Personalized Recommendation of Integrated Social Data across Social Networking Sites

Yuan Wang[1], Jie Zhang[2], and Julita Vassileva[1]

[1]Department of Computer Science
University of Saskatchewan, Canada
[2]School of Computer Engineering
Nanyang Technological University, Singapore
[1]{yuw193, jiv}@cs.usask.ca [2]{zhangj}@ntu.edu.sg

Abstract. We have developed a dashboard application called "SoC-Connect" for integrating social data from different social networking sites (e.g. Facebook, Twitter), which allows users to create personalized social and semantic contexts for their social data. Users can blend their friends across different social networking sites and group them in different ways. They can also rate friends and/or their activities as favourite, neutral or disliked. We compare the results of applying five different machine learning techniques on previously rated activities and friends to generate personalized recommendations for activities that may be interesting to each user. The results show that machine learning can be usefully applied in predicting the interest level of users in their social network activities, thus helping them deal with cognitive overload. A visualization technique that has been shown to work well in previous work is applied to display personalized recommendations.

1 Introduction

Social Networking Sites (SNSs) have changed how people communicate: nowadays, people spend more time on SNSs than ever, and prefer communication via SNSs over emails [1]. Despite the diversity of SNSs and the fact that social media enriches people's lives, current SNSs have the limitation of poor user data interoperability [2]. User-generated contents, users' online activities, and their friendships are scattered over different places. It becomes increasingly inconvenient for users to manage their social data and constantly check many sites to keep track of all recent updates. People may also keep different accounts on the same SNS in order to protect their privacy or other purposes. In addition, users are often overwhelmed by the huge amount of social data, especially friends' activities (status updates).

In this paper, we present an approach for recommending social activities in a dashboard application called "SoCConnect", described in [3], for integrating social data from different SNSs (e.g. Facebook, Twitter), which allows users to create personalized social and semantic contexts for their social data. More specifically, through SoCConnect, users can blend their friends across different

SNSs to become an "integrated" friend account in SocConnect. Users can create groups for their friends who may share some common features and do some activities together. In the current work, we add the functionality that allows users to rate friends and/or their activities as favourite, neutral or disliked. To relieve users' cognitive overload, we also apply different machine learning techniques to learn their preferences on activities based on their interactions with SCcConnect and to provide personalized recommendations of activities that are interesting to them. Evaluation results show the good performance of these techniques and especially good for some of them. A visualization technique developed in our previous work [4] is also used to display the personalized recommendations.

Section 2 presents the functionalities of SoCConnect. The approach for personalized recommendation of social networking activities is described in Section 3, followed by an experimentation in Section 4 to evaluate the performance. Related work on social data integration and recommendation is presented in Section 5. Finally, Section 6 summarizes the contributions of our work and discusses future research directions.

2 SoCConnect Dashboard

In this section, we provide a brief description about the functionalities of our dashboard application SoCConnect, the results of user studies supporting our design decisions for the functionalities, and the implementation of the system.

2.1 Functionalities

SoCConnect retrieves users' friends information and their activities on different SNSs. It provides three functional categories, "managing friends", "rating friends and activities", and "personalized recommendation of activities".

The first functional category, "managing friends" contains two functions: blending friends and grouping friends. In most cases, there is some level of overlap between the sets of a user's friends on different social networking sites. Our system allows the user to merge the different accounts of a friend across SNSs, to create a single "integrated" (or "blended") friend account for this friend in the user's SoCConnent dashboard. The second function is to group friends. Users can put their friends, both individual SNS accounts and "integrated" accounts, into groups. This function allows users to express the context and semantics of friendships, which could be the shared characteristics, interests or activities between friends.

The second functional category, "rating friends and activities" allows users to rate friends or friends' activities as favourite or disliked. The favourite activities are bookmarked, which can be revisited more easily. By rating, users are able to specify a semantic characteristic (currently limited to postive/negative) of their relationships with their friends and express their preferences on activities that they find more or less interesting and valuable.

The third functional category, "personalized recommendation of activities" recommends activities that may be interesting to users, making use of the previous ratings and the information about friend groups.

