Context-Aware Personal Information Retrieval From Multiple Social Networks

Xiaogang Han, Wei Wei, Chunyan Miao, Jian-Ping Mei and Hengjie Song

Abstract

People use a variety of social networking services to collect and organize web information for future reuse. When such contents are actually needed as reference to reply a post in an online conversation, however, the user may not be able to retrieve them with proper cues or may even forget their existence at all. In this paper, we study this problem in the online conversation context and investigate how to automatically retrieve the most context-relevant previously-seen web information without user intervention. We propose a Context-aware Personal Information Retrieval (CPIR) algorithm, which considers both the participatory and implicit-topical properties of the context to improve the retrieval performance. Since both the context and the user's web information are usually short and ambiguous, the participatory context is utilized to formulate and expand the query. Moreover, the implicit-topical context is exploited to implicitly determine the importance of each web information of the targeting user in the given context. The experimental results using real-world dataset demonstrate that CPIR can achieve significant improvements over several baselines.

Index Terms

Context-aware information retrieval, Information reuse, Social networking services.

I. Introduction

Information reuse and re-finding are very common behaviors in both desktop-based personal information management systems [1][2] and Web search engines [3]. Many information-related activities involve referring to and integrating previously-seen information. For example, when replying to questions on question answering websites or posts on Social Networking Services (SNSs), the user may need such previously-seen information as reference to support reuse. Previous studies [1] have shown that 58%-81% of web page access are re-visits to pages previously seen. Traditionally, people organize and store the interested information in their desktop and then use classical information retrieval technologies to retrieve them for reuse. With the exponential growth of Web 2.0 services, people tend to utilize SNSs to collect and share previously-seen information [4][5][6]. Typical examples of such services include microblogging (e.g., twitter), social network (e.g., Fackbook) and social bookmarking (e.g., Delicious).

We use Personal Web Information (PWI) to indicate the previously-seen information that have been collected and shared by a user on different SNSs. Practically, it is a very important and challenging task to make connections between the user's context and his PWIs automatically and implicitly [1][7], especially when the PWIs spread across multiple SNSs. For example, a film lover, who has reviewed a classical movie on Facebook a few years ago, can provide potentially valuable comments to his friend's recent post about that movie on Twitter. However, the user may not be able to retrieve the review with proper cues or may have totally forgotten it.

In this paper, we formulate the problem of automatic personal web information retrieval. The study addresses how to build a query by implicitly capturing the individual user's *information need* in the on-line conversation context and how to retrieve the user's most relevant PWIs to facilitate information reuse. However, the task is non-trivial. Although the social aggregation services (SASs) can gather the web information of individual users from different SNSs and the conversations in SASs provide an ideal environment to simulate user contexts, there still exists the following challenges. First, the posts in the

Xiaogang Han (hanxiaogang@ntu.edu.sg), Wei Wei (corresponding author, weiwei8329@gmail.com), Chunyan Miao (ASCYMiao@ntu.edu.sg) and Jian-Ping Mei (JPMei@ntu.edu.sg) are with the School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798. Hengjie Song (song@ii.ist.i.kyoto-u.ac.jp) is with Kyoto University.

conversations are usually short and ambiguous [8], which cannot provide sufficient cues for PWIs retrieval. In addition, since the users' documents in SNSs are noisy and complex [9][10], the pairwise relevance measuring alone cannot capture the similarities between the query generated from the context and the user PWIs. Figure 1 shows an example conversation[1] extracted from FriendFeed.

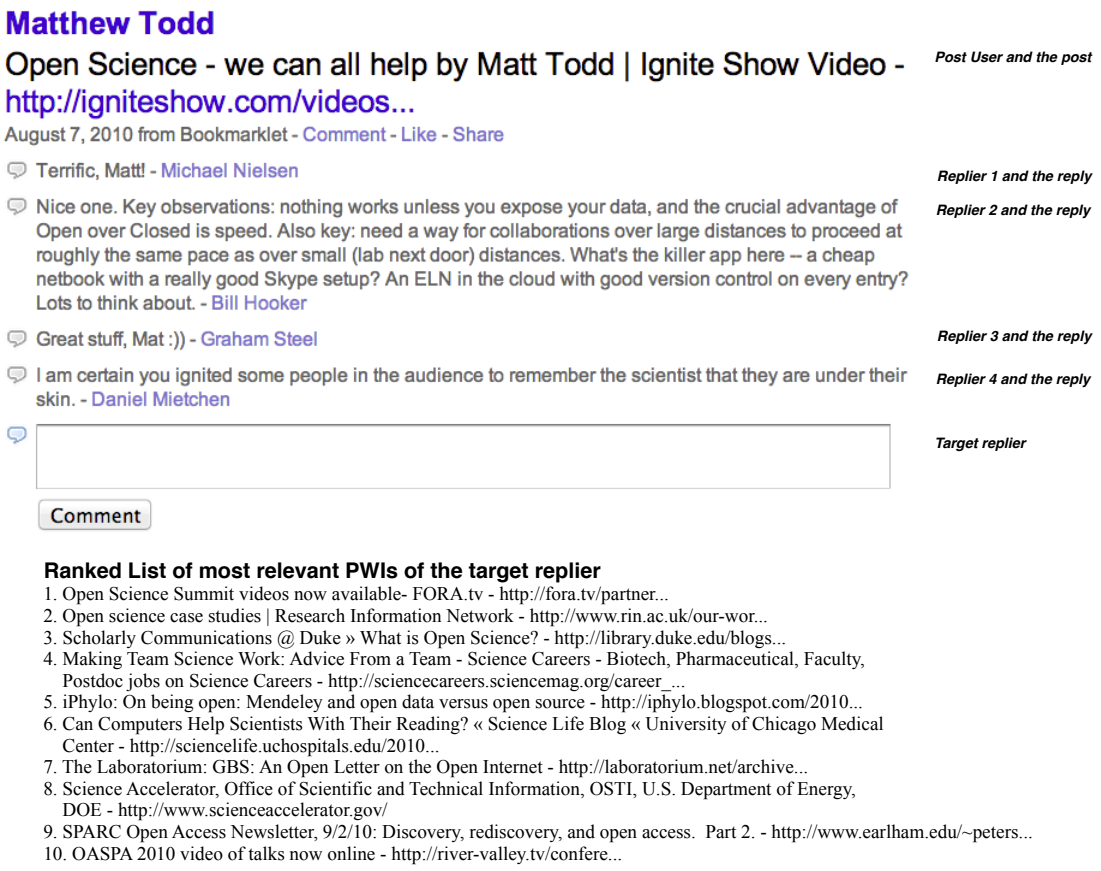

Fig. 1. A conversation on FriendFeed, with the ranked list of top 10 most relevant PWIs of the targeting replier from our algorithm. In this conversation, *Matthew Todd* wrote a post on "open science" and four repliers have responded to the post. Our objective is to automatically retrieve context-relevant documents from the targeting replier's own PWIs before the targeting replier composing the response message for reuse. The most relevant ones retrieved by our algorithm are shown at the bottom.

