SocConnect: A Social Networking Aggregator and Recommender

A Thesis Submitted to the

College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the degree of Master of Science

in the Department of Computer Science

University of Saskatchewan

Saskatoon

By

Yuan Wang

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science

176 Thorvaldson Building

110 Science Place

University of Saskatchewan

Saskatoon, Saskatchewan

Canada

S7N 5C9

ABSTRACT

Users of Social Networking Sites (SNSs) like Facebook, MySpace, LinkedIn, or Twitter face two problems 1) their online social friendships and activities are scattered across SNSs. It is difficult for them to keep track of all their friends and the information about their friends online social activities. 2) they are often overwhelmed by the huge amount of social data (friends' updates and other activities).

To solve these two problems, this research proposes an approach, named "SocConnect". SocConnect 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. "SocConnect" also can recommend unread friend updates to the user based on user previous ratings on activities and friends, using machine learning techniques. The results from one pilot studies show that users like SocConnect's functionalities are needed and liked by the users. An evaluation of the effectiveness of several machine learning algorithms demonstrated that , and machine learning can be usefully applied in predicting the interest level of users in their social network activities, thus helping them deal with the "network" overload.

Acknowledgements

First of all I would like to express my sincere thanks to my supervisor, Dr Julita Vassileva for her support, guidance, patience, and financial assistance throughout my entire two and half years of study. I would like to thank the members of my advisory committee: Dr. Ralph Deters, Dr. Jim Greer, and Dr Anh van Dinh for their valuable advices and insightful suggestions. I would like to thank Dr. Jie Zhang for his continuous support of my work; It was a pleasure to work with him. I also would like to thank Ms. Jan Thompson, Graduate Correspondent at the department of Computer Science, who has been very helpful throughout my study at University of Saskatchewan and very kind. Finally, I would like to thank my parents and wife for their unconditional love and selfless support.

CONTENTS

List of Tables

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	Application Programming Interface
CF	Collaborative Filtering
DAO	Data Access Object
IR	Information Retrieval
RDBMS	Relational Database Management System
RIA	Rich Internet Application
SNS	Social Networking Site
SNSAcc	Social Networking Site Account
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency

Chapter 1

Introduction

The advent of web 2.0 technology, especially social networking sites, has changed the way people communicate. Clara Shih, in her book "The Facebook Era" [32], observes that social media including Facebook[1] has transformed the socio-cultural landscape - people's behaviours, attitudes, interactions, and relationships. People spend more time on social networking sites than ever, and prefer communication via social networking sites over emails [17]. Every successful social networking site (SNS) has its unique features. Facebook allows a large number of third party applications to build on its APIs (Application Programming Interfaces). Twitter[2] offers micro-blogging and an asymmetric following relation between users. MySpace[3] has a large user community interested in music. LinkedIn[4] focuses on career and professional networking.

Despite the diversity of SNSs and the fact that social media enriches people's lives, the end user faces many challenges and limitations of current SNSs. Two of these challenges and limitations motivate this research.

1.1 Motivations

1.1.1 The "Walled Garden" Problem for Social Networking Sites

"A walled garden is an analogy used in the telecommunications and media industries when referring to carrier or service provider control over applications and content/media on platforms (such as mobile devices) and restricting convenient access to non-approved applications or content" [10]. In the context of SNS, "walled garden"

---

[1] www.facebook.com
[2] www.twitter.com
[3] www.myspace.com
[4] www.linkedin.com

is about the SNSs companies such as Google, Facebook, or Twitter have control over user's data. With the explosion of SNSs, it is also common that one user engages with multiple SNSs. In July 2009, Anderson Analytics conducted an online survey over 11,000 SNS users. The results show a high overlap of user populations of Facebook, Twitter, and LinkedIn.[5] User-generated contents, users' online activities, and their friendships are scattered over different SNSs. It becomes increasingly inconvenient for users to manage their social data and constantly check many sites to keep track of all recent updates. Even worse, people may have different accounts on the same social networking site in order to protect their privacy or for other purposes.

1.1.2 The "Networks Overload" Problem

Another problem of SNS is information overload. These users of multiple SNSs see a great number of status updates and other kinds of social data generated by their network friends everyday. In this thesis, "social data" denotes status updates, posts of photos, links, ratings, likes, retweets, i.e. all new items that appear as in the stream of updates in a SNS. The innovation of SNS has constantly increased the richness of their social data. This causes a significant information overload to users. Christian Kreutz in his blog described this specified kind of information overload as "network overload".[6]

This overload is caused by two reasons: first, there is too many new social data appearing constantly on SNS; second, this social data does not have explicit context. The first reason is fairly intuitive, the second one needs some explanations.

SNSs generates huge amount of social data. However, lots of these data do not have explicit context. For example, the way the word "friend" is used in Facebook does not reflect the true meaning of the word in colloquial English. On Facebook, a user's "friends" may include co-workers, college mates, and people who the user barely knows but was too polite to decline befriending. It is thus important to have a way of distinguishing these people. Another example addresses the different purposes

---

[5]http://www.readwriteweb.com/archives/who_uses_social_networks_and_what_are_they_like_part_1.php

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

that SNS have. For example, the user's interactions with friends on last.fm[7] may have different contexts from her interactions with friends on LinkedIn. On last.fm, the users' interactions mostly relate to music, but on LinkedIn, the interactions are more formal and mostly relate to business networking and career development. Users and their friends on different social networking sites may also have different kinds of relationships. For example, Facebook friends are mostly people whom the user already knows [23], but users may have not met most of their Twitter friends in person. Without explicit context, it becomes very difficult to handle the huge amount of social data and too complex for users to make sense of the data. The contexts may include the type of social bound (the provenance, closeness, symmetry, etc.) of relationships (family, colleagues and friends in personal life), the common interests they share, the closeness of friendships, and the location of friends.

The "network overload" becomes more serious when the social data of the user is integrated across different SNSs into one place by a social aggregator application [8]. A social network aggregator is the application pulls together content from multiple social network service into a single location. The number of updates will increase significantly in this case. One way to deal with information overload is by providing recommendations for interesting social updates, which allows the user to focus her attention more effectively and deal with the "information glut".

This research proposes an approach called "SocConnect" (short for social connect) which attempts to address these two problems: "walled garden" and "network overload". SocConnect should not only provide the functionality to integrate social data across SNSs, but also should provide the functionality to allow users to organize their social data across SNSs. Thus, user can define social contexts of their social data. In further usage, the context could help user to browser his or her social data. Moreover, SocConnect should be able to learn the user's preference and recommend new unread social data to user base the preference.

---

[7]www.last.fm
[8]http://en.wikipedia.org/wiki/Social_network_aggregation#Social_network_aggregators

1.2 Thesis Outline

The structure of remaining chapters of this thesis as follows: Chapter 2 presents how SocConnect addresses the "walled garden" problem. Chapter 3 focuses on how the SocConnect approach addresses "network overload" problem. Chapter 4 describes the implementation and demonstration of SocConnect. Chapter 5 presents a field study to evaluate SocConnect's functionalities with users and reports the results. At last, Chapter 6 summarizes and concludes the contributions and presents directions of future work.

CHAPTER 2

AGGREGATING DATA ACROSS SNS

This chapter describes how the SocConnect approach addresses the "walled garden" problem. This chapter focuses on the aggregator aspect. Section 2.1 presents a review of existing work in the area of social network aggregators and user data interoperability in SNSs. Section 2.2 describes the architecture of SocConnect aggregator. Section 2.3 describes the proposed functionalities in SocConnect for aggreagating and managing social data across SNS. Section 2.4 demonstrates SocConnect's user interface for each proposed functionality. Section 2.5 presents a pilot study conducted to collect user background and elicit user requirements to make sure that the proposed functionalities are needed.

2.1 Literature Review

2.1.1 Social Network Aggregators

A social network aggregator is an application that integrates different user's social data across different SNSs and present together. Currently, many social network aggregators are available to users on the Internet. In 2007, the Mashable (mashable.com) listed 20 popular social network aggregators in one of its articles.[7] Based on their platforms, social network aggregators can be classified as web and desktop applications. In web aggregators, users need to register and create a new account for the aggregator, and provide their SNSs accounts information to the aggregator. In desktop aggregators, users normally do not need to create an account. Desktop-like aggregators have been emerging on mobile platforms recently. Based on their functions, social aggregators can be divided into three groups: write-only, read-only, and write and read. Write-only and read-only aggregators usually are lightweight and

web-based. They allow users to publish or read the same status update to multiple SNSs. Write and read aggregators provide both write and read functions. There are many well-known social network aggregator: TweetDeck[1], Hootsuite[2], Seesmic[3].

Digsby integrates Instant Message (IM), email, and Social networks sites services together. When the user receives any new information from these services, a notification tool will alert the user and let her perform actions "delete" or "reply" with simple clicks.

HootSuite is social network aggregator that supports organizations in their brand management. Organizations can use HootSuite to publish news to various SNSs; it supports team collaboration: multiple users can share one or a set of SNS accounts to publish new content, it also can schedule updates, assign tasks among team members, internationalize content, and monitor the organization name mentioned in different SNSs. HootSuite is a available for different mobile and desktop platforms.

Seesmic is a standard social network aggregator which can connect with Twitter, Facebook, LinkedIn, and Google Buzz, it is available for mobile, desktop and web platforms. It supports both read and write functions.

TweetDeck, as its name indicates, started as a Twitter client which is still its main functionality, and evolved along the way to include Facebook, MySpace, LinkedIn, Foursqaure, and Google Buzz. It supports both read and write functions. It is available for mobile and desktop platforms.

Two things to notice: these social network aggregators are constantly adding new services and features, and one aggregator's functionality across different platforms, such as web, desktop, or mobile, may not be the same. All of them represent different feeds from different SNSs in parallel tabs thus increasing the information overload of the user. They do not provide a true integration of the feeds.

2.1.2 User Data Interoperability

User data interoperability allows to move and combine a given user's data across different systems. In order to achieve user data interoperability, there needs to be

---

[1]http://www.tweetdeck.com/
[2]http://hootsuite.com/
[3]http://seesmic.com/

Table 2.1: Comparison of the main known SNS Aagregators

Name	Platform(s)	Functions	SNS Support	Note
Digsby	Desktop	Read/Write	Facebook/Twitter/MySpace	Instant Message-style
HootSuite	Web/Mobile	Read/Write	Facebook/Twitter/MySpace/LinkedIn	Collaborative Publication
Seesmic	Desktop/Web/Mobile	Read/Write	Facebook/Twitter/MySpace/LinkedIn/Google Buzz	
TweetDeck	Desktop/Web/Mobile	Read/Write	Facebook/Twitter/MySpace/LinkedIn/Google Buzz	
SocConnect	Desktop	Read/Write	Facebook/Twitter	Blend,Group, Tag, and Recommend

a way to mapping the user's identification across systems and handle authentication across systems to gain the user's permission, and finally able to invoke the Application Programming Interface (API) provided by other systems to access user data. Therefore, user data interoperability requires identification and authentication management, and standardization of API. Standards like OpenID[5] and OAuth[4] have emerged from open web community to fulfill these requirements. OpenID is a solution for universal identification management, and OAuth is an open protocol about how to request and handle user authentication between systems. These two standards have been wildly accepted.

Berkovsky et al [12] state four major challenges for user data interoperability. The following list presents these challenges in the context of SNS.

1. Systems are not designed to share their user models: The merging of Web 2.0 and successful open API stories motivate SNSs to release open API. However, different SNSs have different priority and perspectives on open API development and release. For example, Facebook has put open API as its high priority: it has a clear roadmap of its API releases, an annual developer conference, and official library to facilitate the third party development. LinkedIn, in contrast, is relatively slow on the open API track.

2. Users' privacy: Exposing user data through API is a sensitive issue. In August 2009, the Canadian government requested Facebook to improve its user privacy protection, especially on API. User data should be behind locks. Users can allow only trustworthy applications to access their data [19].

3. Practical and technical considerations: Almost every large SNS faces scalability issues. These sites have their API traffic limit. Moreover, API method calls have other limitations. For example, A Twitter API call can only retrieve maximum

200 tweets (user updates). These considerations need to be dealt with when integrating social data.

4. Algorithms to translate one application's schema to another's: One important requirement for integrating social data across different social networking sites is a unified ontology to represent social data [17]. SNSs have their own syntaxes and terms for representing social data. Ontologies serve as a shared and common understanding of a domain that can be communicated between human users and widely spread software systems [20].

The academic and open web community have put great effort to design ontologies or other forms of schema for the representation of social data. There are several major standards, including FOAF[1], XFN[9], GUMO[2], Activity Streams[**?**], and MediaRSS[3]:

- FOAF[4], the *friend of a friend project* aims to define a RDF (resource description framework) vocabulary to describe relations between people;

- XFN[5], the *XHTML Friends Network* is a micro-format to represent a person's relations on the web;

- The *activity stream*[6] is atom-based standard format to represent a user's activities on social web applications [29];

- GUMO[7], the *general user model ontology* is an OWL (web ontology language) based ontology to describe user's characteristics and other information [22].

- SIOC, the *Semanticall-Interlinked Online Communities*, is a semantic ontology that aims to solve the user data interoperability in online communities, such as blogs and forums. [13]

- MediaRSS [8] is a RSS-based schema from Yahoo to describe rich media elements, such as audio, images, or video, on the Internet

---

[4] www.foaf-project.org
[5] www.gmpg.org/xfn/
[6] www.activitystrea.ms
[7] www.gumo.org
[8] http://search.yahoo.com/mrss

These standards have solid foundations; some of them have already been adopted by social networking sites and other IT companies. For example, the activity stream has been embraced by Facebook, Google Buzz, Windows Space Live, and MySpace. However, syntax differences among SNSs still exist and translation is still needed. My research does not contribute to user data interoperability, but uses existing standards as foundation.

Therefore, the schema, or ontology, should be able to allow users to express the context of social data. There are two solutions for the expression of contexts. One common way is a top-down approach that pre-defines sets of vocabularies to describe different types of social contexts. However, social contexts contain too many dimensions and too many possible variables along each dimension, of which only a few may be relevant to any given user. The process of selecting the relevant value in each dimension from a pre-defined ontology would be too hard for the user. The second solution is to let users themselves express social contexts using tags and attach these tags to the social data. This solution is more flexible and feasible, and we use it in our work.

Bojars et al. [13] have developed 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, I 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. My approach also provides a personalized recommendation of friends' activities from different SNSs that are interesting to users.

The next section proposes a schema to describe the social data across different SNSs.

2.2 Proposed Schema to Integrate Social Data Across SNSs

To represent the heterogeneous social data across SNSs, a unified schema is required. As described in the previous section, a variety of standards and ontologies serve this purpose, such as FOAF, activityStream, SIOC project (Semantically-Interlinked Online Communities) , etc. However, any single one of these standards and ontologies cannot fully meet the requirements of SocConnect. FOAF's scope is the users and their relations among themselves, activityStream focuses on describing the user's online activity. SIOC project's scope is mainly on blogs and forums. Therefore, an adapted schema is developed based on FOAF and activitystream.

The philosophy behind Activity Stream is that the essential elements of SNSs include actors and their activities. Every user is an actor; every movement of an actor is an activity, such as adding a new friend, publishing a new blog article, and commenting on others' articles. Each activity has a type, such as Twitter update, Twitter retweet, sharing a link or a Facebook photo. The type of an activity represents the feature of this activity.

Social data is inherently "URI-based"; almost every piece of social data has its URI (Unique Resource Identifier). For example, each Facebook user has his or her own facebook homepage as URI, each Twitter update has a permanent address (such as http://twitter.com/username/status/999), and each Flickr[9] photo has its URL. This makes social data easy to be interlinked. A design of the schema for representing social data can easily take advantage of this feature of social data.

The proposed schema is presented in Figure 2.1. There are five entities in the ontology: SNS account (SNSAcc), integrated account (person), activity, tag, and group.

- SNSAcc: represents a user account on a SNS. Each SNSAcc has a source which is a SNS, such as Facebook, Twitter, and MySpace. The source indicates what kinds of data are collected by the SNS. For example, Facebook keeps lots of information for each user. On another hand, Twitter only stores very simple user information.

---

[9]www.flickr.com
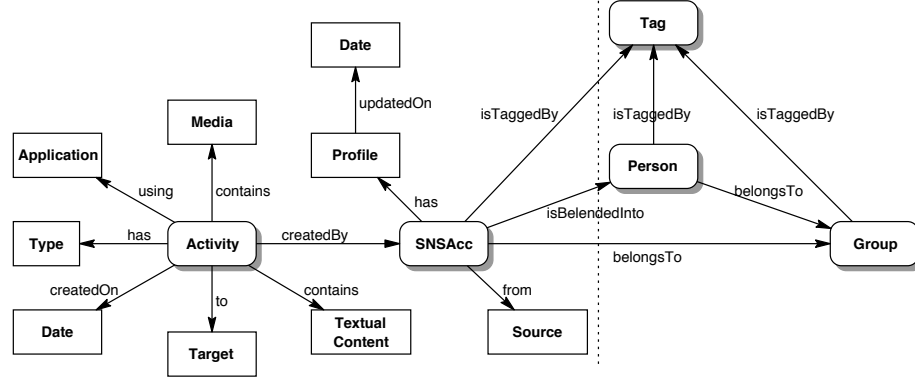
Figure 2.1: Schema for Social Data

- Person: represents a person who holds one or more SNS accounts. For example, a user on Facebook also can have a Twitter account. These two SNSAccs can be blended together. The word "person" may not be the best choice to describe the concept. For example, organizations, or companies, like CNN can have accounts on Twitter or Facebook, and they are not really "persons".

- Activity: represents generic information about activities appearing on SNSs. Activities can be user status updates, events like a new friend added by the user, or a new third party application used by the user.

- Tag: represents a user-generated label. Tags are used to represent contextual information of social data according to the users' own preferences [22].

- Group: represents a user-defined group for keeping friends together. A member of a group can be a SNSAccount or a Person.

The entities are interlinked among each other. Each SNSAcc has a set of activities belonging to the user's SNS account. A person may have a set of SNSAccs and a number of activities associated with each SNSAcc. A group may contain a number of persons and SNSAccs as its members. One SNSAcc can belong to multiple persons or groups, and one person can also belong to more than one groups. The domain objects of SNSAcc, person and group can have a set of tags. The Activity class is the core of this domain. Each activity has a SNSAcc as its actor. Activities of users or their friends incrementally fill social networks with contents. SNSs are essential sources of activity streams. Users and their friends are the actors of the activities.

11

2.3 Proposed Functionality to Allow Users to Add Context

Based on the schema presented in the previous section, an approach has been developed to integrate social data from different SNSs. This approach proposes four categories of functionalities: First, connecting different SNSs and loading users' social data; second, allowing users to manage their friends and assign context to their social data; third, filtering social data, and the fourth type of functionality is recommendation. Recommendation is proposed as a way to deal with "network overload" and is discussed in Chapter 3. Here the first three functional categories are discussed since they are related to the problem of "walled garden", and provide a way for users to combine and organize their social data from different SNSs.

2.3.1 Loading Social Data

The first functional category, "loading social data": SocConnect uses authentication methods provided by different SNSs and invokes their APIs to retrieves users' friends information and their activities on these sites. There are three authentications methods used by current SNSs: basic authentication, OAuth, and custom authentication.

Basic authentication is to ask the SNS user to give her SNS username and password to the external application, e.g. SocConnect. Basic authentication is easy to implement but because it puts user's security in danger, it is considered as an anti-pattern. OAuth, as mentioned before, is open and secure authentication mode. Another alternative is custom authentication is a special authentication method that only works for one SNS. Many SNSs provide multiple authentication methods: for example Twitter provides both basic authentication and OAuth; Facebook provides both OAuth and custom authentication. SocConnect used both basic and custom authentications.

After authentications, SocConnect invokes some APIs provided by SNSs. After get raw data (in XML or JSON) from the SNSs, SocConnect translates it into the schema described above.

2.3.2 Managing Friends

The second 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 SNSs. This approach allows the user to merge the different accounts of a friend across SNSs, to create a single "person". This function is a unique feature of this user-centric approach; as stated in [25], there are not other social network aggregators that allow linking friends corresponding to the same physical persona across different SNSs. The friend can have different user accounts on different sites, but the user knows that they refer to the same person (something that no data mining algorithms can find out accurately). It is up to the user to create the mapping between her friend's accounts across different sites and assign an integrated account to represent the same friend. In this way, the user can have an integrated view of all activities of this friend, no matter which SNSs the activities come from. Compared to the social network aggregators that only present social data at the same place, SocConnect provides users with the possibility to integrate deeper the scattered social data.

