Social Network Analysis on Location-Based Recommender Systems

written by
Bardia Khalesi

Supervisors:
Maarten van Someren
Abdallah El Ali

A thesis submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of MSc
in
Artificial Intelligence

August 2012
Abstract

The growth of social networks in the last few years and their importance in societies lead to directing a large amount of attention to social network analysis on the one hand, and personalized recommendation systems on the other hand. However, most of recommender systems focus on personal profiles of users to estimate the recommendations and do not consider social groups to which users can be related. In this thesis, we aim to investigate the impact of social groups on preference of individuals in social networks, and consequently on the effectiveness of recommender systems. Therefore, we propose a unified framework to incorporate a number of collaborative filtering systems based on different social groups to assess the social influence of each of them on the behavior of users. Experimental results show that social groups play an effective role in making relevant suggestions. Furthermore, people share their information about venues via their comments. This information is spread through the network and used by other users to decide to attend any of available venues or not. Thus, to benefit from this flow of information in social networks, we try to utilize natural language processing methods to extract users’ opinion from their comments. In this manner, our informed recommender illustrates the influence of information cascading and also an admissible evidence of positive/negative relationships’ impact in making more accurate recommendations.
Acknowledgements

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, all praises to the omnipresent God, for answering my prayers for giving me the strength to plod on despite my constitution wanting to give up and throw in the towel, thank you so much Dear Lord.

I would like to express my deepest gratitude to my supervisor, Dr. Maarten van Someren, for his excellent guidance, caring, patience, and providing me with the room to work in my own way. I attribute the level of my Masters degree to his encouragement and effort and without him this thesis would not have been completed or written. One simply could not wish for a better or friendlier supervisor.

Special appreciation goes to my co-supervisor, Mr. Abdallah El Ali, for his supervision and constant support. His invaluable help of constructive comments and suggestions throughout the experimental and thesis works have contributed to the success of this research.

I offer my sincerest gratitude to my wife, Sanaz Najiyan Tabriz. She was always there cheering me up and stood by me through the good times and bad.

I would also like to thank my parents, Mr. Alireza Khalesi and Mrs. Roughiyeh Rassooli Tazangi, and my sister, Arshiya Khalesi. They were always supporting me and encouraging me with their best wishes.

Last but not the least, Sincere thanks to all my friends especially Hossein Kazemi, Karun Rao, and Kamran Massoudi, who as good friends, were always willing to help and give their best suggestions during my study. Thanks for the friendship and memories.

To all my teachers, especially Dr. Gwenn Englebienne, and those who indirectly contributed in this research, your kindness means a lot to me. Thank you very much.
6 Conclusions and Future Work 38

Bibliography 40

Appendix A 44
List of Figures

1.1 Social network graph ................................................. 3
1.2 Closures .............................................................. 6
1.3 Positive and Negative Relationships ................................. 7
3.1 The procedure of sentiment analysis ............................... 20
4.1 The first 60 records of the dataset (red nodes are users and blue ones are venues) ............................................ 23
4.2 Friends vs. Non-friends .............................................. 24
4.3 Local vs. Distant venues ............................................ 25
4.4 The flowchart of the baseline algorithm ........................... 26
4.5 Plot of Mean Absolute Error for the baseline recommender system respect to regularization coefficient ............................................. 27
4.6 The flowchart of the unified collaborative filtering algorithm ........................................... 28
4.7 Plot of Mean Absolute Error for the unified recommender system respect to regularization coefficient ............................................. 28
4.8 Plot of Mean Absolute Error for the informed recommender system respect to regularization coefficient ............................................. 30
4.9 Variation of Mean Absolute Error for the baseline recommender system (MAE=0.7432) ............................................. 32
4.10 Variation of Mean Absolute Error for the social recommender system (MAE=0.6906) ............................................. 32
4.11 Variation of Mean Absolute Error for the informed recommender system (MAE=0.6859) ............................................. 33
Chapter 1

Introduction

1.1 Introduction

The rapid development of web applications and mobile devices has brought about a huge source of data for online social network services to deal with. This, in consequence, leads to delivering a myriad of results in response to users’ queries. Thus, it is very critical to filter the data in a way to avoid overwhelming users by a number of results not appealing to them. In this fashion, recommender systems are efficient tools designed to overcome the information overload problem by providing users with the most relevant content [1].

Basically, traditional recommender systems predict user preferences—often represented as numeric ratings—for new items based on the user’s past ratings on other items (content-based methods or collaborative filtering). In this manner, traditional recommender systems do not consider the social influence at all. However, by revealing the effect of social ties on tastes, preferences, and activities of individuals [2], a number of attempts have recently been made in order to utilize the social network’s features to improve the accuracy and the performance of their recommender systems. For example, Seth exploited the concepts of strong and weak ties in the context of social networks to benefit the effect of relationships on the users’ decisions [2], and He and Chu have done some experiments to assess not only the effect of friends on the users’ interests, but also the influence of friends of friends on this matter [3]. Such sort of techniques, in which social ties are used to predict ratings, is referred as social filtering [4].
In addition to users’ profiles and social ties, social network websites contain other valuable sources of information that can be incorporated into social filtering to improve the performance and accuracy of the recommender systems. To exemplify, most online social networks, like Foursquare\(^1\), capture the geographic location data of users and their point-of-interests (POI) as well. In Foursquare, a user’s location is recognized by the navigational data streams of GPS devices or the location of the network provider. Then, the user can check in predefined venues in his neighborhood to specify his POIs. In accordance with the impact of the users’ context on predicting recommendations [5], recent research has revealed that the geographic feature plays a very effective role in the preference of the users. For instance, Ye et al. have conducted a research on the effect of physical distance between users and their POIs as well as their social relationships on the recommendation process. Their experimental results exhibit a significant improvement in prediction of users’ POIs [6, 7].

Moreover, one of the advantages of online social networks is that users can share their viewpoints with each other by putting comments. Being aware of others’ opinion can be very helpful in decision making [8]. According to Groh and Ehmg, more often, people do not decide totally logically and based on mathematical weighing advantages and disadvantages; they make decision heuristically based on others’ advices [4]. Berjani and Strufe also point out the importance of the comments’ content and information that they can convey in [9]. They illustrate that status messages, like “Had a great party at Club Z”, can express the location s/he was, and the fact that whether the user liked the place. Therefore, by considering the phenomena of information cascading through a social network, and also positive and negative relationship in social networks [10], we can see the importance of analyzing comments in predicting the preference of users.

1.2 Concepts Definition

Social network analysis (SNA) is the study of social relations among a group of people, in which the focus is on relationships between individuals rather than their characteristics. Therefore, the best way of representing social networks is a graph that consists of a set of nodes, corresponding to individuals within the network, and a set of edges (connections or links), which represent relationships between the individuals, such as

\(^1\)www.foursquare.com
friendship and organizational position (e.g. Figure 1.1). Therefore, we can exploit graph theory in analyzing social networks to define metrics and concepts, namely the clustering coefficient (e.g. the similarity weight) and cliques (e.g. closures—defined in section 1.2.2). Moreover, since the structure of networks in human societies have been changed due to being provided with novel means of communication, such as the World Wide Web, huge online social networking services have been discovered to be the most accurate medias reflecting the real-life social relationships of people [11]. In this manner, another advantage of graph-based methodologies is the fact that they are powerful and effective in analyzing of huge online social networks.

With regard to the above, in this section, we describe a number of concepts and phenomena in the field of the social network analysis helpful to our work.

1.2.1 Homophily

The homophily principle describes the tendency of individuals to be similar to their friends [10]. Easley et al. argue that people tend to socialize and team up with similar others. To illustrate, a group of friends are, typically, similar in age, their interests, beliefs, and opinions. Even though they may have some connections out of this group, those links also relate each of them to other groups of people with whom they have other characteristics in common. Therefore, the homophily principle explains the formation of networks and the effect of networks’ contexts on their structures, for example, when
two persons become friends through a common friend, or because of attending the same venues (e.g. school, party). Easley et al. also state that there are strong interactions between intrinsic and contextual effects on the formation of any single link. For this reason, people may modify their behaviors to bring them more closely into alignment with the behavior of their friends [10].