Through data analysis described in Section V, it is found that (i) the replies and PWIs of all users participating in a conversation provide additional information that may be used to expand the query. For example, a football fan usually posts football related news on different SNSs. Therefore, in a football related conversation, all football related documents of all repliers can be mined to expand the context; and (ii) the users (e.g., the initiator, the existing repliers, and the targeting replier) involved in the same conversation share common interests - the topic of the conversation. From this perspective, there exists a subset of the PWIs of all participating users, which is related to the topic of a conversation. As a result, the utilization of these implicit relationships make it possible to reveal the subset, and thus obtain the list of relevant PWIs for the targeting user.

Based on the analysis, a two-step ranking approach called CPIR is developed to solve the problem and obtain better retrieval results. We treat an online conversation as a session and the post created by the initiator in a session as the initial query. Firstly, the participatory context, which includes the replies and PWIs of all participating users, is exploited. KL-Divergence [11] model with a customized smoothing method is developed to find the most relevant replies and PWIs of the existing repliers to expand the initial query. Moreover, the implicit-topical context, in which a graph-based algorithm is employed to

[1]The conversation is from http://friendfeed.com/science-2-0/ea2aadbb/open-science-we-can-all-help-by-matt-todd-ignite

deduce the implicit relationships among the PWIs of all participating users, is introduced to rank those documents so that documents with higher similarities with each other are assigned with higher scores. The relevance score of each PWI measured by its similarity with the query and the importance score of each PWI obtained from the graph-based algorithm are combined subsequently to produce the final ranking score for each PWI of the targeting replier. In our study, FriendFeed[2], a well-known social aggregation service which aggregates users' PWIs from different SNSs, is used as the testbed. Given a replier who is trying to reply a post in a conversational session in FriendFeed, the concerned problem is how to efficiently capture the context of the session and retrieve the most relevant PWIs of the targeting replier for reuse. The contributions of our work are summarized as follows:

- We formulate the problem of automatic context-aware personal information retrieval across multiple SNSs and investigate how to retrieve the most relevant user PWIs by exploiting the various context information.
- We introduce a participatory context to expand the query with two levels of participator information. A customized smoothing method based on KL-divergence is developed to measure the similarities between the documents in order to expand the query.
- We introduce implicit-topical context to exploit the implicit relationships between the PWIs of all the users involved in the session for deducing the subset of user documents that are most important to the session.

The rest of the paper is organized as follows. The related work is discussed in Section II. In Section III the concerned problem is stated. Our proposed method called CPIR is presented in detail in Section IV. Then the experimental results and analysis will be given in Section V. Finally, we conclude the paper and outline possible future work in Section VI.

II. Related Work

In this section, we briefly review the related work from three aspects.

A. Personal information retrieval across multiple social networks

Personal information indexing and re-finding have been studied extensively in the literature of extending human memory [12]. Staff I've Seen [1] was developed to facilitate information re-use on the desktop by indexing the documents that the user has seen and utilizing various contextual cues such as date, document type, and author to assist retrieval in the search interface. SenseCam [13] captures people's everyday life using a wearable camera to support people's memory for the past and personal events. However, the scope of these work are limited to desktop or a single device.

The information re-finding problem in social networks is very challenging due to the information fragmentation problems [14]. As the user's data is distributed on different web sites, the diversity and heterogeneity among platforms degrades data connectivity and cleanness [15]. Moreover, the social networks also bring new features via so called collective intelligence, which are significantly different from traditional personal information management systems. For instance, the social tagging generated by the community in Folksonomy [16] and knowledge aggregated in community question answer websites [17] have provided extra metadata about entities in which each individual user is involved. Finally, evidence shows that the user's documents on different platforms share a common vocabulary on multiple social bookmarking systems [18][19] and multiple social network sites [20]. Therefore, the data needs to be aligned before integration.

[2]FriendFeed: www.friendfeed.com

B. Context-based query generation and retrieval approaches

Normally, information retrieval systems utilize solely the terms in the query and the document collection to decide the relevance, while the information in the specific context is ignored [21]. Automatic query generation [22] is the task of identifying the most representative texts in the context for information retrieval. Query expansion is the process of enriching the original query by adding extra content words deduced from the context [23][24]. In order to implicitly query relevant resources based on user's current computing activities, [25][26] utilize *tfidf*-based *cosine* similarity to rank candidate resources based on their similarities with the document in the context. More recent approaches such as [27] represent the meaning of words as interval type-2 fuzzy sets that constrain an abstract emotion space and utilize Jaccard Index to calculate the similarity between two words. In our work, the query is generated by considering the post, replies to the post, and the PWIs of all the participating users, which captures richer information regarding both the context and the user.

Retrieving relevant documents from a collection of candidate documents can be considered as a classification problem or a ranking problem. We treat the personal information retrieval problem as a ranking problem in which two properties of relevance are explored and combined. Fuzzy combination methods have been shown [28] to consistently outperform simple linear combination methods. However, since the linear combinators showed a stable performance [28], we utilize a linear combinator in our experiment.

C. Social aggregation services

Gupta et al. [29] studied the FriendFeed service in terms of social aggregation properties and user activity patterns. They found that approximately 73% of the users have subscribed to two or more services. Celli et al. [30] provided a descriptive analysis on FriendFeed and identified the distinction between weak and highly dedicated users via clustering analysis. They concluded that for the group of highly dedicated users, FriendFeed provided a perfect platform for hosting fruitful conversations based on the sharing of a large amount of information from various SNSs. Garg et al. [31] examined the evolution of the FriendFeed network and found that membership ages, proximity between users, and commonness in aggregated services are the primary factors for relationship formation. Such patterns support our assumption that the users involved in the same conversation share common interests across multiple SNSs.

III. Problem Statement

Given a session and the targeting replier for whom we will make recommendation, the context-aware personal information retrieval problem studied in this paper is to model the context of the session and generate a query so as to retrieve the most relevant PWIs from the user's document collection.

For clarity, the key notations used in this paper are listed in Table I. In particular, the definition of *Session* is as follows:

Definition 1: A **Session** ($\mathcal{S}$), is an on-line conversation with an initial post p and a set of replies $\mathcal{R} = \{r_1, \cdots, r_i, \cdots, r_n\}$.

Examples of $\mathcal{S}$ include conversations in SNSs, such as a tweet posted on Twitter with replies from fellow followers, or a question posted on a question answering website with answers from other users.

