Making SENSE: Socially ENhanced Search and Exploration

Tom Crecelius[†], Mouna Kacimi[†], Sebastian Michel[‡], Thomas Neumann[†],
Josiane Xavier Parreira[†], Ralf Schenkel[†], Gerhard Weikum[†]

[†]Max-Planck-Institut für Informatik
Saarbrücken, Germany

[‡]École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

socialsearch@mpi-inf.mpg.de

ABSTRACT

Online communities like Flickr, del.icio.us and YouTube have established themselves as very popular and powerful services for publishing and searching contents, but also for identifying other users who share similar interests. In these communities, data are usually annotated with carefully selected and often semantically meaningful tags, collaboratively chosen by the user who uploaded an item and other users who came across the item. Items like urls or videos are typically retrieved by issueing queries that consist of a set of tags, returning items that have been frequently annotated with these tags. However, users often prefer a more personalized way of searching over such a 'global' search, exploiting preferences of and connections between users.

The SENSE system presented in this demo supports hybrid personalization along two dimensions: in the social dimension, a search process is focused towards items tagged by users explicitly selected as friends by the querying user, whereas in the spiritual dimension, users that share preferences with the querying user are preferred. Orthorgonal to this, the system additionally integrates semantic expansion of query tags to improve search results. SENSE provides an efficient top-k algorithm that dynamically expands the search to related users and tags. It is based on principles of threshold algorithms, folding related users and tags into the search space in an incremental on-demand manner, thus visiting only a small fraction of the social network when evaluating a query. The demonstration uses three different real-world datasets: a large set of urls from del.icio.us, a large set of pictures from Flickr, and a large set of books from librarything, each together with a large fraction of the corresponding social network of these sites.

1. INTRODUCTION

With the social revolution in Web 2.0, online communities have established themselves as very popular and powerful services for publishing and searching contents, turning users from mere consumers into information providers. Users can store their personal content, share it with other people, explore other users' contents, and identify other users sharing similar interests. Popular examples of such online communities are YouTube, MySpace, Facebook, del.icio.us, Flickr, LibraryThing and Friendster.

Social tagging has emerged as an important asset to explore the fast-growing communities in order to identify interesting content and users [8, 6, 13, 12]. In these communities, data is usually annotated with carefully selected and often semantically meaningful tags, collaboratively chosen by the user who uploaded an item and other users, or by ratings or comments expressing their opinions about items. To improve the search experience, different kinds of social relations have been explored in the recent literature. For example social relations between tags have been used to enhance searching and ranking in social communities [1, 3, 15]. Similarly, social relations between users have been used for query routing in peer-to-peer networks [4, 9].

Items like urls or videos are typically retrieved by issueing queries that consist of a set of tags, returning items that have been frequently annotated with these tags. However, users often prefer a more personalized way of searching over such a 'global' search, exploiting preferences of and connections between users. Search tasks in such systems can be classified in three different categories: social, spiritual, and global searches, expressing different kinds of information needs. First, *social* searches are targeted towards information from the social context of the user, i.e., information that was contributed by explicit friends (possibly including transitive friends, i.e., friends of friends). They are suitable for getting information from friends which you know and you trust, irrespective of their interests, rather than from unknown users which you never heard of before. Another pattern for social searches is, for instance, finding items about users themselves, like photos, where users are more likely to be pleased when seeing their friends than seeing people they don't know. Second, *spiritual* searches seek information within the user's interests already expressed in the system, which should be directed towards her implicit friends, which are other users with a similar behaviour such as high overlap in tag usage, bookmarked pages, or commenting and rating activity. Due to their behavioral affinity, we call implicit friends of a user her *brothers in spirit* and, thus, the notion of *spritual search*. This personalized search that asks for recommendation-style results is very common in online communities, for example asking for books tagged by other users with similar interests, or searching for restaurants tagged by users from the same area and with similar preferences for food. Last, *global* searches are neither social nor spiritual, so they consider information by all users equally important, disregarding any social relationships or common interests (maybe because the user asks for something that does not match her usual preferences). Examples for global searches are exploiting the "wisdom of the crowds" by asking the best book to give as a present to somebody else (whose preferences don't match those of the user) or searching a text book on Java when you tagged only travel guides before; in

