PAPEER: Bringing Social Networks into Research

Davide Frey
INRIA Rennes, France
davide.frey@irisa.fr

Anne-Marie Kermarrec
INRIA Rennes, France
Anne-Marie.Kermarrec@inria.fr

Vincent Leroy
INSA de Rennes, France
vincent.leroy@irisa.fr

ABSTRACT

We propose PAPEER, a novel user-centric gossip-based paper indexing platform, through which users can search for scientific papers, share them with their collaborators as well as find new people to collaborate with on new research topics. Different from web-based tagging platforms, PAPEER's decentralized architecture makes locally stored tags and content available to other users with similar research interests, through a personalized interest-based social network. The same social network constitutes the basis for PAPEER's ability to recommend the coworkers who are most interested in a given paper, or who are most likely to be interested in a given paper or topic. We validate PAPEER's recommendation approach by means of an experimental evaluation on a Delicious data trace with 13000 users.

1. INTRODUCTION

PAPEER is a novel application designed to improve productivity by boosting the collaborative nature of the work revolving around research papers. PAPEER provides a natural interface to search for scientific papers not only in the usual repositories normally accessible through a Google query, but also in the shared repositories of colleagues and other researchers with similar research interests.

Consider the following example. John has a strong research record in peer-to-peer distributed systems. Recently, however, he has become interested in the emerging domain of social networks and is willing to apply his expertise to new challenging problems in this area. As a first step, John needs to retrieve relevant papers and find links to relevant conferences. A google search will certainly help. However, John is going to be able to access more result faster as soon as PAPEER starts to identifying John's new interest and *connecting* him with people who have been working on social networks, while giving preference to those that have the strongest connection with his previous background on peer-to-peer.

As John reads papers and finds interesting ideas, he starts to think he should share some of these papers with current and past coworkers who may also be interested in them. Given John's long curriculum, he knows a large number of potentially interested people, but he doesn't want to spam all of them! Again, PAPEER is able to help John by choosing people from his past coworkers and contacts who are most likely to be interested in the proposed papers.

The ability to share and access information in a completely peer-to-peer fashion constitutes the basis for the features of PAPEER that make the above scenario possible. People using PAPEER can in fact associate tags and or comments with each paper they access. PAPEER uses these tags to automatically organize papers into a folksonomy, which serves as a basis for (i) retrieving new papers on a given topic, (ii) finding new people who are interested in a given paper or topic, (iii) *recommending* people who are likely to be interested in a given paper or topic.

The design of PAPEER builds on our previous experience in the context of GOSSPLE and gossip-based protocols. The use of a gossip-based approach to maintain this dynamic overlay gives PAPEER the ability to naturally adapt to highly dynamic environments resulting from dynamic interests, new activities, and churn, thereby supporting roaming between different networks as well as disconnected operation. Our preliminary evaluation of PAPEER based on a 13,000-user Delicious data trace confirms its ability to provide highly relevant user recommendations, and motivates our current development of a full-fledged prototype implementation.

The paper is structured as follows. Section 2 presents an overview of PAPEER's features. Section 3 details PAPEER's user recommendation mechanism. Section 4 presents a preliminary evaluation by simulation. Section 5 places our work in the context of related efforts, while Section 6 concludes the paper by presenting our future directions.

2. PAPEER

In this section we present PAPEER's interface from a user's perspective, namely its main concepts as well as the operations it supports.

2.1 Personal Library

The personal library constitutes the central abstraction of PAPEER. Essentially, it consists of entries $< p, t, c, f >$, where p is a set of metadata representing a paper, t is a set of tags, or keywords, that the user has assigned to the paper, c is a comment the user associated with the paper, and f is a (link to a) local copy of the paper.

2.2 Implicit vs Explicit Social Networks

The main feature of PAPEER is the ability to enable collaborative work among the users in the same social network. PAPEER's notion of social network comprises both an *explicit* and an *implicit* network. The former contains references to other users who have been explicitly selected by the user (e.g. ala facebook friends) or that have been automatically extracted from available data because they are tied to the user by real-world relationships (e.g. coauthors). The latter, on the other hand, contains references to users who are not in the explicit network but whose profiles are particularly close to that of the user. This implicit network is particularly important when searching for new information [3].

2.3 Navigating PAPEER's Libraries

By combining the personal library with the social network abstraction, PAPEER is able to automate the most common activities revolving around research papers. This includes retrieving a paper matching a given query, or obtaining information about other users who may be interested in the same paper. To make this possible PAPEER offers a set of search facilities that operate on a user's *federated library*, that is on the union of the personal libraries of the users in the same social network. Depending on the search operation, and on the user preferences, the federated library may comprise the personal libraries in the explicit social network, those in the implicit one, or both.

Querying for a Paper.