In a Session ($\mathcal{S}$), p denotes the initial post. $\mathcal{U}$ indicates the set of users involved in $\mathcal{S}$, i.e., $\mathcal{U} = \{u_p, u_1, \cdots, u_i, \cdots, u_n, u_t\}$. u_p is the creator of p, u_t is the targeting replier to whom we will make recommendations, and u_i ($u_i \notin \{u_p, u_t\}$ and $i = 1...n$) is an existing replier. $\mathcal{R} = \{r_1, \cdots, r_i, \cdots, r_n\}$ is the set of existing replies, assuming that each user gives solely one reply. $\mathcal{D} = D_p \cup \{D_{1i}\}_{i=1}^{n} \cup D_t$ is the entire collection of PWIs of all the users within $\mathcal{S}$, where D_p, D_t and D_i ($D_i \notin \{D_p, D_t\}$) are the sets of PWIs of u_p, u_t and u_i respectively. In addition, $D_i = \{d_j\}_{j=1}^{m_i}$ is the PWIs of u_i, where m_i is the number of u_i's PWIs.

The content of all the documents in $\mathcal{S}$ are represented by the Vector Space Model [32], i.e., $\mathbf{v}_d = [w_{1,d}, \cdots, w_{k,d}]^T$, in which each term is weighted by its *tfidf* score [33].

TABLE I
SYMBOLS AND DEFINITIONS

Symbol	Definition
$\mathcal{S}$	Session
p	Initial post in $\mathcal{S}$
Q	Expanded query
$\mathcal{R}$	Collection of existing replies, $\mathcal{R} = \{r_i\}_{i=1}^{n}$
$\mathcal{U}$	User collection within $\mathcal{S}$, $\mathcal{U} = \{u_p, u_1, ..., u_n, u_t\}$
D_i	PWIs set of i-th replier, $D_i = \{d_j^i\}_{j=1}^{m_i}$
$\mathcal{D}$	Collection of PWIs of all the users involved in $\mathcal{S}$, $\mathcal{D} = \{D_p, D_1, ..., D_n, D_t\}$

IV. CONTEXT-AWARE PERSONAL INFORMATION RETRIEVAL

In this section, we present the details of the CPIR method.

A. Overview

CPIR is mainly decomposed into two sub-steps: (i) query formulation and expansion: building query by extracting sufficient information from the session; and (ii) PWIs ranking: extracting the most relevant PWIs of the targeting replier according to the query. The overview of CPIR is illustrated in Figure 2.

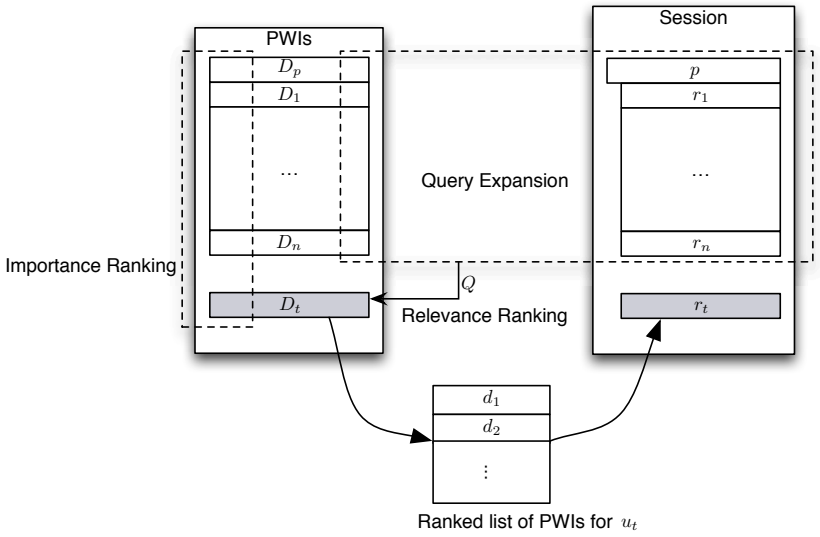

Fig. 2. Diagram of Context-aware Personal Information Retrieval Framework. The framework takes a session and users' PWIs as input. It works by first building a query Q from the session using the information in both the session and users' PWIs. A graph-based algorithm is employed to derive the importance scores for each PWI of u_t, which is then combined with the relevance scores to obtain the final ranked list of documents for u_t.

In the first step, the participatory context is used to formulate and expand the query by considering both the replies and the PWIs of all participating users. An intuitive idea is to use the initial post p and the existing replies $\mathcal{R}$ as the context to build the query since they are the basic available context information in $\mathcal{S}$. However, the posts in SNSs are usually short and ambiguous and thus are not sufficient to characterize different properties of the session. Therefore, the PWIs of the creator and the existing repliers are utilized to obtain richer information. The main reason is that users participating in the same session usually share common interests related to p, and they might have posted similar documents in other SNSs before. Therefore, the PWIs of the creator and the existing repliers can be used as the complementary context information.

In the second step, the shared interests among the targeting replier, the creator, and the existing repliers are considered, which forms the implicit-topical context. One of the common interests of all the participating users is the topic of the conversation they are involved in. By implicitly inferring the subset of documents on the topic in an unsupervised manner, the relevant PWIs of the targeting user can be collected.

In the remaining parts of this section, we will describe the two steps in detail and then combine the ranking results from these two steps to obtain the final ranking scores for each PWI of the targeting user.

B. Utilizing Participatory Context for Query Expansion

As aforementioned, the context of the session $\mathcal{S}$ consists of an initial post p, existing replies, and the PWIs of the creator and the existing repliers. To accurately model the context of session $\mathcal{S}$ so as to build a comprehensive query for covering the different properties of context, we need to extract as many cues as possible from them.

In order to model the context, the query Q is built by modeling the session at two levels. Formally, the initial p is treated as the basic query. We first combine the replies of existing repliers with p, since these replies are the responses to p and thus can provide extra information about $\mathcal{S}$. As different replies have different levels of correlation with p, the replies are weighted according to their similarities with p. The expanded query is calculated as follows:

$$\mathbf{v}_{Q_{p+R}} = \alpha \mathbf{v}_p + (1-\alpha) \sum_{r_i \in \mathcal{R}} sim(\mathbf{v}_p, \mathbf{v}_{r_i}) \cdot \mathbf{v}_{r_i} \tag{1}$$

where α $(0 \leq \alpha \leq 1)$ is a trade-off parameter to control the contribution of replies. $sim(.,.)$ is the similarity metrics to measure the relevance between the initial post p and the replier r_i $(r_i \in \mathcal{R})$.

Among the different similarity measures, it has been shown that probabilistic methods like KL-divergence [34] can obtain better results than vector space based measures [35][36], especially for short texts [37] like the PWIs discussed in this paper. However, as the vocabulary in PWIs is sparse, smoothing techniques are always introduced to take the entire vocabulary into consideration to compare two distributions. We introduce the translation-based language model [38] with WordNet[3] as an external source to expand the documents before calculating their similarities.

