Improving Search In Social Networks by Agent Based Mining*

Anıl Gürsel
Department of Computer Science
University of Tulsa
anilgursel@gmail.com

Sandip Sen
Department of Computer Science
University of Tulsa
sandip@utulsa.edu

Abstract

The popularity of social networks have burgeoned in recent years. Users share and access large volumes of information on social networking sites like Facebook, Flickr, del.icio.us, etc. Whereas a few of these sites have generic, impersonal searching mechanisms, we have developed an agent-based framework that mines the social network of a user to improve search results. Our Social Network-based Item Search (SNIS) system uses agents that utilize the connections of a user in the social network to facilitate the search for items of interest. Our approach generates targeted search results that can improve the precision of the result returned from a user's query. We have implemented the SNIS agent-based framework in Flickr, a photo-sharing social network, for searching for photos by using tag lists as search queries. We discuss the architecture of SNIS, motivate the searching scheme used, and demonstrate the effectiveness of the SNIS approach by presenting results. We also show how SNIS can be utilized for expertise location.

1 Introduction

Recently, there has been a rapid increase in the number of users signing up and actively using social networking websites such as *Facebook, Flickr, MySpace*, etc. For example, during peak periods, photos viewed on *Flickr* reaches up to 12,000 per second. With many-fold increase in the number of subscribers and items posted/shared by them, users are confronted with an abundance of information and options, which can lead to a delay in locating items of interest. Search facilities provided by social networking sites are almost always impersonal, i.e., given a search query, the same results are reported for all users. As few users access results shown beyond the first couple of pages, it is imperative that the ranking of the results is key to user satisfaction. In the context of social networks, it is particularly critical to personalize the search process. We believe that using the topology and neighborhood of a user in the search process, e.g., by focusing on items posted by friends, can improve user satisfaction and hence the precision of the search process.

Aimeur and Onana [Aimeur and Onana, 2006] allow users to limit the recommendation collection process to a set of manually selected contacts and to assign trust values to each such contact. Recommendations from manually selected contacts were shown to be better than the ones made by traditional collaborative filtering. In addition, previous research [Lerman, 2006] has shown that people tend to like items that their friends like and are interested in the activities of acquaintances compared to that of people they do not know. User-designated friends in social networks, therefore, can be reliable sources of recommendations.

We seek to develop a searching mechanism for social networks that will cater to the preferences of a user by tracking indirect past ratings of that user. The goal is to rank results of search queries to highlight recently posted items by friends and peers in the social network that will be of particular interest to a user. To rank user preferences from past ratings, we identify topic-specific preferences of the user for his/her friends' items, i.e., we allow a user to have different preferences for the items corresponding to different topics posted by a given friend. For example, a user may like photos of nature but not like photos of pets posted by a friend. Such topic-specific item preference is mined by a combination of a Naive Bayes learning scheme and by elaboration on tag lists associated with items by using folksonomies.

To aid the learning of topic-specific user preferences, we use a category identification process that utilizes textual content. In particular, we focus on tags, which can be used by users to describe the content of items such as photos and videos. To test our proposed approach, we designed and implemented an agent-based photo searching system for *flickr.com*. The *Flickr* online service allows users to share and tag photos as well as comment on photos of other users. We mined the history of comments written by users for learning user preferences. Comments signal user interest in photos. We believe that commenting on a photo correspond to user interest, either positive or negative, in that photo. We do not analyze the contents of the comments to refine searches as a large majority of the comments on *Flickr* are found to be positive in nature.

Related to the issue of finding the right information is the problem of finding the right person who possesses the re-

*A DOD-ARO Grant #W911NF-05-1-0285 partially supported this work.

quired knowledge. For example, consider a scenario that a person travels to a city and is looking for an ethnic restaurant that he will like. Though restaurant recommendations abound on the internet, these ratings are from anonymous people. Since taste changes from person to person, it is important to get suggestions from somebody that has the information and is trusted on this issue by the user. No user in the immediate neighborhood, however, may have the requisite expertise, and, hence, the search might have to be expanded. Because of performance constraints, the search needs to be limited to some neighborhood in the network. Research on expertise locator mechanisms mostly assume that users provide their expertise levels. Moreover, they do not consider the strength of the connections between the person seeking assistance and the expert. Social networking websites with pre-built links between users can facilitate location of trusted experts. We show, with experiments on flickr.com data, that our recommendation mechanism, which calculates topic-based confidence between users, can be utilized for harnessing information in social networks to develop expertise locator systems.