PVLDB '08, August 23-28, 2008, Auckland, New Zealand
Copyright 2008 VLDB Endowment, ACM 978-1-60558-306-8/08/08

that case, a spiritual search may return travel guides on Indonesia instead of books on programming languages. Our model uses an integrated scoring model for global, social and spiritual searches with tunable parameters to steer query evaluation towards one of these types. Note that the notions of authority and trust are generally orthogonal to these search categories.

With the high dynamics of online communities with items and tags being added at very high rates and user profiles changing rapidly, social and spiritual searches cannot rely on standard evaluation algorithms and precomputed scores like other search applications. The SENSE system presented in this demo provides an efficient top-k algorithm for these settings that incrementally explores the space of (explicit or implicit) friends and accumulates scores of items on the fly as they are encountered, limiting the expansion to a minimal number of related users. Orthogonal to this, the system additionally integrates semantic expansion of query tags to improve search results. The algorithm can efficiently compute the best matches to social and spiritual searches and, falling back to a threshold algorithm on precomputed index lists, also for global searches, even in huge communities with high dynamics [11].

The remainder of this paper is structured as follows. In Section 2, we give an overview of the hybrid scoring model used for global, spiritual and social searches. In Section 3, we introduce the architecture of the system and highlight important aspects of the top-k algorithm at its core. In Section 4, we give detailed information on the demonstration itself.

2. DATA AND SCORING MODEL

2.1 Social Network Model

This section introduces our general social network model. As shown in Figure 1, the set of nodes $N = U \cup D \cup T$ in the network represents *users* U, *documents*[1] D and *tags* T. Additionally, social networks exhibit various relationships, both among the nodes of the same type and between nodes of different types. These relationships are represented by edges in the graph.

Three main relations exist between nodes of the same type. First, *friendship* edges model explicit and implicit friends of users. Here, *explicit* friends are those explicitly, manually selected by the user, while *implicit* friends (or *brothers in spirit*) share preferences or interests with the user. The strength of these relationships is reflected by a *friendshipStrength*. Second, *similarity* captures the semantic closeness between tags. It can be computed using different methods based on semantic relationships or tag-usage statistics in the social network. Third, *linkage* represents the relationship between documents. It can be given by hyperlink graph for web pages or simply derived from the similarity between document tags.

Additional three relations exist between nodes of different types. First, *content* connects documents with tags at least one user used on this document. Second, *tagging* associates tags to users who used them at least once. Third, *rating* links users to documents which they annotated with a tag or explicitly rated (but ratings are not further exploited in this paper). Note that tagging actions, i.e., the tags which a specific user assigned to a specific document, are not represented explicitly in this graph, but split into tagging (user-tag) and content (tag-document) edges.

2.2 Integrated Scoring

In our model, we consider a *query* $Q(u, q_1 \dots q_n)$, issued by a query initiator u with a set of tags $q_1 \dots q_n$. Result documents should contain at least one of the query tags and be ranked according to a *score*. In contrast to standard IR query models, our scoring function can be tuned towards the different search processes in social systems. Scores are *user-specific*, i.e., they depend on the social and/or the spiritual context of the query initiator, depending on the configuration of the model. The querying user can decide if her information need is spiritual, social or global (which is the default).