The KL-divergence between p and r_i, $D_{KL}(\mathbf{v}_p||\mathbf{v}_{r_i})$ is calculated as follows:

$$D_{KL}(\mathbf{v}_p||\mathbf{v}_{r_i}) = \sum_{w_{i,p}} P'(w_{i,p}|\mathbf{v}_p) \log \frac{P'(w_{i,p}|\mathbf{v}_p)}{P'(w_{i,p}|\mathbf{v}_{r_i})} \tag{2}$$

in which $P'(w|\mathbf{v})$ is the expanded distribution. $P'(w|\mathbf{v})$ is calculated as follows:

$$P'(w|\mathbf{v}) = \sum_{w' \in \mathbf{v}} f(w'|w) P(w'|\mathbf{v}) \tag{3}$$

where $P(w'|\mathbf{v})$ denotes the *tfidf* score of w' in $\mathbf{v}$ and $f(w'|w)$ is the translation probability of word w to word w' calculated using WordNet sense similarity.

The similarity between p and r_i can be calculated as:

$$sim(\mathbf{v}_p, \mathbf{v}_{r_i}) = e^{-|\frac{1}{2}(D_{KL}(\mathbf{v}_p||\mathbf{v}_{r_i}) + D_{KL}(\mathbf{v}_{r_i}||\mathbf{v}_p))|} \tag{4}$$

To further expand the query, we also consider the PWIs of the creator and the existing repliers. However, not all of them are incorporated into query Q which otherwise will result in a very long query. Instead, the top k most relevant ones are selected to expand $Q_{p+\mathcal{R}}$. The expanded query can be represented as,

$$\mathbf{v}_Q = \beta \mathbf{v}_{Q_{p+\mathcal{R}}} + (1-\beta) \sum_{d \in D_i} sim(\mathbf{v}_d, \mathbf{v}_{Q_{p+\mathcal{R}}}) \cdot \mathbf{v}_d \tag{5}$$

[3]http://wordnet.princeton.edu/

where β $(0 \leq \beta \leq 1)$ is a trade-off parameter and $D_i \in \mathcal{D}$ but not D_t.

C. PWIs Ranking

1) Utilizing Implicit-Topical Context for Importance Ranking: As aforementioned, the users involved in the same session $\mathcal{S}$ share common interests, at least including the topic of S. We employ a Markov random walk model to rank all the PWIs of a user u based on the implicit relationships between the web information of all the users in $\mathcal{S}$ and find a subset of u's PWIs that are most relevant to the topic of the session.

Let $G(N, E)$ be a graph of documents where $N = \{n_1, \cdots, n_{|N|}\}$ is a set of vertices and $E = \{e_1, \cdots, e_{|E|}\}$ is a set of edges. In G, a vertex $n_i \in N$ is an PWI $d \in D_i$ $(D_i \in \mathcal{D}, \text{and } D_i \neq D_p)$. The transition probability matrix of G is represented by $P = [p_{ij}]$, in which each transition probability from node n_i to node n_j is given by,

$$p_{ij} = \frac{sim(\mathbf{v}_{n_i}, \mathbf{v}_{n_j})}{\sum_k sim(\mathbf{v}_{n_i}, \mathbf{v}_{n_k})} \tag{6}$$

where $sim(.,.)$ is defined as Eq. (4).

To overcome the "dangling link" while conducting a random walk on graph G, the similarity scores between the generated query Q and each PWI in G as the reset probability are used to allow the random walk process jumping from a node to an arbitrary node with a small probability. For node n_i, the reset probability x_i is calculated as follows:

$$x_i = sim(\mathbf{v}_{n_i}, \mathbf{v}_Q) \tag{7}$$

In order to make the sum of all the elements in $\mathbf{x}$ equal to 1.0, each $x_i \in \mathbf{x}$ will be normalized by $x_i = \frac{x_i}{\sum_j x_j}$. With the transition matrix P and the reset probability vector $\mathbf{x}$, the stationary eigenvector π can be computed iteratively using the power method [39].

2) Final Ranking for User PWIs: The ranking obtained from the random walk model denotes the importance of the document in the collection of PWIs. The similarity between the expanded query Q and each document measures the relevance of the document for the session. We use a linear combination of these two ranking scores to obtain the final score for each candidate $d_i \in D_t$, which is calculated as follows:

$$Score(d_i) = \lambda cosine(d_i, Q) + (1 - \lambda)\pi_i \tag{8}$$

where λ $(0 \leq \lambda \leq 1)$ is a combining parameter. Note that $cosine$ measure is used to calculate the similarities between the Q and the user document d_i. It is primarily due to the fact that Q is usually quite longer than PWI since it is the combination of multiple PWIs, and thus $cosine$ is more suitable in this case. Finally, the top ranked PWIs ($d_i \in D_t$) are selected as the recommendation results to the targeting replier. The algorithm is described in Algorithm 1.

V. Experiments and Analysis

A set of experiments were designed to evaluate the retrieval algorithm. In this section, we describe our analysis on the dataset and discuss the performance of the proposed CPIR algorithm by comparing it with other baselines.

A. Experiment settings

1) Data Description: Celli et al. [30] provided a FriendFeed dataset[4], which was collected by monitoring the data stream on FriendFeed from 01/08/2010 to 30/09/2010. Note that the set of comments for

[4]FriendFeed Dataset: http://larica.uniurb.it/sigsna/data/

ALGORITHM 1: Context-aware Personal Information Retrieval (CPIR)

input : $p,\ R,\ \mathcal{D} = \{D_p, D_t, D_1, \cdots, D_n\}$
output: Document set $D_t^* = \{d_i\}_{i=1}^{K},\ d_i \in D_t$, which contains the top-k relevant documents with regarded to the topic of Session $\mathcal{S}$

Query Expansion:
$v_{Q_p} \leftarrow v_p$;
foreach $r_i \in \mathcal{R}$ **do**
  Compute $sim(v_p, v_{r_i})$ according to Equation (2), (3) and (4);
  $v_{Q_{p+R}} \leftarrow \alpha v_{Q_p} + (1-\alpha) sim(v_{Q_p}, v_{r_i}) \cdot v_{r_i}$;
end
foreach $r_i \in \mathcal{R}$ **do**
  $D_i^* \leftarrow \arg\max_{X, d \in D_i} sim(v_{Q_{p+R}}, v_d)$;
  foreach $d \in D_i^*$ **do**
    Compute $sim(v_{Q_{p+R}}, v_d)$ according to Equation (2), (3) and (4);
    $v_Q \leftarrow \beta v_{Q_{p+R}} + (1-\beta) sim(v_{Q_{p+R}}, v_d) \cdot v_d$;
  end
end
Importance Ranking:
Construct the probability matrix $\mathbf{P}$ according to Equation (6);
Construct the reset probability vector $\mathbf{x}$ according to Equation (7);
Compute iteratively $\pi^{(t+1)} = \gamma \mathbf{P} \pi^{(t)} + (1-\gamma)\mathbf{x}$;
Final Ranking:
foreach $d \in D_t$ **do**
  Score$(d_i) = \lambda$cosine$(d_i, Q) + (1-\lambda)\pi_i$;
end
$D_t^* \leftarrow \arg\max_K$ Score$(d),\ (d \in D_t)$;
Return D_t^*;

a single entry only contains the initiator's comments, while the comments provided by other users are not included. In order to obtain complete conversations, we re-extracted all the conversations in the dataset via the FriendFeed API[5].

