UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS

Guilherme Reinaldo Corrêa

POI prediction based on user selection influences

FLORIANÓPOLIS
2019

GUILHERME REINALDO CORRÊA

POI PREDICTION BASED ON USER SELECTION INFLUENCES

Course Assignment submitted to Universidade Federal de Santa Catarina, as required to obtain the Bachelor's Degree in Computer Science

FLORIANÓPOLIS
2019

GUILHERME REINALDO CORRÊA

POI PREDICTION BASED ON USER SELECTION INFLUENCES

This thesis was judged appropriate to the obtainment of the title of bachelor in Computer Sciente, being approved in it's final form by the examining committee:

Florianópolis, November 28th 2019

Prof. Renato Cislaghi, Dr
Course coordinator

Examiners:

Prof. Vânia Bogorny, Dr.
Adviser
Universidade Federal de Santa Catarina

Prof. Angelo Augusto Frozza
Instituto Federal Catarinense - IFC

Banca: Prof. Ronaldo dos Santos Mello, Dr.
Universidade Federal de Santa Catarina

Acknowledgements

Eu agradeço primeiramente meus pais, Cid e Jacira, por terem me dado a melhor educação que o dinheiro deles podia comprar, e por sempre me incentivarem a concluir meus objetivos. Agradeço também minha irmã, Ana Beatriz, por sempre ter estado ao meu lado e por ter me falado as coisas que eu precisava ouvir. Agradeço meu tio Júlio por ter sido meu orientador extra-oficial ao longo da faculdade, e por ter sido um guia pra mim. Agradeço também minhas avós Maria e Benilda por sempre terem me acolhido em suas casas de braços abertos.

Agradeço minha amiga Luisa por ter me aguentado durante toda a execução desse trabalho, por tudo que ela me ensinou e por ter sido minha companheira tanto nos altos quanto nos baixos.

Agradeço Luiz Arthur e Aninha por todas as boas lembranças que temos juntos, e pela nossa amizade sobreviver as intempéries do tempo.

Agradeço meus colegas de curso, em especial Ricardo Boing e João Guilherme Colombo, por toda a nossa jornada ao longo da faculdade, e no final acabamos nos formando juntos. Agradeço o pessoal do Memes Fiat por suas companhias e risadas que deixaram a faculdade um pouco mais fácil.

Agradeço também a minha orientadora Vânia Bogorny por não ter desistido de mim e por ter aguentado meus atrasos e entregas não-feitas e ainda sim ter me orientado até o final.

Resumo

Tendo em vista a dificuldade de escolha gerada pela grande diversidade de estabelecimentos encontrados nas cidades atualmente e a crescente aderência da população a redes sociais baseadas em localização, realiza-se uma pesquisa aplicada exploratória quali-quantitativa com enfoque indutivo sobre os frameworks existentes de recomendação e de predição de pontos de interesse. Foram identificados 5 fatores que influenciam usuários de redes sociais baseadas em localização a visitarem novos estabelecimentos: geografia, tempo, amizade, personalidade e escolhas feitas por usuários similares. Baseado nesses frameworks e nas influências identificadas, desenvolve-se um modelo unificado que será capaz de prever que estabelecimentos cada usuário visitará com base em parte de sua trajetória individual. Para tanto, divide-se a trajetória de cada usuário em duas partes, a primeira para análise e a segunda para validação, de modo que, ao aplicar o modelo desenvolvido à primeira parte da trajetória, idealmente chega-se na segunda. O modelo é aplicado à base de dados da rede social Gowalla. Diante disso, verifica-se que dentre os 5 fatores identificados, a escolha de usuários similares e a amizade obtiveram a melhor acurácia, e a união dos 5 fatores apresentou resultados melhores que cada fator individualmente, porém, diferentemente dos outros frameworks estudados, a melhora não foi significativa, o que impõe a constatação de que o problema precisa ser estudado mais a fundo.

Palavras-chave: Previsão de pontos de interesse. Trajetórias semânticas. Redes sociais baseadas em localização.

Abstract

Taking into account the difficulty of choice generated by the great diversity of points of interest currently found in cities and the population increasing adherence to location-based social networks, a qualitative and quantitative exploratory applied research with inductive reasoning is performed on existing points of interest prediction and recommendation. 5 factors that influence users from location-based social networks to visit new points of interest were identified: geography, time, friendship, personal taste and choices made from similar users. Based on the identified influences and the frameworks, an unified model that is capable of predict what points of interest each user will visit based on their individual trajectory is developed. To validate the model, each user trajectory is divided in two parts, the first for analysis and the second for validation, so that by applying the developed model to the former, ideally the latter is achieved. Finally, the model is applied to a Gowalla social network dataset. Facing the results, it is stated that among the five influential factors identified (geography, time, friendship, personal taste and similarity with other users), the similarity with other users and the friendship achieved the best prediction accuracy, and the unified model presented better results than each factor individually, however, unlike the other studied frameworks, the improvement was negligible, which imposes that the problem needs further research.

Keywords: POI prediction. Semantic trajectories. Location-based social networks.

List of Figures

List of Tables

List of abbreviations and acronyms

$LBSN$	Location-based social network.
POI	Point of interest.
SN	Social network.
OSN	Online social network.
CF	Collaborative filtering.
CB	Content-based.
USG	User similarity (U), Social (S) and Geographical (G) influences framework.
$PSMM$	Periodic Social Mobility Model framework.
$UTE-SE$	User-based CF with Temporal smoothing Enhancement - Spatial influence Enhanced by popularity framework.
$UTPcube$	User-Time-POI cube.
$MHMM$	Mixed Hidden Markovian Model framework.
$GTAG$	Geographical-Temporal Influences Aware Graph framework.
$CAPRF$	Content-aware point of interest recommendation framework.
$GeoSoCa$	Geographic, Social and Categorical correlations.
$ORec$	Opinion-based point-of-interest recommendation framework.
$IEMF$	Intrinsic and Extrinsic Model framework.
MLE	Maximum-Likelihood Estimation.
EM	Expectation-Maximization algorithm.

Contents

1 Introduction

The evolution of the web 2.0 and the popularization of social networks allowed the creation of many communities within a virtual environment, and permitted that people met around the world through interests and friends in common, no matter where in the world those people live. The social networks are of great utility and provide information about the social link between users. There are many types of social networks, and one of them is known as Location Based Social Networks (LBSN). This kind of social network provides a unique experience to their users, where they can voluntarily check-into a place they have visited and share their experience within that place with their friends and strangers, including information about that place like its name, service provided, weather condition and many other. As social beings that live in community, this information with semantic richness is capable of influencing the decision on whether visitings a point of interest (POI) rather than others. Figure 1 shows the user-user friendship and user-POI check-in activity in a LBSN in the form of a graph, where u_n is a user and l_n is a POI of the LBSN. A friend link indicates a real life connection between u_n and u_m, and a check-in is a timestamp t connecting u_n and l_m, meaning that user u_n visited the POI l_m at t.

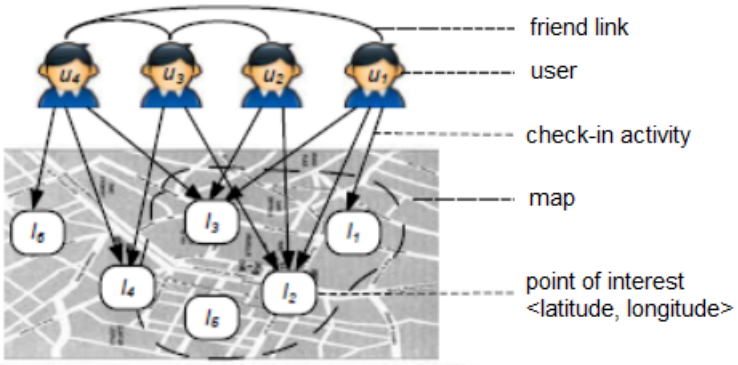


Figure 1 – *LBSN* in form of a graph [Ye et al. 2011]

This kind of virtual environment generates a large amount of data about an individual's location on a daily basis, and this data allows looking at human mobility and behavior in a more analytic way. The analysis of POIs visited by users is of great importance when looking for populational behavior patterns. Each individual has different tastes, personalities and routines, and these characteristics are determinant in the choice of a place to visit (i.e., a POI).

1.1 Problematic

Nowadays, with the advance of globalization and technology, the crescent number of people living in cities and consequently the increase on the amount of options to do what we like or even to perform the simplest activities on our daily routine has become more of a hard task than a synonym of freedom. This phenomenon is called the paradox of choice [Schwartz 2004]. Take the example of buying a pair of jeans, found in his book *SCHWARTZ, B. The paradox of choice: Why more is less? 2004, p. 1*:

> *"About six years ago I went to the Gap to buy a pair of jeans. I tend to wear my jeans until theyre falling apart, so it had been quite a while since my last purchase. A nice young salesperson walked up to me and asked if she could help.*
>
> *- I want a pair of jeans 32 - 28, I said. Do you want them slim fit, easy fit, relaxed fit, baggy, or extra baggy? she replied.*
>
> *- Do you want them stonewashed, acid-washed, or distressed? Do you want them button-fly or zipper-fly? Do you want them faded or regular?*
>
> *I was stunned. A moment or two later I sputtered out something like, I just want regular jeans. You know, the kind that used to be the only kind. It turned out she didn't know, but after consulting one of her older colleagues, she was able to figure out what regular jeans used to be, and she pointed me in the right direction.*

For some people, the task of buying a pair of jeans is simple and straight: I just want a pair of jeans. On the counterpart, for others, this amount of options would be essential to choose a pair of jeans, like the regular Gap customer the salesperson would be used to sell for. More than that, for some people, the place they went over to buy a pair of jeans would not matter much as they just wanted a pair of jeans, and for others it would matter a lot. But how to identify and separate those two categories of individuals?

A pair of jeans is a metaphor that can represent every single task a person may be interested in pursuing during its daily routine. Despite the fact that every person has similar needs, they choose to satisfy those needs differently, and some tasks may interest more to ones and less to others. To be able to differentiate every individual, it is important to take a look on his/her daily routine, every place they visit, which POIs they tend to visit and with whom. A POI is a widely known concept that represents a point location that is useful or interesting for someone, and it is a concept used in many fields of study. In our case, a POI is a place that a person tend to visit during its daily trajectory, and may be its home, workplace, a bakery, park, among others. With the modernization of smart phones and of the remote tracking via satellite technology, there is plenty of space-temporal

data from diverse objects, like cars, animals and even people, available to be collected and analyzed. This data allowed a number of studies on trajectories in a very wide field of application. Data provided by tracking devices are called *raw data*, and trajectories extracted from *raw data* are called *raw trajectories* [Bogorny et al. 2012]. However, *raw trajectories* are limited in the way that they do not provide much semantic information about the location, as they provide only information about the geometry of the trajectory. To attain richer knowledge on the analysis of the data, it is necessary to add semantics to the trajectory, which can be achieved by providing extra information like the name of the visited place, what kind of place it is, and many other information that may be relevant to the study.

Figure 2 illustrates 3 examples of trajectories: Trajectory 1 is a *raw trajectory*, that contains only points located in time-space. Trajectory 2 is a semantic enriched trajectory with information of POIs for a tourism application. Trajectory 3 is an example of *semantic trajectory* enriched with POIs for a traffic application. Both trajectories 2 and 3 are semantic trajectories constructed from the raw trajectory 1, however they consist of different POIs. This is because the semantics of the trajectory depends on what is being analyzed.

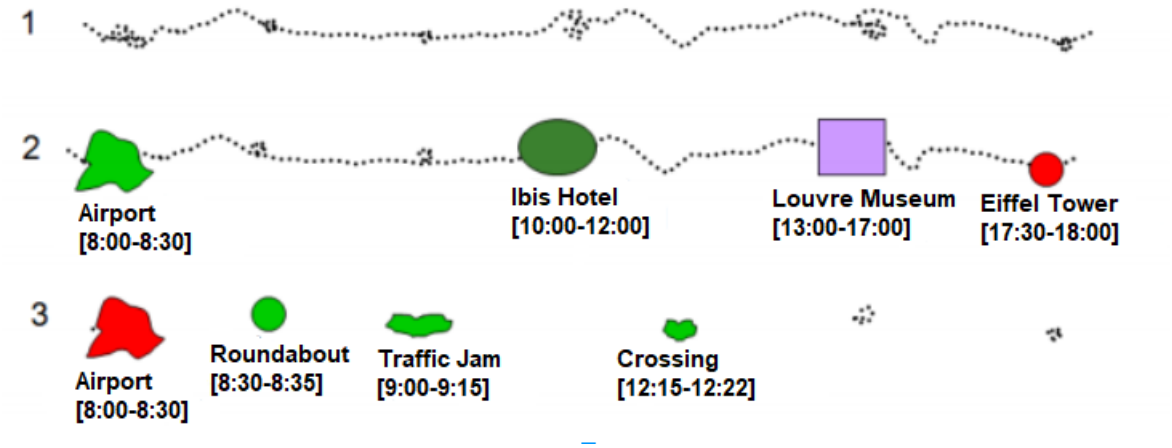


Figure 2 – Examples of *Brute* and *semantic trajectories* [Bogorny et al. 2012]

The LBSNs provide rich data about their users' trajectories, and through those semantic trajectories it is possible to observe many individual and collective behaviors. Although the concept of semantic trajectories is recent, many studies have been made, mainly on POI recommendation [Ye et al. 2011, Yuan et al. 2013, Yuan, Cong and Sun 2014, Gao et al. 2015, Zhang and Chow 2015, Zhang, Chow and Zheng 2015, Li et al. 2017] and location prediction [Cho, Myers and Leskovec 2011, Cheng, Ye and Zhu 2013]. The first tries to identify places for their users to visit based on the most diverse characteristics that a place may present, and the second one, based on his or her routine, environment and other characteristics, tries to correctly identify the next location the user is going to visit given his or her check-in history. Those characteristics are decisive factors when a user

chooses a POI to visit or seek for a service, and they are called influences. After studying the existing frameworks on the area of POI recommendation and location prediction, we identified five factors that influence human choices:

1. **Geographic influence:** relates to distance and accessibility;
2. **Temporal influence:** relates to periodicity;
3. **Social influence:** relates to friendship and social encounters;
4. **Personal influence:** relates to personal tastes and goals;
5. **Collateral influence:** relates to behavior similarity between users;

Each studied framework addresses only a subset of the identified influential factors. Also, each framework models the considered influential factors differently (e.g. different frameworks have different formula to calculate the same influence). Having all the things mentioned above in mind, the objective of this work is introduced in the next section.

1.2 Objective

The general objective is to predict the POIs each user is going to visit based on his or her check-in history taken from a LBSN record.

1.2.1 Specific Objectives

Considering the presented general objective, the specific objectives are the following:

- Score each unvisited POI using the five human mobility influential factors for every user with models based on the current state-of-the-art LBSN recommender and prediction systems;
- For each influence and the combined influence, rank the top-N unvisited POIs the user is most likely to visit;
- Check the accuracy of each model, and then compare their results.

1.3 Methodology

The methodology used to accomplish the objective described earlier is:

- Study the state-of-the-art on recommendation algorithms to identify which factors are relevant to successfully recommend venues and to predict the location of LBSN users;

- Compare the influential factor models from each recommendation and location prediction framework with the others of the same category in order to identify the best approach;
- Divide the dataset into train and test;
- Define a model to score every influential factor for each unvisited POI has on a user based on the studied frameworks for POI recommendation and location prediction;
- Apply the model for each user's check-in activity on the train dataset and rank each unvisited POI;
- Compare the results of the model with the test dataset;
- Expose the conclusions over the analysis of the results.

1.4 Organization

The present work is further divided as follows:

- Chapter 2: Base concepts and brief explanation of the human mobility influential factors learned from the state-of-the-art recommendation systems;
- Chapter 3: Introduction to the state-of-the-art algorithms and a brief explanation of how they address the recommendation problem;
- Chapter 4: formal definition of the human mobility influential factors extracted from the algorithms studied;
- Chapter 5: application of the proposed definitions on a real dataset and analysis of the results obtained;
- Chapter 6: conclusions and future work.

2 Basic Concepts and Influences

In section 2.1, 2.2 and 2.3 we define concepts about trajectories, LBSNs and Recommender systems, respectively. Then, in section 2.4 we discuss deeper the influential factors on human mobility.