1.2.2 Closures

*Clique* is one of the essential definitions of graph theory in analyzing the topological characteristics of online social networks. Theoretically, clique is a subset of an undirected graph’s vertices, in which every two vertices in the subset are connected by an edge. Therefore, *closure* is referred to the creation of an edge between two vertices that brings about a new clique in the graph (network).

In regard to the homophily principle, three kinds of closures are defined in social network analysis that explain the possibility of creation of connections between individuals by virtue of establishing trust between people with different kinds of relationships. These are as follows:

**Triadic Closure** One of the most basic principles, derived directly from the homophily principle, is *triadic closure*. Triadic closure states the following:

> There is a higher likelihood for two persons in a social network to become friends when they have a common friend with each other [10].

In other words, the inclination (the clustering coefficient / the similarity weight) of two persons will be higher to make friends when they have more common friends. This is very intuitive because, first of all, when they have a common friend, there is a better *opportunity* for them to meet each other such as attending a party held by their common friend. Secondly, having a common friend makes people *trust* each other more since they both trust the friend. Finally, there can be an *incentive* for the common friend to bring his friends together, for example, to do an activity more efficiently like a business [10].
In this work, we use the possibility of making friendship between two users via evaluating the similarity between them (as the amount of trust) to measure the effect of friends-of-friends on users' ratings.

**Membership Closure** Another interpretation of the homophily principle can be stated as follows:

> It is more probable that a person becomes involved with a particular activity when a friend of him is already involved in it [10].

This means that, for example, the target user will visit locations checked in by his friend(s) with higher probability than other arbitrary places because he trusts his friends more. This is the most fundamental idea of our social filtering algorithm because by taking any real or possible connection between two users into account, the probability of making this closure has to be predicted in making venue recommendations.

**Focal Closure** By taking the two last principles, we can deduce the third principle as follows:

> When two persons have a number of activities in common with one another, there is a higher possibility that they will make friends at some point in the future [10].

Of course, multiple effects can operate simultaneously on the formation of a single link. Therefore, by considering activities (e.g. venues) as another kind of actors in the graph, we generally can say that the probability of the formation of a link between two persons in a social network will be increased by increasing the number of common neighbors [10]. You can see the schema of closures in figure 1.2.

Thus, we can define focal group as a group of individuals with whom the target user can form a connection to make a focal closure. Moreover, We calculate the probability of forming this connection based on the similarity between the target user and this group of individuals to assess their influence on his decision making.
1.2.3 Information Cascade

Easley et al., in [10], argue that people sometimes make decisions based on observing others’ actions regardless of their own private information. Easley et al. maintain that because of an underlying human tendency to conform and information inferred from others’ behavior, they may choose to abandon their private information and follow a crowd [10]. To illustrate, in the case of choosing a restaurant, if there are two restaurants next to each other, and one of them is more crowded, one can reach the point that these people know something about its quality and choose that restaurant as well. In such situations, it is said that an information cascade has occurred.

However, in this way, people may not make the best decision, or even decide wrongly. Easley et al. point out that cascades can be wrong or be based on very little information [10]. Yet, people can disprove them as soon as they gain stronger information. Therefore, cascades are also fragile [10].

All in all, information cascade is one of the features of social networks that brings about social influence on individuals’ behaviors.

1.2.4 Positive and Negative Relationships

In social networks, we can consider relationships either positive (e.g. friendship) or negative (e.g. opposition). While negative relationships can be effective as a social influence as well as positive relationships, the majority of research has focused just on positive relationships. According to Easley et al., in [10], the notion of structural balance is one of the basic frameworks for understanding the tension between these two kinds of relationship. Correspondingly, this model is defined in a clique where everyone is aware
of (connected to) others. The crucial idea—based on theories in social psychology—is that by considering a clique with three vertices, the triangle is referred as balanced if it has one or three +’s, while the one with zero or two +’s is referred as unbalanced [10]. To exemplify, when two persons are enemies, but they have a common friend, there would be an incentive for their friend to try to get those persons to become friends or to turn against one of them in favor of the other one. Figure 1.3 depicts these four plausible configurations of +’s and -’s relationships. This principle is true in other situations as well. For example, when someone had a bad experience of participating in an event, but one of his friends got a positive impression from there, the state is unbalanced. However, that friend can change his mind and persuade him to go there again by virtue of existing a trust relation between them, or vice versa. Hence, the state will become balanced when one of them change his mind about that event. That is, we investigate how a recommender system can take advantage of this property of social networks’ graph to boost its performance. In this work, to identify the polarity of relationships between users and venues, we analyze opinions expressed by users, in their comments, about locations they have visited.

1.2.5 Informed Recommender System

According to [12], informed recommender is a recommendation system that bases recommendations on consumers product reviews. To do this, it requires to extract users’
opinions about products from their comments expressed in free-form text to generate product recommendations [12].

1.3 Problem definition

In this thesis, we address two features of social networks of which the effects on the quality of ratings prediction have not been investigated in any personalized recommender systems based on social filtering.

First and foremost, according to social network analysis, in addition to social groups of friends and friends of friends, there exists another effective group in social networks, called focal group, in which there are individuals who do not have any connection and any common friend with each other, but they have more than a certain number (say two) of common POIs together [10]. This factor has been considered implicitly in collaborative filtering systems when the system finds similar users based on common checked-in locations. Although, in calculating the similarity between users, the importance and the effectiveness of members of this kind of social group are considered the same as friends. Thus, our first contribution is to assess the level of the influence of each of these social groups explicitly on the preference of the target user.

Secondly, we investigate the effect of information cascading in social networks on the performance of recommender systems. Therefore, we extract information about users’ feelings and opinions toward places they have checked in by mining the content of their comments. Then, we consider opinions of individuals related to the target user (members of any of the social groups) in predicting his ratings on the target venues. To do this, we exploit natural language processing techniques to analyze the sentiment of users’ comments to classify these comments into two classes of positive and negative ones [13]. In this way, we equipped our method with text-mining techniques to extract useful information from review comments to improve its accuracy and performance by virtue of being more informed about various contexts of the social network [12].

Besides, it should be mentioned that we have obtained our dataset from the website of Foursquare (one of the most representative location-based social networks [7]).
Chapter 1. *Introduction*

The rest of this paper is structured as follows: In chapter 2, we discuss previous work done in this field of study. In chapter 3, we discuss our approach to address the questions of this paper. Next, in chapter 4, we describe our experiments and represent their results. After that, we discuss the obtained results and what we have achieved through our experiments in chapter 5. Finally, we conclude this paper, in chapter 6, and suggest some promising directions for future work.
Chapter 2

Related work

Traditional recommender systems are mainly based on the collaborative filtering techniques that use the user-item matrix to find similar users and items for making the predictions. In this way, there are two major categories of collaborative filtering methods: memory-based and model-based. Memory-based methods are also divided into two subcategories; some of the memory-based methods, known as user-based, focus on the similarity between users to predict user preferences to items [3, 6, 7, 14], while others are item-based, in which predictions are made by computing a weighted combination of user ratings on similar items [15]. Additionally, there are a number of methods proposed to integrate these two types of methods in order to take advantages of either of them [16].

On the other hand, model-based approaches employ statistical and machine learning techniques, like Bayesian networks [3] and the latent factor model [17], to learn patterns of rating behaviors of users.