Table II summarizes the basic information about the conversations and their replies, including the number of repliers involved in these conversations and the aggregated PWIs for those users. From these conversations, we select the post-reply pairs written in English[6] and the repliers of which have at least 50 PWIs.

TABLE II
BASIC STATISTICS OF THE FRIENDFEED DATASET

Number of conversations	56,460
Number of replies	749,369
Number of repliers	14,443
Number of PWIs	637,320

In order to construct manual annotation results for evaluation, we randomly sampled 105 post-reply pairs. Those replies are posted by 73 unique users. Each user has 316 PWIs on average. Two volunteers manually labeled the 23,046 PWIs of the repliers as either relevant or irrelevant for the given conversations.

Before applying the models to the documents, tokenization and part-of-speech tagging are performed to eliminate noisy terms. In addition, stop words are removed and terms are stemmed using Porter Stemmer [40].

2) Data Analysis: Figure 3(a) shows the distribution of the number of repliers per conversation. Figure 3(b) shows the distribution of unique replies per conversation. We can see that 98% of conversations

[5]FriendFeed API: http://friendfeed.com/api/

[6]The language detection tool: guess-language (https://github.com/dsc/guess-language) is used for language detection

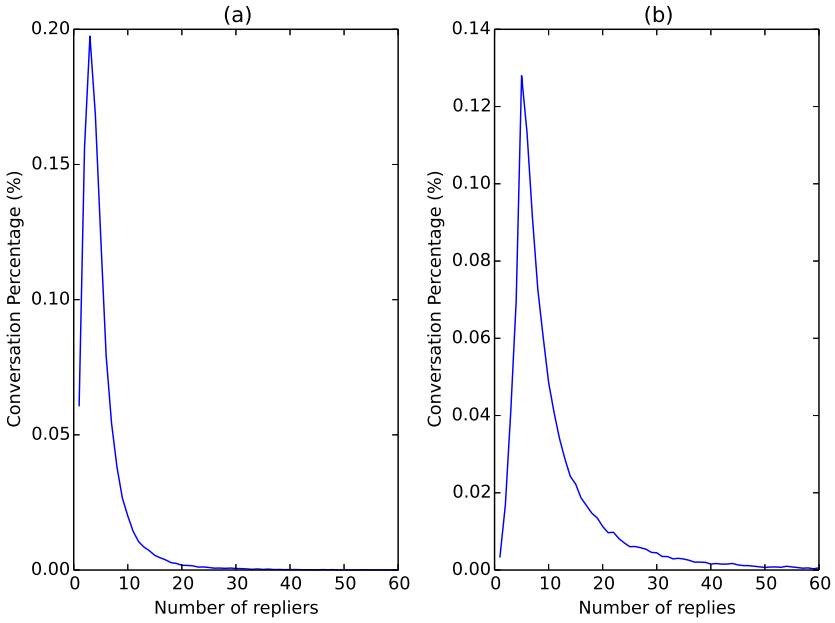

Fig. 3. Conversation engagement - (a) Distribution of the number of repliers per conversation in the corpus, (b) Distribution of the number of replies per conversation in the corpus.

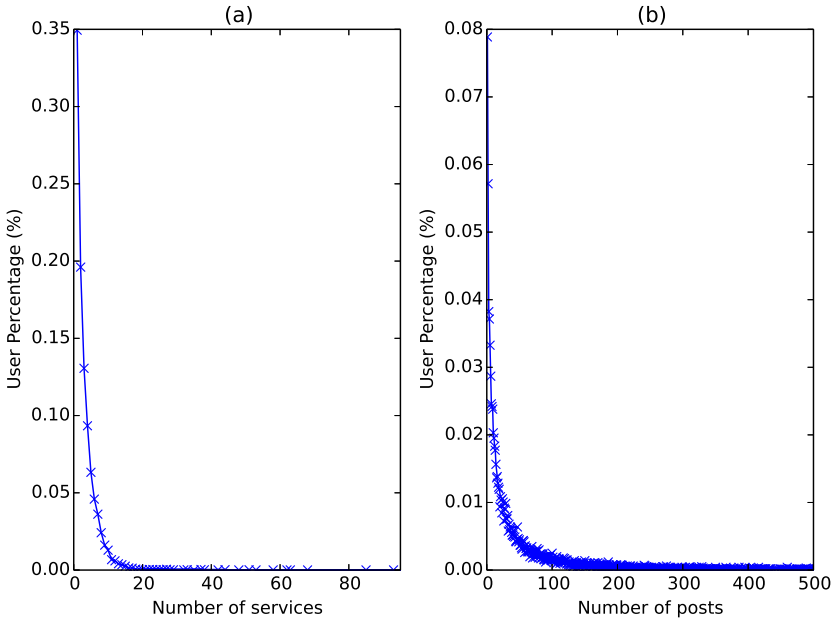

Fig. 4. User engagement - (a) Distribution of the number of services per user in the corpus, (b) Distribution of the number of PWIs per user in the corpus.

have at least three replies and 78% of conversations have at least three unique repliers. This confirms the feasibility of utilizing the conversations to model the task environment so as to retrieve past information.

The distribution of the number of aggregated services for the repliers is depicted in Figure 4(a), which shows that more than 65% of the users use at least two services. It confirms that the documents are extracted from diverse information sources. The distribution of the number of aggregated PWIs for the users is depicted in Figure 4(b). It indicates that 63% of the users posted more than 10 PWIs. These observations motivate us to utilize the PWIs of the users in the conversation to expand the query and

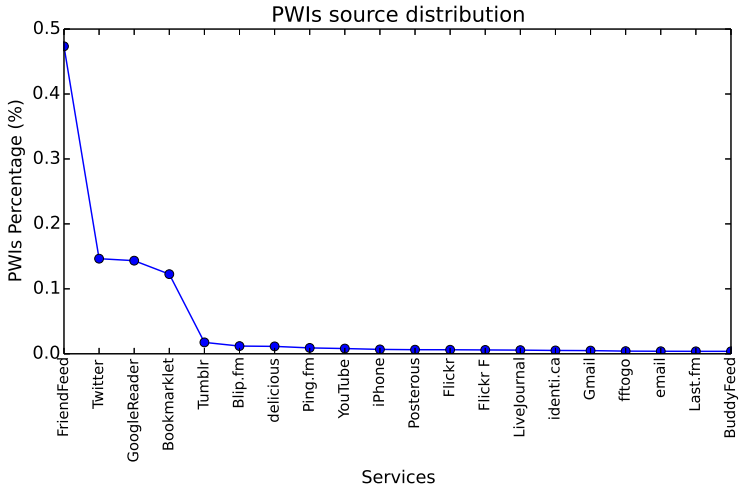

Fig. 5. Top 20 services in terms of the number of PWIs.

improve the retrieval performance.