The second function in the "managing friends" category is to group friends. Users can put their friends, both individual SNS accounts and blended "person" accounts, into groups. This function allows users to express the contexts of friendships, by specifying a dominant which are the shared characteristics or interests between friends. For example, a user John who is a graduate student in Computer Science has a friend, Ben. Ben is John's buddy from a scuba-diving club, and he is also a computer scientist. Ben and John are both interested in Erlang programming and often share their findings and ideas using Twitter. They use Facebook to share their diving pictures, news about diving club events, and general news about their lives. In this approach, John will first map the two Bens he knows - the one from Twitter and the one from Facebook. Next, he will define one group for his diving friends and add Ben (the Facebook Ben) into this group. He will also define an "Erlang" group and add Ben (the Twitter Ben) into it. John has another friend Vivian, she talks about Erlang programming both on Twitter and Facebook, so John can blend her accounts on these SNSs and then add the integrated person into the "Erlang" group.

2.3.3 Filtering Social Data

The third functional category, "filtering social data" also has two functions. A filter can be created for social data according to tags provided by users. Users can tag friends (both individual social network site accounts and integrated accounts), groups, and individual social update. After tagging, users can browse social data based on these tags. For example, John attaches the tag "scuba-diving" to his friend Ben's Facebook account. If John now wants to view social data about scuba-diving, information about Ben's recent activities from Facebook will be presented to John. Tagging allows the user to add richer context description to their friends, in addiction to the that achieved by grouping.

Another function is to allow users to browse social data based on groups. Users can view the activities of the members in the groups which they are interested in. Note that the function of filtering social data by tags and that of filtering social data by groups are different, and both are necessary. Normally, the number of groups created by a user is not excepted to be very large. Otherwise, it will become difficult for the user to manage or look after all her groups. Tagging friends provides a flexible way for the user to view activities of only a few friends for whom she does not want to create a separate group.

2.4   Demonstration

This section provides several screenshots to demonstrate the user interface of Soc-Connect. This interface was an early prototype implementing the main functionalities proposed of the approach rather than the ultimate interface for the application. Suppose that a user Jane has accounts on both Facebook and Twitter. SocConnect retrieves Jane's social data on these two sites. The social data of her friends can then be managed and filtered by her SoCConnect dashboard based on her personal needs or interests. We step through an example to show more specifically what Jane can do with the application. The social networking site accounts of the actual users in the screenshots are blacked out to protect their privacy.

Jane can use SocConnect to blend her friends who have social networking site
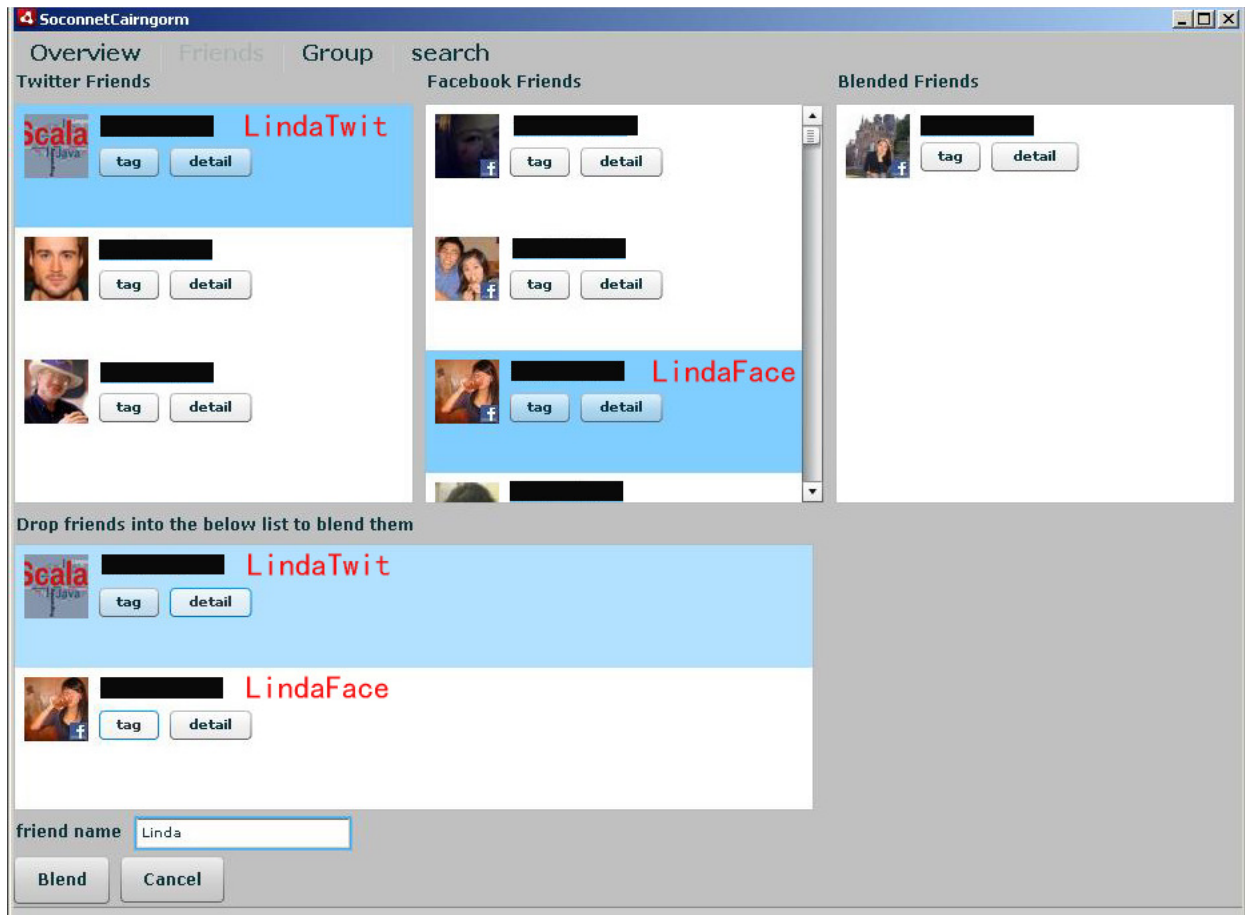
Figure 2.2: Blending Friends

accounts on both Facebook and Twitter. As shown in Figure 2.2, there are three lists in the upper part. The left list contains Jane's friends on Twitter and the middle one contains her friends on Facebook. Jane drags her friend Linda's Twitter account "LindaTwit" from the left list and Linda's Facebook account "LindaFace" from the middle list to the lower list. By clicking the "Blend" button shown in the bottom of the figure, Linda's accounts in the lower list are joined into a "blended" person. Jane gives a name "Linda" for the blended person. The third list in the upper-right part of the screen shows the list of all Jane's "blended" persons. Linda will be added to the list.

Jane can also use SoCConnect to group her friends together. As shown in Figure 2.3, the interface for this function is similar to the interface for blending friends. To add members into a group, Jane can drag her friends' accounts from the three lists in the upper part of the figure and drop them into the list in the lower part.

Figure 2.3: Grouping Friends

She drags her friends in New Jersey into the lower list, including John and Bob from the Twitter list and Amy from the Facebook list. She also drags the blended person Linda into this list from the list of blended persons. She gives the name "friends@NJ" to the group and clicks the button of "Create a new group" in the bottom of the screen. A new group is then created for Jane, and the list of Jane's groups is shown in the right most list in the lower-right part of the screen. A user can also put her friends in different groups, e.g. John can be both a member of Jane's "friends@NJ" group and her "friends@SK" group.

The function of grouping friends provides a flexible way for users to organize their friends by contexts. It also allows users to filter only social data from the members of a particular group. For example, Jane can check news from friends@NJ by clicking the group name listed in the right most list of Figure 2.4 marked by "Groups". A

list of the members in this group will appear in the middle list, and the updates from these members will appear in the left most list.
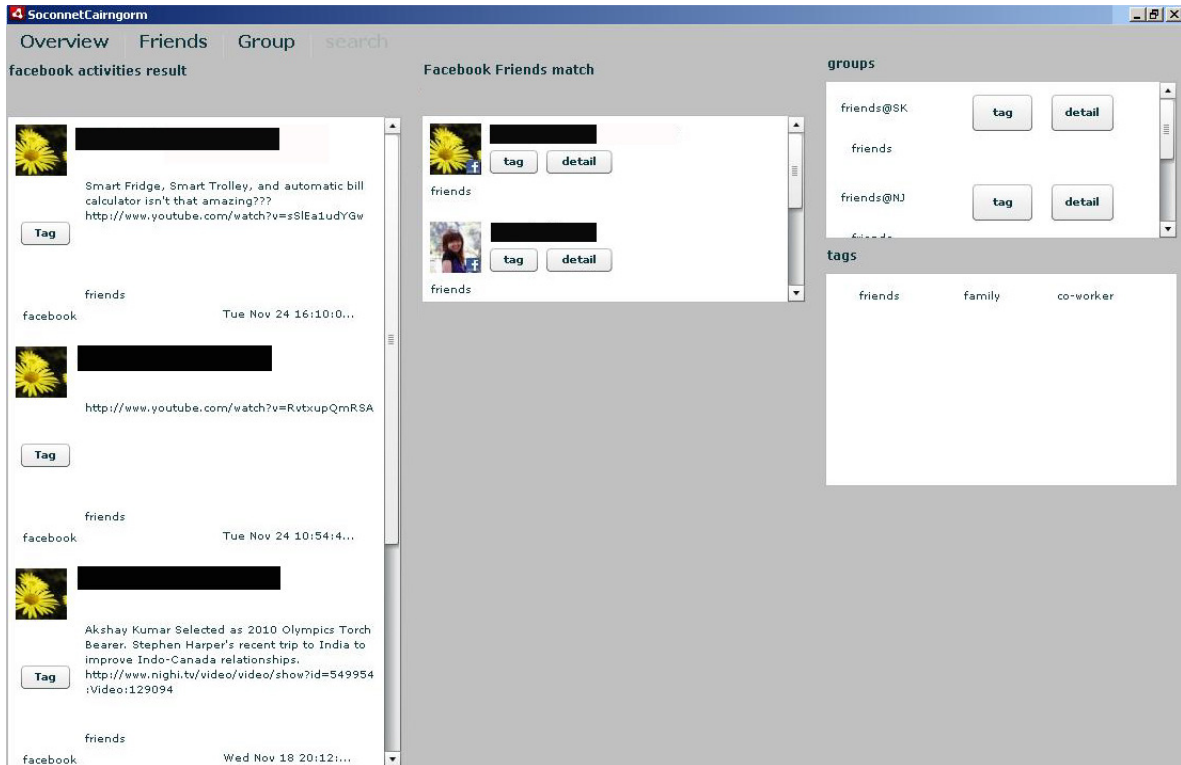
Figure 2.4: Filtering Social Data

To allow for more expressive representation of context information, users can add tags to their friends and groups. They can choose any of these tags as a keyword for filtering, and the application will display the social data that relates to the tag. As shown in Figure 2.4, Jane can add a tag to her friend John by clicking the button "tag" beside John's icon. A separate window pops up as shown in Figure 2.5. Jane can choose an existing tag from the list of tags or add her own tag. In this case, Jane adds her own tag "Diving" to John and clicks the "Add" button (Figure 2.5).

The list of all Jane's tags is shown in the right most list marked by "Tags" in Figure 2.4. Jane can view the activities of all her friends that relate to diving by clicking the tag "Diving". All her friends who are tagged by "Diving" will appear in the middle list, and the updates from these friends will appear in the left-most list.

17
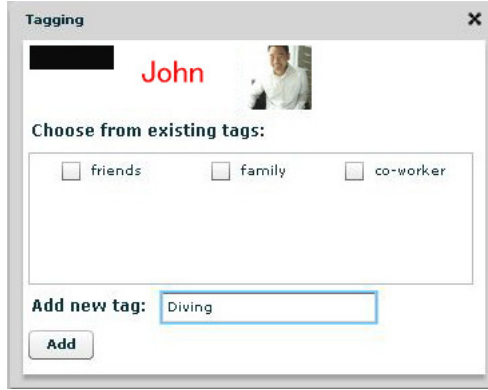
Figure 2.5: Tag a Friend

2.5 The Pilot Study

2.5.1 Goals

To evaluate the functionalities choice made in Section 2.3, a pilot study was conducted. The goal was to verify the need for the proposed functionalities and their contribution, and also to get initial feedbacks on a prototypical user interface.

This study involved 16 subjects (all were students). Table 2.2 summarizes the demographic information about these subjects.

Table 2.2: Demographic Information about Subjects

Subjects	Computer Science	Non-CS	Total
Female	5	3	8
Male	5	3	8
Total	10	6	16

2.5.2 Methods

A simple prototype of the SoCConnect approach was used in the pilot study, as described in Section 2.4. The study had three phases: interview, testing, and suggestion. The interview phase aimed to collect user background on social networking sites usage. The testing phase involved getting the user to use the prototype , the suggestion phase involved answering a questionnaire discussed below. The full document of the pilot study is attached in Appendix B.

In the interview phase, the subjects were asked questions related to the Soc-Connect functionalities.[10] The first set of questions aimed to learn about the subjects' past experience on using social networking sites. For example, subjects were asked which SNS they visited more frequently (i.e. more than once in every week). The users were asked about approximately how many hours they spent on SNS and how many friends they had on each site. The questions in the rest of interview were adapted to their answers. For example, if a subject had mostly used Facebook and Twitter, the next questions were set in the context of these two SNSs.

The other three sets of questions in the interview phase were used to evaluate the necessity of the main functionalities, including blending friends, grouping friends, and filtering social data by tags and groups. For evaluating the necessity of the functionailties of blending friends, the following questions were asked:

1. Do you keep friends on different social networking sites for different purposes?

2. Do you have some friends who have accounts on several social networking sites? If so, how many roughly?

3. Have those friends been active on these sites?

4. Do they mostly have identical activities on these sites?

5. If they mostly have identical activities on these sites, do you want to view their activities in one place?

The functionality of blending friends is necessary only if users have some friends who have accounts on different (at least two) social networking sites. The positive answer of Question 2 (Q2) is then the prerequisite of having this function. But, even if a subject has the same friends on different sites, the subject may still not want to blend these friends if she keeps friends on different SNSs for different purposes or contexts. Therefore, the negative answer of Q1 is also the prerequisite. To argue that the functionality is actually necessary and useful, users' friends have to be active on different sites (Q3) and users should feel that it is valuable to view friends' identical activities in one place (Q4 and Q5).

---

[10]Note that during this interview phase, detailed explanations were provided if subjects were unclear about the questions or misunderstand some questions.

The questions related to the necessity of having the grouping friends function were as follows:

6. Some social networking sites allow you to put some friends into a list (a group). Have you ever used this function?

7. Do you have some friends who share similar interests, preferences, or demographic information, or do some activities together?

8. Do you want to create a group for these friends?

9. Do you also want to include in groups some friends on different sites?

Some SNSs (i.e. Twitter) provide the grouping functionality. If users have already made good use of this functionality (Q6), this becomes a positive indication for the functionility of grouping friends. However, this functionality is fairly new to most SNSs. It is likely that most SNS users have not paid much attention to this function. The other questions also provide estimation for the necessity of the functionality . It is very likely that users have friends who share some commonalities (Q7) if the users have many friends, this question was asked, in order to guide subjects to being focused on their friends who are in common for the next question (Q8). Question Q9 is related to the special functionality of grouping users' friends (who are on different SNSs). This functionality allows to put those friends in one group, which is impossible via a single SNS such as Facebook or Twitter.

For the functionality of tagging and filtering social data, the following questions were asked:

10. Have you ever had difficulty in browsing through your friends' updates, and have you been overwhelmed?

11. Do you want to organize your updates and your friends' updates into categories by tagging them?

12. Do you want to view your friends' activities (updates) by groups?

If a subject has many friends, the answer to Question 10 is likely positive. Thus, this question and the questions related to the subject's past experience on using

SNSs to provide indication for the necessity of having the functionality of filtering social data, especially in the case where the subject's friends on different SNSs are now gathered in one place (as in SocConnect). Question Q11 investigates subjects' preferences about organizing the status updates on SNSs by tagging friends. The final question (Q12) provides indication whether the functionality of grouping friends will be helpful for subjects' navigation of activities.

The purpose of the interface testing phase was to evaluate the usability of the interface of SocConnect. In this phase, subjects were asked to perform some tasks using the functions offered by the application. The six tasks included logging into the application, blending two friends, creating a group of friends, tagging one friend, and filtering by tags. Subjects' actions, such as whether subjects could success-fully accomplish those tasks and how much time they took for each task, had been recorded.

In the suggestion (feedback) phase, subjects were asked to provide suggestions or feedback about SocConnect. For example, they could provide feedback about which part of the current interface needs to be improved and which other functionalities should also be provided by SocConnect. The feedback could be useful for refining the application.

2.5.3 Results

Based on the subjects' answers to the questions about their past experience on using SNSs, Figure 2.6 summarizes the number of SNSs they frequently used (for at least once a week) by them. All subjects have frequently used more than one social sites. Most of the subjects have frequently used 2, 3 or 4 social sites. Figure 2.7 summarizes for each subject the total number of the subject's friends on all frequently used SNSs. Most of the subjects have more than 50 friends in total. Almost a half of the subjects have at least 100 friends. These results are very encouraging and motivating for our approach that integrates users' social data (including friends and friends' updates) on different social sites.

For evaluating the necessity of the "blending friends" function, proposed in Soc-Connect, Questions 1-5 were asked during the interview phase. Figure 2.8 is the
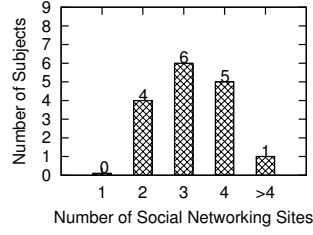
Figure 2.6: Number of Frequently Used Sites
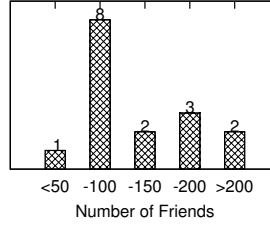
Figure 2.7: Total Number of Friends

summary of the number of each subject's friends who have user accounts on more than two social sites (Q2). As can be seen, only two subjects do not have such friends. More than a half subjects have at least 7 such friends. Several subjects (25% of all subjects) have more than 20 such friends. This result suggests a strong need for the function of blending friends.
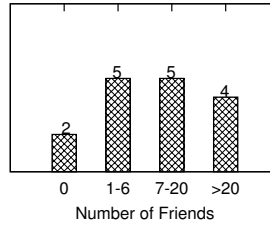
Figure 2.8: Number of Friends Having Accounts on More than Two Sites

The summary of the answers for Questions 1 and 3 to 5 is presented in Table 2.3.[11] The results of Questions 3-5 indicate the subjects' strong desire for the functionality of blending friends. Note that Q4 is overly strict. In fact, three subjects provided the answer of "50% similar and 50% identical". Within these three subjects, two subjects still provided positive answers to Q5 and only one subject was not sure whether

---

[11] For the two subjects who do not have friends with accounts on more than two social sites as pointed in Figure 2.8, we assume that they do not support the functionality of blending friends and will be negative for the questions 1 and 3-5.

the function is important. The result of Q1 is not significant even though most of the subjects do not keep friends on different social networking sites for different purposes. The reason that almost a half of the subjects keep friends on different sites for different purposes is because many of them also use social sites that are most popular in their own country, for example, Orkut[12] of India and Xiaonei[13] of China. They often keep friends in their own countries on these sites and friends in other countries on Facebook or Twitter. For most of these subjects, some of their friends still have accounts on different sites (see the result of Q2 in Figure 2.8). For example, their friends may have accounts on both Orkut and Facebook. These subjects want to view their friends activities of those friends in one place (see the result of Q5 in Table 2.3).

Table 2.3: Results Related to Blending Friends Function

	Yes		No	
Questions	Num	Percent	Num	Percent
Q1	7	43.75%	9	54.25%
Q3	13	81.25%	3	18.75%
Q4	11	68.75%	5	31.25%
Q5	12	75%	4	25%

As expected, from the subjects' answers to Q6, only three subjects have used the new function of grouping friends offered by Facebook or Twitter. All these three subjects provided positive answers to Q7, Q8 and Q9, which indicates that they like the function of grouping friends and think that the function is necessary. The subjects' answers to the questions (Q7, Q8 and Q9) related to the function of grouping friends were also very positive in support of this function, as can be seen from Table 2.4. Only one subject (out of 16) was consistently against this function.

The subjects' answers to Q12 are summarized in Table 2.5. The majority (81.25%) of the subjects support this function of allowing them to view their friends' activities by groups. They thought that this function of filtering social data by groups

---

[12]www.orkut.com
[13]www.renren.com

Table 2.4: Results Related to Grouping Friends Function

Questions	Yes		No	
	Num	Percent	Num	Percent
Q7	15	93.75%	1	6.25%
Q8	15	93.75%	1	6.25%
Q9	14	87.5%	2	12.5%

would provide much convenience for reading friends' updates. This result is further supported by the answers to Q10 that most of the subjects were overwhelmed by the number of their friends' updates. However, the result of Q11 (whether they want to tag friends and view their activities by tags) is not so significant. One reason may be that many (31.25%) of the subjects were still not overwhelmed by their friends' updates on one single social networking site. However, the number of updates will be much increased when integrating them.