2.1 Trajectories

The understanding of a trajectory is the one defined by [Spaccapietra et al. 2008]:

Definition 1 (Trajectory)**.** A trajectory is the user defined record of the evolution of the position (perceived as a point) of an object that is moving in space during a given time interval in order to achieve a given goal, i.e.

$$trajectory : [t_{begin}, t_{end}] \rightarrow space.$$

Where t_{begin} is the instant when the trajectory starts, and t_{end} the instant when trajectory ends.

With the advance of the global navigation satellite systems, it became possible to record its position at a defined time interval or distance travelled, e.g., every 1 second, or every 10 meters. This record contains its carrier location and time and can be seen as a footprint generated by the travelling object. The sequence of this recorded points from an individual is called *moving object trajectory*, which allows us to track an individual, as well as analyzing its trajectory.

Data generated by mobile devices are spatio-temporal data called *raw trajectories*, or as defined by [Furtado et al. 2016]:

Definition 2 (raw trajectory)**.** Raw trajectories' refer to trajectory sample points in their original format, i.e., sequences of spacetime points $\langle((x_1, y_1), t_1), ..., ((x_n, y_n), t_n)\rangle$, where x and y represent a location in space and t corresponds to the time dimension.

Raw trajectories are nevertheless limited, considering that they only carry information about the physical properties of the traveling object, and they do not provide any extra information to the context of their location that might be interesting in an application's point-of-view. However, the simple analysis of raw trajectories on a broader aspect can provide useful information: the traveling object is not necessarily always moving, which means a trajectory can be also semantically segmented by defining a temporal sequence of time sub-intervals where alternatively the object position changes and stays

fixed. The first is called *moves* and the latter *stops*, originally introduced by [Spaccapietra et al. 2008]. Having that in mind, trajectories can be seen as a sequence of stops separating two different moves or as a sequence of moves between a stop and the next one. These stops and moves can be further enriched with any additional information that may be useful for an application, forming *semantic trajectories*, or as defined by [Furtado et al. 2016]:

Definition 3 (Semantic trajectory)**.** A semantic trajectory A is a sequence of stops $\langle a_1, ..., a_n \rangle$ with each stop in the form of a tuple $((x, y), [t1, t2], type)$, where (x, y) is the centroid of all points of the subtrajectory identified as the stop, representing the space dimension, t1 and t2 are the start and end time of the stop, respectively, corresponding to the time dimension, and type is the category of the place where the stop was detected, characterizing the semantics.

More attributes can be added to the stop: the activity of the stop, the goal of the trajectory, or many others, as long as they provide spatio-temporal information. It is important to notice that the *moves* of a user is not provided by most location-based social networks, so only *stops* information is going to be relevant for this work.

2.2 Location-Based Social Networks

Social Networks (SN) are networks of interpersonal relationships (e.g., friendship, family, common interests and shared knowledge) between two or more individuals [Zheng et al. 2011,Zheng 2012]. LBSN is a type of Online Social Network (OSN), where people can share not only their relationship to others but also places they have visited and activities they have performed. According to [Kefalas, Symeonidis and Manolopoulos 2013], a LBSN has 4 basic entities, which are *users*, *locations*, *groups* and *activities*, and they also relate to each other, as shown in Figure 3.

1. On the right side of Figure 3 there are four graphs labeled as **Unipartite graphs:**

   a) *Group graph*: The Group graph is a group-group unipartite graph, which consists of the relations among groups (i.e. RecSys, KDD, and LBSN conference).

   b) *User graph*: The User graph is a user-user unipartite graph, which indicates the social relations among the five users. Each node represents a user connected with another user.

   c) *Location graph*: The Location graph is a location-location unipartite graph, which presents relations among locations. Each location is represented as a node and is connected with another location.

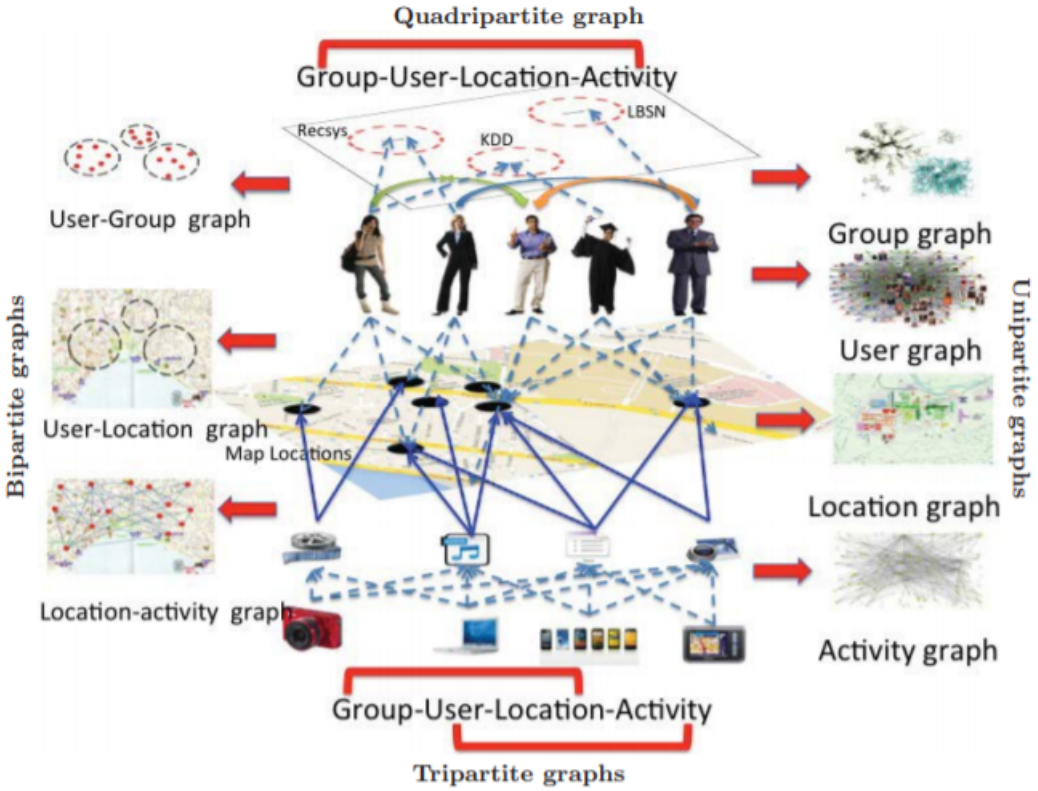


Figure 3 – [Kefalas, Symeonidis and Manolopoulos 2013] Group, User, Location, Activities and their correlations on in LBSNs

d) *Activity graph*: The Activity graph is an activity-activity unipartite graph, which presents relations between activities. Each node represents an activity, which users have performed in the past.

2. On the left side of Figure 3, there are three graphs labeled as **Bipartite graphs:**

   a) *User-Group Graph*: User-group is a bipartite graph that indicates the groups where users belong to.

   b) *User-Location Graph*: User-location is also a bipartite graph presenting locations that users have visited. There are two types of nodes. One type of node represents the user, whereas the second represents the location.

   c) *Location-Activity Graph*: Location-activity graph is a bipartite graph that consists of two types of nodes, i.e. the activity that is performed in a given location.

3. On the bottom of Figure 3, there are two tripartite graphs labeled as **Tripartite graphs**, which are the following:

   a) *Group-User-Location Graph*: The group-user-location graph is a tripartite graph, which presents information about what locations have been visited by users who belong in specific groups.

b) *User-Location-Activity Graph*: The user-location-activity graph is also a tripartite graph that indicates what activities have been performed in a specific location by the users.

4. On the top of Figure 3, there is one quadripartite graph labeled as **Quadripartite graph**, which is the following:

   a) *Group-User-Location-Activity Graph*: The group-user-location-activity graph is a quadripartite graph that includes all four dimensions. In this way, all knowledge about user preferences for activities and groups in POI's is available.

As it can be seen, the applications of LBSN can be very wide, however, in this study we are interested in understanding the influential factors a user has when visiting a POI, so the only relevant graphs are the *user graph*, the *location* graph and the *user-location* bipartite graph.

2.3 Recommender Systems

2.3.1 Traditional Recommender Systems

The recommender problem is a recurrent issue that has been addressed by many companies differently since early 1990s, when the recommender system Tapestry was created and its developers coined the term Collaborative filtering. Facebook, for example, recommended ads to users according to their profile and to the ones their friends liked.

In 2005, [Adomavicius and Tuzhilin 2005] defined the recommendation problem as it follows. U is the collection of all users and S is the collection of all items (e.g. stores, news, locations). let *v:* $U \times S \rightarrow R$ be an utility function that measures the effect of recommending an item s to a user u, where R can be a sorted collection. For each user $u \in U$, item $s' \in S$ is chosen to maximize the user's utility:

$$\forall u \in U, s'_u = arg\max_{s \in S} v(u, s) \tag{2.1}$$

As pointed out by [**?**], the traditional recommender systems mine the two-dimensional relationship between users and items, and recommend items to users through collaborative filtering (CF), content-based (CB) or hybrid recommendation:

1. **Collaborative Filtering Recommendation:** Collaborative filtering finds the user preference through its ratings on different items, next it finds users with the same preference to the target user, and then infer the score of an item to the target user according to the similar user ratings on that item. There are two types of Collaborative Filtering techniques:

- *Item-based CF*: Analyses the user's rating on different items to speculate the similarity among items, and then recommend items based on this similarity.
- *User-based CF*: Analyses different users' ratings on an item to speculate the similarity among items, and then recommend items based on this similarity.

Although CF's simplicity and acceptance in the most variable fields, it faces a cold start problem with both items and users, as new items almost never get rated and new users rarely rate anything, so there are no recommendations for that items or users.

2. **Content-Based Recommendation:** CB recommendation first finds the user preferences through its rating on items, and then recommends items whose contents match the user preference. It differs from collaborative filtering because the first uses other user information to determinate the item rating towards the target user, and the second uses only the target user preference through its historical information.

3. **Hybrid Recommendation:** Combines both CF and CB to avoid their disadvantages: A recently added item will rarely receive any rating through CF so it may never be recommended, but can be recommended via CB through the user's profile. A user that has such uniqueness that no similar users are found via CF would never get recommendations, although he or she still can be recommended via his or hers rating history. A user with little history may be difficult to recommend to via CB, but some similarity may be found through CF and so on.

2.3.2 LBSN Recommender Systems

LBSN recommender systems are an extension to the traditional recommender systems, so they also use CF and CB, but the difference is that the items are restricted to the four elements of a LBSN. It can recommend friends (users that play the role of an item in collaborative filtering) to users (user unipartite graph), POIs to users (user-location bipartite graph), activities performed on a specific location to users (user-location-activity tripartite graph), activities performed on a specific location to users that belong to a determined group (group-user-location-activity quadripartite graph) and so on. This study is interested in the user bias towards a location via his or her check-in record. In this case, there are many other filters that can be applied to increase the quality of the recommendation that are based on the influential factors that will be further discussed on the next section.

2.4 Influential Factors on Human Mobility

In this section we discuss the factors that may influence and individual to visit a location. The symbols that are used in this section are presented in Table 1. There are many factors that can be taken into account when searching a place to visit:

Table 1 – Table of Symbols

Symbol	Description
U	the User set
u_{i}	a user $\in$ U
L	the POI set
l_{i}	a POI $\in$ L
C	the check-in set
$c_{\mathrm{i,j}}$	check-in from user i $\rightarrow$POI j $\in$ C

1. **Geographic Influence**

   According to Tobler's First Law of Geography [**?**], *"Everything is related to everything else, but near things are more related than distant things"*. The influence of geography in human mobility is an old and well known concept, and also very intuitive. It is much more likely for a person to search for a clothing store nearby his or her house, workplace or in the way between them than to go to the next city. Although Tobler's Law has never been actually proven, it is a widely accepted and philosophical concept, and this influence is verified in many studies in POI recommendation and location prediction [Ye et al. 2011, Cho, Myers and Leskovec 2011, Yuan et al. 2013, Cheng, Ye and Zhu 2013, Yuan, Cong and Sun 2014, Zhang and Chow 2015, Li et al. 2017] in the form of a power law distribution for distances below 100 km^2, as illustrated on Figure 4. Even before Tobler, [Stouffer 1940] proposed the theory of intervening opportunities, suggesting that the distance is not the only matter that counts, but also the density. The probability of a user u visiting a POI p far away from home decreases proportionally to the number of POIs that would provide the same service between u and p. If there are many pizza places in 3 block radius from home, it is very likely that the user would choose one of those places if she or he wanted to go out and eat a pizza rather than one that is 15 blocks away. Regardless of the theory adopted, it is undeniable that geography has an influence on choosing a POI.

2. **Temporal Influence**

   Habits and routines are basically activities executed periodically, so it is natural to think that time also has an influence to the user's choice over a place to visit. As stated by [Yuan et al. 2013], *"Users' activities are often influenced by time. For example, a user is more likely to go to a restaurant rather than a bar for lunch at*

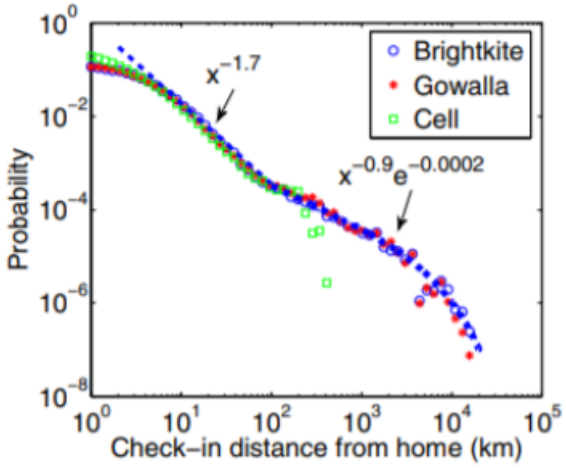


Figure 4 – [Cho, Myers and Leskovec 2011] Probability for a check in to occur as function of the check-in distance

noon, and is more likely to go to a bar rather than a library at midnight". Depending on the time of the day, some services are more requested than others. Culturally, people tend to have lunch around noon, so it is expected that most of the visited POI around noon would belong to the category Food. Also, people are usually in their homes at night and at work during the day. People go to the park during the day, and are very unlikely to go at night because it can be dangerous. As noticed, there are many time constraints that can be inferred depending on the category of the POI, suggesting that choosing one has a strong temporal influence, and many recommender systems are based on that influence.

To verify the periodicity of the human movement behavior, [Cho, Myers and Leskovec 2011] divided the daily check-ins of several users into two states named Home and Work and calculated the probability from all check-ins to be generated either by the Home or Work state as a function of the time of the day.

3. **Social Influence**

Humans are social beings. As a consequence, it is expected that people that are close to each other tend to visit the same places. It is more likely to a user to visit some highly recommended place from a friend than from a total stranger. This influence is verified by [Cho, Myers and Leskovec 2011], and their study shows that it increases with the geographical distance between two friends. This makes sense specially when we think that people usually know their neighborhood well, so most nearby places will be known to them already, but when they visit a distant friend they won't know many nearby places and their suggestion would weight more on the decision of which place to visit. However, their research also showed that only a small fraction of user check-ins from the used databases were previously checked-in by a friend, indicating that friendship has low influence on check-in decision. In contrast, [Zhang

and Chow 2015] affirmed from their research that social influence plays a key role on the performance for recommending POIs. Despite the doubts on social influence being a factor that could be taken into account to successfully indicate a place for one to visit, not all LBSN have explicit friendship links, and some of the LBSN implement them differently: many consider friendship as directed, i.e. user a can be friend of user b without user b being friends of user a, and others consider friendship as reflexive, meaning that if a is a friend of b, then necessarily b is a friend of a.

4. **Personal Influence**

   People have preferences and personalities. Consequently, they perform some activities more often compared to others. A person that likes shopping would have more check-ins into shopping-related POIs than a person that does not. Many LBSN divide their registered POIs into categories, and the category a point of interest belongs to reflect their usual business activity and nature. It is expected that people would have biases towards some categories they identify themselves the most. If a user has a large number of check-ins into some category and there is some popular place into that category that he has not yet visited, there is a great chance that this place would be a very good recommendation for him or her.

5. **Collateral Influence**

   Even though every single person is different, some people are more alike than others. People share tastes, preferences and visit the same kinds of places although they do not necessarily know each other. The collateral influence addresses this kind of issue, also called user similarity, employed by many POI recommender systems in the form of collaborative filtering. Suppose a user that regularly visits three vegetarian restaurants, let us call it user a, indicating a's taste on vegetarian food. Then suppose user b has only visited steakhouses. Clearly users a and b do not frequent the same kind of restaurant, so user b would not be a good reference to indicate restaurants to user a and vice versa. Now, suppose user c has visited the same three vegetarian restaurants and also a fourth vegetarian restaurant r, which he goes frequently. Since users a and c share the vegetarian food taste, and user c likes r, then there is a high chance that restaurant r is going to be a good indication for user a. This is specially uneffective for cold-start users, which are users that just started to use the recommender system and thus have a few check-in data available.