2 Related Work

Searching for information in the infosphere has become an essential service that we all rely on. Many search tools and recommender mechanisms have been developed to locate information on the Internet. Recommender systems typically suggest information items (movies, music, books, blogs, news images, web pages) that are likely to be of interest to the user. E-commerce sites, for example, use recommender systems to help their customers find products and services that might be of interest to the customer [Schafer *et al.*, 1999].

Given the popularity of social networks and their acceptance and innovative use by the general populace, improved search and location schemes will have a significant impact on our level of satisfaction for using these services. A social network connects a set of entities via social relationships and facilitates information exchange [Jamali and Abolhassani, 2006]. In literature, social networks are typically represented by *nodes* and *edges*, which correspond to individuals (or actors) and relationships among them, respectively.

A successful social network-based recommendation system will need to identify the influential nodes. Though effective mechanisms have been developed to characterize the importance of nodes in a network [Brin and Page, 1998], these algorithms typically calculate the global importance of a node. For searching and recommendation of items in a social network, the importance of nodes relative to a specific user will be of primary interest than its global measure. Sinha and Swearingen [Sinha and Swearingen, 2001] found that the quality of recommendations from friends are better than that from traditional recommender systems. Lerman analyses the popular social news aggregator $Digg$ [Lerman, 2006] and demonstrates that users like stories submitted by friends. $Flickr$ allows searching images by tags and can also rank search results by their interestingness values. Lerman *et al.* show that the precision value of returned images increase if the results are filtered so that only the images from contacts, or contacts' contacts, are returned [Lerman *et al.*, 2007].

Social network structures can be used to form average ratings of an item from one's friends instead of averaging ratings from all users. FilmTrust [Golbeck and Hendler, 2006] lets users assign a trust value for each friend, and then computes the weighted average rating of a movie.

An important observation from the above applications is that the source of recommendation is important to a user. Most recommendation systems are black boxes where information about the recommendation process is not revealed. Users, however, prefer transparency and explanations about how the recommendations were derived. Moreover, recommendations from people with similar profiles or known users are preferred over those from others. In social networks, users self-select other users that they know and trust. The key insight for our work is that a user agent can determine, by mining past user activities in the social network, the likelihood that the user will like an item given the query which the item matched and the source of the item.

Social webs provide a fruitful medium for expertise identification. An individual, however, might only be a few steps away from a needed expert but fail to realize it [Adamic and Adar, 2005]. Expertise locator systems allow a user to be aware of one's friends' friends and so on by expanding the social circle when information is needed. The main concern is the time efficiency of the underlying algorithm for this search process. A social network usually contains hundreds of thousands of nodes and millions of edges, yet the response time to a search should be within a few seconds.

Yu and Singh [Yu and Singh, 2003] propose an approach to finding an expert in social networks where an agent learns its user's profile and its acquaintance models based on an evaluation of answers and the referrals that led to them. ReferralWeb [Kautz *et al.*, 1997] uses search engines to find individuals with a given expertise and identifies social relationships by co-occurrence of names in close proximity in any document publicly available on the net.

Users cannot be expected to completely reveal their expertise and especially knowledge because of privacy issues. In addition, people will be reluctant to help people they do not know, and mentioning expertise might be a burden for average users. The expertise of people, however, can be derived from their activities in social networks, e.g., uploads, comments, ratings, contacts. We believe that if each node in a social network can locally identify the most related person for a query, then the person with the required knowledge can be quickly identified.

3 Architecture

3.1 Agent based mining of social networks

In our approach, each user has an associated SNIS agent. This agent observes the user's activities and, in particular, the ratings and comments provided by the user to items retrieved from the social network. For the Flickr domain, the SNIS system scans photos posted by all of the user's contacts and gathers statistics about their categories and user comments (which represent user interest).