Another reason may be that tagging requires effort. The subjects were not sure whether they want to spend much time on tagging rather than browsing through a long list of updates. Some subjects also felt that not many updates were important (and were not worth the effort of tagging them). They preferred to tag only important ones for them to be able to revisit later. Perhaps, this function will be more demanding when users have more and more friends. And, the function of filtering social data is certainly more useful for approaches like ours that integrate a user's social data on several different SNSs.

Table 2.5: Results Related to Filtering Social Data

Questions	Yes		No	
	Num	Percent	Num	Percent
Q10	11	68.75%	5	31.25%
Q11	9	54.25%	7	43.75%
Q12	13	81.25%	3	18.75%

In the interface testing phase, we closely observed the process of performing the six tasks (logging into the application, blending two friends, creating a group of friends,

tagging one friend, and searching a tag) by each subject. They succeeded in finishing all tasks within a reasonable time period (less than 30 seconds for each task). Some subjects were confused by the interface for blending friends. This function requires them to drag two or more friends from the friends list and drop them into a separate box. Another cause of confusion is that the friends in the list from Facebook and Twitter were not sorted by name, so it was hard to find a particular user in each of the list for blending. This finding task in an unsorted list had to be perform twice (once to find the friend in the Facebook friends list, and once - in the Twitter friend list), which was work-intensive and inconvenient for people with long friends lists. Therefore a new requirement was found - that the lists of friends available for blending had to be sorted and even a search function would be useful. Also the operation of "drag and drop" is not straightforward especially for those subjects who are not in Computer Science (thus less familiar with interface design). After some wondering around and trying other ways, they all finally managed to accomplish the task.

After the interface testing phase, we asked the subjects for feedback about the interface and suggestions for improvement. They all thought that the functions provided through the interface were intuitive. They suggested to provide more textual explanation for the function of blending friends, to avoid confusion. They also suggested to include other functions, such as a "Help" function to provide more detailed help information about the application and its interface. In terms of tag recommendations, they suggested that user profiles, shared activities and the most important keywords in updates are useful for generating meaningful tags. All these feedback and suggestions were beneficial for refining and extending the functionality and interface of the SocConnect implementation.

CHAPTER 3

DEALING WITH NETWORK OVERLOAD

This chapter focuses on the recommender aspect of SocConnect. Section 3.1 serves as a literature review for related works on recommender systems. Section 3.2 explains how the recommendation functionality is incorporated in SocConnect. Section 3.3 presents a visualization technique used to present the recommendation results. Section 3.4 presents a study comparing how different machine learning algorithms perform on particular selections of features in social data to generate recommendations in SocConnect.

3.1 Literature Review

3.1.1 Recommender Systems

There is a lot of research in the area of recommender systems dating back from the mid Ninety. There are two main types of recommender systems: content-based (or feature-based) and collaborative (social). Content-based recommenders analyze features of the content in the set and match them to features of the user (e.g. preferences, interests), based on a user model developed by analyzing the previous actions of the user. The problem with these recommenders is that creating models of users is time consuming and domain-dependent. Collaborative (social) recommenders [30] avoid these problems, since they work by statistically correlating users based on their previous choices. Based on the assumption that people who have behaved similarly in the past will continue to do so, these recommenders suggest content, rated highly by a user, to similar users who have not seen the content yet. While collaborative recommenders do not require domain specific design, they suffer from the "cold start" problem, because they need a lot of ratings to correlate user's previous choices

and find similar users. There are also hybrid recommender systems, which deal with the "cold start" problem by starting with content-based recommendation, and once sufficient amount of ratings has been accumulated, collaborative recommendation algorithms are applied.

Collaborative (social) recommend systems are widely used to recommend movies, books, or other shopping items in e-commerce sites. More recently, recommender systems have been applied in SNSs. There are still relatively few academic works in this area. SoNARS [15] recommends Facebook groups. It takes a hybrid approach, combining results from collaborative filtering and content-based algorithms. 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. Many existing SNSs use social network analysis to recommend friends to users. This, however, does not help in dealing with information overload. This research focuses on recommending status updates. Status update is different from items like movies, book, or shopping goods in two ways: first, the number of status updates arrive in large volumes, and are only relevant for very short time; second, a status update is more personal and aims at a small audience. Due to these two features, a collaborative recommendation approach is not a good solution: collaborative filtering works well for a large group of similar users and requires previous ratings.

Another related work [16] proposed and implemented a content-based recommend, but just for Twitter. Their work focuses on recommending URLs that appear in the tweets by people who are in user's network (followed by the user). Their approach is based the assumption: that the user's friends (who are followed by the user, and called "followees" in Twitter context) belong to a neighbourhood which shares one or more interests. In contrast, this research can recommend all kinds of status updates, with or without URL, text, and image. The SocConnect approach does not assume that the followed friends form one neighbourhood with shared interest(s). Often, users have to follow other users due to a social norm, because of the relationship they have with the other users e.g. they have to add as friends list their bosses or their mothers.

3.1.2 Text Recommendation

Another relevant area text recommendation in the field of Information Retrieval and Personal Information Management, since much of the social data is textual, e.g. many status updates can be considered as text documents. Text recommendation usually has four steps [11]: (1) recognizing user interest and document value; (2) representing user interest; (3) identifying other documents of potential interest; and (4) notifying the user - possibly through visualization.

To recognize user interest, there have to be a measurable interest indicator or indicators. Previous related work uses implicit interest indicators, explicit interest indicators, or a combination of both [11]. Explicit interest indicators, such as rating, allow the user to give the system direct feedback about the how much she likes the document. Explicit interest indicators are fairly reliable and easy to implement, however they are intrusive because they can interrupt the user's normal pattern of browsing and reading [18]. On another hand, implicit interest indicators, such as display time and number of mouse clicks, could indicate user preference and requires no user action, but their effectiveness is not certain and it is context-dependent [11]. Currently SocConnect uses the explicit interest indicators.

To recognize document value, there has to be a model to present each document. Text document representation is a major research topic in the field Information Retrieval(IR). The common models of representing text documents are Vector Space Model (VSM) [31], Standard Boolean Model (BIR) [24], and Probabilistic Model [33]. Among them, vector space model is the most widely used one and the one used in SocConnect.

A vector space model represents a document or documents by the terms occurring in the document with a weight for each term. The weight represents the importance of the term in the given document. The most common two ways to calculate the weight are Term Frequency (TF) and Term Frequency - Inverse Document Frequency (TF-IDF).

TF is simply counting how many times each term occurs in the given document,

defined as follows:

$$\text{TF}_i = \frac{N_i}{\sum_i N_i} \tag{3.1}$$

TF-IDF takes into account not only the importance of the term in the given document but also the general importance of the term across all documents, based on the number of documents containing this term. It can be defined as follows:

$$\text{TF-IDF}_i = \text{TF}_i \times \lg \frac{|A|}{|A_i|} \tag{3.2}$$

where $|A|$ is the total number of documents, and $|A_i|$ is the number of documents containing the term.

Cosine Similarity Measure is a popular method to calculate the similarity between vector spaces. It can be defined as follows:

$$\text{Similarity} = Cosine \ \Theta = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{3.3}$$

where A and B two vector space models, $A \cdot B$ is dot product of vector space model A and B, and $\|A\| \, \|B\|$ is the product of the magnitude of vector A and B.

3.2 Proposed Way of Recommending Updates from Social Networks

To relieve the user's network overload, SocConnect provides personalized recommendations of activities to individual users according to a prediction generated using their ratings on previous social data. Thus, SocConnect approach is a content-based recommendation, rather than collaborative. This section presents several machine learning techniques that are used to predict users' preferences on activities, a list of potential non-textual and textual features for representing each activity.

3.2.1 Learning User Interests on Activities

As mentioned Chapter 2 Section 3.1.2, this research uses the explicitly interest indicator to determine user's interest on activity. In SocConnect, users directly express their preferences on activities and friends by using the function of rating activities as "favourite" or "disliked". The users' ratings of their friends are also used in predicting users' interests in activities posted by these friends. Based on the ratings,

SocConnect can learn users' preferences and predict whether they will be interested in new similar activities. Machine learning techniques are often used for learning and prediction. SocConnect applies the classic techniques of Decision Trees, Support Vector Machine [28], Bayesian Networks, and Radial Basis Functions [27]. 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. Bayesian Belief Networks is a commonly used Bayesian learning technique. The method of Radial Basis Functions belongs to the category of instance-based learning to predict a real-valued function. Support Vector Machines have shown promising performance in binary classification problems.

3.2.2 Features for Representing Activities

All machine learning techniques listed above require a set of features describing the data. Social data is semi-structured, it contains both highly structured metadata, such as an activity has an actor, and unstructured text data, such user bio information and the message in status update. In order to effectively recommend social data, both structured and unstructured data need to be addressed in the machine learning. In the following subsections will refer the structured data as non-textual and unstructured as textual.

Non-Textual Features

Table 3.1 summarizes a list of relevant non-textual features and some of their possible values. Each activity has an actor (creator). SocConnect allows a user to rate friends as favourite or disliked; by default friend rated neutral. Using actor and actor's rate features, SocConnect 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 (i.e. favourite, neutral, or disliked friends). As discussed in Section 2.2, each activity has a type, for example, upload an album, share a link, retweet (more examples see in Table 3.1). SocConnect also take into account the SNS sources of activity, 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, SocConnect can find out whether a user is only interested in activities from particular SNS sources. Different applications used to generate those activities are also useful to consider. For example, if a user's friend plays "MafiaWars" on Facebook but the user does not, the status updates generated from the "MafiaWars" application may be annoying to the user.

Table 3.1: Non-Textual Features of Activities for Learning

Non-Textual 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

Textual Features

SocConnect also considers the textual content of activities, even though many activities, such as video uploads, do not have any textual content. The purpose of having textual features is to investigate whether text analysis will contribute to the personalized recommendation of social activities.

In the text analysis part, SocConnect first removes the stop words and URL links in each activity. Two vector spaces are then calculated for each activity; one is using TF and another one is using TF-IDF. The reason of using both weight calculation algorithms is to investigate whether the commonality (IDF value) of terms plays a role in the data mining process in the context of analysis social data.

Having the vector spaces for each activity and given training data containing a set of activities rated by a user as favourite, neutral or disliked, SocConnect sums up the weight values for each term in all the favourite, neutral and disliked activities, respectively. The results are three vectors over the training data, for the favourite, neutral and disliked activity sets respectively. Each vector consists of the total weight

of each term in all activities of the corresponding set (either favourite, neutral or disliked activity set). SocConnect then calculates the cosine similarity between a vector representing each activity and the three vectors representing the favourite, neutral and disliked activity sets, denoted as S_F, S_N and S_D, respectively. Each of these similarity values can represent a textual feature for activities. The range of the similarity is from 0 to 1, where 0 means totally irrelevance, and 1 means totally relevance (include exactly same)

SocConnect also use one combined textual feature C for an activity. Two ways can be used to represent this feature. One way is to use a numeric value presents difference between the two similarity values, $C = S_F - S_D$. Another way is to use a nominal value to represents the difference as favourite, neutral, or disliked, as follows:

$$C = \begin{cases} \text{favourite} & \text{if } 0.33 < S_F - S_D \leq 1 \\ \text{neutral} & \text{if} -0.33 \leq S_F - S_D \leq 0.33 \\ \text{disliked} & \text{if } -1 \leq S_F - S_D < -0.33 \end{cases} \qquad (3.4)$$

In summary, socConnect can have four potential textual features for representing activities, including S_F, S_N, S_D and the combined one C, as listed in Table 3.2. Note that the combined feature C can have a numeric value $(S_F - S_D)$ or a nominal one. Also note that the values of each feature summarized in Table 3.2 can be calculated based on either TF or TF-IDF.

Table 3.2: Textual Features of Activities for Learning

Textual Features	Possible Values
S_F	$\in [0, 1]$
S_N	$\in [0, 1]$
S_D	$\in [0, 1]$
C	$S_F - S_D \in$ [-1, 1]; or nominal interest levels: $\in$ {favourite, neutral or disliked}

3.3 Adaptive Presentation of Recommendations in Visualization

As pointed out in Section 3.1.2 the four steps of text recommendation [11]. After identifying other documents (in this case updates) of potential interest, the system should notify the user -possibly through a visualization.

Webster and Vassileva [34] tested a visualization technique with different colour metaphor indicate the levels of interestingness of the post in an online community called Comtella-D 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. Thus, SocConnect uses this visualization technique for the recommendation result.

Table 3.3: Interest Level and Colour Presentation

Interest Level	Colour
Favourite	Persimmon
Neutral	Maroon
Disliked	Thyrian purple

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 (See Fig. 3.1). Colours in a spectrum that can be distinguished by people with the most common type of colour-blindness (red-green) to distinguish,[1] is used to represent if an activity is recommended or unrecommended according to the predicted interest level calculated for the activity (Table 3.3). 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 in bright orange background, yellow text) and "cold" item (dark purple background, blue text).

---

[1]Images can be tested for appearance with simulated colour blindness at: http://www.colblindor.com/coblis-color-blindness-simulator/
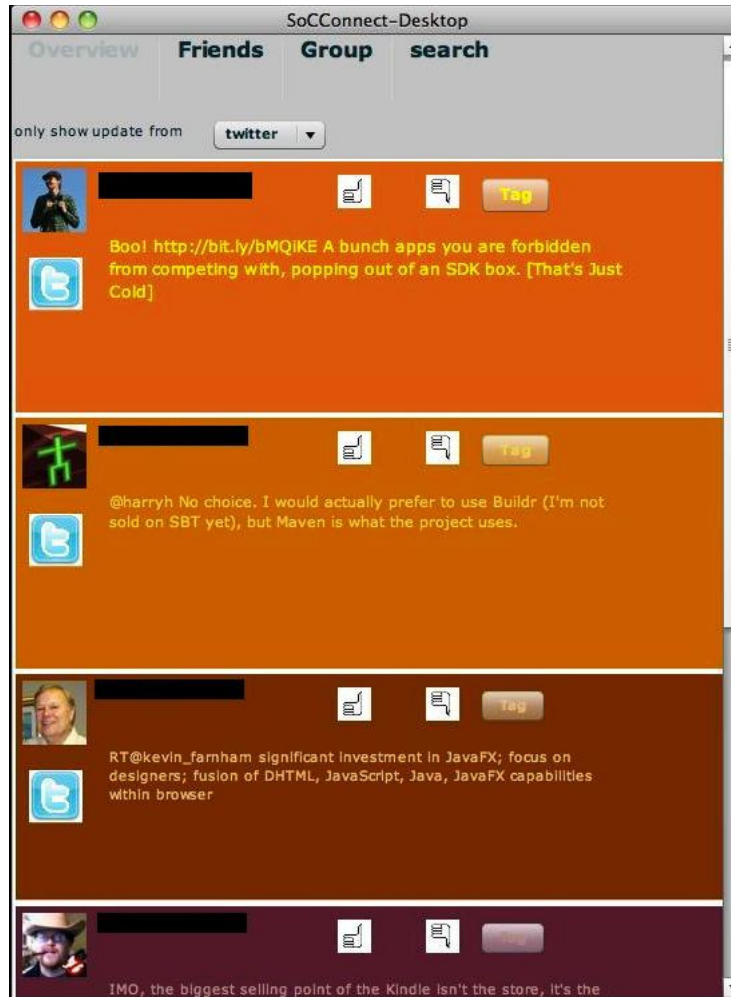
Figure 3.1: An Exampleof Visualization

3.4 Evaluation of Different Algorithms Applied to Social Data

3.4.1 Goals

The goal of this study is to evaluate the the accuracy of the prediction of the four machine learning techniques with different sets of activity presentation features. We carried out experiments to evaluate 1) the performance of the four machine learning techniques for learning user preferences on social activities and 2) the performance of personalized recommendations when different features are used to represent social activities.

3.4.2 Experimental Setup

Social data streams from ten subjects were used in the evaluation. Five of the subjects are from Saskatoon, Canada, and the other five are from New Jersey, USA. A half of them are students and the other half are workers. Six of the subjects are experienced users of Facebook and Twitter. For each of these subjects, we collected from Facebook and Twitter 200 recent activities of their friends. The other four subjects are relatively new users of Facebook and Twitter. For each of them, we collected around 100 recent activities of friends. Thus, in total, we collected around 1,600 user activities. We asked all subjects to rate their friends and activities, presented to them in a spreadsheet. On average, they rated 38% of their friends as favourite or disliked friends and 45% of the activities as favourite or disliked. Thus, the data sample is quite diverse. A 10-fold cross validation was performed on the collected data from each subject, and the average performances of the machine learning techniques over the activities of all subjects are reported in the following sections.

3.4.3 Performance when Using only Non-Textual Features

We first used only the set of non-textual features summarized in Table 3.1. Fig. 3.2 shows the performance of the four machine learning techniques. Although the performance difference among these techniques is not significant, support vector machine (SVM) provides the best performance, and it correctly classifies 69.9% of instances in the testing data. RBF performs the worst (68.4%). The performance of Decision Tree and that of Bayesian Belief Networks are about the same, which is around 69.5%. So, these machine learning techniques generally do not show good performance when only the non-textual features are used for representing activities.
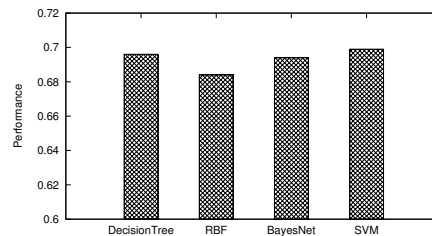

Figure 3.2: Performance when only Non-Textual Features are Used

3.4.4 Performance when Using only Textual Features

We then evaluated the performance of personalized recommendations on social activities when only the textual features summarized in Table 3.2 are used. In this set of experiments, we first tested the performance when the combined feature C is used. All the four machine learning techniques perform the same and achieve 64.9% of correct prediction. In addition, there is no difference when TF or TF-IDF is used as term weight. Using this feature alone shows even worse performance than using the non-textual features.
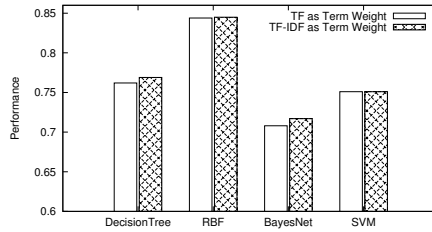
Figure 3.3: Performance when Three Textual Features are Used

We then tested the performance when the other three textual features (S_F, S_N and S_D) are used. The results are plotted in Fig. 3.3 when TF and TF-IDF are calculated for term weight respectively. We can see that now RBF performs the best (84.5% of correct prediction). RBF is known as generally showing good performance when the values of features are continuous, as it predicts a real-valued function. Decision Tree is the second best and has the performance of 76.9%. SVM is better than Bayesian Belief Network in this case. We can also see that there is still no much performance difference between TF and TF-IDF. From the evaluation results presented in this section, it is also clear that the performance when the three textual features are used is significantly better than that when the combined textual feature C is used and also better than the performance when non-textual features are used.

3.4.5 Using both Non-Textual and Textual Features

We further evaluated the performance of personalized recommendations on social activities when non-textual and textual features are both taken into account. We first use the combined feature C and the non-textual features. As described in

Table 3.2 in Section 3.2 , four different ways can be used to calculate the value for the feature C of an activity, listed as follows:

- TF+Numeric: weight of term is calculated using TF and feature value is calculated by $S_F - S_D$;

- TF+Nominal: weight of term is calculated using TF and feature value is represent a nominal value by $S_F - S_D$ in the three interest levels;

- TF-IDF+Numeric: weight of term is calculated using TF-IDF and feature value is calculated by $S_F - S_D$;

- TF-IDF+Nominal: weight of term is calculated using TF-IDF and feature value is represent by a nominal value $S_F - S_D$ in the three interest levels.

Table 3.4: Performance when Using C and Non-Textual Features

Methods	DecTree	RBF	BayesNet	SVM
TF+Numeric	0.777	0.793	0.773	0.764
TF+Nominal	0.712	0.704	0.711	0.716
TF-IDF+Numeric	0.780	0.794	0.761	0.749
TF-IDF+Nominal	0.718	0.698	0.713	0.718

The performance of each method is summarized in Table 3.4. We can see that the methods without mapping to interest levels produce the better performance than those with mapping. There is no much difference between "TF-IDF+numeric" and "TF+numeric" or between "TF-IDF+Nominal" and "TF+Nominal". Thus, calculating term weight using TF-IDF does not provide much contribution to the personalized recommendation of social data. The performance when using both the combined feature C and the non-textual features (up to 79.4%) is much better than that using each alone (up to 69.9% with non-textual features and 64.9% with only the combined feature C).