6. **Sentimental Influence**

   People usually like to share their good and bad experiences when they go somewhere. Not only LBSNs, but social networks in general and some mobile applications have largely facilitated the diffusion of user sentiment and experiences towards a location. Places that provide people a good time tend to attract more and more people and

may even become popular places, while places that people have bad experiences tend to repulse other people to come around. Take a dangerous neighborhood for example. Individuals that have a previous knowledge that some street or alley have a great incidence of robbery would try to avoid passing by that place. Or the other way around: people are having a good time on some food event happening on their neighborhood park. Others seeking for a good time would eventually go there and have a good time as well. Sentiment influence analysis, or also called user opinion influence, is however not well developed yet, as it requires human-text interpretation, and there are not many conclusive studies on this influence on POI recommendation.

3 State of the art Frameworks on POI recommendation and location prediction

3.1 Introduction

In this chapter we present a study on several of the existing frameworks on POI recommendation and location prediction to discover how they address the recommendation/prediction problem. Table 2 summarizes the year the framework has been published, the datasets used as base to the frameworks, what influential factors are considered, the geographic granularity level the data is extracted from (i.e. city, state, country or global), statistics of the dataset like number of users, number of POIs and number of check-ins, the restriction applied to the data (pre-processing filters), if the goal is recommendation or prediction and then the type of evaluation. Some frameworks also consider friendship links and user tips or other types of data that contain relevant information about POIs and users such as tweets.

The evaluation of POI recommendation frameworks consist basically in selecting a part of the data as training set and the other as test set. The training set is used by each framework to learn the user preferences, returning the top-N locations to be recommended for every user by the own framework criterion. Then, the test set is used to verify the *precision* and *recall* of the algorithm, as expressed below:

$$Precision@N = \frac{1}{n}\sum_{i=1}^{n}\frac{|Si(N) \cap \mathcal{T}_i|}{K} \tag{3.1}$$

$$Recall@N = \frac{1}{n}\sum_{i=1}^{n}\frac{|Si(N) \cap \mathcal{T}_i|}{|\mathcal{T}_i|}, \tag{3.2}$$

where $S_i(K)$ is a set of top-K unvisited locations recommended to user i that he has not visited in the training set, and $\mathcal{T}_i$ is the set of locations that are visited by user i in the testing set. Because of the check-in matrix data sparsity, both precision and recall are very low in POI recommendation and the frameworks are interested in comparing with other frameworks and some baseline methods. Also, many of the authors are interested in observing the weight of each influential factor individually on their dataset. In Table 2, the *precision* and *recall* values are related to the dataset that reached the best precision values. Note that every dataset has different characteristics, so better performance values from an algorithm does not reflect that it is superior to others.

For location prediction, there are many evaluation metrics that can be taken into account. [Cho, Myers and Leskovec 2011] use 3 kinds of metrics, which are: (1) The average log-likelihood of the check-ins on the unseen test set; (2) predictive accuracy, meaning that the exact location is predicted; (3) Expected distance error, which is a minimum distance of the exact location and the predicted. [Cheng, Ye and Zhu 2013] use category prediction accuracy and location prediction accuracy, which recognize a prediction as correct as long as the true location is at least 400 meters distant to the the top-k (k = 1, 2, 3) returned locations.

In the next sections of this chapter, a brief explanation to each framework is presented considering three categories: (1) Framework, (2) Data analysis and (3) Conclusions made by the authors of each framework.

Framework		USG	PSMM	UTE-SE	MHMM	GTAG	CAPRF	GeoSoCa	ORec	IEMF
Year		2011	2011	2013	2013	2014	2015	2015	2015	2017
Datasets	Foursquare	✓		✓		✓	✓	✓	✓	✓
	Gowalla		✓	✓	✓	✓				✓
	Brightkite		✓							
	Whrll	✓								
	Yelp							✓	✓	
	Twitter						✓			
	Others		✓							
Influence	Geographic	✓	✓	✓	✓	✓		✓	✓	✓
	Temporal		✓	✓	✓	✓				
	Social	✓	✓					✓	✓	
	Collateral	✓		✓		✓				
	Personal				✓		✓	✓		✓
	Sentimental						✓		✓	
Geographic Granularity Level	City	–			✓			✓	✓	–
	State	–		✓		✓	✓			–
	Country	–	✓	✓		✓				–
	Global	–						✓		–
Statistics	# of Users (K)	159	17	12	23	12	4	75	77	55
	# of POIs (K)	150	–	30	818	30	6	137	346	112
	# of Check-ins (K)	–	10900	651	6600	651	134	819	–	2700
Data Restriction	Min check-ins per user	5	10/day	5	58	5	2	–	–	–
	Min check-ins per POI	0	–	5	0	5	–	–	–	–
	Other Restrictions	–	–	–	–	–	[1]	–	–	–
Goal	POI recommendation	✓		✓		✓	✓	✓	✓	✓
	Location prediction		✓	✓	✓					
Evaluation	precision@5	0.061	–	0.031	–	0.040	0.019	0.105	0.071	0.070
	recall@5	0.110	–	0.010	–	0.014	0.026	0.074	0.060	0.053

Table 2 – Framework comparison

3.2 USG

3.2.1 Framework

The USG framework is one of the earliest frameworks that makes POI recommendation and it is based on three kinds of influence:

1. **User similarity (U)**, or collateral influence as called by this study, which is handled simply via user-based collaborative filtering, where the similarity is calculated via cosine similarity; [1]

2. **Social influence (S)**, also handled via collaborative filtering, where friends that have closer social ties and similar check-in behavior have greater influence on the user;

3. **Geographical influence (G)**, which the authors claim to be the first study to consider this kind of influence. They use a Naive-Bayes probabilistic model considering that the check-in over distance probability fits a power-law distribution [2], then they calculate the probability of an unvisited POI l_j belong to the user u visited spot set L_u.

The fusion of those three influences result in their final framework $S_{i,j}$ that calculates the probability score that user i has on POI j, which can be expressed as:

$$S_{i,j} = (1 - \alpha \ - \beta)S_{i,j}(U) + \alpha S_{i,j}(S) + \beta S_{i,j}(G), \tag{3.3}$$

where $S_{i,j}(U), S_{i,j}(S)$ and $S_{i,j}(G)$ are collateral, social and geographical influence scores of user i over POI j, and α and β are weighting parameters.

3.2.2 Data analysis

The authors [Ye et al. 2011] apply their framework on a Foursquare and a Whrll dataset separately, both with check-ins gathered over a month. The Foursquare dataset has 153,577 users, 96,229 POIs and a check-in matrix with density of 4.24×10^{-5}. The Whrll dataset has 5,892 users, 53,432 POIs and a check-in matrix with density of 2.72×10^{-4}, as can be seen in Table 2. The geographic granularity level and from where in the world the check-in activity data was extracted from was not revealed. Users with less than 5 check-ins were discarded from their data. The precision and recall measures are then

[1] Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. If two vectors are equal, then the angle between them is zero, so the value of the cosine is 1.

[2] Probabilistic distribution function that decays very quickly for lower values of x and very slowly for higher values.

compared with other algorithms, such as Random Walk [3], user-based collaborative filtering and friend-based collaborative filtering.

3.2.3 Conclusions

The authors have concluded with their experiment that, through the way they defined the problem, geographical influence shows a more significant impact on the effectiveness of POI recommendations than social influence. Also, Random Walk have shown very poor evaluation results, indicating that it is not suitable for POI recommendation in LBSNs. Finally, because a lot of POIs did not have enough visitors, item-based CF was not an effective approach. This last conclusion agrees with our previous comment that CF tends to never recommend new locations because not many users have actually visited the POI.

3.3 PSMM

3.3.1 Framework

PSMM, or Periodic & Social Mobility Model is a framework proposed by [Cho, Myers and Leskovec 2011] and is used to perform location prediction based on two states "home" and "work", that depend on the time of the day (i.e. people are at work during the day and at home in the night) and has three components:

1. a **Geographical influence** model of spatial locations that a user regularly visits when he or she is in the home/work state using a 2-dimensional time-independent Gaussian distribution based on the user check-in distribution during the day (work) and night (home);

2. a **Temporal influence** model of temporal movement between these locations, that is modeled as a probability distribution over the home/work state of the user with a truncated Gaussian distribution;

3. a **Social influence** model of movement that is influenced by the ties of the social network (i.e., meeting friends). The probability that user u will perform a social check-in at a certain place x_i is determined by two factors: how long since a friend w has checked in, and the distance of w's check-in to x_i, both factors decaying as a power-law over time and distance, respectively.

As stated by the authors [Cho, Myers and Leskovec 2011], PSMM is a two-state mixture of Gaussians with a time-dependent state prior. The temporal part of the model governs

[3] Succession of random steps on some mathematical space, like a user-POI graph.

the transition between home/work states and then depending on the state geographic location of the check-in is generated the time-varying mixture of two time-invariant 2-dimensional Gaussian distributions. The parameters of the model are fitted using Expectation-Maximization. Every check-in is classified as "home", "work" or "outlier", and then the outliers may be further classified into "social" or remain "outlier".

3.3.2 Data analysis

The authors use three datasets for this study: Gowalla, Brightkite and Cell-phone call data from europe. The check-in data is collected between Feb. 2009 and Oct. 2010 for Gowalla and Apr. 2008 to Oct. 2010 for Brightkite. The total number of check-ins for Gowalla is 6.4 million and 4.5 million for Brightkite, rendering the largest check-in data, as can be seen in Table 2. Gowalla and Brightkite also contain an explicit social network. In Gowalla the friendships are undirected and in Brightkite they are directed, but only bi-directional edges are considered for simplicity. There are 196,591 nodes, 950,327 edges in Gowalla and 58,228 nodes, 214,078 edges in Brightkite. The cell phone dataset consists of nearly two million users and 450 million phone calls over the course of 455 days yielding nearly 900 million check-ins with a spatial accuracy of about 3km. Social network ties are created between pairs of people that have both called each other at least five times (10 calls total), yielding a network on 2 million nodes and 4.5 million edges. Only users that had more than 10 check-ins per day were considered, meaning that only 6,233 Brightkite, 10,997 Gowalla, and 853,812 Cellphone users were evaluated.

3.3.3 Conclusions

Even though location-based social networking services are very different from cell phone tower location data, [Cho, Myers and Leskovec 2011] have found many common patterns of human mobility across their datasets, such as the power-law distribution model for geographical movement under 100Km, where there is a change of slope in the probability of a check-in to occur in function of the distance from home (Figure 4), suggesting that the short range and periodic movement is not impacted by social influence while long-distance travel is more influenced by the social network ties.

3.4 UTE-SE

3.4.1 Framework

The UTE-SE framework exploits a mix of 3 influences:

1. **Temporal Influence:** Instead of using a user-POI check-in matrix, the authors adopt a user-time-POI cube (or UTP cube) to represent a user u checking into a POI

l at time slot t. The similarity between two users is high if they always check-into the same location at the same time, but if they check-in the same place in different times the similarity between the two time slots is considered.

2. **Collateral Influence:** user similarity is treated via simple collaborative filtering through extended cosine similarity enhanced by temporal behavior, meaning that the cosine similarity is calculated over the UTP cube, instead of the user-POI check-in matrix.;

3. **Geographical Influence:** The willingness of a user moving from one location to another is modeled through a power-law distribution in the distance function between the two locations, and then given a user u and his/her historical POIs L_u , they calculate $P(l|L_u)$ as the ranking score for each candidate POI l, and then recommend the top ranked POIs to the user based on the Bayesian probability model, which is further enhanced by the POI popularity at the given time.

The unified framework that gives the probability score $c_{u,t,l}$ for a user u checking into a POI l at time t is then represented by:

$$c_{u,t,l} = \alpha \times \overline{c}^{(t)}_{u,t,l} + (1 - \alpha) \times \overline{c}^{(s)}_{u,t,l}, \tag{3.4}$$

where $\overline{c}^{(t)}_{u,t,l}$ and $\overline{c}^{(s)}_{u,t,l}$ are the probability score for temporal and geographical influence, respectively, and α is a tuning parameter.

3.4.2 Data analysis

The authors [Yuan et al. 2013] used two datasets in their experiment: (1) Foursquare, with Singapore check-in data from Aug. 2010 to Jul. 2011 and (2) Gowalla, with data from California and Nevada between Feb. 2009 and Oct. 2010. All users with less than 5 check-ins and all POIs with less than 5 users that checked-in were removed from the data, resulting in 194,108 check-ins made by 2,321 users at 5,596 POIs and a check-in matrix density of 6.35 $\times 10^{-3}$ for the Foursquare dataset and 456,988 check-ins made by 10,162 users at 24,250 POIs and a check-in matrix density of 9.85 $\times 10^{-4}$ for the Gowalla dataset. Note that, as can be verified in Table 2, UTE-SE and GTAG use the same datasets for their experiment.

3.4.3 Conclusions

This study has verified that check-in behavior is influenced by the day of the week and maybe the month of a year. Moreover, it was verified that POI popularity varies as function of the time of the day. Through the evaluation of other algorithms against UTE-SE, [Yuan et al. 2013] concluded that social friend links contribute little for

the accuracy of POI recommendations. This work also verified that Random-walk based method and item-based CF have low performance compared to user-based CF for POI recommendations.

3.5 MHMM

3.5.1 Framework

The MHMM stands for *Mixed Hidden Markovian Model*[4], where the authors [Cheng, Ye and Zhu 2013] divide each user check-in data into a semantic trajectory of stops, where each stop is in the form of $\langle uid, time, latitude, longitude, POI, category \rangle$, where *uid* is the id of the check-in tuple, *time* is the timestamp of the check-in, *latitude* and *longitude* are the geographic coordinates of the POI and *category* is the category that the POI belongs to, one the nine top-level categories from Gowalla: Community, Entertainment, Food, Night life, Outdoors, Shopping, Travel, Events, and None. Then, the whole city of New York is divided into 1km x 1km squares, where they mark the most dominant category, which will be useful to calculate the geographic influence. Their goal is to make location prediction: given a test semantic trajectory $T = c_1 c_2 ... c_n$, they want to predict the location of the next check-in record c_{n+1}. The problem is decomposed into two sub-problems: predicting the category of the user activity at a determined time, and then predict the location of the user activity given the predicted category. This framework is based on three influences:

1. **Personal Influence:** To predict the category of the user, they consider the user semantic trajectory as a sequence of check-ins into one of the nine categories. Then, they map the categories into a set of M hidden states via Hidden Markovian Model, where the ideal value of M is also calculated by their framework. For each semantic trajectory, the model calculates the most probable next underlying state of the user, then the probability of the user next check-in being at a category c given the current state.

2. **Geographic influence** and **Temporal Influence:** The authors then mix the Hidden markovian model (MHMM) with temporal and spatial covariates, making the underlying state also depend from the time of the day, the day of the week, and from the spatial information. The latter is accomplished by collecting all check-in records each 1km x 1km square and compute a probability distribution of the check-in categories.

[4] Statistical model in which the system being modeled is assumed to be a sequence of possible events with hidden states, where the probability of the next event depends only on the state attained in the previous event.

3.5.2 Data analysis

The authors have inspected 13 million check-in records of over 230,000 users from Gowalla for 12 months, from September 2009 to August 2010. The average check-in times of a user during the 12-month period is 58, and users with less than the average are not considered. After the data selection of 23,040 users with a total of 6,634,176 check-in records at 817,683 POIs remain, amount that is approximately half of the original dataset. There are 9 categories of POIs in Gowalla: Community, Entertainment, Food, Night life, Outdoors, Shopping, Travel and Events. Every user is treated by a daily check-in sequence, resulting in a total of 1,054,689 sequences. It holds the second largest check-in records, as can be seen in Table 2, behind only PSMM, which was compared for performance by the authors of MHMM.