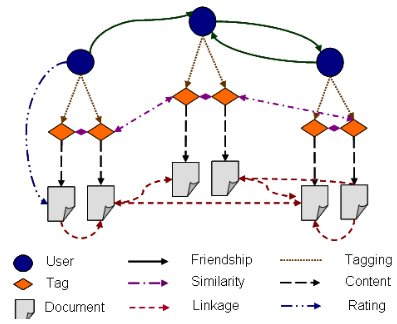


Figure 1: Social Network Model

Modelling Friendship Strengths. The core of our scoring is formed by three different quantizations for user-to-user similarities (also called *friendship strengths*), corresponding to the three different searches in communities. Each similarity can be implemented in different ways, and our current implementation allows to switch between different definitions of the similarities at runtime. The *spiritual* friendship similarity $F_{sp}(u, u')$ of two users u and u', tuned towards spiritual searches, is computed using a combination of syntactic measures such as overlap of tag usage, bookmarked pages, or commenting and rating activity. The *social* friendship similarity $F_{so}(u, u')$, applied for social searches, is based on social measures like the inverse distance of u and u' in the social network graph, but may additionally include syntactic measures (like the spiritual friendship similarity). The *global* friendship similarity $F_{gl}(u, u') = \frac{1}{|U|}$, used for global searches, gives equal weight to all users. All similarities are normalized such that $\sum_{u' \in U} F(u, u') = 1$ for all u.

The actual friendship similarity used to evaluate a query is a linear mixture of these three similarities:

$$F_u(u') = \alpha \cdot F_{sp}(u, u') + \beta \cdot F_{so}(u, u') + (1 - \alpha - \beta)\frac{1}{|U|}$$

The parameters α and β, $0 \leq \alpha, \beta \leq 1$, can be configured by the user (typically by selecting predefined configurations corresponding to spiritual ($\alpha = 1, \beta = 0$), social ($\alpha = 0, \beta = 1$), and global ($\alpha = 0, \beta = 0$) searches; however also nontrivial combinations are reasonable).

Score for Tags. To compute the score $s_u(d, t)$ of a document d with respect to a single tag t relative to the querying user u, we use a scoring function in the form of a simplified BM25 [10] score:

$$s_u(d, t) = \frac{(k_1 + 1) \cdot |U| \cdot sf_u(d, t)}{k_1 + |U| \cdot sf_u(d, t)} \cdot idf(t)$$

where k_1 is a tunable coefficient (just like in standard BM25) and $idf(t)$ is the inverse document frequency of tag t, instantiated as

$$idf(t) = \log \frac{|D| - df(t) + 0.5}{df(t) + 0.5}$$

with $df(t)$ denoting the number of documents that were tagged with t by at least one user. Unlike the original BM25 formula, our model has no notion of document lengths; the number of tags assigned to a document does not vary as much as the length of text documents.

[1] We use the term 'document' which is familiar from scoring models in IR instead of the more general term 'item'.

The social-aware term frequency $sf_u(d,t)$, our replacement for the standard term frequency (tf) known from text IR, weights tags by the friendship similarity of the query initiator and the user who added the tag to the document. More formally, denoting by $tf_u(d,t)$ the number of times user u used tag t for document d, we define the social-aware term frequency $sf_u(d,t)$ for a tag t and a document d, relative to a user u, as

$$sf_u(d,t) = \sum_{u' \in U} F_u(u') \cdot tf_{u'}(d,t).$$

Tag Expansion. Even though related users are likely to have tagged related documents, they may have used different tags to describe them. It is therefore essential to allow for an expansion of query tags to "semantically" related tags. To avoid topic drift problems [5], we adopt in our scoring model the *careful expansion* approach proposed in [14] that considers, for the score of a document, only the best expansion of a query tag, not all of them. More formally, we introduce the *tag similarity* $tsim(t_1,t_2)$ for a pair of tags t_1 and t_2, $0 \leq tsim(t_1,t_2) \leq 1$. The final score $s_u^*(d,t)$ of a document d with respect to a tag t and relative to a querying user u, considering tag expansion, is then defined as

$$s_u^*(d,t) = \max_{t' \in T} tsim(t,t') \cdot s_u(d,t')$$

Our current implementation provides several alternatives to compute the similarity between two tags: In addition to *SocialSimRank* from [3], we exploit the co-occurrence of the tags in the entire document collection by estimating conditional probabilities:

$$tsim(t,t') = P[t|t'] = \frac{df(t)}{df(t \wedge t')}$$

where $df(t \wedge t')$ is the number of documents that have been tagged by both tags (but possibly by different users).