We then use the combination of the three textual features (S_F, S_N and S_D) and the non-textual features. The results are plotted in Fig. 3.4 when TF and TF-IDF are calculated for term weight respectively. Again, there is no much performance

difference between TF and TF-IDF. RBF performs the best (81.4%). Decision Tree and SVM perform similarly (around 80%). Bayesian Belief Network is the worst in this case (around 75.2%).
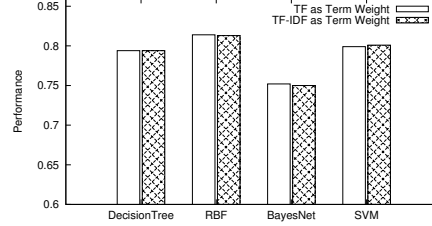
Figure 3.4: Using S_F, S_N, S_D and Non-Textual Features

We compare the performance between different textual features when the textual features are integrated with the non-textual features. In this comparison, we choose the best performance of the combined feature C. The result obtained is similar as that when only textual features are used, as shown in Fig. 3.5. In most of the cases, the three textual features provide better results than the combined feature. Bayesian Belief Network is the exception. The result concludes that it is generally better to use the three features separately instead of combining them.
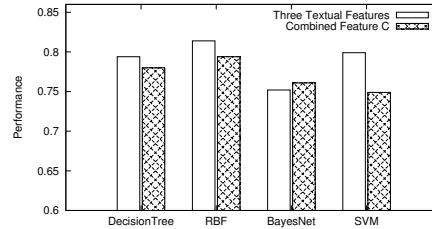
Figure 3.5: Performance Comparison between Textual Features

3.4.6 More Analysis

To further analyze the obtained evaluation results, we also plot the performance of personalized recommendations when using only non-textual features, when using only textual features of S_F, S_N and S_D, and when using both, respectively in Fig. 3.6. We can see that in general, the best performance of the machine learning algorithms is produced when both non-textual and textual features are used. Thus, both non-textual and textual features contribute to the personalized recommendations of social activities. Note that RBF is exceptional. Its performance when using both non-

textual and textual features is worse than that when using only textual features. Integrating discrete values of non-textual features degrades its performance. We analyzed the evaluation results using two factor ANOVA (analysis of variance) test with replication with 0.05 p-value and the analysis shows that the difference between the performance of the combined approach and the other two approaches (textual and non-textual) is statistically significant. The ANOVA analysis did not show significant difference in the performance of the four tested machine learning algorithms. The combined text and non-text features approach yielded significantly better results with all four algorithms.
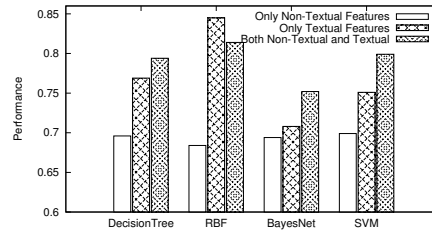
Figure 3.6: Performance Comparison for Different Features

Using Weka's feature selection function, we can see which features are more important for individual users. We summarize in Fig. 3.7 the number of subjects for whom each feature was the most important one in the prediction. In this experiment, non-textual features and the three textual features (S_F, S_N and S_D) are used because they produce the best performance for most of the machine learning algorithms.

For most of the users, the three textual features are important. This implies that most of the users are interested in the textual content of their friends' activities. "Activity Type" is also important for most of the users. For half of the users, "Application" is important. "Actor Type" is important for three 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 the textual content of activities. Some users care about the type of their friends' activities. Some users care more about the applications that generate the activities, which are usually the games they are playing. And, some users care about their close friends' activities. The implication

is that learning the user type may be useful in personalized recommendation of activities. We leave this for future work.
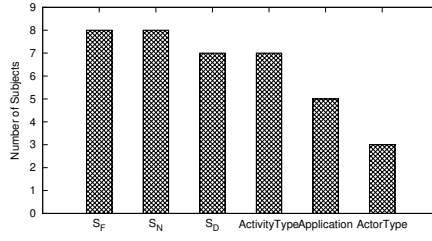
Figure 3.7: The Most Important Features

3.4.7 Limitations

The size of dataset is limited, if we could recruit more participants, the evaluation results may differ. But the goal of this evaluation was get a baseline evaluation on the performance of data mining techniques.

3.4.8 Conclusions

Several important conclusions can be drawn from the evaluation results presented in the this section: a) a combination of both non-textual and textual features contributes to the personalized recommendation of social activities; it performs significantly better than only textual or only non-textual features across all four algorithms; b) the best performance (84.5%) is produced by RBF using only the textual data, indicating that good performance can be achieved for the personalized recommendation of social activities; c) From current dataset, calculating term weight using TF-IDF does not show much advantage for textual features;

Chapter 4

Implementation

A prototype of SocConnect has been implemented to demonstrate the feasibility of the approach. This Chapter states how SocConnect is implemented in technical details. Section 4.1 focuses on the architecture of SocConnect. Section 4.2 describes the underlying technologies SocConnect based on. Sections 4.3 describes a series of efforts to improve SocConnect's implementation.

4.1 Architecture of SocConnect

In a nutshell, SocConnect follows traditional client and server (CS) architecture. In the front, a rich internet application (RIA) implemented in Adobe AIR acts as client interface and a RESTful (Representational State Transfer) Web Server as server.

The alternative architectures could be 1) a pure web application using HTML, CSS, and JavaScript as user interface, or 2) a pure desktop application without server-side. Each choice has its pros and cons.

Pure web application architecture needs no installation by the user, and requires least computation resources from the user. However, this architecture does all computations on the server side. More importantly, almost every SNS API has traffic limitations, for example Twitter's API currently traffic limit is 150 per hour for basic authentication. Thus, API traffic limit is a bottleneck for the pure web application architecture.

Pure desktop application architecture requires no external server and can easily work around API traffic limit. However, it requires the most computation resources from the client. Moreover, for research purposes, pure desktop application architecture makes evaluation and data analysis more difficult.

A client-server architecture balances the pros and cons of both pure web and pure

desktop architectures. Some of the computations, mainly the SNS API invoking, happens at client-side as a solution to API traffic limits. Other computations including recommendation happens on the server side. Usage data can be collected on the server side, and can be easily analysed.

4.2 The Full Stack Of Technology

The following list shows the full stack of technologies in SocConnect: Adobe Flex for client side desktop application, the Play framework as the web framework, Apache Lucene for text analysis, and Weka for data mining.

1. Adobe Flex

2. Play framework

3. Apache Lucene

4. Weka

4.2.1 Adobe Flex

Adobe Flex is an open source framework to build rich internet applications (RIA), which can be deployed on browser and desktops. For SocConnect, Flex is used for developing the client desktop application. To use Flex desktop application, the user must have pre-installed the Adobe Integrated Runtime (AIR). AIR is available for Window, Mac, Linux, even on some version of Android mobile platforms and Blackberry Playbook Tablet with possible modification. Alternatives techniques are Microsoft Sliverlight, Java SWT or AWT, and etc. The reasons for choosing Adobe Flex are: 1) that is open source, 2) cross-platform, 3) it is easy to develop some advanced visual features, such as drag and drop, view transitions, and customized skin for different visual components.

4.2.2 Play framework

On the server side, the intention was to keep all implementations based on Java for integrated development environment. For many years, Java developers have been

blessed and cursed with various web frameworks, such as Structs, Spring MVC, Wicket, Google Web Toolkit, Java Server Face, Roo, ZK, and many others. For SocConnect, the "play" framework was chosen [6] . The reasons for that are 1) play has a well integration with Hibernate[?], a relational persistence for Java and .Net, Memcache[?], a distributed memory object caching system, Selenium [?], a web application testing system, etc. 2) play fastens the development circle by reducing the recompile and re-deploy phases.

4.2.3 Apache Lucene

Apache Lucene [?] is a open source text search engine library written entirely in Java. SocConnect does not use Lucene's text search functionality, but the text analysis and computing vector space parts are based on Lucene's functionalities.

4.2.4 Weka

Weka [8]is a open source data mining software and library in java developed by the University of Waikato. It consists of a collection of machine learning algorithms for data mining tasks. Weka has capabilities for data pre-processing, classification, regression, clustering, and association rules [21]. Even in Java based data mining technology, Weka is not the only choice. There are other alternatives: RapidMiner, Apache Mahout, and other. Weka is easy to use in a Java program. RapidMinder is an integrated solution, and has its own server. It is less easy to integrate with existing Java code. Aapche Mahout is a very new project, at the point of writing, it is only at version 0.3.x. Currently, its implementation use Apache Hadoop, a map-reduce based distributed computing framework. Hadoop can only handle the data in its file system (HDFS Hadoop Distributed File System) format. For SocConnect, data exists either in memory or database. Writing and reading data via file is too costly.

4.3 System Implementation

To ensure a good user experience of the the prototype, more features were added, some existing features were improved, and many bugs were fixed. During the time,

Twitter dropped its support of the basic authentication, [1] and Facebook released its Graph API [2], with data representation different from Facebook's Old REST API [3]. SocConnect was updated according to these changes: Facebook and Twitter authentications workflows and user interfaces in SocConnect were changed to adapt the new authentication models of Facebook and Twitter APIs, two tool-kits were developed to make the communication with new APIs and data parsing easier.

4.3.1 Design of the User Interface

One suggestion received during the pilot study was that the interface of SocConnect needed more polish. A lot of effort was put to improve the interface. Figure 4.1a shows the new interface of SocConnect.

In the previous prototype, the user could not edit or delete a blended person or a group. These functionalities were added. Figure 4.3 is a screenshot of the " editing a group screen".

4.3.2 Motivating and Weighting Ratings

As mentioned previously, one of the goals of this user study was to test the effectiveness of the recommendation mechanism. The recommendations are based on the users' ratings of their friends' updates.(like/dislike) Without users' ratings, the recommendation mechanism cannot provide any recommendations. Related works [18], [26] show that users do not like to rate, because rating requires more attention, cognitive processing to make a decision, and thus, it interrupts the normal browsing. Therefore, to ensure sufficient ratings are entered by the participants, some motivations are necessary. This was addressed, similar to other work [26], by providing a reminder to users to rate the updates they receive. (See Figure 4.4)

4.3.3 Deployment of the Server

The SocConnect server was deployed on the built-in Play framework HTTP server. "The built-in play HTTP server can serve thousands of HTTP requests per second." [4]

---

[1] http://dev.twitter.com/pages/basic_auth_shutdown
[2] http://developers.facebook.com/docs/api
[3] http://developers.facebook.com/docs/reference/rest/
[4] http://www.playframework.org/documentation/1.1RC3/deployment#standalone
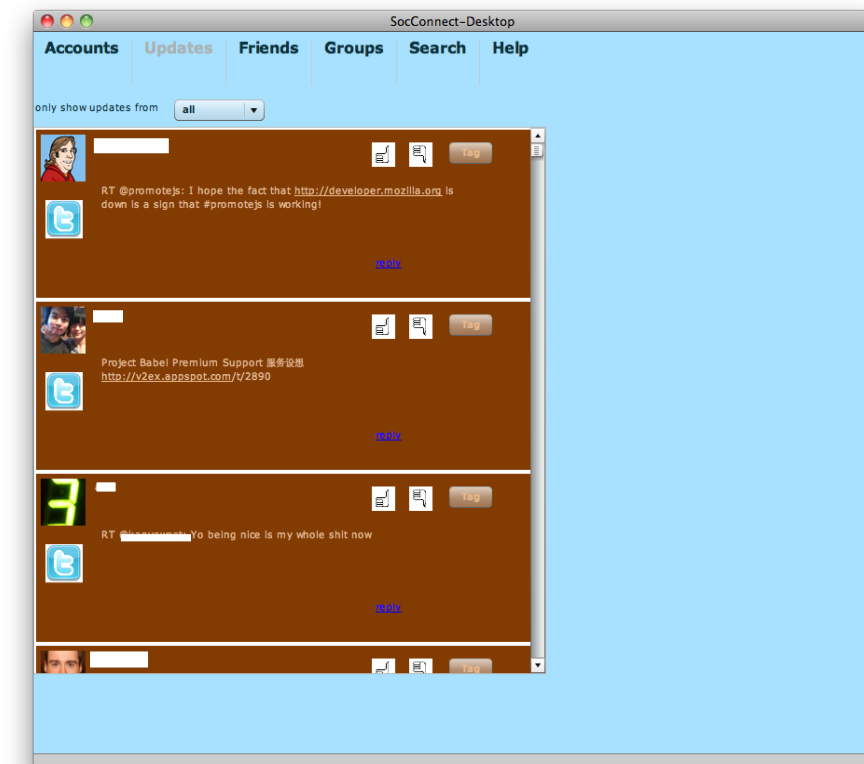
(a) Interface of SocConnect home
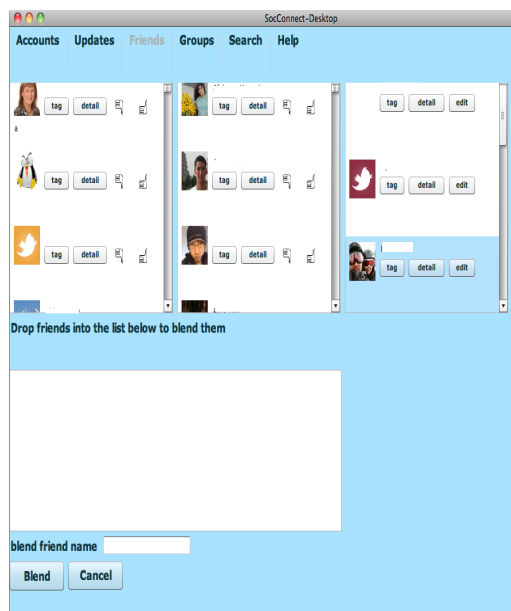
(b) Interface of SocConnect blend

Figure 4.1: SocConnect's Interface
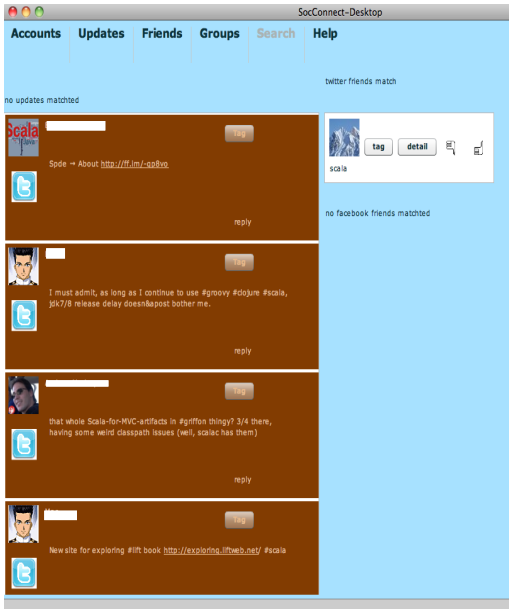
Figure 4.2: Interface of SocConnect search
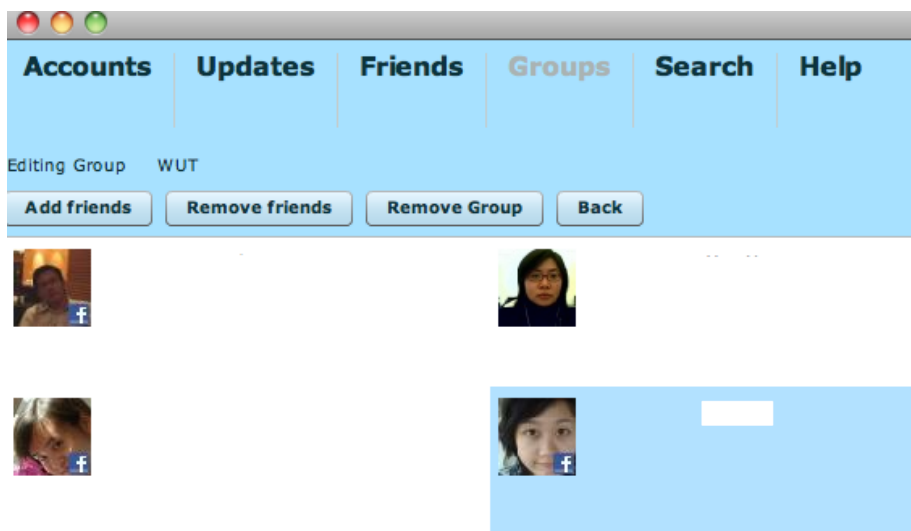
Figure 4.3: Editing a group
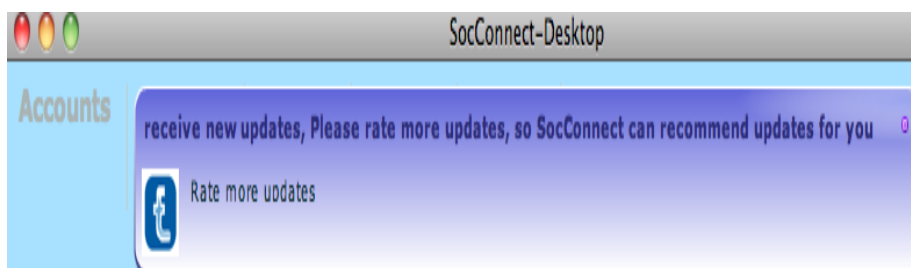
Figure 4.4: Reminder for Rate Update

Moreover it uses a more efficient threading model. Whereas Java EE standard servlet container uses one thread per request, Play's thread pool size is the number of processors plus one. So for a thousand concurrent HTTP requests, the servlet container needs one thousand thread while, Play only consumes two to four threads. The server runs on a ubuntu-10.04.1 OS based machine located at department lab. The Server uses Apache log4j 1.2 as the logging tool and HypersQL (hsqldb.org) file-based database. All the interactions between the SocConnect client and server, such as user login, rate, blend and group friend, tag and etc, are stored in the database.

Chapter 5

Evaluation of SocConnect

This chapter describes an evaluation of the SocConnect SNS aggregator and recommender with users in a field study. The chapter contains four parts describing the goals and methods of the user study, the preparation for the study, and the results of the study.

5.1 Goals

This study is a follow up of the two studies that were conducted earlier and described in Chapter 2 Section 2.5 and Chapter 3 Section 3.4. The pilot study aimed to gauge the users' the reactions to the proposed SocConnect's functionalities; the second study evaluated the recommendation algorithm candidates. The results from these two studies were positive and constructive. However, neither of these stuides evaluated the user experience with the complete SocConnect System. For example, the social update recommendation algorithm was evaluated "on paper" using a list of updates that were retrieved came directly from Facebook and Twitter rather than in the real SocConnect's system.

The overall goal of this field study is to evaluate the SocConnect's functionalities in real use over a period of time. The main hypothesis is that SocConnect provides users, with useful functions to manage their updates from Facebook and Twitter and with useful recommendations for update that can be of interest.

5.2 Methods

The method of this field study is to recruit Facebook and Twitter users to download the SocConnect dashboard application on their computer after signing consent to

participate in the study, and use it in an uncontrolled environment for two weeks. The SocConnect server logs the actions of users interacting with SocConnect. After two weeks of usage, the participants fill a user satisfaction survey. The survey has three sections: basic information, functionality feedback, and general questions. The basic information section collects the participants' contact information, Twitter and Facebook usernames, estimated frequency of Twitter and Facebook usages, whether they use other client applications to watch Twitter and Facebook beside the original websites (www.facebook.com and www.twitter.com), whether the participants have used any social network aggregator before, and whether the participants prefer to view Facebook and Twitter updates together. The functionality feedback section is organized into several sub-sections, each of which collects participants' feedback on the major SocConnect's functionalities: blending friends, grouping friends, tagging friends, searching by tags, SocConnect recommendation, rating updates. The general questions section collects participants' overall user satisfaction with Soc-Connect along the dimensions "Like the functionality", "The functionality is necessary", "The functionality is easy to use". For these questions, the answers are Likert-scale with 5 options, from "strongly agree"" to "strongly disagree". For the functionalities "recommendation" and "rate", there are a few additional questions. In the recommendation subsection, participants are asked how much they agree with the recommendation results, and how intuitive they find the visualization color. The rate subsection asks how much participants are willing to rate to gain a better recommendation results, how easy it is to decide to rate one activity, and whether the participant has a consistent rating criterion. The full questionnaire is attached in Appendix C.

5.3   Preparation

This section presents how the study was prepared. The preparation involved two tasks: setting up experimental environment, and recruitment of participants.

Figure 5.1: Help Section

5.3.1 Experimental Environment Preparation

The SocConnect client installer is stored on Google Storage (code.google.com/apis/storage). Users need to go http://socconet.appspot.com, read and agree with consent form (See Appenix A), and then download the client installer.

A google document,[1] which is publicly available, and google code project [2] was created. The google document served as user manual, the google code project had an issue tracker, the users could create new issues to report bugs in SocConnect.

An online survey [3] was created using an online survey service called "surveymonkey"[4].