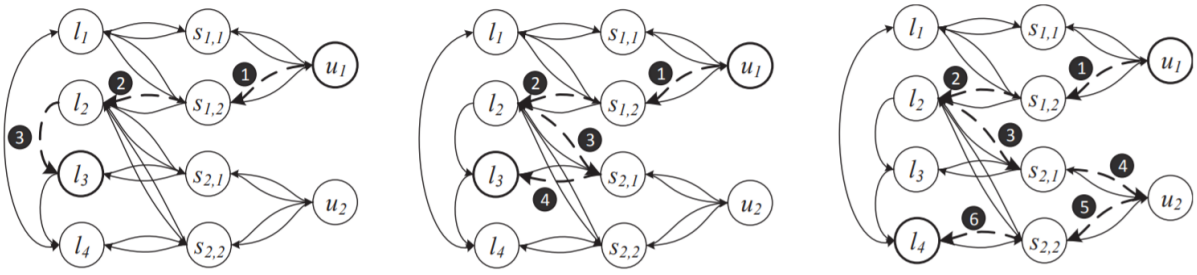
3.5.3 Conclusions

The authors [Cheng, Ye and Zhu 2013] verified that individuals have a bias towards specific categories according to the time of the day, and that filtering possible locations taking into account that bias has great impact on location prediction. This means that personal influence overlaps temporal influence. They had also found out that many 1km x 1km squares have a dominant category, which is a category that represents more than 50% of the POIs from that square.

3.6 GTAG

3.6.1 Framework

Geographical-Temporal Influences Aware Graph (GTAG) is a graph-based approach introduced by [Yuan et al. 2013] to the POI recommendation problem. It has three types of nodes (user node, POI node, and session node) and two types of links (check-in link and POI link). Each user is represented by a user node that connects to a set of session nodes of the user, which represents a time slot (e.g., an hour in a day). A session node of a user is connected to a POI node if the user of the session node visits the POI in the corresponding time slot, like represented on Figure 5, l_n represents a POI node, $s_{a,b}$ represents a session node, where a is the index of the time slot and b is the index of the session node among others with the same time slot, and u_m represents a user node. A pair of edges (between user node and session node, and between session node and POI node, respectively) form a check-in link, which represents a check-in record of the user. Two POI nodes are connected through a POI link if they are located geographically near each other, representing the **Geographical Influence**. The authors consider that (1) Users interests vary with time, and a user may visit different POIs at different time. The temporal interests of a user in



(a) 3-step propagation (b) 4-step propagation (c) 6-step propagation

Figure 5 – [Yuan, Cong and Sun 2014] Examples of valid propagation paths of lengths 3, 4, and 6, from a user node u to a location node. The numbered nodes in black color indicate the propagation steps in each path.

a time is reflected as the POIs he or she visited in that time, so each day is divided in 24 time slots; (2) The check-in of a user in the time closer to the target time are more relevant, and thus more important for recommendation, therefore more recent edges, i.e. have occurred later, have larger weights and (3) If two users have similar behavior over a time, they tend to visit the same POIs at that time, so there are edges from POIs to session nodes, enabling the exploitation of other users temporal interests for recommendation, exposing the **Temporal** and **Collateral Influence**. The basic idea of the framework is to use Breadth-first [5] preference propagation, which consists in injecting the initial preference on the target user node u, and then propagate the preference to candidates POI nodes through various paths over the graph. The resulting paths are selected by these 3 restrictions, that determines that a valid path must be of length 3, 4 or 6, as shown in Figure 5:

1. There is no repeated node in a path, so there are no loops.
2. The path can contain only one visited POI node and session node of the target user, which avoid long paths.
3. The path terminates when an unvisited POI node is met. Without this constraint, the preference would be propagated from an unvisited POI to another unvisited POI, which will amplify the uncertainty in recommendation.

The POIs with N greatest weight sum along the path among all POI paths are accounted for recommendation.

3.6.2 Data analysis

The two datasets contain check-in records from Foursquare and Gowalla, respectively. As it can be seen in Table 2, the UTE-SE framework uses the exact same two

[5] Like breadth-first search, in opposition of depth-first search.

Content Information	Facets of check-in actions
POI properties	What is this POI about?
User Interests	Am I interested?
Sentiment Indications	How good is this POI?

Table 3 – [Gao et al. 2015] Check-in Actions with respect to Content Information

datasets with the same restriction. The Foursquare dataset contains 342,850 check-ins made in Singapore between Aug. 2010 and Jul. 2011. The Gowalla dataset contains 736,148 Gowalla check-ins made within California and Nevada between Feb. 2009 and Oct. 2010. The users who checked in less than 5 POIs and the POIs which have less than 5 users checked in were removed. After preprocessing, the Foursquare dataset contains 194,108 check-ins made by 2,321 users at 5,596 POIs, and the Gowalla dataset contains 456,988 check-ins made by 10,162 users at 24,250 POIs.

3.6.3 Conclusions

The authors have concluded that temporal influence should be taken into account by POI recommendation algorithms, since there was a significant better performance from algorithms that explore temporal influence from the ones that do not. Indeed, it was verified that increasing the time slot length would increase the precision at the cost of recall.

3.7 CAPRF

3.7.1 Framework

CAPRF stands for content-aware point of interest recommendation framework, where the authors [Gao et al. 2015] retrieve content information (Table 3) from user tips related to check-ins from twitter and perform *low-rank matrix factorization* [6] method for POI recommendation in order to exploit both **Sentimental influence** and **Personal influence**. The basic POI recommendation model approximates user u_is latent interests in an unvisited v_j by solving the following optimization problem:

$$\begin{aligned}\min_{u_i,H,V_j\geq 0} \mathcal{J} = &\frac{1}{2}||\hat{W}\odot(C-UHV^T)||_F^2 + \frac{\lambda_1}{2}||A-UG||_F^2 \\ &+ \frac{\lambda_2}{2}||B-V(G-D)||_F^2 + \delta||D||_1 \\ &+ \frac{\alpha}{2}(||U||_F^2 + ||H||_F^2 + ||V||_F^2 + ||G||_F^2),\end{aligned} \tag{3.5}$$

where $\hat{W} = W = \eta * S$ is a non-negative sentiment-enhanced function, W is a Weighting matrix to the user sentiments, S are the sentiment indications, $\odot$ is the symmetric tensor

[6] A technique used to compress a sparse matrix into smaller data

product operator, C is a check-in matrix, U is the user interest matrix, H is a data-dependent matrix for model flexibility, V is the POI latent properties matrix, A is the user interest content and B is the POI-property content, both extracted from the tweets, G and $\hat{G}$ are overlapping word latent topics whose relationship $G - \hat{G} = D$, where the former is in user context related to user-interest content, and the latter is in POI context related to POI-property content, $||\theta||_p$ is the p-norm of θ and $\eta, \lambda_1, \lambda_2, \alpha$ and δ are control weights. The parameters from the function are estimated via gradient descent.

3.7.2 Data analysis

The check-in data from this work is gathered from Foursquare distributed over California from May 2008 and Sep. 2013, totalizing 4,287 users, 134,556 check-ins and 5,878 POIs. Only users with at least 2 check-ins in 2 distinct POIs were considered. On top of that, 19,741 tips and 56,718 comments were extracted from the corresponding tweets over those check-ins. The first framework to use data from a non location-based social network. Also the first to perform sentiment analysis, as can be verified in table 2.

3.7.3 Conclusions

The authors had verified that sentimental influence provides an increase of performance against the algorithms that do not take it into account. Furthermore, through comparison of other frameworks, it seems that personal influence have a greater impact than sentiment influence.

3.8 GeoSoCa

3.8.1 Framework

GeoSoCa stands for geographic, social and categorical correlations, and is based solely on the 3 influences:

1. **Geographical Influence:** The distribution form from the check-in POIs of a user is learned based on a kernel density estimation, which include 3 steps: (1) pilot estimation based on a kernel density estimation with a fixed bandwidth, weighting every check-in in relation to the frequency the user has visited it over all of his or her check-in record (2) local bandwidth determination from the pilot estimation and (3) adaptive kernel estimation for geographical relevance score.

2. **Social Influence:** The social influence relevance scoring process also consists of 3 steps: (1) social aggregation which is simply the user rating over a location based on his friends rating over that same location (2) distribution estimation of social

frequency or rating, modeling social check-in frequency as a power-law distribution (verified through their data) and (3) social relevance score computation, which ranks the check-in frequency on a POI over all check-in frequency.

3. **Personal Influence:** Another 3 steps incorporate the personal influence, or user bias towards a category: (1) weighting popularity by categorical bias, which extracts the categories the user is more inclined to visit as an arithmetic mean (2) distribution estimation of categorical popularity, approaching the categorical popularity to a power-law distribution (verified through their data) and (3) categorical relevance score computation, similar to the social relevance score computation. Finally, all influences are merged into a unified preference score $s(u, l)$ based on the product rule:

$$s(u, l) = f_{Geo}(l|u) \cdot F_{So}(x_{u,l}) \cdot F_{Ca}(y_{u,l}), \tag{3.6}$$

where $f_{Geo}(l|u)$ is the adaptive kernel estimation f Geo (l|u) of the check-in distribution, $F_{So}(x_{u,l})$ is the social relevance score computation and $F_{Ca}(y_{u,l})$ is the categorical relevance score computation of user u on an unvisited POI l.

3.8.2 Data analysis

This study made by [Zhang and Chow 2015] is based on a Foursquare dataset with 4,163 users with 32,512 social links, 121,142 POIs belonging to 35 categories and 483,813 check-ins over the world but focusing in two cities, which were not specified. A Yelp dataset with 70,817 users with 303,032 friend links, 15,579 POIs belonging to 591 categories and 335,022 check-ins over the state of Arizona, USA. This framework has also the best precision value compared to the other analyzed frameworks (Table 2).

3.8.3 Conclusions

In this work, [Zhang and Chow 2015] observed that the social information had the best performance among the three influences on the Foursquare dataset, contradicting the thought that social influence was not significant to POI prediction, as concluded by [Cho, Myers and Leskovec 2011] (for small distances) and [Yuan et al. 2013]. However, it performed worse than the other two influences on the Yelp dataset.

3.9 ORec

3.9.1 Framework

ORec, or opinion-based point-of-interest recommendation framework, proposed by [Zhang and Chow 2015] is a framework to extract the polarity from the user tips by

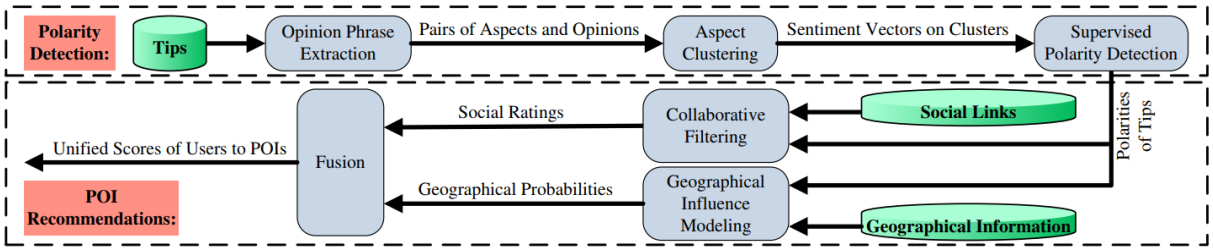


Figure 6 – [Zhang, Chow and Zheng 2015] ORec overview

extracting opinion phrases from tips and determining the sentiment orientation (positive or negative) of an opinion word depending on the associated aspect (i.e. atmosphere, taste, quality, price and their synonyms), and then process the data for POI recommendation. ORec explores 3 influential factors and is divided in two major parts, illustrated on Figure 6:

- **Sentimental Influence:** Handled in the polarity detection part, that generates the polarities of tips for the POI recommendations part. This part has three steps:
  1. Extracting pairs of aspects and opinions from tips;
  2. Grouping aspects into clusters and aggregating opinions of a tip into a sentiment vector in the cluster space;
  3. Training a classification model to predict the polarity of new tips on the polarity labeled tips.
- **Geographical and Social Influences:** The POI recommendations part exploits the tip polarity with social links and geographical information to estimate the score of a user to check into an unvisited POI in order to recommend POIs for the user and also has three steps:
  1. using collaborative filtering techniques to combine tip polarities with social links to estimate the social rating of users to new POIs;
  2. modeling geographical influence by integrating sentimental influence and geographical information of POIs to predict the probability of users visiting new POIs;
  3. fusing the social influence and geographical influence into a unified score for POI recommendations.

3.9.2 Data analysis

[Zhang, Chow and Zheng 2015] use two datasets from two different cities in his work:

1. Los Angeles, which contain 30,208 users, 142,798 POIs, 244,861 tips, 349,985 friend links and a check-in matrix density of 5.68 $\times 10^{-5}$

2. New York City, which contains 47,240 users, 203,765 POIs, 388,594 tips, 810,672 friend links and a check-in matrix density of 4.04 $\times 10^{-5}$.

This data was extracted from both Foursquare and Yelp, and the latter also has 113,993 tips labeled by ratings for 15,585 POIs from various categories including Food, Nightlife, etc. The authors use the tips extracted from Yelp on the Foursquare dataset.

3.9.3 Conclusions

The greatest contribution [Zhang, Chow and Zheng 2015] have made with ORec is to successfully fuse sentimental influence with personal and geographical influence. Their method to extract the polarity by linking sentiment orientation to aspects from user tips greatly improves the enhancement of POI-property and user interest information.

3.10 IEMF

3.10.1 Framework

IEMF, or Intrinsic and Extrinsic model framework, created by [Li et al. 2017], proposes to resume the recommendation problem into a Matrix-Factorization collaborative filtering of two factors:

1. **Intrinsic:** Which is, in other words, **personal influence**;

2. **Extrinsic:** The role that the environment plays on the user interest to visit a location, meaning **geographical influence**.

The framework first identify user activity areas, which cluster user local check-in history over a minimum distance. The user is considered to have unrestricted access to every POI within the user area, so any check-in within would be motivated by intrinsic interest. The chance for the user to visit a POIs outside the user activity area is modeled by extrinsic interest. To deal with data sparsity, the authors propose to deal with unobserved feedback (i.e. user never checked in into that POI) as a mix of positive and negative values and then learn optimal values from missing entries. Integrating all the previous steps, the authors reduce the problem to a optimization function that is solved by a stochastic gradient descent approach.

3.10.2 Data analysis

The data gathered by [Li et al. 2017] are from both Gowalla and Foursquare dataset, from Jan. 2009 to Aug. 2010 and from Dec. 2009 to Jun. 2013, respectively. The former has 52,216 users with 2,577,336 check-ins on 98,351 POIs, and the latter has 2,551 users with 124,993 check-ins on 13,474 POIs.

3.10.3 Conclusions

Recent works focused on squared error based loss were approaching the data-sparseness problem by treating the unobserved feedbacks as negative. [Li et al. 2017] had success when introduced the adaptive learning of the missing feedback. Furthermore, it was verified through comparing with traditional recommender systems that the distinction from this systems and the LBSN recommender system is necessary.

4 Proposed Model

In this chapter we model each influence based on the frameworks explained on chapter 3, and then a unified model that is able to rank the influence factors on each user. The goal of the model is to make POI prediction to the users by analyzing the training dataset, and then comparing the predicted POIs to the test dataset.

4.1 Geographic Influence

Since LBSNs work in a way that users have to go to a POI to perform a check-in, geography plays a very important role when a user is interested to visit a POI. POIs that are closer to the user daily routine are more likely to be visited than others that are far away. The previously discussed frameworks address the geographic influence differently: USG as a Naive-Bayes probability function, PSMM as a Gaussian distribution, and so on. [Cho, Myers and Leskovec 2011] also stated that the probability of a user to check-into a POI is inversely proportional to the square of the distance, and we observe that this is similar to how gravity works. According to Newton's law of universal gravitation, the gravitational force g between two objects is:

$$g = \frac{Gm_1m_2}{r^2}, \tag{4.1}$$

where G is the gravitational constant, m_1 and m_2 are the mass of the objects 1 and 2, respectively, and r is the distance between the center of mass of m_1 and m_2. This equation shows that the gravitational force between two objects decay with the square distance and are proportional to each of the objects mass. Similarly, the probability of a user visit an unvisited POI given a visited POI is proportional to the check-in count of the two POIs (the number of check-ins to the visited POI represents the probability that at any time the user is at that POI, and the number of check-ins to the unvisited POI represents the attractiveness the POI has among the other nearby unvisited POIs) and inversely proportional to the square of the distance between the two POI.Assuming that check-in count plays a similar role to the mass of an object in space, we propose that the geographic influence generated by an unvisited POI can be approached to the sum of the force exerted by each of the user visited POIs to the user unvisited POI, or more formally, the geographic score $S_g(u, l)$ of an unvisited POI l has on the user u is:

$$S_g(u,l) = \sum_{j \in L_u} \frac{\alpha c_j \cdot c_l}{d^2}, \tag{4.2}$$

where L_u is the set of the POIs visited by user u, c_j is the number of check-ins user u has on POI j and c_l is the number of check-ins the unvisited POI l have from all users. To simplify the calculation, α is going to be 1, because it's a multiplicative constant that would be applied to every score equally and consequently can be neglected.

