December 2016

Three Essays on Friend Recommendation Systems for Online Social Networks

Jiaxi Luo
University of Wisconsin-Milwaukee

THREE ESSAYS ON FRIEND RECOMMENDATION SYSTEMS FOR ONLINE SOCIAL NETWORKS

by

Jiaxi Luo

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Management Science

at

The University of Wisconsin-Milwaukee

December 2016

ABSTRACT

THREE ESSAYS ON FRIEND RECOMMENDATION SYSTEMS FOR ONLINE SOCIAL NETWORKS

by

Jiaxi Luo

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Dr. Atish P. Sinha and Dr. Huimin Zhao

Social networking sites (SNSs) first appeared in the mid-90s. In recent years, however, Web 2.0 technologies have made modern SNSs increasingly popular and easier to use, and social networking has expanded explosively across the web. This brought a massive number of new users. Two of the most popular SNSs, Facebook and Twitter, have reached one billion users and exceeded half billion users, respectively.

Too many new users may cause the *cold start* problem. Users sign up on a SNS and discover they do not have any friends. Normally, SNSs solve this problem by recommending potential friends. The current major methods for friend recommendations are profile matching and "friends-of-friends." The profile matching method compares two users' profiles. This is relatively inflexible because it ignores the changing nature of users. It also requires complete profiles. The friends-of-friends method can only find people who are likely to be previously known to each other and neglects many users who share the same interests. To the best of my

knowledge, existing research has not proposed guidelines for building a better recommendation system based on context information (location information) and user-generated content (UGC).

This dissertation consists of three essays. The first essay focuses on location information and then develops a framework for using location to recommend friends--a framework that is not limited to making only known people recommendations but that also adds stranger recommendations. The second essay employs UGC by developing a text analytic framework that discovers users' interests and personalities and uses this information to recommend friends. The third essay discusses friend recommendations in a certain type of online community – health and fitness social networking sites, physical activities and health status become more important factors in this case.

Essay 1: Location-sensitive Friend Recommendations in Online Social Networks

GPS-embedded smart devices and wearable devices such as smart phones, tablets, smart watches, etc., have significantly increased in recent years. Because of them, users can record their location at anytime and anyplace. SNSs such as Foursquare, Facebook, and Twitter all have developed their own location-based services to collect users' location check-in data and provide location-sensitive services such as location-based promotions. None of these sites, however, have used location information to make friend recommendations.

In this essay, we investigate a new model to make friend recommendations. This model includes location check-in data as predictors and calculates users' check-in histories--users' life patterns--to make friend recommendations. The results of our experiment show that this novel model provides better performance in making friend recommendations.

Essay 2: Novel Friend Recommendations Based on User-generated Contents

More and more users have joined and contributed to SNSs. Users share stories of their daily life (such as having delicious food, enjoying shopping, traveling, hanging out, etc.) and leave comments. This huge amount of UGC could provide rich data for building an accurate, adaptable, effective, and extensible user model that reflects users' interests, their sentiments about different type of locations, and their personalities. From the computer-supported social matching process, these attributes could influence friend matches. Unfortunately, none of the previous studies in this area have focused on using these extracted meta-text features for friend recommendation systems.

In this study, we develop a text analytic framework and apply it to UGCs on SNSs. By extracting interests and personality features from UGCs, we can make text-based friend recommendations. The results of our experiment show that text features could further improve recommendation performance.

Essay 3: Friend Recommendations in Health/Fitness Social Networking Sites

Thanks to the growing number of wearable devices, online health/fitness communities are becoming more and more popular. This type of social networking sites offers individuals the opportunity to monitor their diet process and motivating them to change their lifestyles. Users can improve their physical activity level and health status by receiving information, advice and supports from their friends in the social networks. Many studies have confirmed that social

network structure and the degree of homophily in a network will affect how health behavior and innovations are spread. However, very few studies have focused on the opposite, the impact from users' daily activities for building friendships in a health/fitness social networking site.

In this study, we track and collect users' daily activities from Record, a famous online fitness social networking sites. By building an analytic framework, we test and evaluate how people's daily activities could help friend recommendations. The results of our experiment have shown that by using the helps from these information, friend recommendation systems become more accurate and more precise.

TABLE OF CONTENTS

LIST OF FIGURES

Location-sensitive Friend Recommendations in Online Social Networks

LIST OF TABLES

Location-sensitive Friend Recommendations in Online Social Networks

Friend Recommendations Based on User-generated Contents

Friend Recommendations in Health/Fitness Social Networking Sites

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my advisors, Professor Atish Sinha and Professor Huimin Zhao, who have instilled in me a spirit of adventure in regard to research and scholarship, and provided me with useful insights and feedback for my doctoral dissertation. Without their guidance and help, this dissertation would not have been possible.

I would like to sincerely thank my other committee members, Professor Steven France and Professor Xiaojing Yang, for their insightful comments and encouragement, and also for raising difficult questions, which forced me to consider different perspectives and widen my research focus.

I would like to thank Mrs. Gail Schemberger for her help in greatly improving the writing quality of the dissertation document.

Finally, I express my profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this dissertation. This accomplishment would not have been possible without them.

Author

Jiaxi (Jesse) Luo

Location-sensitive Friend Recommendations in Online Social Networks

1. Introduction

A social networking site (SNS) is an online service, platform, or site that is designed to facilitate the building of social networks or social relations among people who, for example, share interests, activities, backgrounds, or real-life connections. In recent years, SNSs have exploded in popularity. Facebook attracted more than one billion users who signed up in 2013. In the same year, Twitter exceeded 500 million users (Dudley-Nicholson 2013).

To help new users deal with the "*cold start*" problem - cold start is defined as giving recommendations to new users who have no preference on any items, or recommend items that no user of the community has seen yet (Lam et. al 2008), and to help old users further expand their social networks, SNSs have started to employ friend recommendation systems. Recommending people on SNSs is becoming one of the essential tasks of such sites. New users can find real-life friends already known to them or people who share similar interests to begin to build their social networks. Old users can expand their friendships and find new interests.

The major methods for friend recommendations, such as "friend-of-friend" and profile matching, have been proposed and used for some time. Much of the recent research has focused on these methods (Al Hasan et al. 2006; Benchettara et al. 2010; Chen et al. 2009b; Guy et al. 2009a; Jensen et al. 2002; Lichtenwalter et al. 2010; Quercia and Capra 2009). These studies have focused on methods that suggest people whom the user already

knows in real life (Guy et al. 2009b) or methods that match people based on their profiles, ignoring the changing nature of user profiles. However, more comprehensive recommendation methods, such as methods that recommend new friends who are previously unknown but share similar interests and backgrounds, may be more valuable to users than methods that merely rediscover existing friends (Chen et al. 2009b), especially in situations such as traveling to a new city and seeking a date. Unfortunately, little research has focused on recommending strangers to users in SNSs.

Basic profile matching has some disadvantages for stranger recommendations: (1) It does not comprehensively analyze a user's life pattern and interests since the information available for the matching is restricted to the user profiles provided by the SNS; (2) new users may not have complete profiles; and (3) old users may forget to update their profiles. Due to these problems, basic profile matching may not yield a good recommendation (Zheleva et al. 2010).

Friend-of-friend is a very efficient and economical method for recommending existing friends because it analyzes entire social networks and finds overlapping links of friends, implying a real-life connectivity of users (Al Hasan et al. 2006; Lichtenwalter et al. 2010). However, this method is not useful for recommending unknown users to each other because two strangers sharing the same interests will probably not have any common friends. Thus, the friend-of-friend method would most likely miss this kind of recommendation.

With the recent advances in location-aware mobile devices (e.g., GPS-enabled portable devices, smart phones, tablets, and wearable devices), wireless communication

technologies (e.g., 3G, LTE, and Wi-Fi), map services (e.g., Google Maps, Microsoft Bing Maps, and Yahoo! Maps), and spatial database management systems, location-based social networking applications have been moving at a fast pace. A survey by the technology research firm RNCOS suggests that the market of mobile location technologies will grow at an annual compound rate of 20% (Carroll 2010). The increasing popularity of location-based applications enables people to conveniently log the locations they have visited with spatial-temporal data. Such real-world location histories imply users' favorites and bring us opportunities to understand the correlation between users and locations (O' Madadhain et al. 2005; Shi 2013). This motivates us to strive to address the following research questions: 1) How will location information imply users' interests and lifestyles? 2) How could those implied interests and lifestyles help improve friend recommendation performance?

In this study, we propose a new method for building a more comprehensive friend recommendation system for location-based SNSs. In our method, the system first records users' check-in data. Then, the location information is transformed into check-in history distributions, physical geographic data, and types of frequently visited locations. The system calculates the similarity/dissimilarity between two users by comparing their demographic attributes, social-tie attributes, and location attributes. Next, using data mining techniques, it classifies a pair of users as potential friends or otherwise. Finally, the system sorts the probability output of the classification to make a top-M friend recommendation. The results from evaluation with real-world data show that by adding location information, our proposed method significantly improves friend recommendation performance.

The rest of this paper is organized as follows. In the next section, we review related research. The details of our method are presented in Section 3. In Section 4, we describe our evaluation and present the results. Finally, we discuss the contributions, implications, and limitations of this study.

2. Related Work

2.1 Friend Recommendation Systems

According to the recommended objects, recommendation systems in online social networks can be categorized into two types: item recommendation systems and friend recommendation systems (Adomavicius et al. 2005; Adomavicius and Tuzhilin 2005). Item recommendation systems suggest interesting items such as movies, songs, books, and other products, to a user. Friend recommendation systems recommend to a given user homogeneous users in the same social network in order to help the user discover expertise, potential friends, old acquaintances, etc.

Item recommendation systems have been extensively studied (Arazy et al. 2010; Chen 2013; Christidis and Mentzas 2013; Deng et al. 2013; Gavalas and Kenteris 2011; Park et al. 2012; Sankaradass and Arputharaj 2011). Three kinds of filtering methods in item recommendation systems have been proposed: collaborative filtering methods, content-based filtering methods, and hybrid methods. Collaborative filtering methods rely on the interactions between users and items such as: How frequently does a user buy/browse an item? How does a user rate an item? Content-based filtering methods focus on the

attributes of an item without considering interactions between users and items. Hybrid methods combine both collaborative and content-based filtering methods.

Friend recommendation systems have been much less studied, despite their increasing importance to both users and service providers in SNSs (Tian et al. 2010b). Friend recommendation systems could help new users who start off in SNSs without friends (Adomavicius and Tuzhilin 2005; Park et al. 2012). When a new user signs up with an SNS, the user has no friend and, therefore, cannot enjoy sharing activities with other users. The user may feel bored and leave the platform. Recommending suitable friends to new users is essential. Existing users may also find friend recommendation systems beneficial. Finding users who share similar interests and habits could broaden their friend networks, enrich their social activities, allow them to share contents to more people, enhance loyalty to the website, and improve their satisfaction with the SNS.

Friend recommendation systems are equally beneficial to the service providers. Friend recommendation systems boost the social network densities and bring higher active interactions among users, providing natural and valuable channels for the propagation of news, advertisements, and trends, which could be transformed into great market potential. By helping users strengthen their social connectivity, service providers can increase their market share of their services.

Recently, leading SNSs, such as Facebook and LinkedIn, have added “people you may know” features to their homepage, which suggest new connections (Scellato et al. 2011). These features use users’ contact information, profiles, or common friends to make friend recommendations. Based on these attributes, the recommendation results will tend to

include only people the user already knows in real life. However, in some situations, recommendation systems that cover a more comprehensive range of users, recommending strangers who share similar interests, will be more valuable to users. For example, when a user travels to a new city, it will be valuable for the user to meet new friends who share similar habits and lifestyles because they know more about the city and could give advice that caters to the user's interests in the city.

Another example is an online dating website, which allows individuals to make contact and communicate with each other over the Internet, usually with the objective of developing a personal, romantic relationship. Obviously, recommending someone who shares similar interests but is previously unknown is far more desirable than recommending a friend who is already known by the user(Menon et al. 2003).

Although more comprehensive recommendation systems could be very useful in online social networks, there has been little research specifically on recommending unknown people or on making inclusive recommendations. Most existing friend recommendation systems use simple strategies, such as suggesting a "friend-of-friend", e.g., Facebook (Chen et al. 2009b), or trying to match users' profiles.

Profile matching methods compare users based on demographic attributes. For example, Facebook collects users' age, gender, educational background, job positions, favorite items, etc., in online SNSs (Chen et al. 2009b). However, there are several issues with profile matching. First, the demographic attribute sets depend on what the platform provides and may not be comprehensive enough. Second, new users may not complete their profiles, so the demographic characteristics could be sparse and have many unfilled

values. Third, the interests and demographics of existing users may change, but those changes might not be reflected in their profiles.

"Friend-of-friend" or the social-tie method compares users' friend networks, such that two users with many overlapping friend links would have enhanced chances to become friends (Adamic and Adar 2003; Jeh and Widom 2002; Liben Nowell and Kleinberg 2007; Newman 2001). However, the analysis is also not comprehensive because the idea of social-tie recommendation is based on the intuition that in real life, two people who share many friends may also be friends. To recommend known people in a social network provides some value; however, it does not include people who are unknown but share similar habits. A system that includes the latter could provide even more value—an opportunity that the service provider will not want to miss.

2.2 Location-based Information

Today, mobile phone vendors are increasingly producing smartphones that are capable of incorporating a Global Positioning System (GPS). Location adds a complementary value to the product and significantly broadens its applicability to new kinds of services and usage scenarios (Khurri 2009). Many innovative applications and location-based services (LBSs), such as Foursquare, Loopt, etc., were released after 2009. As of April 2012, Foursquare reported it had 20 million registered users and more than two billion check-ins. People like to check in and post their thoughts in different places. The data are then collected by the service providers. This geo-temporal information will be very useful for businesses because users' outdoor movements in the real world could imply more

information about their interests and preferences compared to their online activities (Chen 2009; Ren 2014; Zheng et al. 2011).

For instance, if a person frequently goes to stadiums and gyms, it implies that the person might like sports. Likewise, if a user frequently travels to mountains, it might imply that the user is interested in hiking. According to the first law of geography (Tobler 1970), "everything is related to everything else, but near things are more related than distant things." In other words, people who have similar location histories might share similar interests and preferences. The more similar location histories they share, the more correlated these two users might be. People who visit the same restaurants and shopping malls would tend to share some similar tastes. Visitors traveling to the same lakes and valleys would likely share similar styles of tourism.

.

In turn, the geographical regions visited by users might imply a similar profile. As a consequence, people's location histories cannot only help us understand the similarity between individuals but also reveal the correlations among geographic locations.

The significance of location in friend recommendation systems is also shown in computer-supported social matching process theory. As one of the six major attribute categories in this theory, we believe it is helpful if we collect users' check-in data and place it within our analytic framework. The extracted information will help in making a comprehensive friend recommendation.

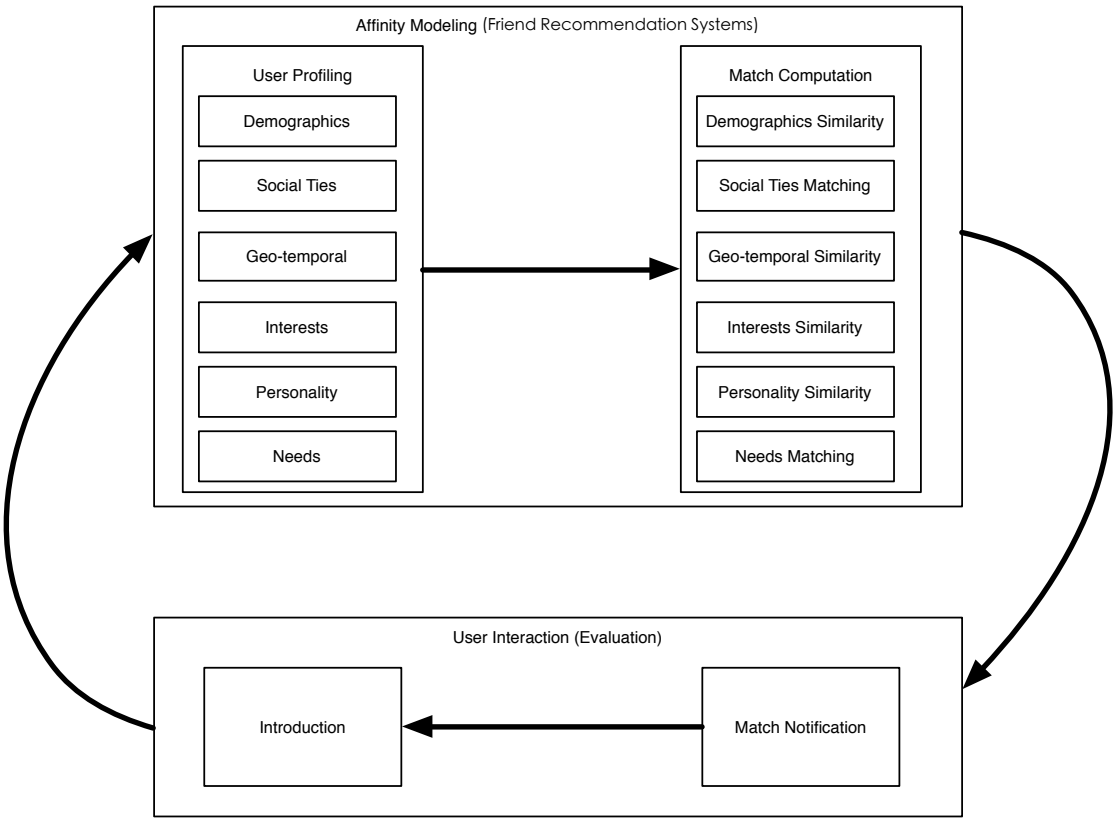


Figure 2-1 Computer-supported Social Matching Process

The computer-supported social matching process model was proposed by Terveen and Mcdonald (2005). This model consists of four steps: modeling, matching, introducing, and interacting. Mayer et al. (2010) more clearly represent these steps by splitting them into two parts: affinity modeling and user interaction. Affinity modeling is the process of gathering data from users to build profiles that enable the system to compute social matches. User interaction includes the interactions between the system and the user necessary to collect data, send a match notification, and facilitate the introduction and interaction between matched users.

Social matching systems calculate user affinities by weighting the similarities between users over a set of user attributes. According to Mayer et al. (2010), there are six different

types of user attributes:

- Demographics (geographical background, educational background, etc.);
- Social Ties (friends, co-worker, relatives, etc.);
- Interests (hobbies, favorites, music, books, etc.);
- Geo-temporal Patterns (frequently visited places, mobility traces, proximity patterns, etc.);
- Needs (partner, help, knowledge, etc.);
- Personality (extraversion, neuroticism, agreeableness, conscientiousness, openness, etc.)

From this model, we can see the basic profile matching is actually using demographic attributes, and the "friend-of-friend" is using social ties attributes. But, in fact, much more could be done to make comprehensive friend recommendations by using additional or other user attributes. In this essay, we will focus on using geo-temporal pattern attributes.

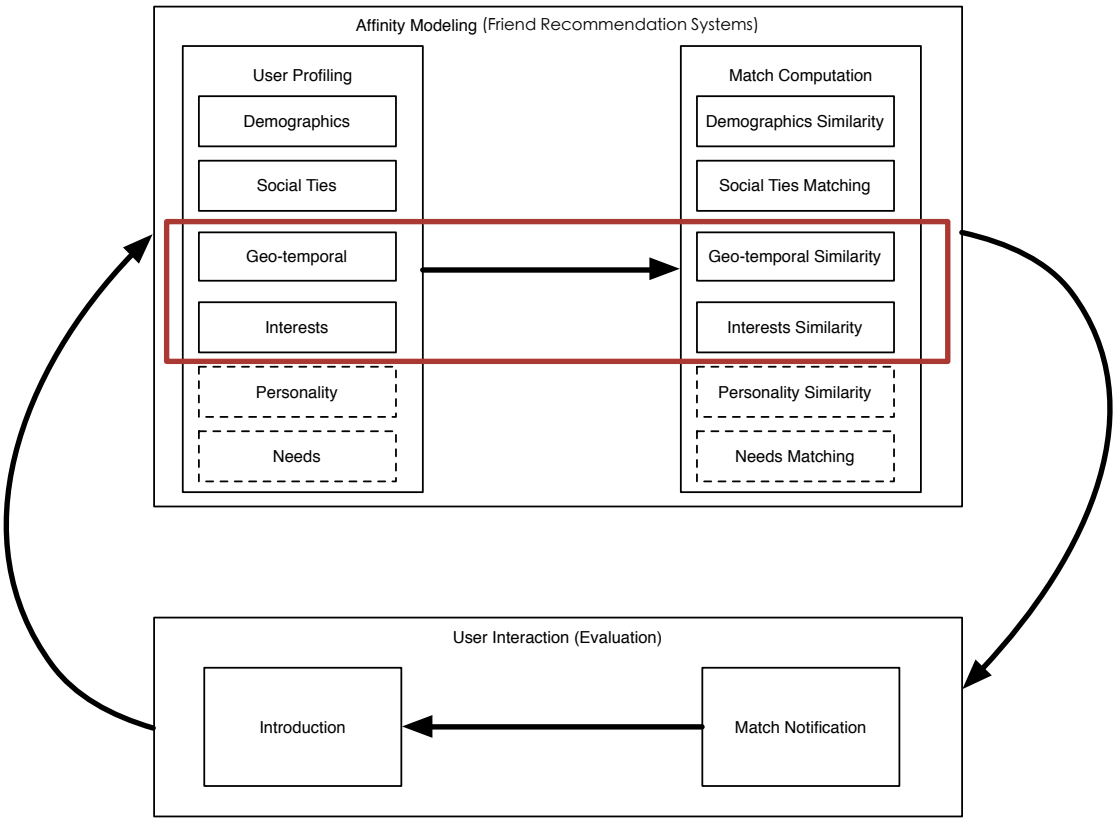


Figure 2-2 Location-sensitive Social Matching Process

3. Model

3.1 Overview

In this study, we propose a novel model that includes users' location information for discovering users' shared interests and lifestyle patterns to make recommendations for unknown people. The system overview is shown in Figure 3-1.

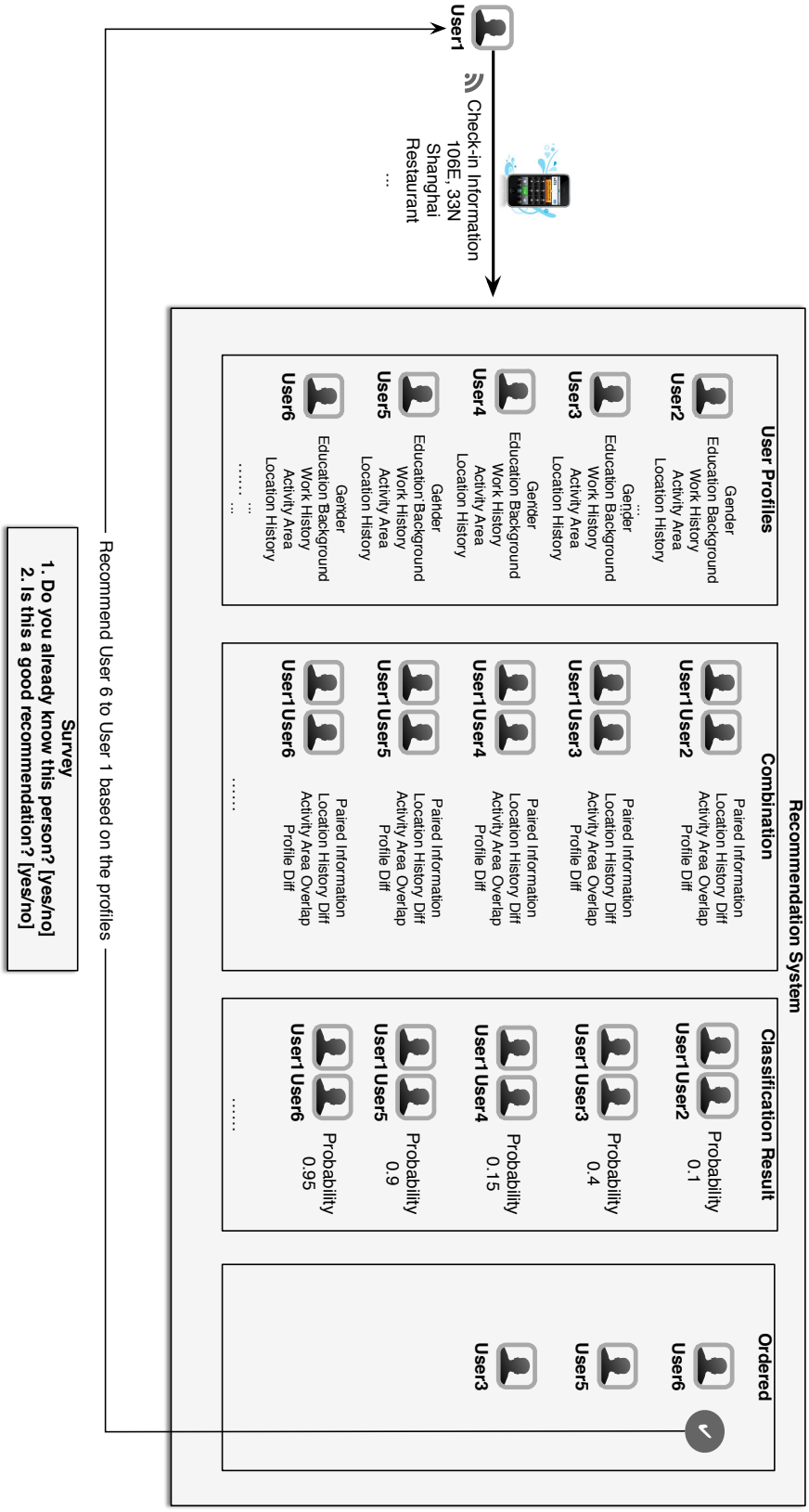


Figure 3-1 Overview of the Model

Our recommendation model is divided into five steps:

1) The system first collects users' location information and puts it into our location analytic framework.
2) In the location analytic framework, the location information is divided into three parts: the geographic attributes, the point of interest (POI) attributes, and check-in history distribution.
3) We then calculate the similarity of attributes between each pair of users.
4) Data mining classifiers are employed for classifying friends. The dependent variable is whether two users are linked or not, i.e., whether they are friends.
5) The classification results are sorted based on probability estimates. The users who have the highest top M probabilities are included in the recommendation list.

3.2 Location Analytic Framework

To handle the location information, we develop a location analytic framework that divides location data into three parts: the geographic attributes, point of interest attributes, and check-in history distribution.

1) Geographic attributes

Normally, location-based services will provide the record for geographic attributes such as longitude, latitude, and altitude. Physical location may have some implications in friend recommendations. For example, two people sharing the same hometown could have similar kinds of experiences growing up, could be involved in the same events, study in the same school, etc. Also, two users living in the same neighborhood have a greater chance of meeting each other and enriching their activities online or offline.

2) Point of interest attributes

Point of interest (POI) is a specific point location that people may find useful or interesting. It can be a building, tourism spot, hotel, restaurant, etc. that people might be going to. There are thousands of points of interest in each city throughout the world. Points of interest can be distinguished by their name, type of location, street location, etc. Points of interest have some components describing their details. It can be a description, an image, or a latitude and longitude. To attract visitors, usually a POI provides interesting information about itself. It can include the number of check-ins, which would indicate how popular this POI is.

POI data could provide some implications about users' preferences. For example, if a user visits Chinese restaurants frequently, the user probably likes Chinese food. Two users who visit the same POIs could also share similar lifestyles.

3) Check-in history distribution

We are also interested in users' check-in histories, which provide a chance to systematically analyze users' check-ins. We record the frequency of each POI type and build the distribution. To simplify our model, we do not consider other dimensions in the distribution, such as time and check-in sequence. The similarity of two check-in distributions would describe similar lifestyles, which would imply similar interests.

Then, we calculate the similarity of location attributes between two users.

1) Geographic attributes

We employ Haversine Formula to calculate the distance between two geographic coordinates (http://www.movabletype.co.uk/scripts/latlong.html). We also calculate the overlaps of the activity areas. To simplify the calculation, we find the farthest east, farthest west, farthest north, and farthest south of users' check-in histories, and then assume the area is a rectangle. The overlap is easy to calculate:

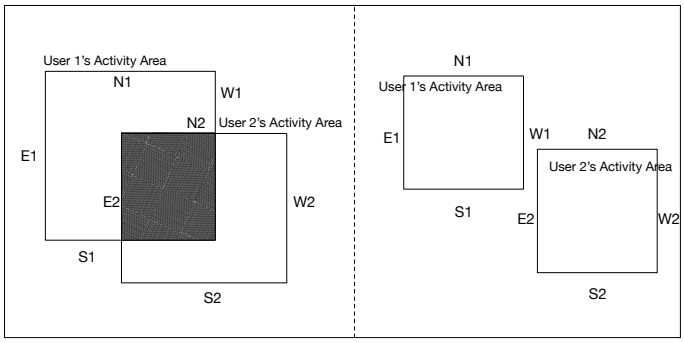


Figure 3-2 Calculation for Activity Area Overlap

$$A(u_1, u_2) = \begin{cases} (W_1 - E_2) \cdot (S_1 - N_2) & if\ W_1 > E_2\ and\ S_1 > N_2 \\ 0 & if\ W_1 \leq E_2\ \ or\ \ S_1 \leq N_2 \end{cases}$$

2) Point of interest attributes

To determine the similarity of POI information, we calculate the shared types of POIs in two users' check-ins. For example, check-in history of user #1 indicates this user has checked-in at two restaurants, three gyms, and five parks, and user #2 has checked-in at three restaurants, one gym and two parks. Therefore, the two users would have shared two restaurants, one gym, and two parks.

3) Check-in history distribution

We calculate the check-in history distribution similarity by employing Kullback-Leibler Divergence (Kullback S; Leibler, R.A. 1951). In the discrete case, let f and g be two probability mass functions in a discrete domain ID, with a finite or countable infinite number of value. The Kullback-Leibler divergence between f and g is:

$$D(f|g) = \sum_{x \in ID} f(x) log \frac{f(x)}{g(x)}$$

In our situation, f(x) and g(x) are the users' check-in histories based on different categories of POIs. The K-L divergence is not symmetric, so we calculate both D(f|g) and D(g|f) as the similarity of two users' check-in histories.