In addition, in the context of social networking systems, the principle of Homophily is one of fundamental notions describing how a network’s structure is formed based on its surrounding contexts, and how its connections affect users’ preferences and activities by cascading information through the network [10]. Accordingly, research has shown the positive effect of social connections on the performance of recommender systems that are based on collaborative filtering methods. To exemplify, Ma et al. in [17] use the group of friends to predict the preference of the target user; He and Chu consider the influence of friends of friends as well as friends’ in order to gain a more accurate results [3]. However, since they apply a model-based approach, they could not evaluate the
influence of each of these social groups separately, and they just interpreted these effects generally from the difference of the final results. Plus, the coverage of their proposed method is lower than of the traditional CF as a result of ignoring individuals out of the mentioned social groups related to the target users.

Moreover, geographical context of location-based online social networks is another factor that has recently been applied in recommender systems. As an example, Barjani and Strufe propose a model for a location recommender system that utilizes the check-in patterns of users to make personalized predictions [9]. Yet, in this work, Barjani and Strufe consider locations as conventional items, and they assess the correlations between places regarding users’ activities instead of the geographical specifications of venues [7]. In other words, they calculate the users’ ratings based on users’ check-in frequency on locations, and then they use a content-based CF to make recommendations. However, Ye et al. attempt to exploit Geo-spatial characteristics of data in their recommender system. In [6], they propose a method, in which the similarity weight between friends is computed by the physical distance between them. Nevertheless, they do not obtain promising results because the distance between friends’ location can be very arbitrary, and they overlook a part of their dataset due to eliminating non-friend users from their dataset. Then, Ye et al., in [7], introduce a unified collaborative that incorporated geographical influence along with user preference and social influence (that are, practically, traditional CF and Friend-based CF respectively), while, in this work, the geographical similarity weight is calculated by the distance between venues. In this manner, user preference and social influence overlap with one another that this leads to doubling the influence of a number of individuals on the prediction of ratings.

Furthermore, whereas the wealth of information included in users’ comments has been disregarded by most recommender systems in social networks [8], deducting information from free-text reviews, by which recommender systems can be fortified, has attracted scholars’ attention in the field of Information Retrieval (IR) and Natural Language processing (NLP). In this matter, the first phase is to analyze sentences to retrieve expressed opinions of users, such as in [18, 19]. Then, based on this analysis, users’ ratings are estimated, like in [20]. In the second phase, the obtained ratings should be applied to a collaborative filtering system. To the best of our knowledge, the first work which integrates opinion mining with a recommendation system is [21]. In this work, the authors
use a relative-frequency-based method to identify features of movies’ reviews. Unfortunately, they do not describe the method employed for CF. In addition, since comments in social networks are very short, classification algorithms based on term frequencies do not provide satisfactory results [12]. Another case study has been presented in [12]. In this work, the recommender makes suggestion just based on reviews and ranked them by considering the level of users’ skills. Therefore, they do not take account the effect of social relationships at all. Moreover, instead of extracting users’ preferences automatically, they make users search their product by making queries. Additionally, Jakob et al., in [8], show improvement in the accuracy of movie recommendations by extracting movie aspects as opinion targets and use them as features for the collaborative filtering. Yet, since they use a traditional CF, their work still has the corresponding defects. However, in [22], authors focus on improving recommendation accuracy in a restaurant review scenario. Their experiments show that using textual information results in better general or personalized review score predictions than those derived from the numerical star ratings given by the users. But again, in the personalized recommender, Ganu et al. do not considered the pertaining social groups and just specify the five closest person through a KNN algorithm in their CF.

Succinctly, all of the mentioned approaches above suffer from defects in their computations because of missing or overlapping part of their knowledge about the corresponding social groups in their social networks.

Hence, our study can be distinguished from all prior works in three aspects: Firstly, besides the effect of friends, friends of friends, and the distance between users and venues, we evaluate the influence of a new group of individuals in social networks—called focal group—on the process of predicting users rating, which has not been considered in previous works explicitly. Secondly, we split the data into three discrete groups of friends, friends of friends, and focal for the target user and identify the effect of each of these social groups particularly on his/her ratings on the target items. Thirdly, since social networks are not specialized in particular topics, focusing on one subject, say restaurants or cinemas, is not useful for their recommender systems. Hence, in this work we train our recommender by labeled comments about the most prominent topics in Foursquare to have a more informed recommender in order to augment the accuracy of the recommendations.
Chapter 3

Approach

In this chapter, we provide background information on some of the methods and concepts pertinent to our approach, and then we give a more detailed description of our approach.

Our approach investigates the effect of different social groups on recommendations beside the influence of geographical and textual context of social networks. To do this, it is needed to apply a method that gives us the capability of computing the similarity weight between pairs of individuals for each of the features separately. In this matter, user-based methods of collaborative filtering seem to be the applicable approach since the similarity weight calculated in item-based methods cannot reflect the social ties between users, and model-based methods require considering all the features together as dependent probability and latent variables to make the predictions. Therefore, we use and extend the method proposed by Ye et al. in [7]. Their method has three major parts: One part is traditional User-based collaborative filtering (see section 3.1) using cosine similarity between users. The second part is Friend-based collaborative filtering (FCF) (see section 3.2), called social influence and in which Ye et al. introduced a similarity weight between friends based on the number of their common friends and point-of-interests (POIs). We will exploit this weight of similarity in our approach to calculating the similarity for friends-of-friends and focal groups of each user. And the last part is Geo-location-based collaborative filtering (GCF) (see section 3.3), in which they came up with a power-law probabilistic model to calculate the similarity between individuals’ POIs based on their observations in their dataset. As our dataset analysis
Chapter 3. Approach

has revealed that users’ POIs have a power-law distribution, we use this model in our approach as well.

Additionally, we employ text-mining techniques (see section 3.5) to extract users’ opinions from their free-text comments on venues and to identify if they are positive or negative statements [8]. Then, using the concept of positive/negative relationships in networks, we specify the sign of weights between users and venues\(^1\) and predict the rating score that a user is likely to give to the target venues. Thus, consequently, we will be able to make more judicious recommendations.

As mentioned previously, our work is based on data from the website of Foursquare. The dataset has been collected through a snowball sampling method to maintain the structure of the network (see section 4.1). Besides, due to the lack of location ratings in Foursquare, we are just able to figure out whether a user visits a venue or not by searching through his profile information. Also, since we do not have access to the real check-ins lists of users, we cannot rely on the frequency of check-ins done by them. Thus, we use a 2-point rating to represent users’ ratings in the first approach [6]. We denote 1 to the checked in venues and 0 otherwise. However, in the second approach, since we can assay users’ attitude toward a venue by analyzing sentiment of their comments, we use a 3-point rating (\([-1, 0, 1]\)) to represent whether a user has checked in the venue and s/he likes or dislikes it, or s/he has not checked in there before.

3.1 User-based Collaborative Filtering

User-based Collaborative Filtering (UCF) is a method for making automatic predictions (filtering) based on the idea that users’ preferences can be determined by investigating the preference of similar users. Therefore, in the first step, UCF should find similar users. To do this, UCF defines the similarity weight between a pair of users by adopting one of the similarity measures such as vector cosine. For example, the similarity weight between users \(i\) and \(k\) is calculated as follows.

\[
w_{i,k} = \frac{\sum_{l \in L} r_{i,l} r_{k,l}}{\sqrt{\sum r_{i,l}^2} \sqrt{\sum r_{k,l}^2}}
\]  

\(^1\)The sign of relations between users is always positive since there is just defined the friendship relations in social networks
where $L$ is the set of items rated by both of the users, and $r_{i,j}$ represents the rating of the user $i$ on the item $j$. In other words, the similarity between users is obtained by measuring the angle between check-in vectors of the given pair of users.

After that, UCF estimates the rating of the target user on the target item based on the ratings of the top-K similar users in the neighborhood of the target user. This is done via the weighted sum of the ratings of similar users as follows.

$$\hat{r}_{i,j} = \frac{\sum_{u_k \in \hat{U}_i} r_{k,j} w_{i,k}}{\sum_{u_k \in \hat{U}_i} w_{i,k}}$$

(3.2)

where $\hat{U}_i$ consists of the top-K similar users in the neighborhood of the user $i$. In the rest of this paper, we will refer to this algorithm as the “traditional CF”.