4.2 Temporal Influence

Time has also a great influence when it comes to choose a location to visit. People have strong temporal behaviour. There is time to eat, time to work, basically time for everything, and since POIs are managed by people, it is natural that they also have to follow a timely routine. Restaurants open at Lunch and Diner, Night clubs only open at night, Shopping have greater activity when people are not working and so on. People have a relatively fixed daily routine and perform similar tasks at the same time of the day, as stated by [Yuan et al. 2013]. More than that, as verified by [Cho, Myers and Leskovec 2011], people usually commit to activities on a weekly basis. At working days people do work-related activities, like having lunch at a restaurant because they do not have time to cook, at weekends people tend to spend time with their family or friends, every Wednesday night a person may go to the gym and many others. Based on the fact that both people and POIs have a temporal behavior, it is natural to assume that people will have a greater chance to visit POIs that have a schedule that matches their own. For temporal influence, it is clear that people tend to have two types of temporal behaviour: daily and weekly.

To model the daily temporal behavior, the model proposed by us divides a day into M time slots, and every time slot is mapped into the respective check-in percentage that occurred in that time slot, for every user and POI. Then, the difference between the POI l and the user u temporal behavior can be calculated as:

$$\overline{S}_d(u,l) = \sum_{t_d=0}^{M} |c_l^d(t_d) - c_u^d(t_d)|, \tag{4.3}$$

where $c_l^d(t_d)$ is the check-in percentage of the total number of check-ins for the POI l that occurred in the time slot of a day t_d, $c_u^d(t)$ is the check-in percentage of the total check-ins of user u in the same time slot. To simplify, in this study the M is arbitrarily set to 24 to represent the 24 hours in a day.

The weekly temporal behavior is calculated similarly to the daily temporal behavior,

except that the time slot is the day of the week:

$$\overline{S}_w(u,l) = \sum_{t_w=0}^{N} |c_l^w(t_w) - c_u^w(t+w)|, \tag{4.4}$$

where $t_w = 0$ is Sunday, $t_w = 1$ is Monday, and so on until $N = 6$, which is Saturday, $c_l^w(t_w)$ is the check-in percentage of the total number of check-ins on POI l that occurred in the day of the week t_w and $c_u^w(t_w)$ is the check-in percentage of the total number of check-ins of user u in the same day.

So the user temporal influence score $S_t(u,l)$ over a POI is defined as:

$$S_t(u,l) = \beta \cdot (1 - \overline{S}_d(u,l))^2 + (1-\beta) \cdot (1 - \overline{S}_w(u,l))^2 \tag{4.5}$$

Where β is a tuning parameter. Note that each difference is squared, so larger discrepancies are scored worse.

4.3 Social Influence

Since humans are social beings, a person who had a great experience on a POI will certainly share it with his or her friends, influencing them to visit that POI eventually. The same thing happens with negative experiences. Certainly, friends are a strong influence to a person, and they tend to be a good source for recommendation. People tend to visit a place highly rated by a friend, usually the ones that are closer or have a greater influence on them. People also like to go to a place with their friends to have a good time together. Since the database holds no record of user ratings, the user rating on a POI is going to be represented by user check-in frequency on that same POI. [Zhang, Chow and Zheng 2015] noted that some recommender systems assume that the check-in count is equivalent to the user rating on that POI, however it does not reflect how much the user enjoyed the POI:

"These methods assume that the check-in frequencies of users to POIs directly reflects the preference levels, which may not be true in reality. For example, when a user checks in a POI only once, the user may like the POI very much or conversely."

In the same year, on another article, [Zhang and Chow 2015] analyzed the social check-in frequency on a Foursquare dataset and the rating distribution of a user into a POI on a Yelp dataset to calculate the social aggregation of a POI l on a user u, defined below, and they have found that both social check-in frequency and rating fits a power-law distribution, indicating that there is at least a distribution correlation between them. The social aggregation of a POI l and a user u proposed by [Zhang and Chow 2015] is calculated as:

$$x_{u,l} = \sum_{u' \in U} S_{u,u'} \cdot R_{u',l}, \tag{4.6}$$

where $S_{u,u'}$ is 1 if there is a social link between u and u' or 0 if not, $R_{u',l}$ is the check-in count (for their Foursquare dataset) or rating (for their Yelp dataset) of u' on l. The authors go further by exploiting the power-law distribution and calculate the social relevance score taking into account the relative position of the social aggregation score among all users. This work simplifies the Social influence score $S_s(u, l)$ by normalizing the social aggregation score to the friend check-in count, and then weighting it to the similarity between the two friends:

$$S_s(u, l) = \sum_{u' \in F} \frac{R_{u',l} \cdot w_{u,u'}}{c_{u'}}, \tag{4.7}$$

where F is the set of u's friends, $c_{u'}$ is the total number of check-ins of u', and $w_{u,u'}$ is the similarity between user u and u'. The similarity between the two friends is going to be used instead of the user proximity, since there is no way of calculating the latter. The similarity between the two friends can be calculated as the *cosine similarity* between them:

$$w_{u,u'} = \frac{\sum_{l \in L} c_{u,l} c_{u',l}}{\sqrt{\sum_{l \in L} c_{u,l}^2} \sqrt{\sum_{l \in L} c_{u',l}^2}} \tag{4.8}$$

When $u = u'$, then both friends have the exact same check-in amount on the same POIs, and $w_{u,u'}$ is at it's maximum value 1. When both u and u' have not visited any POI in common, then $w_{u,u'}$ is zero.

4.4 Personal influence

As explained in the previous sections, the personal influence considers the user's taste and personal choice when choosing a POI to visit. It reflects the user preferences on a specific category of POI that provides a certain type of service, like a user that visits several restaurants that belongs to the Vegetarian Food category means that the user likes vegetarian food, and consequently is more inclined to visit other vegetarian restaurants as well. [Zhang and Chow 2015] use the bias $B_{u,c}$ of a user u towards a category c, which is roughly the check-in count of user u into category c, and the popularity $P_{c,l}$ of the POI l among other POIs of category c, which is the percentage of check-ins from category c that

were performed into l. Then, the categorical (personal) popularity $y_{u,l}$ of an unvisited POI for a user proposed is given by:

$$y_{u,l} = \sum_{c \in C} B_{u,c} \cdot P_{c,l} \tag{4.9}$$

Note that a POI can belong to more than one category, i.e. a vegetarian restaurant belongs to the category Vegetarian Food, which itself is a subset of the category Food, meaning that the POI belongs to both categories, consequently $B_{u,c} = 0$ if the POI does not belong to any subset of category c, and $B_{u,c} > 0$ if the POI belongs to a subset of c. The same idea is used for the POI categorical popularity $P_{c,l}$. That is why the categorical popularity is a sum from the product of the categorical bias and the POI categorical popularity. [Zhang and Chow 2015] go further by calculating the distribution of the categorical popularity learned from all users, similarly to the social influence score. To simplify, the personal influence score proposed by us will be calculated by normalizing the categorical bias with all user's average bias B_c to that category:

$$S_p(u,l) = \sum_{c \in C} \frac{B_{u,c} \cdot P_{c,l}}{B_c}, \tag{4.10}$$

where $B_c = \frac{\sum_{u \in U} B_{u,c}}{|U|}$.

4.5 Collateral Influence

Collateral influence is, in simple words, the similarity between two users considering the POIs both have visited. The more POIs visited they have in common, the greater is the similarity between them, and the more similar two users are, the greater the probability that a POI that was visited by user u and unvisited by the user u' to be a good recommendation for u'. The model proposed is no different than the classic user-based collaborative filtering technique, which can be calculated as:

$$S_c(u,l) = \frac{\sum_{u' \in U} w_{u,u'} \cdot c_{u,l}}{\sum_{u' \in U} w_{u,u'}}, \tag{4.11}$$

where $w_{u,u'}$ is the similarity between the users u and u', $c_{u,l}$ is the check-in count of user u on POI l and U is the set of all users from the dataset. $w_{u,u'}$ can be any similarity equation, but for this study the *cosine* similarity is considered due to its simplicity, as presented in equation (4.8).

4.6 Unified Influences

Each influence score formula previously introduced allows us to rank all POIs for every user, but only by one influence at a time. To calculate the combined score, all influences must be taken into account. [Zhang and Chow 2015] for example multiply each of the influences to generate the final score. The approach selected was based on [Ye et al. 2011], [Yuan, Cong and Sun 2014] and many others, where they calculate the combined score by summing all of them, each multiplied by a factor called tuning parameters. They vary the tuning parameters to discover approximately how much each influence was responsible for the model accuracy. To calculate the unified influence score, first the score for each influence is going to be normalized within the range of $[0, 1]$ for every user u, where 0 represents the lowest score a POI p has on u, and 1 the highest. Then, the unified influence formula $S(u, l)$ is defined as:

$$S_(u, l) = a \cdot S_g(u, l) + b \cdot S_t(u, l) + c \cdot S_s(u, l) + d \cdot S_p(u, l) + e \cdot S_c(u, l), \tag{4.12}$$

where a, b, c, d and e are the tuning parameters for geographic, temporal, social, personal and collateral, respectively, and $a + b + c + d + e = 1$.

4.7 Comparison to the State-of-the-art

In this section we give a comparison between our model and the state-of-the-art frameworks for each influence individually.

1. **Geographic Influence:**

   The authors [Ye et al. 2011] use a power-law model to describe the geographic model which is used in a naive-Bayes network. The geographic score is sorted by the probability of an unvisited POI to belong to the user visited POIs set using the network. [Cho, Myers and Leskovec 2011] use the home and work check-in position covariance matrices and the means of users check-in locations when she is in home and work state to feed a 2-dimensional time-independent Gaussian distribution that returns the user geographic score. [Yuan et al. 2013] fit the model the same way [Ye et al. 2011], however they use a Bayesian network instead of naive-Bayes. [Cheng, Ye and Zhu 2013] use a mixed hidden Markovian model with temporal and spatial covariates followed by a Monte Carlo Bayes estimation to calculate the geographic score. [Yuan, Cong and Sun 2014] use a graph-based approach mixing geographic, temporal and collateral influences, where the vertices are users, POIs and session nodes using a breadth-first preference propagation method. The POI nodes are connected to nearby POIs, so the user preference can transition from a visited to an

unvisited POI depending on the edge weight. [Gao et al. 2015] are the only authors that do not consider the geographic influence. [Zhang and Chow 2015] and [Zhang, Chow and Zheng 2015] calculate the geographic influence using an adaptive kernel estimation method on their data. [Li et al. 2017] calculate a mixture personal and geographic influences by maximizing the stochastic gradient descent learned from their data. This work innovates the geographic influence by proposing that it could be approximated by an analogy to the law of universal gravitation.

2. **Temporal Influence:**

   [Cho, Myers and Leskovec 2011] model the temporal influence as a truncated Gaussian distribution parameterized by the time of the day, and the model created for each day of the week. [Yuan et al. 2013] divides a day into 24 time slots and include the time slot into the user-POI check-in matrix, turning it into a user-POI-time cube, and then the temporal score is retrieved via collaborative filtering. [Cheng, Ye and Zhu 2013] calculate both geographic and temporal influences in the same process using the MHMM. [Yuan, Cong and Sun 2014] incorporate the temporal influence by connecting the POI nodes to the POI visitors session nodes, so a user can transition from a POI node to another user session node and then to an unvisited POI. Our work uses the same approach used by [Yuan et al. 2013] by dividing a day into 24 time slots to map each user and POI check-ins, however we do not integrate it into the check-in matrix. Instead, we calculate the differences between the user and the POI activity during the day, and do the same for each day of the week.

3. **Social Influence:**

   [Ye et al. 2011] calculate the social influence via enhanced collaborative filtering, where they consider common POIs and friends to calculate the similarity between two friends. [Cho, Myers and Leskovec 2011] enhance the temporal and geographic model by incorporating non-periodic check-ins and considering them social check-ins. The social influence is modeled as a power-law probability function that decreases with the distance between the user u and the friend v last check-in and how long has been since v has checked in. [Zhang and Chow 2015] first calculate the social aggregation of the candidate POI, which is the sum of all check-in on that POI from all the user friends. Then, they fit their data to a power-law function, and finally the social influence score is calculated with a cumulative distribution function. [Zhang, Chow and Zheng 2015] calculate their social influence score with a user-based collaborative filtering using the users rating instead of check-in count. We adopt the cosine similarity to calculate the similarity between the user and his or her friends, and then multiply by percentage of the friend total check-ins that belong to the candidate POI.

4. **Personal Influence:**

[Cheng, Ye and Zhu 2013] calculate the most likely location to be visited from the category distribution generated by the MHMM. [Gao et al. 2015] extract the user preference by analyzing the semantics from the tips written by the users. [Zhang and Chow 2015] uses the user bias towards the categories of the candidate POI multiplied by the POI popularity, which is calculated as the percentage of the check-ins to the candidate POI categories that belong to the POI itself. Then, they fit their data to a power-law function and calculate the personal influence score with a cumulative distribution function. [Li et al. 2017] calculate the personal influence together with the geographic influence by maximizing the stochastic gradient descent. Our approach is similar to [Zhang and Chow 2015] as we use both the user bias towards the categories of the candidate POI and the candidate POI popularity, however we do not fit our data to a power-law distribution and instead we calculate the general category bias learned from all users, that is used together with the user bias to determine the personal influence score.

5. **Collateral Influence:**

   [Ye et al. 2011] use a simple user-based collaborative filtering that uses common POIs between the users as the measure of similarity. [Yuan et al. 2013] use the enhanced UTP cube to get the similarity between the users and [Yuan, Cong and Sun 2014] incorporate the collateral influence through their graph by connecting the session node to their user node, so the graph can transition through $visitedPOI \longrightarrow anotherusersessionnode \longrightarrow anotherusernode \longrightarrow anothersessionnode \longrightarrow unvisitedPOI$. Our work use the exact same approach as [Ye et al. 2011].

5 Evaluation and Results

In this chapter we perform the experimental evaluation with the Gowalla dataset. We introduce the dataset and the baseline models in sections 5.1 and 5.2, respectively. The explanation on how we conduce the experiment is presented in Section 5.3. Section 5.4 shows a visual representation of the user scores, an important step to understand the experimental results analyzed in Section 5.5.

5.1 Dataset

The data used in the experiment is the Gowalla Dataset gathered by [Liu et al. 2014]. The unprepared dataset contains over 130 thousand users with 4.4 million friendship links, 2.7 million POIs and 36 million check-ins over the whole world. This dataset was chosen because both check-in records and friendship link records are vast and a category tree structure with over 266 categories, which allows us to deeply explore user preference. Only POIs with at least 10 user visitors within the state of New York and users that have visited at least 10 of those POIs in their whole lifetime with at least 10 friends were used on the influence learning. Most of the studied frameworks consider only users and POIs with at least 5 check-ins. However, their interest is to make POI recommendation, which requires that they address fresh users (i.e. users with few check-ins). Since this study is focused on prediction, it is interesting that users have more check-ins so the model can learn the user check-in behavior better. Also, less users increase the algorithm speed. The check-in data for each user were divided in two groups: *train* and *test*, where the former contains approximately 70% of the user check-ins and is used to calculate the user influences. The latter contains the remaining check-ins, and is used to validate the learning model. The percentages of the training and test dataset division is what is generally used on the state-of-the-art. The users that did not visit at least 3 unvisited POIs in the test dataset were also discarded. The filtered dataset contains 1,250 users, 3,459 spots over 266 categories and 311,167 check-ins combining train and test datasets. All POIs that are already visited by the user receive a score of zero.

5.2 Tuning parameters

There are a total of seven tuning parameters used in this work. α from equation 4.2 is considered to be 1, since α multiplies each score for each user and its value would not make a difference to the experiment, and it is used in equation 4.2 to make an analogy to the law of the universal gravitation. The other is β from equation 4.5 that is used to

weight the daily and weekly temporal scores. Due to time difficulties, a deep analysis of β could not have been done, and β was considered to be 0.5. The other five are a, b, c, d, e from the equation 4.12. How we calculate them is explained further.