Our location analytic framework is summarized in Figure 3-3

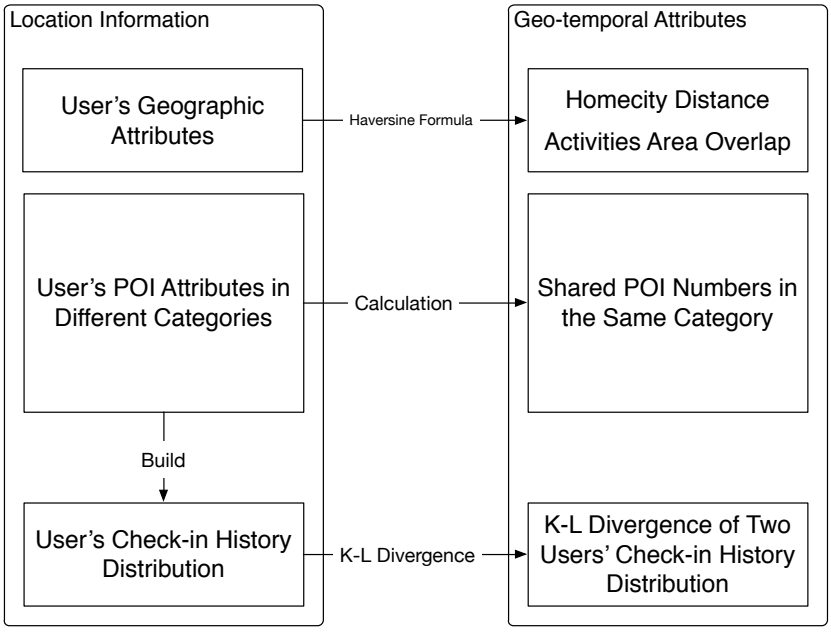


Figure 3-3 Location Analytic Framework

3.3 Classification

After we have calculated the similarity between each pair of users, we employ probabilistic classifiers to make classifications. The non-probabilistic classifiers in data mining always provide a strict output. For example, in this case, they only give as output whether two users are friends or not. However, in this study, we need to sort the probabilities and find the most probable friends. So non-probabilistic data mining algorithms are not suitable for this study. We use Naïve Bayes, Bayesian Network, and Logistic Regression in this study.

After the classification, our model collects all probability output. We sort the probability and make the top M friend recommendation.

4. Experiment

4.1 Data

To attract software programmers to develop plugins or applications for the SNSs, platform providers release application programming interfaces (APIs) to developers. The APIs are normally capable of collecting some restricted data from the platform when users accept. I applied for developer membership in Foursquare and Facebook.

I started to collect training data in October 2011 based on the Foursquare's public API. I wrote a Java program to scan all public timelines in Twitter. Then I sent friend requests to randomly selected users, some of whom accepted the request and some of whom rejected it. After the acceptances, we were able to collect these users' profiles, check-ins, text information, etc. I tried to keep collecting all the check-in information they posted.

From October 2011 to February 2013, I received 997 users' acceptances and 6,417 check-in information. The friendship network was recorded as well. There were 4,074 pairs of friends. The connectivity is not high and, on average, one user had four to five friends. From Foursquare, I built a POI type database that contained nine major types and 420 sub-categories. The database was hierarchical and tree-structured.

We did not only collect data from Foursquare to make the experiment. We collected data from Facebook too. In the Facebook platform, we were able to extract users' demographic data. The attributes we obtained include: users' name, gender, friend count, tip count (which indicated how active they are in the social network), religion, political, birthday, educational background, work positions, language spoken, and favorite sports. As we discussed before, some profiles were not complete, and many values were blank.

Using the data from the SNS and our location analytic framework, we developed the user model shown in Table 4-1:

Demographic Attributes	
Gender	Male: 617, female: 353
Age	Range: 18 - 64, mean: 30.1
Religion	There are 30 different religions.
Political view	There are 24 different political views.
Highest education	High school: 88, College: 241, Graduate school: 56
Work type	There are 38 different types of work.
Favorite sports	There are 63 different sports.

Languages	There are 51 different languages.
Tip count	The number of tips the user has.
Tip-likes count	The number of "likes" the user's tips have received.
Location Attributes	
Check-in count	The number of check ins the user has. Range: 0 - 173, mean: 6.5
Home city	The home city of the user.
Art and entertainment check-ins	The number of check-ins at art and entertainment locations. Range: 0 - 83, mean: 1.1
College check-ins	The number of check-ins at college locations. Range: 0 - 24, mean: 0.4
Food check-ins	The number of check-ins at food locations. Range: 0 - 49, mean: 2.5
Professional check-ins	The number of check-ins at professional locations. Range: 0 - 33, mean: 0.7
Nightlife check-ins	The number of check-ins at nightlife locations. Range: 0 - 43, mean: 0.7
Outdoor check-ins	The number of check-ins at outdoor locations. Range: 0 - 56, mean: 0.9
Shop check-ins	The number of check-ins at shop locations. Range: 0 - 41, mean: 1.6
Travel check-ins	The number of check-ins at travel locations. Range: 0 - 15, mean: 0.7
Residence check-ins	The number of check-ins at residence locations. Range: 0 - 6, mean: 0.4
Area	The physical geographic area (longitude and latitude) covering the check-ins of the user.

Table 4-1 Attributes in the Collected Dataset

We then calculated the similarity values between users. For two integer attributes, such as difference in friend count, tip-like count, tip count, and check-in count, we used the Jaccard coefficient (Salton and McGill 1983):

$d(a,b) = |\frac{a-b+\delta}{(a+b)+\delta}|$, where δ is a small smoothing factor and was set to 0.001 in our evaluation

And the difference was the relative difference, which was between 0 and 1.

For POI category attributes, we measured how many similar check-ins two users share using another type of Jaccard coefficient (Kuo et al. 2013; Scellato et al. 2011; Schifanella et al. 2010; Wang et al. 2011; Xu et al. 2011):

$$d(a,b) = \frac{|a \cap b|}{|a \cup b|}$$

For the check-in history distribution, we calculated the K-L divergence (Dahlhaus 1996; Kullback and Leibler 1951). Table 4-2 summarizes the similarity/dissimilarity measures we used.

Demographic similarity/dissimilarity	
Gender	The genders of two users. Male and female: 49.05%, two males: 38.32%, two females: 12.63%
Age difference	Range: 0 - 46, mean: 4.33
Same religion	Whether two users have the same religion: False: 22.2%, True:0.2%, Unknown: 77.6%
Same political view	Whether two users have the same political view: False: 16.5%, True:0.1%, Unknown: 83.4%
Same education	Whether two users have the same highest education: False: 76.73%, True: 23.27%
Same work type	Whether two users have the same work type: False: 51.98%, True: 0.4%,

	Unknown: 47.6%
Same favorite sport count	The number of sports two users both like. Range: 0 - 3, mean: 0.6
Same language count	The number of languages two users both speak. Range: 1 - 3, mean: 1.8
Tip count difference	The relative difference between two users' tip counts
Tip-likes count difference	The relative difference between two users' total tip-likes counts
Social-tie similarity/dissimilarity	
Friend count difference	The relative difference between two users' friend counts
Common friends	The number of common friends two users share. Two measures are used, one is in the collected data set only, and the other in the Foursquare platform.
Location similarity/dissimilarity	
Check-in count difference	The relative difference between two users' total check-in counts
Home city distance	The physical distance between two users' home cities. Range: 0 - 1.9k km
Common art and entertainment check-ins	The number of check-ins two users both have at art and entertainment locations. Range: 0 - 40, mean: 0.2
Common college check-ins	The number of check-ins two users both have at college locations. Range: 0 - 11, mean: 0.05
Common food check-ins	The number of check-ins two users both have at food locations. Range: 0 - 34, mean: 0.8
Common professional check-ins	The number of check-ins two users both have at professional locations. Range: 0 - 15, mean: 0.1
Common nightlife check-ins	The number of check-ins two users both have at nightlife locations. Range: 0 - 19, mean: 0.1
Common outdoor check-ins	The number of check-ins two users both have at outdoor locations. Range: 0 - 33, mean: 0.1

Common shop check-ins	The number of check-ins two users both have at shop locations. Range: 0 - 23, mean: 0.5
Common travel check-ins	The number of check-ins two users both have at travel locations. Range: 0 - 13, mean: 0.1
Common residence check-ins	The number of check-ins two users both have at residence locations. Range: 0 - 6, mean: 0.1
Area overlap	The overlap between the physical geographic areas covering the check-ins of two users.
Check-in history distribution difference	The K-L divergence between two users' check-in history distributions. There are two attributes, because K-L divergence is not symmetric.

Table 4-2 Similarity/dissimilarity Measures Derived

4.2 Experiment Design

Our model transfers a recommendation question into a classification question after the calculation of similarities between pairs of users. Each classification record is a pair of two users and their similarity attributes. The dependent variable is whether two users were friends or not. In this experiment, we had three variables to control: connectivity of the friend network, attribute groups, and how many friends to recommend.

First, we want to simulate the real-world online social networking connectivity. Our data set is limited and has a relatively sparse friend network density in which only 1% of links are friend links. To make a better simulation, we tried to select links in our test data set. By controlling the proportion of friend/non-friend links in the link set, we have social networks with different densities of connections. Here is a simple example for different densities of a five-user network: We select the links of A-B, B-C, C-D, D-E, E-A as friend links, and in this case, the proportion of friend/non-friend links is 1:1. In our data set, I select 1:1, 1:2, 1:5, and 1:10 as the proportion of friend/non-friend links in social

networks. After the selection process, we have four different data sets, which have different numbers of users. The 1:1 data set has 835 users, the 1:2 data set has 891 users, the 1:5 data set has 936 users, and the 1:10 data set has 957 users. The number of users could impact the experimental results, which we will explain later.

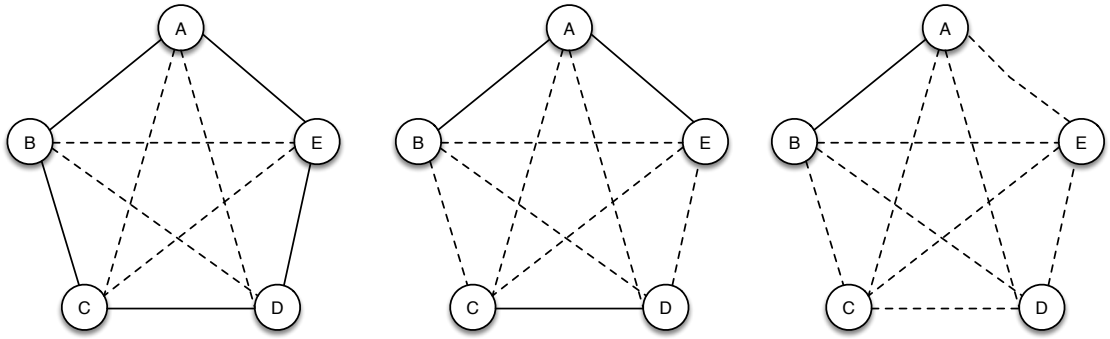


Figure 4-1 Example of proportion of friend: non-friend

Second, we want to examine the performance between different attribute sets. To compare our model with the existing profile matching recommendation methods or "friend-of-friend" recommendation method, we select different groups from Table 4-3:

Group	**Attribute Data set**
Group 1	Demographic Attributes Only
Group 2	Demographic Attributes + Location Attributes
Group 3	Demographic Attributes + Social Ties Attributes
Group 4	Demographic Attributes + Location Attributes + Social Ties Attributes

Table 4-3 Test Attribute Groups

We compare Group 1 with Group 2 to see if the location attributes help in simple profile matching recommendations, and we compare Group 3 with Group 4, which could prove whether the location attributes help in "friend-of-friend" recommendations.

Finally, we evaluate our experimental results by changing how many friends we want to recommend. Recommending too few friends may reduce the chance of users finding a friend, but recommending too many friends might look like a random guess and make it difficult for users to select.

4.3 Results

The experiment platform we use is Weka 3.6.10; in the classification test settings, we use 10-fold cross validation; and we first use the accuracy as the evaluation result. By definition, we have:

$$accuracy = \frac{number\ of\ true\ positive + number\ of\ true\ negative}{number\ of\ records\ in\ test\ data\ set}$$

And, because our friend/non-friend network data set was very biased when the proportion went from 1:1 to 1:10, the classifiers could put all classification output to negative to get a better accuracy. In a 1: P proportion friend network, we calculate the base accuracy rate as:

$$base\ accuracy = \frac{P}{1+P}$$

To alleviate the effect of classification bias, we make the evaluation cost sensitive. The settings of the cost matrix are:

Proportion	Cost Matrix
1:1	$\begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}$
1:2	$\begin{vmatrix} 0 & 2 \\ 1 & 0 \end{vmatrix}$
1:5	$\begin{vmatrix} 0 & 5 \\ 1 & 0 \end{vmatrix}$
1:10	$\begin{vmatrix} 0 & 10 \\ 1 & 0 \end{vmatrix}$

Table 4-4 Settings of Cost Matrix

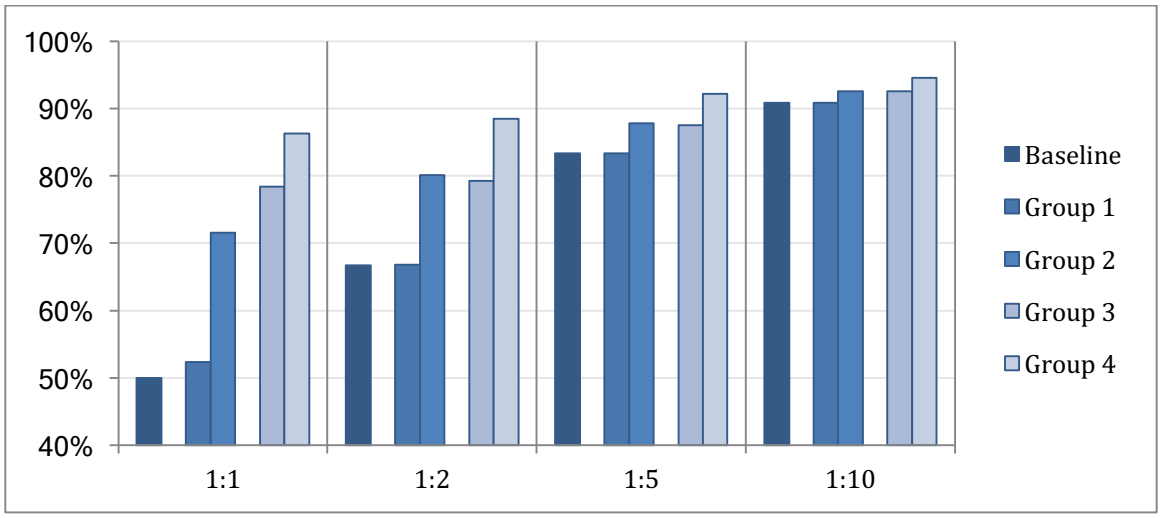


Figure 4-2 Accuracy of Friend Recommendation

Attribute Sets	1：1	1：2	1：5	1：10
Baseline Accuracy	50%	66.7%	83.3%	90.9%
Group 1	52.4%	66.8%	83.3%	90.9%
Group 2	71.6%	80.1%	87.8%	92.6%
Group 3	78.4%	79.3%	87.5%	92.6%
Group 4	86.3%	88.5%	92.2%	94.6%

Table 4-5 Accuracy of Friend Recommendation

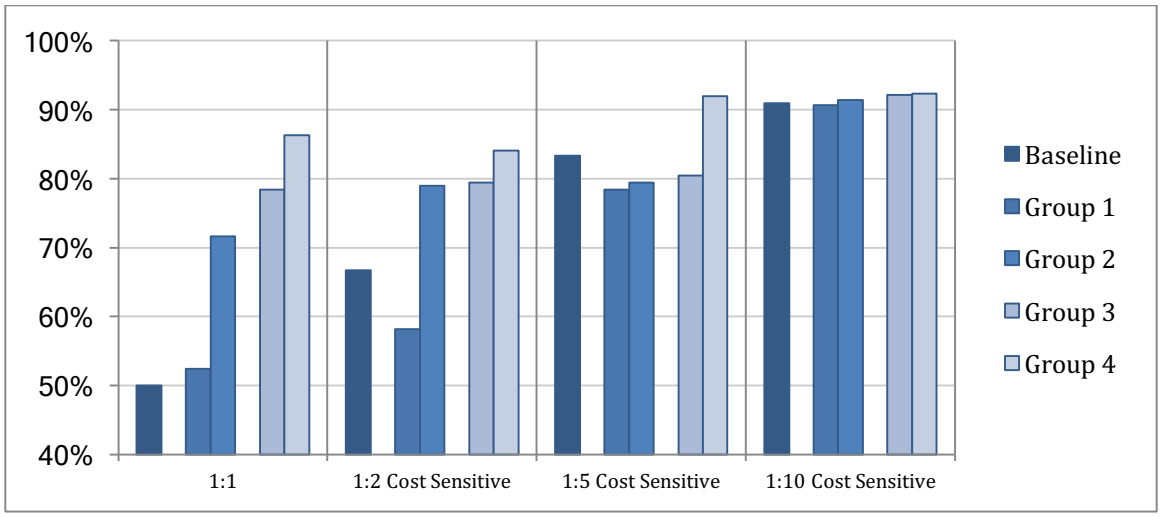


Figure 4-3 Accuracy of Friend Recommendation in Cost-sensitive Case

Attribute Sets	1: 1	1: 2	1: 5	1: 10
Baseline Accuracy	50%	66.7%	83.3%	90.9%
Group 1	52.4%	58.15%	78.44%	90.67%
Group 2	71.6%	78.96%	79.45%	91.36%
Group 3	78.4%	79.44%	80.49%	92.10%
Group 4	86.3%	84.09%	91.99%	92.30%

Table 4-6 Accuracy of Friend Recommendation in Cost-sensitive Case

From the accuracy results, we can see that in the Group 1, with only demographic attributes, the recommendation results could be just as similar as the random guess (baseline accuracy). The sparse profile attribute sets do not help. And with location information, the accuracy significantly improves, with all accuracy output having higher values than Group 1. For Group 4 and Group 3, the same thing happens. With location attributes, the accuracy outputs are considerably higher in Group 4, which suggests that location attributes help in social-tie recommendations.

To further evaluate the recommendation performance, we need to simulate the Top M recommendation process and then to calculate the precision. By using the classification probability results from the Weka output, we sort and select the top M users. Finally, we

calculate the correct rates to recommend a true positive friend (i.e., predict as a friend someone who is actually a friend).

To evaluate the precision of our algorithms, we need to calculate the baseline of precision and the optimal line of precision. Theoretically, the calculation for the optimal precision only depends on the connectivity of social networks and the number of friend recommendations. But because our data set is very sparse, we need to consider each user's friend links.

Assume we have n users, and in a 1: P proportion network, for each user i, we have friend link number F_i, non-friend link number N_i, and we want to recommend M friends in the list.

Baseline Precision:

For each user, if the total number of links $F_i + N_i$ is less than the number of recommendations M, then all friend links would be in the recommendation list, so the precision is F_i/M. Otherwise, the number of possible ways to select M links is C_{F+N}^{M}. The number of possible ways to select x friend links and M-x non-friend links is: $C_F^x \times C_N^{M-x}$. The expected precision of random Top M recommendation for this user is:

$$Baseline\ Precision_i = BP_i = \frac{\sum_{j=0}^{M} j \cdot C_{F_i}^{j} \cdot C_{N_i}^{M-j}}{C_{F_i+N_i}^{M} \cdot M}$$

And the average baseline precision for the data set is:

$$Baseline\ Precision = (\sum_{1}^{n} BP_i) \div n$$

Optimal Precision:

For each user, the selected friend link number will be: min (F_i, M), so, for Top M Recommendation, the optimal precision is:

$$Optimal\ Precision = (\sum_{1}^{n} \frac{\min(F_i, M)}{M}) \div n$$

So, for Top3, Top5 and Top 10 Recommendations, we will have the optimal precisions:

	Number of Users	Optimal Precision in Top 3 Recommendation	Optimal Precision in Top 5 Recommendation	Optimal Precision in Top 10 Recommendation
1:1 Data Set	835	56.846307%	42.562874%	24.395210%
1:2 Data Set	891	53.273475%	39.887767%	22.861953%
1:5 Data Set	936	50.712251%	37.970085%	21.762821%
1:10 Data Set	957	49.599443%	37.136886%	21.285266%

Table 4-7 Optimal Precisions in Top 3,5,10 Recommendations

And the baseline precisions will be:

	Number of Users	Baseline Precision in Top 3 Recommendation	Baseline Precision in Top 5 Recommendation	Baseline Precision in Top 10 Recommendation
1:1 Data Set	835	26.613807%	27.989746%	32.297519%
1:2 Data Set	891	14.776371%	17.759222%	18.170754%
1:5 Data Set	936	5.350665%	6.971847%	9.244579%
1:10 Data Set	957	2.190556%	2.901769%	4.220156%

Table 4-8 Baseline Precisions in Top 3,5,10 Recommendations

When we have the baseline and the optimal precisions, we can also calculate the relative positions of our recommendation precisions. The formula for the relative positions will be:

$$Position = \frac{Recommendation\ Precision - Baseline\ Precision}{Optimal\ Precision - Baseline\ Precision}$$

Then we normalize all the results for the top 3 recommendations and place them in the same chart:

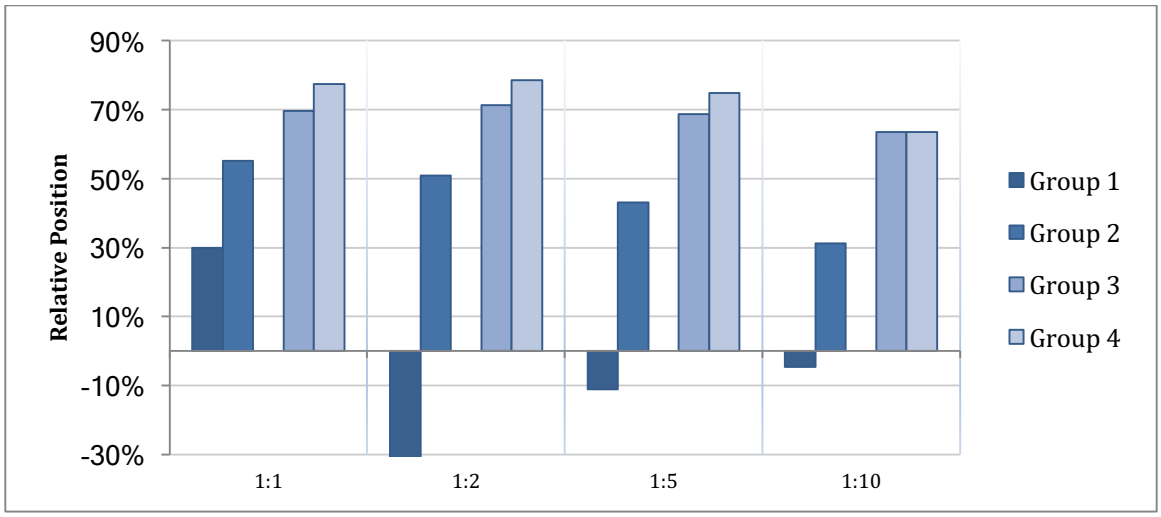


Figure 4-4 Top 3 Friend Recommendation Precisions

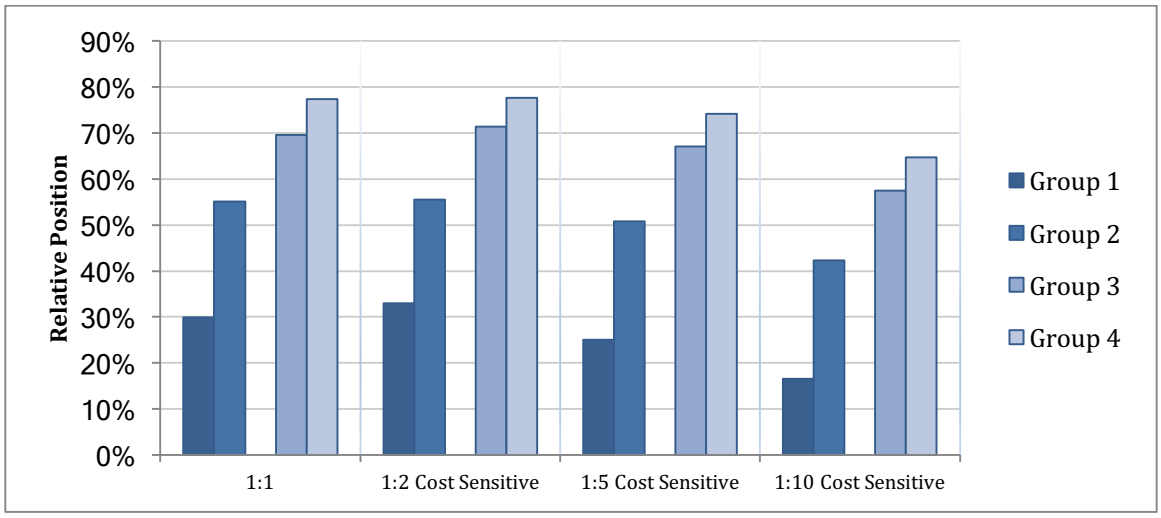


Figure 4-5 Top 3 Cost-Sensitive Friend Recommendation Precisions

	1:1	1:2	1:2 Cost Sensitive	1:5	1: 5 Cost Sensitive	1:10	1:10 Cost Sensitive
Group 1	28.96%	-33.04%	32.95%	-11.17%	25.02%	-4.62%	16.61%
Group 2	55.11%	50.83%	55.49%	43.16%	50.78%	31.23%	42.33%
Group 3	58.59%	63.65%	63.74%	63.5%	61.58%	58.79%	40.68%
Group 4	77.42%	78.52%	77.65%	74.80%	74.17%	63.56%	64.73%

Table 4-9 Relative Positions of Top 3 Friend Recommendations

From Figure 4-4 and 4-5, we found that for the Top 3 friend recommendation precision, as discussed before, the location attributes also improve the performance. The Group 2 results are significantly better than the Group 1 results, and Group 4 has slightly less

improvement but still significantly superior results compared to Group 3. Also, we can see that cost-sensitive classification results are more reasonable than the higher-biased data set results. When we only have demographic attributes, cost-sensitive classifiers have much higher precision than the non-cost-sensitive cases. Another trend we observed is that when the proportion goes up, the relative position goes down. The reason could be that when data sets get larger, as in Top 3 friends recommendation, it is harder to reach the optimal line. We saw the trends when we manipulated the Top M friend recommendations as follows.

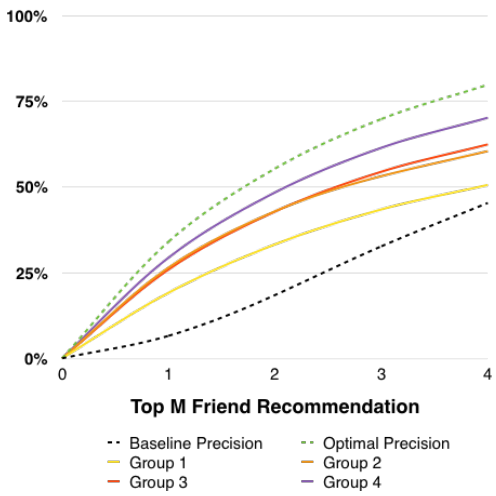
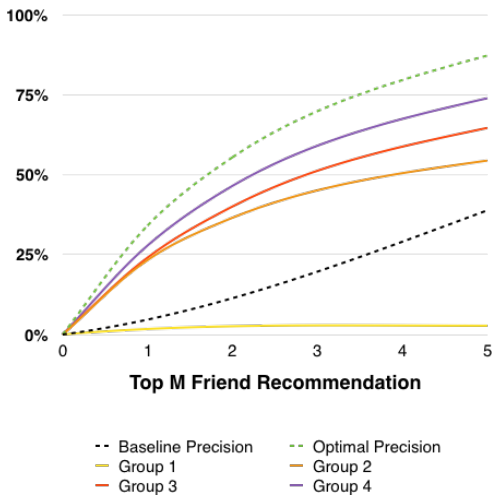
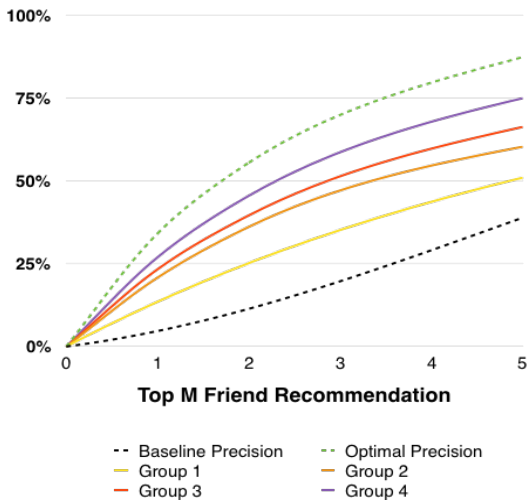
To make the evaluation more comprehensive, we calculated and generated the chart of precision based on recommendation number. The x-axis of the performance chart is the top M friend recommendations, and the y-axis is the ratio of the true friend links (the number of actual friend links that have been recommended as friend links by the system) to the length of the recommendation list (M). Because the friend links are different for each user, we report the average value.

The maximum value of x-axis is related to the total number of links we had in the test data set. The highest value is the maximum number of links for a user, which could be more than a hundred. So, to get an applicable maximum number, we selected the average friend links and added a bit more. For example, in 1:1 data set, we have 2,037 friend links, 2,037 non-friend links, and 835 users, so the average friend links will be (2,037 + 2,037) / 835 ≈ 5.

Because we are not going to the maximum number in x-axis, we will not reach the 100% value in y-axis. And the maximum precision our recommendation will have, depend on

the accuracy result of classification, which means that it cannot reach 100% and get flat after some value of x.