### 3.2 Friend-based Collaborative Filtering

As we mentioned before, according to the homophily principle, friends tend to behave similarly. As we will see in the next chapter, our data analysis illustrates that friends have more common location than non-friends (Figure 4.2).

For this reason, groups of friends can be exploited as a valuable source of good recommendations for CF systems [7]. Therefore, it is needed to adapt CF methods to discover the most interesting common items among users’ friends in social networks. In this fashion, friend-based collaborative filtering (FCF) is defined, in [7], as a user-based collaborative filtering, in which the similarity weight is computed based on ratios of common friends and venues instead of the cosine similarity. The formula is as follows.

$$w_{i,j} = \alpha \frac{|F_i \cap F_j|}{|F_i \cup F_j|} + (1 - \alpha) \frac{|L_i \cap L_j|}{|L_i \cup L_j|}$$

(3.3)

where $\alpha$ is a tuning parameter, ranging within $[0, 1]$, and $F_k$ and $L_k$ denote the friend set and POI set of user $u_k$ respectively [7]. In our approach, we apply this similarity weight for groups of friends of friends and focals. However, in case of estimating the similarity between friends, according to Akcora et al., this way of calculating the similarity weight ignores the friendship relationship between the target user and his friends [23] due to the fact that users are not the member of their own friends list; so, when we join their list of friends, the direct edge between them will be overlooked. To solve this problem,
the simplest solution can be adding one unit to the number of common friends. So, our formula for evaluating similarity between pairs of friends would be as follows.

\[
    w_{i,j} = \alpha \frac{|F_i \cap F_j| + 1}{|F_i \cup F_j|} + (1 - \alpha) \frac{|L_i \cap L_j|}{|L_i \cup L_j|}
\]

(3.4)

where the first term is called \textit{common friends ratio} and the second term is \textit{common location ratio}.

From the viewpoint of graph topology, since the graph is undirected and so the friend relation is symmetric, the numerator of common friends ratio refers to the number of paths between the pair of friends, and the denominator is the count of all possible paths, through which the target user can reach his friend.

Besides, we evaluate the parameter \( \alpha \) through analyzing our dataset (see section 4.2). Yet, in case of the focal group, we consider \( \alpha = 0 \) since people, in this group, do not have any common friend with the target user.

Eventually, the prediction will be computed for each social group in the same way of the user-based CF, via Equation 3.2.

### 3.3 Geolocation-based Collaborative Filtering

As mentioned earlier, spatial analyses on real datasets have revealed a \textit{geographical clustering phenomenon} in user check-in activities [7]. This means that people usually incline to visit places close to their location (e.g. see section 4.2). Therefore, distance is an influential factor in venue recommendations. These experiments depict that the check-in probability follows a power-law distribution\(^2\). In this manner, Ye et al. proposed a model, in [7] as follows, to compute the likelihood that the user \( u_i \) would check in POI \( l_j \).

\[
    Pr[l_j|L_i] = \prod_{l_y \in L_i} a \times d(l_j, l_y)^b
\]

(3.5)

where \( L_i \) denotes the visited POI set of the user \( u_i \), and \( l_j \in L - L_i \). In addition, \( d(l_j, l_y) \) refers to the distance between \( l_j \) and \( l_y \). Moreover, \( a \) and \( b \) are parameters of the power-law distribution that are computed via a linear regression method\(^3\).

\(^2\)\( y = a \times x^b \)

\(^3\)Least square error method [24] (please go through [7] for more details)
Chapter 3. Approach

3.4 Social Collaborative Filtering

In this section, we explain our social CF algorithm, which is inspired by the fusion framework proposed by Ye et al. in [7]. In this way, we also use a linear combination method to integrate ratings assessed by the above-mentioned recommender systems to estimate the target user’s rate on the target item more precisely. This integration can lead to increasing both coverage (recall) and precision of the system by virtue of considering more factors that are influential in users’ decision making.

By denoting $S_{s,i,j}$ and $S_{g,i,j}$ as the check-in probability scores of the user $u_i$ at a POI $l_j$, corresponding to CF method based on the social and geographical influence, respectively [7], we can calculate the final rate by using a linear combination as follows.

$$S_{i,j} = \alpha \times S_{s,i,j} + \beta \times S_{g,i,j} \quad (3.6)$$

where the weighting parameters $\alpha$ and $\beta$ indicate the relative importance of each of these influential factors in estimating users’ preferences [7]. Moreover, we define the social influence as a combination of the influence of the three social groups of friends, friends of friends, and focals. Thus, we can define $S_{s,i,j}$ as follows.

$$S_{s,i,j} = \alpha \times S_{fr,i,j} + \lambda \times S_{fof,i,j} + \gamma \times S_{foc,i,j} \quad (3.7)$$

To find these coefficients, we apply ridge regression method (regularized least squares) in order to avoid over-fitting [24].

Accordingly, to compute these check-in probability scores, first, we estimate the check-in probabilities for the list of the target items (e.g. venues) based on each of the features by applying the appropriate collaborative filtering methods (e.g. Equations 3.2 and 3.5). After that, in the case of each of the features, we divide the obtained probabilities by the highest one, as the normalization term, to compute the corresponding scores. To clarify, consider $p_{g,i,j}^q$ as the probability—based on the geographical influence—that the user $u_i$ may visit a POI $l_j$ [7]. Hence, we can write it as follows.

$$p_{i,j}^q = Pr[l_j|L_i] = \prod_{l_y \in L_i} a \times d(l_j,l_y)^b$$
So,

\[ S_{i,j}^g = \frac{p_{i,j}^g}{Z_i^g}, \text{ where } Z_i^g = \max_{l \in L-L_i} \{p_{i,j}^l\} \]

The pseudocode is represented in algorithm 3.1.

**Algorithm 3.1** Unified Collaborative Filtering

**Require:** users must have at least one Mayorship

```plaintext
users_list & venues_list ← Read(dataset)
for user_i ∈ users_list do
   for venue_j ∈ target_venues_list do
      p_{i,j}^{fr} ← FCF(user, list of friends, venue)
      p_{i,j}^{fof} ← FCF(user, list of friends of friends, venue)
      p_{i,j}^{foc} ← FCF(user, list of focals, venue)
      p_{i,j}^g ← geo_CF(user, venues_list, venue)
      S_{i,j}^{fr} ← \frac{p_{i,j}^{fr}}{Z_i^{fr}}, \text{ where } Z_i^{fr} = \max_{l \in L-L_i} \{p_{i,j}^{fr}\}
      S_{i,j}^{fof} ← \frac{p_{i,j}^{fof}}{Z_i^{fof}}, \text{ where } Z_i^{fof} = \max_{l \in L-L_i} \{p_{i,j}^{fof}\}
      S_{i,j}^{foc} ← \frac{p_{i,j}^{foc}}{Z_i^{foc}}, \text{ where } Z_i^{foc} = \max_{l \in L-L_i} \{p_{i,j}^{foc}\}
      S_{i,j}^g ← \frac{p_{i,j}^g}{Z_i^g}, \text{ where } Z_i^g = \max_{l \in L-L_i} \{p_{i,j}^l\}
   end for
weighting_parameters ← linear_regression(S_{i,j}^{fr}, S_{i,j}^{fof}, S_{i,j}^{foc}, S_{i,j}^g)
S_{i,j} ← calculate ratings by Equations 3.6 and 3.7
```

where, generally, CFs can be demonstrated in algorithm 3.2.