2.2 Motivation for these Functionalities

We conducted a user study to evalute our design decisions for the functionalities of SoCConnect. [1]A total number of 16 subjects (all students) were involved in this study, distributed over both gender and major (Computer Science or Non-CS). They were asked questions related to the functionalities during interviews. We provide here only the most relevant results.

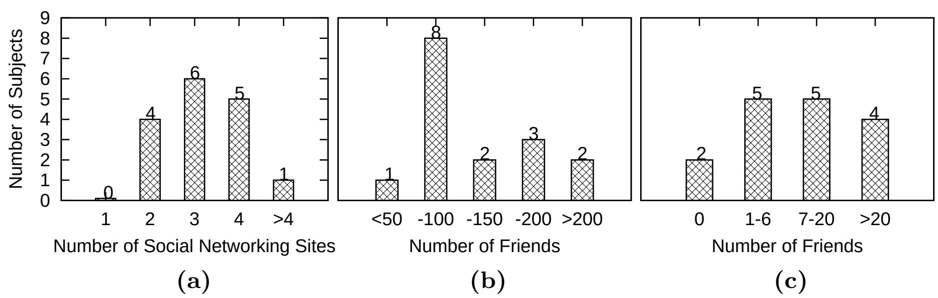

Fig. 1. (a) Number of Frequently Used SNSs; (b) Total Number of Friends on All SNSs; (c) Number of Friends Who have Accounts on More Than Two SNSs

All subjects have frequently used more than one SNSs (see Figure 1(a)). Most of them have frequently used more than two SNSs. Most of them also have more than 50 friends in total (see Figure 1(b)). Almost a half of the subjects have at least 100 friends. While it can be argued that this mini-study involved only students, this group presents the majority of users on most SNSs. For example, users of age 18-35 represented collectively 90% of the users on Facebook in 2008.[2] An e-business report from 2009 shows that 75% of the adults aged 18-25 have accounts on a SNS.[3]

The subjects were asked about the number of their friends who have user accounts on more than two SNSs. Only two subjects do not have such friends (see Figure 1(c)). More than a half of the subjects have at least 7 such friends. Several subjects (25% of all subjects) have more than 20 such friends. 81.25% of subjects answered that these friends were active on different sites and most of the friends have identical activities on these sites. 75% of subjects want to view these friends' activities in one place. These results confirm a strong need for the

[1] The study was approved by the Behavioural Ethics Research board of the University of Saskatchewan

[2] http://social-media-optimization.com/2008/05/social-network-user-demographics/

[3] http://emarketer.com/Article.aspx?R=1006882

function of blending friends. A significant majority (about 90%) of subjects have some friends who share similar interests, preferences or demographic information, or do some activities together. They want to create a group for these friends and include in groups some friends on different sites. These results support strongly our function of grouping friends.

In order to check if users would be willing to describe semantically their relationship and the content on their SNSs, we asked the subjects whether they want to tag friends and activities. Only half of them (54.25%) provided a positive answer. Tagging requires cognitive effort. The subjects were not sure whether they want to spend much time on tagging. Some subjects also feel that not many updates (friends' activities) are important. They prefer to tag only important activities or friends to revisit later. Instead of tagging with a word or a phrase, it requires less effort to mark an activity or friend as "favourite" or "disliked". Moreover, users of Twitter are familiar with this way of marking updates that they may wish to revisit later. This is why we provide the function of allowing users to add friends and activities as favourite or disliked, instead of a tag of any possible phrases.

The majority (68.75%) of the subjects said that they feel overwhelmed by the number of their friends' updates in one SNS. The number of updates will increase significantly when the friends' accounts across different SNSs are integrated by an application like SoCConnect. Thus, it is necessary to provide recommendations to help users navigate through their long list of friends' updates.

2.3 Semantics of SNS Data

To represent the semantics of social for generating recommendations, we design a generic ontology consisting of four main classes: SNS account (SNSAccount), integrated account (person), activity, and group. "SNSAccount" represents a user account on a SNS. "Person" represents a person who holds one or more SNS accounts. "Activity" represents generic information about activities appearing on SNSs. Each activity has a type. It can be a user update, e.g. a new friend added by the user, or an update by a third party application, e.g. a game such as FarmVille (farmville.com), and MafiaWars (mafia-wars.com) or other applications for Facebook, or specific clients (e.g. Tweetie, Twitdroid) or applications (e.g. Bit.ly) for Twitter. The activity may contain text and different types of media, such as pictures, videos and links. It may also have a target identifying the targeted user. "Group" represents a user-defined group for keeping friends together. A member of a group can be a SNSAccount or a Person.

