDCG (Normalized Discounted Cumulative Gain) and and p@k (precision at top k) measures for each query, and finally averaged over all 50 queries.

The precision scores for our systems were NDCG(15)=0.70, p@1=0.76, p@5=0.70, p@10=0.67. These are very high levels of precision, especially when compared to the state-of-the-art enterprise search engine deployed in IBM, whose precision we measured to be much lower (NDCG(15)=0.48, p@1=0.44, p@5=0.4, p@10=0.38).

Our second evaluation, that of the quality of the related people results, could not be done using the same standard technique. This is because the judges found it virtually impossible to decide whether a person whom they didn't know was relevant to a given query. Rather, we had to ask the person himself whether he considers himself relevant: For each query we took the top 100 people. To each person related to any of the queries we sent an email with a list of the queries to which the system believed him related, mixed with other queries selected randomly. We asked each person to rate on a Likert scale of 1 to 5 how relevant they think each query is to them. 612 people from 116 IBM locations replied to our request with their rating. We measured the NDCG of the related user list using the 1-5 scale of user feedback as different relevance levels. The results show a high level of agreement between the retrieved people and the system's ranking: average NDCG(10)=0.77, NDCG(30)=0.74.

Finally, our third evaluation measured whether tags retrieved by the social search system are indeed related to the queries. The evaluation was done using Normalized Google Distance [3], which measures the similarity of meaning between terms, based on the number of hits returned by a search engine. For each query we found the list of related tags returned by our system. For each query-tag pair, we performed three queries in IBM's enterprise search engine: a) the query alone, b) the tag alone and c) the query and tag together. The number of results for each of the queries along with the total number of documents in the search index defines the NGD between the query and the tag. NGD scores are in the range [0,1], which we translated to discrete relevance levels for the purpose of NDCG calculations. The NDCG of the top k tags shows high agreement between the relevance level of retrieved tags with the system's ranking: NDCG(10)=0.681, NDCG(30)=0.745, NDCG(50)=0.777.

To summarize, this work describes a social search solution using a unified approach in which all system entities are searchable and retrievable. Our results reveal that related users and tags seems to be very valuable, as complementary facets to regular search results. Moreover, user comments and tags gathered from Web 2.0 applications are highly valuable in identifying high quality content in the corpus.

4. REFERENCES

[1] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW*, pages 501–510, 2007.

[2] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev. Beyond basic faceted search. In *WSDM*, 2008.

[3] R. Cilibrasi and P. M. B. Vitányi. Automatic meaning discovery using google. In *Kolmogorov Complexity and Applications*, 2006.

[4] P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita. Using annotations in enterprise search. In *WWW*, pages 811–817, 2006.

[5] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarks improve web search. In *WSDM*, 2008.