Figure 4-6 shows the performance charts for different proportions with/without cost-sensitive classification:

	Not Cost Sensitive	*Cost Sensitive*
1:1		
1:2		

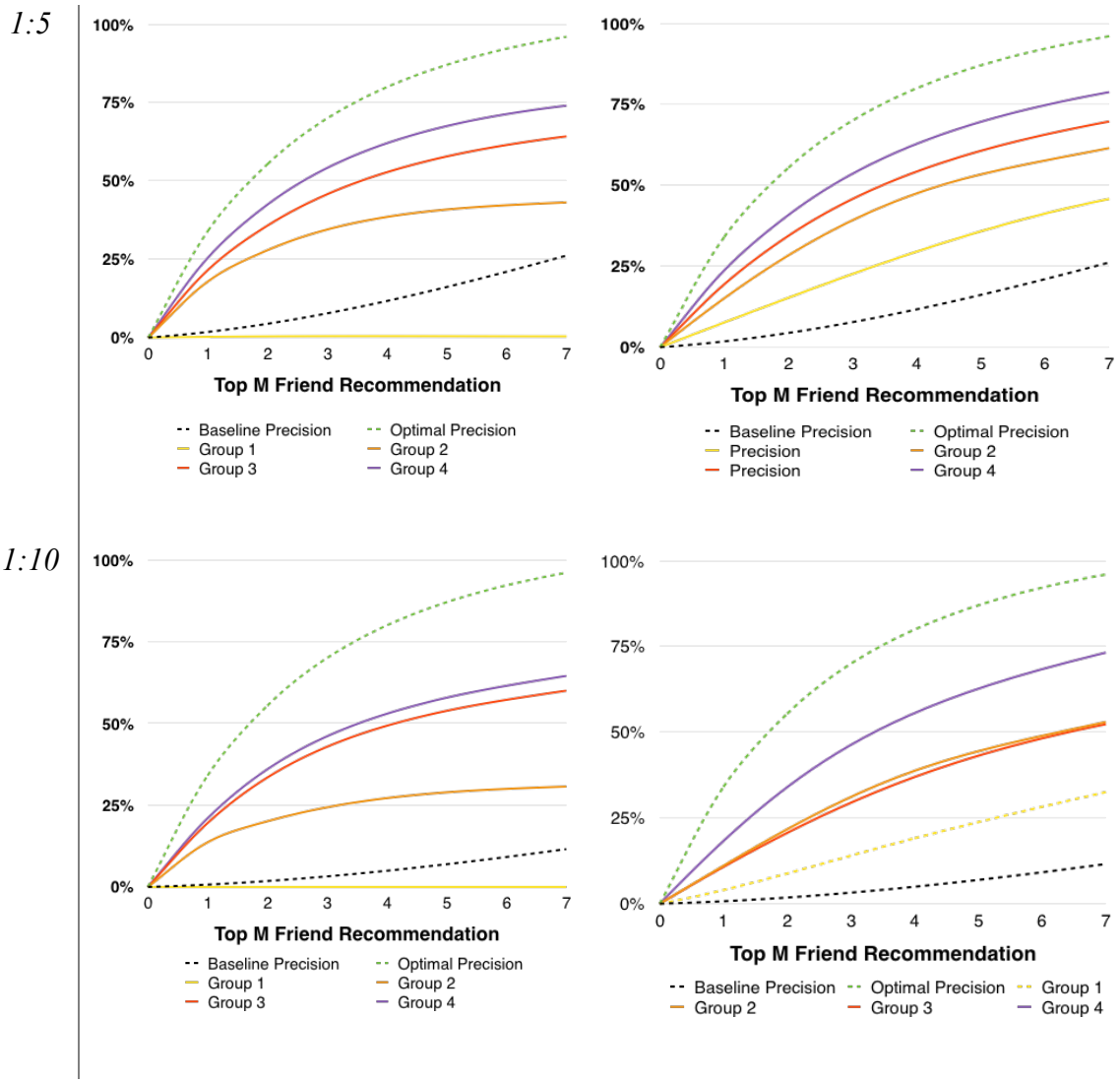


Figure 4-6 Performance Charts for Friend Recommendations

From the performance charts, we can see the evaluation more clearly, and we can see the trends when the recommendation numbers are changed. For a high connectivity SNS such as our 1:1 data set, if we recommend more than three users in the list, the performance of recommendations would not be much better than a baseline performance. And in a sparser SNS, we could recommend more users to reach the highest performance. For example, in 1:5 and 1:10 data set, we would recommend approximately six to seven users.

5. Discussion

In this study, we proposed a novel model for a comprehensive friend recommendation system. By following the guidelines of a computer-supported social matching process, the geo-temporal attribute sets were applied in our model. We developed a location analytic framework, and in this framework, the location data were systematically analyzed. The results of our experiment show that in both profile matching recommendation and "friend-of-friend" recommendation, by adding our location attributes, the performance of recommendations significantly improved.

We make several contributions in this essay with respect to both research and practice:

1) From the standpoint of academic research, to the best of my knowledge, this is the first study that uses location information to make comprehensive friend recommendations. Previous research focused on how to select well-defined demographic profiles or how to improve the efficacy for social-tie friend recommendations. But we have studied, discussed and discovered that different location attributes could imply people's habits and lifestyles. The experimental results show that well-structured location attributes could achieve higher accuracy and better precision outputs for friend recommendations.
2) In our essay, we built a model to test the computer-supported social matching process. For this process, Terveen and McDonald (2010) provided a guideline for a more complete friend recommendation. This process has six types of attributes, and in this study, we verified how the geo-temporal attributes work. For future research, we could focus on other types of attributes.

3) From the standpoint of practice, we built an applicable location analytic framework that systematically summarizes location features with three categories: physical location attributes, POI attributes, and check-in history attributes. The implications of these three categories of attributes are discussed. (1) Physical location infers people may be involved in similar types of environments. (2) POI indicates users' interests in different types of locations. (3) Check-in history provides a chance to systematically analyze users' lifestyles.
4) We implemented a protocol for a location-sensitive friend recommendation system. In this protocol, we collected users' demographic, social tie, and location data, put them in the attribute sets, and then calculated the similarity between users. After that, we classified our records, sorted the probability of classification outputs, and then made the recommendation. The system is relatively easy to implement. A social networking site could follow our steps and quickly create a comprehensive friend recommendation system.
5) We provided a suitable method to evaluate the recommendation performance, not only for its accuracy. We also found the recommendation precision depends on the number of users in the list. In our results, we could see if the connectivity density in the social network is high. If it is, it is better to recommend less people, otherwise the performance may not be good, and for a sparse social networking website, recommending six to seven users could be a reasonable solution.

Our study suffers from several limitations:

1) The data we collected were quite sparse and distributed throughout the world, which meant that users shared very few friend links. The low densities of

connectivity in our data sets created some difficulties for the evaluation. We had to select the links to simulate the higher density SNSs. However, the friend links were repeatedly used, and the results, therefore, could be biased.

2) The evaluation is limited. For a good recommendation system, we not only want to know the accuracy or precision but also the satisfaction of its users. To estimate the satisfaction, we have to do a survey after the recommendations, the findings of which we may use in a long-term research project. From the survey, we may then know whether the recommendation really does provide a good suggestion.
3) The attributes we used could be more complete; most of users didn't fill in their religious and political views attributes in Facebook. Even after I collected and processed the data manually, many null values remained in the profiles.
4) Finally, the dependent variable is defined by the friend links we found from the data set, which means two users are already friends in the SNSs. The implication is that these two users are a match, but it is not known whether these users will become friends. Future research should examine the long-term results whereby two users who are previously not friends become friends later.

There are three possible areas for future research:

1) Following the computer-based social matching theory, we have more attribute sets to discover, such as interests, personality, and needs. We could find clues of them from all possible user-generated contents and develop a suitable framework for them.

2) We could develop long-term research on collecting data. We could examine the activities of a user after the user received a recommendation, for example, whether the user tended to link to the person after the recommendation or not. This would provide better ways to evaluate the recommendation system.
3) We can collect data from a certain city or area, which would likely provide a higher density of friend link networks. We can also change the degree of profiles completion, which could provide better results for the basic profile matching process.

References

Adamic, L. A., and Adar, E. 2003. "Friends and Neighbors on the Web," *Social Networks* (25:3), pp. 211-230.

Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. 2005. "Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach," *ACM Transactions on Information Systems (TOIS)* (23:1), pp. 103-145.

Adomavicius, G., and Tuzhilin, A. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *Knowledge and Data Engineering, IEEE Transactions on* (17:6), pp. 734-749.

Al Hasan, M., Chaoji, V., Salem, S., and Zaki, M. 2006. "Link Prediction Using Supervised Learning," *SDM 06: workshop on link analysis, counter-terrorism and security. 2006*.

Arazy, O., Kumar, N., and Shapira, B. 2010. "A Theory-Driven Design Framework for Social Recommender Systems," *Journal of the Association for Information Systems* (11:9), pp. 455-490.

Benchettara, N., Kanawati, R., and Rouveirol, C. 2010. *A Supervised Machine Learning Link Prediction Approach for Academic Collaboration Recommendation*. New York, New York, USA: ACM.

Carroll, J. 2010. "Location Is the New Intelligence," *CA Magazine*, pp. 1-2.

Chen, I. B. X. 2009. "A Framework for Context Sensitive Services: A Knowledge Discovery Based Approach," *Decision Support Systems* (48:1), p. 10.

Chen, J., Geyer, W., Dugan, C., Muller, M., and Guy, I. 2009. *Make New Friends, but Keep the Old: Recommending People on Social Networking Sites*. ACM.

Chen, X. L. H. 2013. "Recommendation as Link Prediction in Bipartite Graphs: A Graph Kernel-Based Machine

Learning Approach," *Decision Support Systems* (54:2), p. 10.

Christidis, K., and Mentzas, G. 2013. "A Topic-Based Recommender System for Electronic Marketplace Platforms," *Expert Systems With Applications* (40:11), pp. 4370-4379.

Dahlhaus, R. 1996. "On the Kullback-Leibler Information Divergence of Locally Stationary Processes," *Stochastic Processes and their Applications* (62:1), pp. 139-168.

Deng, Z.-H., Wang, Z.-H., and Zhang, J. 2013. "Robin: A Novel Personal Recommendation Model Based on Information Propagation," *Expert Systems With Applications* (40:13), pp. 5306-5313.

Dudley-Nicholson, J. 2013. "Australians Now Using Social Media in Bedrooms and Toilet Cubicles," http://www.news.com.au).

Gavalas, D., and Kenteris, M. 2011. "A Web-Based Pervasive Recommendation System for Mobile Tourist Guides," *Personal and Ubiquitous Computing* (15:7), pp. 759-770.

Guy, I., Ronen, I., and Wilcox, E. 2009a. *Do You Know?: Recommending People to Invite into Your Social Network*. New York, New York, USA: ACM.

Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., and Ofek-Koifman, S. 2009b. *Personalized Recommendation of Social Software Items Based on Social Relations*. New York, New York, USA: ACM.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. "The Weka Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter* (11:1), pp. 10-18.

Jeh, G., and Widom, J. 2002. *Simrank: A Measure of Structural-Context Similarity*. ACM.

Jensen, C., Davis, J., and Farnham, S. 2002. *Finding Others Online: Reputation Systems for Social Online Spaces*. New York, New York, USA: ACM.

Khurri, A., and Luukkainen, S. 2009. "Identification of Preconditions for an Emerging Mobile Lbs Market," *Journal of Location Based Services* (3:3), pp. 188-209.

Kullback, S., and Leibler, R. A. 1951. "On Information and Sufficiency," *The Annals of Mathematical Statistics*).

Kuo, T.-T., Yan, R., Huang, Y.-Y., Kung, P.-H., and Lin, S.-D. 2013. *Unsupervised Link Prediction Using Aggregative Statistics on Heterogeneous Social Networks*. ACM.

Liben Nowell, D., and Kleinberg, J. 2007. "The Link‐Prediction Problem for Social Networks," *Journal of the American Society for Information Science and Technology* (58:7), pp. 1019-1031.

Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V. 2010. *New Perspectives and Methods in Link Prediction*. New York, New York, USA: ACM.

Mayer, J. M., Motahari, S., Schuler, R. P., and Jones, Q. 2010. "Common Attributes in an Unusual Context: Predicting the Desirability of a Social Match," *Proceedings of the fourth ACM conference on Recommender systems*), pp. 337-340.

Menon, A. M., Deshpande, A. D., Perri III, M., and Zinkhan, G. M. 2003. "Trust in Online Prescription Drug Information among Internet Users," *Health Marketing Quarterly* (20:1), pp. 17-35.

Newman, M. 2001. "Clustering and Preferential Attachment in Growing Networks," *Physical Review E* (64:2), p. 025102.

O' Madadhain, J., Hutchins, J., and Smyth, P. 2005. "Prediction and Ranking Algorithms for Event-Based Network Data," *ACM SIGKDD Explorations Newsletter* (7:2), pp. 23-30.

Park, D. H., Kim, H. K., Choi, I. Y., and Kim, J. K. 2012. "A Literature Review and Classification of Recommender Systems Research," *Expert Systems With Applications*).

Quercia, D., and Capra, L. 2009. "Friendsensing: Recommending Friends Using Mobile Phones," *the third ACM conference*), pp. 273-276.

Ren, H. Y. Y. Q. R. Y. M. 2014. "Human Mobility Discovering and Movement Intention Detection with Gps Trajectories," *Decision Support Systems* (63), p. 12.

Salton, G., and Michael, J. 1983. "Introduction to Modern Information Retrieval,").

Sankaradass, V., and Arputharaj, K. 2011. "An Intelligent Recommendation System for Web User Personalization with Fuzzy Temporal Association Rules," *European Journal of Scientific Research*), pp. 1-9.

Scellato, S., Noulas, A., and Mascolo, C. 2011. *Exploiting Place Features in Link Prediction on Location-Based Social Networks*. New York, New York, USA: ACM.

Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. 2010. *Folks in Folksonomies: Social Link Prediction from Shared Metadata*. ACM.

Shi, Z. W., Andrew B. 2013. "Network Structure and Observational Learning: Evidence from a Location-Based Social Network," *Journal of Management Information Systems* (30:2), p. 27.

Terveen, L., and McDonald, D. W. 2005. "Social Matching: A Framework and Research Agenda," *ACM Transactions on Computer-Human Interaction (TOCHI)* (12:3), pp. 401-434.

Tian, Y., He, Q., Zhao, Q., Liu, X., and Lee, W.-c. 2010. "Boosting Social Network Connectivity with Link Revival," *CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*), pp. 589-598

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region," *Economic geography* (46), p. 234.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. 2011. *Human Mobility, Social Ties, and Link Prediction*. ACM.

Xu, B., Chin, A., Wang, H., and Zhang, L. 2011. "Social Linking and Physical Proximity in a Mobile Location-Based Service,", pp. 99-108.

Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. *"Addressing cold-start problem in recommendation systems," In Proceedings of the 2nd international conference on Ubiquitous information management and communication (ICUIMC '08).* ACM, pp. 208-211.

Zheleva, E., Getoor, L., Golbeck, J., and Kuter, U. 2010. "Using Friendship Ties and Family Circles for Link Prediction." Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 97-113.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., and Ma, W.-Y. 2011. "Recommending Friends and Locations Based on Individual Location History," *ACM Transactions on the Web* (5:1), pp. 1-44.

Friend Recommendations Based on User-generated Contents

1. Introduction

Social networking sites (SNSs) are Internet sites where people can interact freely, sharing and discussing information about each other and their own lives, and which use multimedia such as personal words, pictures, videos, audios, and context information (for example, location). The development of SNSs has taken longer than 10 years, from the earlier versions of SNSs, such as Myspace and Friendster, to today's SNSs, such as Facebook and Twitter. The number of users is increasing rapidly--to more than a billion today. SNSs have now become an integral part of people's daily lives, profoundly impacting individuals, organizations, and society as a whole. Social network users try to stay connected with acquaintances and find new friends. More than half of adult users use social networks at the office, and almost a third of young adults use them in the bathroom.

From time to time, these SNSs have collected a tremendous volume of user-generated contents (UGCs). All these contents reflect different aspects of users' lifestyles and patterns. The rapid development of smart mobile devices and wearable devices has enabled even more context information, such as location information and health information, to be collected. An Australian survey counted 34% of social network users logged on at work, 13% at school, 18% in the car, while 44% used social networks in bed, 7% in the bathroom, and 6% in the toilet (Dudley-

Nicholson 2013). SNSs have collected a lot of data on people over a long period of time, and the contents are comprehensive and complete.

These UGCs serve as a gold mine that is yet to be tapped for various business and consumer intelligence applications. Many researchers and business analytics professionals have been attracted to UGCs and have focused on exploring a variety of ways to use UGCs. Academic researchers have tried to discover users' behavior patterns, trends, and activities, and then integrate this information into existing social behavior theories. Business analytics professionals have tried to increase sales through using personalized promotions based on these UGCs and engaging in customer relations management by addressing issues that arose for users from different social network channels (Woolridge 2011).

Unfortunately, very little research has used UGCs to make friend recommendations. Friend recommendation systems are one of the most essential parts of social network sites. These systems try to recommend people based on shared similar interests and backgrounds, thus helping SNSs avoid the cold-start problem, increase network speed, and boost the quality of users' activities. The existing friend recommendation systems use simple profile matching or friend network matching to recommend friends, but according to the theory of computer-supported social matching process (Terveen and McDonald 2005), there are many more attributes that could be used in this process.

There are six different kinds of user attributes in Terveen and McDonald (2005) theory. They are: demographics, social ties, geo-temporal, interests, personality, and

needs. In the first essay of my study, I used location information (geo-temporal) to build a friend recommendation system. From the results of experiment, we found that our recommendation system improved the overall performance compared to other state–of-the-art systems. In this essay, we take the next step by proposing a novel text analytics framework. In this study, we extract users' writing styles and document readability, sentiment scores in different locations, and auto-recognized personality scores by processing user-generated texts. With the help of these attributes, our friend recommendation system further improves the accuracy and precision of recommendations. This framework provides the first example of applying personality and interest attributes to friend recommendation systems based on text mining.

The rest of this essay is organized as follows. In the next section, we discuss related work, including existing text analysis methods, research on readability, sentiment scores, and auto-recognized personality. The third section describes our model: the process of text analysis, attribute generation, and record pair-wising. In section 4, we discuss the experiment and results. The last section discusses the implications for academics and business, the limitations of this work, and future research.

2. Related Work

2.1 Friend Recommendation Systems

Based on Adomavicius and Tuzhilin (2005 research (2005), the existing recommendation systems could be divided into two categories based on the

recommended objects. The first and the most common systems try to recommend products, such as movies, songs, books, articles, and blogs. They are quite useful in e-commerce websites, like Amazon.com, but in SNSs, most of them are just part of users' weblogs. The second category of recommendation systems recommends friends. These systems are essential in SNSs since they recommend homogeneous users within the same networks in order to help users discover potential friends or old acquaintances.

The item/product recommendation systems have been very well developed. A large amount of research has focused on how to make recommendations based on reviews, customized tags, number of "likes" or "stars," and friends' suggestions. Relatively, friend recommendation has not been highlighted in prior research, even though it is very useful to both users and businesses in social network sites (Tian et al. 2010b).

For users, a better friend recommendation system can help users avoid the cold start problem, increase network speed, and boost social network activities. For example, with the help of friend recommendations in movie social networks, users can quickly find potential friends and discuss their common interests on scientific fiction movies, such as "Star Wars," or love stories, such as "Gone with the Wind." And based on their common interests, users can have more fun and be amused by their friends. The higher the similarity between friends, the more likely they are to take the time to enjoy the friendship.

For businesses, highly active interaction among users provides natural and valuable channels for the propagation of information and trends, which can be transformed into greater market potentials. Hence, it is desirable for service providers to help users strengthen their social connectivity and thus increase service market value.

To help provide better quality of friend recommendations, Terveen and McDonald (2005) proposed a computer-supported social matching process model.

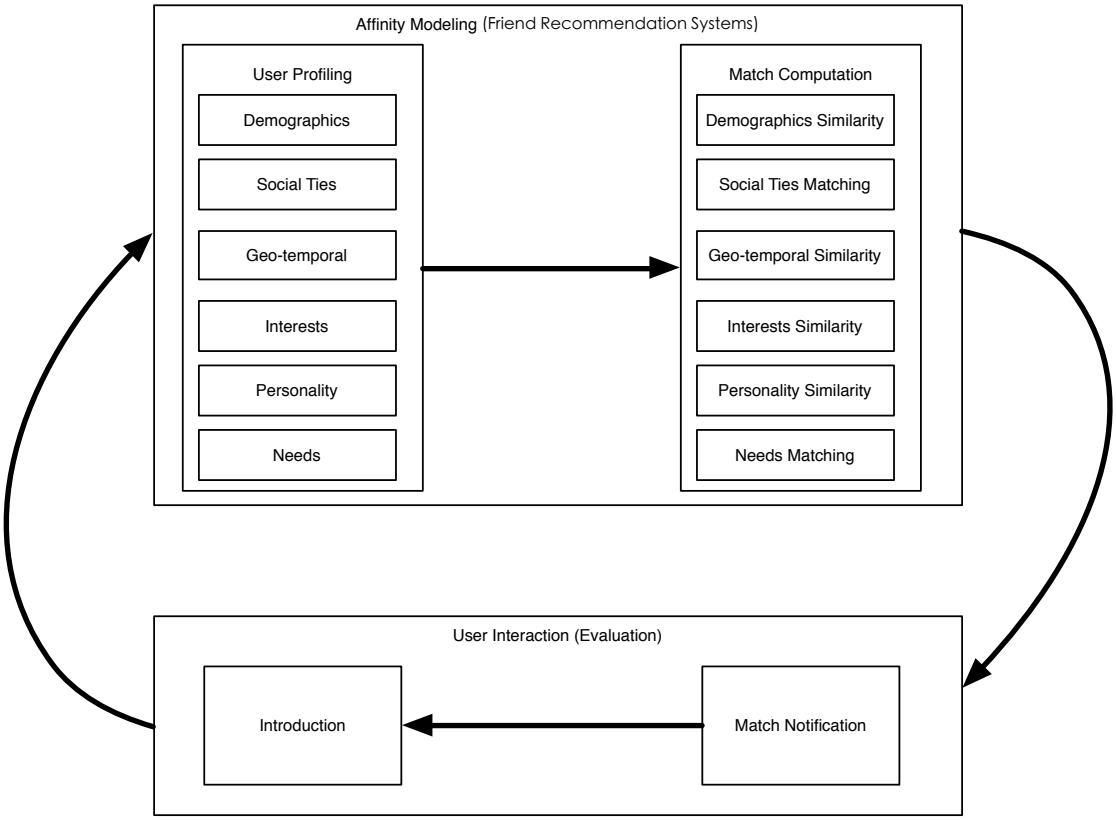


Figure 2-1 Computer-supported Social Matching Process

This model consists of four steps: modeling, matching, introducing, and interacting. (Mayer et al. 2010) more clearly represented these steps by splitting the process

into two parts: affinity modeling and user interaction. Affinity modeling is the process of gathering data from users to build profiles that enable the system to compute social matches. User interaction includes the interactions between the system and the user that is necessary to collect data, send a match notification, and facilitate the introduction and interaction between matched users.

Social matching systems calculate user affinities by weighting the similarities between users over a set of user attributes. According to Mayer et al. (2010), there are different types of user attributes:

- Demographics (geographical background, educational background, etc.)
- Social Ties (friends, co-worker, relatives, etc.)
- Interests (hobbies, favorites, music, books, etc.)
- Geo-temporal Patterns (frequently visited places, mobility traces, proximity patterns, etc.)
- Needs (partner, help, knowledge, etc.)
- Personality (extraversion, neuroticism, agreeableness, conscientiousness, openness, etc.)

The leading social network sites, such as Facebook and LinkedIn, use the common attributes. Their friend recommendation systems provide a list of people you may know, based on analyzing users' profiles and existing friend networks. So, based on Computer-supported Social Matching Process, they use demographic attributes and social ties as predictors to make friend recommendations. Still, a lot of gaps and potential exist. In my first essay, I provided a novel model for using geo-temporal

patterns as attributes in friend recommendation systems. In this essay, I further extract from UGCs personality and interest attributes as text features.

2.2 Text Features in User-generated Contents

In machine learning and pattern recognition, a feature is an individual measurable heuristic value of a phenomenon being observed that describes one aspect of an item. In our situation, a text feature will be one aspect of a user's interests or personality. Content analysis of text has long been an interesting research area in sociology and business. Researchers have discovered meta-information from different documents that range from shallow to insightful. A list of feature variables of text has been proposed in this literature. In this study, we broadly divide them into the following major feature types:

The shallow meta-information, which can easily be seen directly from the documents, may also have a strong impact on describing users' personality. Features include:

- *Document Length*: These features are simply the measures of the document text, such as number of words, number of sentences, and number of lines/paragraphs. Use of small numbers of words or small numbers of sentences could imply these people are straightforward and like to use imperative sentences or mandatory sentences. In contrast, people who like to use many words could be attentive and tender people.

- *Writing Style*: Very similar to document length features, these features measure the average syllables per word, average words per sentence, and percent of complex words. They describe the writing styles of the user. A more complex writing styles may imply a person who has a higher educational background, likes to read complicated books, or is of an older age. Using simple words or short sentences could suggest a person who is younger and of a less complex nature.
- *Readability*: There are several indexes or scores to measure the readability of a document. For example, the Fog Score, developed by Gunning (1952), is well-known and has a simple formula for calculation. The index specifies the number of years of formal education a reader of average intelligence would need to understand a text on the first reading, for example, scores such as 18 for unreadable, 14 for difficult, and 8 for childish. The Flesch-Kincaid grade level score rates text based on the U.S. grade school level. A score of 8.0 means that the document can be understood by an 8^{th} grader. A score of 7.0 to 8.9 is considered to be optimal.

We use text-mining techniques to extract meta-features from whole documents. In this category, we have the following features:

- *Sentiment*: By using natural language processing, text analysis, and computational linguistics techniques, we could recognize the polarity of opinion in text. From well-established general polarity cues in an existing word list, each word in a text can sometimes be annotated for its polarity

strength in a range. With sentiment analysis, we could find documents that have positive or negative ideas in different contexts, for example, in different locations. Using sentiment features, we could identify people who like or dislike a certain type of location, which helps to identify users' interests.

- *Subjectivity*: Similar to sentiment features, subjectivity is also a kind of opinion-mining technique. By using text-mining algorithms, we could automatically rate the text as more subjective or more objective.
- *Personality*: Past literature has shown that psycho-linguistic attributes, frequency-based analysis at lexical level, emotive words and other lexical clues such as number of first person or second person words could help in automatic personality detection. In this study, we use the Big Five, a widely exploited scheme for Personality Recognition from Text. It shows consistency across age and gender, and its validity remains the same when using different tests and languages. The features in Big Five are:
  - Openness to experience (tendency for non-conventional, abstract, symbolic thinking vs. preference for non-ambiguous, familiar, and non-complex things)
  - Conscientiousness (tendency for long-term planning vs. impulsive and spontaneous behavior)
  - Extraversion (tendency for active participation in the world around vs. concentration on one's own feelings)
  - Agreeableness (tendency for eagerness to cooperate and help vs. self-interest)

- Neuroticism (tendency to experience negative feelings and being overemotional vs. emotional stability and calmness)

By using these text features, we could extract users' interests and personalities, and from the social matching process, we could extract interest attributes and personality attributes, all of which will help to discover the degree of user matching. We propose a text analytic framework, which digs into UGCs, and extracts these attributes to make friend recommendations. The following section describes how this framework works.

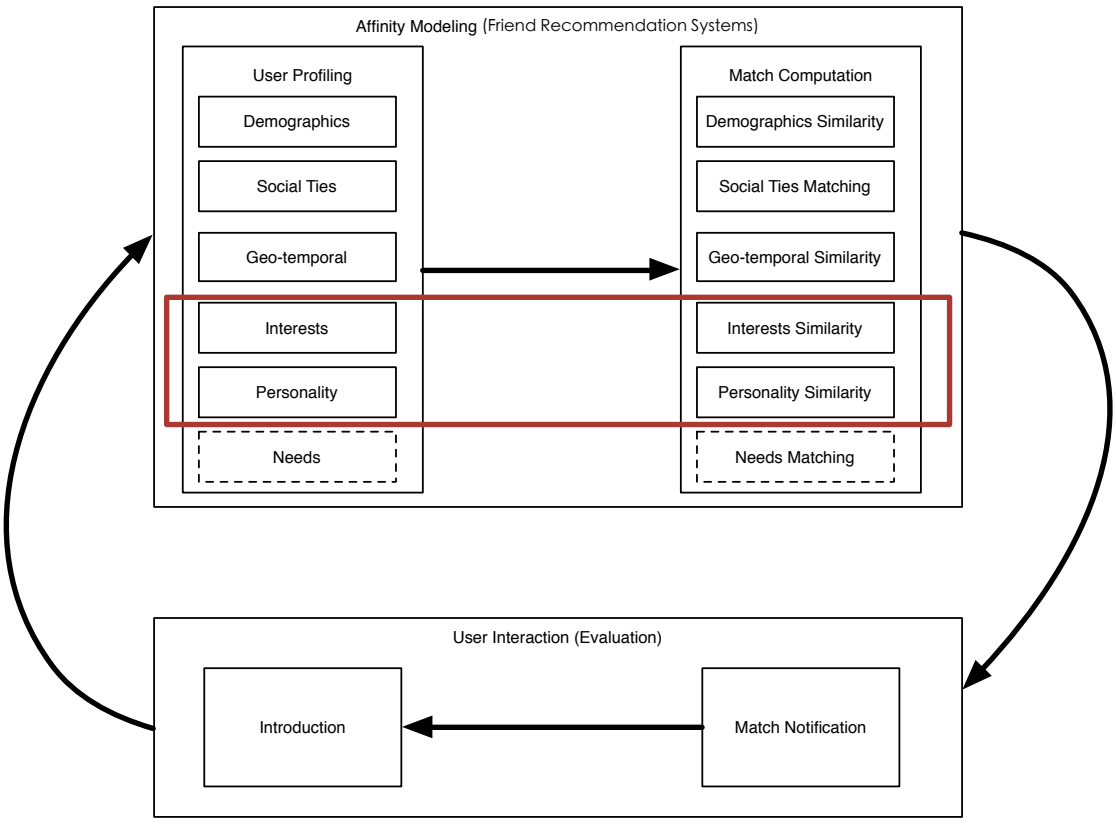


Figure 2-2 Computer-supported Social Matching Process with Text Features

3. Model

Based on the Computer-supported Social Matching Process Theory, the process of our text analytic framework is:

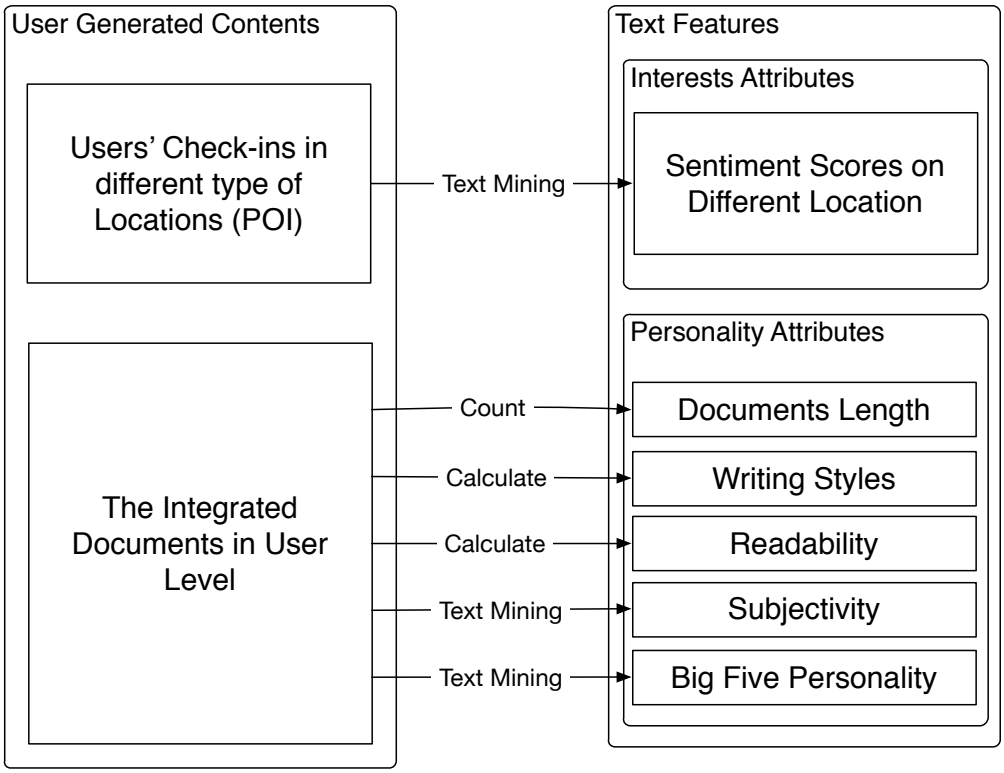


Figure 3-1 The Text Analytic Framework

1) Separate UGCs into different categories based on users' check-in location type and also integrate the UGCs into one document for analysis.