In the offline mining process, first, the likelihood of a photo belonging to each category is determined. Each item has an associated ordered list of tags. We believe that not all tags attached to an item are of equal importance. The higher the rank of a tag in the list associated with an item, the more effect on the category identification it is expected to have. Hence, we weigh each tag according to its position in the list. Let a tag t_j^i be in the j^{th} position of the list i. Then the weight of t_j^i on the category determination algorithm is calculated with a decreasing function of j. Given a set of tags for an item i, $tags(i)$, we calculate the probability of an item i belonging to any category according to the number of tags in $tags(i)$ that are also included in the dictionary for that category.

The SNIS agent calculates the preference of an user by computing probabilities of the form $Pr(\ likes(u_a, i) \mid i \in posted(u_b))$, which corresponds to the probability of an item i that is posted by u_b being liked by the target user u_a. These probabilities are then used to order search results. $Pr(\ likes(u_a, i) \mid i \in posted(u_b))$ is calculated by applying the Bayes rule, which necessitates the calculation of, along with other terms, (a) the probability of an item, belonging to a particular category and liked by user u_a, being posted by user u_b and (b) the probability of user u_a liking an item belonging to a particular category. The learning process used by SNIS agents mines the social network activities of the user to approximate these probabilities.

When the user presents a search query as a list of tags, the associated SNIS agent first retrieves a set of matching items from the items posted by users in the vicinity of this user. These items are ranked using the probabilities mentioned above, and the top few items are then recommended to the user.

3.2 Modules of SNIS

The SNIS agent architecture has the following modules:
Preference Learner: This module mines past activities of the user in his/her social network to identify the user preferences for items of different topics/categories posted by the user's contacts.
Query Processor: This module takes a list of tags as a search query and then performs a directed search in the social network. The search proceeds until either a predetermined number of items have been returned or a depth cutoff in the network is reached.
Category Identifier: This module determines the likelihoods of the given item belonging to different categories. This determination is based on the tags associated with the returned item. A Naive Bayes approach is used in the category determination process and the corresponding probabilities are mined offline.

Given a list of tags associated with an item i, $tags(i)$, the probability of an item i belonging to each category is calculated based on the number of tags in $tags(i)$ that are also included in the dictionary for that category. Let $lookup(d, t)$ be a predicate that returns 0 or 1 depending on whether the tag t is included in dictionary d or not.

C is the set of categories. The probability of item i belonging to a category c_x is calculated as

$$Pr(cat_i = c_x) = \frac{\sum_{k=1}^{l_i} w(t_k^i) * lookup(d_{c_x}, t_k^i)}{\sum_{c_y \in C} \sum_{k=1}^{l_i} w(t_k^i) * lookup(d_{c_y}, t_k^i)} \quad (1)$$

where cat_i refers to the category of i. The weight of a tag t_j^i, situated in the j^{th} position of the list $tags(i)$, $w(t_j^i)$, as used by the category determination algorithm is calculated using a decreasing function of j that can be tuned for each domain.
Recommendation/Ranking: Given the likelihoods of a matching item belonging to different categories, the mined preferences of the user for these categories as well as for the user who posted this photo are used to determine the likelihood that the user will like that item. All items in the list returned by the search process are evaluated and the list is ranked by the corresponding probability values.

We calculate the probability $Pr(\ likes(u_a, i) \mid i \in posted(u_b)\ )$ by using Bayes Theorem:

$$\frac{Pr(\ i \in posted(u_b) \mid likes(u_a, i)\ )\ Pr(\ likes(u_a, i)\ )}{Pr(\ i \in posted(u_b)\ )} \quad (2)$$

We now derive the probabilities in the expression above except $Pr(\ likes(u_a, i)\ )$ which will be eliminated in the following steps. We assume that the preference for an item depends both on the owner and the content of the item. $Pr(i \in posted(u_b) \mid likes(u_a, i))$, therefore, can be expanded as

$$\sum_{c_x \in C} [Pr(i \in posted(u_b) \mid cat_i = c_x, likes(u_a, i)) \; Pr(cat_i = c_x \mid likes(u_a, i))] \quad (3)$$