**Algorithm 3.2** Collaborative Filtering

```plaintext
target_user related_users_list & target_venues_list are given
for venue ∈ target_venues_list do
   for user ∈ related_users_list do
      w_{i,k} ← similarity weight between the target user and user
      if venue ∈ user’s check-ins then
         r_{k,j} ← 1
      else
         r_{k,j} ← 0
      end if
   end for
\hat{r}_{i,j} = \frac{\sum_{u_k \in U_i} r_{k,j} w_{i,k}}{\sum_{u_k \in U_i} w_{i,k}}
```

Although the order of complexity of this algorithm is \(O(n^2)\), it can be scaled down by being executed in parallel with smaller numbers of target venues for each run.
3.5 Sentiment Classification

Sentiment analysis is one of the application of natural language processing that aims to classify a given text by extracting subjective information from it and determining whether the overall opinion of the writer is positive or negative toward an entity [25]. In this way, we can gain very useful information about the opinions of users about venues in Foursquare since it does not have any rating system to represent how users rate venues.

However, identifying structured information from free-text comments is a challenging task in social networks for several reasons [22]. First and foremost, comments are written in an informal language associated with a lot of grammatical and spelling errors. Secondly, comments are very short sentences since users try to write briefly usually through mobile applications. Thirdly, particular domains have their particular languages; hence, learning specific phrases in social networks, like Foursquare, that include different domains is very difficult. Last but not least, comments are written in different languages. In addition, due to the limitation of language tools, especially in mobile phones, comments are written in different languages but with English alphabets just in the way that words are pronounced in the original languages. Although there are a number of words in languages that is used the same or have the same roots. For instance, the words “excellent” in English and “excelente” in Spanish have the same meaning and root. Even in languages with different roots like English and Farsi, there are a number of words with the same meaning such as the word “bad”.

Therefore, we need to employ feature-level (i.e. word-level) opinion mining on free-text comments to identify the polarity of opinionated user statements [8] in Foursquare. In addition, in this paper, we mostly focus on comments written in English, yet our dataset also contains a number of comments in Dutch, and Spanish. To do this, there are two prominent supervised machine learning methods, which can be applied, to sentiment classification [13, 25]: Naïve Bayes (NB), and support vector machines (SVM) [24]. According to previous research, these two classifiers perform almost the same [25], sometimes SVM outperforms NB [13]; however, in some attempts, NB has shown better result than SVM [26].

In this matter, we map the classified comments into the corresponding ratings, and then, we generate predictions by incorporating this information into our recommendation
system [8]. In order to incorporate sentiment analysis into our social collaborative filtering, for simplicity, we have decided to perform Naive Bayesian classification [24] to classify users’ comments based on their sentiment.

In the first step, we require to annotate part of users’ comments manually as training and testing sets. We treat each sentence as a document and represent features by the most prevalent single word terms or unigrams occurring within the sentences [26]. In this manner, we eliminate stopwords (uninformative words) from the sentences. However, we do not remove negation words, like “never” and “not”, from the text of the comments to preserve the originality of the sentences and their sentiments. After that, we perform stemming and then tokenize the sentences to extract the feature vectors [27]. Then, we train our classifier by our annotated comments to identify the polarity of users’ opinions in their comments. The procedure is visualized in Figure 3.1. Therefore, by considering

![Figure 3.1: The procedure of sentiment analysis](image)

the class of each comment as the user’s rating, we recalculate the rating of the target user on the target venues by applying our social collaborative filtering one more time.
Therefore, by incorporating sentiment analysis into the social collaborative filtering, we will have an informed recommender system.

### 3.6 Evaluation Metric

There are a number of types of measures for evaluating the quality of a recommender system [15]. One of statistical accuracy metrics is Mean Absolute Error (MAE) between ratings and predictions, which is employed widely to measure the predictive accuracy of recommendation algorithms [3, 9, 15, 17, 28]. MAE is an average of the absolute errors between recommendations and their actual values specified by users. Formally, MAE is computed as follows.

\[
MAE = \frac{\sum |S_{i,j} - r_{i,j}|}{N}
\]

where \(r_{i,j}\) denotes the rating user \(i\) gave to item \(j\), and \(N\) is the number of testing instances [17]. The smaller the MAE, the better the inference [3].

We use MAE to evaluate the performance of our proposed approach in comparison with the base-line method.
Chapter 4

Experiments

In this chapter, we describe several experiments to evaluate the performance of our social recommendation approaches. In the following, Section 4.1 describes our dataset and the crawler by which we have collected the dataset. In Section 4.2, we analyze our dataset to get insight about how users and locations are associated with each other in the social network [6]. Section 4.3 describes the experiments addressing the questions of this thesis.

4.1 Data Collection

We crawled the website of Foursquare to obtain our dataset. To build up our crawler, we used the API provided by Foursquare. Besides, we utilized the snowball sampling method to collect data since research has shown that, topologically, social networks are scale-free networks (graphs) and so their degree distribution follows a power-law [11]. In this matter, a method of sampling is required that can conserve this structural characteristic of social networks. The snowball sampling method is the most efficient method to crawl online social networks since it moderates the degree exponent and brings about heavier tail by virtue of picking up the high-degree nodes [11]. In this method, the crawler is given a hub as the initial seed node, and then it performs a breadth-first search. The crawler will stop by the time of exceeding a certain predefined number of levels.

Correspondingly, our dataset consists of 71,028 users and 537,129 venues. Each record of users contains the user’s id, address, and lists of friends, mayorships, dones, and
tips. Plus, each record of venues contains the venue’s id, address, list of visitors and comments. However, we could not obtain the complete list of users’ check-ins due to privacy issues. Hence, our dataset suffers from a high level of sparsity. By defining the parameter of sparsity for a user-item matrix as $1 - \frac{\|\text{nonzero entries}\|}{\|\text{total entries}\|}$ [15], we figured out that the sparsity level of our dataset is 99.93%.

To illustrate, we visualized a part of our dataset (given the large size of the dataset, we were only able to visualize the first sixty records) by means of Gephi [29] (an open source software for graph and network analysis) that includes 6648 nodes and 8164 edges where 89.7% of the nodes are user nodes (users and their friends) and the rest 10.3% of nodes are venue nodes. Figure 4.1 exhibits the visualization of this part of the dataset. To alleviate this sparsity, we considered lists of mayorships, dones, and venues of the users’ comments as their POIs.

4.2 Data Analysis

Analyzing the dataset showed that users share almost no common visited place with each other in our dataset. However, it was a bit better in case of friends (this is shown
in figure 4.2). As you can see in this figure, more than 90% of friends share no common location with one another, and the rest (less than 10% of friends) just have ten percent of their checked-in venues in common. Thus, the number of social friends who contribute to the prediction via their common checked-in locations is much smaller than what was expected. In figure 4.2, common location ratio (clr) between users $u_i$ and $u_j$ is a metric to measure the degree of overlap in visited locations, which is defined as $\text{clr}_{ij} = \frac{|L_i \cap L_j|}{|L_i \cup L_j|}$ [6]. Therefore, to calculate the similarity weight between users in friend-based collaborative filtering (see section 3.2), the tuning parameter, in Equation 3.4, is valued by $\alpha = 0.9$ to put more weight on common friend ratio.

Nevertheless, the analysis on the physical distance between venues that a user has checked in revealed the fact that users mostly intend to participate in venues within a short distance i.e. in their neighborhoods [7]. As it can be seen in figure 4.3, a significant percentage of check-ins happens in less than 10 km from each other. On the other hand, there is another peak on the right part which shows a high number of check-ins. A reasonable explanation for this part of the figure can be that when people go on trips, they check in a number of locations [7]. Though, from the left part, it can be inferred that they may also check in locations close to where they stay during their vacations. According to Ye et al. in [7], this analysis indicates the strong influence of geographical feature of POIs. Beside this analysis, we calculated the parameters of power-law distribution of users’ check-ins via performing linear regression upon our dataset. As the...
result, we got $a = 0.31805$ and $b = -0.09495$ for Equation 3.5, which will be applied in our experiments to estimate the probability of users’ check-in activities.