The most basic operation supported by PAPEER is to search for a given paper in the federated library, either by title or keywords. Upon finding a paper, a user can immediately add it to his/her personal library, optionally downloading a copy and printing it.

Tagging and Commenting on a Paper.

A very important part of the operation of PAPEER is the ability to associate tags and comments with papers in the personal library, thereby supporting its collaborative nature. Tags may be keywords associated with the paper, but, most importantly, they may also be nicknames given to papers within a particular research group. For example, within our team, we often refer to [2] as the Yahoo paper, while often find it hard to remember its exact title. Nicknames such as this make sense in PAPEER exactly because of the personalized nature of its social network.

Navigating from papers to people.

Even though searching and tagging are important features of PAPEER, its most distinctive feature is the ability to exploit and strengthen the connection between users sharing the same research interests. After searching for a paper in the federated library, PAPEER offers the opportunity to access two lists of users. The first is the list of users who have the paper in their own personal libraries. The second contains instead users that are likely to be interested in the paper.

Different from a basic recommendation system recommending papers to users, PAPEER has the ability to discover which users are potentially interested in a given paper, by observing their past tagging behaviors. In our example scenario, this allows John to forward interesting social network papers only to the set of colleagues that are interested in areas that are closely related to social networks.

Navigating from keywords to people.

Finally, PAPEER also offers the possibility to associate directly people and research topics, in the form of tags. Users can thus type in keywords representing research topics, such as "distributed systems" or "social networks" and be prompted with a list of users who are related with such topics.

3. USER RECOMMENDATION

PAPEER is able to exploit and strengthen the connections between people sharing the same research interests. In the following, we concentrate on its most distinctive features. First, we describe how PAPEER maintains its social network and then how it is able to exploit it to recommend the right users who should read a given paper.

3.1 Maintaining The Social Networks

To maintain the implicit social network, we use the approach described in [4]. Each peer relies on a gossip-based clustering protocol deployed over a random peer sampling protocol, itself gossip-based [7]. A gossip protocol consists in a periodic exchange of information between pairs of nodes. In the case of peer sampling and clustering, this information consists of a *view* of a subset of the network.

The random peer-sampling provides each node with a random, continuously changing, view of the network. The clustering protocol uses this constantly changing view to construct the personalized implicit social network of the node. To achieve this, each time a node n encounters another node q, it has to evaluate its *similarity* with q in order to assess q's eligibility to belong to its own implicit social network.

Cosine similarity.

In PAPEER, our similarity metric consists of the widely used *cosine similarity.* Similar to what is done in [4], we define the *paper cosine similarity* between two users by taking into account the set of papers they tagged.

$$PaperCos(u_1, u_2) = \frac{|Paper(\{u_1\}) \bigcap Paper(\{u_2\})|}{\sqrt{|Paper(\{u_1\})| \times |Paper(\{u_2\})|}}$$

3.2 Choosing the Right Users for a Paper

Most existing recommendation systems focus on recommending items to users by selecting what items would be good for a given user based on the user's past preferences. In the context of research however, we are often interested in finding the right people for a given topic. Locating which people are likely to be interested in a given paper, or project is therefore key to fostering collaboration, while avoiding spamming uninterested people.

Proximity-based recommendation.

The personalized nature of PAPEER's social networks already provides a basis for choosing the right users to recommend a given paper to. If John decides to forward an interesting paper, he is likely to forward it to people close to him in terms of their research interests. However, forwarding even a very interesting paper to all of his contacts would probably be almost indistinguishable from spamming. Then, how to best select the right subset of users for a given paper? We answer this question by proposing a proximity-based recommendation protocol that builds on the notions of personalized network and similarity defined above.

Given a paper p and a user u, let P be the set of profiles of the users in u's social network who have this paper in their personal libraries and who are therefore known to be interested in the paper. Also, let S represent, the explicit or the implicit social network being considered. The goal is to find the set of users in $S \backslash P$ who are most likely to share an interest in paper p. To achieve this, we compute, for each user $u \in S \backslash P$ the average of the similarities between u and each of the profiles in P. The users such that this average is larger than a threshold δ are considered to be potentially interested in p.

To compute the value of the threshold δ, we introduce the notion of *set-similarity* for a set of user profiles, P. For each profile, $p_i \in P$, let $\overline{p_i}$ be the average of the similarities between p_i and each of the profiles in $P \setminus p_i$, ignoring the tagging information associated with paper p. Then the *set-similarity*, $\sigma(P)$, of the set P is the average of these $\overline{p_i}$ values. Based on this definition, we then use a threshold $\delta = k \cdot \sigma(P)$, where $k > 0$ is the *threshold multiplier.*

4. PRELIMINARY EVALUATION

We evaluate PAPEER's *user recommendation ability* by means of simulations using a real data trace with 13000 users crawled from the social bookmarking system Delicious in January 2009.