We can rewrite Equation 3 by applying Bayes rule to the second conditional probability

$$\sum_{c_x \in C} [Pr(i \in posted(u_b) \mid cat_i = c_x, likes(u_a, i)) \; \frac{Pr(likes(u_a, i) \mid cat_i = c_x)\ Pr(cat_i = c_x)}{Pr(likes(u_a, i))}] \quad (4)$$

The probability of an item i posted by a user also depends on the content, since a user's posting habits might be biased towards some categories. So, the denominator of Equation 2, $Pr(i \in posted(u_b))$, can be written as follows:

$$\sum_{c_x \in C} [Pr(i \in posted(u_b) \mid cat_i = c_x)\ Pr(cat_i = c_x)]. \quad (5)$$

After substituting 3, 4, and 5, and simplifying, $Pr(i \in u_b \mid likes(u_a, i))$ can be written as:

$$\frac{\sum_{c_x \in C} [Pr(i \in posted(u_b) \mid cat_i = c_x, likes(u_a, i)) \; Pr(likes(u_a, i) \mid cat_i = c_x)\ Pr(cat_i = c_x)]}{\sum_{c_x \in C} Pr(i \in posted(u_b) \mid cat_i = c_x)\ Pr(cat_i = c_x)} \quad (6)$$

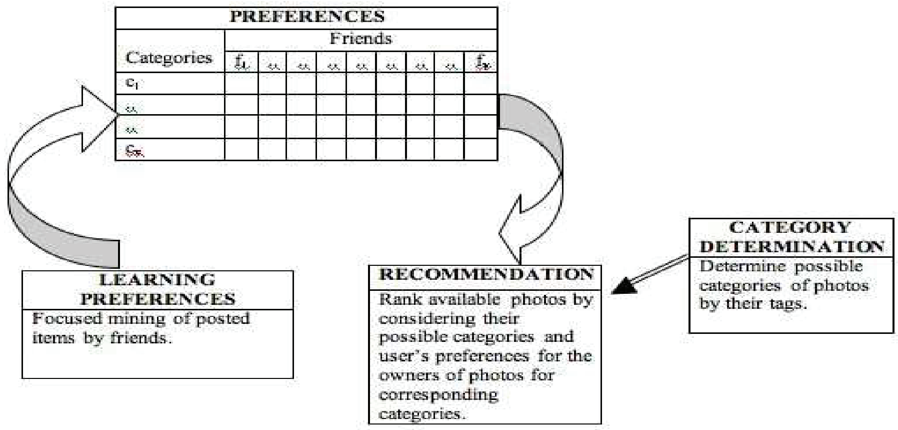

Figure 1: The SNIS process of refining search results to recommend items of likely interest to the user.

The unknown probabilities in 6 are computed as follows:

• $Pr(i \in posted(u_b) \mid cat_i = c_x, likes(u_a, i)) =$ (Number of photos posted by u_b that belong to c_x and are commented on by u_a) / (Number of all photos that are in c_x and are commented on by u_a).

• $Pr(likes(u_a, i) \mid cat_i = c_x) =$ (Number of all photos that are in c_x and are commented on by u_a) / (Number of all photos that are in c_x).

• $Pr(i \in posted(u_b) \mid cat_i = c_x) =$ (Number of photos posted by u_b that are in c_x) / (Number of all photos that are in c_x).

Figure 1 shows the interaction between some of the modules in a SNIS agent as well as the representation of the preference information mined for a user.

4 Experimental Framework and Results

4.1 Flickr: Photos, friends, tags

We have evaluated the quality of our *SNIS* agent-based system on the popular photo sharing website *flickr.com (Flickr)*, which allows users to upload photos, tag photos with descriptive words, and write comments on them to express their opinion. Moreover, it also allows users to designate others as friends, thereby forming a social network. *Flickr* provides an API that allows us to write a program to gather information from others' accounts. For these reasons, the *Flickr* system has been studied by many researchers [Lerman *et al.*, 2007].