The top services in terms of the number of PWIs are depicted in Figure 5. We can see that the major portion of PWIs are aggregated from the popular services including FriendFeed, Twitter, and Google Reader.

3) Evaluation Metrics: The performance of our algorithm is evaluated against six widely used metrics - precision at rank 1 (P@1), precision at rank 5 (P@5), recall, mean average precision (MAP), mean reciprocal rank (MRR), and F-measure [41]. The top 50 retrieved PWIs for each post-reply pair are used for evaluation.

4) Baselines: Our algorithm is compared with three baselines. They include:

- using the initial query to build the session query, denoted as $post$;
- using $p{+}R$ to build query, with *cosine* and *KL* to measure similarities between two vectors, denoted as $prcos$ and $prkl$ respectively;
- using $p{+}R{+}D$ to build query, but with *cosine* as similarity measure, denoted as $prcos{-}prdcos$.

Those baselines are compared against the relevance ranking scheme in CPIR, denoted as $\text{CPIR}_{\lambda=1}$.

Moreover, we combine each of the baselines with the importance scores calculated from the random walk model and obtain another set of baselines, which are denoted as $post{+}graph$, $prcos{+}graph$, $prkl{+}graph$, and $prcos{-}prdcos{+}graph$ respectively. These baselines are compared against CPIR.

TABLE III
RETRIEVAL RESULTS OF EXPANDING THE SESSION QUERY, COMPARED WITH BASELINES

Algorithms	P@1	P@5	Recall	MAP	MRR	F-measure
post	0.6476	0.4667	0.7639	0.4395	0.7608	0.2197
prcos	0.6476	0.4914	0.8009	0.4651	0.7720	0.2316
prkl	0.7143	0.5162	0.8263	0.4949	0.8176	0.2390
prcos-prdcos	0.7619	0.5505	0.8272	0.5290	0.8336	0.2414
$\text{CPIR}_{\lambda=1}$	**0.8000**	**0.5562**	**0.8529**	**0.5466**	**0.8620**	**0.2497**

TABLE IV
RETRIEVAL RESULTS OF COMBINING RELEVANCE AND IMPORTANCE

Algorithms	P@1	P@5	Recall	MAP	MRR	F-measure
post+graph	0.6571	0.4762	0.7881	0.4618	0.7693	0.2306
prcos+graph	0.6762	0.4933	0.8191	0.4835	0.7857	0.2394
prkl+graph	0.7238	0.5200	0.8457	0.5035	0.8232	0.2471
prcos-prdcos+graph	0.7714	0.5524	0.8489	0.5401	0.8412	0.2494
CPIR	**0.8000**	**0.5619**	**0.8619**	**0.5528**	**0.8618**	**0.2543**

B. Retrieval Performance

Table III shows the comparison between $\mathsf{CPIR}_{\lambda=1}$ and the corresponding baseline algorithms. $\mathsf{CPIR}_{\lambda=1}$ achieves a significant improvement over the baseline methods with respect to all the six metrics. Expanding the initial query with the replies in the conversation enhanced the context cues, while adding selected PWIs further captured the context information. It is also observed that our KL-based measure outperforms $cosine$-based measure to calculate document similarities.

Table IV compares the retrieval results of CPIR and the corresponding baselines. CPIR, which combines $\mathsf{CPIR}_{\lambda=1}$ and graph ranking, obtains the best performance compared with the combination of graph ranking with other baselines. When compared with the corresponding baselines without graph ranking in Table III, graph-based ranking algorithm can select the PWIs with common topics among the users involved in the same conversation so as to further improve the performance. Specifically, CPIR outperforms $\mathsf{CPIR}_{\lambda=1}$ in terms of five out of the six evaluation metrics.

TABLE V
PARAMETER SETTINGS

	Value	Description
α	0.6	p and R combination controller
β	0.6	pr and D_{p+R} combination controller
k	15	number of top PWIs for $\mathsf{CPIR}_{\lambda=1}$
λ	0.75	relevance ranking and importance ranking combination controller

A specific example conversation with the retrieval results obtained from CPIR and two baselines is shown in Figure 6. The post and selected replies are shown on the top left, while the manually labeled relevant PWIs of the targeting user are shown on the top right. The retrieval results by *prcos*, *prcos-prdcos*, and CPIR are shown at the bottom, in which the indexes of the matched PWIs are marked in **Bold**. It is observed that CPIR found 6 matched PWIs, while *prcos* and *prcos-prdcos* found 1 and 3, respectively.

We also analyzed the distribution of the retrieved documents in different social network sites. The top five social network sites with the largest number of retrieved documents are shown in Figure 7. It clearly shows that the number of retrieved documents is proportional to the total number of documents in those platforms.

C. Parameter Settings

Table V shows the main parameters and their settings in our experiments. The optimal parameter for each variable are obtained by fine tuning. For example, The optimal α in our experiment is 0.6, which implies that the weights of the terms in p is slightly more important than the terms in R. The optimal k is 15, which means the best performance is obtained by only extending pr with the selected documents in D_{p+R}.

One of the most important parameters in our experiments is λ, which controls how to combine the ranking scores from the query expansion results with the ranking scores from the random walk model.

Post	**Manually labeled relevant PWIs**
ekins older scientists more likely to prefer non collaborative lab notebooks as opposed to wiki jcbradley thoughts acs_boston **Replies** - confounded somewhat by effect of tenure some older scientists are very into open foo but recognize that earlier in their careers they might have been more cautious - i think it would be hard to get meaningful statistics on this numbers are small and how do you define scientist undergrads don t typically have problem with sharing notebook again have to clearly define what sharing means but most of them don t become scientists by most definitions - bill makes an excellent point that younger scientists may not have the same choice as older ones - but this is all going to change no those who begin their career with this technology will always use it in 30 years that will be everyone	0. discusses lab notebook requirements and discusses the scrap of paper disallowed mistake acs_boston acsrdf2010 any literature on such 1. biomed central is going to support open data 2. open science microformats initial thoughts jessy s acceptable 3. why i blog because i need thoughts out of my system like with getting things done write it down know it s organized 4. introducing the mcprinciples for open cheminformatics 5. biomed central is going to support open data via 6. dutch intelligence service raises awareness with universities of the risk of espionage opendata science 7. linked open data and pavlova 8. molecular networks is discussing all the benefits of open standards and open source acs_boston are they changing their own north too 9. sharing detailed research data is associated with increased citation rate