3 Personalized Recommendations in SocConnect

One common problem of social networking site is information overload[4]. As indicated in our user studies, most of the activities from friends are not very

[4] http://www.stormdawg.com/2009/10/12/social-networking-and-information-overload/

important or interesting. Christian Kreutz in his blog calls this problem "network overload"[5].

everything in the list because as indicated in our user studies, most of the activities from friends are not very important or interesting. SocConnect aims to provide a personalized recommendation on activities to individual users according to a prediction generated using their preferences on previous social data. In this section, we will present a comparison of several machine learning techniques that can be used to predict users' preferences on activities and the approach selected for visualization of the personalized recommendations.

3.1 Learning User Preferences on Activities

Users directly express their preferences on activities by using the function of rating activities as favourite or disliked activities. Based on the ratings, SocConnect can learn users' preferences and predict whether they will be interested in new similar activities from friends. Machine learning techniques are often used for learning and prediction. SocConnect applies the classic techniques of Decision Trees, Support Vector Machine [5], Naive Bayes, Bayesian Networks, and Radial Basis Functions [6]. In brief, decision tree learning is one of the most widely used techniques to produce discrete prediction about whether a user will find an activity interesting. It classifies an instance into multiple categories. Naive Bayes Classifier and Bayesian Belief Networks are the two commonly used Bayesian learning techniques. The method of Radial Basis Functions belongs to the category of instance-based learning to predict a real-valued function. Support Vector Machines have been shown promising performance in classification problems. The implementation of these techniques bases Weka 3.7.0. The performance of these techniques on learning users' preferences on their social network activities will be presented and compared in Section 4. The one providing the best performance will be used by our system.

Table 1. Features of Activities for Used Learning

Features	**A Set of Possible Values**
Actor	actor's SNS account ID
Actor Type	favourite; neutral; disliked
Activity Type	upload album; share link; upload a photo; status upload; use application; upload video; reply; twitter retweet; etc
Source	Facebook; Twitter; etc
Application	foursquare; FarmVille; etc
Rating	favourite, neutral, disliked

[5] http://www.crisscrossed.net/2009/10/15/network-overload-the-burden-to-deal-with-too-many-social-network-sites/

To work with the above learning techniques, an activity needs to be represented by a set of features. Table 1 summarizes a list of relevant features and some of their possible values. Each activity has an actor (creator). SocConnect allows a user to add friends into a favourite or disliked list. Using these two features, we will be able to learn whether a user tends to be always interested in some particular friends' activities or activities from a particular type of friends. As discussed in Section 2.3, each activity has a type. We also take into account the sources which activities come from, such as Facebook and Twitter, since often users have a particular purpose for which they predominantly use a given SNS, e.g. Facebook for fun, Twitter for work-related updates. From this feature, we can find out whether a user is only interested in activities from particular social networking sites source. Different applications used to generate those activities are also useful to consider. For example, if a user's friend plays "MafiaWars" on Facebook but this user does not, the status updates generated from the "MafiaWars" application may be annoying to the user. We leave out the textual content of activities. One reason is that many activities, such as video uploads, do not have any textual content. Another reason is that activities may contain non-Latin language characters and the current meta-data of activities cannot reflect which language the actor is using, which makes text analysis difficult and expensive.

After learning from a user-annotated list of activities from his or her friends, each of which is represented by a set of the feature values, our system is able to predict whether a new activity from a friend will be considered as "favourite", "neutral" or "disliked" by the user. We assign an approximate weight to the new activity as follows:

$$w = \begin{cases} 0.5 & \text{if predicted as favourite;} \\ 0 & \text{if predicted as neutral;} \\ -0.5 & \text{if predicted as disliked.} \end{cases} \tag{1}$$

These predictions are based on the features of each activity. The next section presents how the social context, expressed by the user by grouping friends in SocConnect, influences the recommendations.