   Many UGCs in social networks, such as Twitter and Foursquare, have a limitation in the number of characters because of the fact that short messages are propagated more readily. That presents some difficulties in text analysis. In this study, our solution is to integrate more than one pieces of text into a document.

Based on the location type, we integrate texts into different categories. For example, the most popular location-based service, Foursquare, has nine major point of interest (POI) types: Art, College, Food, Professional, Nightlife, Outdoors, Shop, Travel, and Residence. To analyze these nine different documents, we could extract users' interests in different locations. Then, for users' personality extraction, we also need to combine all comments from a certain user into one document. One document for one user will be eligible for analysis with enough number of words.

2) Count the document's length features in the integrated document.

   Counting document length is quite straightforward. By splitting documents according to stop marks (periods) and blank spaces, we can get the number of words and number of sentences.

3) Calculate the writing style features in the integrated document.

   Niels Ott's research study (Ott and Meurers 2011) provided a Perl package for calculating the number of syllables. The calculation is not entirely accurate but has about 90% accuracy. From the Java Fathom Java package, we can calculate three features: average number of syllables per word, average number of words per sentence, and percentage of complex words (i.e., words of three or more syllables).

4) Calculate the readability scores for the integrated document.

   The readability scores are defined as the grade level at which readers need to read and understand the document. Much research has discussed methods on how to calculate readability scores, such as Automated Readability Index

(ARI) (Senter and Smith 1967), Coleman-Liau Index (Coleman and Liau 1975), Flesch-Kincaid Readability Test (Kincaid et al. 1975), and Gunning Fog Index (Gunning 1952).

The formula for Flesch-Kincaid Test is:

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6(\frac{total\ syllables}{total\ words})$$

The formula for Gunning Fog Index (Gunning 1952)is:

$$0.4[\left(\frac{words}{sentences}\right) + 100\left(\frac{complex\ words}{wirds}\right)]$$

In this study, we use the Gunning Fog Index and Flesch-Kincaid Test to score the text features for readability.

5) We use the text-mining package Opinion Finder from the University of Pittsburgh to analyze the text document and extract the subjectivity scores. Subjectivity could be used as an explanation for what influences and informs people's judgments about truth and reality. Opinion Finder uses a rule-based subjectivity classifier, which relies on manually crafted rules to tag sentences in a document as subjective or objective with high precision and low recall. We then calculate the percentage of subjective sentences with sentiment scores in the range of 0.0 to 1.0, with 1.0 meaning all subjective and 0.0 meaning all objective.

6) We use auto-recognized personality techniques to calculate the Big Five Personality scores.

Poria et al. (2013) proposed a new architecture for recognizing personality scores by using common sense knowledge with associated sentiment polarity

and affective labels. They designed five SMO (Sequential minimal optimization)-based supervised classifiers for the Big Five personality traits (John and Naumann 2008). The evaluation results in this study yielded a precision score of around 0.6-0.7. We follow their algorithm by using LIWC (Linguistic Inquiry and Word Count) and MRC Psycholinguistic Database, and combine them with the common sense knowledge-based features extracted by septic computing techniques. Finally, we get the Big Five personality scores.

7) For each type of location, we calculate the sentiment scores of the documents by using sentiment analysis techniques.

   Sentiment is the attitude, opinion, or feeling toward a certain object, such as a person, organization, product, or location. By using text-mining techniques and natural language processing, we could get the polarity of the text, such as positive, negative, or neutral. We use AlchemyAPI in this study to analyze the overall document to determine if it is generally more positive or more negative in certain types of locations.

   After the analysis of user-generated text, we have attributes of users' interests and attributes of personality. Then, the text features are put into our recommendation model. The process is very similar to the one in my first essay. Figure 3-2 shows the model.

Our recommendation system has the following process:

1) The system analyzes all users' demographic attributes, social-tie attributes, location attributes, and then combines them with the attributes extracted from the above framework, which are interest attributes and personality attributes.
2) The system compares a user's attributes with all other users' attributes, generates the similarities between two users, and then generates pairwise records.

   We use Jaccard coefficient (Salton and Michael 1983) in this study, which means the distance between two users is:

   $$d(a,b) = |\frac{a - b + \delta}{(a + b) + \delta}|$$

3) We employ data mining techniques to classify our records into two categories: Friend or Not Friend. We want to use probability of the classification results as the output.
4) The system sorts the outputs and then selects the top M users as the recommendation list for the user.

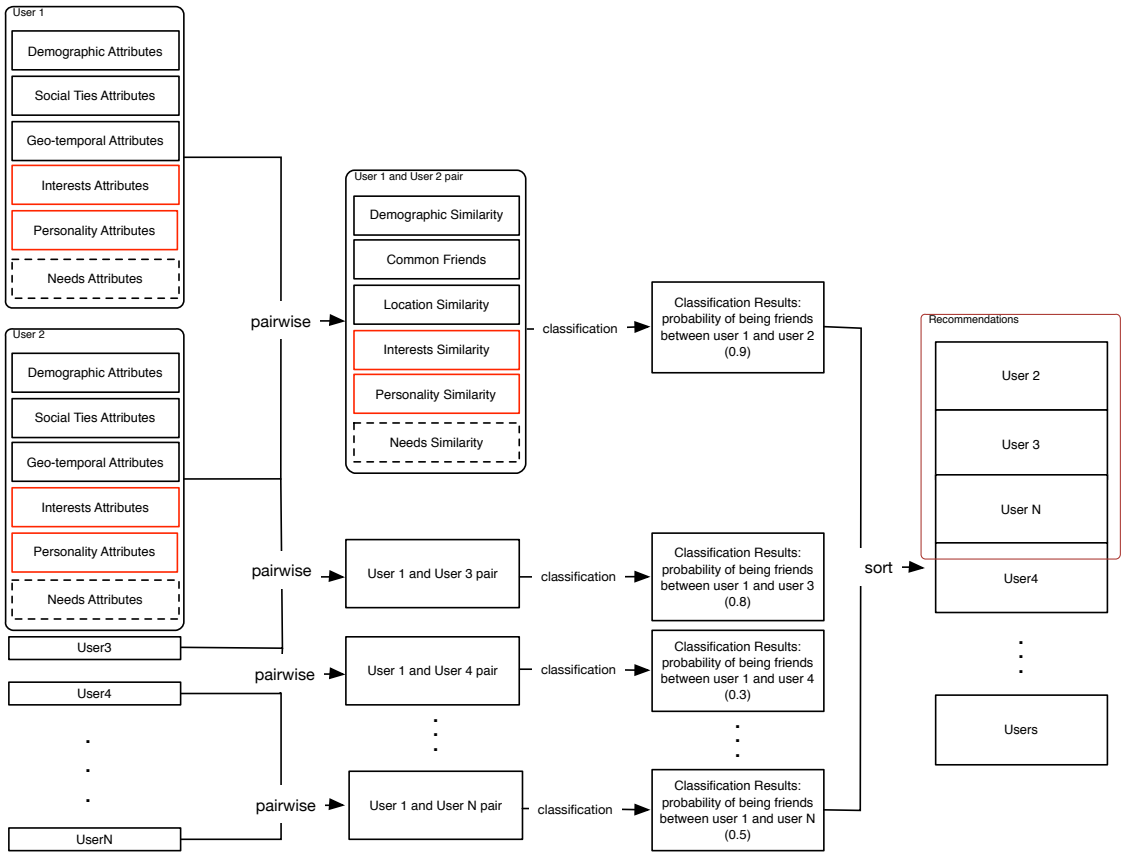


Figure 3-2 Recommendation Model

4. Experiment

4.1 Data Collection

Most online social networks have opened their platforms and enabled programmers to develop applications for them. The advantages of providing these APIs helped us visit their data. By applying as a developer for Foursquare, Twitter, and Facebook websites, I programmed and extracted our test data from those three social networks.

I started to collect the experiment data in October 2011. Based on the Foursquare open APIs, I wrote a Java program to scan the entire public timeline and then randomly found users who checked in. Due to authorization and privacy limitations, I sent a friend request to them first. After their acceptance, I was able to collect all the check-in information they posted. From October 2011 to February 2013, I collected 998 users and 6,417 check-in records. The friendship network was also recorded. There were 4,074 pairs of friends. The connectivity and density of the network is very low, averaging around four to five friends to one user. From Foursquare, I also built the POI dataset, which contains 420 different subtypes under nine major types.

Using this model, I collected data not only from Foursquare but also from Facebook and Twitter. From the Foursquare platform, I found that 754 of the users I collected had Facebook and Twitter accounts. I used the account id from Foursquare and then connected to the Facebook and Twitter websites. From the Facebook website, I got the demographic information of users, such as their religion, political orientation, age, educational background, work background, language, and favorite sports. When some of the data was not on the website, I manually visited each user's page and grabbed as much data as I could. From the Twitter website, I got a large amount of user-generated text information. Combined with the comments of check-ins from Foursquare, the text data set was large, including 17,890 pieces of text. So, on average, one user could have 18 pieces of text for creating a document.

Table 4-1 shows the attributes we collected from the social network websites.

Demographic Attribute		
ID	String	Unique identifier of the records
Gender	Type	User's gender: Male: 617, and Female: 353
Age	Integer	User's age, the range of age is from 18 years old to 64, the mean value is 30.055, and stand deviation is 5.543
Religion	Type	User's religion. There are 30 different religions. Major ones are Islam: 23 and Catholicism: 42.
Political	Type	User's political view. There are 24 different political views in the data set. Major ones are Liberal: 18 and Democracy: 24.
Highest education	Type	Describe user's education background. High School: 88, College: 241, and Graduate School:56
Work position	Type	Describe user's work position now. There are 38 different types of work.
Favorite sports	Type	Describe user's favorite sports. There are 63 different types.
Language	Type	Describe what language the user speaks. There are 51 different types.
Tip count	Integer	How many tips does the user have?
Tip-likes count	Integer	How many likes have this user's tips got?
Geo-temporal Attributes		
Check-in count	Integer	How many check ins does the user have? Range: 0 to 173, mean: 6.477
Home city	Type	The home city of user.
Art and entertainment	Integer	How many check-ins are in art and entertainment locations? Range: 0 to 83, mean: 1.082.

check-ins		
College check-ins	Integer	How many check-ins are in college locations? Range: 0 to 24, mean: 0.412.
Food check-ins	Integer	How many check-ins are in food locations? Range: 0 to 49, mean: 2.504.
Professional check-ins	Integer	How many check-ins are in professional locations? Range: 0 to 33, the mean value is 0.698.
Nightlife check-ins	Integer	How many check-ins are in nightlife locations? Range: 0 to 43, mean: 0.71.
Outdoors check-ins	Integer	How many check-ins are in outdoors locations? Range: 0 to 56, mean: 0.862.
Shop check-ins	Integer	How many check-ins are in shop locations? Range: 0 to 41, mean: 1.575.
Travel check-ins	Integer	How many check-ins are in travel locations? Range: 0 to 15, mean: 0.689.
Residence check-ins	Integer	How many check-ins are in residence locations? Range: 0 to 6, mean: 0.403.
Area	Double	The check-in areas in the physical geographic longitude and latitude.
Interest Attribute		
Art and entertainment sentiment	Double	What are the sentiment scores for the documents on art and entertainment locations?
College sentiment	Double	What are the sentiment scores for the documents on college locations?
Food sentiment	Double	What are the sentiment scores for the documents on food locations?
Professional sentiment	Double	What are the sentiment scores for the documents on professional locations?

Nightlife sentiment	Double	What are the sentiment scores for the documents on nightlife locations?
Outdoors sentiment	Double	What are the sentiment scores for the documents on outdoors locations?
Shop sentiment	Double	What are the sentiment scores for the documents on shop locations?
Travel sentiment	Double	What are the sentiment scores for the documents on travel locations?
Residence sentiment	Double	What are the sentiment scores for the documents on residence locations?
Personality Attributes		
Number of words	Integer	Number of words in the user's entire document.
Number of sentences	Integer	Number of sentences in the user's entire document.
Words / sentences	Double	Average number of words per sentences in the user's entire document.
Syllables / words	Double	Average number of syllables per words in the user's entire document.
Percentage of complex words	Double	The percentage of complex words in the user's entire document.
Fog score	Double	The fog index of readability for the user's entire document.
Kincaid score	Double	The score of Flesch-Kincaid Readability Test for the user's entire document.
Subjectivity	Double	The subjectivity score for the user's entire document.
Openness	Double	The personality openness score for the user's entire document.
Conscientiousness	Double	The personality conscientiousness score for the user's entire document.

Extraversion	Double	The personality extraversion score for the user's entire document.
Agreeableness	Double	The personality agreeableness score for the user's entire document.
Neuroticism	Double	The personality neuroticism score for the user's entire document.

Table 4-1 Attributes Collected from Social Network Websites

We then calculated the similarity values between two users. For two integer values, we used the Jaccard coefficient (Salton and McGill 1983). The formula for calculating the relative difference is:

$d(a,b) = |\frac{a-b+\delta}{(a+b)+\delta}|$, where δ is a small smoothing factor and was set to 0.001 in our evaluation

Demographic Attribute		
Gender type	Type	Describe two users' genre type: Male and Female: 49.05%, Two Males: 38.32% and Two Females: 12.63%
Age difference	Integer	User's age difference., Range: 0 to 46, mean: 4.33.
Religion difference	Boolean	Do two users have different religious views: False: 22.2%, True:0.2%, Unknown: 77.6%.
Political difference	Boolean	Do two users have difference political views: False: 16.5%, True:0.1%, Unknown: 83.4%.
Share same education	Boolean	Do two users have the same highest education: False: 76.73%, True: 23.27%.
Share same work type	Boolean	Do two users have the same work type: False: 51.98%, True: 0.4%, Unknown: 47.6%.

Share sport counts	Integer	How many sports do both users like? Range: 0 to 3, mean: 0.561.
Share language counts	Integer	How many languages do both users speak? Range: 1 to 3, mean: 1.758.
Tip count difference	Double	The relative difference between two users' tip counts.
Tip-likes count difference	Double	The relative difference between two users' total tip-likes counts.
Social-tie Attributes		
Friend count difference	Double	The relative difference between two users' friends counts.
Common friends	Integer	How many friends two users share in the social networks? We have two attributes, one is only in the data set we have, and the other is in the Foursquare platform.
Geo-temporal Attributes		
Check-in count difference	Double	The relative difference between two users' total check-in counts.
Home city distance	Double	The physical distance of two users' home city. Range: 0 to 1.9k km.
Arts and Entertainment check-ins share	Integer	How many check-ins have the two users made in Arts and Entertainment locations in history? Range: 0 to 40, mean: 0.172.
College check-ins share	Integer	How many check-ins have the two users made in college locations in history? Range: 0 to 11, mean: 0.049.
Food check-ins share	Integer	How many check-ins have the two users made in food locations in history? Range: 0 to 34, mean: 0.845.

Professional check-ins share	Integer	How many check-ins have the two users made in professional locations in history? Range: 0 to 15, mean: 0.148.
Nightlife check-ins share	Integer	How many check-ins have the two users made in nightlife locations in history? Range: 0 to 19, mean: 0.092.
Outdoors check-ins share	Integer	How many check-ins have the two users made in outdoors locations in history? Range: 0 to 33, mean: 0.136.
Shop check-ins share	Integer	How many check-ins have the two users made in shop locations in history? Range: 0 to 23, mean: 0.469.
Travel check-ins share	Integer	How many check-ins have the two users made in travel locations in history? Range: 0 to 13, mean: 0.148.
Residence check-ins share	Integer	How many check-ins have the two users made in residence locations in history? Range: 0 to 6, mean: 0.088.
Area overlap	Double	The check-in area overlaps between two users in the physical geographic longitude and latitude.
Check-in history distribution similarity	Double	The K-L divergence between two users' check-in history distributions. We have two attributes, because K-L divergence is not symmetric.
Interest Attribute		
Arts and Entertainment sentiment difference	Double	What's the relative difference of sentiment scores between two users' documents on Arts and Entertainment locations?
College sentiment difference	Double	What's the relative difference of sentiment scores between two users' documents on college locations?
Food sentiment	Double	What's the relative difference of sentiment scores between two users'

difference		documents on food locations?
Professional sentiment difference	Double	What's the relative difference of sentiment scores between two users' documents on professional locations?
Nightlife sentiment difference	Double	What's the relative difference of sentiment scores between two users' documents on nightlife locations?
Outdoors sentiment difference	Double	What's the relative difference of sentiment scores between two users' documents on outdoors locations?
Shop sentiment difference	Double	What's the relative difference of sentiment scores between two users' documents on shop locations?
Travel sentiment difference	Double	What's the relative difference of sentiment scores between two users' documents on travel locations?
Residence sentiment difference	Double	What's the relative difference of sentiment scores between two users' documents on residence locations?
Personality Attributes		
Number of words difference	Double	Relative difference in number of words for two users' documents.
Number of sentences difference	Double	Relative difference in number of sentences for two users' documents.
Words / sentences difference	Double	Relative difference in average number of words per sentences between two users' documents.
Syllables / words difference	Double	Relative difference in average number of syllables per words between two users' documents.
Percentage of	Double	Relative difference in the percentage of complex words between two users'

complex words difference		documents.
Fog score difference	Double	Relative difference in the fog readability index between two users' documents.
Kincaid score difference	Double	Relative difference in the Kincaid readability index between two users' documents.
Subjectivity difference	Double	Relative difference in subjectivity scores for two users' documents.
Openness difference	Double	Relative difference in personality openness scores for two users' documents.
Conscientiousness difference	Double	Relative difference in personality conscientiousness scores for two users' documents.
Extraversion difference	Double	Relative difference in personality extraversion scores for two users' documents.
Agreeableness difference	Double	Relative difference in personality agreeableness scores for two users' documents.
Neuroticism difference	Double	Relative difference in personality neuroticism scores for two users' documents.
Dependent Variable		
is friend	Boolean	Do two users connect – i.e., are they friends – in the social network?

Table 4-2 Similarity Calculation for Two Users

4.2 Experiment Design

Our recommendation question then changes to a classification question that can be addressed by employing data mining techniques. The next step is to simulate the real world social network density. Due to the limitations of collecting users' data, the friendship network in our data set is relatively sparse. To simulate the real world density of friendship network, we try to select links in our test data set. By controlling the proportion of friend/non-friend links, we have social networks with different densities. Figure 4-1 shows a simple example of different densities of five users' networks. And in our situation, we set the network proportion of friend: not friend as 1:1, 1:2, 1:5, and 1:10.

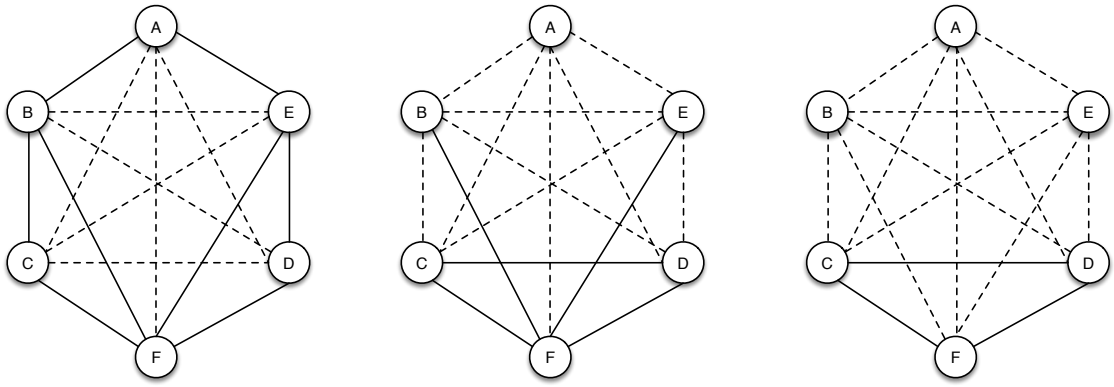


Figure 4-1 Examples of the Proportion of Friend: Not Friend

We also need to consider the attribute sets to compare with state-of–the-art recommendation systems, which only use demographic attributes and social-tie attributes. In this experiment, we also compare the text attributes with essay 1's location- based friend recommendation system. To make the comparison, we design

eight different test groups, which include the settings in essay 1 as shown in Table 4-3:

Group	Attribute Data set
Group 1	Demographic Attributes Only
Group 1_text	Demographic Attributes + Interests Attributes + Personality Attributes
Group 2	Demographic Attributes + Geo-temporal Attributes
Group 2_text	Demographic Attributes + Geo-temporal Attributes + Interests Attributes + Personality Attributes
Group 3	Demographic Attributes + Social-Tie Attributes
Group 3_text	Demographic Attributes + Social-Tie Attributes + Interests Attributes + Personality Attributes
Group 4	Demographic Attributes + Social-Tie Attributes + Geo-temporal Attributes
Group 4_text	Demographic Attributes + Social-Tie Attributes + Geo-temporal Attributes + Interests Attributes + Personality Attributes

Table 4-3 Test Attributes Groups

Groups 1 to 4 are the groups we used in essay 1. Groups 1 and 3 are basic profile matching and social-tie matching in the current friend recommendation systems. And Group 2 and Group 4 have location information added to them. In this essay, we propose several text features that include interest attributes and personality attributes. We add text attributes to each of previous groups to make the evaluation.

The classifiers we use for generating the recommendations are also important. In our model, we need to know the probability of classification in the output. That means we need our classifier to be probabilistic. We used Bayesian Network, Naive Bayes, and Logistic Regression.

4.3 Results

Weka is a popular toolkit for machine learning written in Java and developed by the University of Walkaton, New Zealand. We used it as the experiment platform for our study. We used the default settings in Weka and the accuracy as the result of evaluation. We had different outputs in Weka, such as precision, ROC, recall, and confusion matrix. Only accuracy is discussed here. The definition of accuracy is:

$$accuracy = \frac{number\ of\ true\ positive + number\ of\ true\ negative}{number\ of\ records\ in\ test\ data\ set}$$

The baseline accuracy, like random guess, will be considered in the biased data set. We have the friend networks from 1:1 to 1:10, which means the positive and negative proportions are also 1:1 to 1:10. So, in a 1:1 network, the random guess accuracy will be 50%, and in a 1:2 network, the classifiers will lean towards giving a negative output, so the random guess accuracy rate is 2/3 = 66.6%. To alleviate the effect of classification bias, we also perform cost-sensitive tests. The settings of the cost matrix are shown in Table 4-4.

Proportion	Cost Matrix
1:1	$\begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}$
1:2	$\begin{vmatrix} 0 & 2 \\ 1 & 0 \end{vmatrix}$
1:5	$\begin{vmatrix} 0 & 5 \\ 1 & 0 \end{vmatrix}$
1:10	$\begin{vmatrix} 0 & 10 \\ 1 & 0 \end{vmatrix}$

Table 4-4 Settings of Cost Matrix

Table 4-5 and Figure 4-2 show the results of the accuracy test after the experiment.

	1：1	1：2	1：5	1：10
Baseline Accuracy	50%	66.7%	83.3%	90.9%
Group 1	52.4%	66.8%	83.3%	90.9%
Group 1_text	62.6%	75.3%	84.9%	91.9%
Group 2	71.8%	80.1%	87.8%	92.6%
Group 2_text	77.6%	83.4%	89.4%	92.9%
Group 3	78.4%	79.3%	87.5%	92.6%
Group 3_text	87.727%	89.658%	92.6362%	94.7383%
Group 4	86.3%	88.5%	92.2%	94.6%
Group 4_text	89.7%	91.1%	93.5%	94.8%

Table 4-5 Results of Accuracy Test

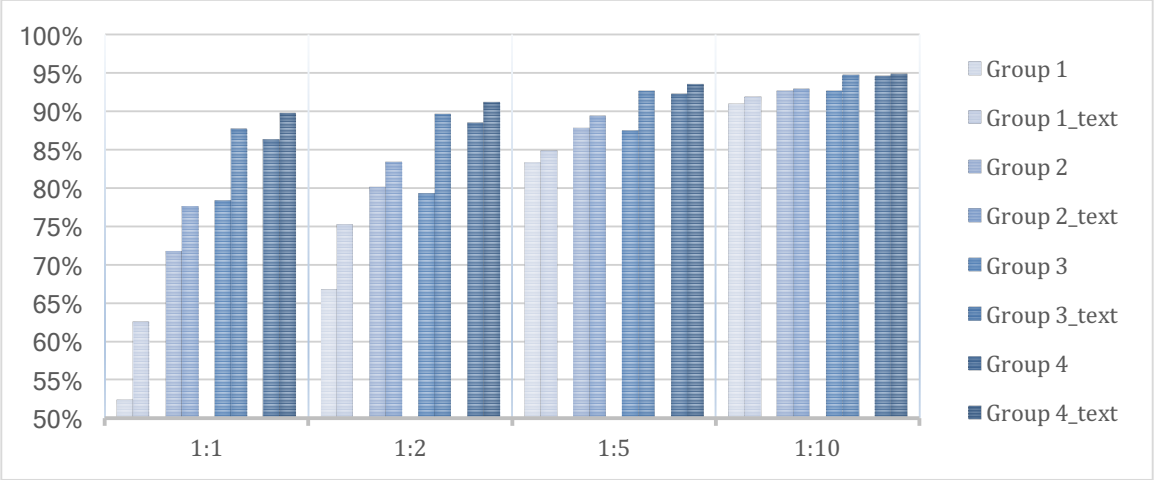


Figure 4-2 Results of Accuracy Test

From the accuracy test results, we can see that with only the demographic attributes, the accuracy is like a random guess. The reason may be that our demographic data is quite sparse. Most Facebook users did not complete their profile information as we had assumed, and some of the attributes may have been out of date. With the interest attributes and personality attributes extracted from text, users' information became much clearer and our accuracy results significantly improved. The Group 2 attribute sets used the location attributes, which, as stated in essay1, improved the

accuracy and, compared to Group 1_text, had better performance than text features. In Group 2_text, the interest attributes and personality attributes further improved the results. The social-tie attribute sets had performance similar to location attribute sets. Group 2 and Group 3 had similar accuracy. Interest attribute sets and personality attribute sets seemed to provide higher improvement than in Group 2. We believe the reason is that the interest attributes are extracted from documents based on location and have higher correlations with location attributes, which could weaken the improvement. In Group 4, with demographic attributes, social-tie network attributes, and location attributes, the text features (Group 4_text) only provide a slight improvement.

Table 4-6 and Figure 4-3 show the accuracy results when we also made an experiment in cost-sensitive classification.

	1：1	1：2	1：5	1：10
Baseline Accuracy	50%	66.7%	83.3%	90.9%
Group 1	52.4%	57.9938%	57.2738%	57.0536%
Group 1_text	62.6%	68.3358%	72.9488%	90.9046%
Group 2	71.8%	77.0723%	79.45%	91.36%
Group 2_text	77.6%	77.4341%	79.5469%	91.5019%
Group 3	78.4%	79.44%	80.49%	92.1%
Group 3_text	87.727%	89.658%	92.6362%	94.7383%
Group 4	86.3%	84.09%	91.99%	92.3%
Group 4_text	89.7%	90.9835%	92.6708%	93.8463%

Table 4-6 Results of Cost-Sensitive Accuracy Test

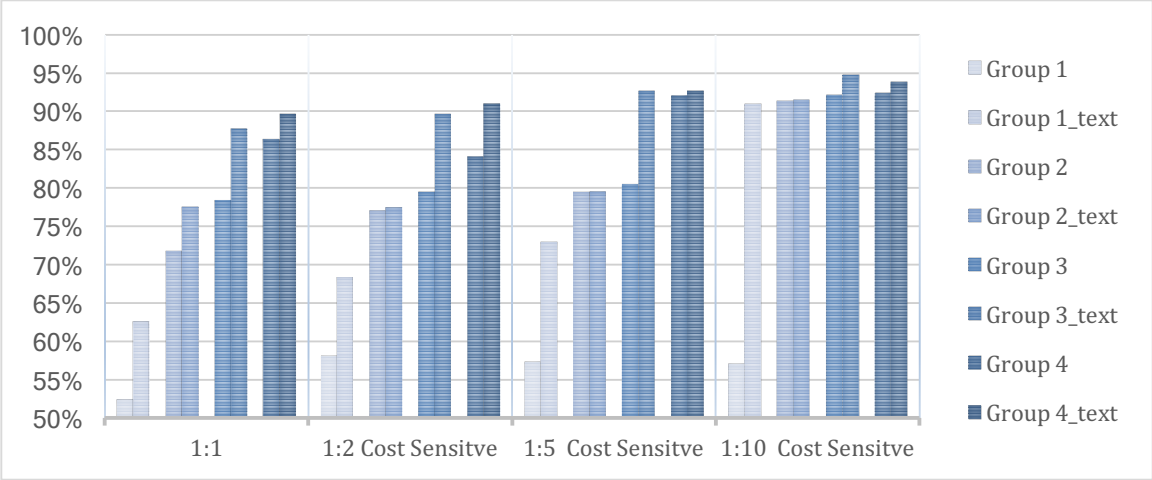


Figure 4-3 Results of Cost-Sensitive Accuracy Test

From the results we can see the same trends as the cost insensitive situations. And according to the cost matrix, the classifiers like to classify a negative result as positive result, which causes the accuracy to decrease. But in a precision test, we can see that the cost-sensitive results are better.

To further simulate the recommendation results, we try to use the top M precision. By using the classification probability results from the Weka output, we sort and recommend the top M users and calculate the correct rate for recommending a friend that is an actual friend in the dataset.

To clearly examine the relative positions for different groups of classifiers, we need to calculate the baseline precisions and the optimal line positions. In realistic scenarios, the calculation for the baseline and optimal precisions only depends on the connectivity of social networks and the number of friend recommendations. But, in our data set, the scarcity of friend links forced us to consider each user's friend links. The calculations are:

Assume we have n users, and in a 1: P proportion network, for each user I, we have friend link number F_i, non-friend link number N_i, and we recommend M friends in the list.

Baseline Precision:

For each user, if the total number of links $F_i + N_i$ is less than the number of recommendations M, then all friend links would be in the recommendation list, so the precision is F_i/M. Otherwise, the number of possible ways to select M links is C_{F+N}^{M}. The number of possible ways to select x friend links and M-x non-friend links is $C_F^x \times C_N^{M-x}$. The expected precision of random top M recommendation for this user is:

$$Baseline\ Precision_i = BP_i = \frac{\sum_{j=0}^{M} j \cdot C_{F_i}^{j} \cdot C_{N_i}^{M-j}}{C_{F_i+N_i}^{M} \cdot M}$$

And the average baseline precision for the data set is:

$$Base\ Precision = (\sum_{1}^{n} BP_i) \div n$$

Optimal Precision:

For each user, the selected friend link number will be: min (F_i, M), so for Top M Recommendation, the optimal precision is:

$$Optimal\ Precision = (\sum_{1}^{n} \frac{\min(F_i, M)}{M}) \div n$$

Table 4-7 shows the optimal precisions for the Top3, Top5 and Top 10 Recommendations.