Top 10 PWIs retrieved by *prcos*
0. discusses lab notebook requirements and discusses the scrap of paper disallowed mistake acs_boston acsrdf2010 any literature on such
1. jeremy talks about computers and wet lab chemistry acsrdf2010 acs_boston
2. great drawing of benzene but not quite the way you re used to it acs_boston
3. rich showed some usage stats on one new during during his talk acs_boston
4. the life scientist room at cliqset
5. more acs_boston slides online if anyone at acs listens we need central repository for this
6. oh and jeremy is profechem so feel free to contact the speaker directly acsrdf2010 acs_boston but not until the end of his talk
7. lovely regex logo on slides by lezan acsrdf2010 acs_boston
8. leonid shows nice get coffee slide the soap and the semantic version acsrdf2010 acs_boston
9. the list of 128 papers that cite one or two of the main cdk papers

Top 10 PWIs retrieved by *prcos-prdcos*
0. discusses lab notebook requirements and discusses the scrap of paper disallowed mistake acs_boston acsrdf2010 any literature on such
1. jeremy talks about computers and wet lab chemistry acsrdf2010 acs_boston
2. great drawing of benzene but not quite the way you re used to it acs_boston
3. the life scientist room at cliqset
4. molecular networks is discussing all the benefits of open standards and open source acs_boston are they changing their own north too
5. rich showed some usage stats on one new during during his talk acs_boston
6. open science microformats initial thoughts jessy s acceptable
7. acsrdf2010 acs_boston
8. collaborative development of predictive toxicology applications
9. more acs_boston slides online if anyone at acs listens we need central repository for this

Top 10 PWIs retrieved by CPIR
0. discusses lab notebook requirements and discusses the scrap of paper disallowed mistake acs_boston acsrdf2010 any literature on such
1. the life scientist room at cliqset
2. molecular networks is discussing all the benefits of open standards and open source acs_boston are they changing their own north too
3. jeremy talks about computers and wet lab chemistry acsrdf2010 acs_boston
4. biomed central is going to support open data
5. biomed central is going to support open data via
6. open science microformats initial thoughts jessy s acceptable
7. rich introduces the problem that science is too specialized for not using online tools just like i just did for cheminformatics acs_boston
8. linked open data and pavlova
9. jeremy discusses that chemistry is really slow with open scientific dissemination behind biology for example acsrdf2010 acs_boston

Fig. 6. An example conversation and its retrieval results (documents are truncated)

Figure 8 shows the effect of varying λ on MAP with step $\Delta\lambda = 0.02$. The best MAP is obtained by setting λ to 0.75. The optimal settings for other parameters are obtained in a similar way.

VI. Conclusion and Future Work

In this paper, we formulated the problem of automatic context-aware personal information retrieval to enhance user memory in the online conversation environment. By analyzing the FriendFeed data, we found that the participating users and their PWIs possibly provide rich information for query expansion and implicit PWIs ranking. We employed such information to develop a two-step algorithm, namely CPIR, so as to solve the retrieval problem by considering both the participatory property and implicit-topical property of the context. In the first step, the query in the session is expanded with extra information from both the replies and the PWIs of the participating users. A customized smoothing method is developed to extract semantic information from texts that are typically short in length. In the second step, a graph-based

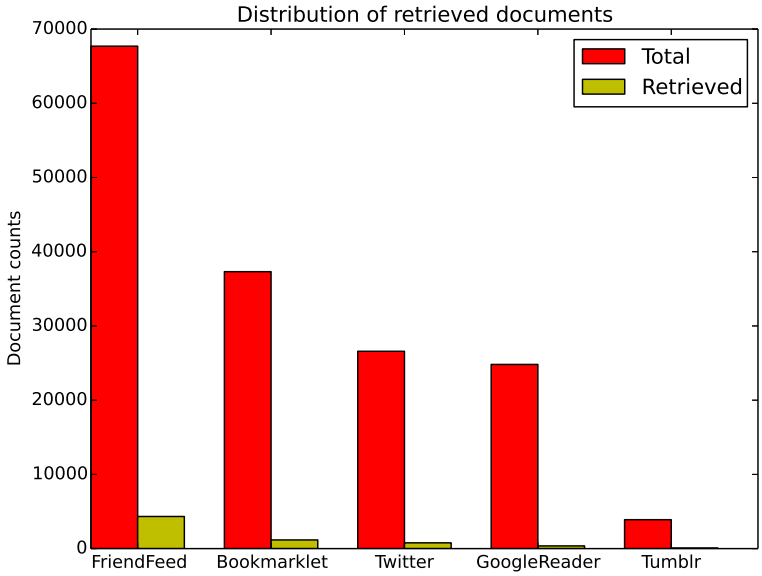

Fig. 7. Distribution of retrieved documents in different SNSs

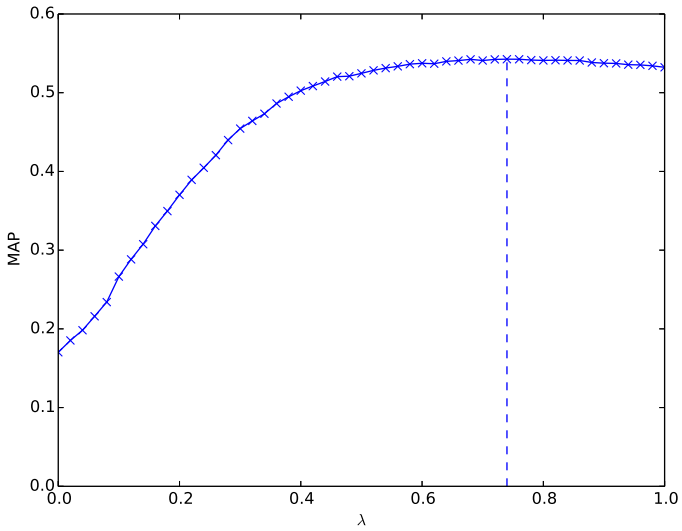

Fig. 8. Effect of varying λ on MAP.

algorithm is employed to reveal the implicit relationship among the PWIs of all participating users and extract the user PWIs which match the topic of the session. The experiments conducted on the real-world dataset demonstrated that CPIR outperforms several baseline methods significantly.

In future, we will investigate how to improve the performance of our proposed method from three different aspects. First, as each conversation might have multiple topics, the importance ranking algorithm in our method can be replaced with clustering-based techniques like fuzzy C-means [42] and self-organizing maps [43] to capture the topical diversity among the collection of participating user PWIs. Given a conversation, this approach makes it possible to extract the most relevant PWIs for a specific user from the most context-relevant clusters. Moreover, we used linear combination in both the query expansion (see Eq. (1) and Eq. (5)) and the final ranking (see Eq. (8)) for their simplicity in this paper.

However, since fuzzy combination methods have been shown [28] to boost the performance, we will integrate fuzzy combination methods into our algorithm and compare it with the linear methods. Finally, as document recency is an important factor in document ranking [44], we plan to take the recency into the algorithm design and study how the time decay factor affects the relevance of the PWIs in a given user context. A longer period of user data collection is preferred to study the shift of user's topic of interests in both short-term and long-term periods.