3.2 Heuristic to Supplement Learning

As described earlier, SocConnect allows users to create groups and add friends into the groups. A group implies the existence of some commonalities among the members of the group or some activities that group members have been doing together. The group information provides an indirect indication about users' preferences on activities. For example, if many activities of members in a given groups are considered as favourite by a user, the activities of the other friends classified by the user in this group will also be likely interesting to the user. Based on this heuristic, we extend the results of machine learning, by adjusting the weight of an activity. More specifically, for a friend in a group, if the number of favourite activities of other group members is larger than that of disliked activities, the weight of each activity from this friend will be increased. Otherwise, the weight will be decreased. Formally, suppose that the number of liked (marked as "favourite") activities of other group members in the group is F, and the number of disliked activities from them is D, then the weight of an activity from the friend will be updated as follows:

$$w = w + 0.5 \times \frac{F - D}{F + D} \tag{2}$$

Note that in extreme cases where every activity of the other group members is considered favourite, the weight of the friend's activity will be increased by 0.5. On another hand, if every activity of the other group members is considered disliked, the weight of the friend's activity will be decreased by 0.5. Also note that w stays the same if every activity of the other group members is considered neutral by the user ($F + D = 0$). For a friend who belongs to several groups, the effect of the heuristic on the weight of the friend's activity will be averaged over these groups.

This extension brings two extra levels of user interests in activities, namely "very favourite" and "very disliked". Together, we have a range of five levels of distinction for user interests, which has been commonly used in many popular rating systems, such as Amazon (amazon.com) and TripAdvisor (tripadvisor.com). The mapping between the interest levels of users in activities and the numerical weight for the activities is summarized in Table 2.

Table 2. Interest Level, Activity Weight and Colour Presentation

Interest Level	**Activity Weight**	**Colour**
Very Favourite	$0.6 \leq w \leq 1$	Persimmon
Favourite	$0.2 \leq w < 0.6$	Tawny
Neutral	$-0.2 \leq w < 0.2$	Maroon
Disliked	$-0.6 \leq w < -0.2$	Burgundy
Very Disliked	$-1 \leq w < -0.6$	Thyrian purple

3.3 Adaptive Presentation of Recommendations in Visualization

The recommendations for the activities that the user may find interesting are integrated in the display of the activities in the activity stream that the user views in the interface of SocConnect. Colour in a spectrum that allows people with the most common type of colour-blindness (red-green) to distinguish,[6] is used to represent if an activity is recommended or unrecommended according to the predicted interest level calculated for the activity (Table 2). In this way the recommendation is unobtrusive, and can be easily ignored, but in the same time, it is intuitively clear for the user since it uses the metaphor "hot" item (displayed

[6] Images can be tested for appearance with simulated colour blindness at: http://www.colblindor.com/coblis-color-blindness-simulator/

in bright orange background, yellow text and large font) and "cold" item (dark purple background, blue text and small font). The metaphor allows representing a spectrum of recommendations with a larger number of values than 5, but we have picked 5 colours to represent transitions from hot through neutral (earth colour) to cold.

We have tested a visualization of items with different levels of interestingness using this metaphor with users in a study in previous work [4] and it was shown to work very well in quickly focussing user attention to the recommended items, while still allowing them to explore all items. This kind of visualization has been successfully deployed in the Comtella-D system in four classes with over hundred students for 2 years. That is why we decided to use it in SocConnect.

Fig. 2. An Example of Visualization

4 Evaluation

We carried out another study to evaluate the performance of the five machine learning techniques on predict user preferences on social activities. Twelve subjects were involved in our evaluation. Five of them are from Saskatoon, Canada, and the other seven are from New Jersey, USA. A half of them are students and the other half are workers. We collected from the subjects the recent Facebook and Twitter activities from their friends. Ten of the subjects are experienced users of Facebook and Twitter. For each of these subjects, we collected 100 recent activities of friends. The other two subjects are relatively new users of Facebook and Twitter. For each of them, we collected around 50 recent activities of friends. We asked all subjects to rate their friends and activities. On average, they rated 38% of their friends as favourite or disliked friends and 45% of the activities as favourite or disliked, thus representing quite a diverse data sample.