Score for Queries. Finally, the score for an entire query with multiple tags $q_1 \ldots q_n$ is the sum of the per-tag scores:

$$s_u^*(d, q_1 \ldots q_n) = \sum_{q_1 \cdots q_n} s_u^*(d, q_i)$$

3. SYSTEM ARCHITECTURE AND ALGORITHMS

Figure 2 shows an overview of the architecture of SENSE. Data from social communities are imported and precomputed into database-backed data structures used for query evaluation. Given a query which was entered by a user through a Tomcat servlet, the top-k-aware query processor uses this information to compute the best results for a query, which are again returned through Tomcat.

Data Structures. Like other systems which apply Threshold algorithms for top-k processing, SENSE uses precomputed index lists. However, unlike usual applications (for example for text retrieval) which store precomputed scores in these lists, we cannot do that as our scores depend on the querying user, her social context and the configuration of the scoring function, so a precomputation would be inflexible and way too huge for reasonably sized communities. Instead, we maintain the following lists (kept in a database with indexes for efficient access):

For each tag t, we maintain a list DOCS(t), containing documents d tagged by at least one user and corresponding global tag frequencies $TF(d,t)$ (i.e., the number of users which tagged d with t), odered by descending $TF(d,t)$. For each user u and each tag t she used, we maintain a list USERDOCS(u,t) with the unsorted set of documents d tagged with t by u. Note that such data structures exist in real social systems anyway and are therefore no additional overhead for SENSE.

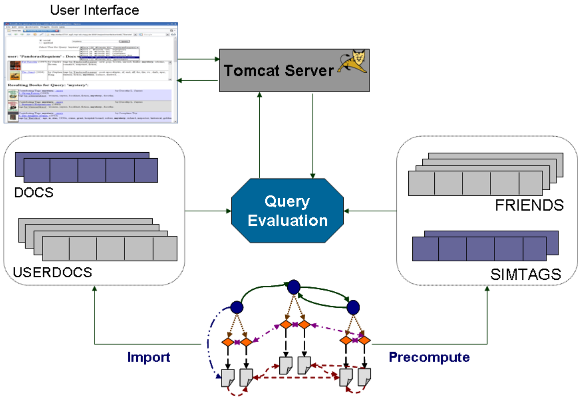


Figure 2: Architecture of SENSE

Precomputed friendships strengths between users are stored in FRIENDS_SP(u) and FRIENDS_SO(u) for spiritual and social friendships, respectively, which contain, for a user u, all related users u' and their similarity in descending order. Finally, SIMTAGS(t) contains for a tag t all similar tags t' with their similarity in descending order. These lists are precomputed and updated regularly. There may be different instances of these lists, corresponding to different implementations of the strengths and similarities.

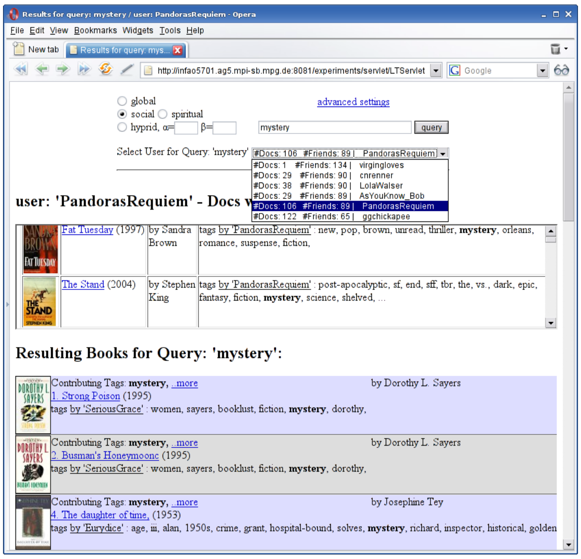
Algorithm. SENSE implements the *ContextMerge* algorithm, an efficient threshold algorithm for evaluating the top-k results for a query in social networks [11]. As the score of a document depends on the user who initiates the query, standard top-k algorithms relying on precomputed per-tag scores for each document [7, 2] cannot be applied here. Instead *ContextMerge* incrementally builds social-aware term frequencies by considering users that are related to the querying user in descending order of friendship similarity, computes upper and lower bounds for the social score from these frequencies, and stops the execution as soon as it can be guaranteed that the best k documents have been identified.