5.3 Baseline models

As previously discussed, the accuracy of the methods was calculated by two metrics: precision@N and recall@N, where N is the number of top-ranked recommendations. The precision@N calculates the fraction of the predicted POIs that the user actually visited on the test dataset, while recall@N calculates the fraction of the correctly predicted POIs among all POIs that were previously unvisited on the train dataset. Because there are 5 influences being considered, the number of individual combinations among them is $2^5 - 1$, which gives a total of 31 combined influences to compare. One of them, the collateral score, which is already used by many recommender systems and can be used as a baseline model itself. Since we did not have access to any of the state-of-the-art frameworks, the only real baseline model is given by the collaborative filtering. As there are many combinations of influence, only each individual and the combined score were actually tested.

5.4 POI prediction

The raw dataset contains the following data:

1. **User**: each user has a set of friends and of visited POIs.
2. **POI**: each POI have the geographic coordinates, the category and the set of visitor users.
3. **Check-ins**: each check-in links a user to a POI through a timestamp.
4. **Category**: each category has a set of subcategories.

To perform the prediction, for every pair of user u and unvisited POI l from the train dataset we calculate the score for the five influences through their respective formula introduced previously. After calculating the scores, all the user unvisited POI are ranked by their own individual influence score. For each individual influence, the POIs that have achieved the top-N scores are predicted. With the predicted POIs for each individual influence, we compare to the POIs the user has visited in test dataset. With both the predicted POIs from the training set and the visited POIs from the test set, the precision@N and recall@N are calculated through equations 3.1 and 3.2. The previous steps are performed for N $\in$[1, 50], and the results are exposed further.

5.5 Experimental Results

The first step of the experiment is to calculate the precision and recall for every individual influence. Then, we calculate the precision and recall for the combined influence. To do that, the optimal tuning parameters to maximize the accuracy of the model must be found. There are a few ways to estimate them: The first is to naively give random values to the tuning parameters a few times and run the experiment and consider the one that scored best, which would not be the best approach to discover optimized values and it would take lots of iterations, however it could give an idea of what influences would account more to predict check-ins to unvisited POIs. There are many parameter estimation algorithms on the literature such as Maximum-Likelihood estimation (MLE), used by [Yuan, Cong and Sun 2014] to estimate GTAG parameters, also Expectation-Maximization (EM), used by [Cho, Myers and Leskovec 2011] together with MLE to estimate different parameters. Since there are 5 parameters to be found, the optimization problem is too complex. The solution is to use the harmonic mean between precision and recall for each individual factor as a multiplicative constant to their respective score fraction on the combined model. To tune the weights, we elevate each score to a weighting constant γ, and then find the optimal value for γ. This way, each score maintains its relative importance, while we can tune the result to get the best precision and recall values. More formally:

$$a = \left(\frac{\sum^{N} \overline{F_{geo}}}{N} \right)^{\gamma}, \tag{5.1}$$

$$b = \left(\frac{\sum^{N} \overline{F_{tem}}}{N} \right)^{\gamma}, \tag{5.2}$$

$$c = \left(\frac{\sum^{N} \overline{F_{soc}}}{N} \right)^{\gamma}, \tag{5.3}$$

$$d = \left(\frac{\sum^{N} \overline{F_{per}}}{N} \right)^{\gamma}, \tag{5.4}$$

$$e = \left(\frac{\sum^{N} \overline{F_{col}}}{N} \right)^{\gamma}, \tag{5.5}$$

where N is the number of recommended POIs, γ is a tuning parameter and $\overline{F_{geo}}$, $\overline{F_{tem}}$, $\overline{F_{soc}}$, $\overline{F_{per}}$ and $\overline{F_{col}}$ are the average harmonic mean calculated between precision@N and recall@N for geographic, temporal, social, personal and collateral influences, respectively.

The harmonic mean $F@N$ between $precision@N(p@N)$ and $recall@N(r@N)$ is defined as:

$$F@N = 2 \cdot \frac{p@N \cdot r@N}{p@N + r@N}, \tag{5.6}$$

and the average Harmonic mean $\overline{F}$ is:

$$\overline{F} = \sum^{N} F@N \tag{5.7}$$

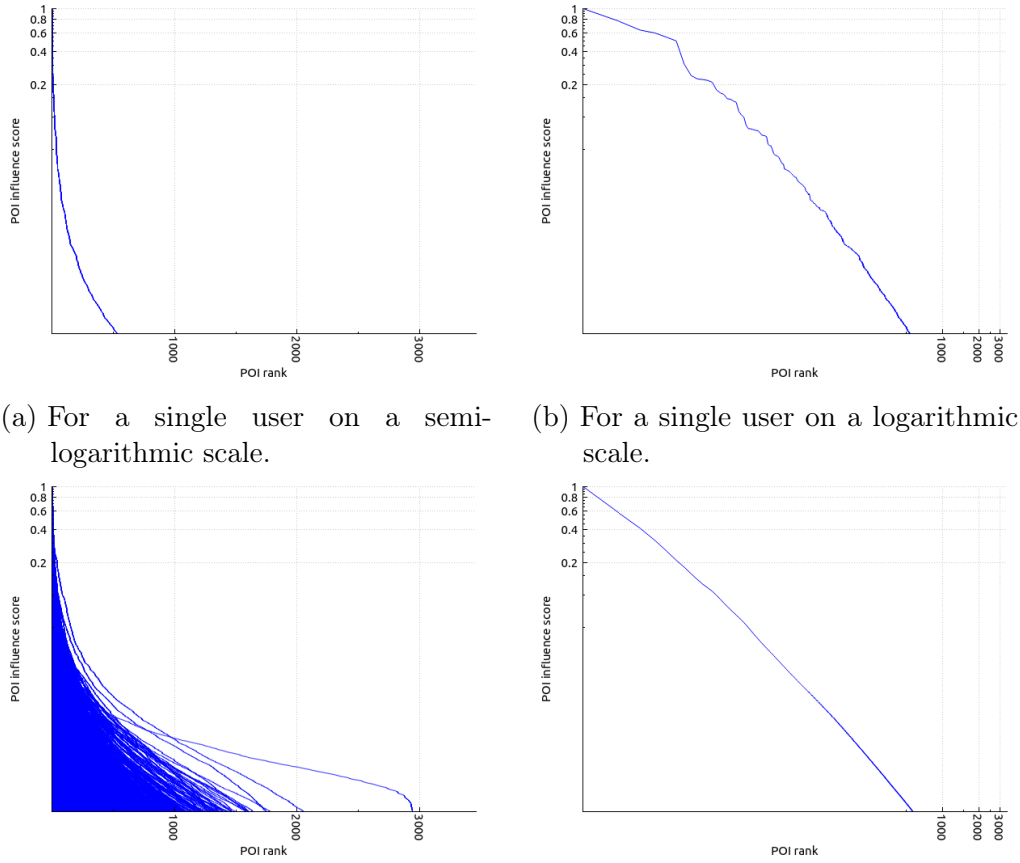
The actual values for each tuning parameter are exposed on section [?].

5.6 User scores

In this Section we show some visual feedback from each of the influential factor experiment. The scores are showed in different scales to give an idea of their behavior. The data generating a line on a log-log function is necessary but not sufficient to conclude that the data follows a power-law distribution. A line on a semi-log function indicates that the data follows a exponential distribution.

5.6.1 Geographic Influence

Figure 7a shows the user rank in relation to the geographic score for a random user on a semi-logarithmic scale. As we can see, the user geographic influence peaks on the first few ranked POIs, and it decreases very rapidly. This is due to the fact that the geographic score decreases with the square of the distance between the unvisited POI and the POIs visited by the user. Figure 7b shows the same information on a logarithmic scale. Figure 7c shows the geographic influence for all users on a semi-logarithmic scale. We can observe that the decrease ratio vary with each user. Some users even have plateaus of scores, where the POI influence score tends to stabilize, decreasing slowly as the rank decreases. The variation between the decrease ratio for different users can be explained by a number of reasons: 1) the user has checked-in in a few, distant POIs in an area with small POI density, as constrained check-in area will have very high ratings on nearby POIs and very small on far away POIs. This is aggravated if the area has a few POIs, because only those few will have high scores while the rest have insignificant scores. The data show that many POIs receive scores in the order of 10^{-10}, while others in the order of 10^{-1}, which represents a huge difference. 2) All POIs visited by the user receive the score zero, so some users have less POIs that actually scored. The plateaus are an interesting phenomenon: it most likely indicates that the user has a scattered check-in behavior, where many unvisited POIs have a similar average distance to the user visited POIs. Figure 7d



(a) For a single user on a semi-logarithmic scale.

(b) For a single user on a logarithmic scale.

(c) For all users on a semi-logarithmic scale.

(d) For the average user on a logarithmic scale.

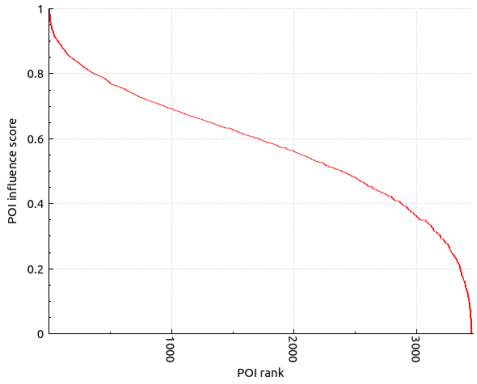
Figure 7 – POI geographic influence score in relation to the POI ranking.

shows the geographic influence as a function of the POI rank for the average user[1] on a logarithmic scale. The line on the log-log scale indicates that the geographic score can be fitted to a power-law distribution.

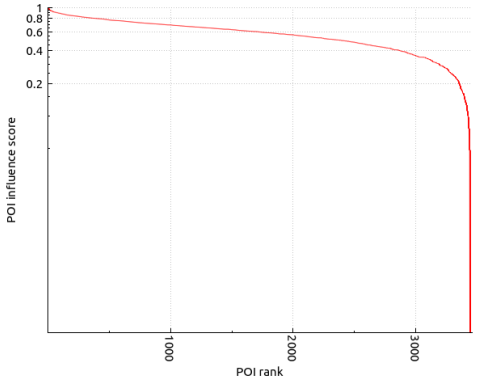
5.6.2 Temporal Influence

Figure 8 shows the POI rating over the POI score for the same user from Figure 7 in different points of view. As we can see in Figure 8a, differently from the geographic influence, the temporal influence score does not reduce rapidly with the POI rank. It has a steep decrease at first, but then it tends to stabilize in the middle ranks and then another rapid decrease at the end. This shows that there are three kinds of temporal relationships between users and POIs: The first is represented by the top-ranked POIs, which are the few that have a truly similar temporal behavior with the user, which are the minority. The second one is represented by the middle ranked POIs, that are the ones that have some activities at the same time but the behavior is not really alike, which represents most of the POIs. This happens because the influence calculates the difference between both

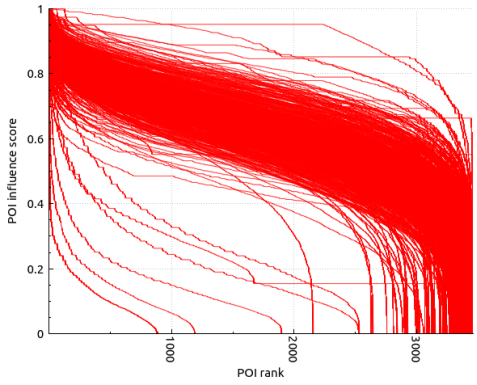
[1] The average user is a fictional user that holds the average score between all users



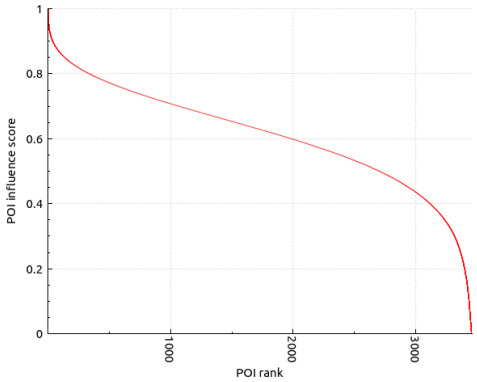
(a) For a single user on a linear scale.



(b) For a single user on a semi-logarithmic scale.



(c) For all users on a linear scale.

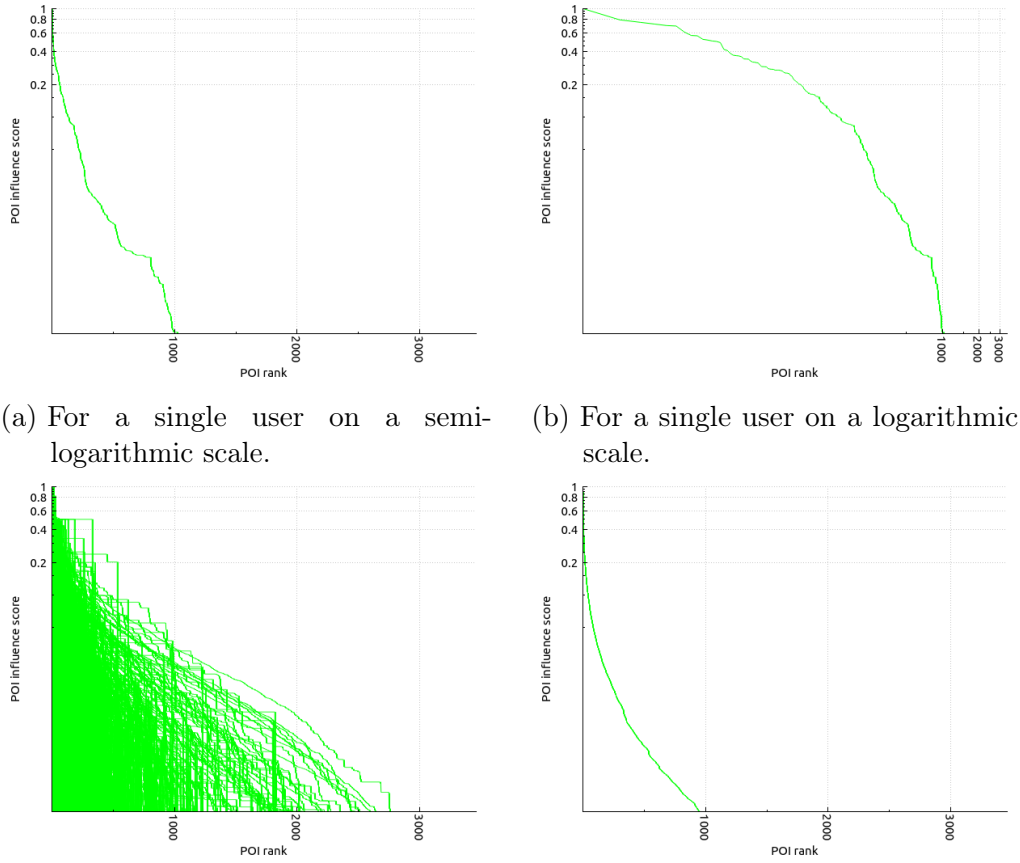


(d) For the average user on a linear scale.

Figure 8 – POI temporal influence score in relation to the POI ranking.

the user and the POI schedule at each time slot and at each day of the week. Since all POIs are located in the state of New York, all POIs belong to the same time zone and the check-in behavior between the user and the POI time slots will overlap most of the time, saved a few exceptions. The same is valid for the day of the week parcel: Users and POIs have some check-ins on the same day, even though their greatest days of check-in activity may not match, resulting in an average temporal score for the two of them. The opposite behavior explains the third group, which are the worst-ranked POIs. Figure 8b shows the same data displayed on a semi-logarithmic scale for comparison with other influences.

Figure 8c shows the POI influence score w.r.t the POI ranking for all users on a linear scale. As we can see, the three kinds of temporal relationships between users and POIs happen to most users. This only shows that the proportion of POIs and users with real similar behavior and with real different behavior represent the minority, and most POIs and users have an average check-in behavior. Figure 8d shows the POI influence score w.r.t the POI ranking for the average user on a linear scale.



(a) For a single user on a semi-logarithmic scale.

(b) For a single user on a logarithmic scale.