We randomly selected 10 root users from those listed on *Flickr*'s *interestingness* page ensuring that these users have posted a relatively higher number of comments. For each root user, we visited his/her friends' accounts and gathered information about their photos, e.g., tags and comments, that were uploaded or posted between *January 1, 2008* and *April 30, 2008*. Our database includes information of 4025 users, who have together posted 121953 photos and written 30040 comments. For our experiments, the data from the first 3 months, from *January 1, 2008* to *March 31, 2008*, is used for training the prediction system (mining the photo preferences of individual users), and the last month's data, from *April 1, 2008* to *April 30, 2008*, is used for testing the system. We have used 10 different photo categories. A photo in our system can belong to one or more of the following categories: *Animal, Art, City, Entertainment, Nature, News and Politics, People, Science and Technology, Sports, Travel and Places*. A more fine-grained set of categories can also be used since our system does not depend on a pre-determined, fixed, set categories. For each category, we built a dictionary that contains a set of category-related tags.

To form a dictionary for a category c, we first visited *Flickr*'s tag-based search to find photos that are related to c in content and added tags of photos in search results to our dictionary of c after filtering out extraneous and unrelated tags manually. We then queried this search tool with some common tags of photos returned from previous search results and repeated this process. After the dictionary c reaches a certain size, *Flickr*'s folksonomy (related tags search) is utilized to extend the dictionary and stabilize it. In this step, for each tag in the dictionary, related tags are collected. The related tags that have high count values, which is a measure of relatedness, are added to the dictionary. There are 1213 unique tags in our dictionary database and since tag clusters cannot be precisely separated, some tags might occur in more than one dictionary. A dictionary on average consists of 133 tags. Approximately 80% of photos in our database having at least 3 tags could be categorized by our approach. This percentage can be further improved if folksonomies are incorporated into the category determination process. Though folksonomies are utilized in SNIS for the formation of dictionaries, they are not used in the category determination process for extending the tag list in retrieved photos with related tags.

4.2 Analysis

We now state and verify the key hypothesis about user behavior in *Flickr* that justifies our approach:

Hypothesis 4.1 *A user has different preferences for the same friend based on item topics (photos in the case of* Flickr*).*

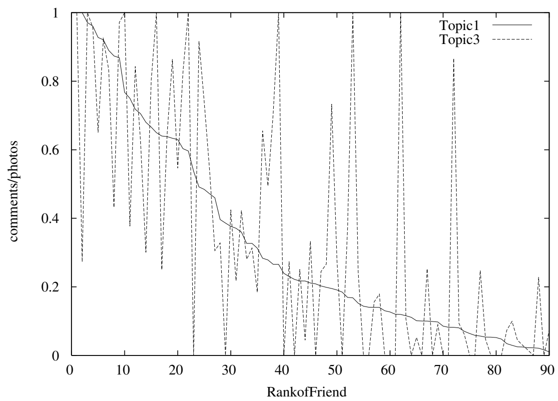

Figure 2: Number of comments per photo by a representative user on photos posted by friends on two different topics.

We present, in Figure 2, the percentage of photos of friends that are commented on by a representative *Flickr* user for two different topics. On the *x-axis*, the friends are placed in descending order based on percentage of commented photos belonging to Topic 1 by the target user. The figure tells us that the preference of a user for photos posted by a friend is different for different topics.

4.3 Performance Metrics

Given the set of Recommended and Favorite Items (items commented on by the user) the evaluation criteria used for our search mechanisms are two well-known metrics used in the *Information Retrieval* literature to measure the quality of retrieved information:

Precision is the fraction of recommended items that are liked by the user:

$$Precision = \frac{\mid Recommended\, Items \bigcap Favorite\, Items \mid}{\mid Recommended\, Items \mid}.$$

Recall is the fraction of liked items that are recommended by the recommender system:

$$Recall = \frac{\mid Recommended\, Items \bigcap Favorite\, Items \mid}{\mid Favorite\, Items \mid}.$$

4.4 Results

Lerman *et al.* [Lerman *et al.*, 2007] showed that search performance in *Flickr* can be improved by filtering results by user's contacts or a larger social network that includes those contact's contacts. We claim that search performance can be further improved if these filtered results are ranked according to the preferences of the user for his/her contacts. We use the ranking mechanism described in the previous section to rank the search results. To demonstrate the enhancement of our system, we conducted experiments with 5 different tag sets where 1, 3, and 5 tags are chosen from each tag set for querying, and present results averaged for 5 different users.