To compute the top-k results for a query $q_1 \ldots q_n$ submitted by a user u, *ContextMerge* sequentially scans, for all query tags, the DOCS lists and the USERDOCS lists of the friends of u in an interleaved way, maintains a list of candidate documents seen during the scans and a list of current top-k candidates, and terminates as soon as none of the candidates can move to the top-k. To improve efficiency, *ContextMerge* can additionally perform random accesses to the index lists to lookup the values for selected documents. Note that the algorithm can be further optimized if the query is purely social or spiritual (i.e., $\alpha + \beta = 1$), thus, no DOCS lists need to be opened as the execution can be limited to the context of u. By contrast, to process global queries ($\alpha = \beta = 0$) there is no need to consider any lists of friends, so just the DOCS lists are read and *ContextMerge* behaves like a standard top-k algorithm. However, the interesting case is for mixed search where document scores are computed using both global and social or spiritual components.

Tag expansion adds another dimension that *ContextMerge* needs to combine with the user-expansion dimension. However, it would be very inefficient to directly include the lists of all similar tags in the processing. Instead, *ContextMerge* incrementally adds lists for similar tags to the processing on the fly in the style of [14].

4. DEMO DESCRIPTION

The demo presents a full implementation of SENSE, a hybrid personalized search system where the user can perform spiritual, social, or global searches or hybrid combinations of them in online communities. We showcase three different social communities at the demo: a subset of Flickr with 10 million pictures and more than 50,000 users, a subset of librarything.com with more than 6 million books and about 10,000 users, and a subset of del.icio.us with more than 400,000 bookmarks and about 10,000 users. A visitor of the demo can first provide a query consisting of one or multiple tags using the interface shown in Figure 3. SENSE generates a list of candidates to be the initiators of the query, based on the overlap of the tags they used with the query tags and the number of (explicit and implicit) friends they have in the community. If the visitor is a member of one of the communities and happens to be in the collection available at the demo, she can even choose herself as a query initiator. Otherwise, she chooses an initiator among the list of candidates provided by the system. This allows to study the influence of the match of the query and user profile (huge vs. few or even no tag overlap) and the size of the user's friend network on query performance and result quality. The query is then evaluated as spiritual, social or global search (using buttons in the interface that set the parameters to predefined values) or as hybrid query by explicitly specifying values for the parameters α and β. The demo additionally allows to select among different definitions of friendship and tag similarities.


Results for query: mystery / user: PandorasRequiem - Opera
File Edit View Bookmarks Widgets Tools Help
New tab
Results for query: mys...
http://infao5701.ag5.mpi-sb.mpg.de:8081/experiments/servlet/LTServlet
Google
advanced settings
query

○ global
● social ○ spiritual
○ hyprid, α= β=

mystery

Select User for Query: 'mystery' #Docs: 106 #Friends: 89 | PandorasRequiem

#Docs: 1 #Friends: 134 | virgingloves
#Docs: 29 #Friends: 90 | cnrenner
#Docs: 38 #Friends: 90 | LolaWalser
#Docs: 29 #Friends: 89 | AsYouKnow_Bob
#Docs: 106 #Friends: 89 | PandorasRequiem
#Docs: 122 #Friends: 65 | ggchickapee

user: 'PandorasRequiem' - Docs w

	Fat Tuesday (1997)	by Sandra Brown	tags by 'PandorasRequiem' : new, pop, brown, unread, thriller, **mystery**, orleans, romance, suspense, fiction,
	The Stand (2004)	by Stephen King	tags by 'PandorasRequiem' : post-apocalyptic, sf, end, sff, tbr, the, vs., dark, epic, fantasy, fiction, **mystery**, science, shelved, ...