A 10-fold cross validation was performed on the collected data from each subject, and the average performance of the machine learning techniques over the activities of all subjects are reported in Figure 3. Although the performance difference among these techniques is not very significant, support vector machine (SVM) provides the best performance, and it correctly classifies 74.1% of instances in the testing data. RBF performs the worst (70%). The performance of Naive Bayes and that of Bayesian Belief Network are about the same (around 72.6%). Decision Tree performs a little better (71.4%) than RBF.

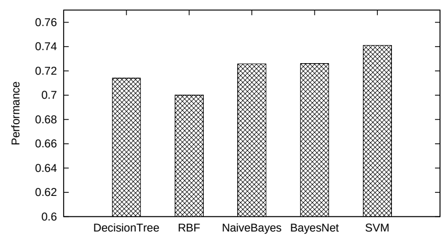

Fig. 3. Performance Comparison among Machine Learning Techniques

By looking closely into the predicted results, we found that many instances were misclassified by only one interest level, i.e. from "favourite" to "neutral" or from "disliked" to "neutral" and vice versa. We consider these as smaller mistakes. We summarize in Figure 4 the percentage of more serious misclassification from "favourite" to 'disliked" and vice versa. We can see that only a very few (less than 3%) activities have been misclassified from "favourite" to "disliked" and vice versa. SVM consistently shows its best performance in this case. Overall, the experimental results confirm the good performance of machine learning techniques in learning social networking users' preferences on their friends' activities. SVM is particularly recommended in this context.

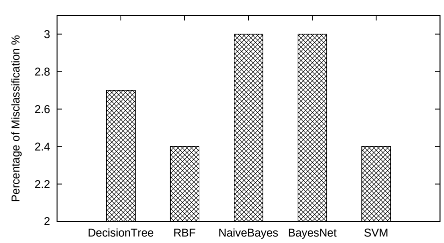

Fig. 4. Percentage of More Serious Misclassification

We also performed the validation on only 50% of collected data. More specifically, for each subject, we randomly selected 50% of collected instances. For each half of the data, we performed the same 10-fold cross validation to test the performance of the machine learning techniques. We repeated this process for 10 times to get the average performance when using only 50% of collected data.

Results shown in Figure 5 indicate that the performance when using 50% of data is consistently lower than that when using all data for the five machine learning techniques. This implies that the performance of personalized recommendation on social activities can be much improved when more data is collected from users, as users continuously use our system.

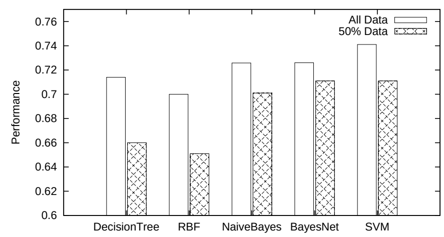

Fig. 5. Performance When Using All Data verses Performance When Using 50% Data

Using Weka's feature selection function, we can see which features are more important for individual users. We summarize in Table 3 the number of subjects for whom each feature was the most important one in the prediction. For all users, the feature "Actor" is the most important. "Actor Type", "Activity Type" or "Application" are more important for different users. The source of activities (i.e. whether they come from Twitter or Facebook) turns out to be not important. This interesting difference represents the diversity of social networking users' criteria in judging whether an activity is interesting to them, reflected in their ratings. Some users mainly care about their close friends' activities. Some users care more about the applications that generate the activities, which are usually the games they are playing. The implication is that learning the user type may be useful in personalized recommendation of activities. We leave this for future work.

Table 3. The Most Important Features

Features	Actor	Actor Type	Activity Type	Application	Source
Number of Subjects	12	4	3	3	0

5 Related Work

There have been some attempts to create personal portals that aggregate a user's accounts on different social networking sites, for example, the Seesmic Desktop (seesmic.com), power.com and the social web browser Flock (flock.com). They allow the user to view her pages on different social networking sites in one place. In this way, the users do not have to login to many different sites to view the updates of their friends. However, these applications do not allow users to blend or group their friends from different places.[7] They provide just a single-login

[7] http://news.cnet.com/8301-17939_109-10109878-2.html

interface in which users can switch between different tabs, one for each social networking site.