Figure 3 compares the performances of two searching mechanisms. In these experiments only the top 20 ranked items are returned from any search query. The first scheme, *s1*, filters the results by user's contacts while the second

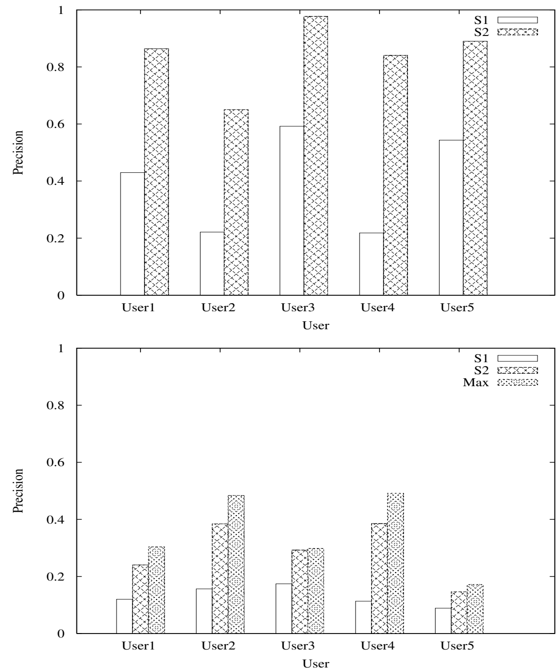

Figure 3: Precision (top) and recall (bottom) values of search results for systems *s1* and *s2*. *s1* restricts results to photos of contacts. *s2* restricts results to photos of contacts and ranks them according to preferences. *max* corresponds to the maximum possible recall values for 20 recommendations.

scheme, *s2*, further ranks the results according to the user's preferences for his/her contacts. The plot on top displays that *s2* enhances the precision values for the first 20 results dramatically. The bottom plot shows the corresponding recall values for this experiment. In this figure, *max* represents the maximum possible recall value that can be achieved by considering only 20 results. It can be seen that recall values for *s2* are very close to the maximum possible values. These results show that the SNIS agent-based search method can use user preference for specific users given search topics to further enhance the quality of the search results.

Expertise locator

To locate an expert on a category, it is inefficient to query every connected peer. Besides, sending a query to every other peer can pose other complications, e.g., this can cause a glut of requests which, in turn, can lead to experts ignoring pertinent requests. In addition, people are reluctant to answer queries from strangers [Kautz *et al.*, 1997]. Accordingly, for a query (list of tags), it can be useful to identify the set of peers who has posted matching items of interest If an expertise locator mechanism can locally identify the right set of peers to direct a query, the system can quickly retrieve relevant information. Social networks contain a rich repository of information to aid this process. In this experiment, we examine whether our system is able to identify right peers for a given set of tags. Thus, instead of filtering photos by contacts and then ranking them, as in search experiments, we first restrict the number of contacts to some constant *k*, fetch the

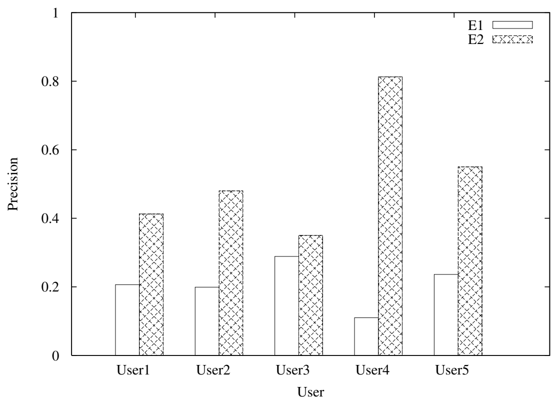

Figure 4: Locating expertise: search restricted to 10 contacts.

matching photos from these k users, and then rank them. The difficulty of this problem is identifying the k peers to contact. We use the probability values described in Section 3.2 to identify the k peers who are most likely to have photos of interest to our user given the search tags.