4.1 Experimental Setting

Each experiment starts by selecting a random user from the data trace, and building a personalized social network for this user. For simplicity, we consider only the implicit social network in this evaluation. To this end, we select the set of N users whose profiles are closest to the one of the selected user according to the PaperCosineSimilarity as defined in Section 3.1. This effectively reproduces the social network built through the gossip-based clustering protocol described in Section 3.1.

Once this process has been completed, we select a random item from those tagged by at least $u = 20$ users in the constructed social network and randomly select 30% of its taggers as recommendation targets after removing their corresponding tagging information. The goal of the experiment is to recommend the item to as many recommendation targets as possible based on the information associated with the remaining 70% of taggers.

Metrics and baseline.

Our main metric to evaluate the quality of PAPEER's recommendation is *recall. Recall* expresses the ability of the user recommendation system to recommend the paper to the users in the paper's recommendation target. The value of recall, is therefore the ratio between the number of recommended users who are in the dissemination target, and the

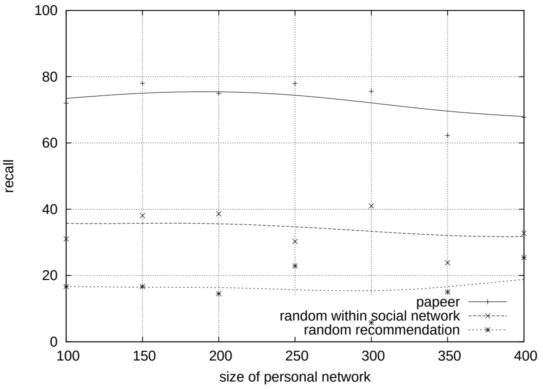

Figure 1: Recall with personal networks of increasing size.

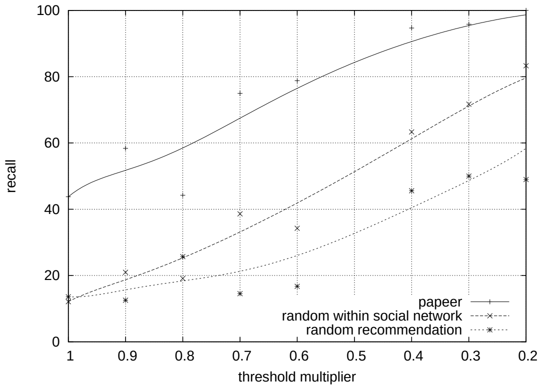

Figure 2: Recall with a personal networks of 200 entries and with decreasing threshold values.

total number of users in the dissemination target.

As a second metric, we also consider a lower bound on *precision* defined as the ratio between the number of recommended users who are in the dissemination target and the total number of recommended users.

To evaluate the quality of the results obtained by PAPEER's user recommendation strategy, we compare them against those obtained by a random recommendation process (among all users in the system) and a random recommendation process operating on the social network (among the users in the personalized network).[1] More specifically, in each experiment we first extract the set R of users chosen by PAPEER. Then we randomly extract $|R|$ users from the entire network as well as $|R|$ users from the social network and we compare the performances of these two random recommendation processes against that of PAPEER.

4.2 Results

Our results are depicted in Figures 1 and 2. Figure 1 shows the values of recall obtained by PAPEER with a threshold multiplier value of 0.7 and by the two random recommendation processes with implicit social networks of increasing size.

Results clearly show the importance of both the personalization of the social network and of PAPEER's user recommendation strategy. A completely random recommendation process is in fact only able to recommend the considered paper to less than 20% of the users in the recommendation target set. A random extraction from the social network improves this figure bringing recall values to almost 40%. This demonstrates the impact of personalization of the social network. Both random approaches, however, are significantly outperformed by PAPEER, which is able to recommend the considered papers to over 70% of the users in the recommendation target. This illustrates the impact of PAPEER's user recommendation strategy. Finally, the figure shows that the performance of both personalized approaches appears to decrease slightly if the social network becomes too large. This is because too large a social network inevitably ends up including irrelevant users whose profiles are not necessarily related with one another, thereby negatively impacting the quality of the recommendation results. This illustrates further the impact of personalization of the social network.

Precision figures are also in favor of PAPEER which scores an average of 10% against the 2% of the completely random approach and the 4% of a random recommendation from the social network. As mentioned above, these low absolute values have to be considered as lower bounds. A correct evaluation of the recommendation system's precision would require a complete user study and is outside the scope of this paper.