4.3 Experiments

In this section we give details about our experiments. The goal of our experiments is to evaluate the effect of social groups and information cascading in the network on the accuracy of personalized recommendations. Wherefore we construct a simple recommender system as our baseline to be compared with our proposed approaches. Then, we build our social recommender system to figure out the importance of each social group in making a recommendation. Lastly, we exploit sentiment analysis to have an informed recommender system to estimate ratings of users on venues.

To evaluate the performance of these recommender systems, for each user in the dataset, we remove all locations in which s/he has mayorship, and then have recommendation algorithms predict the ratings of the user for those locations.
4.3.1 Baseline Recommender System

Our baseline consists of two components; one is the traditional collaborative filtering (see section 3.1) and the other one is a collaborative filtering based on geolocation\(^1\) (see section 3.3).

As it is depicted in Figure 4.4, first, each of these parts computes check-in probability scores independently for all users in the training set. Thus, primarily, we can evaluate their performance by calculating MAE metric for each of them. This experiment shows that the performance of the traditional collaborative filtering was very weak ($MAE = 0.9902$), however, geolocation-based collaborative filtering performed very well ($MAE = 0.5679$). Having calculated rating scores of each part, we estimated the importance coefficients $\alpha$ and $\beta$ in Equation 3.6 in order to integrate these two parts linearly. To do this, we exploited ridge regression method, implemented in \textit{mlpy} package [30]. This method has a regularization coefficient that must be specified manually. Since ridge regression method is a \textit{quadratic regularizer} [24], it has just one extremum point. So, we estimated this coefficient by tweaking. Figure 4.5 shows that to have minimum error

\(^1\)We use Google map service to convert addresses into geographical points (i.e. latitude and longitude) and to calculate the distance between pairs of POIs.
for the performance of baseline recommender the value of the regularization coefficient should be set to 0.245. Consequently, we got $\alpha = -1.6174$ and $\beta = 0.6309$.

![Figure 4.5: Plot of Mean Absolute Error for the baseline recommender system respect to regularization coefficient](image)

Figure 4.5: Plot of Mean Absolute Error for the baseline recommender system respect to regularization coefficient

After calculating final ratings scores, we compared obtained ratings with the actual ratings of users and achieved $MAE = 0.7433$.

### 4.3.2 Social Recommender System

In this phase, we investigate the performance of our proposed a unified collaborative filtering, explained in section 3.4, and of each part of it individually to assess the effectiveness of social groups explicitly. Figure 4.6 gives a clearer picture of this process.

Primarily, we examine components’ performance separately. Inasmuch as the geolocation-based collaborative filtering is exactly the same as the one in the baseline, its performance is equal to what we have obtained in the baseline experiment. Surprisingly, social filtering components performed very weak. while Friends-based CF’s performance was very weak ($MAE = 0.9876$), it still perform better than others. To illustrate, CF based on Friends-of-Friends could not predict ratings precisely ($MAE = 0.9901$). Plus, although CF based on focals did not perform strongly ($MAE = 0.9961$), this experiment has showed the impact of focals group in making recommendations.
Besides, since social filtering includes three CFs based on three different social groups related to individuals in a social network, we had to estimate four importance coefficients $\alpha$, $\lambda$, $\gamma$, and $\beta$ in Equation 3.7, corresponding to CFs based on Friends, Friends-of-
We do this by means of ridge regression method.

Next, with regularization coefficient of 0.44 (see figure 4.7), we got $\alpha = -0.6834$, $\lambda = -0.6929$, $\gamma = -0.7713$, and $\beta = 0.7913$, and, as a result, the error of our unified CF turned to be better than of the baseline ($MAE = 0.6906$).

In addition, in order to have a clearer idea about the effect of social filtering’s components, we retried this experiment while each time we eliminated one of the components based on Friends-of-Friends and focals, and once both. The obtained results are represented in Table 4.1 From these results, we can see that both components improve the performance, especially the one based on Friends-of-Friends.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Friends+Geolocation)-based CF</td>
<td>0.7554</td>
</tr>
<tr>
<td>(Friends+Focals+Geolocation)-based CF</td>
<td>0.7148</td>
</tr>
<tr>
<td>(Friends+Friends-of-Friends+Geolocation)-based CF</td>
<td>0.7119</td>
</tr>
</tbody>
</table>

Table 4.1: Obtained MAEs by eliminating social filtering components

4.3.3 Informed Recommender System

As it is mentioned in section 3.5, to analyze the sentiment of users’ comments, primarily, we labeled 1281 sentences as positive and 819 ones as negative. Then, we assigned 75% of each class of comments to construct the training set and the rest (25% of positive and negative comments) for the testing set. Next, we were supposed to prepare the comments. One of the preparation steps was removing very frequent and informative words from the sentences. We retrieved the list of such words from the website of “ranks.nl”\(^2\). In the next step, we had to reduce words to their stems (stemming) and then identify each word as a token of the feature vector. To implement this, we used natural language toolkit (\textit{nltk}) [27]. Finally, we exploited the Naive Bayesian classifier, implemented in this package, to indicate the polarity of comments’ sentiments. In this way, we trained the classifier on the training set, and then we evaluated its accuracy by the testing set. This test showed that the classifier was 90% accurate (100 of the most informative words are listed in appendix 4).

\(^2\)http://www.ranks.nl/resources/stopwords.html
Chapter 4. Experiments

Figure 4.8: Plot of Mean Absolute Error for the informed recommender system respect to regularization coefficient

After figuring out users’ ratings from their comments on venues, we had our informed CF predict users’ ratings. This time, Friends-of-Friends-based component slightly outperformed others; it had $MAE = 0.9929$ while Friends-based CF and Focals-based had $MAE = 0.9976$. Geolocation-based error still was the same in the last experiments.

Then, we calculated corresponding importance coefficients via ridge regression with regularization coefficient of 0.43 (see figure 4.8) such as $\alpha = -0.7039$, $\lambda = -0.7107$, $\gamma = -0.7744$, and $\beta = 0.7933$. As the result, informed CF had a better performance ($MAE = 0.6859$) than the social CF.

Furthermore, we repeated the components examinations and got results as follows in Table 4.2. One more time, this confirms the benefit of considering the effect of social groups and also information cascading on users’ check-in behavior.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Friends+Geolocation)-based CF+SA</td>
<td>0.7530</td>
</tr>
<tr>
<td>(Friends+Focals+Geolocation)-based CF+SA</td>
<td>0.7111</td>
</tr>
<tr>
<td>(Friends+Friends-of-Friends+Geolocation)-based CF+SA</td>
<td>0.7081</td>
</tr>
</tbody>
</table>

Table 4.2: Obtained MAEs by eliminating social filtering components associated with Sentiment Analysis (SA is not a separate component. It just represents exploiting of Sentiment Analysis in ratings predictions)
4.4 Stability Evaluation

At the end, to measure the stability of our proposed approach, we use another metric called Root Mean Square Error (RMSE) to calculate the variation of prediction error which is evaluated by MAE. The RMSE is defined as:

\[
sd = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}
\] (4.1)

where \(\{x_1, x_2, \ldots, x_N\}\) are the computed values and \(\bar{x}\) is the mean value of them.