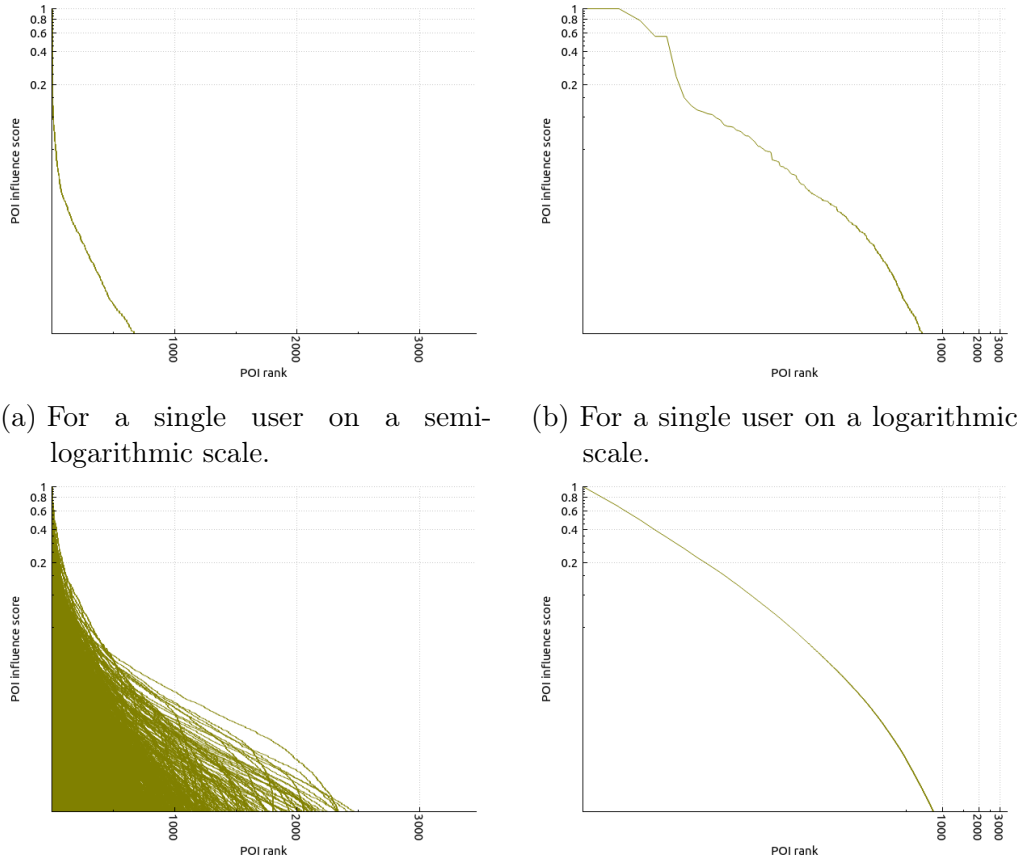
(c) For all users on a semi-logarithmic scale.

(d) For the average user on a semi-logarithmic scale.

Figure 9 – POI social influence score in relation to the POI ranking.

5.6.3 Social Influence

Figure 9a shows the POI social influence score in relation to the POI ranking for the same single user in a semi-logarithmic scale, and Figure 9b in a logarithmic scale. The social score does not fit very well into a power-law distribution as the geographic. However it is a better fit compared to the temporal one, meaning that the geographic score can be better described as a power-law distribution, while the social score can be poorly described as one and temporal score can not be described at all. Figure 9c shows that many users have multiple POIs with the same influence score. This happens because many users have few friends, and as can be seen in Figure 4.7, if the friend has visited many spots once (i.e. $R_{u',l}$ is 1), as the other two variables are constant for each friend, then all of those once visited POIs are going to receive the same score. Also, if a POI is not visited by any friend then its rating would be zero, leaving most scores for those POIs as zero, which explains why the curve decreases quicker in the first few POI ranks for the average user, as shown in Figure 9d.



(a) For a single user on a semi-logarithmic scale.

(b) For a single user on a logarithmic scale.

(c) For all users on a semi-logarithmic scale.

(d) For the average user on a logarithmic scale.

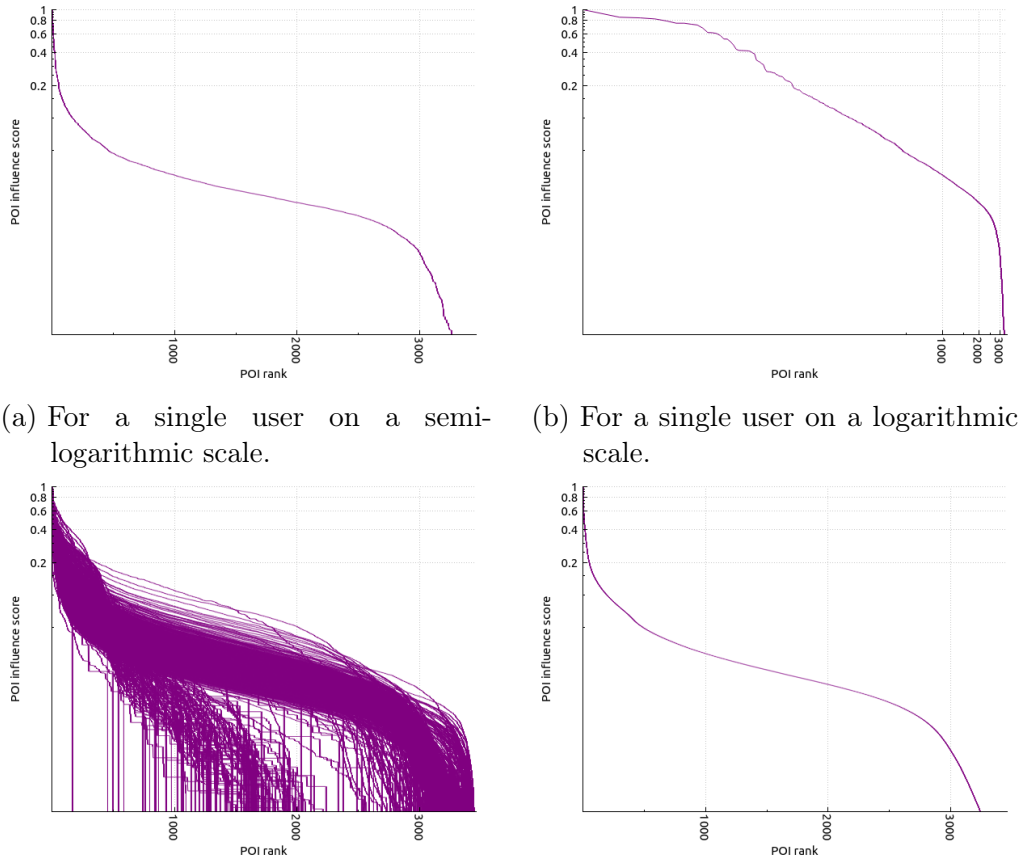
Figure 10 – POI personal influence score in relation to the POI ranking.

5.6.4 Personal Influence

The personal influence scores higher POIs that belongs to categories that the user has a greater check-in number to POIs belonging to that category than the average, as defined in equation 4.9. Further than that, it also scores POIs with a higher popularity among all POIs belonging to the same category. The first statement implies that the fewer check-ins a category has, the greater is the potential score of a POI belonging to that category. Assume two hypothetical categories c and d, with biases $B_c = 0.5$ and $B_d = 0.05$, respectively. A user u with all of his check-ins belonging to POIs on category c would have a bias weight of 2, while a user v with all of his check-ins belonging to POIs on category d would have a bias weight of 20. The second statement implies that POIs belonging to categories with fewer POIs will receive a greater score in general, since the popularity of a POI p is calculated as the percentage of all check-ins to the category of p that belongs to p. This means that the more uncommon and unpopular the POI category is, the higher is the POI score to a user that have a strong bias to that category. Figure 10a shows the POI personal influence score with relation to the POI in a semi-logarithmic scale. As we can see, only a few POIs have really high personal influence score. Figure 10b show

the POI personal influence score with relation to the POI ranking for the same user in a logarithmic scale. Figure 10c shows the POI personal influence score in relation to the POI ranking for all users in a semi-logarithmic graph and Figure 10d shows the average POI personal influence in relation to its ranking. The figure indicates that the personal score also fits a power-law distribution. The further implications of this could be verified in a future work. In overall, 10 shows that personal influence score has a similar shape to the geographic score.

5.6.5 Collateral Influence



(a) For a single user on a semi-logarithmic scale.

(b) For a single user on a logarithmic scale.

(c) For all users on a semi-logarithmic scale.

(d) For the average user on a semi-logarithmic scale.

Figure 11 – POI collateral influence score in relation to the POI ranking.

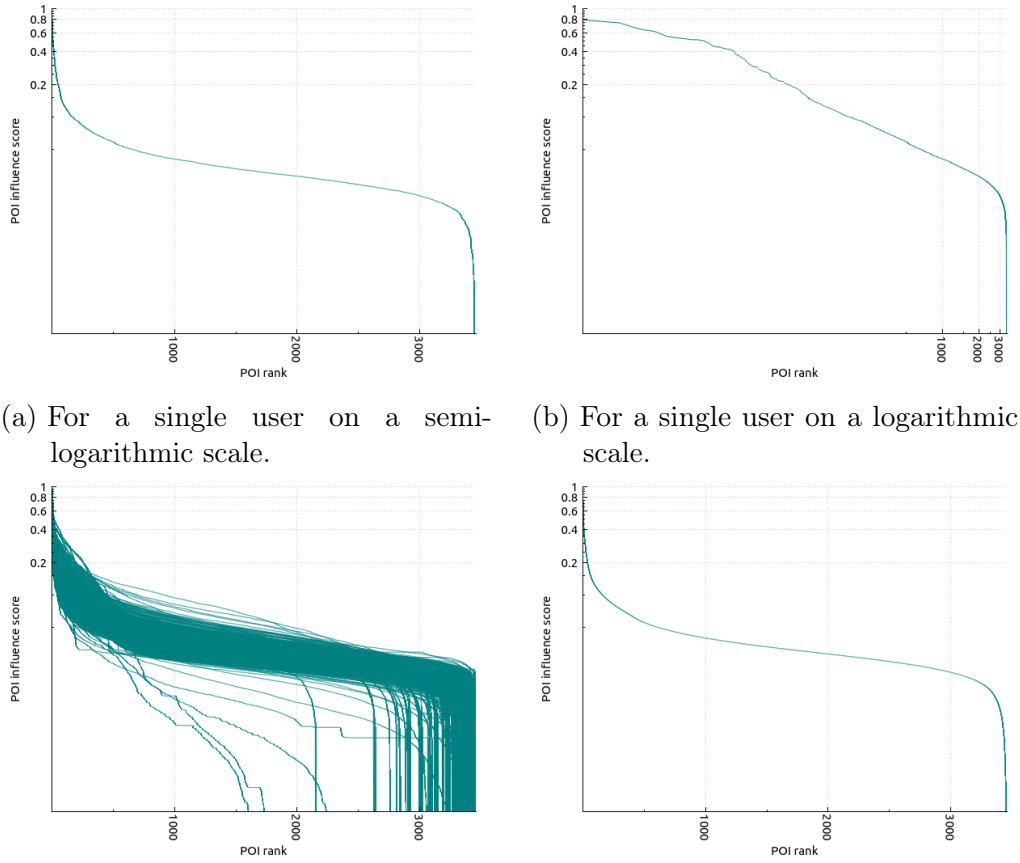
Figures 11a and 11b show the POI collateral influence score in relation to the POI ranking for the same user in both semi-logarithmic and logarithmic scales, respectively. Similarly to the temporal influence, the collateral influence has three clear regions: the few top ranked-POIs, whose score decreases quickly until it stabilizes at the middle and a quick decay at the end. Unlike the temporal influence, however, this influence gives much clearer high scores. The high scores can be mostly explained by the fact the more two users have in common, the less different their check-in behavior is, and consequently the

differences are more evident. This represents the absolute minority of users. Most users will have few visited POIs in common, so they will have a large amount of unvisited POIs to recommend at a low score, which represents the majority of the cases. At last, POIs that are too distant from the user activity area will be frequented by users that have few to none common visited POIs, so unpopular, distant POIs will receive score zero.

Collaborative filtering is a common approach used by recommender systems. Although this work is not focused on recommendation, the technique can be adapted to prediction, as we will show later. Figure 11c shows the POI combined score in relation to its rank for all users in a semi-logarithmic graph. Curiously, the graph shows that many users can be approximated by the one shown in Figure 11a. There are clearly two distinct curves. This can be elucidated by the fact that the less POIs n user has visited, the less users he or she will receive recommendations from. So this dataset has many users that have a few check-ins and consequently receive recommendations from less users, which will result in a smaller pool of recommended POIs by collateral influence. Also, because the collateral influence score is based on the amount of check-ins similar users have on a POI, POIs with a large number of check-ins will be most likely to be recommended to any user. The inverse is also true: POIs that have a small number of check-ins will be less likely to be recommended to any user. Finally, Figure 11d shows the average collateral influence score in relation to the POI ranking on a semi-logarithmic graph, which shows a little to no difference from the user from Figure 11a. This indicates that despite the fact that cases like the ones from Figure 11c that follow the second curve happen, they represent the minority of the cases and have little to no impact to the average collateral score.

5.6.6 Unified Influence

The last analysis is the POI unified influence score in relation to the POI ranking. The optimal value to the exponent γ from equations 5.1 to 5.5 that will be replaced in equation 4.12 is $\gamma = 1.95$. This is the approximated converged value of γ that gives the maximum precision and recall to the combined score, calculated through multiple iterations with different values of γ. This way, the final weights to each individual influence are $a = 0.071, b = 0.027, c = 0.220, d = 0.067$ and $e = 0.615$. This means that collateral score alone accounts for more than 60% of the combined score. Figure 12a and 12b represent the POI unified influence score in relation to the POI ranking for the same single user. It is possible to notice that the combined rating is very close to the collateral influence score, the exception is that middle rankings got higher values. The other influences help to suppress the downsides of the collateral influence. Figure 12c shows the unified influence score for all users. Here it is evident that the combined influence gave higher scores to the POIs that were considered zeros in the collateral influence score. This is a reinforcement that the combined score gives better recommendation.



(a) For a single user on a semi-logarithmic scale.

(b) For a single user on a logarithmic scale.

(c) For all users on a semi-logarithmic scale.

(d) For the average user on a semi-logarithmic scale.

Figure 12 – POI unified influence score in relation to the POI ranking.

Figure 12d shows the average unified influence score in relation to the POI ranking for all users in a semi-logarithmic scale, which is similar to the average collateral influence score shown in figure 11d. This is expected, since the collateral influence has the greatest weight on the unified score calculation.

5.7 Experiment Results

In this section we analyze the prediction accuracy values obtained from the top $N = 1$ to $N = 50$ recommended POIs based on each of the 5 individual influences, plus the combined influence. Figures 13 shows the precision values for the top-N predicted POIs and 14 shows the recall values. The temporal influence got the worse score. This is because the temporal influence accounts only for the similarity between the user and the POI check-in amount over the same time slot, not mattering where in the world the POI is or if the user is effectively able to go visiting. It also does not reflect any of the user taste or bias to actually make the POI be more attractive to the user than other POIs. Indeed, the experiment has shown that, for most users, the actual POI temporal score was very similar in all ranks and did not decrease fast enough to represent a great change

between ranks. The state-of-the-art algorithms have managed to condensate the temporal influence in a more precise model, such as PSMM ([Cho, Myers and Leskovec 2011]), UTE-SE([Yuan et al. 2013]), MHMM ([Cheng, Ye and Zhu 2013]) and GTAG([Yuan, Cong and Sun 2014]). The result effectively has shown that our approach to the temporal influence was not very suitable for the prediction problem.

The personal influence had also low accuracy results, however the results were better than the ones we got from temporal score. The fact that Gowalla has so many categories has turned the prediction harder. Personal influence score has also the same problem that the temporal influence has: it does not take into account the distance between the POI and the user activity.

The Geographic influence was more precise than the personal score, however the recall had worse results than the personal score for lower N values. Geographic score does not take into account the density of POIs around the areas, so there may be actually fewer POI choices in a greater area, which is why it could not perform so well.

The social score got a surprisingly high relative score compared to the other proposed models. This indicates that friend taste is a good recommendation for users, and a user will more likely visit POIs that were visited by a friend.

The collateral score, which is a name wrapper for collaborative filtering, got absolutely the best score between the individual influences, with more than double precision than the second best. This proofs that it is a good recommending and predicting approach, specially if the dataset has the users with low check-in amount filtered out. Users are a good prediction hint to those users who behave similarly.

The combined score had, as expected, the best accuracy. However, the accuracy gain from the other influences combined performed worse than expected. The combined score barely outperformed the collateral score alone. This is probably because each influence was modeled with great independence from each other, so they did not work well when used collectively. Indeed, many of the influences performed too poorly, which disqualified them as good prediction metrics.