5.3.2 Recruitment of Participants

Invitations were sent through email and the study was advertised on Twitter and Facebook. Figure 5.2 and Figure 5.3 show the advertisements on Facebook and Twitter. Twenty-two participants have responded to the advertisements and installed the

---

[1]https://docs.google.com/document/edit?id=1-NszHuuQ0HLdNsuf7rw-MzKLSakGE_AO3YdKc7bqD08&hl=en
[2]http://code.google.com/p/socconect/
[3]http://www.surveymonkey.com/s/socconnect
[4]http://www.surveymonkey.com

Figure 5.2: Advertising on Facebook

Figure 5.3: Advertising on Twitter

application.

5.4 Results

This section presents the results of the user study. The results come from two sources: SocConnect server logs and the online survey. The section contains three parts: overview of the results, the detailed report on the usage and participant feedback for each function SocConnect provided, and discussion.

5.4.1 Overview

During the two weeks study, twenty participants installed and used SocConnect. Thirteen of them used both Twitter and Facebook; seven only used Facebook. However, ten of twenty participants only used SocConnect once. Eight of these ten participants basically had no interactions with SocConnect whatsoever. Figure 5.4 shows the connection times of all participants, which SNS they used in SocConnect, and whether they answered the survey.
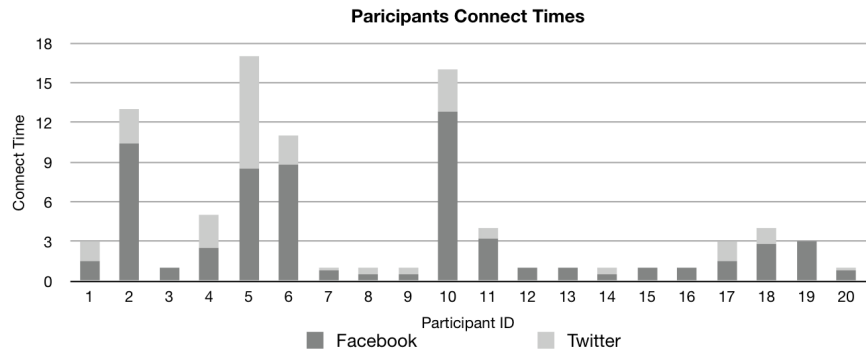
(a) Participant Connection Times and SNS Usage Percentage
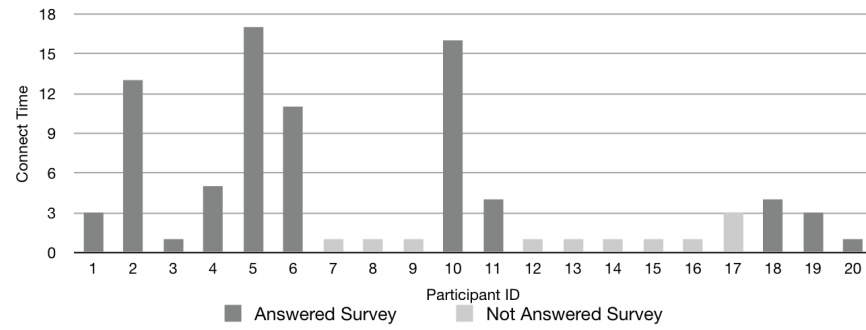
(b) Participant Connection Times and whether answered the survey

Figure 5.4: Participants Connect Times

Various reasons may have to contributed these eight participants' not using the system. A Facebook API bug related issue maybe have been the deciding factor: In the Facebook Graph API News feed[5] ,the author of a comment could be "null'. This API bug was not very common and was not documented, but it made Soc-Connect (before version 0.9.5.1) unable display any Facebook updates in certain circumstances. During the development and testing, that bug was never captured and was not documented on any Facebook API documentation page. Until one week in the the user study, that bug had been not noticed, after that a newer version of SocConnect was released to fix that bug. However, some participants never updated their SocConnect after the fix. From the SocConnect server logs, it can be seen that eight participants have never retrieved any Facebook updates in SocConnect, so one can assume that SocConnect encountered that bug or other similar bugs in these participants' cases. Moreover, five of these eight participants only use Face-book, which means that SocConnect could not display any updates from them. This easily explains why these participants never had any interactions with SocConnect. Of course, it is also possible that these participants did not encounter the bug, but changed their mind and decided not to participate in the study.

After the study, eleven participants answered the survey. As can be seen from Figure 5.4b, most of them were active users, who accessed SocConnect several times over the two weeks, so they comprise a valid sample for the study. Therefore the remaining nine users who accessed SocConnect only once and did not fill the survey were eliminated from the study. In the remaining participants, some of them only logged once or twice. However, they do present some real world users, therefore, their data still included in the study. Based on their usage of Facebook and Twitter, the eleven participants could be divided into three groups: only using Facebook, using both Facebook and Twitter but using Twitter less frequently, and using both Face-book and Twitter frequently. Two participants only used Facebook, six participants used both Twitter and Facebook, but used Twitter less frequently and followed few people (less than ten people). Some of the participants used Facebook very intensely and have many friends (one participant had almost a thousand friends). Another

---

[5]http://developers.facebook.com/docs/api

Function Actions of Questionnaire Responders
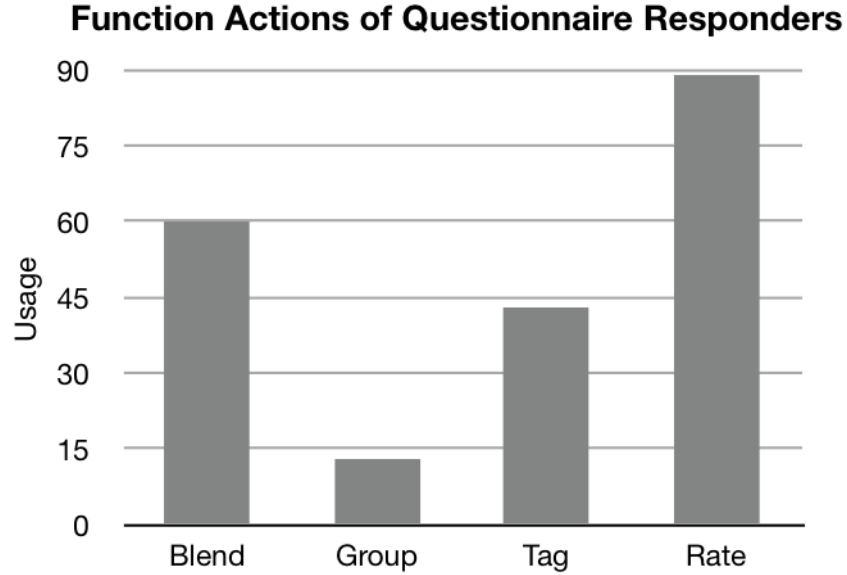
Figure 5.5: Overview SocConnect function usages

three participants used both Facebook and Twitter frequently and had many friends on both Facebook and Twitter; they had also experience with other social network aggregators.

Figure 5.5 gives an overview of the usage of SocConnect's main functions of the eleven participants answered the questionnaire. There are 61 times "blending" friends actions, 14 "group" actions, 43 "tag" actions, and 89 "rates" actions. Notably, one participant used "blending" friends 50 times, one participant "tag" 26 times, and one participant rated 44 times. The result is understandable since "tag" and "rate" require less effort than "blend" or "group".

The following subsections describe the usage data and user feedback about each function that SocConnect provides: blend, group, tag, search, recommend, and rate.

5.4.2 Blend Function Results

Figure 5.6a presents the usage of the "blend" function and Figure 5.6b shows the participants' feedback about this function.

For the two participants who only used Facebook, the "blend" function does not bring value and they never used it. Yet, both of them were aware of the function.

For the six participants who used Twitter less regularly than Facebook, one of them was not aware of this function and thought it was not easy to use. But the rest

liked the function, thought it was necessary to have this function, and most of them thought it is easy to use. Three participants stated that blending friends requires too much effort. Participants commented they wish SocConnect could give a hint about which two accounts should be blended together.

For those three participants who used both Facebook and Twitter regularly, all of them were aware of the "blend friends" function, but two were neutral on whether they like this function, another one liked the function. One participant commented that he uses "Facebook for personal friends and Twitter for Business", so the blend function did not provide any value to him.

5.4.3 Group Function Results

Figure 5.7a presents the usage of the "group" function by the eleven participants who responded to the survey, and Figure 5.7b shows the participants' feedback about this group function.

For the participants who only used Facebook, both of them were aware of the function and thought that it was easy to use. One participant commented that grouping was similar to "blend" since she did not use Twitter.

For the six participants who used Twitter less regularly, all of them were aware of the group function, and four of them liked it; for the another two participants who did not like it, one thought that the group function is similar to the blend function, and the other participant commented that he found no explanation about the group function.

For the three participants who used both Facebook and Twitter regularly, one participant preferred keeping Facebook for personal friends, and Twitter for business. The other two participants liked the group function, and thought it was necessary.

5.4.4 Tag and Search Functions Results

Figure 5.8a presents the usage of the "tag" function, and Figure 5.8b shows the participants' feedback about the function.

The participants' attitudes about the tag function were very divided. Two participants strongly liked the "tag" function, four liked it, four were neutral, one disliked

Blend Actions of Questionnaire Responders

(a) Blend function usages

Blend Feedback

- Strongly agree
- agree
- neutral
- disagree
- strongly disagree
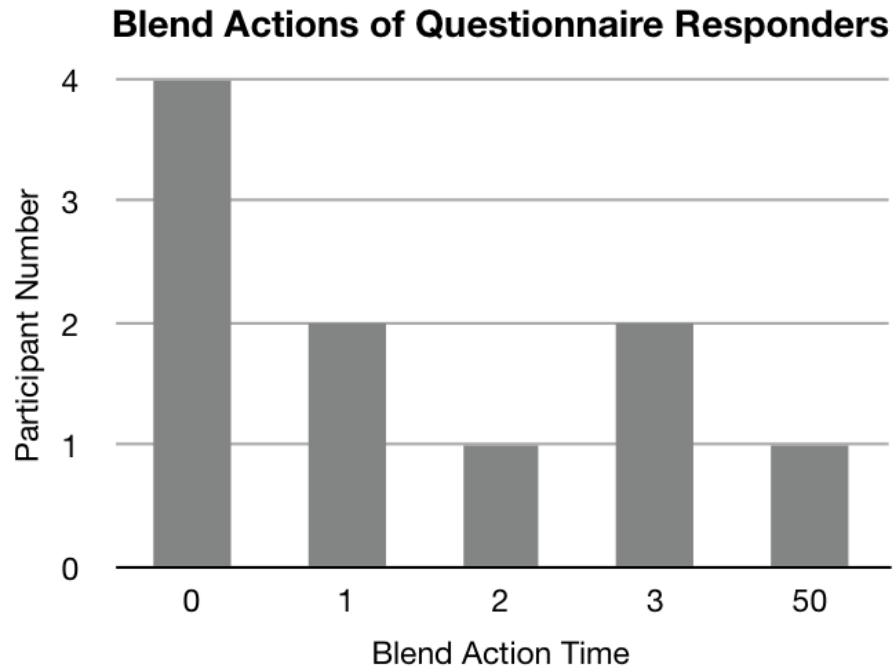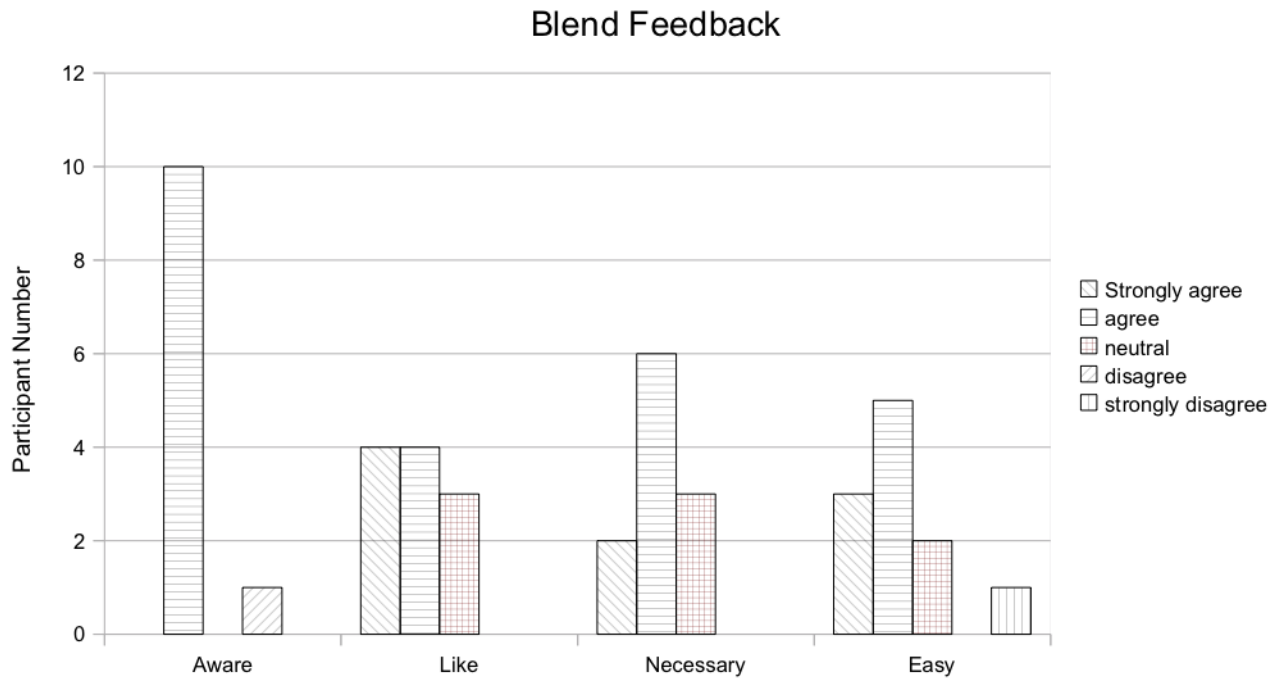
(b) Blend function feedback

Figure 5.6: Blend Function Usage and Feedback

(a) Grouping Function Usages
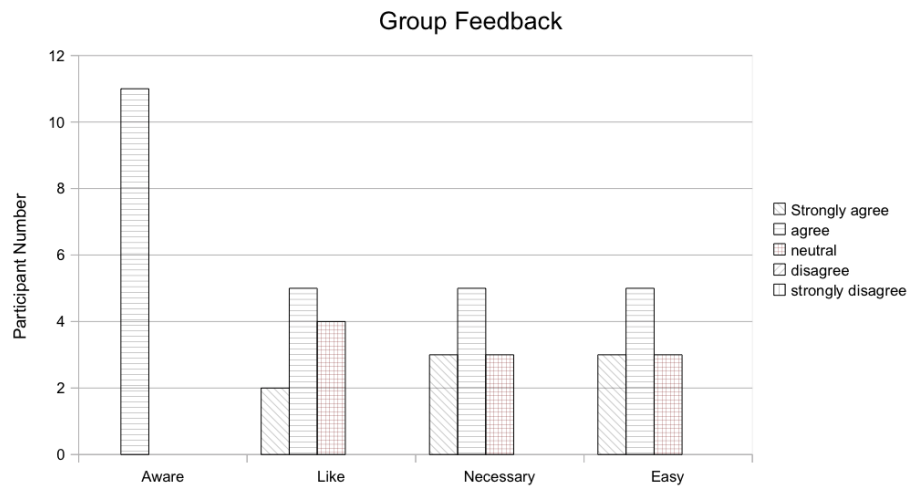
(b) Grouping Function Feedback

Figure 5.7: Group Function Usage and Feedback

the function. Three participants strongly agreed that "tag" function was necessary, two agreed it was necessary, four were neutral, two disagreed that it was necessary. Overall participants found that the function was easy to use. One participant stated that tag was her favorite function in SocConnect.

Figure 5.9 shows the participants' feedback on the search function. Unlike "tag", participants generally like the search function and thought it was necessary. Two participants thought that full-text search would be a nice function.

5.4.5 Rate and Recommendation Functions Result

Figure 5.10 shows the participants' feedback on the recommendation function of SocConnect. Three participants were not aware of this function, the largest number among all functions. One participant strongly liked the recommendation function, six participants liked it, the other three participants were neutral. Four participants strongly agreed that the recommendation function was necessary and another four participants agreed that it was necessary. Only two participants stayed neutral on whether the recommendation function was necessary. One of them was a participant used both Facebook and Twitter frequently, another participant was used Twitter less frequently than Facebook. Both of them commented that SocConnect did not generated any recommendations for them. Nevertheless the recommendation function received the most positive feedback on whether users liked this function and whether this function was necessary. This shows that users in the field study need recommendations on the updates they receive. However, the participants' feedback on the quality of recommendation was less positive, as can be seen in the last two histograms in Figure 5.10: only participant strongly agreed that the reco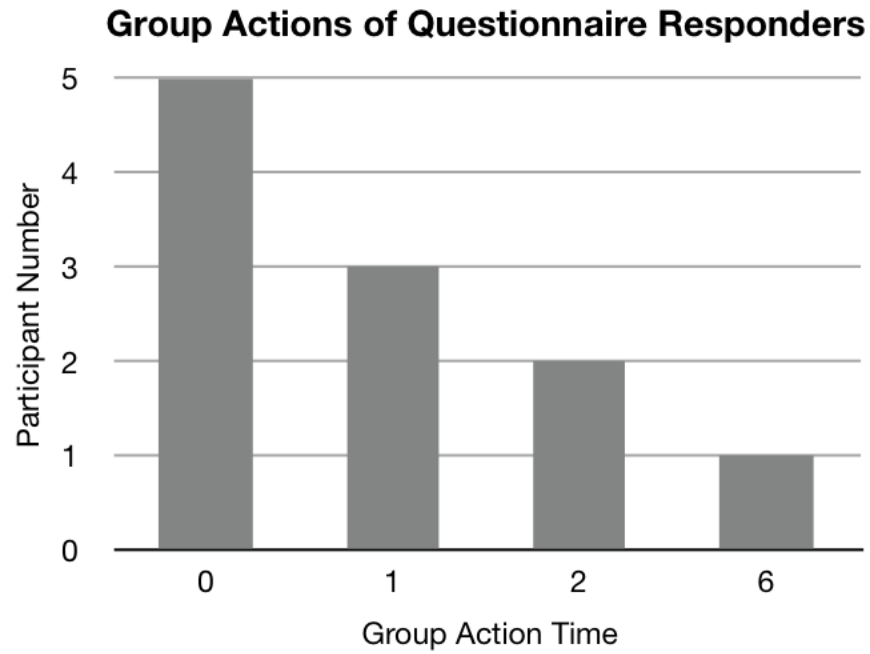mmendation result reflected her previous ratings, three participants agreed, while the remaining six participants stayed neutral on this question. The participant who strongly agreed with the received recommendation only rated once, so the generated recommendations were all neutral. Therefore, either the neutral presents her attitudes to those updates or the participant did not answered that question carefully. One participant rated forty-four times, but he did not recognize the recommendations SocConnect generated through the visualization. It is possible the highlighted up-
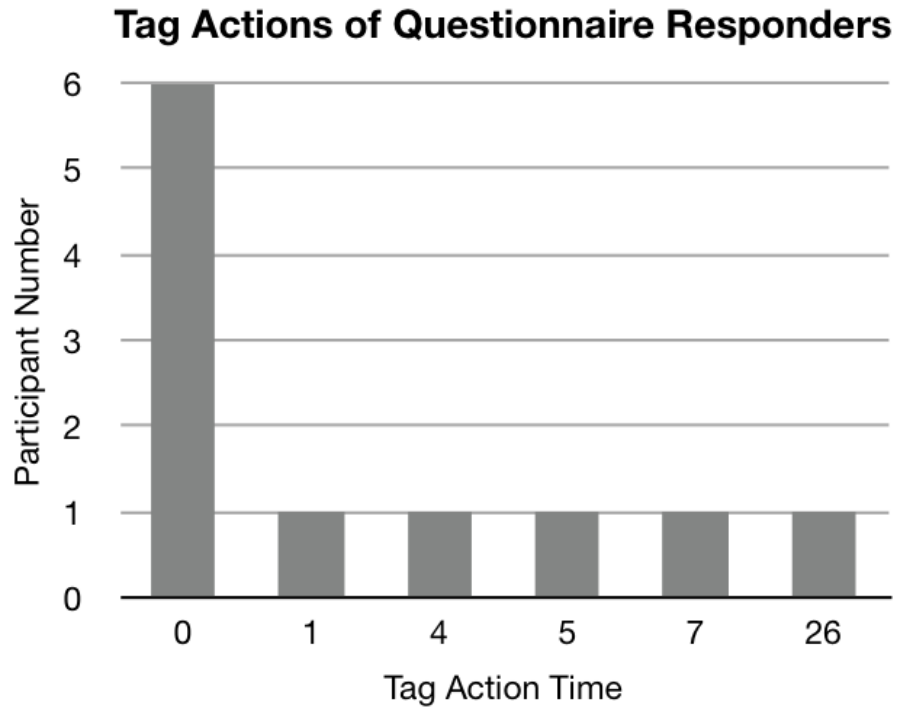
(a) Tag function usage
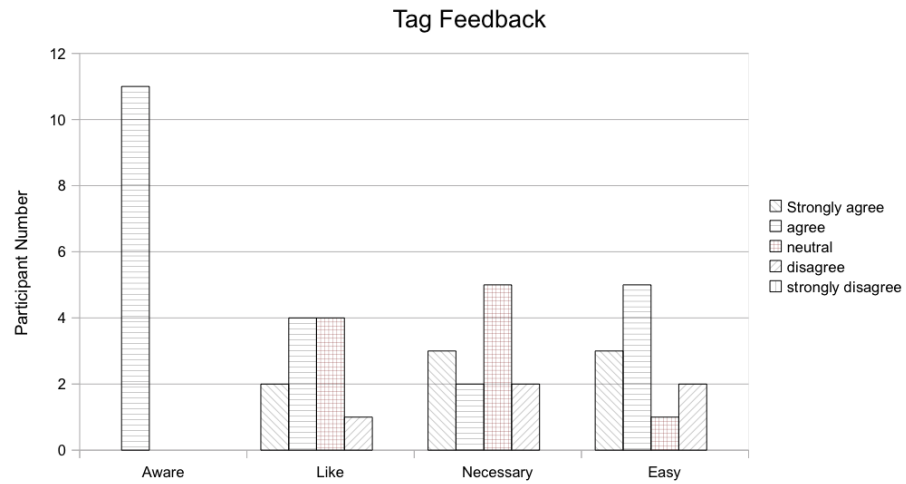
(b) Tag function feedback