In addition, we sampled a small dataset as the testset by crawling the website of Foursquare. Then, we considered each user who had at least two mayorships along with his related social groups as a subset. Next, we evaluated the performance of our recommender systems on the subsets by calculating MAE of each subset. Finally, we used RMSE to measure the standard deviation of MAEs. The results are shown in Table 4.3. As it can be seen, the variation of MAE is quite low which expresses the reliability of the results and stability of the recommender systems. Figures 4.9, 4.10, and 4.11 depict the MAE variations of Baseline, Social CF and Informed CF respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.7432</td>
<td>0.0668</td>
</tr>
<tr>
<td>Social CF</td>
<td>0.6906</td>
<td>0.0697</td>
</tr>
<tr>
<td>Informed CF</td>
<td>0.6859</td>
<td>0.0664</td>
</tr>
</tbody>
</table>

Table 4.3: Standard deviation measured for discussed recommender systems
Figure 4.9: Variation of Mean Absolute Error for the baseline recommender system (MAE=0.7432)

Figure 4.10: Variation of Mean Absolute Error for the social recommender system (MAE=0.6906)
Figure 4.11: Variation of Mean Absolute Error for the informed recommender system (MAE=0.6859)
Chapter 5

Discussion

Some important observations become feasible through the experimental evaluation of the above-mentioned collaborative filterings. For ease of comparison, we present the mean absolute error of collaborative filtering discussed in this work in Table 5.1.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geolocation-based CF</td>
<td>0.5679</td>
</tr>
<tr>
<td>Traditional CF</td>
<td>0.9902</td>
</tr>
<tr>
<td>Friends-based CF</td>
<td>0.9876</td>
</tr>
<tr>
<td>Friends-based CF+SA</td>
<td>0.9976</td>
</tr>
<tr>
<td>Friends-of-Friends-based CF</td>
<td>0.9901</td>
</tr>
<tr>
<td>Friends-of-Friends-based CF+SA</td>
<td>0.9929</td>
</tr>
<tr>
<td>Focals-based CF</td>
<td>0.9961</td>
</tr>
<tr>
<td>Focals-based CF+SA</td>
<td>0.9976</td>
</tr>
<tr>
<td>(Friends+Geolocation)-based CF</td>
<td>0.7554</td>
</tr>
<tr>
<td>(Friends+Geolocation)-based CF+SA</td>
<td>0.7530</td>
</tr>
<tr>
<td>(Friends+Focals+Geolocation)-based CF</td>
<td>0.7148</td>
</tr>
<tr>
<td>(Friends+Focals+Geolocation)-based CF+SA</td>
<td>0.7111</td>
</tr>
<tr>
<td>(Friends+Focals-of-Friends+Geolocation)-based CF</td>
<td>0.7119</td>
</tr>
<tr>
<td>(Friends+Focals-of-Friends+Geolocation)-based CF+SA</td>
<td>0.7081</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.7432</td>
</tr>
<tr>
<td>Social CF</td>
<td>0.6906</td>
</tr>
<tr>
<td>Informed CF</td>
<td>0.6859</td>
</tr>
</tbody>
</table>

Table 5.1: Comparison of the MAEs of the investigated methods. Here, Baseline = Geolocation-based CF+Traditional CF, Social CF = (Friends+Focals+Friends-of-Friends+Geolocation)-based CF, and Informed CF = (Friends+Focals-of-Friends+Geolocation)-based CF+SA (SA is not a separate component; it is mentioned to indicate the use of Sentiment Analysis in ratings predictions).

The first observation is that geolocation-based CF performs very well in predicting ratings. This proves the importance of geographical factor in users’ decision making.
Evidence from data analysis revealed that how geographical specification of venues affect users’ preferences and, in result, the performance of recommenders.

Secondly, traditional CF has a very weak performance; this can suggest that cosine similarity is not appropriate to calculate the similarity weight for friends because feature vectors of users should consist of just one sort of features. In other words, to construct the feature vector either the list of friends or the list of checked-in venues can be considered. However, to measure the closeness of two people both factors are influential. In addition, considering all users in the network to make the user-item matrix consumes a lot of memory and process that makes it impossible for huge social networks like Foursquare.

Another observation is that, the components of social CF (CFs based on Friends, friends-of-friends, and focals) did not perform well. According to homophily principle and our data analysis, friends are supposed to have more common friends and visited locations with each other, and so it is expected that Friends-based CF predicts ratings accurately. But, its error ratio is very high and close to traditional CF. Meanwhile, the result of the CF based on friends-of-friends was even worse. To explain this, the first reason, we have come up with, is that our dataset is too sparse. Since we do not have access to the check-ins list of users in Foursquare, we miss a remarkable part of information about users’ POIs, which is very critical to calculate both similarity weight between the target user and related individuals and to estimate his ratings.

The second reason can be argued with regard to sampling method. Although Snowball sampling preserves the structure of the network related to the given seed, the further levels it samples, the more nodes must be crawled, and, as a result, much information will be lost in the boundaries of the network. This leads to stronger sparsity and, therefore, higher error ratio since the recommender cannot cover more POIs in the sampled network. This can be validated through an experiment on a smaller dataset. This time, the error rate of our social filtering was lower ($MAE = 0.6225$) that confirms our claim regard to stronger sparseness in the boundaries of the dataset.

Besides, examining the CF based on focals social group (foc-based CF), we observed a very low coverage. However, this time, this result was expected since there are few number of such individuals are in the network, and also they have lower similarity weights than other social groups because they do not have any friends in common with the target user.
All in all, further experiments have shown that, first and foremost, Friends-based CF is not an efficacious method singly. However, when other components are integrated into Friends-based CF as a social filtering, the recommender produces more accurate predictions than a recommender system contains traditional CF. For instance, even a combination of CFs based on Friends and Focals brings about a lower error rate. Hence, our proposed unified collaborative filtering which includes geolocation-based, friends-of-friends-based and focals-based CFs, outperforms our baseline recommender as the state-of-the-art, as it is represented in Table 5.1.

Last but not the least, our informed CF performs adequately good although social CF’s components still showed to be erroneous. In addition to the aforementioned reasons, a number of reasons can be enumerated for this observation. First, in social networks, people put comments in different languages from all over the world. The problem of classifying a multilingual corpus is still an open search area. Second, comments are written about different kinds of venues, such as restaurants, gyms, hospitals etc. Thus, the informed recommender cannot focus on one specific domain to learn the sentiments of specific words and collocations used in that domain. Third, comments often contain both positive and negative sentences. For example, in sentence “Our family’s favorite Chinese restaurant. Service is a little slow but food is awesome.”, the user has positive sentiment about a restaurant and its food while he has negative attitude toward the service there. This makes it difficult for recommender, especially in a social network like Foursquare which does not have rating system, to make suggestions.

Moreover, surprisingly, this time, fof-based CF made better recommendations than other components. From the social network analysis point of view, this can be discussed based on the theories of strength of ties in social networks proposed by Granovetter\textsuperscript{1}, and triadic closure, discussed in section 1.2.2.

Granovetter, in [31], argues that people more often gain their novel information about new events or locations, like about new jobs, from distant friends [32]. Since this information is newer, it seems that it is more attracting to individuals. Then, the information cascaded from friends-of-friends can be more influential in users’ decision making.

\textsuperscript{1}According to Granovetter, there is a strong tie between close friends, and a weak tie between acquaintances (distant friends).
Therefore, we can explain our observation as follows. With regard to triadic closure, there is a high possibility that two friends of a user will make friends with each other. Hence, by considering this possibility as a weak tie, and the real friendship connection as a strong tie, we can claim that our experimental results justify the fact that information received from acquaintances plays a very influential role in individuals’ decision making procedure. Plus, since usually the number of friends-of-friends is higher than the number of friends, much information can be gained from them.

Succinctly, individuals get influenced by different social groups through both observing their behaviors, such as check-ins and receiving information from their comments. This is that our informed CF produced more accurate predictions than other CFs discussed here.
Chapter 6

Conclusions and Future Work

In this thesis, we attempt to evaluate the influence of different social groups on preference and check-in behavior of users along with the effect of geographical properties of venues. Our idea is that in addition to the influence of friends, the effect of other related social groups in farther levels, namely friends-of-friends and focals, should be considered in the recommendation. In this manner, we propose a unified framework of collaborative filtering that encompasses the influence of all social groups in a social network.