	User Numbers	Optimal Precision in Top 3 Recommendation	Optimal Precision in Top 5 Recommendation	Optimal Precision in Top 10 Recommendation
1:1 Data Set	835	56.846307%	42.562874%	24.395210%
1:2 Data Set	891	53.273475%	39.887767%	22.861953%
1:5 Data Set	936	50.712251%	37.970085%	21.762821%
1:10 Data Set	957	49.599443%	37.136886%	21.285266%

Table 4-7 Optimal Precisions in Top 3, 5, 10 Recommendations

And Table 4-8 shows the baseline precisions.

	User Numbers	Baseline Precision in Top 3 Recommendation	Baseline Precision in Top 5 Recommendation	Baseline Precision in Top 10 Recommendation
1:1 Data Set	835	26.613807%	27.989746%	22.297519%
1:2 Data Set	891	14.776371%	17.759222%	18.170754%
1:5 Data Set	936	5.350665%	6.971847%	9.244579%
1:10 Data Set	957	2.190556%	2.901769%	4.220156%

Table 4-8 Baseline Precisions in Top 3,5,10 Recommendations

When we have the baseline and the optimal precisions, we can also calculate the relative positions of our recommendation precisions. The formula for the relative positions is:

$$Position = \frac{Recommendation\ Precision - Baseline\ Precision}{Optimal\ Precision - Baseline\ Precision}$$

Then we normalized all the results for top 3 recommendations and put them in the same chart as shown in Figure 4-4, Figure 4-5 and Table 4-9.

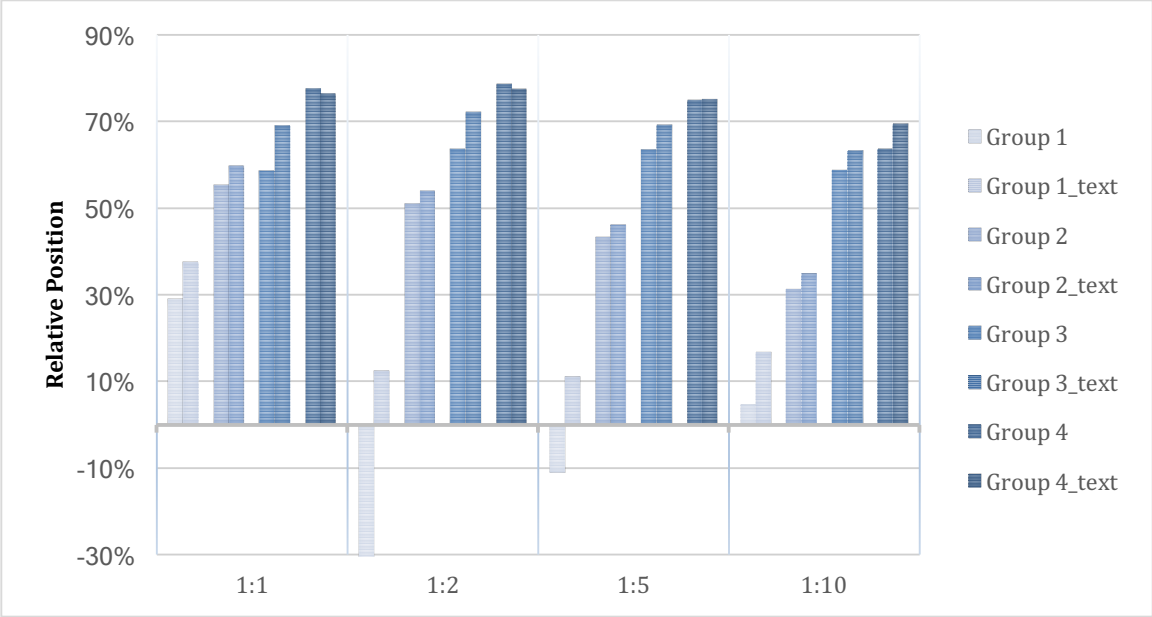


Figure 4-4 Top 3 Friend Recommendation Precisions

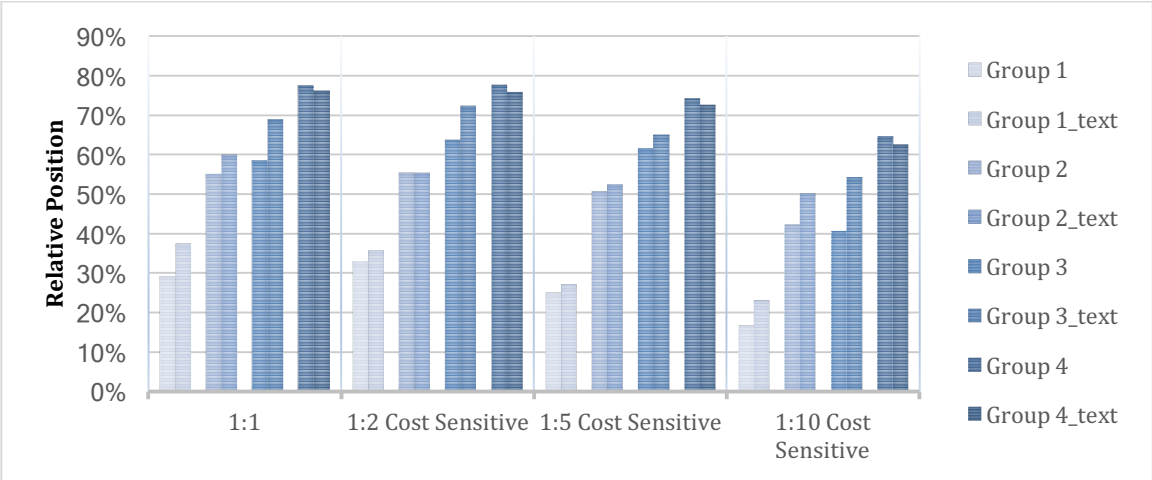


Figure 4-5 Top 3 Cost-Sensitive Friend Recommendation Precisions

	1:1	1:2	1:2 Cost Sensitive	1:5	1: 5 Cost Sensitive	1:10	1:10 Cost Sensitive
Group 1	28.96%	-33.04%	32.95%	-11.17%	25.02%	-4.62%	16.61%
Group 1_text	37.48%	12.44%	35.86%	11.18%	27.14%	16.77%	23.24%
Group 2	55.11%	50.83%	55.49%	43.16%	50.78%	31.23%	42.33%
Group 2_text	59.73%	53.94%	55.30%	46.14%	52.28%	34.91%	50.08%

Group 3	58.59%	63.65%	63.74%	63.5%	61.58%	58.79%	40.68%
Group 3_text	68.84%	72.01%	72.30%	68.99%	65.06%	63.19%	54.30%
Group 4	77.42%	78.52%	77.65%	74.80%	74.17%	63.56%	64.73%
Group 4_text	76.23%	77.26%	75.90%	74.96%	72.60%	69.29%	62.50%

Table 4-9 Relative Positions of Top 3 Friend Recommendations

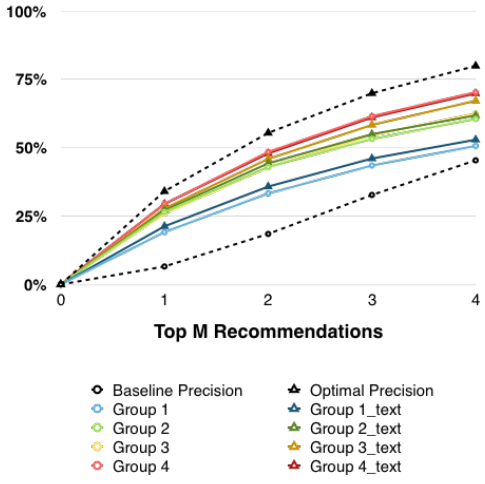
From Figure 4-4 and 4-5, we found that for the top 3 friend recommendation precisions, as discussed before, the text features influence the precision. The trends were quite similar as seen in the accuracy results. All text attribute groups had superior performance over the non-text feature groups except Group 4_text. The Group 4 and Group 4_text precisions are very similar. In the cost non-sensitive groups, the demographic attribute groups showed poor performance, which was lower than the baseline precision. But when we used the cost-sensitive matrix to alleviate the bias, the performance got much better. The demographic groups basically had a 10%-20% better performance. And in the best case, which was Group 4, the recommendation precision relative position went to 70% of optimal precision. Another trend we observed was when the proportion goes up, the relative position goes down. The reason could be that when data sets get larger, as in top 3 friends recommendation, it is harder to reach the optimal line. We would see the trends when we manipulate the top M friends recommendations in following analysis.

To make the evaluation more complete, we also generate performance charts for precision based on the number of recommendations, M. The x-axis of the performance chart is the number of links we recommended, and the y-axis is the ratio of the number of true friend links to the number of friend links in the recommendation list (M). Because the friend links are different for each user, we report the average value here.

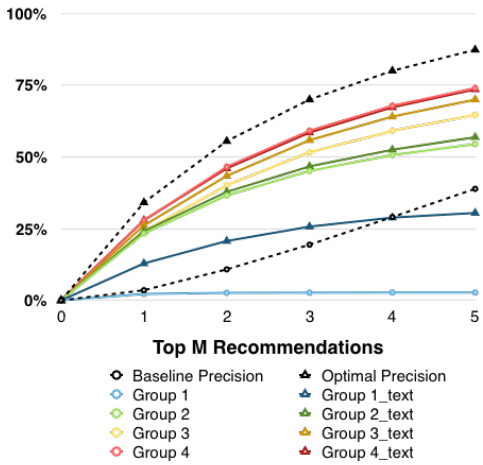
The maximum value of x-axis is related to the total links we have in the test data set. The highest value could be the maximum links for a user and could exceed hundreds. So, to get an applicable maximum number, we selected the average friend links and added a bit more. For example, in 1:1 data set, we had 2,037 friend links, 2,037 non-friend links, and 835 users, so the average friend links would be (2,037 + 2,037) / 835 ≈ 5.

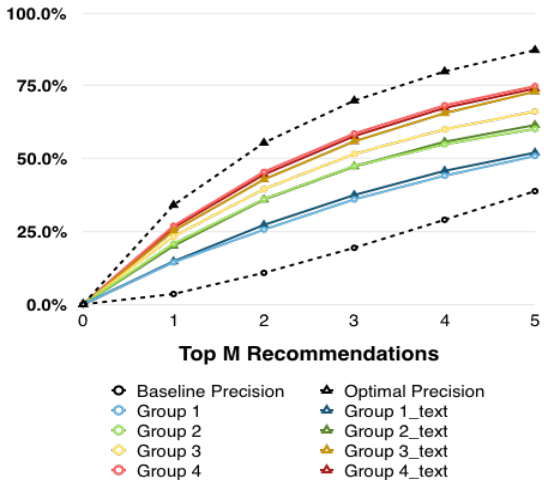
Because we are not going to the maximum number in x-axis, we will not reach the 100% value in y-axis. And the maximum precision our recommendation will have depends on the accuracy of the classification, which means that it cannot reach 100% and becomes flat after some value of x.

Figure 4-6 shows the performance charts for different proportions with/without cost-sensitive classification:

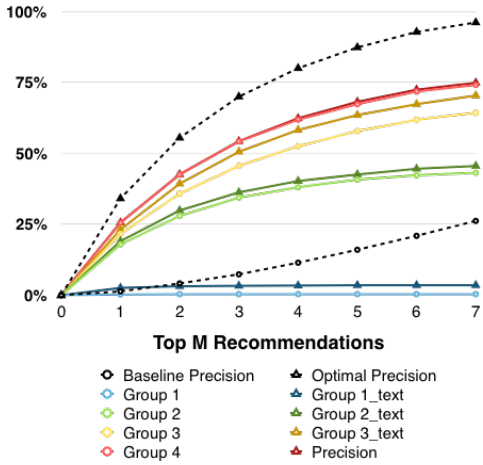
	Not Cost Sensitive	*Cost Sensitive*
1:1		

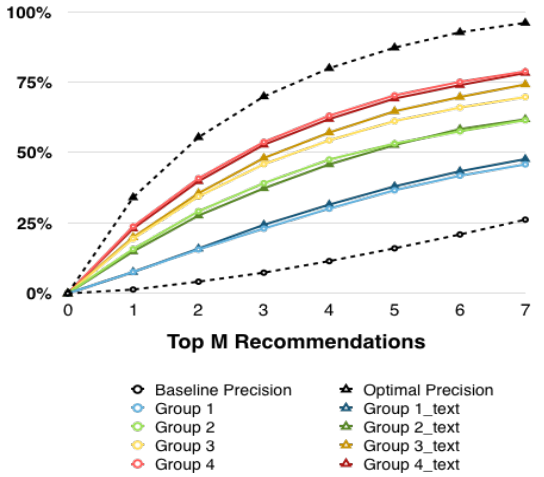
1:2


100%
75%
50%
25%
0%
0
1
2
3
4
5
Top M Recommendations
Baseline Precision
Group 1
Group 2
Group 3
Group 4
Optimal Precision
Group 1_text
Group 2_text
Group 3_text
Group 4_text


100.0%
75.0%
50.0%
25.0%
0.0%
0
1
2
3
4
5
Top M Recommendations
Baseline Precision
Group 1
Group 2
Group 3
Group 4
Optimal Precision
Group 1_text
Group 2_text
Group 3_text
Group 4_text

1:5


100%
75%
50%
25%
0%
0
1
2
3
4
5
6
7
Top M Recommendations
Baseline Precision
Group 1
Group 2
Group 3
Group 4
Optimal Precision
Group 1_text
Group 2_text
Group 3_text
Precision


100%
75%
50%
25%
0%
0
1
2
3
4
5
6
7
Top M Recommendations
Baseline Precision
Group 1
Group 2
Group 3
Group 4
Optimal Precision
Group 1_text
Group 2_text
Group 3_text
Group 4_text

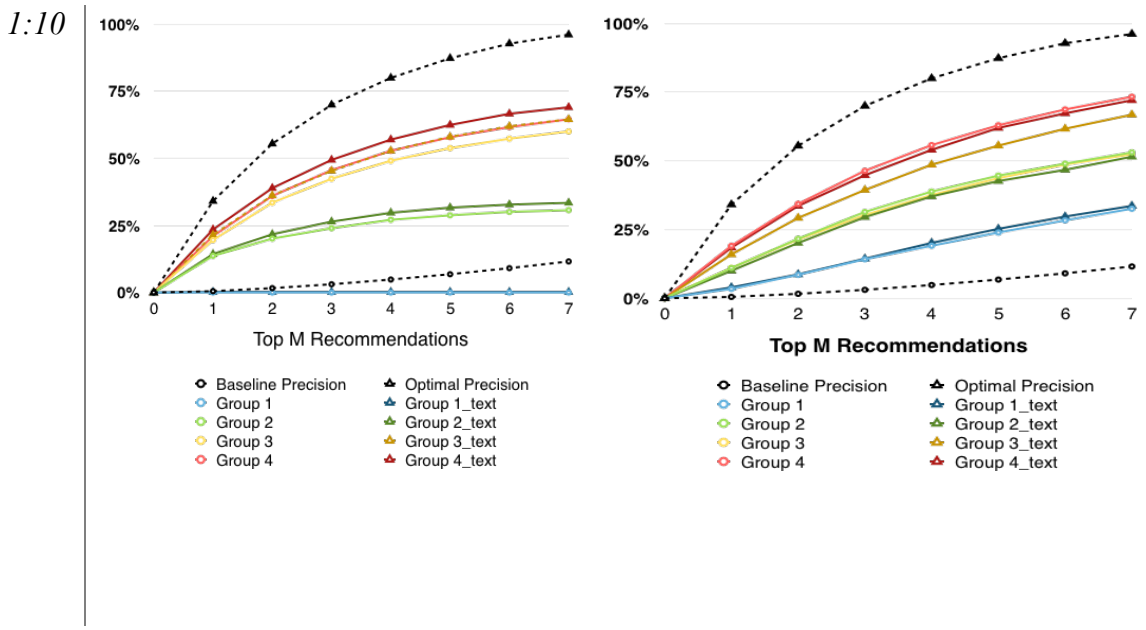


Figure 4-6 Performance Charts for Friend Recommendations

From the performance charts, we can see the evaluation more clearly, and we can spot the trends when the recommendation numbers are changed. For a high connectivity SNS such as our 1:1 data set, if we recommended more than three users in the list, the recommendation performance is hardly better than baseline performance. And in a sparser SNS, we could recommend more users to reach the highest performance. For example, in 1:5 and 1:10 data sets, we could recommend around six to seven users.

5. Discussion

In this study, based on the computer-supported social matching process theory, we added interest attributes and personality attributes to a friend recommendation system by extracting text features from UGCs. Although a lot of studies have focused on UGCs, very few of them used UGCs to make friend recommendations. The text features from UGCs could be used to build an appropriate user-topic model and could explain users' interests and personality. By creating an elegant text analytic

framework, we analyzed users' text documents to extract their interest attributes and personality attributes. After that, we calculated the similarity between two users and made friend recommendations. The experimental results show that the interest attributes and personality attributes could significantly improve the recommendation performance, especially in a sparse network. The results have some implications for both research and practice.

First of all, to the best of my knowledge, this is the first paper to study a user-topic model in friend recommendation systems. The earlier studies tried to dig into UGCs and extract emotion to predict trends or for information propagation. The topic-user model they built was used for document recommendation or expert finding. But this study shows this model could also be used for friend recommendations. This study is a first attempt and we could further improve the algorithm with a more appropriate natural language processing method, a more matched lexicon, or a better feature set.

Secondly, this study gives an example of a computer-supported social matching process that shows the importance of interest attributes and personality attributes. By applying this process to a friend recommendation system, we can discover more interesting attributes to help find new friends. The current friend recommendation systems in social networks are too simple in that they only find people who already know each other, but they have difficulty in finding people with similar habits. The more comprehensive attribute sets will solve these problems and bring more active users into social networks.

Not only for academic area but in the real world, our recommendation system model and its implementation could be used for the major social network websites. Our text analytic framework could also help these websites make better use of UGCs. UGCs are very useful for improving sales, which rely on accurate and personalized marketing and promotion. Having more users with a higher density in friend networks could help social networks maintain high level of activity.

A business using our model could easily extend or modify the feature sets for recommendations. Our model is comprehensive but a business could always change it to adapt to a particular social network such as an expert finding network. In this instance, document length and readability could be very important, while interests in the location may not be useful. In another example, a travel social network site will prefer location-related features but may not use document length.

Our research has several limitations. The data we collected are from around the world, which makes it difficult when people use not only English but also other languages. Because of this, we had to select the text. This also means a sparser social network. Further research could constrain the data to only one state or one city, which may provide a more interpretive result. We also did not use a specific lexicon for Twitter, so the natural language processing result could be improved. Further research should discuss how a different dictionary and lexicon will affect the results of text analytics. The evaluation results are also based on existing social networks, which are more likely to recommend to users' friends already known in real life, while it is also important to recommend strangers. Our recommendation

system is more comprehensive and capable of recommending people who share similar life patterns and habits. It would be interesting to conduct a survey on the satisfaction of recommendations, which may reveal more about the usefulness of the model. Finally, our model used five of six attribute sets in the computer-supported social matching process. Future research could bring need attributes into the recommendation system, which may be very useful for question-answering social networks such as quora.com.

References

Adomavicius, G., and Tuzhilin, A. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *Knowledge and Data Engineering, IEEE Transactions on* (17:6), pp. 734-749.

Coleman, M., and Liau, T. L. 1975. "A Computer Readability Formula Designed for Machine Scoring," *Journal of Applied Psychology* (60), pp. 283-284.

Dudley-Nicholson, J. 2013. "Australians Now Using Social Media in Bedrooms and Toilet Cubicles," http://www.news.com.au.

Gunning, R. 1952. "The Technique of Clear Writing".

John, O., and Naumann, L. 2008. "Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues," *Handbook of personality: Theory and research* 3(2008), pp. 114-158.

Kincaid, J., Jr., F. R., RL, R., and BS, C. 1975. "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Research Branch Report*).

Mayer, J. M., Motahari, S., Schuler, R. P., and Jones, Q. 2010. "Common Attributes in an Unusual Context: Predicting the Desirability of a Social Match," *Proceedings of the fourth ACM conference on Recommender systems*), pp. 337-340.

Ott, N., and Meurers, D. 2011. "Information Retrieval for Education: Making Search Engines Language Aware," *Themes in Science and Technology Education* (3), pp. 9-30.

Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., and Howard, N. 2013. "Common Sense Knowledge Based Personality Recognition from Text," *Context based Expert Finding in Online Communities using Social Network Analysis* (8266:Chapter 42), pp. 484-496.

Salton, G., and Michael, J. 1983. "Introduction to Modern Information Retrieval,".

Senter, R. J., and Smith, E. A. 1967. "Automated Readability Index,".

Terveen, L., and McDonald, D. W. 2005. "Social Matching: A Framework and Research Agenda," *ACM Transactions on Computer-Human Interaction (TOCHI)* (12:3), pp. 401-434.

Tian, Y., He, Q., Zhao, Q., Liu, X., and Lee, W.-c. 2010. "Boosting Social Network Connectivity with Link Revival," *CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*), pp. 589-598

Woolridge, A. 2011. "Too Much Buzz: Social Media Provides Huge Opportunities, but Will Bring Huge Problems," *Economist*).

Friend Recommendations in Health/Fitness Social Networking Sites

1. Introduction

Over the past decade, smartphones have drastically changed many aspects of people's everyday lives. Thanks to innovative digital techniques, such as cloud computing technologies, machine learning, global positioning systems, and pervasive computing technologies, people are now able to connect to the Internet and track their activities/health indicators anytime and anywhere. From the financial industry to the entertainment industry, from social networking sites to the healthcare industry, the connectivity of smartphones is widely and deeply advancing the world, especially through social networking activities.

A social networking site (SNS) is defined as a platform to build social network connections between people who share similar interests, activities, stories, etc. Recently, in 2015, the major social networking site provider Facebook reached about 1.5 billion active users—up from more than 1 billion active users in 2013 (www.statista.com). Other platforms have shown similar increases. Twitter has one billion active users in the world and 48.2 million in the U.S. (Twitter.com 2016, Bennett 2014); LinkedIn has 433 million users in 2016 (Smith 2016); and Under Amour had 140 million users in 2015 (Pai 2015). The huge amount of content generated by these users has become a trusted source of user information and can

contribute to many areas, such as e-commerce, the travel industry, and especially, health/fitness activities (Anderson et al. 2011).

How social networking sites could influence healthcare has been well-researched recently (Alshaikh et al. 2014). Health campaigns are more and more based on "network interventions" (Valente 2012; Jiang, Zhu and Wang 2015). Peer and social networks have long been thought to be important influencers on behavior change during adolescence (Ennett and Baumann 1994), an argument that aligns with the assertion that social networks have important effects on health activities and health innovations across a lifetime (Smith 2008; McPherson et al. 2001; Christakis 2007). Internet users are eager to find information on health topics, including exercise (38% in 2008, up from 21% in 2002) and weight loss (33% in 2008) (Fox and Jones 2010). Daw et al. (2015) showed that degree of homophily across various relationship types and behaviors or interests contributes positively to health outcomes. However, to the best of our knowledge, no research has so far investigated how health activities could impact relationship-building on social networking sites.

To build connections between one user and other users on social networking sites, platform providers typically employ friend recommendation systems. Friend recommendation is one of the most fundamental tasks during the development of a social networking platform. Friend recommendation systems could help newly registered users to initialize their relationship networks and find people who share similar interests. Previous users also need to develop their existing networks to find

more friends. The existing algorithms for recommendation are simple static profile matching and link network matching (friend-of-friend). Both methods are well developed and broadly used in most social networking sites. However, according to the Terveen and McDonald's computer-supported social matching process (2005), there are more attributes that can be used in friend recommendation systems, for example, users' daily physical activities and health-related records.

Thanks to the rapid growth of smartphone and wearable device technologies such as personal area network technologies and sensor technologies, today, 60% of U.S. adults are able to regularly track their weight, diet, steps walked, and exercise routine, and more than half of them additionally track other health status indicators or symptoms such as blood pressure, blood sugar, headaches, and sleep patterns (www.pewinternet.org 2013). Compared to traditional methods for tracking this health-related data such as notes and spreadsheets, mobile applications could be much better tools (Darwish and Hassanien 2011; Jiang, Zhu, and Wang 2015).

The iOS platform started to provide Health App with the platform's major upgrade in 2014, and the Android platform released Google Fit at nearly the same time. Both applications could integrate users' health indicator information and fitness/sports data and show them on an easily understood dashboard. Beyond these two pre-installed applications, many third party applications provide more specific solutions with health/fitness-related social networking sites. For example, the third largest sportswear company, Under Armour, announced its fitness network application, Record, in 2013, and today, it has more than 30 million users.

On these kinds of social networking sites, users share their physical activities, exercise routines, and health indicators online with their online companions, and discuss their fitness and healthcare. Users seek the benefits of social support and peer pressures from their online friends. Social networks actively leverage principles of social support in novel ways and allow users to engage in fitness challenges with one another by sharing workout routines (Nakhasi et al. 2014). This shared information from them, as with other user-generated content, will become quite a valuable data resource to explain users' life patterns and interests (Alshaikh et al. 2014).

This study will bring users' health- and fitness-related features to the existing friend recommendation system framework and then try to create a more comprehensive method to help users find friends. The rest of this essay is organized as follows. In the next section, we introduce related work on health and fitness social networking sites, pervasive computing with tracked health data, and friend recommendation systems. The proposed model and details are discussed in section 3. We have performed an experiment based on our collected data from online health/fitness social networking sites, and the results are documented and discussed in section 4. And finally, we discuss the contributions and limitations of our work and identify a set of future research directions.

2. Related Work

2.1 Pervasive computing and health/fitness data

Pervasive computing, or ubiquitous computing, is a concept whereby processing is made to appear anywhere and anytime (Fritz et al. 2014). Based on this computation model, computer scientists invented many different digital devices that are involved in users' daily lives. Pervasive technologies use varied strategies for shaping people's behavior and activities. Most notably are those described by Fogg: *self-monitoring* and *conditioning* (Fogg 2003). Self-monitoring is one of the most prevalent pervasive technology strategies, although technologies often employ multiple strategies (Tollmar et al. 2012).

A variety of monitoring devices have been researched and evaluated for their pervasive influence on users' physical activities and behaviors. As Table 2-1 shows, smartphones are one of the most common devices that enable users' self-monitoring capability. Both Apple and Android platforms have motion coprocessors that support several functions such as collecting sensor data from integrated accelerometers, gyroscopes, and compasses and simulating users' activities like walking, running, swimming, etc. Another kind of user movement tracking component is a wearable device, which includes three main categories: activity trackers, such as Garmin and Misfit; smart bands, such as Jawbone and Microsoft Band; and smart watches, such as Apple Watch and Samsung Gear. There are also other clothing or accessories incorporating sensors and computers. These devices provide similar or even more functions than a smartphone. Thanks to these portable

personal technologies, wearable devices are able to track users' heart rates, sleep patterns, etc.

Devices		Communication Techniques	Sensor Techniques	Health Indicators	Fitness Indicators	Products
Smartphones		3G/4G Bluetooth	Motion Coprocessor Accelerometer Gyroscope Compass Camera	Heartrate Sleep Quality	Walking Running Climbing	iPhones Android Phones Windows Phones
Wearable Devices	Other Activity Trackers	Bluetooth	GPS Motion Coprocessor Pulse Sensor	Sit Position Heartrate Nutrition Intakes	Walking Running Climbing Swimming Workouts	Garmin Misfit
	Smart Bands	Bluetooth Zigbee	GPS Motion Coprocessor Pulse Sensor	Heartrate Sleep Quality	Walking Running Climbing Swimming Workouts	Microsoft Band Jawbone
	Smart Watches	Bluetooth	GPS Motion Coprocessor Pulse Sensor	Heartrate Sleep Quality	Walking Running Climbing Swimming Workouts	Apple Watch Samsung Gear

Table 2-1 Self-monitoring Devices Summary

Recently, using human computer interaction and other ubiquitous computing methods, smartphones and wearable devices attempt to persuade using various representations of sensed activity data. For example, UbiFit combined activity sensing with an understandable visualization of activity (Consolvo et al. 2008). In this paper, the authors found that the visualization helped participants maintain activity levels by providing positive feedback. Other systems attempt to persuade through coaching and advising metaphors. For example, there are several virtual coach apps such as Flowie, which contextually analyze users' activities and identify types of feedback that are most promising for motivation. Laura, a system with a similar goal, used an animated relational agent as an exercise advisor (Bickmore et al. 2005). Participants increased their walking by almost two times during the trial period.

Tracking gadgets are often narrow in the activity that can be sensed, leading to the need for integrating data from multiple sources to get a clearer view of health and fitness. Systems that require more effort on the part of users to track activities are less likely to be successfully adopted. For example, research by Ahtinen et al. (2009) and Jiang, Zhu and Wang (2015) showed that manual entering of health data was troublesome and led to falling use of wellness applications. These authors looked at the effects of health information "mashups" that integrate data from multiple sensors and sources and discovered that these combinations allowed people to gain deeper insights into their wellness. As we discussed before, Apple and Google both followed these research guidelines in developing their integrated health application platforms. Users can install any other third-party sports or

nutrition app for specific usage, and then, these applications will follow a well-designed interface to write their sensors' data into Apple's or Google's health app. After that, users will be allowed to monitor, consolidate, and share their health/fitness data on health/fitness social networking sites.