For these experiments, we use $k = 10$ and measure the precision performance of two systems: *E1*, where the peers are selected randomly, and *E2*, where peers are selected according to learned preferences of users for the topics of the given tags. For expertise locator systems, locating one person with the specific knowledge is important, rather than finding many people with some knowledge. We, therefore, focus on the precision values of *E1* and *E2*. We used 5 different queries with 5 tags for this experiment and averaged the precision results for 5 different users. Figure 4 shows that our system significantly improves the precision results over random contact selection. The average precision level reaches up to 0.8 for some users.

5 Discussion

We show that past user activities in social networks can be used to learn topic-specific user preferences and use a Social Network-based Item Search (SNIS) system to locate items in social networks. Our mechanism not only filters results by contacts [Lerman *et al.*, 2007] but also ranks them according to user preferences for the users who posted those items. Experiments demonstrate that both the precision and recall values of search results are dramatically enhanced compared to only filtering by contacts.

We hope to apply the SNIS approach to other social media sites such as *del.icio.us* and blog sites. To improve the robustness as well as performance of our system, the category dictionaries can be extended with more related tags. We can also consider a more detailed set of categories. It is also possible to use folksonomies in the category determination process. By extending the tag list of an item with related tags and weighting the tags appropriately might improve the quality of category determination process. On the other hand, it might also add some noise and degrade the category determination for some items. The effects of folksonomies on the category determination, therefore, need to be carefully analyzed. Moreover, not all tags in the list of an item have equal importance. Currently, our system gives more weight to the tags higher in the list. A smarter, adaptive weighting mechanism and a system that eliminates noisy tags can enhance the category determination scheme, which in turn, can produce better recommendation results.

One immediate extension to *SNIS* can be to provide recommendations from contacts further afar in the social network. By doing so, people might discover friends' friends that have very similar interests for some item types. It will also improve the expertise locator mechanism. Gathering information about many more users might also produce more reliable results for the existing experiments.

References

[Adamic and Adar, 2005] Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[Aimeur and Onana, 2006] Esma Aimeur and Flavien Serge Mani Onana. Better control on recommender systems. In *CEC-EEE '06: Proceedings of the The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services*, page 38, Washington, DC, USA, 2006. IEEE Computer Society.

[Brin and Page, 1998] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[Golbeck and Hendler, 2006] J. Golbeck and J. Hendler. Filmtrust: Movie recommendations using trust in web-based social networks. In *Proceedings of IEEE Consumer Communications and Networking Conference*, 2006.

[Jamali and Abolhassani, 2006] Mohsen Jamali and Hassan Abolhassani. Different aspects of social network analysis. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 66–72, Washington, DC, USA, 2006. IEEE Computer Society.

[Kautz *et al.*, 1997] Henry Kautz, Bart Selman, and Mehul Shah. Referral web: combining social networks and collaborative filtering. *Commun. ACM*, 40(3):63–65, 1997.

[Lerman *et al.*, 2007] K. Lerman, A. Plangrasopchok, and C. Wong. Personalizing results of image search on flickr. In *AAAI workshop on Intelligent Techniques for Web Personlization*, 2007.

[Lerman, 2006] K. Lerman. Social networks and social information filtering on digg. *ArXiv Computer Science e-prints*, December 2006.

[Schafer *et al.*, 1999] J. Ben Schafer, Joseph Konstan, and John Riedi. Recommender systems in e-commerce. In *EC '99: Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166, New York, NY, USA, 1999. ACM.

[Sinha and Swearingen, 2001] Rashmi R. Sinha and Kirsten Swearingen. Comparing recommendations made by online systems and friends. In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.

[Yu and Singh, 2003] Bin Yu and Munindar P. Singh. Searching social networks. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 65–72, New York, NY, USA, 2003. ACM Press.