Figure 5.8: Tag Function Usage and Feedback
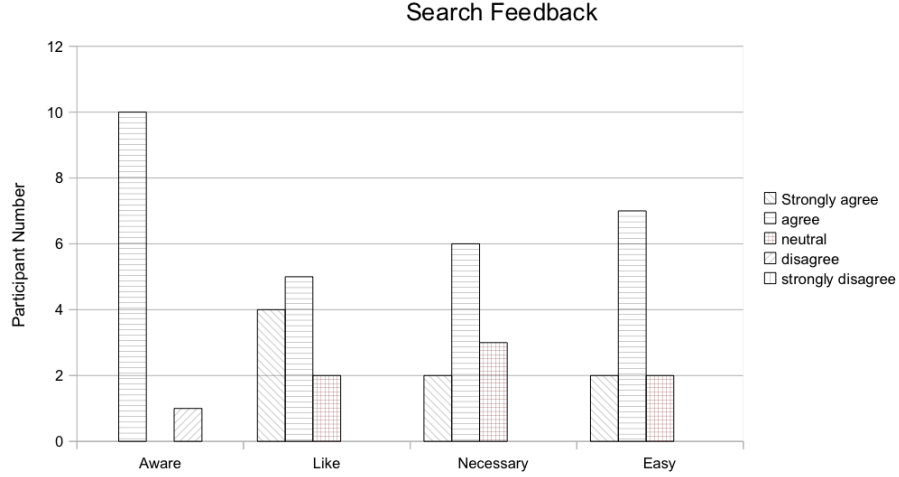
Figure 5.9: Search Feedback

dates were buried among the neutral updates, because that participant received five hundred and eighty-nine updates during the study. However, nobody disagreed with the statement, which gives some credit to the quality of the recommendations generated.

The recommendation visualization (described in Chapter 3, Section 3.3), and displayed Figure 3.1) seems not very intuitive. Only half of the participants found the colour intuitive. A bigger problem was that the highlighted updates were very likely to be buried away many updates requiring the user to scroll a lot to find highlighted items. Because users receive many updates, the recommended updates may be overlooked. The two participants, who rated the most, did not notice that SocConnect has generated recommendation for them. One participant suggested to separate the updates from the recommendations.

Further, the recommendation function was not transparent enough for the users. The current implementation of SocConnect reminds users to rate more updates to receive recommendations.The recommender algorithm requires at least ten ratings on ten different updates at the moment, before it can generate predictions for the user's liked activities. However, the user has no idea how many ratings are required, and whether the recommendation function is already working for him or her. One participant states this non-transparent should be fixed in future.

Figure 5.12 presents the usage of the "rate" function and figure 5.11b shows
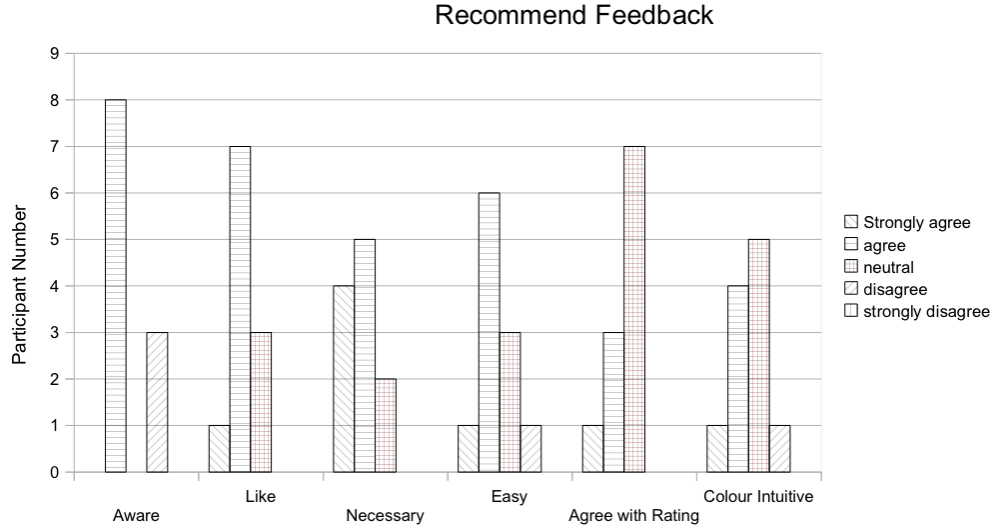
Figure 5.10: Recommendation Function Feedback

the participants' feedback on the "rate" function. Even though "rate" was the most commonly used function (see Figure 5.5), Only four participants rated more than ten times, and only six of the participants were willing to rate to get recommendations. Therefore the usage of the rate function shows a strong skewed distribution, shown in Figure 5.12. While the most participants agreed with the statements that they were aware the rating function, that they like it, that it was necessary and easy to use, slightly more than half of them (6) were willing to rate, and these six participants thought it was easy to decide how to rate.

5.4.6 General Feedback

Generally, the participants enjoyed using SocConnect and were willing to use it in the future. The participants suggested adding LinkedIn, Hi5, and renren (a Chinese SNS) in the future. Four participants suggested adding a "retry" function, so that users can directly respond to the updates from SocConnect.

5.4.7 Discussion

The overall goal of this field study was to evaluate the SocConnect's functionalities in real use over a period of time. The main hypothesis was that SocConnect provides users with useful functions to manage their updates from Facebook and Twitter and

61
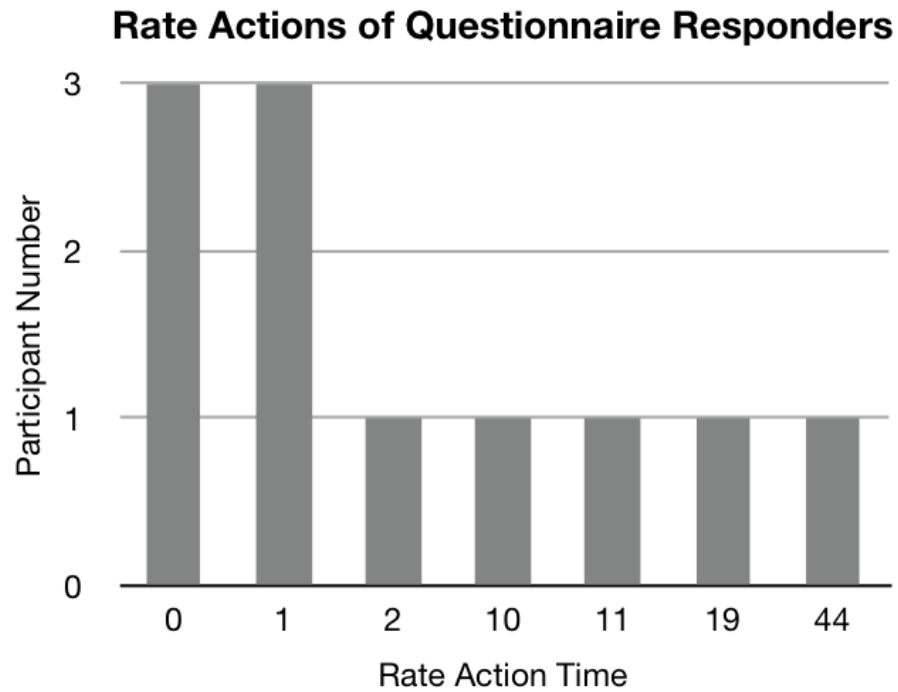
(a) Rate function usage
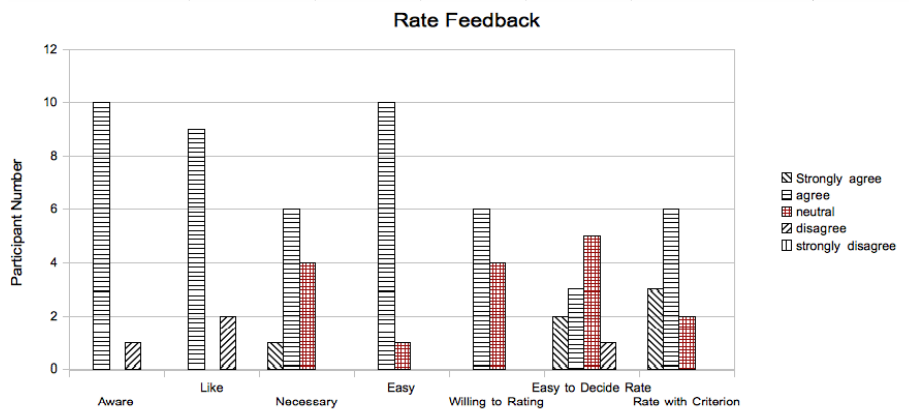
(b) Rate function feedback

Figure 5.11: Rate Function Usage and Feedback
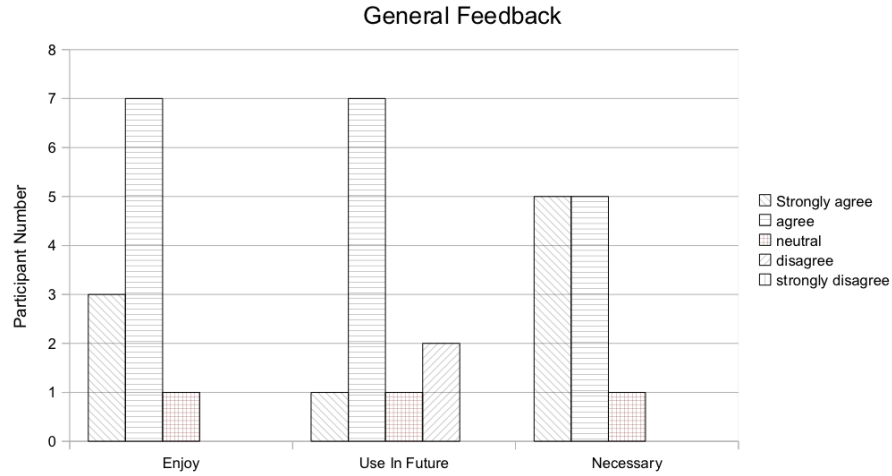
Figure 5.12: General Feedback on SocConnect

with useful recommendations for social updates that can be of interest. The results confirm the first hypothesis, that SocConnect provides users with a set of useful functions. Each functionality, except tag, was found useful, necessary and easy to use by the majority of the participants. The tag function was highly liked by only two participants. Interestingly, this result confirms our finding from the pilot study that users were reluctant to tag. The general feedback to the SocConnect system was overly positive.

Regarding the second hypothesis, the participants found that the recommendation functionality was useful in general , it was necessary and easy to use, but only few participants (4/11) agreed with the recommendations generated for them and found the colour visualization of the recommendations intuitive (5/11). This may have been due to the fact that participants did not provide a sufficient number of ratings to train the recommender. Yet, it points to a weakness in making participants aware that they need to rate in order to receive recommendations.

The results pointed to directions of improvement of some of the functionalities currently offered by SocConnect. To improve the blend function, SocConnect should suggest candidates to be blended. The possible implementation of this suggestion could be based on the friends' name, email address and other profile information. Some participants do not like to tag, but most of them like the search function, so possibly a full-text search on friends and updates across SNS would be a nice

function.

Recommendation is a wanted function, however, recommendation should not rely solely on ratings. Rating is a common activity on social sites with shared articles and blogs, but not so much on SNS (apart from specific signs of liking to the social application, like "Likes" or "Shares" which users do on their friends updates in Facebook, and "Favorite" and "Retweet" which users do in Twitter with update they find valuable. If data about like retweets, and favorites were obtained in SocConnect instead of rating, it would have allowed a less obtrusive, and possibly a better quality training data set for the recommender.). Rating is not be sufficient as the only source of user feedback for social recommendation.

The visualization of recommendations will also have to be redesigned. The visualization seems not effective, and the recommendation result may be overlooked by the users. An alternative is to show the recommended updates in a separated list. The new Google's Priority Inbox [6] uses this approach. This approach can clearly present the recommendation results to users, and users can provide further feedbacks on these recommendations. Yet the users may find it difficult to pay attention to two separate lists, one for recommended items, and one for general items, and it may in fact increase their information overload. More studies are needed in the area of visualizing recommendations.

5.4.8 Limitations and Challenges

There were two challenges in this field study.

1. SocConnect is a desktop application. Before using it, users had to install it. For every new version release, users had to re-install it. This caused a significant inconvenience that a web-based application could have avoided.

2. Twitter API was unstable from time to time. Sometimes, SocConnect could not display the users' Twitter data due to the unavailability of Twitter's API server. It was beyond SocConnect's control, but it could have decreased the user satisfaction and discouraged user participation.

---

[6]http://mail.google.com/mail/help/priority-inbox.html

There were three limitations in the study: the limited study period, and the selection of participants, the lack of control and possibility to observe the participants during the study.

1. For practical reasons, the third study lasted only two weeks. Considering that the user may not use SocConnect very often, two weeks is not a long period. In particular, the data mining algorithm required a bootstrap time to collect enough ratings from the user to form the training dataset. If the users did not give many ratings during the two weeks period, their recommendations were likely not good, since the machine learning algorithm could not make an accurate prediction.

2. It turned out to be too difficult to attract users who engage both on Facebook and Twitter to participate the study.The ideal participant in the study, would not only use both Twitter and Facebook regularly, but would also have previous experience using some existing social network aggregator to be able to compare the experience with SocConnect. However, to recruit participants with that experience was quite difficult. Therefore it is not possible to analyse "post-mortem" what exactly happened in particular case, e.g. why the user suddenly quit the application after just a few seconds. The reason might have been something out the control of SocConnect, e.g. a failure in the Twitter or Facebook APIs, which could have been easily caught during a lab experiment, and could have been removed from the data-set.

3. The field study was done in an uncontrolled environment. It was impossible to record from the side how user interacted with SocConnect and where they faced difficulties and had questions,even though most of their actions were recorded on the server. If the study was done on a controlled environment, e.g. in the lab, it would have been possible to know exactly the reasons for users not logging into the system, and to see what difficulties they encountered. Yet such a study would haven been limited in time duration and realism.

CHAPTER 6

SUMMARY AND CONTRIBUTIONS

This research aims to design an approach for personalized and intelligent integration of the data from SNSs. The approach attempts to solve two problems faced by users: "walled garden" and "networking overload" by providing functions like: blend, group, tag, search, and personalized recommendations.

6.1 Summary

The "walled garden" problem is that the user's data is scattered across different SNSs, it is difficult for user to browse her data across them. The "networking overload", a term invented by Christian Kreutz to describe the information overload on SNS, which comes in many forms: user may have too many friends; they generate many activities: adding new friends, joining new groups, posting status updates.

There have been many related works to solve these two problems. For the "walled garden" problem, the academic and open web community have developed standards for user data interoperability. Many social network aggregators have been released on the Internet. For the "networking overload", there have been already some approaches for recommending groups for users on Facebook.

SocConnect is a user-centric approach to social network aggregation, that empowers the user to attach contextual data to her friends and activities. It learns from the user's past actions and recommends social data (updates) that may be of interest to the user.

To represent the heterogeneous social data across SNSs, a unified schema is designed based on activitystream[1] and FOAF[2]. Inspired and based on the previous

---

[1]http://activitystrea.ms/
[2]http://www.foaf-project.org/

related work, SocConnect proposes four groups of functionalities: loading social data, managing friends, filtering social data, and recommending social data. "Loading social data" uses authentication methods provided by different SNSs and invokes their APIs to retrieves users' friends information and their activities on these sites. "Managing friends" contains two functions: blending friends and grouping friends. Filtering social data contains functionalities of tagging social data and searching by tag. Recommending social data contains the functionalities of rating friends and activities. The recommendation aims to learn the user's interests bases on her previous ratings by applying machine techniques and to predict her liking of new coming activities.

Three studies were conducted in this research. The pilot study focused on evaluating the functionalities design. The second study evaluated the accuracy of different recommendation algorithms as applied to social activities data. The field study evaluated SocConnect with users in real use. The first study results show the subjects were in favour of the functionalities proposed in SocConnect (except tagging which seem to be overkill according to some of the participants). The second study shows as some of the recommendation algorithms could reach fairly high accuracy (over 80%). For the field evaluation study, twenty participants installed SocConnect. Eleven of them used it for two weeks and filled a questionnaire. The results of the final study showed that the participants found the new functions provided by SocConnect necessary, easy to use and they liked using them.

6.2 Contributions

This work has three contributions:

1. Integration of social data from different SNSs. 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, tag their friends, groups, and social updates, and indicate as well as their interest level (favourite, neutral or disliked) for updates.

2. Personalized recommendation of activities that may be interesting to individual users.

3. Suggestion of a particular machine learning method for user preferences that has the best performance among five compared methods (SVM).

A new personal dashboard application, SocConnect, was developed to demonstrate the approach, which 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.

6.3 Future Work

6.3.1 Web Version of SocConnect

The current SocConnect dashboard only has a desktop version. The desktop approach has certain advantages: it could easily work around SNSs API traffic limit, and the computations shared by both client and server sides. But client-server approach causes the following obstacles to adoption and testing: 1. A user has to install SocConnect first. 2. Releasing a new version is difficult, since it requires users to download and reinstall the application. 3. Bug report on client side is difficult.

Web version is a better alternative to solve these problems. To implement a web version, the client side needs to re-implement; the server side require less, or no change. Currently SocConnect only connects Twitter and Facebook. If SocConnect could integrate more social networking sites, its social data integrating functions, such as blending and group friends, would be more powerful. The question is to choose which social networking sites. There are some candidates: LinkedIn, Google Buzz, foursquare and Last.fm due their popularities. However, LinkedIn and last.fm have their own domains: LinkedId focuses on career, last.fm focuses on music. These domains would brings more vocabularies into the SocConnect schema.

6.3.2 Choosing Functions Based on Users' Goals

Beside organizing and recommending social data, there are many other possible functions that can be provided by SocConnect. The design of new functions should

be based on users' goals on SNSs. From previous works[14], there are eight major goals that users pursue on: looking for new relations, maintaining social ties, finding information, debating, time-killing, profile surfing, sharing content, and maintaining contact with family. If an accurate model of users' goals could be designed, SocConnect adapt the functions it offers to different users depending on their goals.

6.3.3 Implicit Interests Indicator in Learning User Preference

The current recommendation approach is entirely depends on user's rating, the most common explicit interests indicator. Explicit interest indicators are easily to implement, but they disturb the user's normal browsing pattern. Users usually do not rate often. Without sufficient number of ratings, the recommendation algorithm cannot generate high quality results. The field study results clearly proves this point. Implicit interest indicator could enhance the recommendation algorithm. The possible implicit interest indicators are whether the user reads the update, whether she replies the update, and the browsing time on a update.

6.3.4 Alternative Recommendation Algorithm

The current recommendation algorithm uses data mining techniques. Data mining offers a robust way to find potential relation between activities' features and users' rating on these activities. Data mining requires a bootstrap time, so the data mining algorithm cannot generate any recommendations for a new user. However data mining techniques are computationally expensive. Every time the user uses SocConnect, the number of the activities from his or her friends will increase. It means the dataset for that user is getting larger and larger, the required memory for recommendation is getting larger and larger, and the time for generating recommendation is getting longer and longer. Many social networking sites companies and other organizations use different technologies for data analysis. Twitter uses Pig[3]. [4], Facebook uses Hive [5] [6].