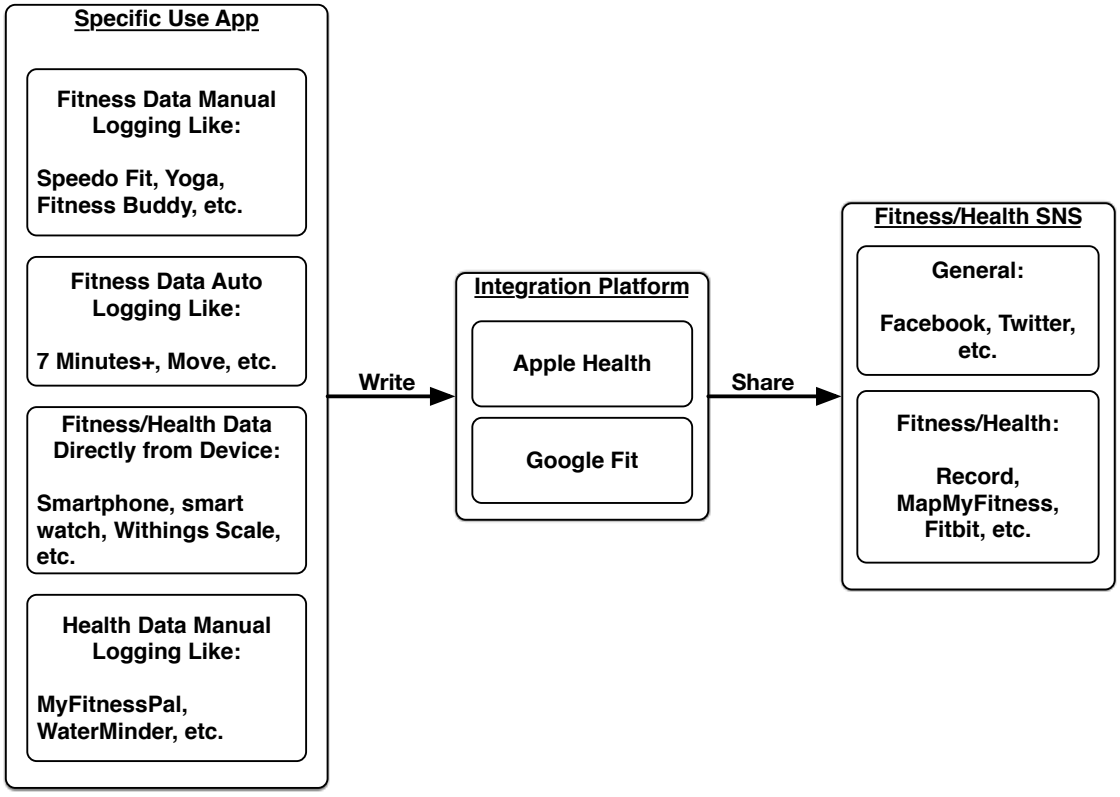


Figure 2-1 Mashups of Fitness/Health Data

Integrated health data such as counts of physical exercise, energy consumed for each activity, number of steps walked, length of time walked, user's heart rate, user's achievements, frequency of the user's workout, and user's sleep patterns, could not only affect the wellness outcomes but could also help to describe a user's lifestyle. Hirsch et al. (2014) pointed out that health and fitness data are powerful tools to investigate patterns of physical activity across large geographic and

temporal scales. We believe that by capturing and acknowledging everyday activities in an accessible and non-invasive manner and by facilitating the sharing and comparison of that information among peers, pervasive computing devices and health fitness apps could help to find more appropriate friends who share a similar physical activity level.

2.2 Health- and fitness-related social networking sites

Besides the data collection aspects of pervasive computing technologies, researchers have also considered their social aspects. Health-related social networking sites are starting to play an important role in people's daily lives by allowing them to monitor their food intake, fitness exercises, etc. In Balatsoukas et al. (2015)'s view, social support, peer pressure, and information sharing in online communities may affect health behaviors. If there are positive and sustained effects, then social networking technologies could increase the effectiveness and efficiency of many users' health and fitness routines. Peer-to-peer communication is an important feature for health and fitness applications. It enables users to discuss health matters and fitness routines with people who have similar conditions and then receive support and advice. The major health- and fitness-related social networking sites Record from Under Armour, and MapMyFitness are both very popular for this reason. Health and fitness application users benefit even more from social networking notifications from friends by interacting with other users online and, when agreed upon, even meeting face-to-face. Peer–to-peer communication

allows members with the same interests to support each other even if they do not live in the same city.

Health and fitness social networking sites can also help a person to stay motivated, which is an important element for medication compliance or sportive activities. In studies by McCullagh et al (1993) and Passer (1982), participants report social reasons for engaging in physical activity. These reasons include affiliation, being part of a team, and social status. A person might be encouraged to stay fit if his friend encourages him to do so while having discussions on a social networking site. This social support will enhance confidence and encourage users to persist. Workout records will also be shared and published on the timeline. Friends will try to exercise together if they see their friends work out every day. This pressure from friends will provide more motivation than from a coach. Also, there are several applications that provide challenge or compete features whereby users can select their friends to do a race. The system will trace and compare joined users' fitness records and give rewards or honors to the winners periodically. In summary, the use of social networking technologies can promote activities by allowing users to check the status of their friends or to plan daily exercise or weekly activities, such as a cycling tour (Smith et al. 2011).

Many studies have focused on the positive influence of social effects on health outcome, and some researchers indicate that link strength and user similarity are also positively related to fitness performance. People who share similar physical activities would be more likely to work out together or compete. Simpkins et al.

(2013) showed that friendships are an important component of people's health. Liza et al. (2013) showed that social network factors, such as online connections, physical proximity, network relationship roles, and exercise strength, will impact all pre-, during, and post-physical activity routines. To further improve health outcomes, social networking site providers let users find more friends who share greater similarities through friend recommendation systems.

2.3 Friend Recommendation Systems

Two types of recommendation systems exist in online social networking sites and these are based on what substances are recommended (Adomavicius et al. 2005). In an e-commerce site like Amazon.com, product recommendation systems that try to recommend, for example, movies, songs, and books, are extremely common. On social networking sites, recommendation systems will suggest articles, blogs, users' posts, etc. On the other hand, link recommendation or friend recommendation systems will try to recommend homogenous users to build friendships for people.

Item or product recommendation systems have been well studied. Much research and implementations have focused on how to make recommendations based on reviews, customized tags, number of "likes" or "dislikes", review stars, friends' comments, etc. Compared to item or product recommendation, friend recommendation has not been emphasized in recent research even though it is a very fundamental task in social networking sites (Tian et al. 2010a).

For the platform users, a more efficient friend recommendation system could help people overcome the so-called "cold-start" problem. This means that when a new user registers on a social networking site, without any links to other users, the user requires a long time to explore and find other users who share similar interests. Also, a friend recommendation system could provide a more convenient network building experience and encourage sharing activities among users. In some general social networking sites, for example, Facebook.com, with more friend-links, users could be motivated to share posts, pictures, and discussions if they received more friends' likes and comments. The greater similarity between these friend-links, the more users would tend to take the time to enjoy social networking activities.

Friend recommendation systems could help social networking sites from a business perspective as well. Social networking sites often feed a business' marketing strategy by letting users discover and share information from the company. To support the discovery and sharing of activities, platform providers need better connectivity and higher active interactions among users. Hence, social networking sites could benefit from a friend recommendation system that would attract more users to their sites. Larger numbers of users could greatly heighten the value of the platform provider.

To improve the quality of friend recommendation results, some researchers have studied the friend-of-friend algorithm (which Facebook uses) (Chen et al. 2009a), or the profile matching algorithm. Both of these methods have advantages and disadvantages.

The profile matching method is quite straightforward. The algorithm tries to collect users' demographic attributes in online social networking sites. For example, in Linkedin.com, researchers could collect users' age, gender, educational background, job position, skill sets, etc. The algorithm could calculate the similarity between two users' profiles and then make recommendations. However, there are some problems with the profile matching method. First, most of these attributes are not comprehensive and were preset by the platform provider. If Linkedin.com doesn't provide users' job positions, then one could not use or analyze it. Second, a new user may not have a completed profile, which means that several attributes of this user are empty and may never be filled. Profile matching also ignores the changing nature of the user. For example, old users may change their position and forget to update it in the social networking site, and this will affect the recommendation results. To summarize, profile matching has better performance on a highly connected and active social networking site.

The friend-of-friend or social tie matching algorithm tries to match two users' linking networks. In this method, two users with more inter-related friend-links have a greater chance to become friends. This method is very efficient for people who want to find all real-life friends on social networking platforms, but it presents difficulties in finding people who share similar interests but do not know each other.

Friend recommendation systems have been developed on several different types of social networking sites, especially on more general purpose platforms, but

interest-based social networking sites, for example, health and fitness social networking sites, need a more specific algorithm for their sites.

Terveen and McDonald (2005) have proposed the computer-supported social matching process model to provide a more in-depth view of how people build their social links. This model points out there are six different types of attributes that can be used to start a social matching process. These attribute categories are (Mayer et al. 2010):

- Demographics (geographical background, educational background, etc.)
- Social ties (friends, co-workers, relatives, etc.)
- Interests (hobbies, favorites, music, books, etc.)
- Geo-temporal patterns (frequently visited places, mobility traces, proximity patterns, etc.)
- Needs (partner, help, knowledge, etc.)
- Personality (extraversion, neuroticism, agreeableness, conscientiousness, openness, etc.)

Social matching systems try to calculate users' affinities by comparing the similarities between the above sets of attributes. We have investigated several attributes before, such as in profile matching when we used demographics attributes, and in the friend-of-friend system when we used users' social ties. We have built models in essays 1 and 2 to evaluate users' similarities by using geo-temporal data and interest/personality data. For health/fitness social networking sites, we could collect more related data, such as frequency of users' physical

activities, average time spent, energy consumed, etc. These daily activities actually reflect users' interests, life patterns, and needs. In this essay, we will build a recommendation model that embeds users' physical activities to further improve the friend recommendation system of social networking sites.

3. Model

To help health and fitness social networking site users find more friends with similar interests, we would like to create a health and fitness activity recommendation framework. As we discussed in Section 2.1, by using pervasive computing devices such as smartphones, smart bands, and smart watches, people's daily activities, health status indicators, and physical exercise indicators are computed, tracked, visualized, and recorded. According to the tracked data type, we could categorize health data into two groups. The first one is health indicators, which include average heart rate, average heart cadence rate, sleep hours, sleep patterns, weight, body mass index, etc. The second category is physical activity indicators. Based on the type of fitness exercise tracked, we could have one category that has distance-related records, such as speed and time, and the other category has only heart rate, energy consumed, etc.

Based on the computer-supported social matching process, we believe that people's health indicator data could become a dynamic source for demographic attributes, and people's physical activities data could become a source for interest attributes (Figure 3-1). Health indicator data is defined as "a characteristic of an individual, population, or environment which is subject to measurement and can be

used to describe one or more aspects of the health of an individual or population." Almost two-thirds of trackers monitor their health indicators every day and share this data online. According to some previous research, people are likely to find friends with similar body types. Overweight people have fewer friends, and normal weight people like to find friends with similar weights (de la Haye et al. 2011). From Simpkins et al. (2013)'s view, higher BMI people were more likely to have closer friendships, or conversely, less likely to have weaker, non-reciprocated friendships.

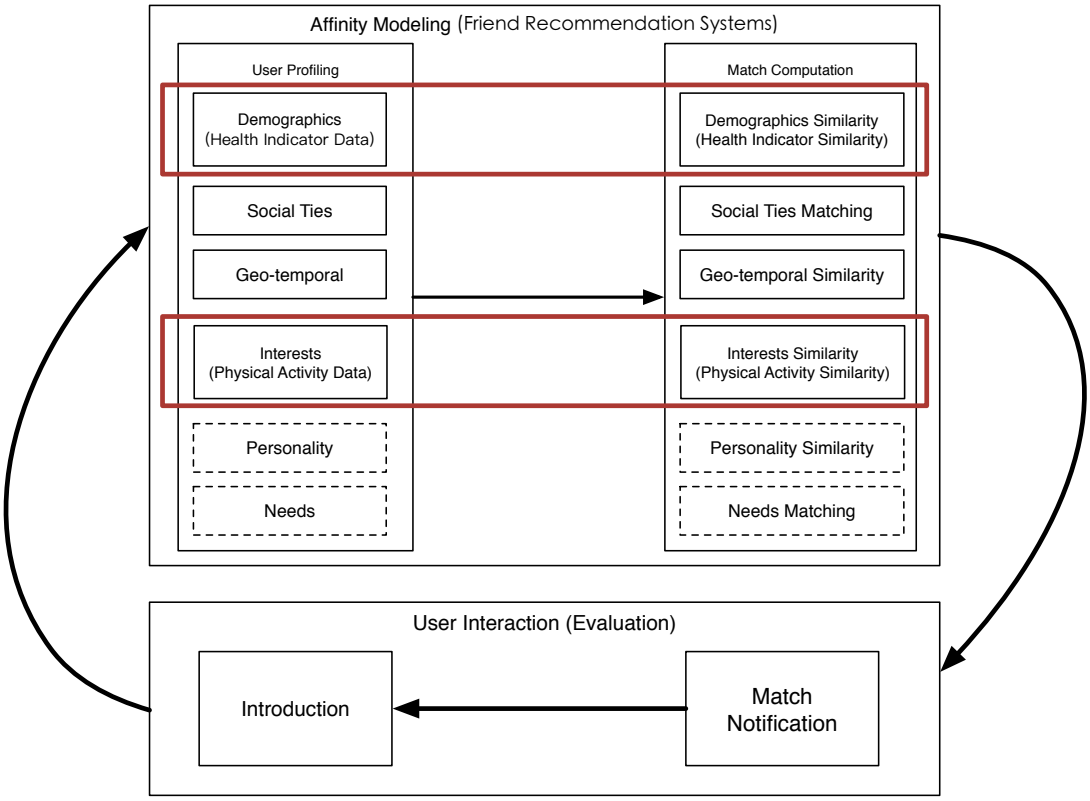


Figure 3-1 Computer-supported Social Matching Process with Health and Fitness Features

Physical activity indicator data is data about people's daily physical exercises and workouts, such as walking, running, swimming, and working out with machines. Regular physical activity has long been regarded as an important component of a healthy lifestyle. Recently, this impression has been reinforced by new scientific evidence linking regular physical activity to a wide array of physical and mental health benefits (Dishman 1992; Hagberg 1990; King et al. 1989; Marcus et al. 1992; Morris et al. 1990; Paffenbarger et al. 1986; Powell et al. 1987). The fact that higher levels of physical activities are associated with people having more friends and having friends who support physical activity suggests that promoting activity with friends could be helpful (Russell and Tom 2004). Besides the activity level, the physical activity types are important too. For example, a jogging lover likes to become friends with other jogging lovers, and mountain climbers like discussions with other mountain climbers.

Our health and fitness analytic framework, based on the computer-supported social matching process theory, is summarized in Figure 3-2.

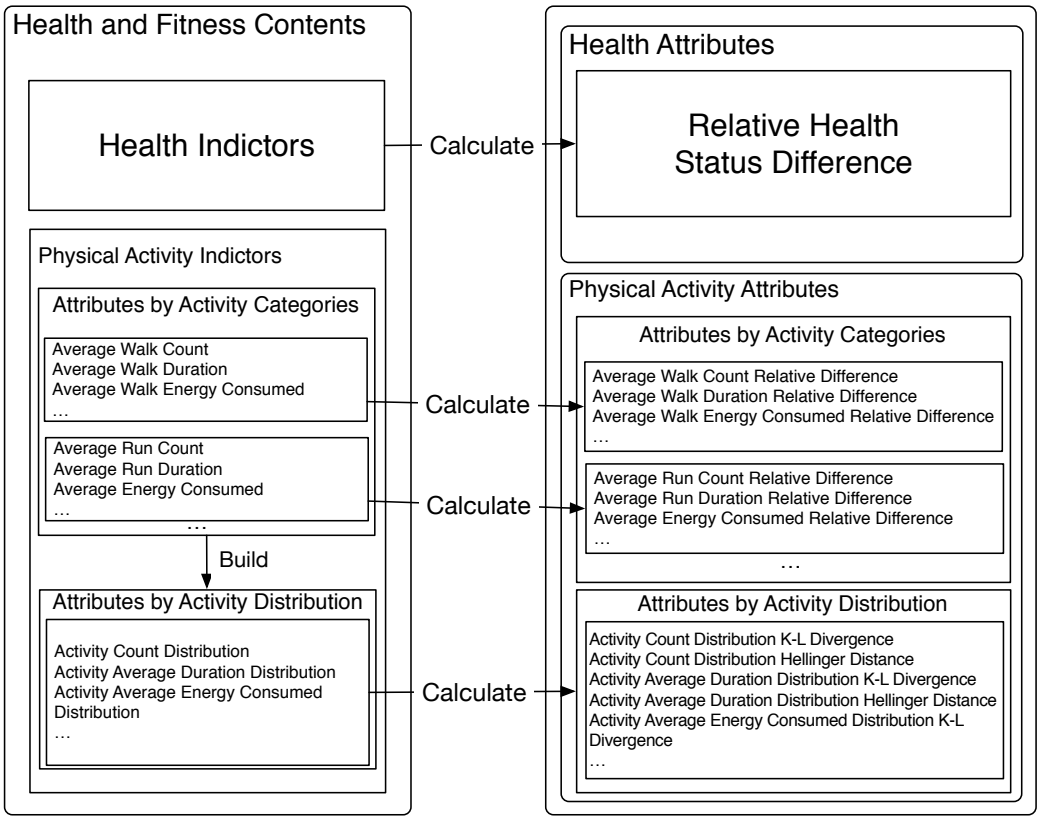


Figure 3-2 Health and Fitness Analytic Framework

By using the health indicators and physical activity indicators, we could calculate the similarities between two users' daily life patterns. The health status difference and physical activity difference will be put into our recommendation model. Figure 3-3 shows the process model:

1. All data collected by the social networking sites will be input into our system, including users' demographic attributes, social-tie attributes, and location attributes, and *the system* then combines *the data* with the attributes extracted from the health and fitness analytic framework in Figure 3.2, which are health indicators and physical activity indicators.

2. Our system will then compare a user's attributes with all other users' attributes, generate the similarities between two users, and then record pairwise similarities.

   We use the Jaccard coefficient (Salton and Michael 1983) in this study, which measures the distance between two users as:

$$d(a,b) = |\frac{a-b+\delta}{(a+b)+\delta}|$$

3. Besides the individual attributes for each type of physical activity, our system will also calculate the Kullback-Leibler divergence (K-L divergence) and Hellinger Distance in the histogram distribution level. In information theory, the K-L divergence could measure the difference between two probability distributions P and Q, and Hellinger Distance is used to quantify the similarity between two probability distributions. By using distribution divergence and similarity, we could dramatically reduce the amount of attribute sets and shorten the model building time. The K-L divergence is calculated as follows:

$$D_{KL}(P|Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

   In this formula, P(i) and Q(i) means the probability for the activity i in all activities, for user P and user Q. From the formula, we can find the K-L divergence is not symmetric and we will calculate both $D_{KL}(P|Q)$ and $D_{KL}(Q|P)$. The K-L divergence will always be greater than zero and equal to zero only if P=Q almost everywhere.

The calculation for Hellinger Distance will be:

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$$

The Hellinger Distance has a range from zero to one, H(P,Q) = 0 only if P=Q everywhere and H(P,Q)=1 if P assigns probability zero wherever Q assigns a positive probability, and vice versa.

4. We will employ data mining techniques to classify our records into two categories: *Friend* or *Not Friend*. We want to use the probabilities of the classification results as the outputs.
5. The system sorts the outputs and then selects the top-M users for the recommendation list for this user.

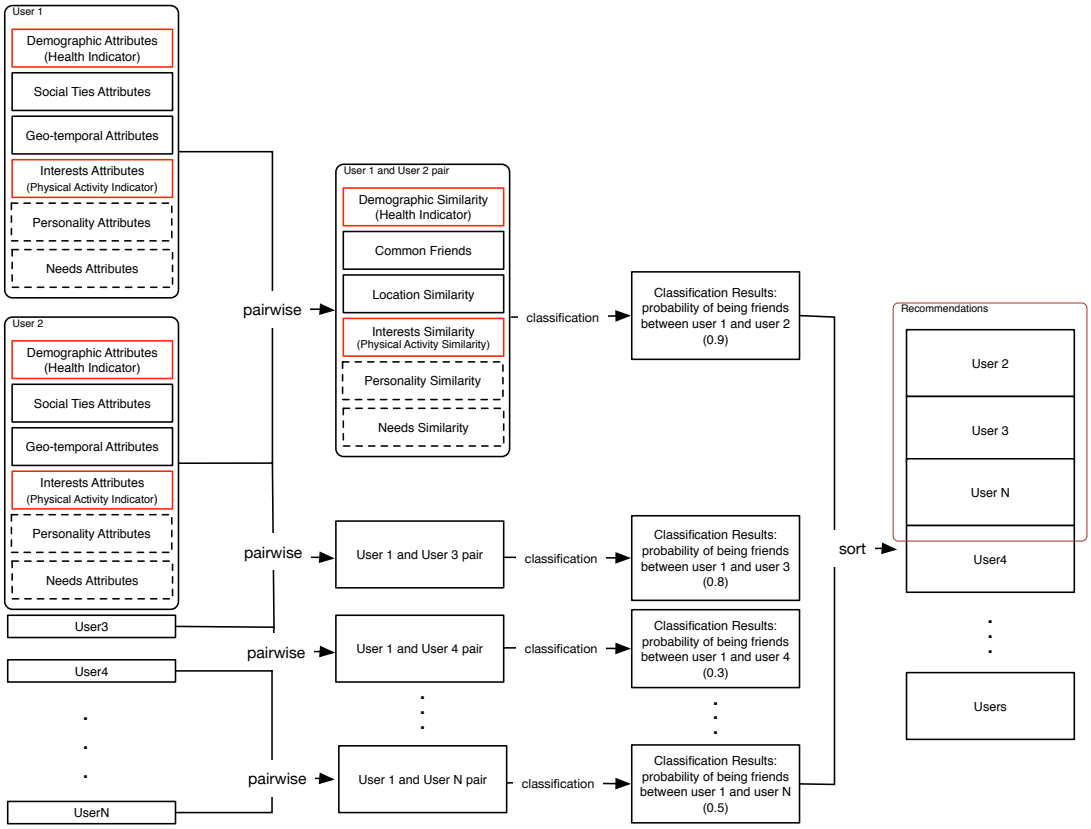


Figure 3-3 Recommendation Model

4. Experiment

4.1 Data Collection

To validate our friend recommendation model, we did a lot of research on different kinds of fitness and health social networking sites, such as *MapMyFitness.com, MyFitnessPal.com,* and *Health Mate* from Withings. After researching them, we decided to collect data from UA Record from Under Armour (https://record.underarmour.com). The comparison of the major health/fitness social networking sites is summarized in Table 4-1.

	Healthmate	MyFitnessPal	Fitbit	Record
Measure	Activity & Sleep Pattern Heartrate Weight & Fat mass Air quality Steps Nutrition	Calories Consumed Exercise Calories Burnt Nutrition Weight Loss	Calories Burnt Food Plan Drink Weight Sleep	Activities (Calories Burnt, Duration, Heartrate, Distance) Course Route Weight Sleep
Badges/Achievement	Yes, Badges for Walking Distance	No	Yes	Yes, achievements for different kind of sports
Challenge to Friends	Yes, by email	No	No	Yes
Sharing	No	Yes, you can share the weight loss trends to your friends	Yes, you can share your steps and you can see the top charts	Yes, full sharing features includes picture, messages, and physical activities you just workout
API	Not opened	Yes, you need to apply for limited usage.	Yes, you can access parts of data from API	Yes, well designed and documented API
Comments	Healthmate from Withings is the SNS for its own health measurement devices, and only have limited social networking features.	MyFitnessPal is focusing on food plan and health lifestyles.	Fitbit is a good physical activity social networking site.	Records from Under Armour is a very popular fitness social networking site built from previous MapMyFitness app.

Table 4-1 Summary of Major Health/Fitness Social Networking Sites

UA Record is the world's first 24/7 connected health and fitness system. It tracks users' steps, sleep patterns, and nutrition and logs different kind of workouts, from swimming to running. After the authorization, users will then automatically share their real-time statistics, including pace, distance, and calories burned. The UA Record system also supports users who wish to challenge their friends and connect and synchronize to pervasive devices. The best feature of this system is its application programming interfaces (APIs) that help developers design applications

for the platform. More than 17 major categories and 700 sub-categories will log and record with routes and mappings--a valuable data source for research of fitness and health social networking sites.

We collected the users' profiles, social ties, health indicators, and fitness indicators from UA Records for the period July 2014 to August 2015. During this one-year period, we had 1,089 users, with 25,310 pairs of friends among them. On average, one user has around 46 friends in our dataset. And we had 166,639 workouts within 17 major sport categories and 5,839 achievements, so a user had 166 workout records on average. The demographics we collected had users' age, the time of joining the platform, gender, country, region, and hobbies. We also had counts, energy, duration, distance, speed average, steps, and pace attributes for 17 major fitness categories. The amounts of users' achievements and health indicators were also collected. Table 4-2 summarizes the attributes used.

Demographic Attributes	
Gender	Male: 636, female: 453
Age	Range: 20 - 42, mean: 29.635
Region	There are 163 different regions.
Locality	There are 714 different localities.
Country	There are 58 different countries, USA is the major country with 795 records, and UK has 85 records.
Hobbies	There are 266 different types of hobbies.
Health Indicator Attributes	

Average heart rate	The average heart rate of the user. Range: 12-186, mean: 131.096.
Average heart cadence rate	The average heart cadence rate of the user. Range: 7-99, mean: 75.007.
Average energy consumed	The average energy consumed of the user. Range: 79-248, mean: 173.005.
Achievement Attributes	
Number of achievement	The number of achievements the user earned in the platform. Range: 0-20, mean: 5.361.
Number of personal record	The number of achievements (personal records) the user earned in the platform. Range: 0-20, mean: 2.084.
Number of King of Mountain and Queen of Mountain	The number of achievements (King of Mountain or Queen of Mountain) the user earned in the platform. Range: 0-10, mean: 0.186.
Number of Guru	The number of achievements (Guru) the user earned in the platform. Range: 0-5, mean: 0.108.
Number of fastest time	The number of achievements (fastest time) the user earned in the platform. Range: 0-7, mean: 0.137.
Number of sprint King and spring Queen	The number of achievements (sprint King or spring Queen) the user earned in the platform. Range: 0-9, mean: 0.135.
Fitness Sport Attributes	
Generic Sports	Generic sport counts of the user. Range: 0-402, mean: 3.129.
	Generic sport energy consumed of the user. Range: 0-316000kcal, mean: 4.359kcal.
	Generic sport total duration of the user. Range: 0-884.43hours, mean: 4.23hour.
	Generic sport total distance of the user. Range: 0-621.8km, mean: 4.056km.

	Generic sport average speed of the user. Range: 0-13.102mile/hour, mean: 0.102mile/hour.
	Generic sport total steps of the user. Range: 0-883,613, mean: 1,452.585.
Indoor Sports	Indoor sport counts of the user. Range: 0-187, mean: 1.988.
	Indoor sport energy consumed of the user. Range: 0-941165.696kcal, mean: 4885.213kcal.
	Indoor sport total duration of the user. Range: 0-243.65hours, mean: 2.58hour.
Walk	Walk counts of the user. Range: 0-1213, mean: 34.451.
	Walk energy consumed of the user. Range: 0-2147483.647kcal, mean: 448985.56kcal.
	Walk total duration of the user. Range: 0-2760.95hours, mean: 35.44hour.
	Walk total distance of the user. Range: 0-14809.945km, mean: 140.157km.
	Walk average speed of the user. Range: 0-27.449mile/hour, mean: 0.921mile/hour.
	Walk total steps of the user: Range: 0-5,626,292, mean: 54,552.973.
Winter Sports	Winter sport counts of the user. Range: 0-37, mean: 0.163.
	Winter sport energy consumed of the user: Range: 0-178857.632kcal, mean: 572.309kcal.
	Winter sport total duration of the user: Range: 0- 143.63hours, mean: 0.367 hour.
	Winter sport total distance of the user: Range: 0-390.659km, mean: 0.559km.
	Winter sport average speed of the user: Range: 0-6.706mile/hour, mean: 0.033mile/hour.
Bike Ride	Bike Ride counts of the user: Range: 0-1516, mean: 22.084.
	Bike Ride energy consumed of the user: Range: 0-2147483.647kcal, mean: 70323.398kcal.
	Bike Ride total duration of the user: Range: 0-1612.48hours, mean: 31.31hour.

	Bike Ride total distance of the user: Range: 0-44996.605km, mean: 540.978km.
	Bike Ride average speed of the user: Range: 0-32.08mile/hour, mean: 0.815mile/hour.
Gym	Gym counts of the user: Range: 0-350, mean: 7.581.
	Gym energy consumed of the user: Range: 0-702995.68kcal, mean: 11341.052kcal.
	Gym total duration of the user: Range: 0- 276.14hours, mean: 6.02 hour.
Indoor Winter Sport	Indoor winter sport counts of the user: Range: 0-11, mean: 0.015.
	Indoor winter sport energy consumed of the user: Range: 0-26099.792kcal, mean: 40045.759kcal.
	Indoor winter sport total duration of the user: Range: 0- 9.16hours, mean: 0.013hour.
Machine Workout	Machine workout counts of the user: Range: 0-291, mean: 3.428.
	Machine workout energy consumed of the user: Range: 0-1468935.456kcal, mean: 8115.385kcal.
	Machine workout total duration of the user: Range: 0- 568.55hours, mean: 3.44hour.
	Machine workout total distance of the user: Range: 0-2659.650km, mean: 13.898km.
	Machine workout average speed of the user: Range: 0-26.822mile/hour, mean: 0.367mile/hour.
	Machine workout total steps of the user: Range: 0-235927, mean: 413.303.
Swim	Swim counts of the user: Range: 0-203, mean: 0.701.
	Swim energy consumed of the user: Range: 0-423751.336kcal, mean: 1146.75kcal.
	Swim total duration of the user: Range: 0- 185.85hours, mean: 0.55hour.
	Swim total distance of the user: Range: 0-456.488km, mean: 1.274km.
	Swim average speed of the user: Range: 0-1.836mile/hour, mean: 0.02mile/hour.
Run	Run counts of the user: Range: 0-1278, mean: 54.046.

	Run energy consumed of the user: Range: 0-2147483.647kcal, mean: 135913.925kcal.
	Run total duration of the user: Range: 0- 4,138.88hours, mean: 64.22hour.
	Run total distance of the user: Range: 0-39768.681km, mean: 455.171km.
	Run average speed of the user: Range: 0-54.456mile/hour, mean: 1.536mile/hour.
	Run total steps of the user: Range: 0-25851650, mean: 131230.179.
Program Workout	Program workout counts of the user: Range: 0-682, mean: 3.405.
	Program workout energy consumed of the user: Range: 0-1136068.968kcal, mean: 5783.164kcal.
	Program workout total duration of the user: Range: 0- 425.37hours, mean: 3.17 hour.
Weight Workout	Weight workout counts of the user: Range: 0-748, mean: 6.63.
	Weight workout energy consumed of the user: Range: 0-2147483.647kcal, mean: 11068.366kcal.
	Weight workout total duration of the user: Range: 0- 1,769.27hours, mean: 6.57hour.
Indoor Bike Ride	Indoor Bike Ride counts of the user: Range: 0-569, mean: 2.941.
	Indoor Bike Ride energy consumed of the user: Range: 0-93608.632kcal, mean: 91.959.633kcal.
	Indoor Bike Ride total duration of the user: Range: 0- 1,282.76hours, mean: 3.37hour.
	Indoor Bike Ride average speed of the user: Range: 0-34.869mile/hour, mean: 0.753mile/hour.
Indoor Swim	Indoor swim counts of the user: Range: 0-237, mean: 1.388.
	Indoor swim energy consumed of the user: Range: 0-281235.928kcal, mean: 2141.643kcal.
	Indoor swim total duration of the user: Range: 0- 143.37hours, mean: 1.11hour.
	Indoor swim total distance of the user: Range: 0-354.4km, mean: 1.624km.