Figure 2, on the other hand, presents the impact of the choice of the recommendation threshold on the results. As expected, recall values increase as we decrease the value of the threshold multiplier. However, this obviously comes at the cost of a decrease in precision, which goes, in the case of PAPEER, from values around 10% with multipliers larger than 0.7 to values around 4% with a multiplier value of 0.2. This decrease in precision is also evident if we observe the decrease in the performance difference between PAPEER and the random strategies as

[1] Note that a random selection among the users in the personalized network intuitively could already lead to reasonable results since those users are likely to have similar interests.

the multiplier value decreases. A smaller threshold causes, in fact, Papeer to recommend the considered paper to a larger number of users, and thus increases the number of users selected by the random approaches, making it more likely for them to pick users within the recommendation target.

5. RELATED WORK

Social networking and collaborative tagging systems have attracted the attention of researchers, engineers and users, thanks to their flexibility and ease of use. Websites like Delicious, Flickr, or LastFM allow users to tag URLs, pictures, or music, and support search operation as well as recommendation features. Bibliography management systems like citeseerx, citeulike, or Bibsonomy [1] have also joined this trend offering features like collaborative tagging, or recommendation. However, different from Papeer, all of these systems are characterized by a web-based architecture, which makes it difficult to support the personalized perspective — e.g. with tags that are only meaningful in a local setting — offered by Papeer.

In addition, the recommendation features offered by these systems almost always address the recommendation of items to users. Papeer, on the other hand, takes a complementary approach and focuses on identifying which other users may be *recommendable* for a given paper, thus supporting search for collaborators and other user-centric scenarios.

Alongside web-based systems, a number of peer-to-peer solutions have also addressed the management of bibliographic data. However, these almost always fall short of their web-based counterparts. Bibster [6] is peer-to-peer system designed to support the exchange of bibliographic information among researchers, while Edutella [8] is a similar system supporting the exchange of educational material. Albeit similar to Papeer, neither system supports any form of personalization or recommendation features. In addition, Bibster is ontology-based in contrast to the highly dynamic folksonomy-oriented approach adopted in Papeer. PDBMS [5] follows a peer-to-peer social-network-oriented approach similar to ours, although its architecture is based on a Distributed Hash Table. However, it also does not support any form of recommendation. To the best of our knowledge, Papeer is thus the first system integrating collaborative tagging and a user-centric perspective with novel user recommendation features, in the context of bibliographic management systems.

6. CONCLUSIONS AND DIRECTIONS

We presented Papeer, a user-centric, gossip-based bibliography-management application. Based on the combination of explicit and implicit social networks, Papeer provides a personalized perspective to tagging and querying, and supports recommendation features such as the ability to identify which users are most likely to be interested in a given paper or topic.

We evaluated paper's user recommendation system on a 13000-user data trace crawled from Delicious. Results show that the combination of a personalized social network with Papeer's similarity-based recommendation system are able to provide good recall values and motivate further experiments.

In this respect, we are currently working on an open-source implementation of Papeer, which will provide real-world data to fine tune Papeer's recommendation process. In addition, we are also evaluating the extension of Papeer with different approaches to recommendation, based on tag similarity, and on most-frequent item sets. Finally, we plan to integrate techniques to deal with privacy and potential attacks.

7. REFERENCES

[1] www.bibsonomy.org.

[2] S. Amer-Yahia, M. Benedikt, L. Lakshmanan, and J. Stoyanovich. Efficient network aware search in collaborative tagging sites. In *Proc. Very Large Data Bases (VLDB'08)*, volume 1, pages 710–721, 2008.

[3] M. Bender, T. Crecelius, M. Kacimi, S. Miche, J. Xavier Parreira, and G. Weikum. Peer-to-peer information search: Semantic, social, or spiritual? *Bulletin of Computer Society Technical Committee on Data Engineering*, 2007.

[4] M. Bertier, R. Guerraoui, A.-M. Kermarrec, and V. Leroy. Toward Personalized Query Expansion. In *Social Network Systems 2009*, Nuremberg Germany, 2009.

[5] A. Datta. A p2p bibliographic content management system: Realizing decentralized infrastructure for social softwares. Technical report, Nanyang Technological University, Singapore.

[6] P. Haase, B. Schnizler, J. Broekstra, M. Ehrig, F. van Harmelen, M. Menken, P. Mika, M. Plechawski, P. Pyszlak, and R. Siebes. Bibster–a semantics-based bibliographic peer-to-peer system. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):99–103, December 2004.

[7] M. Jelasity, R. Guerraoui, A.-M. Kermarrec, and M. van Steen. The peer sampling service: experimental evaluation of unstructured gossip-based implementations. In *Proc. of the 5th international conference on Middleware*, pages 79–98, New York, NY, USA, 2004. ACM/IFIP/USENIX.

[8] W. Nejdl, B. Wolf, C. Qu, S. Decker, M. Sintek, A. Naeve, M. Nilsson, M. Palm&#233;r, and T. Risch. Edutella: a p2p networking infrastructure based on rdf. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 604–615, New York, NY, USA, 2002. ACM Press.