Resulting Books for Query: 'mystery':

Contributing Tags: **mystery,** ..more by Dorothy L. Sayers
1. Strong Poison (1995)
tags by 'SeriousGrace' : women, sayers, booklust, fiction, **mystery**, dorothy,

Contributing Tags: **mystery,** ..more by Dorothy L. Sayers
2. Busman's Honeymoonc (1995)
tags by 'SeriousGrace' : women, sayers, booklust, fiction, **mystery**, dorothy,

Contributing Tags: **mystery,** ..more by Josephine Tey
4. The daughter of time, (1953)
tags by 'Eurydice' : age, iii, alan, 1950s, crime, grant, hospital-bound, solves, **mystery**, richard, inspector, historical, golden

Figure 3: Screenshot of the search interface

The system then produces a set of results (i.e., documents) together with an explanation for each result, consisting of the most influential users which contributed the highest scores to the results, together with the corresponding tag(s). The interface additionally provides a 'drill-down' option to examine the complete set of users which contributed to the score of a result. The visitor can additionally browse the documents and tags of these users just like in the community systems themselves. This allows the user to get some idea why these results were generated. Query evaluation usually takes less than one second for our demo collections. Beyond these features that would also be available in a production version of such a system, we additionally show visualizations of the search process itself. First, SENSE can show an excerpt of the social network around the querying user, highlighting users that have been considered by ContextMerge during query execution. Figure 4 shows an example for this. This helps to demonstrate how ContextMerge incrementally opens user lists and visits only a small fraction of the social network when evaluating a query. A similar interface is available to browse the social network. Second, lists of similar tags can be displayed for each query tag, and again those which were considered during query execution are highlighted.

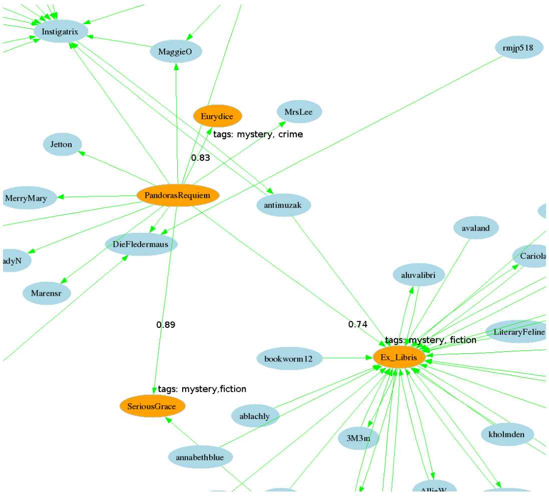


Figure 4: Excerpt of the social network as shown in the user interface

5. REFERENCES

[1] S. Amer-Yahia et al. Challenges in searching online communities. *IEEE Data Eng. Bull.*, 30(2):23–31, 2007.

[2] V. N. Anh and A. Moffat. Pruned query evaluation using pre-computed impacts. In *SIGIR*, 2006.

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

[4] M. Bender et al. Peer-to-peer information search: Semantic, social, or spiritual? *IEEE Data Eng. Bull.*, 30(2):51–60, 2007.

[5] B. Billerbeck and J. Zobel. Questioning query expansion: An examination of behaviour and parameters. In *ADC*, 2004.

[6] M. Dubinko et al. Visualizing tags over time. *ACM Transactions on the Web*, 1(2), 2007.

[7] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *J. Comput. Syst. Sci.*, 66(4):614–656, 2003.

[8] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

[9] J. Pouwelse et al. Tribler: A social-based peer-to-peer system. In *IPTPS*, 2006.

[10] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR*, 1994.

[11] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. Parreira, and G. Weikum. Efficient top-k querying over social-tagging networks. In *SIGIR*, 2008.

[12] Special section on social media and search. *IEEE Internet Computing*, 11(6), 2007.

[13] Special issue on data management issues in social sciences. *IEEE Data Engineering Bulletin*, 30(2), 2007.

[14] M. Theobald, R. Schenkel, and G. Weikum. Efficient and self-tuning incremental query expansion for top-k query processing. In *SIGIR*, 2005.

[15] S. Xu et al. Using social annotations to improve language model for information retrieval. In *CIKM*, 2007.