	Indoor swim average speed of the user: Range: 0-7.27mile/hour, mean: 0.042mile/hour.
Other Activity	Other activity counts of the user: Range: 0-218, mean: 1.945.
	Other activity energy consumed of the user: Range: 0-311419.304kcal, mean: 4772.66kcal.
	Other activity total duration of the user: Range: 0- 307.34hours, mean: 2.94hour.
	Other activity total distance of the user: Range: 0-1546.892km, mean: 5.325km.
	Other activity average speed of the user: Range: 0-13.947mile/hour, mean: 0.172mile/hour.
Indoor Hike	Indoor Hike workout counts of the user: Range: 0-23, mean: 0.026.
	Indoor Hike workout energy consumed of the user: Range: 0-93608.632kcal, mean: 91.959kcal.
	Indoor Hike workout total duration of the user: Range: 0- 99.287hours, mean: 0.103hour.
Class Workout	Class workout counts of the user: Range: 0-241, mean: 2.72.
	Class workout energy consumed of the user: Range: 0-433592.104kcal, mean: 4792.501kcal.
	Class workout total duration of the user: Range: 0- 291.68hours, mean: 2.73hour.
	Class workout total distance of the user: Range: 0-5781.053km, mean: 5.526km.
	Class workout average speed of the user: Range: 0-3108.093mile/hour, mean: 2.89mile/hour.
Hike	Hike counts of the user: Range: 0-126, mean: 1.084.
	Hike energy consumed of the user: Range: 0-8272966.496kcal, mean: 4368.61kcal.
	Hike total duration of the user: Range: 0- 7413.16 hours, mean: 2.25hour.
	Hike total distance of the user: Range: 0-1071.899km, mean: 7.584km.

	Hike average speed of the user: Range: 0-7.604mile/hour, mean: 0.182mile/hour.
	Hike total steps of the user: Range: 0-25851650, mean: 131230.179.
Indoor Run	Indoor Run counts of the user: Range: 0-340, mean: 4.857.
	Indoor Run energy consumed of the user: Range: 0-1618078.32kcal, mean: 10994.218kcal.
	Indoor Run total duration of the user: Range: 0- 753.84hours, mean: 4.33hour.
	Indoor Run total distance of the user: Range: 0-5643.567km, mean: 33.011km.
	Indoor Run average speed of the user: Range: 0-73.648mile/hour, mean: 0.727mile/hour.

Table 4-2 Attributes in the Collected Dataset

We then calculated the similarity/dissimilarity between every pair of users with respect to each attribute. For numeric attributes, such as friend count, tip count, tip-like count, and check-in count, we used the Jaccard coefficient (Salton and Michael 1983):

$d(a,b) = |\frac{a-b+\delta}{(a+b)+\delta}|$, where δ is a small smoothing factor and was set to 0.001 in our evaluation, a and b are the values of two users' attributes.

We then summarized the similarity/dissimilarity measures we used (Table 4-3).

Demographic Attributes	
Gender_type	Female-female: 17.28% Male-female: 48.63% Male-male: 34.09%
Age relative difference	Range: 0-0.355, mean: 0.116
In different region	There are 17118 in the same region, 517537 in different region.

In different country	There are 272055 in the same region, 320361 in different region.
City distance	Range: 0-19.955km, mean: 4.711km
Join day difference	Range: 0-113days, mean: 19.736days
Share hobbies	Range 0-5 hobbies, mean: 0.001
Health Indicator Attributes	
Relative heart rate difference	The relative average heart rate difference between users. Range: 0-0.8749, mean: 0.023.
Relative heart cadence rate difference	The relative average heart cadence rate difference between users, Range: 0-0.868, mean: 0.007.
Relative energy consumed difference	The relative average energy consumed difference between users, Range: 0-0.517, mean: 0.001.
Achievement Attributes	
Relative achievement number difference	The relative achievement number difference between users, Range: 0-1, mean: 0.558
Relative personal record difference	The relative personal record number difference between users, Range: 0-1, mean: 0.486
Relative number of King of Mountain and Queen of Mountain difference	The relative KoM or QoM number difference between users, Range: 0-1, mean: 0.151
Relative number of Guru difference	The relative guru achievement number difference between users, Range: 0-1, mean: 0.134
Relative number of fastest time difference	The relative number of fastest time record difference between users, Range: 0-1, mean: 0.156.

Relative number of sprint King and spring Queen difference	The relative number of sprint King or Queen achievement number difference between users, Range: 0-1, mean: 0.15
Fitness Sport Attributes	
Generic Sports Difference	Generic sport counts relative difference between users: Range: 0-1, mean: 0.315.
	Generic sport energy consumed relative difference between users: Range: 0-1, mean: 0.294.
	Generic sport total duration relative difference between users: Range: 0-1, mean: 0.292.
	Generic sport total distance relative difference between users: Range: 0-1, mean: 0.135.
	Generic sport average speed relative difference between users: Range: 0-1, mean: 0.078.
	Generic sport total steps relative difference between users: Range: 0-1, mean: 0.02.
Indoor Sports	Indoor sport counts relative difference between users: Range: 0-1, mean: 0.345.
	Indoor sport energy consumed relative difference between users: Range: 0-1, mean: 0.332.
	Indoor sport total duration relative difference between users: Range: 0-1, mean: 0.338.
Walk	Walk counts relative difference between users: Range: 0-1, mean: 0.71.
	Walk energy consumed relative difference between users: Range: 0-1, mean: 0.712.
	Walk total duration relative difference between users: Range: 0-1, mean: 0.718.
	Walk total distance relative difference between users: Range: 0-1, mean: 0.707.
	Walk average speed relative difference between users: Range: 0-1, mean: 0.552.
	Walk total steps relative difference between users: Range: 0-1, mean: 0.613.
Winter Sports	Winter sport counts relative difference between users: Range: 0-1, mean: 0.061.
	Winter sport energy consumed relative difference between users: Range: 0-1, mean: 0.058.

	Winter sport total duration relative difference between users: Range: 0- 1, mean: 0.059.
	Winter sport total distance relative difference between users: Range: 0-1, mean: 0.031km.
	Winter sport average speed relative difference between users: Range: 0-1, mean: 0.027.
Bike Ride	Bike Ride counts relative difference between users: Range: 0-1, mean: 0.555.
	Bike Ride energy consumed relative difference between users: Range: 0-1, mean: 0.556.
	Bike Ride total duration relative difference between users: Range: 0-1, mean: 0.557.
	Bike Ride total distance relative difference between users: Range: 0-1, mean: 0.541.
	Bike Ride average speed relative difference between users: Range: 0-1, mean: 0.485.
Gym	Gym counts relative difference between users: Range: 0-1, mean: 0.553.
	Gym energy consumed relative difference between users: Range: 0-1, mean: 0.55.
	Gym total duration relative difference between users: Range: 0- 1, mean: 0.555
Indoor Winter Sport	Indoor winter sport counts relative difference between users: Range: 0-1, mean: 0.004.
	Indoor winter sport energy consumed relative difference between users: Range: 0-1, mean: 0.004.
	Indoor winter sport total duration relative difference between users: Range: 0- 1, mean: 0.004.
Machine Workout	Machine workout counts relative difference between users: Range: 0-1, mean: 0.387.
	Machine workout energy consumed relative difference between users: Range: 0-1, mean: 0.38.
	Machine workout total duration relative difference between users: Range: 0- 1, mean: 0.384.
	Machine workout total distance relative difference between users: Range: 0-1, mean: 0.285.
	Machine workout average speed relative difference between users: Range: 0-1, mean:

	0.192.
	Machine workout total steps relative difference between users: Range: 0-1, mean: 0.024.
Swim	Swim counts relative difference between users: Range: 0-1, mean: 0.075.
	Swim energy consumed relative difference between users: Range: 0-1, mean: 0.075.
	Swim total duration relative difference between users: Range: 0- 1, mean: 0.069.
	Swim total distance relative difference between users: Range: 0-1, mean: 0.058.
	Swim average speed relative difference between users: Range: 0-1, mean: 0.057.
Run	Run counts relative difference between users: Range: 0-1, mean: 0.728.
	Run energy consumed relative difference between users: Range: 0-1, mean: 0.736.
	Run total duration relative difference between users: Range: 0- 1, mean: 0.739.
	Run total distance relative difference between users: Range: 0-1, mean: 0.732.
	Run average speed relative difference between users: Range: 0-1, mean: 10.549.
	Run total steps relative difference between users: Range: 0-1, mean: 0.655.
Program Workout	Program workout counts relative difference between users: Range: 0-1, mean: 0.32.
	Program workout energy consumed relative difference between users: Range: 0-1, mean: 0.309.
	Program workout total duration relative difference between users: Range: 0- 1, mean: 0.317.
Weight Workout	Weight workout counts relative difference between users: Range: 0-1, mean: 0.487.
	Weight workout energy consumed relative difference between users: Range: 0-1, mean: 0.475.
	Weight workout total duration relative difference between users: Range: 0- 1, mean: 0.48.
Indoor Bike Ride	Indoor Bike Ride counts relative difference between users: Range: 0-1, mean: 0.292.

	Indoor Bike Ride energy consumed relative difference between users: Range: 0-1, mean: 0.007.
	Indoor Bike Ride total duration relative difference between users: Range: 0- 1, mean: 0.289.
	Indoor Bike Ride average speed relative difference between users: Range: 0-1, mean: 0.204.
Indoor Swim	Indoor swim counts relative difference between users: Range: 0-1, mean: 0.201.
	Indoor swim energy consumed relative difference between users: Range: 0-1, mean: 0.197.
	Indoor swim total duration relative difference between users: Range: 0-1, mean: 0.199.
	Indoor swim total distance relative difference between users: Range: 0-1, mean: 0.135.
	Indoor swim average speed relative difference between users: Range: 0-1, mean: 0.133.
Other Activity	Other activity counts relative difference between users: Range: 0-1, mean: 0.382.
	Other activity energy consumed relative difference between users: Range: 0-1, mean: 0.365.
	Other activity total duration relative difference between users: Range: 0- 1, mean: 0.38.
	Other activity total distance relative difference between users: Range: 0-1, mean: 0.163.
	Other activity average speed relative difference between users: Range: 0-1, mean: 00.161.
Indoor Hike	Indoor Hike workout counts relative difference between users: Range: 0-1, mean: 0.009.
	Indoor Hike workout energy consumed relative difference between users: Range: 0-1, mean: 0.007.
	Indoor Hike workout total duration relative difference between users: Range: 0- 1, mean: 0.009.
Class Workout	Class workout counts relative difference between users: Range: 0-1, mean: 0.367.
	Class workout energy consumed relative difference between users: Range: 0-1, mean:

	0.353.
	Class workout total duration relative difference between users: Range: 0- 1, mean: 0.359.
	Class workout total distance relative difference between users: Range: 0-1, mean: 0.035.
	Class workout average speed relative difference between users: Range: 0-1, mean: 0.035.
Hike	Hike counts relative difference between users: Range: 0-1, mean: 0.254.
	Hike energy consumed relative difference between users: Range: 0-1, mean: 0.245.
	Hike total duration relative difference between users: Range: 0-1, mean: 0.249.
	Hike total distance relative difference between users: Range: 0-1, mean: 0.23.
	Hike average speed relative difference between users: Range: 0-1, mean: 0.225.
Indoor Run	Indoor Run counts relative difference between users: Range: 0-1, mean: 0.426.
	Indoor Run energy consumed relative difference between users: Range: 0-1, mean: 0.426.
	Indoor Run total duration relative difference between users: Range: 0-1, mean: 0.425.
	Indoor Run total distance relative difference between users: Range: 0-1, mean: 0.365.
	Indoor Run average speed relative difference between users: Range: 0-1, mean: 0.344.
Activity Distribution K-L Divergence	Activity counts distribution K-L Divergence (User1|User2) : Range: 0-7.5, mean: 2.977
	Activity counts distribution K-L Divergence (User2|User1) : Range: 0-7.5, mean: 3.081
	Activity energy consumed distribution K-L Divergence (User1|User2) : Range: 0-90, mean: 36.677
	Activity energy consumed distribution K-L Divergence (User2|User1) : Range: 0-90, mean: 36.782
	Activity duration distribution K-L Divergence (User1|User2) : Range: 0-13.5, mean: 5.234
	Activity duration distribution K-L Divergence (User2|User1) : Range: 0-13.5, mean:

	5.348
	Activity distance distribution K-L Divergence (User1|User2) : Range: 0-156, mean: 63.181
	Activity distance distribution K-L Divergence (User2|User1) : Range: 0-156, mean: 64.475
	Activity average speed distribution K-L Divergence (User1|User2) : Range: 0-8, mean: 3.634
	Activity average speed distribution K-L Divergence (User2|User1) : Range: 0-8, mean: 3.738
	Activity steps distribution K-L Divergence (User1|User2) : Range: 0-7, mean: 3.861
	Activity counts distribution K-L Divergence (User2|User1) : Range: 0-7, mean: 3.866
Activity Distribution Hellinger Distance	Activity counts distribution Hellinger Distance : Range: 0-1, mean: 0.719
	Activity energy consumed distribution Hellinger Distance : Range: 0-1, mean: 0.714
	Activity duration distribution Hellinger Distance : Range: 0-1, mean: 0.728
	Activity distance distribution Hellinger Distance : Range: 0-1, mean: 0.667
	Activity average speed distribution Hellinger Distance : Range: 0-1, mean: 0.667
	Activity steps distribution Hellinger Distance : Range: 0-1, mean: 0.533

Table 4-3 Similarity/Dissimilarity Measure Derived

4.2 Evaluation Procedure

To evaluate our model, we used Weka (Hall et al. 2009), an open source platform that embeds a collection of machine learning algorithms for data mining tasks. In our experiment, we converted the friend recommendation problem into a classification problem. Each instance would pair two users, and their features are the similarity/dissimilarity attributes. The dependent variable is whether the two

users were friends or not. To evaluate different networking settings, we tried to manipulate three factors: connectivity of the friend network, attribute groups, and the number of friends to recommend to a user (M).

1) Networking connectivity is one of the most important properties in social networking sites. It is defined as how many friends a user will have on average in the platform. Because we had a relatively sparse network of 1,089 users in which one user only had around 25 friends, we tried to simulate different levels of social networking connectivity. We randomly sampled the links in our dataset. By controlling the proportion of friend/non-friend links, we created four social networks with different densities of connection. We built four datasets, with 1:1, 1:2, 1:5, and 1:10 as the proportion of friend/non-friend links. For the three imbalanced datasets, we performed both a cost-sensitive classification (using the instance weighting method in Weka (Hall et al. 2009) with a cost ratio of 2:1, 5:1, and 10:1 respectively) and a regular cost-insensitive classification.
2) We tried to compare our proposed health and fitness friend recommendation model with the existing simple profile-matching and friend-of-friend methods by varying the attribute set (Table 4-4). We compared Group 1 with Group 2 to see if the health and fitness data helped in the simple profile matching algorithm. We then compared Group 3 with Group 4 to see if the health and fitness data could help in the social tie matching method. We had a large number of attributes in the health and fitness data, which may have slowed down the model building process. To reduce our attribute sets, we

tried to use histogram distribution K-L divergence and Hellinger Distance to replace individual data for each category of physical activities.

Group	**Attributes**
1	Demographic attributes only
2a (Include Activity Attributes by Categories Only)	Demographic attributes + health and fitness attributes (Activity Attributes by Categories Only)
2b (Include Activity Attributes by Histogram Only)	Demographic attributes + health and fitness attributes (Activity Attributes by Histogram Only)
2a&b (Include All Activity Attributes)	Demographic attributes + health and fitness attributes (Activity Attributes by Categories and Histogram)
3	Demographic attributes + social ties attributes
4a (Include Activity Attributes by Categories Only)	Demographic attributes + health and fitness attributes (Activity Attributes by Categories Only) + social ties attributes
4b (Include Activity Attributes by Histogram Only)	Demographic attributes + health and fitness attributes (Activity Attributes by Histogram Only) + social ties attributes
4a&b (Include All Activity Attributes)	Demographic attributes + health and fitness attributes (Activity Attributes by Categories and Histogram) + social ties attributes

Table 4-4 Attribute Groups

3) We also wanted to control the number of friends to recommend for a given user. Recommending too few friends for a user may reduce the chance for a user to find a friend, whereas recommending too many friends may frustrate the user. We also wanted to see the trends that would yield a list with the best number of recommendations in the system.

We used several different classifiers as well. According to our model, we had to have the result list in the probability format, so we used probabilistic classification methods: Bayesian network, naïve Bayes, and logistic regression. We used cross-validation to estimate the performance in each experiment environment variables setting. For each of the four different levels of network connectivity, with cost-sensitive or cost-insensitive classification, with each of the four groups of attribute sets, under different numbers of recommendations, using each of the three classifiers, we performed a 10-fold cross validation 50 times.

4.3 Results

When we collected the data, we collected the friend networks from the UA Record, making it possible for us to evaluate the supervised classification and check the classification result. We had several calculated results from the Weka platform, such as accuracy, ROC, recall, and confusion matrix. We looked at accuracy first:

$$accuracy = \frac{number\ of\ true\ positive + number\ of\ true\ negative}{number\ of\ records\ in\ test\ data\ set}$$

The baseline accuracy would be a random guess. Since we had a biased dataset in 1:2, 1:5, and 1:10 proportions, the classifiers would guess all classification outputs as negative. So, in a 1:1 network, the accuracy baseline would be 1/(1+1)=50%, and in a 1:2 network, it would be 2/(1+2) = 66.7%. To alleviate the effect of classification bias, we also performed cost-sensitive tests. The settings of the cost matrix are shown in Table 4-5.

Proportion	Cost Matrix
1:1	$\begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}$
1:2	$\begin{vmatrix} 0 & 2 \\ 1 & 0 \end{vmatrix}$
1:5	$\begin{vmatrix} 0 & 5 \\ 1 & 0 \end{vmatrix}$
1:10	$\begin{vmatrix} 0 & 10 \\ 1 & 0 \end{vmatrix}$

Table 4-5 Settings of Cost Matrix

Table 4-6 and 4-7 show the results of the accuracy test.

Group	1: 1	1: 2	1: 5	1: 10
Base Accuracy	50%	66.7%	83.3%	90.9%
1	54.8301%	65.405%	82.798%	90.9087%
2a	64.6898%	71.2235%	89.9295%	90.8994%
2b	62.744%	71.0083%	84.3913%	90.9127%
2a&b	65.083%	71.561%	89.6586%	90.9375%
3	83.5579%	84.9875%	89.6721%	93.3059%
4a	84.7234%	85.7751%	89.9295%	93.4255%
4b	84.1979%	85.4919%	89.8991%	93.5879%
4a&b	84.761%	85.8304%	90.1383%	93.6838%

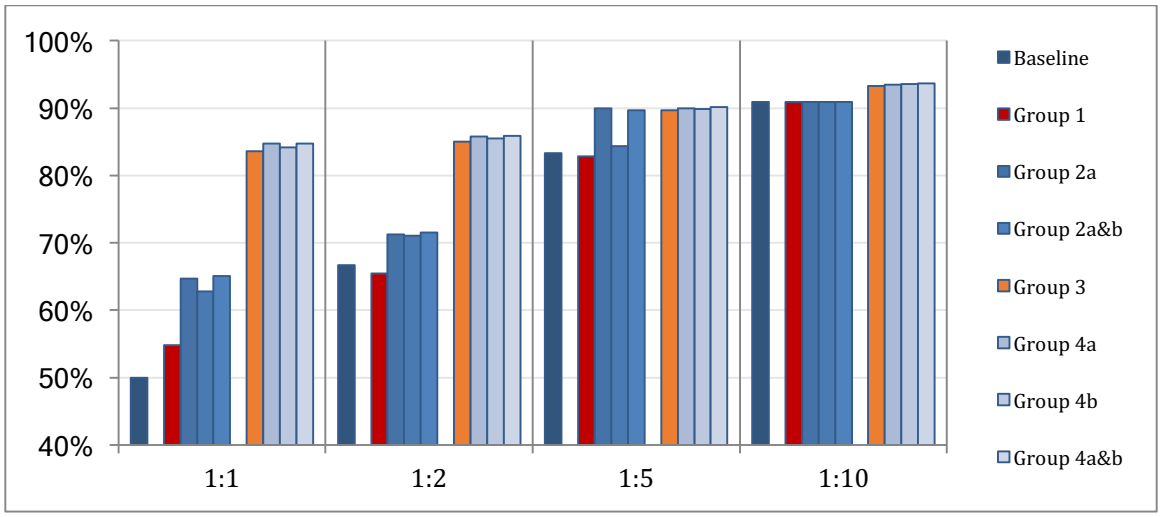


Table 4-6 Accuracy of Friend Recommendation

Group	1：1	1：2	1：5	1：10
Base Accuracy	50%	66.7%	83.3%	90.9%
1	54.8301%	65.5498%	81.8813%	90.894%
2a	64.6898%	71.4948%	82.8204%	90.6422%
2b	62.744%	68.5787%	82.551%	89.8471%
2a&b	65.083%	71.783%	83.2982%	90.7518%
3	83.5579%	85.1613%	89.761%	90.894%
4a	84.7234%	86.0503%	90.054%	93.2933%
4b	84.1979%	86.0635%	90.0994%	92.2801%
4a&b	84.761%	86.2979%	90.3339%	93.5793%

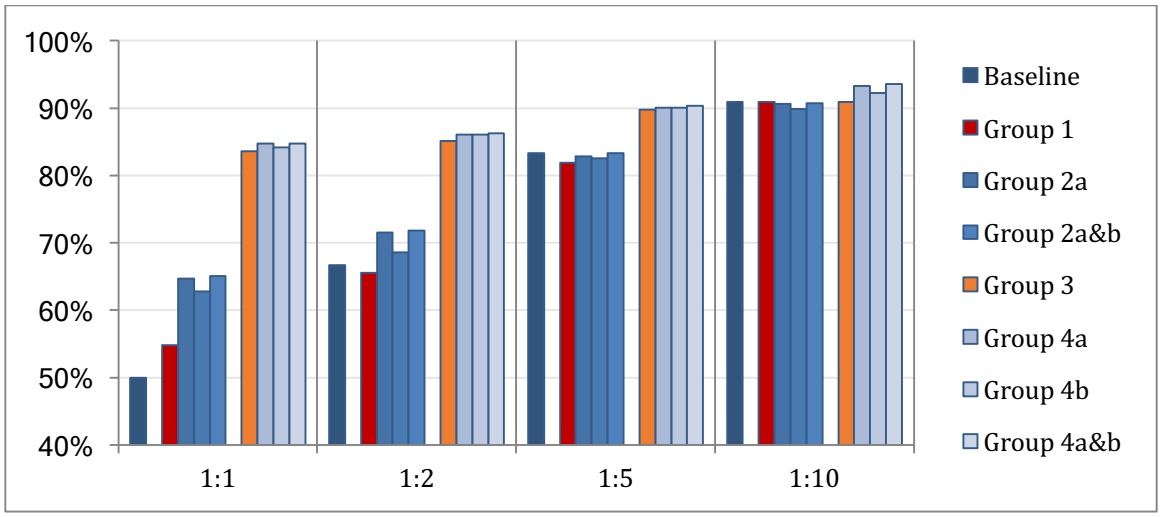


Table 4-7 Accuracy of Friend Recommendation

From the accuracy results, we found that in a more connected network, e.g., 1:1 network or 1:2 network, health and fitness data helped the classification results. The accuracy was improved not only in simple profile matching but also in social tie matching. However, in a sparser network, e.g., 1:5 or 1:10 network, more attributes actually did not improve classification accuracy. Another interesting finding was that after replacing detailed activity attributes of sport categories with activity histogram K-L divergence and Hellinger Distance, the accuracy results did not change a lot.

To further compare the physical activity attribute sets, we recorded and compared their model building time as shown in Table 4-8.

Group	Model Building Time (sec)						
	1: 1	1: 2	1:2 Cost Sensitive	1: 5	1:5 Cost Sensitive	1: 10	1:10 Cost Sensitive
2a	19.45	27.98	26.9	70.38	71.4	131.94	137.1
2b	4.76	8.07	7.28	16.41	15.48	31.26	29.32

Group	Model Building Time (sec)						
	1：1	1：2	1:2 Cost Sensitive	1：5	1:5 Cost Sensitive	1：10	1:10 Cost Sensitive
2a&b	23.41	36.41	32.54	86.7	77.03	159.63	152.86
4a	27.06	36.96	35.18	74.14	71.84	138.12	142.82
4b	6.54	11.3	9.93	20.32	21.3	35.26	36.48
4a&b	28.66	65.67	63.36	108.33	105.89	298.59	299.89

Table 4-8 Model Building Speed Comparison

We found that after reducing the attribute sets, our model building time would be significantly shortened (see Table 4-8). To use less time to reach a similar performance, it would be better to use the attribute groups 2b and 4b as our friend recommendation attribute sets. In the following paragraphs, we use group 2 and group 4 to refer to group 2b and group 4b.

The accuracy represents only the results for the classification process--but not the actual friend recommendation part. To evaluate the recommendation performance, we further simulated the top-M recommendation results and calculated the precision. By using the classification probability results from the outputs, we used a piece of Java program to sort and select the top-M users for the recommendation list. Then, for a given user, the top-M recommendation precision is the proportion of the M-recommended friends that are actually friends of the user. The average of the top-M recommendation precision for all users provides an aggregate performance measure. To further detect the position of our recommendation method, we calculated the baseline of the recommendation, which used the combination calculation for recommendation list. We calculated the

optimal case too, which assumed all friends would be at the top of the recommendation list.

Suppose in a dataset with n users, each user i has F_i friend links and N_i non-friend links. The average precisions of the baseline and the optimal recommender can be calculated as follows.

For each user i, if the total number of links $F_i + N_i$ is less than the number of recommended friends M, then all friend links would be in the recommendation list, so the precision is F_i/M. Otherwise, the number of possible ways to select M links is $C_{F_i+N_i}^{M}$. The number of possible ways to select x friend links and M-x non-friend links is $C_{F_i}^{x} \times C_{N_i}^{M-x}$. The expected precision of random top-M recommendation for this user is therefore:

$$BP_i = \frac{\sum_{j=0}^{M} j \cdot C_{F_i}^{j} \cdot C_{N_i}^{M-j}}{C_{F_i+N_i}^{M} \cdot M}.$$

The average baseline precision for the dataset is:

$$BP = (\sum_{1}^{n} BP_i)/n.$$

For each user i, the number of friend links selected by the optimal recommender will be $\min(F_i, M)$, so, for top-M recommendation, the optimal precision is:

$$OP = (\sum_{1}^{n} \frac{\min(F_i, M)}{M})/n.$$

We selected three cases to represent here, which have M=3, 5, and 10, as shown in Table 4-9.

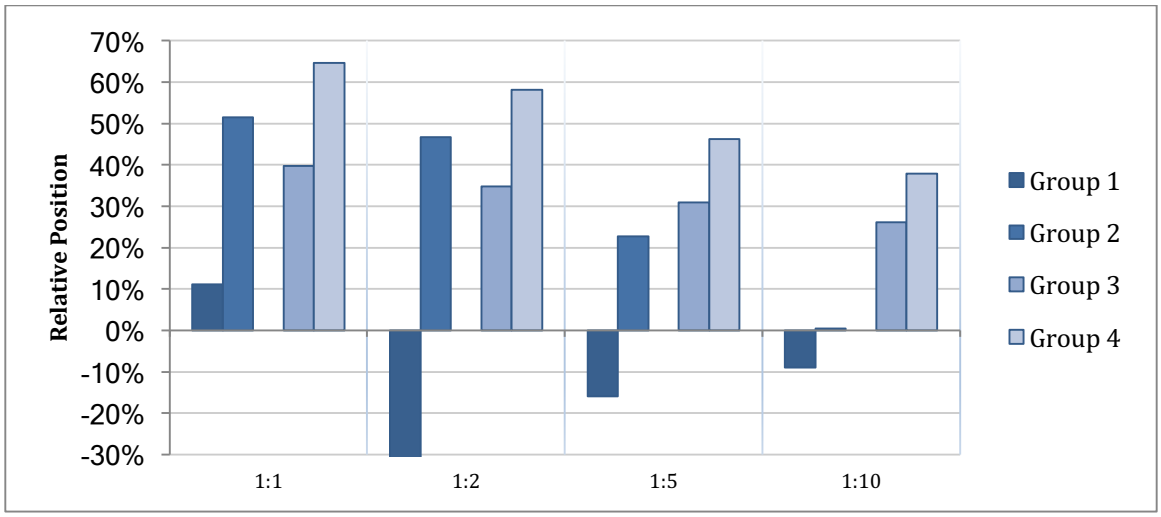
		M		
Dataset		**3**	**5**	**10**
1:1	OP	90.77	87.81	80.63
	BP	35.02	35.29	33.24
1:2	OP	97.52	94.97	87.81
	BP	24.225	24.223	21.28
1:5	OP	96.34	94.97	87.81
	BP	13.42	13.40	7.57
1:10	OP	65.7	64.0	59.86
	BP	0	0	0

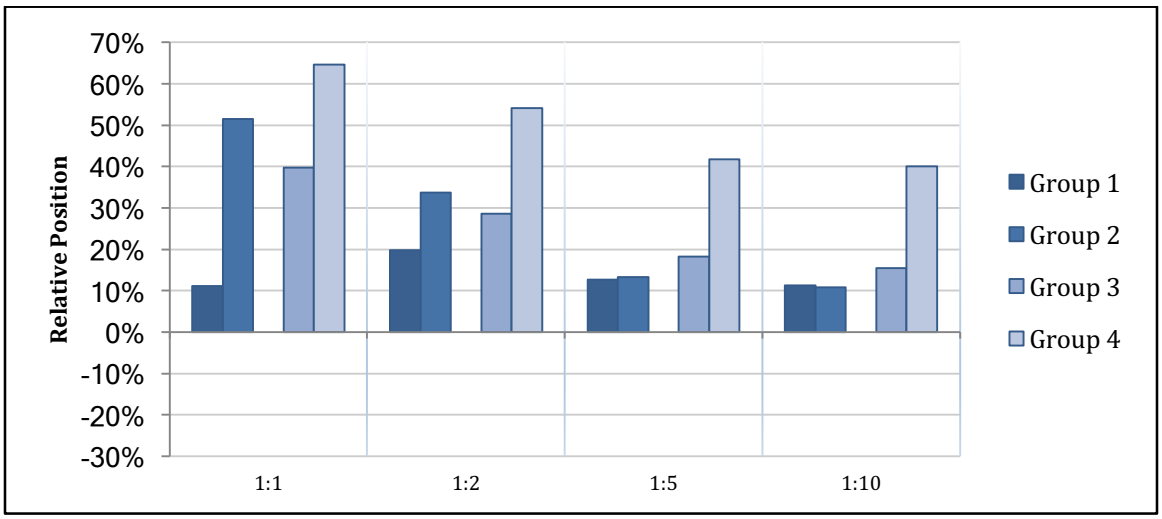
Table 4-9 Baseline and Optimal Precision of Friend Recommendation

When we had the baseline and the optimal precisions, we could also calculate the relative positions of our recommendation precisions. The formula for the relative position is:

$$Position = \frac{Recommendation\ Precision - Baseline\ Precision}{Optimal\ Precision - Baseline\ Precision}$$

Then we normalized all the results for the top 3 recommendations and placed them in the same chart, as shown in Table 4-10.