---

[3]http://pig.apache.org/
[4]http://www.slideshare.net/kevinweil/hadoop-pig-and-twitter-nosql-east-2009
[5]http://www.slideshare.net/zshao/hive-data-warehousing-analytics-on-hadoop-presentation
[6]http://wiki.apache.org/hadoop/Hive

6.3.5 A Large Scale User Study

Due to the limited number of participants in the field study, the effectiveness of the recommendation could not be proved. With a web implementation of SocConnect, it will become possible to conduct a user study with larger number of participants and to investigate better the real-life performance of recommender function of Soc-Connect.

References

[1] Foaf. `www.foaf-project.org`.

[2] Gumo. `www.gumo.org`.

[3] Mrss. `http://search.yahoo.com/mrss`.

[4] Oauth homepage. `http://oauth.net`, 11.

[5] Open id homepage. `http://openid.net`, 11.

[6] Play framework. `www.playframework.org`.

[7] Social network aggregators. `http://mashable.com/2007/07/17/social-network-aggregators`.

[8] Weka. `http://www.cs.waikato.ac.nz/ml/weka/`.

[9] Xfn. `www.gmpg.org/xfn/`.

[10] Walled garden. `http://en.wikipedia.org/wiki/Walled_garden_(technology)`, 11 2010.

[11] Rajiv Badi, Soonil Bae, J. Michael Moore, Konstantinos Meintanis, Anna Zacchi, Haowei Hsieh, Frank Shipman, and Catherine C. Marshall. Recognizing user interest and document value from reading and organizing activities in document triage. In *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, pages 218–225, New York, NY, USA, 2006. ACM.

[12] Shlomo Berkovsky, Tsvi Kuflik, and Francesco Ricci. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18(3):245–286, 2007.

[13] Uldis Bojars, Alexandre Passant, John Breslin, and Stefan Decker. Social network and data portability using semantic web technologies. In *Proceddings of the Workshop on Social Aspects of the Web*, 2008.

[14] Petter Brandtzæg and Jan Heim. Why People Use Social Networking Sites. In Ant A. Ozok and Panayiotis Zaphiris, editors, *Online Communities and Social Computing*, volume 5621, chapter 16, pages 143–152. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[15] Francesca Carmagnola, Fabiana Vernero, and Pierluigi Grillo. Sonars: A social networks-based algorithm for social recommender systems. In *Proceddings of the 17th International Conference on User Modeling, Adaptation, and Personalization*, 2009.

[16] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 1185–1194, New York, NY, USA, 2010. ACM.

[17] Michael Chisari. The future of social networking. In *Proceedings of the W3C Workshop on the Future of Social Networking*, 2009.

[18] Mark Claypool, Mark Claypool, Phong Le, Phong Le, Makoto Waseda, Makoto Waseda, David Brown, and David Brown. Implicit interest indicators. In *In Intelligent User Interfaces*, pages 33–40. ACM Press, 2000.

[19] Adrienne Felt and David Evans. Privacy protection for social networking platforms. In *Proceddings of the Workshop on Web 2.0 Security and Privacy*, 2008.

[20] Dieter Fensel. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce.* Springer-Verlag, 2001.

[21] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[22] Dominik Heckmann, Eric Schwarzkopf, Junichiro Mori, Dietmar Dengler, and Alexander Kroner. The user model and context ontology gumo revisited for future web 2.0 extensions. In *Proceedings of the International Workshop on Contexts and Ontologies: Representation and Reasoning*, 2007.

[23] Cliff Lampe, Nicole Ellison, and Charles Steinfield. A face(book) in the crowd: Social searching vs. social browsing. In *Proceedings of the ACM Special Interest Group on Computer-Supported Cooperative Work*, 2006.

[24] F. Wilfrid Lancaster and Emily Gallup Fayen. *Information retrieval: on-line [by] F. W. Lancaster and E. G. Fayen.* Melville Pub. Co. Los Angeles,, 1973.

[25] Ching man Au Yeung, Ilaria Liccardi, Kanghao Lu, Oshani Seneviratne, and Tim Berners-Lee. Decentralization: The future of online social networking. In *Proceedings of the W3C Workshop on the Future of Social Networking*, 2009.

[26] Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl. Interfaces for eliciting new user preferences in recommender systems. In *The 9th International Conference on User Modeling (UM'2003)*, pages 178–188, 2003.

[27] Tom M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[28] John C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Learning.* MIT Press, 1999.

[29] David Recordon. "blowing up" social networks by going open. `http://www.slideshare.net/daveman692/blowing-up-social-networks-by-going\-open-presentation`, 2008.

[30] Paul Resnick, Neophytos Lacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceddings of the ACM conference on Computer supported cooperative work*, 1994.

[31] Gerard Salton, Andrew K C Wong, and C. S. Yang. *A vector space model for automatic indexing*, pages 273–280. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[32] Clara Shih. *The Facebook Era: Tapping Online Social Networks to Build Better Products, Reach New Audiences, and Sell More Stuff.* Prentice Hall PTR, 2009.

[33] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 2 edition, 1979.

[34] Andrew Webster and Julita Vassileva. Personal relations in online communities. In *Adaptive Hypermeida and Adaptive Web-Based System*, pages 223–233, Dublin, Ireland, June 2006. Springer LNCS.

Appendix A

Appendix: SocConnect Online Consent Form

1. Consect Form 1. 1. You are invited to participate in a study entitled "SocConnect: Personalized Social data aggregator". Please read this form carefully, and feel free to ask the researchers any questions you might have.

Researchers: Yuan Wang and Julita Vassileva, Department of Computer Science (1-306-966-2073), yuw193@mail.usask.ca, jiv@cs.usask.ca

The purpose of the study is to evaluate the understandability and motivational effect of an application, researchers have designed,called SocConnect which supports users in managing their social data. There are no known risks in this study.

As a token of appreciation for your time to participate in this study, you will be given a gift certificate of $ 10 from Amazon.ca (we will reward only the first 25 responders).

For this, you have to answer an online questionnaire with approx. 30 questions about your experience with the application in the end. It should take no more than 20 minutes of your time.

The research data will be anonymized immediately after this study. It will be available only to the researchers. Pseudonyms (alias) will be used to refer to the participants. Any information that could be potentially linked to a specific participant will be removed or altered. The data will be kept by the researchers for a minimum of five years upon the completion of this study in a secure storage. Aggregate results will be used in a research project through NSERC and articles published in peer reviewed scientific conferences and journals.

Your participation is voluntary, and you may withdraw from the study for any reason, at any time, without penalty of any sort. You may refuse to answer individual questions.

If you have any questions concerning the study, please feel free to contact the researchers at any point during or after the experiment.

This study has been approved on ethical grounds by the University of Saskatchewan Behavioural Research Ethics Board with certificate 08-143 on (November, 2009). Any questions regarding your rights as a participant may be addressed to that committee through the Ethics Office (966-2084). You may find out about the results of the study through the MADMUC website (http://madmuc.usask.ca) or by contacting the researchers.

I have read and understood the description provided above; I have been provided with an opportunity to ask questions and my questions have been answered satisfactorily. I give the consent to participate in the study described above, understanding that I may withdraw this consent at any time. A copy of this consent form has been given to me for my records.

You can only proceed if you select "Agree".

- Agree

- Disagree

Figure A.1: Consent Form Webpage

Appendix B

Appendix: SocConnect Pilot Study Interview Questions

You are invited to participate in a study entitled *(here the title of the actual study will be listed, e.g. "Usability evaluation of SoCConnect", or "Social data tagging study")*. Please read this form carefully, and feel free to ask the researchers any questions you might have.

Researchers: **Julita Vassileva**, Department of Computer Science (966-2073), jiv@cs.usask.ca
Yuan Wang, Department of Computer Science**, Jie Zhang**, Department of Computer Science

The purpose of the study is to *(the following paragraph describes in brief the purpose and method of the particular study, in this case Social data tagging study)* evaluate the accurateness of tag recommendation algorithms for social data. In the study, you will be tagging your Twitter and Facebook activity with text annotation on a paper form. The estimate of the total time to participate in this study is 30 to 45 minutes.

There are no known risks in this study.

Findings from the study will be used to evaluate the algorithms for social data recommendation. You will be asked to tag 100 to 200 twitter or/and facebook updates from your friends on these Social networking sites.

The research data will be stored on a password-protected computer system and will be available only to the researchers. Personally identifying information will be destroyed upon completion of data collection, and pseudonyms (alias) will be used to refer to the participants. The data will be kept by the researchers for a minimum of five years upon the completion of this study in a secure storage. The signed consent form will be stored separately from the data.

Aggregate results will be used in a M.Sc. thesis and articles published in peer reviewed conferences and scientific journals. However, any information that can be linked to a specific participant will be removed or altered.

Your participation is voluntary, and you may withdraw from the study for any reason, at any time, without penalty of any sort. You may refuse to answer individual questions. If you withdraw from the study at any time, any data that you have contributed will be destroyed at your request.

If you have any questions concerning the study, please feel free to ask at any point; you are also free to contact the researchers if you have questions at a later time. This study has been approved on ethical grounds by the University of Saskatchewan Behavioural Research Ethics Board on (insert date). Any questions regarding your rights as a participant may be addressed to that committee through the Ethics Office (966-2084). Out of town participants may call collect. You may find out about the results of the study through the MADMUC website (http://madmuc.usask.ca) or by contacting the researchers. The result will be available, but any information that can be linked to a specific participant will not be contained in the result.

I have read and understood the description provided above; I have been provided with an opportunity to ask questions and my questions have been answered satisfactorily. I consent to participate in the study described above, understanding that I may withdraw this consent at any time. A copy of this consent form has been given to me for my records.

_____   _____
(Name of Participant)   (Date)

_____   _____
(Signature of Participant)   (Signature of Researcher)

There are three steps for evaluating SocConnect:

Step 1: Testify the necessity of having the three functions, including blending friends, grouping friends and filtering social data by tagging: This step is done thought an interview phase. Subjects will be asked a set of questions according to their own experience or preferences. The answers for the questions from those subjects will summarized.

Background:

a. How many social networking sites you frequently (more once every two weeks) using, and what are they?
b. Are you using Facebook? If so how many hours do you usually spend on it every week?
1) less than one hour 2) one or three hours; 3) three to seven hours; 4) more than seven

c. How many friends do you have on Facebook?
  1). < 5 2).5-10 3).10-30 4) 30-70; 5) > 70

d. Are you using twitter? If so, how many hours do you usually spend on it every weeks
1) less than one hour 2) one or three hours; 3) three to seven hours; 4) more than seven

e. Are you using any desktop Facebook application? If so, what the names?

f. How many people do you follow on Twitter?
 1). < 5 2).5-10 3).10-30 4) 30-70; 5) > 70

g. Are you using any desktop Twitter application? If so, what the names?

1) Questions about blending friends:
a. How many friends do you have in total (Facebook and Twitter)? (we may adapt to different subjects' social networking sites when asking questions)
b. Do you have some friends who have accounts on both Facebook and Twitter?
c. How many? 1-3 friends, 4-6, or more than 7
d. Do you keep friends for different purposes on different sites?
e. Have those friends been active on both Facebook and Twitter?
f. Do they have similar or identical activities on Facebook and Twitter?
g. If identical, do you want to see their activities at the same place or view them separately on each social networking site?

2) Question about grouping friends:
a. Facebook and Twitter allow you to put some friends into a list (group friends). Have you ever used this function?
b. If used, why putting them together? Similar interests, preferences, doing activities together, or than?
c. Do you also want to include in the groups some friends on other sites?

d. Do you want to view their activities together or go on each social networking sites to view their activities separately?

3). Question about filtering social data
The questions 2d give some sense about the function of filtering social data based on groups. We may also ask some questions about filtering social data using tags
a. Have you ever had difficulty on browsing through a large number of friends' updates? (This question is related to Question1a because the number of friends' updates is dependent on the number of friends a user has)
b. Have you ever been overwhelmed by a large amount of friends' updates?
c. Have you ever come back to find some particular updates from history of all updates?
d. Do you want organize yours or your friends' updates into categories?

Step 2: Evaluate the usability of the interface of SocConnect. In this step, we may ask subjects to perform some tasks, to observe whether the interface can be easily used. For example, we can give them a brief summary of what the software can do at the beginning. Then, we ask subjects to blend friends together, and use other functions. We see how many subjects can successfully finish those tasks and how much time they take.
1.   login to the application
2.   blend friends
3.   create a group
4.   tag friend
5.   tag activities
6.   search a tag

Step 3. Feedback for improving the software

We may finally ask subjects to provide some feedback for improving the software

a. Which parts of the current interface design needs to be improved?
b. Do you think the interface design for the functionalities is intuitive?
c. Which other functions should be added into the software?
d. We will provide recommendations about tags. Which information might be useful for tag recommendations?
e. Do you think we should recommend tag for tagging people or activities, or both.

Appendix C

Appendix: SocConnect Field Study Survey

1. Consent Form

Consent Form

* **1. 1. You are invited to participate in a study entitled "SocConnect: Personalized Social data aggregator".**

  Please read this form carefully, and
  feel free to ask the researchers any questions you might have.

  Researchers:
  Yuan Wang and Julita Vassileva,
  Department of Computer Science (1-306-966-2073),
  yuw193@mail.usask.ca, jiv@cs.usask.ca

  The purpose of the study is to evaluate the understandability and motivational effect of an application, researchers have designed,called "SocConnect" which supports users in managing their social data. There are no known risks in this study.

  As a token of appreciation for your time to participate in this study, you will be given a gift certificate of $ 10 from Amazon.ca (we will reward only the first 25 responders).

  For this, you have to answer an online questionnaire with approx. 30 questions about your experience with the application in the end. It should take no more than 20 minutes of your time.

  The research data will be anonymized immediately after this study. It will be available only to the researchers.
  Pseudonyms (alias) will be used to refer to the participants. Any information that could be potentially linked to a specific participant will be removed or altered. The data will be kept by the researchers for a minimum of five years upon the completion of this study in a secure storage. Aggregate results will be used in a research project through NSERC and articles published in peer reviewed scientific conferences and journals.

  Your participation is voluntary, and you may withdraw from the study for any reason, at any time, without penalty of any sort. You may refuse to answer individual questions.

SocConnect Survey

If you have any questions concerning the study, please feel free to contact the researchers at any point during or after the experiment.

This study has been approved on ethical grounds by the University of Saskatchewan Behavioural Research Ethics Board with certificate 08-143 on (November, 2009). Any questions regarding your rights as a participant may be addressed to that committee through the Ethics Office (966-2084). You may find out about the results of the study through the MADMUC website (http://madmuc.usask.ca) or by contacting the researchers.

I have read and understood the description provided above; I have been provided with an opportunity to ask questions and my questions have been answered satisfactorily. I give the consent to participate in the study described above, understanding that I may withdraw this consent at any time. A copy of this consent form has been given to me for my records.

You can only proceed if you select "Agree".

- ○ Agree
- ○ Disagree

SocConnect Survey

2. Section 1 Background Information

* **1. Your Email address (for contact and Amazon Gift Card (We will not spam you or reveal your email address ever to any third-party.)**

    [   ]

* **2. Your Facebook User Name: (We will not spam or contact your facebook account or reveal your any facebook information to any third-party).**

    The purpose for asking for your Facebook account information is co-relate your questionnaire answers with the data log This will allow us to evaluate your satisfaction of the quality of recommendations generated for you.

    What is my facebook user name?
    When you login facebook, go to your profile page (on upper left side), the url of your profile page ends with your facebook username
    either http://www.facebook.com/userName
    or http://www.facebook.com/profile.php?id=userId

    [   ]

* **3. Your Twitter User Name: (We will not spam or contact your twitter account or reveal any your twitter information to any third-party).**

    The purpose for asking for your Twitter account information is co-relate your questionnaire answers with the data log This will allow us to evaluate your satisfaction of the quality of recommendations generated for you.

    What is my twitter username?
    After you log into twitter, your username would appear on the upper right side.

    [   ]

* **4. How often are you using Facebook, on average?**

    ◌ less than one hour per week

    ◌ less than one hour per day

    ◌ two hours per day

    ◌ more than two hours per day

SocConnect Survey

SocConnect Survey

* 5. Have you used any clients to access Facebook beside the original website (www.facebook.com), for example, mobile facebook clients for iPhone, blackBerry, Android, or any desktop or website application?

- ◯ Yes
- ◯ No

* 6. If Yes, what are the names of applications you using

[   ]

* 7. How often are you using Twitter on average?

- ◯ less one hour per week,
- ◯ less than 1 hour per day,
- ◯ two hours per day,
- ◯ more than two hours

* 8. Have you use any clients to access Twitter beside the original website (www.twitter.com), such as mobile apps, tweetCaster, HootSuite, TweetDeck?

- ◯ Yes
- ◯ No

* 9. Have you used any applications that integrate Twitter and Facebook or other social networking site data, together like SocConnect? For example, Seesmic, HootSuite?

- ◯ Yes
- ◯ No

10. If yes, what are their names

[   ]

* 11. Do you prefer to browse your Twitter and Facebook togther?

Yes, Why   [   ]

No, Why   [   ]

SocConnect Survey

3. Section 2. Functionality Feedback

SocConnect provides several main functions listed below. Please, indicate how much you agree with the following statements about the Blending friends function:

* 1. You are aware of this function (Blending friends).

- jn Agree

- jn Disagree

* 2. You like this function (Blending friends).

- jn Strongly disagree -2

- jn Disagree -1

- jn Neutral 0

- jn Agree 1

- jn Strongly agree +2

* 3. This function (Blending friends) is necessary.

- jn Strongly disagree -2

- jn Disagree -1

- jn Neutral 0

- jn Agree 1

- jn Strongly agree +2

* 4. This function (Blending friends) is easy to use.

- jn Strongly disagree -2

- jn Disagree -1

- jn Neutral 0

- jn Agree 1

- jn Strongly agree +2

5. Do you have any comments on this function (Blending friends)?

SocConnect Survey

4. Grouping friends

SocConnect provides several main functions listed below. Please, indicate how much you agree with the following statements about the Grouping friends function:

* 1. You are aware of this function (Grouping friends).

- ꭒ Agree
- ꭒ Disagree

* 2. You like this function (Grouping friends).

- ꭒ Strongly disagree -2
- ꭒ Disagree -1
- ꭒ Neutral 0
- ꭒ Agree 1
- ꭒ Strongly agree +2

* 3. This function (Grouping friends) is necessary.

- ꭒ Strongly disagree -2
- ꭒ Disagree -1
- ꭒ Neutral 0
- ꭒ Agree 1
- ꭒ Strongly agree +2

* 4. This function (Grouping friends) is easy to use.

- ꭒ Strongly disagree -2
- ꭒ Disagree -1
- ꭒ Neutral 0
- ꭒ Agree 1
- ꭒ Strongly agree +2

5. Do you have any comments on this function (Grouping friends)?

SocConnect Survey

5. Tagging friends

SocConnect provides several main functions listed below. Please, indicate how much you agree with the following statements about the Tagging friends function:

* 1. You are aware of this function (Tagging friends).

- ○ Argee
- ○ Disagree

* 2. You like this function (Tagging friends).

- ○ Strongly disagree -2
- ○ Disagree -1
- ○ Neutral 0
- ○ Agree 1
- ○ Strongly agree +2

* 3. This function (Tagging friends) is necessary.

- ○ Strongly disagree -2
- ○ Disagree -1
- ○ Neutral 0
- ○ Agree 1
- ○ Strongly agree +2

* 4. This function (Tagging friends) is easy to use.