Bojars et al. [7] have been working on the SIOC project (Semantically-Interlinked Online Communities). This project shares similar focus with our work: social network portability and semantic web technologies. They proposed the SIOC ontology, which mainly focuses on users, implicit friendship, and social contents (primarily photos and discussions) in online communities such as online forums and Weblogs where contexts of social data are not so different.

In contrast, we focus mainly on developing a user-centric approach for integrating users' social data (including explicit friendship) on different SNSs, and that allows users to organize their social data and to create their personal contexts for the social data. We also provide personalized recommendation of friends' activities from different SNSs that are interesting to users.

Most recommender systems use collaborative filtering [8–10] based on the sharing of user ratings. While many SNSs deploy algorithms based on the analysis of social network structure to recommend new friends to the user, there haven't been many approaches to recommend contents on SNSs. One such approach is SoNARS. It takes a hybrid approach, combining results from collaborative filtering and content-based algorithms [11]. Dave Briccetti developed a Twitter desktop client application called TalkingPuffin (talkingpuffin.org). It allows users to remove "noise" (uninteresting updates) by manually muting users, retweets from specific users or certain applications. Currently, SocConnect focuses on automatically providing recommendations of social networking activities mainly based on the features of the activities.

6 Conclusions and Future Work

Our work has four contributions: 1) integration of social data from different SNSs; 2) allowing users to define their personal contexts of social data, including their integrated friends who may have SNS accounts on different SNSs, groups of their friends who share commonalities and activities from the users' own perspective, as well as their interest level (favourite, neutral or disliked) for friends and activities; 3) personalized recommendation of activities that may be interesting to individual users; 4) suggestion of a particular machine learning method for user preferences that has the best performance among five compared methods (SVM). A fifth potential contribution is the visualization of personalized recommendations integrated in the interface for viewing the activities, once its benefits are evaluated with users. Together, the personal dashboard application SocConnect provides users with a tool of integrating social data across different SNSs and with the convenience to selectively view friends' activities that are interesting to them.

For future work, next step will be to conduct user studies on the user interface to evaluate the usability of the visualization of recommendations and the appropriateness of the proposed heuristic to supplement machine learning. We are interested in exploring more deeply the relative importance of different features of social networking activities, to further improve the performance of personalized recommendation of activities. Other features that may be worth looking at include textual content of activities and the targeted friends of friends in activities.

7 Acknowledgement

This work has been supported by the Canada NSERC Discovery Grant of the third author.

References

1. Chisari, M.: The future of social networking. In: Proceedings of the W3C Workshop on the Future of Social Networking. (2009)
2. Erétéo, G., Buffa, M., Gandon, F., Leitzelman, M., Limpens, F.: Leveraging social data with semantics. In: Proceedings of the W3C Workshop on the Future of Social Networking. (2009)
3. Wang, Y., Zhang, J., Vassileva, J.: SocConnect: A user-centric approach for social networking sites integration. In: Proceedings of the International Conference on Intelligent User Interface (IUI) Workshop on User Data Interoperability in the Social Web. (2010)
4. Webster, A., Vassileva, J.: Personal relations in online communities. In: Adaptive Hypermedia and Adaptive Web-Based Systems, Dublin, Ireland, Springer LNCS (June 2006) 223–233
5. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In Schoelkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods: Support Vector Learning. MIT Press (1999)
6. Mitchell, T.M.: Machine Learning. McGraw-Hill (1997)
7. Bojars, U., Passant, A., Breslin, J., Decker, S.: Social network and data portability using semantic web technologies. In: Proceddings of the Workshop on Social Aspects of the Web. (2008)
8. Resnick, P., Lacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. In: Proceddings of the ACM conference on Computer supported cooperative work. (1994)
9. Kautz, H., Selman, B., Shah, M.: Referral web: combining social networks and collaborative filtering. Communications of the ACM **40**(3) (1997) 63–65
10. Goo, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J., , Riedl, J.: Combining collaborative filtering with personal agents for better recommendations. In: Proceddings of the sixteenth national conference on Artificial intelligence. (1999)
11. Carmagnola, F., Vernero, F., Grillo, P.: Sonars: A social networks-based algorithm for social recommender systems. In: Proceddings of the 17th International Conference on User Modeling, Adaptation, and Personalization. (2009)