	1:1	1:2	1:2 Cost Sensitive	1:5	1: 5 Cost Sensitive	1:10	1:10 Cost Sensitive
Group 1	11.11%	-32.47%	19.81%	-15.88%	12.68%	-9.05%	11.35%
Group 2	51.53%	46.71%	33.68%	22.69%	13.28%	0.45%	10.87%
Group 3	39.74%	34.73%	28.64%	30.85%	18.26%	26.13%	15.45%
Group 4	64.60%	58.11%	54.067%	46.18%	41.79%	37.94%	40.04%

Table 4-10 Relative Positions of Top 3 Friend Recommendations

We can see from the results of the recommendations that in a more connected network, health and fitness attributes did improve the recommendation

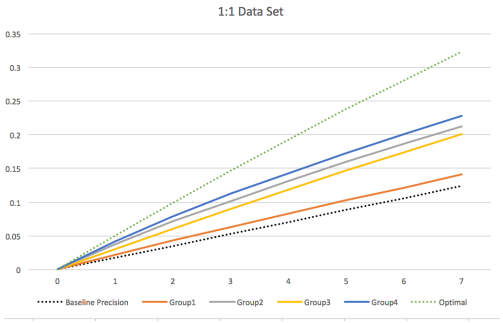
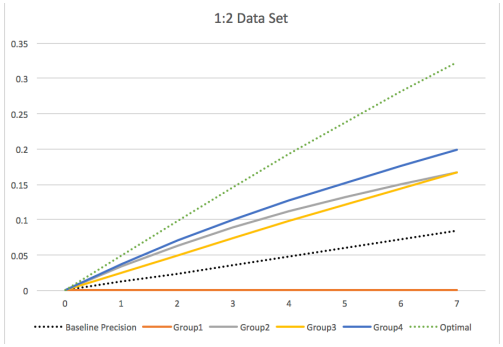
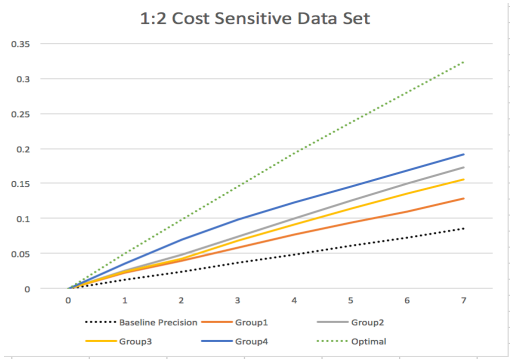
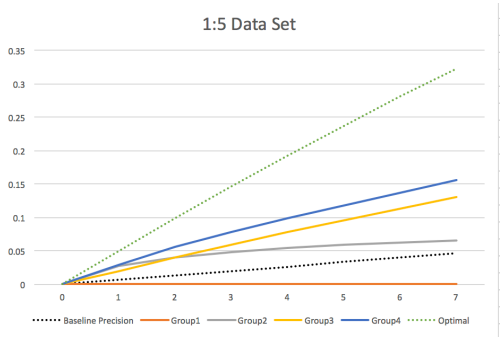
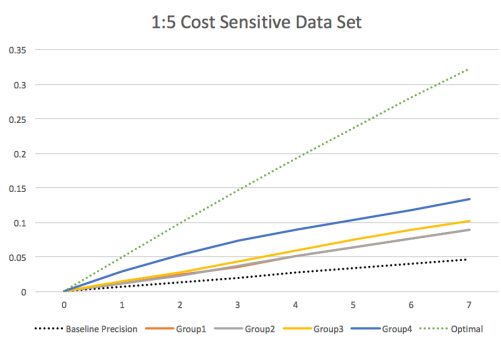
performance, as compared to profile matching and social tie matching. Remarkably, even in a sparser network, we saw improvement as well. If we did not use over-sampling for the imbalanced dataset, we saw that the profile matching performed worse than baseline precision. After the over-sampling process, the results improved.

To make the evaluation more comprehensive, we also produced performance charts for precision based on the classification results. The x-axis of the chart is the number of links we recommended, and the y-axis is the ratio of the true friend links to the length of the recommendation list (M). Because the friend links are different for each user, we report the average value.

The maximum value of x-axis was related to the total links we had in the test dataset. For a user, it could exceed hundreds, so we selected the average friend links and added a bit more to get an applicable maximum number. For example, in the 1:1 dataset, we had 25,310 friend links, 25,310 non-friend links, and 835 users, so the average number of friend links per user would be (25,310 + 25,310) / 1089 ≈ 46.5.

Because we were not going to reach the maximum number in the x-axis, we would not reach 100% in the y-axis. And since the maximum precision that our recommendation would have depended on the accuracy of the classification, the value could not reach 100% and becomes flat after some value of x.

Figure 4-1 shows the performance charts for different proportions and with/without the cost-sensitive matrix.

	Not Cost Sensitive	*Cost Sensitive*
1:1		
1:2		
1:5		

1:10

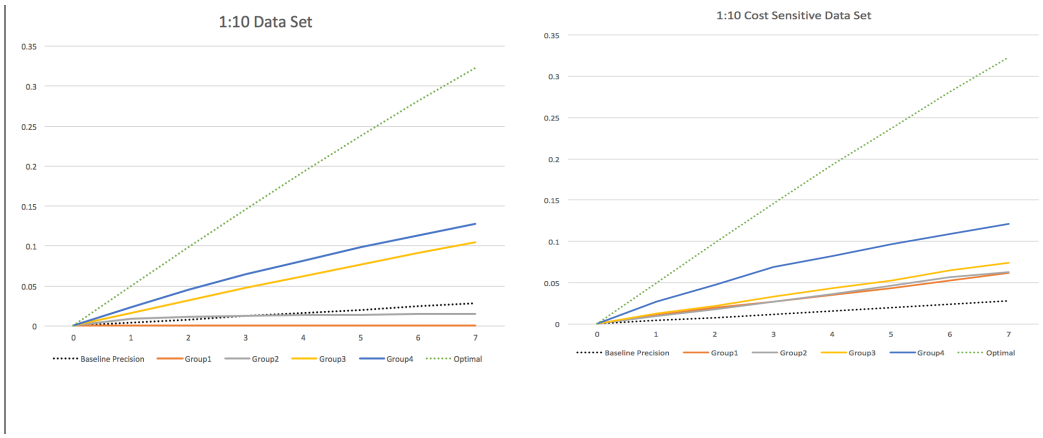


Figure 4-1 The Performance Chart of Recommendations

From the performance charts, we can clearly see that in any length of recommendation list that the group 2 attributes could improve the group 1 attributes performance. And in group 4, health and fitness data could provide better recommendation results. We also see that although a more connected dataset, e.g., the 1:1 dataset, which recommended fewer people, would be more efficient, in a sparser network, we would need to extend the recommendation list to reach our target.

5. Discussion

In this essay, we proposed an advanced model for a friend recommendation system specifically for fitness and health social networking sites. By following the guidelines for a computer-supported social matching process, fitness tracking data and health indicators data were collected and included in our model. We developed a health/fitness analytic framework, in which the fitness and health data were

systematically analyzed. The results from our experiments demonstrated that our model performed quite well and improved profile matching and link matching.

This essay makes a number of contributions with respect to both research and practice:

1) With regard to academic research, to the best of my knowledge, this is the first study to use health indicator information and fitness data in the social networking area. Health and fitness online communities are becoming more and more critical; however, very little research has focused on using health indicators and fitness data to fulfill the requirement of friend recommendations. In our study, we have demonstrated that by using our implemented framework, health data could imply users' lifestyles and interests. The experimental results confirmed that the health indicators and fitness data could significantly contribute to friend recommendation accuracy and precision.
2) In this research, we further tested the computer-supported social matching process. Part of the six categories of attributes in Terveen and McDonald (2010)'s model were selected and used. We verified how the lifestyle attributes could imply users' similarities and help make friend recommendations.
3) With regard to practice, as far as we could tell, very few applications have focused on the usage of health indicators and fitness data. Most of the applications only visualize this data in users' timelines and try to engage

others for physical exercise. However, in our study, we proposed a method to demonstrate how to analyze the data collected from wearable devices and health sensors. The category of fitness workouts, durations, heart rates, running distances, etc., were systematically summarized and helped to improve the social networking building process.

4) We provided an appropriate process not only to evaluate the recommendation performance for data mining accuracy but also to measure recommendation precision based on the number of users in the recommendation list. We analyzed our algorithm in three dimensions: connectivity density, attributes, and recommendation list length, and we found that in more highly connected social networking sites, we do not need to recommend many users, and in a sparser network, we need to recommend six to seven users.

Our study also suffers from the following limitations:

1) Compared to the first two essays, we collected more user records from the UA Records platform than from foursquare.com, but the data was still very sparse. The low density of our dataset influenced the recommendation performance. We tried to use sub-sampling to simulate a more connected network; however, the friend links were repeatedly used and caused biased results.
2) The UA Records platform does not have details of users' profile and demographic information. Thus, we had very few attributes to perform profile matching. For future research, we would take a longer time to select

users who have Facebook accounts, which would enable us to collect more demographic data for friend recommendations.

3) Finally, the dependent variable was based on the friend links we found from the dataset, which means two users were already friends in the social networking sites. The implication is that these two users were a match, but it is not known whether these users would become friends. Future research should examine the long-term results whereby two users who were previously not friends become friends later.

We could possibly improve our work for future research in several ways:

1) To further demonstrate the computer-supported social matching theory, we could analyze users' needs in the social networking sites. Because users' needs are relatively short term, analysis would need to be updated more frequently. We would need to do a more real time-like algorithm to analyze users' attributes.
2) The needs attributes could be represented by the physical activity challenge invitations from one user to other users. It could denote the request for finding workout partners and friends and could possibly help friend recommendations.
3) We could analyze users' activity patterns more carefully and at a finer granularity level. For example, there are some users who are more likely to perform physical exercise in the morning, and there are others who may work out after work. Some users want to engage in sports with more frequency and in shorter time intervals, while others prefer longer activity

times. All these patterns could be categorized more carefully and could improve recommendation performance.

4) We could develop long-term research on collecting data. We could examine the activities of a user after the user has received a recommendation, for example, whether or not the user links to the person after the recommendation. This would provide better ways to evaluate the recommendation system.
5) Due to privacy protection in health and fitness social networking sites, we were not able to collect all categories of health indicator information. In future research, we could try to improve our data collection process, for example, by using Apple's ResearchKit, to request users' signatures for health data collection for research purposes. This would provide better insight into the utility of health indicators in friend recommendation.

References

2013. "Health Fact Sheet," Pew Ressearch Center.

2016. "The Statistics Portal." from http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Adamic, L. A., and Adar, E. 2003. "Friends and Neighbors on the Web," *Social Networks* (25:3), pp. 211-230.

Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. 2005. "Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach," *ACM Transactions on Information Systems (TOIS)* (23:1), pp. 103-145.

Adomavicius, G., and Tuzhilin, A. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *Knowledge and Data Engineering, IEEE Transactions on* (17:6), pp. 734-749.

Ahtinen, A., Mattila, E., Väätänen, A., Hynninen, L., Salminen, J., Koskinen, E., and Laine, K. 2009. "User Experiences of Mobile Wellness Applications in Health Promotion," *PervasiveHealth'09*.

Al Hasan, M., Chaoji, V., Salem, S., and Zaki, M. 2006. "Link Prediction Using Supervised Learning," *SDM06: Workshop on Link* analysis, counter-terrorism and security.

Arazy, O., Kumar, N., and Shapira, B. 2010. "A Theory-Driven Design Framework for Social Recommender Systems," *Journal of the Association for Information Systems* (11:9), pp. 455-490.

Balatsoukas, P., Kennedy, C. M., and Buchan, I. 2015. "The Role of Social Network Technologies in Online Health Promotion: A Narrative Review of Theoretical and Empirical Factors Influencing Intervention effectiveness" *Journal of medical* internet research.

Benchettara, N., Kanawati, R., and Rouveirol, C. 2010. *A Supervised Machine Learning Link Prediction Approach for Academic Collaboration Recommendation*. New York, New York, USA: ACM.

Bennett, S. 2014. "Twitter USA: 48.2 Million Users Now, Reaching 20% of Population by 2018," in: *Social Times*.

Bickmore, T., Caruso, L., and Clough-Gorr, K. 2005. "Acceptance and Usability of a Relational Agent Interface by Urban Older Adults.," *CHI'05*.

Carroll, J. 2010. "Location Is the New Intelligence," *CA Magazine*, pp. 1-2.

Chen, A., Watson, R., Boudreau, M., and Karahanna, E. 2009a. "Organizational Adoption of Green Is &Amp; It: An Institutional Perspective," *ICIS 2009 Proceedings*, p. 142.

Chen, I. B. X. 2009. "A Framework for Context Sensitive Services: A Knowledge Discovery Based Approach," *Decision Support Systems* (48:1), p. 10.

Chen, J., Geyer, W., Dugan, C., Muller, M., and Guy, I. 2009b. *Make New Friends, but Keep the Old: Recommending People on Social Networking Sites*. ACM.

Chen, X. L. H. 2013. "Recommendation as Link Prediction in Bipartite Graphs: A Graph Kernel-Based MachineLearning Approach," *Decision Support Systems* (54:2), p. 10.

Christidis, K., and Mentzas, G. 2013. "A Topic-Based Recommender System for Electronic Marketplace Platforms," *Expert Systems With Applications* (40:11), pp. 4370-4379.

Coleman, M., and Liau, T. L. 1975. "A Computer Readability Formula Designed for Machine Scoring," *Journal of Applied Psychology* (60), pp. 283-284.

Consolvo, S., Klasnja, P., McDonald, D., Avrahami, D., Froehlich, J., LeGrand, L., Libby, R., Mosher, K., and Landay, J. 2008. "Flowers or a Robot Army? Encouraging Awareness & Activity with Personal, Mobile Displays," in: *UbiComp'08*.

Dahlhaus, R. 1996. "On the Kullback-Leibler Information Divergence of Locally Stationary Processes," *Stochastic Processes and their Applications* (62:1), pp. 139-168.

Daw, J., Margolis, R., and Verdery, A. M. 2015. "Siblings, Friends, Course-Mates, Club-Mates: How Adolescent Health Behavior Homophily Varies by Race, Class, Gender, and Health Status," *Social Science & Medicine* (125:C), pp. 32-39.

de la Haye, K., Robins, G., Mohr, P., and Wilson, C. 2011. "Homophily and Contagion as Explanations for Weight Similarities among Adolescent Friends.," *Journal of Adolescent Health* (49), p. 6.

Deng, Z.-H., Wang, Z.-H., and Zhang, J. 2013. "Robin: A Novel Personal Recommendation Model Based on Information Propagation," *Expert Systems With Applications* (40:13), pp. 5306-5313.

Dishman, R. 1992. "Psychological Effects of Exercise for Disease Resistance and Health Promotion," *CRC Press*, p. 28.

Dudley-Nicholson, J. 2013. "Australians Now Using Social Media in Bedrooms and Toilet Cubicles," http://www.news.com.au/.

Ennett, S. T., and Baumann, K. E. 1994. "The Contribution of Influence and Selection to Adolescent Peer Group Homogeneity: The Case of Adolescent Cigarette Smoking," *Journal of Personality and Social Psychology* (67), p. 10.

Fogg, B. J. 2003. *Persuasive Technology*. Morgan Kaufmann.

Fritz, T., Huang, E. M., Murphy, G. C., and Zimmermann, T. 2014. "Persuasive Technology in the Real World: A Study of Long-Term Use of Activity Sensing Devices for Fitness," in: *CHI*.

Gage, E. A. 2013. "Social Networks of Experientially Similar Others: Formation, Activation, and Consequences of Network Ties on the Health Care Experience," *Social Science & Medicine* (95), pp. 43-51.

Gavalas, D., and Kenteris, M. 2011. "A Web-Based Pervasive Recommendation System for Mobile Tourist Guides," *Personal and Ubiquitous Computing* (15:7), pp. 759-770.

Gunning, R. 1952. "The Technique of Clear Writing,".

Guy, I., Ronen, I., and Wilcox, E. 2009a. *Do You Know?: Recommending People to Invite into Your Social Network*. New York, New York, USA: ACM.

Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., and Ofek-Koifman, S. 2009b. *Personalized Recommendation of Social Software Items Based on Social Relations*. New York, New York, USA: ACM.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. "The Weka Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter* (11:1), pp. 10-18.

Hirsch, J. A., James, P., and Robinson, J. 2014. "Using Mapmyfitness to Place Physical Activity into Neighborhood Context," *Frontiers in Public Health*.

Jeh, G., and Widom, J. 2002. *Simrank: A Measure of Structural-Context Similarity*. ACM.

Jensen, C., Davis, J., and Farnham, S. 2002. *Finding Others Online: Reputation Systems for Social Online Spaces*. New York, New York, USA: ACM.

John, O., and Naumann, L. 2008. "Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues," *O. P. John, RW Robins* (8), pp. 114-158.

K. Smith, N. C. 2008. "Social Networks and Health," *Annual Review of Sociology* (34), p. 24.

Khurri, A., and Luukkainen, S. 2009. "Identification of Preconditions for an Emerging Mobile Lbs Market," *Journal of Location Based Services* (3:3), pp. 188-209.

Kincaid, J., Jr., F. R., RL, R., and BS, C. 1975. "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Research Branch Report*).

King, A., Taylor, C., Haskell, W., and DeBusk, R. 1989. "Influence of Regular Aerobic Exercise on Psychological Health," *Health Psychology* (8), p. 19.

Kullback, S., and Leibler, R. A. 1951. "On Information and Sufficiency," *The Annals of Mathematical Statistics*.

Kuo, T.-T., Yan, R., Huang, Y.-Y., Kung, P.-H., and Lin, S.-D. 2013. *Unsupervised Link Prediction Using Aggregative Statistics on Heterogeneous Social Networks*. ACM.

Liben Nowell, D., and Kleinberg, J. 2007. "The Link‐Prediction Problem for Social Networks," *Journal of the American Society for Information Science and Technology* (58:7), pp. 1019-1031.

Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V. 2010. *New Perspectives and Methods in Link Prediction*. New York, New York, USA: ACM.

Marcus, R., Drinkwater, B., and Dalsky, G. 1992. "Osteoporosis and Exercise in Women," *Medicine & Science in Sports & Exercise* (24), p. 6.

Mayer, J. M., Motahari, S., Schuler, R. P., and Jones, Q. 2010. "Common Attributes in an Unusual Context: Predicting the Desirability of a Social Match," *Proceedings of the fourth ACM conference on Recommender systems*, pp. 337-340.

McCullagh, P., Matzkanin, K., Shaw, S., and Maldonado, M. 1993. "Motivation for Participation in Physical Activity," *Pediatric Exercise Science*, p. 9.

McPherson, M., Smith-Lovin, L., and Cook, J. M. 2001. "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology* (27), p. 29.

Menon, A. M., Deshpande, A. D., Perri III, M., and Zinkhan, G. M. 2003. "Trust in Online Prescription Drug Information among Internet Users," *Health Marketing Quarterly* (20:1), pp. 17-35.

Morris, J., Clayton, D., Everitt, M., Semmence, A., and Burgess, E. 1990. "Exercise in Leisure Time: Coronary Attack and Death Rates.," *British Heart Journal* (63), p. 9.

N.A. Christakis, J. H. F. 2007. "The Spread of Obesity in a Large Social Network over 32 Years," *The New England Journal of Medicine* (357), p. 9.

Newman, M. 2001. "Clustering and Preferential Attachment in Growing Networks," *Physical Review E* (64:2), p. 025102.

O' Madadhain, J., Hutchins, J., and Smyth, P. 2005. "Prediction and Ranking Algorithms for Event-Based Network Data," *ACM SIGKDD Explorations Newsletter* (7:2), pp. 23-30.

Ott, N., and Meurers, D. 2011. "Information Retrieval for Education: Making Search Engines Language Aware," *Themes in Science and Technology Education* (3), pp. 9-30.

Paffenbarger, R. S., Hyde, R. T., Wing, A. L., and Hsieh, C. C. 1986. "Physical Activity, All-Cause Mor- Tality, and Longevity of College Alumni," *The New England Journal of Medicine* (314), p. 8.

Pai, A. 2015. "Under Armour's Connected Fitness Apps Now Have 140 Million Users." from http://mobihealthnews.com/45618/under-armours-connected-fitness-apps-now-have-140-million-users

Park, D. H., Kim, H. K., Choi, I. Y., and Kim, J. K. 2012. "A Literature Review and Classification of Recommender Systems Research," *Expert Systems With Applications*.

Passer, M. W. 1982. "Children in Sport: Participation Motives and Psychological Stress," *Quest 33*, p. 13.

Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., and Howard, N. 2013. "Common Sense Knowledge Based Personality Recognition from Text," *Context based Expert Finding in Online Communities using Social Network Analysis* (8266:Chapter 42), pp. 484-496.

Powell, K., Thompson, P., Caspersen, C., and Ford, E. 1987. "Physical Activity and the Incidence of Coronary Heart Disease.," *Annual Review of Public Health* (8), p. 34.

Quercia, D., and Capra, L. 2009. "Friendsensing: Recommending Friends Using Mobile Phones," *the third ACM conference*, pp. 273-276.

Ren, H. Y. Y. Q. R. Y. M. 2014. "Human Mobility Discovering and Movement Intention Detection with Gps Trajectories," *Decision Support Systems* (63), p. 12.

Russell, J., and Tom, B. 2004. "Non-Curricular Approaches for Increasing Physical Activity in Youth: A Review," *Preventive medicine* (39:1), p. 6.

Salton, G., and Michael, J. 1983. "Introduction to Modern Information Retrieval,".

Sankaradass, V., and Arputharaj, K. 2011. "An Intelligent Recommendation System for Web User Personalization with Fuzzy Temporal Association Rules," *European Journal of Scientific Research*, pp. 1-9.

Scellato, S., Noulas, A., and Mascolo, C. 2011. *Exploiting Place Features in Link Prediction on Location-Based Social Networks*. New York, New York, USA: ACM.

Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. 2010. *Folks in Folksonomies: Social Link Prediction from Shared Metadata*. ACM.

Senter, R. J., and Smith, E. A. 1967. "Automated Readability Index,".

Shi, Z. W., Andrew B. 2013. "Network Structure and Observational Learning: Evidence from a Location-Based Social Network," *Journal of Management Information Systems* (30:2), p. 27.

Simpkins, S. D., Schaefer, D. R., Price, C. D., and Vest, A. E. 2013. "Adolescent Friendships, Bmi, and Physical Activity: Untangling Selection and Influence through Longitudinal Social Network Analysis," *Journal of Research on Adolescence* (23:3), pp. 537-549.

Smith, C. 2016. "By the Numbers: 125+ Amazing Linkedin Statistics," in: *DMR Stats*.

Terveen, L., and McDonald, D. W. 2005. "Social Matching: A Framework and Research Agenda," *ACM Transactions on Computer-Human Interaction (TOCHI)* (12:3), pp. 401-434.

Tian, Y., He, Q., Zhao, Q., Liu, X., and Lee, W.-C. 2010a. *Boosting Social Network Connectivity with Link Revival*. New York, New York, USA: ACM.

Tian, Y., He, Q., Zhao, Q., Liu, X., and Lee, W.-c. 2010b. "Boosting Social Network Connectivity with Link Revival," *CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*), pp. 589-598

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region," *Economic geography* (46), p. 234.

Tollmar, K., Bentley, F., and Viedma, C. 2012. "Mobile Health Mashups: Making Sense of Multiple Streams of Wellbeing and Contextual Data for Presentation on a Mobile Device," in: *6th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, pp. 1-8.

twitter.com. 2016. "Twitter Usage - Company Facts." from https://about.twitter.com/company

Valente, T. W. 2012. "Network Interventions," *Science*:337), p. 4.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. 2011. *Human Mobility, Social Ties, and Link Prediction*. ACM.

Woolridge, A. 2011. "Too Much Buzz: Social Media Provides Huge Opportunities, but Will

Bring Huge Problems," *Economist*.

Xu, B., Chin, A., Wang, H., and Zhang, L. 2011. "Social Linking and Physical Proximity in a Mobile Location-Based Service,", pp. 99-108.

Zheleva, E., Getoor, L., Golbeck, J., and Kuter, U. 2010. "Using Friendship Ties and Family Circles for Link Prediction." Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 97-113.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., and Ma, W.-Y. 2011. "Recommending Friends and Locations Based on Individual Location History," *ACM Transactions on the Web* (5:1), pp. 1-44.

Conclusion

In this three-essay dissertation, we focused on one of the essential tasks in online social networks – friend recommendation systems. Such systems can help users find new and more appropriate friends. They are useful for new users to deal with the "cold start" problem and for old users to further expand their friend networks. Having more users with a higher density in friend networks could help social networks maintain high levels of activity. While item recommendation has been extensively studied by researchers and online social network platform providers, friend recommendation system research is still at an early stage. Based on the computer-supported social matching process, we proposed three friend recommendation systems, with different attribute sets and analytic frameworks.

In the first essay, we focused on the location data generated from users' GPS-enabled smart phones. The proposed location analytic framework organizes the massive location check-in data into three categories. The first category consists of users' physical geographic attributes. The physical distance between users could imply users' possibilities to meet each other or to provide useful information to friends. The second category consists of users' POI attributes, which could reflect users' lifestyles and activity ranges. The last category is based on users' check-ins entirely and consists of distribution divergence between two users' check-ins. Our location analytic framework helps friend recommendation systems perform better than simple profile matching or friend-of-friend

matching. The experimental results demonstrate that well-structured location attributes could lead to higher accuracy in friend recommendations.

In the second essay, we studied the use of user generated contents in friend recommendation. UGCs have become very popular and have attracted many researchers and business analytics professionals. Most research has focused on discovering users' patterns from this huge amount of data. Unfortunately, much less has been devoted to using UGCs to make friend recommendations. In this essay, we proposed a text analytic framework to process UGCs for friend recommendation. We analyzed users' posts and check-in documents using various shallow to deep text analytic techniques. The derived measures of document length, writing style, readability, subjectivity, and big five personality could imply the interests and personality of a user. We also performed sentiment analysis of users' different types of check-in documents. Our experiment results show that UGCs are useful for improving friend recommendation accuracy.

The last essay is devoted to friend recommendation in health/fitness social networks. Thanks to the rapid growth of smartphone and wearable device technologies, we were able to collect a lot of users' health indicators and physical activity data. Health indicators could imply users' demographic profile, and physical activities reflect users' interests. The analytic framework targeted three types of health/fitness data. The first type includes users' heart rate, sleep patterns, weight, and height. The second type includes different sport data, such as energy consumed, workout frequency, and durations. The last type consists of our proposed activity distribution divergence and Hellinger distance. Our experimental results show that the health/fitness analytic framework helps to improve friend recommendation performance in health/fitness social networks.

This dissertation makes novel contributions to friend recommendation in social networks and has implications for both research and practice. It also opens up new avenues for interesting future research.

CURRICULUM VITAE

Jiaxi Luo

Place of birth: Changchun, Jilin Province, China

Education

B.A., Group T-University College, Belgium, June 2008
Major: Electronic Engineering

B.A., Beijing Jiaotong University, China, June 2009
Major: Computer Science

Master of Science in Computer Science, Group T-University College, Belgium, June 2009

Dissertation Title: Three Essays on Friend Recommendation Systems for Online Social Networks

Conference presentations

Luo, J., Sinha, A., Zhao, H (2013). Location-sensitive Friend Recommendation in Online Social Networks. In the 23rd Workshop on Information Technologies and Systems (WITS
2013), Milan, Italy.
Luo, J., Sinha, A., Zhao, H (2012). The Effect of Location Information on Friend Recommendation in Online Social Networks. In INFORMS International 2012, Beijing China.
Luo, J., Zhao, H (2011). Stranger Recommendation in Location-based Online Social Networks. In INFORMS 2011 Annual Meeting, Charlotte, North Carolina.

Working papers

Luo, J., Sinha, A., Zhao, H. A Novel Friend Recommendation based on User-generated Contents
Luo, J., Sinha, A., Zhao, H. Friend Recommendation on Fitness Social Networking Sites.
Luo, J. Exam Analysis based on 20 years Chinese National College Entrance Examination.
Luo, J. A Novel Context Data Analysis Model for Online Video Advertisement

Academia experience

Assistant Professor:
Spring 2016, Midwestern State University
MIS 3003: Management Information Systems
• Designed course
• Conducted Lectures
MIS 4163: Business Systems Analysis and Design
• Designed a new course
• Conducted Lectures

Fall 2015, Midwestern State University
MIS 3003: Management Information Systems
• Designed course
• Conducted Lectures
MIS 3203: e-Commerce
• Designed course
• Conducted Lectures
MIS 4663: Special Topic: Mobile App Development
• Designed a new course
• Conducted Lectures

Independent Lecturer:
Fall 2014, University of Wisconsin-Milwaukee
Bus Adm 735: Advanced Spreadsheet Tools (Graduate level)
• Designed course
• Conducted Lectures
• Lead Lab sections

Fall 2014, University of Wisconsin-Milwaukee
Bus Adm 530: Introduction to eBusiness (Senior undergraduate level)
• Designed course
• Conducted Lectures
• Lead Lab sections
Teaching Assistant Mentor:

Fall 2014 – Spring 2015, University of Wisconsin-Milwaukee
Bus Adm 230 and 231: Introduction to Information Management
• Conducted discussion sections
• Designed course
• Help Teaching Assistants

Teaching Assistant Lead:
Fall 2013 – Spring 2014, University of Wisconsin-Milwaukee
Bus Adm 230 and 231: Introduction to Information Management
• Conducted discussion sections
• Designed course
• Graded assignments and project
Teaching Assistant:

Fall 2014, University of Wisconsin-Milwaukee
Bus Adm 536: Business Intelligence
• Conducted discussion sections
• Graded assignments and project

Fall 2012 – Spring 2013, University of Wisconsin –Milwaukee
Bus Adm 335: Visual System Development
• Graded programming assignments
• Project assistant

Fall 2010 – Spring 2011, University of Wisconsin –Milwaukee
Bus Adm 210: Introduction to Business Statistics
• Taught statistics knowledge
• Graded assignments and project
• Conducted discussion sections

Fall 2009 – Fall 2010, Fall 2011 - Spring 2013, University of Wisconsin –Milwaukee
Bus Adm 230: Introduction to Information Technology Management
• Lab discussion sections
• SAP simulation game
• ERP research

Professional activities

Member of the Association for Information Systems (AIS)
Member of INFORMS
Member of IEEE
Student Volunteer, Design Science Research in Information Systems and Technology, 2011
Reviewer of Journal:
• The DATA BASE for Advanced in Information Systems (2014)

Professional certifications

2006 Microsoft Certified Professional (MCP)
2006 Microsoft Certified Solution Developer (MCSD)

Honors

Fall 2014 Chancellor's Graduate Student Awards (CGSA) Business Administration Scholarship
Fall 2013- Spring 2014 Chancellor's Graduate Student Awards (CGSA) Business Administration Scholarship