- ○ Strongly disagree -2
- ○ Disagree -1
- ○ Neutral 0
- ○ Agree 1
- ○ Strongly agree +2

5. Do you have any comments on this function (Tagging friends)?

[   ]

SocConnect Survey

6. Search/Filter by tag

SocConnect provides several main functions listed below. Please, indicate how much you agree with the following statements about the "Search/Filter by tag" function:

* **1. You are aware of this function (Search/Filter by tag).**

  ◌ Agree

  ◌ Disagree

* **2. You like this function (Search/Filter by tag).**

  ◌ Strongly disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strongly agree +2

* **3. This function (Search/Filter by tag) is necessary.**

  ◌ Strongly disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strongly agree +2

* **4. This function (Search/Filter by tag) is easy to use.**

  ◌ Strongly disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strongly agree +2

5. Do you have any comments on this function (Search/Filter by tag)?

SocConnect Survey

7. Recommended updates

SocConnect provides several main functions listed below. Please, indicate how much you agree with the following statements about the Recommended updates function:

* **1. You are aware of this function (Recommended updates).**

  ◌ Agree

  ◌ Disagree

* **2. You like this function (Recommended updates).**

  ◌ Strongly disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strongly agree +2

* **3. This function (Recommended updates) is necessary.**

  ◌ Strongly disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strongly agree +2

* **4. This function (Recommended updates) is easy to use.**

  ◌ Strongly disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strongly agree +2

5. Do you have any comments on this function (Recommended updates)?

* 6. The recommendation generated by SocConnect reflected well your ratings on previous updates.

jn Strongly disagree -2

jn Disagree -1

jn Neutral 0

jn Agree 1

jn Strongly agree +2

* 7. The way recommended or non-recommended updates are displayed with different colours is intuitive.

jn Strongly disagree -2

jn Disagree -1

jn Neutral 0

jn Agree 1

jn Strongly agree +2

SocConnect Survey

8. Rating updates

SocConnect provides several main functions listed below. Please, indicate how much you agree with the following statements about the Rating update function:

* **1. You are aware of this function (Rating update).**

  ◌ Agree

  ◌ Disagree

* **2. You like this function (Rating update).**

  ◌ Strong Disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strong Agree 2

* **3. This function (Rating update) is necessary.**

  ◌ Strongly disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strongly agree +2

* **4. This function (Rating update) is easy to use.**

  ◌ Strongly disagree -2

  ◌ Disagree -1

  ◌ Neutral 0

  ◌ Agree 1

  ◌ Strongly agree +2

5. Do you have any comments on this function (Rating update)?

* **6. You are willing to rate the updates to improve future recommendations.**

   ʝ∩ Strongly disagree -2

   ʝ∩ Disagree -1

   ʝ∩ Neutral 0

   ʝ∩ Agree 1

   ʝ∩ Strongly agree +2

* **7. It is easy to decide how to rate an update**

   ʝ∩ Strongly disagree -2

   ʝ∩ Disagree -1

   ʝ∩ Neutral 0

   ʝ∩ Agree 1

   ʝ∩ Strongly agree +2

* **8. You tend to rate using one criterion (for example, the friend responsible for the update. what kind of update, is it from Twitter or Facebook, the actual content of the update)**

   ʝ∩ Strongly disagree -2

   ʝ∩ Disagree -1

   ʝ∩ Neutral 0

   ʝ∩ Agree 1

   ʝ∩ Strongly agree +2

9. If you agree, please, specify which criterion you tend to use:

Appendix D

Appendix:SocConnect Field Study Raw Results

Blend Actions	0	1	2	3	50	
	0	1	2	3	50	60
The Number of participants	4	2	1	2	1	10
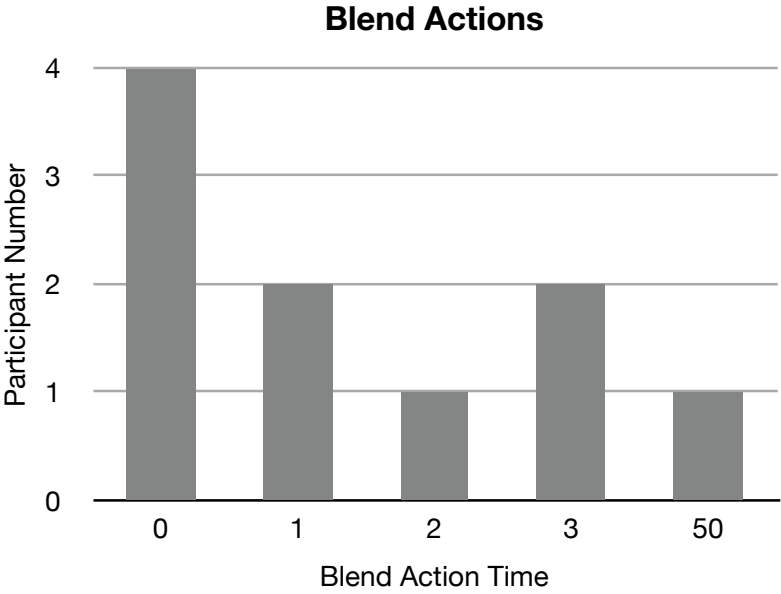
Blend Actions

UserID	Blend	Group	Tag	Rate	
3	0	0	5	10	
5	0	2	0	44	
8	50	1	0	1	
10	1	1	4	19	
15	0	6	26	1	
17	1	2	7	2	
27	3	1	1	11	
28	0	0	0	0	
30	2	0	0	0	
7	3	0	0	1	
B	60	13	43	89	

Overview Function Actions
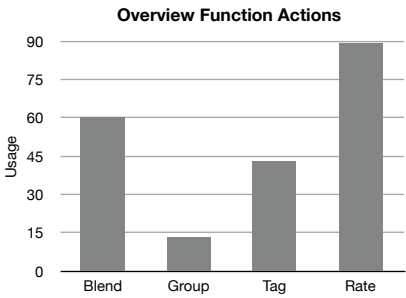
Group Actions	0	1	2	6	
User Number	1	1	1	1	
	1	1	1		
	1	1			
	1				
	1				
Sum	5	3	2	1	11

Group Actions
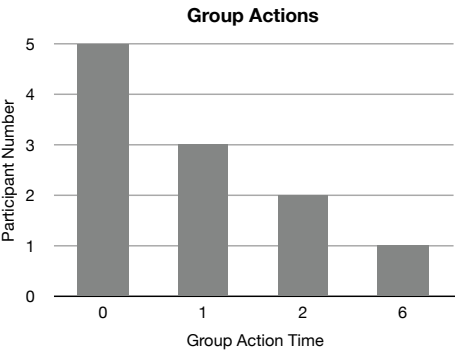
Tag action	0	1	4	5	7	26	
User Number	1	1	1	1	1	1	
	1						
	1						
	1						
	1						
	1						
sum	6	1	1	1	1	1	11
							43

Tag Actions
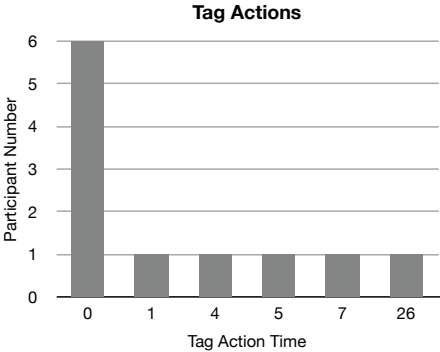
Rate Action	0	1	2	10	11	19	44	
User Number	1	1	1	1	1	1	1	
	1	1						
	1	1						
Sum	3	3	1	1	1	1	1	11
								89

Rate Actions
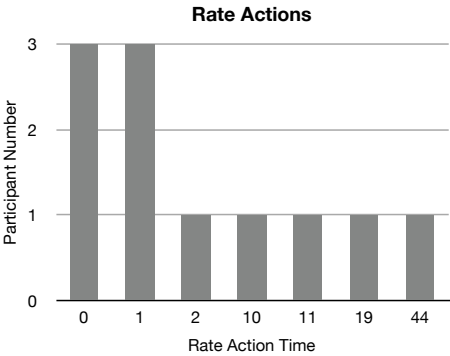
Respondent ID	User ID	email-address	useage facebook	3rd app for fb	prefer integation	facebook friends no	facebook act no
8	6 (excluded)	deleted	less than one hour per week	no	no, dont have a twitter account	118	1
21	28	deleted	less than one hour per week	no	no, dont have a twitter account	74	25

Respondent ID	aware blend	like blend	necessary	easy	comment
8	agree	agree	neutral	agree	
21	agree	agree	agree	agree	

Respondent ID	aware group	like group	necessary	easy	comment
8	agree	agree	agree	agree	
21	agree	neutral	neutral	agree	I think the group

Respondent ID	aware tag	like tag	necessary	easy	comment
8	agree	neutral	neutral	agree	
21	agree	agree	agree	agree	

Respondent ID	aware search	like search	necessary	easy	comment
8	agree	neutral	neutral	agree	
21	agree	agree	agree	agree	

Respondent ID	aware recom	like recom	necessary	easy	comment	recommendation generated by SocConnect reflected well your ratings on previous updates.	color intutive
8	agree	agree	Strong agree	agree		neutral	strong agree
21	disagree	agree	agree	neutral		neutral	agree

Respondent ID	aware rating	like rating	necessary	easy	comment	willing to rate	easy to decide rate	rate using one criterion	criterion tend
8	agree	agree	agree	agree		argee	neutral	neutral	
21	agree	agree	agree	agree		argee	agree	agree	content of upda

Respondent ID	enjoy SocConnect	use in future	add new SNS	which SNS to add	suggest
8	agree	agree	disagree		
21	agree	agree	neutral		I hope the gui c

Respondent ID	User ID	email-address	useage facebook	3rd app for fb	name	usage twitter	3rd app for twitter	name	integration app	name	prefer integation	facebook friends no	facebook act no	twitter friends no	twitter act no
9	10	deleted	less than one hour per day	Yes	Nimbuzz	less one hour per week	no	N/A	no	n/a	Yes, Why - it does gives you the i	216	508	26	102
12	22	deleted	less than one hour per day	No	N/A	less one hour per week	no	N/A	no	n/a	Yes, Why - I never thought of it be	984	0	22	0
15	17	deleted	two hours per day	no	no	less one hour per week	no	N/A	no	n/a	Yes, Why - yes, I like the wa	121	127	14	108
16	11	deleted	less than one hour per day	no	no	less one hour per day	no	N/A	no	n/a	Yes, Why - yes, convenient	16	94	3	107
20	27	deleted	less than one hour per day	no	none	less one hour per day	no	N/A	no	n/a	Yes, Why - save time for me	60	51	16	173
22	5	deleted	less than one hour per day	no	none	less one hour per week	no	N/A	no	n/a	No, Why - I use twitter maini	329	104	4	394

Respondent ID	aware blend	like blend	necessary	easy	comment
9	agree	Strong agree	aggree	agree	
12	agree	agree	agree	strong agee	suggestive ben
15	agree	strong agree	strong agree	strong agee	It is cool, but I w
16	agree	Strong agree	strong agree	strong agee	it helps blendin
20	agree	agree	agree	neutral	
22	disagree	neutral	neutral	strong disagee	I didn't find any

Respondent ID	aware group	like group	necessary	easy	comment
9	agree	strong agree	agree	agree	
12	agree	agree	agree	neutral	
15	agree	strong agree	strong agree	agree	i can group the
16	agree	agree	agree	agree	
20	agree	neutral	neutral	agree	I think the gr
22	agree	Neutral	strong agree +2	strong agree	Never found

Respondent ID	aware tag	like tag	necessary	easy	comment
9	agree	agree	agree	agree	
12	agree	agree	strong agree	strong agree	
15	agree	strong agree	strong agree	strong agree	none. i like the t
16	agree	strong agree	strong agree	strong agree	
20	agree	neutral	neutral	agree	
22	agree	disagree	disagree	disagree	

Respondent ID	aware search	like search	necessary	easy	comment
9	agree	strong agree	strong agree	agree	
12	agree	agree	agree	neutral	
15	agree	strong agree	strong agree	agree	
16	agree	strong agree	strong agree	strong agree	
20	agree	agree	agree	agree	I think the keyw
22	agree	neutral	neutral	agree	

Respondent ID	aware recom	like recom	necessary	easy	comment	recommendation generated by SocConnect reflected well your ratings on previous updates.	color intuitive
9	agree	agree	Strong agree	agree		agree	agree
12	agree	neutral	agree	agree		neutral	neutral
15	agree	strong agree	strong agree	strong agre	its' good	strong agree	neutral
16	agree	agree	agree	agree		agree	agree
20	disagree	neutral	neutral	neutral	haven't got a	neutral	agree
22	disagree	agree	Strong agree	neutral	I believe that	neutral	neutral

Respondent ID	aware rating	like rating	necessary	easy	comment	willing to rate	easy to decide rate	rate using one criterion	criterion tend		
9	agree	agree	agree	agree		neutral	netrual	agree	the friend responsible		
12	neutral	neutral	neutral	agree		neutral	agree	strong agree			
15	agree	agree	strong agree	agree	is overlapping	agree	netrual	strong agree	which update it is from, who it is from		
16	agree	agree	agree	agree		agree	strong agree	agree	what kind of update		
20	agree	agree	agree	agree	I need to click	agree	agree	agree	the actual content of the update		
22	disagree	neutral	neutral	neutral		neutral	netrual	neutral			

Respondent ID	enjoy SocConnect	use in future	add new SNS	which SNS to add	new func	suggest
8	agree	neutral	agree			it would be good to see the video, links i
12	strong agree	agree	agree	LinkedIn,Hi5		GUI needs litle brush up, interface should be little user friendly
15	strong agree	strong agree	agree	linkedIn, delicio	plan an event with pictures or event orga	overall it was good
16	agree	agree	agree	renren	can reply direcelty	
20	agree	agree	neutral			
22	strong disagree	disagree	neutral			Not functions, just improvements, I prefer the way facebook shows the posts, it would be very hard to t

Respondent ID	User ID	email-address	useage facebook	3rd app for fb	name	usage twitter	3rd app for twitter	name	integration app	name	prefer integation	facebook friends no	facebook act no	twitter friends no	twitter act no
23	3	deleted	more than two hours per day	Yes	Digsby	more than two hours per day	yes	digsby	digsby	yes	No, Why - I use Twitter for Busine	314	60	87	278
24	8	deleted	less than one hour per day	yes	iphone	less one hour per day	yes	iphone	no	n/a	Yes, Why - when I want a quick o / No, Why - when I want more deta	252	487	198	1629
4	7	deleted	less than one hour per week	yes	iphone	less one hour per hour	yes	iphone	yes	Twitter app for facebook	Yes, Why - it is much easier to stay up to date with that's happening				

Respondent ID	aware blend	like blend	necessary	easy	comment
23	agree	neutral	neutral	netrual	I guess I just like suggestive bene
24	agree	neutral	agree	agree	
4	agree	strongly agree	agree	agree	I would love to s

Respondent ID	aware group	like group	necessary	easy	comment
23	agree	neutral	neutral	neutral	I guess I just
24	agree	agree	agree	agree	
4	agree	agree	strongly agree	agree	it would have b

Respondent ID	aware tag	like tag	necessary	easy	comment
23	agree	neutral	neutral	agree	
24	agree	neutral	disagree	agree	found it redund
4	agree	agree	neutral	disagree	the idea is great

Respondent ID	aware search	like search	necessary	easy	comment
23	disagree	strong agree	strong agree	neutral	Woops, missed
24	agree	agree	agree	agree	
4	agree	agree	agree	agree	

Respondent ID	aware recom	like recom	necessary	easy	comment	recommendation generated by SocConnect reflected well your ratings on previous updates.	color intuitive
23	agree	agree	agree	agree		agree	disagree
24	agree	neutral	neutral	disagree	System never	neutral	neutral
4	agree	agree	agree	agree		agree	agree

Respondent ID	aware rating	like rating	necessary	easy	comment	willing to rate	easy to decide rate	rate using one criterion	criterion tend		
23	agree	agree	netutral	agree		neutral	disagree	agree	friend and content. Usually friend is the most rel the friend responsible		
24	agree	agree	neutral	agree	I must have not	neutral	neutral	neutral			
4	agree	agree	agree	agree		agree	agree	agree			

Respondent ID	enjoy SocConnect	use in future	add new SNS	which SNS to add	new func	suggest
23	neutral	disagree	neutral		Maybe by just clicking an update it beco	Also maybe even separate rankings instead of coloring? GUI needs little brush up, interface should be little user friendly
12	agree	agree	neutral		it would be good to see the video, links i post update	It would have been nice to see a count down of how
4	agree	agree	agree		It might sound a bit strange, but I would	it would be nice to a have a tag cloud for users and

	Aware	Like	Necessary	Easy
strongly agree		4	2	3
agree	10	5	6	5
neutral		3	3	2
disagree	1	0	0	0
strongly disagree	0	0	0	1
	11	11	11	11

	Aware	Like	Necessary	Easy
strongly agree		2	3	1
agree	11	5	5	8
neutral		4	3	2
disagree	0	0		
strongly disagree		0		
	11	11	11	11

Blend Feedback

strongly agree   agree   neutral   disagree
strongly disagree

Group Feedback

strongly agree   agree   neutral   disagree
strongly disagree
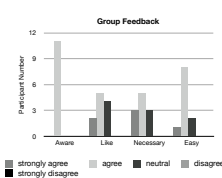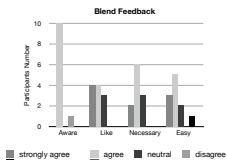
	Aware	Like	Necessary	Easy
strongly agree		2	3	3
agree	11	4	2	5
neutral		4	4	1
disagree	0	1	2	2
strongly disagree	0	0		
	11	11	11	11

	Aware	Like	Necessary	Easy
strongly agree		4	2	2
agree	10	5	6	7
neutral		2	3	2
disagree	1	0	0	0
strongly disagree		0	0	0
	11	11	11	11

Tag Feedback

strongly agree   agree   neutral   disagree
strongly disagree

Search Feedback

strongly agree   agree   neutral   disagree   strongly disagree
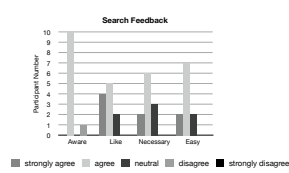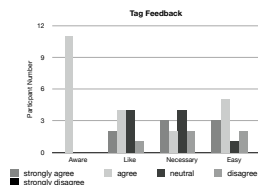
	Aware	Like	Necessary	Easy	Agree with Rating	Colour Intuitive
strongly agree		1	4	1	1	1
agree	8	7	5	6	3	4
neutral	0	3	2	3	7	5
disagree	3			1		1
strongly disagree						
	11	11	11	11	11	11
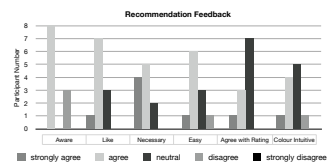
Recommendation Feedback

| Aware | Like | Necessary | Easy | Agree with Rating | Colour Intuitive |

strongly agree   agree   neutral   disagree   strongly disagree

	Aware	Like	Necessary	Easy	Willing to Rate	Easy to Decide Rate	Rate with Criterion
strongly agree				1		2	3
agree	10	9	6	10	6	3	6
neutral			4	1	5	5	2
disagree	1	2				1	
strongly disagree							
	11	11	11	11	11	11	11

Rate Feedback

| Aware | Like | Necessary | Easy | Willing to Rate | Easy to decide rate | Rate with Criterion |

strongly agree   agree   neutral   disagree   strongly disagree
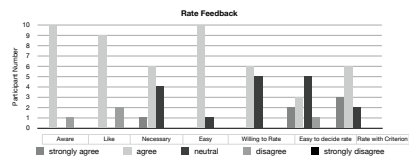
	Enjoy	Use in Future	add new SNS
strongly agree	3	1	
agree	7	7	5
neutral	1	1	5
disagree		2	1
strongly disagree			
	11	11	11

General Feedback

strongly agree   agree   neutral   disagree   strongly disagree
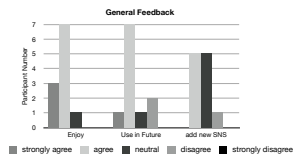
	Like from 1st study	Like from 3rd study	Necessary from 3rd study	
	0.75	0.7	0.7	

	Like from 1st study	Like from 3rd study	Necessary from 3rd study	
	0.94	0.6	0.8	

	Like from 1st study	Like from 3rd study	Necessary from 3rd study	
	0.69	0.5	0.5	

Blend function comparison

Group function comparison

Tag function comparison

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Facebook	1.5	10.4	1	2.5	8.5	8.8	0.8	0.5	0.5	12.8	3.2	1	1	0.5	1	1	1.5	2.8	3	0.8
Twitter	1.5	2.6	0	2.5	8.5	2.2	0.2	0.5	0.5	3.2	0.8	0	0	0.5	0	0	1.5	1.2	0	0.2
Connect times	3	13	1	5	17	11	1	1	1	16	4	1	1	1	1	1	3	4	3	1
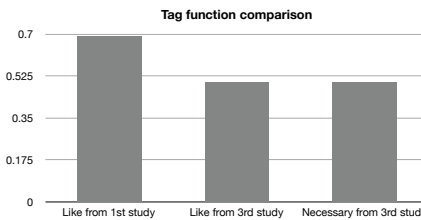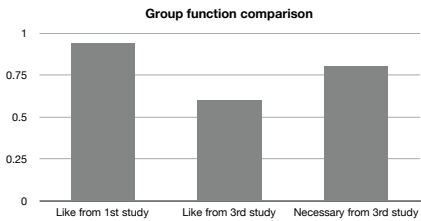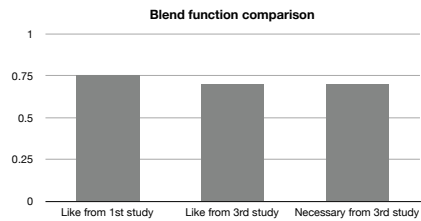
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Answer Survey	3	13	1	5	17	11	0	0	0	16	4	0	0	0	0	0	0	4	3	1	
Not	0	0	0	0	0	0	1	1	1	0	0	1	1	1	1	1	3	0	0	0	11

Paricipants Connect Times