Finally, the relatively low accuracy of the influences was expected, because the trajectory data of users on LBSNs is incomplete since it requires that the user explicitly log their activity in the social network, which does not necessarily happen every time a user visit a POI. Also, many locations a user visits during his or her trajectory are not registered on the social network. Consequently, most users do not have enough data to receive a precise analysis. Because there is almost no gain from using the combined influence, it is best to use the collateral influence alone to predict POIs on a real POI prediction application.

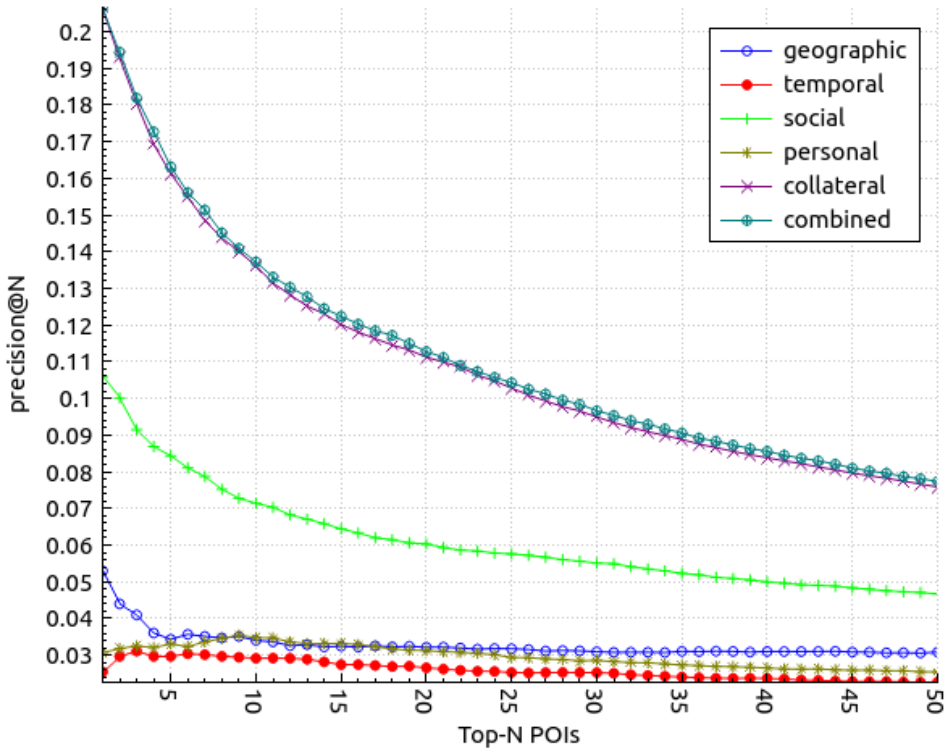


Figure 13 – Prediction precision with respect to given-N values

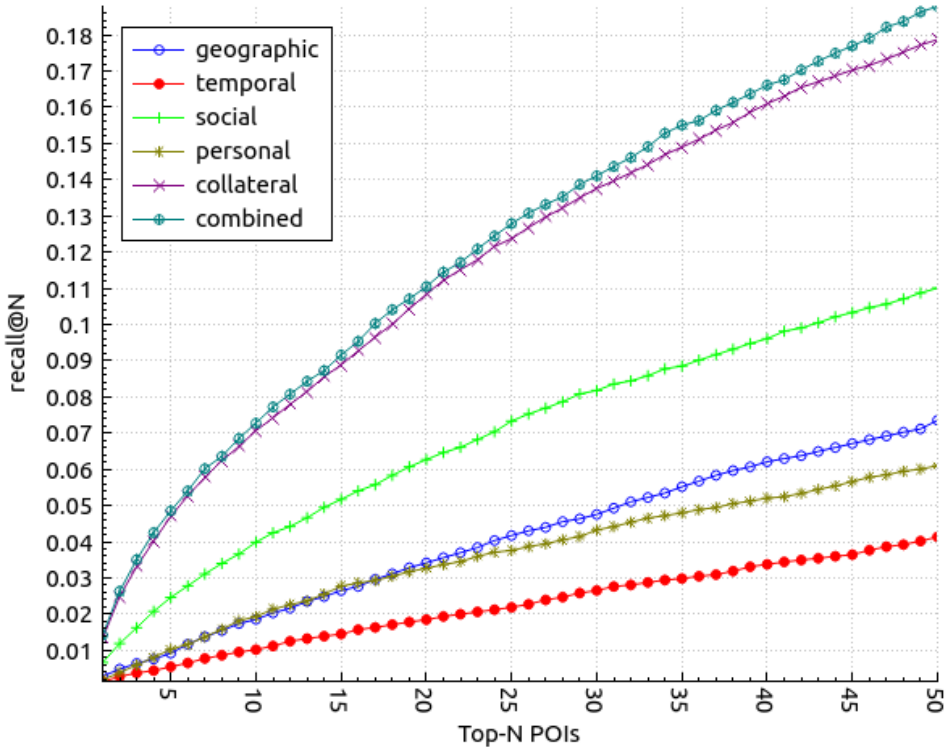


Figure 14 – Prediction recall with respect to given-N values

5.8 Result comparison

In this section we compare the results with other frameworks. Figure 15 shows a graph comparing the accuracy reached by each of the studied frameworks, except PSMM and MHMM, as they use different evaluation metrics. We did not have access to the studied frameworks, so it was not possible to run each algorithm in our own filtered dataset, and the values exposed here are the approximated accuracy values the ones provided by each author, reached on their own dataset and with their own data constraints. All frameworks displayed on Figure 15 are recommender systems, and the precision and recall values are based on a total of 5 recommended POIs.

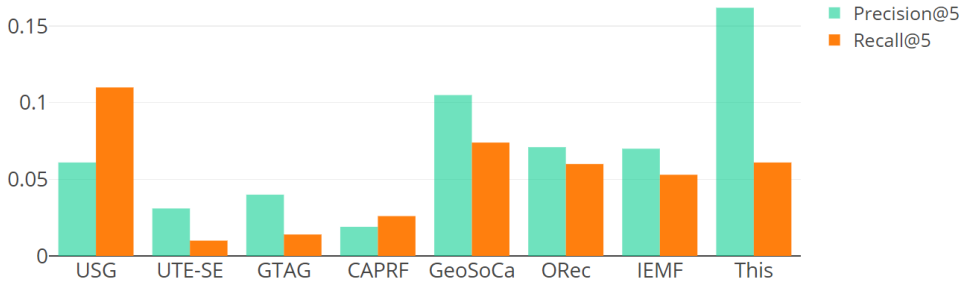


Figure 15 – Accuracy comparison between studied frameworks

As we can see, all frameworks have low precision and recall values. This is due to the check-in matrix sparsity and it is a well-known problem faced by all recommender systems. Among the studied frameworks, GeoSoCa have reached the best precision values (0.105) by examining Geographic, Social and Personal influences, followed by ORec (0.071) examining Geographic, Social and Sentimental influences, and IEMF (0.070), examining Geographic and Personal influences. The lowest values are achieved by CAPRF (0.019) examining Personal and Sentimental influences, UTE-SE (0.031) examining Geographic, Temporal and Collateral influences, and GTAG (0.040), examining Geographic, Temporal and Collateral influences. USG (0.061) have achieved the intermediary precision value examining Geographic, Social and Collateral influences. Our framework have reached the highest precision value by examining Geographic, Temporal, Social, Personal and Collateral influences altogether, however the collateral influence alone have reached better precision values than all other frameworks. This difference can be explained by the higher threshold for user and POI check-in amount. Collaborative filtering yields better results for users with higher number of check-ins, as more similar users can be found and the increased amount of data can differentiate users better.

Geographic influence has been considered by all frameworks except CAPRF, that has the lowest precision among the studied frameworks. This indicates that geography indeed plays an important role on POI recommendation and prediction: although alone it does not determinate which POI is a good candidate, the distance can determinate which

POIs are bad candidates.

The frameworks that have considered the temporal influence have also achieved low precision (UTE-SE and GTAG). This suggests that temporal influence is not suitable for POI recommendation or prediction, as they make recommendations (or predictions) only caring that the user effectively visits the POI, disregarding when the visit happens.

Friendship certainly influence people to visit POIs. Both frameworks that have achieved the highest precision values (GeoSoCa and ORec) considers the social influence. It is also the individual influence with the second highest accuracy on our framework. [Ye et al. 2011] have reported that although friends have good influence on the user check-ins, some friends do not have common taste and the similarity of the check-in behavior is more important than the social ties to make POI recommendation. This is also verified in our work, as the social score have performed worse than the collateral score.

Personal score was only considered by three of the studied frameworks that are on 15. Contradictorily, two of them are in the top 3 accuracy and the other have achieved the lowest values. This indicates that the personal score can be crucial to predict user behavior. In fact, what differentiate users are their personality. The challenge is to successfully extract the user preference from the available data and linking it to other underlying factors that influence user behavior.

At last, collateral influence was considered by three out of the four frameworks with the lowest accuracy (UTE-SE, GTAG and USG) and none of the ones with high accuracy. However, in our work it has achieved the best results. [Ye et al. 2011] have also reported that the collateral influence was responsible for the largest share of their accuracy. The problem with the collateral influence is that it requires a plentiful user data to perform well. This work allow us to conclude that collateral influence is a good way to discover underlying user check-in patterns on a dense dataset.

In general, it is important to consider multiple influential factors when trying to predict user behavior on LBSNs, as each one of them provide a different perspective to the user. There are many underlying factors that makes a user visit a POI, and the more factors are considered, the more precise the prediction can be.

6 Conclusion

In this work, we identified five main influences to the mobility of LBSN users. They are: geographic, temporal, social, personal and collateral. Then, we created a framework unifying all of them in an attempt to predict the user check-in behavior and, among many unvisited locations in a city, discover the most attractive to them based on their previous individual trajectory. After that, we learned how much each influence affects the user decision to visit a previously unvisited location.

After studying some of the current state-of-the-art frameworks on location prediction and recommendation, we discovered many approaches to each of the five influences, and then proposed new models to simulate each of the influences in order to rank unvisited POIs to users, or simply reused the state-of-the-art model. After unifying the influences and applying the resulting model to our train dataset, we have got the top-1 to top-50 locations each user would most likely visit, and then we have compared the results with the test dataset to check our model accuracy, fulfilling the general and each specific objective of this work.

The overall low accuracy of the model was expected due to location-based social network data incompleteness, although some influences scored worse than initially expected. The experiment confirmed that predicting a person's choice towards location preference is indeed a complex problem, and thus the hypothesis that too many choices actually difficult making decisions was confirmed.

Due to the database size, managing computational resources was a challenging task. Many times we were forced to rethink our approach because the complexity got too high and it would take too long for the calculation to finish or because the computer memory was depleted. Also, many analysis were left undone because the deadline to perform this work was short.

This work gathers many of the current state-of-the-art frameworks on POI prediction and POI recommendation, each one of them with unique approaches and insights to the matter. The results also show a little improvement to the consolidated recommendation technique of collaborative filtering when recommending POIs. It has also confirmed that collaborative filtering can be effectively used with POI prediction as well.

6.0.1 Future work

This work focus on five influences. Temporal and Personal influences have shown a very poor performance, thus they could be better studied and remodeled to have their accuracy improved. Geographic influence had an intermediary performance, but only the

distance between the POIs and not the density of the area between them was considered. The effects of the theory of intervening opportunities ([Stouffer 1940]) on the geographic influence was also not tested in this study. Both social and collateral influences have shown good results, the first was newly proposed by us, so it can be tested in the dataset of other LBSNs for comparison.

This work makes comparison between the results of a POI prediction with POI recommender systems because the work has pivoted many times during its execution. A comparison between other recommender systems could provide many useful new information.

The chosen dataset from Gowalla is large, however there are other LBSNs that holds different user, POI and check-in information. The same model could be adapted to a new dataset and the results compared. Further than that, the new information from the datasets may allow us to extract different influences that could also be tested.

Bibliography

2005 ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, 734–749 6 IEEE, 2005.

BOGORNY, V. et al. *Introdução a Trajetórias de Objetos Móveis: conceitos, armazenamento e análise de dados*: Editora Univille, 2012. Quoted 2 times on the pages 11 and 23.

2013 CHENG, H.; YE, J.; ZHU, Z. What's your next move: User activity prediction in location-based social networks. In: *Proceedings of the 2013 SIAM International Conference on Data Mining*, 171–179 2013<https://epubs.siam.org/doi/abs/10.1137/1.9781611972832.19>.

2011 CHO, E.; MYERS, S. A.; LESKOVEC, J. Friendship and mobility: User movement in location-based social networks. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2011. 1082–1090 (KDD '11). ISBN 978-1-4503-0813-7<http://doi.acm.org/10.1145/2020408.2020579>.

ISSN 0360-0300<http://doi.acm.org/10.1145/3154526>.

FURTADO, A. S. et al. Multidimensional similarity measuring for semantic trajectories. *Transactions in GIS*, 280–298 2 20 Wiley Online Library, 2016. Quoted 2 times on the pages 27 and 28.

GAO, H. et al. Content-aware point of interest recommendation on location-based social networks. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. Quoted 5 times on the pages 13, 23, 47, 59, and 60.

2013 KEFALAS, P.; SYMEONIDIS, P.; MANOLOPOULOS, Y. New perspectives for recommendations in location-based social networks: Time, privacy and explainability. In: *Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems*. New York, NY, USA: ACM, 2013. 1–8 (MEDES '13). ISBN 978-1-4503-2004-7<http://doi.acm.org/10.1145/2536146.2536202>.

LI, H. et al. Learning users intrinsic and extrinsic interests for point-of-interest recommendation: A unified approach. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, 19–25 2017. Quoted 6 times on the pages 23, 32, 51, 52, 59, and 60.

LIU, Y. et al. Exploiting geographical neighborhood characteristics for location recommendation. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2014. 739–748 (CIKM '14). ISBN 978-1-4503-2598-1<http://doi.acm.org/10.1145/2661829.2662002>. Quoted on the page 61.

SCHWARTZ, B. *The Paradox of Choice: Why More Is Less - How the Culture of Abundance Robs Us of Satisfaction*: ECCO, 2004. Quoted on the page 22.

SPACCAPIETRA, S. et al. A conceptual view on trajectories. *Data & knowledge engineering*, 126–146 1 65 Elsevier, 2008. Quoted 2 times on the pages 27 and 28.

STOUFFER, S. A. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 845–867 6 5 JSTOR, 1940. Quoted 2 times on the pages 32 and 78.

ISSN 00130095, 19448287<http://www.jstor.org/stable/143141>.

YE, M. et al. Exploiting geographical influence for collaborative point-of-interest recommendation. In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2011. 325–334 (SIGIR '11). ISBN 978-1-4503-0757-4<http://doi.acm.org/10.1145/2009916.2009962>. Quoted 9 times on the pages 11, 21, 23, 32, 40, 58, 59, 60, and 75.

YUAN, Q. et al. Time-aware point-of-interest recommendation. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2013. 363–372 (SIGIR '13). ISBN 978-1-4503-2034-4<http://doi.acm.org/10.1145/2484028.2484030>. Quoted 10 times on the pages 23, 32, 43, 45, 49, 54, 58, 59, 60, and 72.

2014 YUAN, Q.; CONG, G.; SUN, A. Graph-based point-of-interest recommendation with geographical and temporal influences. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2014. 659–668 (CIKM '14). ISBN 978-1-4503-2598-1<http://doi.acm.org/10.1145/2661829.2661983>.

2015a ZHANG, J.-D.; CHOW, C.-Y. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2015. 443–452 (SIGIR '15). ISBN 978-1-4503-3621-5<http://doi.acm.org/10.1145/2766462.2767711>.

2015b ZHANG, J.-D.; CHOW, C.-Y.; ZHENG, Y. Orec: An opinion-based point-of-interest recommendation framework. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2015. 1641–1650 (CIKM '15). ISBN 978-1-4503-3794-6<http://doi.acm.org/10.1145/2806416.2806516>.

ZHENG, Y. Tutorial on location-based social networks. In: CITESEER. *Proceedings of the 21st international conference on World wide web, WWW*, 5 12 2012. Quoted on the page 28.

ZHENG, Y. et al. Recommending friends and locations based on individual location history. *ACM Transactions on the Web (TWEB)*, 5 1 5 ACM, 2011. Quoted on the page 28.