First of all, we evaluate the performances of each components separately to identify the effectiveness of each of them in the recommendation. On the one hand, the experiments show that the traditional CF cannot predict ratings adequately. On the other hand, the geographical feature has the most efficacious role in making recommendations for venues in a location-based social network and thus, geolocation-based CF has a very good performance. In addition, social groups have influence on individuals’ decision making. Although the performance of each of the social filtering components is not good individually, the unified social filtering algorithm outperform the baseline.

Secondly, we used natural language processing techniques to take advantage of users’ comments about their visited venues to fortify our recommender based on the information cascading phenomenon in social networks. The obtained results depict the effect of information from users’ comments on their ratings. That is, our informed recommender system can make recommendations more accurately than our unified social filtering.
In conclusion, the experimental results reveal that considering behavior models of social groups and information flow through social networks have a remarkable impact on the performance and the accuracy of the recommender system.

Therefore, as a complement, one possible direction for future work can be employing multitask learning methods to be able to identify the polarity of users’ opinions in different domains. To clarify, there are several kinds of venues in location-based social networks, on which users write their comments in particular languages. For example, writing about staffs in a restaurant is different from in a hospital or a hotel. In this fashion, experimental results, in [8, 22], exhibit improvement in recommendations when systems focus just on one particular domain such as movies, restaurants etc. Therefore, facilitating multitask learning techniques should enable us to have a system learn languages in different domains, and then perform a general analysis on sentiment of comments in different domains.

Moreover, by exploring through comments, we have noticed that comments contain very rich information about venues in different hours of a day. For instance, there are some comments indicate that a restaurant is too crowded at a specific time of the day, or there are not enough parking spaces somewhere in a town. In this matter, augmenting recommender systems with time context-awareness techniques seem to be very promising.
Bibliography


Appendix A

Most informative features are as follows.

- awesome = True pos : neg = 46.7 : 1.0
- amazing = True pos : neg = 42.9 : 1.0
- delicious = True pos : neg = 33.1 : 1.0
- slow = True neg : pos = 28.6 : 1.0
- best = True pos : neg = 24.0 : 1.0
- love = True pos : neg = 22.1 : 1.0
- excellent = True pos : neg = 17.7 : 1.0
- great = True pos : neg = 13.0 : 1.0
- bad = True neg : pos = 12.3 : 1.0
- fantastic = True pos : neg = 12.2 : 1.0
- smell = True neg : pos = 9.9 : 1.0
- wonderful = True pos : neg = 9.2 : 1.0
- waitress = True neg : pos = 8.9 : 1.0
- town = True pos : neg = 8.8 : 1.0
- didn = True neg : pos = 8.4 : 1.0
- roll = True pos : neg = 7.9 : 1.0
- nasty = True neg : pos = 7.8 : 1.0
- disappointing = True neg : pos = 7.8 : 1.0
- glass = True neg : pos = 7.8 : 1.0
- friendly = True pos : neg = 7.8 : 1.0
- fried = True pos : neg = 7.5 : 1.0
- customer = True neg : pos = 7.5 : 1.0
- shrimp = True pos : neg = 7.0 : 1.0
forever = True \hspace{1em} neg : pos = 6.8 : 1.0
pay = True \hspace{1em} neg : pos = 6.8 : 1.0
parking = True \hspace{1em} neg : pos = 6.5 : 1.0
expensive = True \hspace{1em} neg : pos = 6.4 : 1.0
breakfast = True \hspace{1em} pos : neg = 6.3 : 1.0
don = True \hspace{1em} neg : pos = 6.1 : 1.0
never = True \hspace{1em} neg : pos = 5.9 : 1.0
nice = True \hspace{1em} pos : neg = 5.8 : 1.0
spicy = True \hspace{1em} pos : neg = 5.8 : 1.0
baked = True \hspace{1em} pos : neg = 5.8 : 1.0
absolutely = True \hspace{1em} pos : neg = 5.8 : 1.0
hard = True \hspace{1em} neg : pos = 5.7 : 1.0
inside = True \hspace{1em} neg : pos = 5.7 : 1.0
nothing = True \hspace{1em} neg : pos = 5.7 : 1.0
shitty = True \hspace{1em} neg : pos = 5.7 : 1.0
valet = True \hspace{1em} neg : pos = 5.7 : 1.0
dont = True \hspace{1em} neg : pos = 5.7 : 1.0
shop = True \hspace{1em} pos : neg = 5.3 : 1.0
sauce = True \hspace{1em} pos : neg = 5.2 : 1.0
wing = True \hspace{1em} pos : neg = 5.2 : 1.0
soup = True \hspace{1em} pos : neg = 5.2 : 1.0
not = True \hspace{1em} neg : pos = 5.2 : 1.0
egg = True \hspace{1em} pos : neg = 5.0 : 1.0
highly = True \hspace{1em} pos : neg = 5.0 : 1.0
cheap = True \hspace{1em} pos : neg = 5.0 : 1.0
beef = True \hspace{1em} pos : neg = 4.9 : 1.0
chicken = True \hspace{1em} pos : neg = 4.8 : 1.0
tiny = True \hspace{1em} neg : pos = 4.7 : 1.0
venue = True \hspace{1em} neg : pos = 4.7 : 1.0
turn = True \hspace{1em} neg : pos = 4.7 : 1.0
waited = True \hspace{1em} neg : pos = 4.7 : 1.0
closed = True \hspace{1em} neg : pos = 4.7 : 1.0
wouldn = True \hspace{1em} neg : pos = 4.7 : 1.0
host = True \hspace{1em} neg : pos = 4.7 : 1.0
isn = True  neg : pos = 4.7 : 1.0
average = True neg : pos = 4.7 : 1.0
refill = True neg : pos = 4.7 : 1.0
american = True neg : pos = 4.7 : 1.0
cost = True neg : pos = 4.7 : 1.0
fun = True  pos : neg = 4.5 : 1.0
tea = True  pos : neg = 4.5 : 1.0
bring = True neg : pos = 4.2 : 1.0
doe = True  neg : pos = 4.1 : 1.0
bus = True  neg : pos = 4.1 : 1.0
beach = True pos : neg = 4.1 : 1.0
italian = True pos : neg = 4.1 : 1.0
wine = True  pos : neg = 4.0 : 1.0
pork = True  pos : neg = 4.0 : 1.0
fresh = True pos : neg = 3.9 : 1.0
happy = True pos : neg = 3.9 : 1.0
server = True neg : pos = 3.8 : 1.0
dinner = True pos : neg = 3.7 : 1.0
family = True pos : neg = 3.7 : 1.0
monday = True neg : pos = 3.6 : 1.0
bottle = True neg : pos = 3.6 : 1.0
fail = True  neg : pos = 3.6 : 1.0
okay = True  neg : pos = 3.6 : 1.0
late = True  neg : pos = 3.6 : 1.0
warning = True neg : pos = 3.6 : 1.0
person = True neg : pos = 3.6 : 1.0
kinda = True neg : pos = 3.6 : 1.0
spend = True neg : pos = 3.6 : 1.0
shame = True neg : pos = 3.6 : 1.0
drive = True  neg : pos = 3.6 : 1.0
woman = True neg : pos = 3.6 : 1.0
men = True  neg : pos = 3.6 : 1.0
dj = True  neg : pos = 3.6 : 1.0
wonder = True neg : pos = 3.6 : 1.0
connection = True  neg : pos  =  3.6 : 1.0
mess = True        neg : pos  =  3.6 : 1.0
case = True        neg : pos  =  3.6 : 1.0
bed = True         neg : pos  =  3.6 : 1.0
change = True      neg : pos  =  3.6 : 1.0
wait = True        neg : pos  =  3.6 : 1.0
lady = True        neg : pos  =  3.6 : 1.0
sell = True        neg : pos  =  3.6 : 1.0
brought = True     neg : pos  =  3.6 : 1.0