ACKNOWLEDGMENT

This research is supported by Interactive and Digital Media Programme Office (IDMPO), National Research Foundation (NRF) hosted at Media Development Authority (MDA) of Singapore under Grant No.: MDA/IDM/2012/8/8-2 VOL 01. Wei Wei (weiwei8329@gmail.com) serves as the corresponding author of this article. The authors would also like to thank anonymous reviewers for their comments on improving the quality of this paper.

REFERENCES

[1] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff i've seen: a system for personal information retrieval and re-use," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 2003, pp. 72–79.

[2] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin, "Fast, flexible filtering with phlat," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2006, pp. 261–270.

[3] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts, "Information re-retrieval: repeat queries in yahoo's logs," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 151–158.

[4] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "Ht06, tagging paper, taxonomy, flickr, academic article, to read," in *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, 2006, pp. 31–40.

[5] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 2007, pp. 56–65.

[6] R. Wetzker, C. Zimmermann, and C. Bauckhage, "Analyzing social bookmarking systems: A del.icio.us cookbook," in *Proceedings of the ECAI 2008 Mining Social Data Workshop*, 2008, pp. 26–30.

[7] P. Maes, "Agents that reduce work and information overload," *Communications of the ACM*, vol. 37, no. 7, pp. 30–40, 1994.

[8] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 42–47.

[9] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: a first look," in *Proceedings of the fourth ACM workshop on Analytics for noisy unstructured text data*, 2010, pp. 73–80.

[10] O. Owoputi, B. OConnor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, "Improved part-of-speech tagging for online conversational text with word clusters," in *Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 380–390.

[11] R. M. Fano, "Transmission of information: A statistical theory of communications," *American Journal of Physics*, vol. 29, pp. 793–794, 1961.

[12] H. Bruce, W. Jones, and S. Dumais, "Keeping and re-finding information on the web: What do people do and what do they need?" *Proceedings of the American Society for Information Science and Technology*, vol. 41, no. 1, pp. 129–137, 2004.

[13] A. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood, "Do life-logging technologies support memory for the past?: an experimental study using sensecam," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Ssystems*, 2007, pp. 81–90.

[14] E. L. S. Araujo, Q. Gao and G. Houben, "Linking personal data: towards the web of your digital memories," in *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[15] M. Tungare, P. Pyla, M. Sampat, and M. Perez-Quinones, "Defragmenting information using the syncables framework," in *Proceedings of the 2nd Invitational Workshop on Personal Information Management at SIGIR*, 2006.

[16] V. Robu, H. Halpin, and H. Shepherd, "Emergence of consensus and shared vocabularies in collaborative tagging systems," *ACM Transactions on the Web*, vol. 3, no. 4, pp. 1–34, 2009.

[17] D. Schall and F. Skopik, "An analysis of the structure and dynamics of large-scale q/a communities," in *Advances in Databases and Information Systems*, 2011, pp. 285–301.

[18] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt, "Semantic modelling of user interests based on cross-folksonomy analysis," in *Proceedings of the 7th International Conference on The Semantic Web*, 2008, pp. 632–648.

[19] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 522–525.

[20] F. Abel, S. Araújo, Q. Gao, and G.-J. Houben, "Analyzing cross-system user modeling on the social web," in *Proceedings of the 11th International Conference on Web Engineering*, 2011, pp. 28–43.

[21] X. Shen, B. Tan, and C. Zhai, "Context-sensitive information retrieval using implicit feedback," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005, pp. 43–50.

[22] X. Xue and W. Croft, "Automatic query generation for patent search," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009, pp. 2037–2040.

[23] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu, "Statistical machine translation for query expansion in answer retrieval," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 464–471.

[24] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp, "Incorporating query expansion and quality indicators in searching microblog posts," *Advances in Information Retrieval*, pp. 362–367, 2011.

[25] S. Dumais, E. Cutrell, R. Sarin, and E. Horvitz, "Implicit queries (iq) for contextualized search," in *Proceedings of the 27th Annual International ACM Conference on Research and Development in Information Retrieval*, 2004, pp. 594–594.

[26] J. Shen, W. Geyer, M. Muller, C. Dugan, B. Brownholtz, and D. R. Millen, "Automatically finding and recommending resources to support knowledge workers' activities," in *Proceedings of the 13th international conference on Intelligent user interfaces*, 2008, pp. 207–216.

[27] A. Kazemzadeh, S. Lee, and S. Narayanan, "Fuzzy logic models for the meaning of emotion words," *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 34–49, 2013.

[28] L. I. Kuncheva, ""fuzzy" versus "nonfuzzy" in combining classifiers designed by boosting," *IEEE Transactions on fuzzy systems*, vol. 11, no. 6, pp. 729–741, 2003.

[29] T. Gupta, S. Garg, A. Mahanti, N. Carlsson, and M. Arlitt, "Characterization of friendfeed - a web-based social aggregation service," in *International AAAI Conference on Weblogs and Social Media*, 2009.

[30] F. Celli, F. Di Lascio, M. Magnani, B. Pacelli, and L. Rossi, "Social network data and practices: The case of friendfeed," *Advances in Social Computing*, pp. 346–353, 2010.

[31] S. Garg, T. Gupta, N. Carlsson, and A. Mahanti, "Evolution of an online social aggregation network: an empirical study," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, 2009, pp. 315–321.

[32] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[33] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[34] S. Kullback, *Information theory and statistics*. Courier Dover Publications, 1968.

[35] B. Bigi, "Using kullback-leibler distance for text categorization," in *Advances in Information Retrieval*, 2003, pp. 305–319.

[36] X. Zhou, X. Zhang, and X. Hu, "Semantic smoothing of document models for agglomerative clustering," in *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2007, pp. 2928–2933.

[37] D. Metzler, S. Dumais, and C. Meek, "Similarity measures for short segments of text," in *Advances in Information Retrieval*, 2007, pp. 16–27.

[38] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, pp. 84–90.

[39] J. E. Gubernatis and T. E. Booth, "Multiple extremal eigenpairs by the power method," *Journal of Computational Physics*, vol. 227, no. 19, pp. 8508–8522, 2008.

[40] M. F. Porter, "An algorithm for suffix stripping," *Program: Electronic Library and Information Systems*, vol. 14, no. 3, pp. 130–137, 1980.

[41] C. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[42] F. Chung-Hoon Rhee, "Uncertain fuzzy clustering: Insights and recommendations," *IEEE Computational Intelligence Magazine*, vol. 2, no. 1, pp. 44–56, 2007.

[43] G. Stegmayer, M. Gerard, and D. H. Milone, "Data mining over biological datasets: an integrated approach based on computational intelligence," *IEEE Computational Intelligence Magazine*, vol. 7, no. 4, pp. 22–34, 2012.

[44] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz, "Towards recency ranking in web search," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 2010, pp. 11–20.