Scalable models for points-of-interest recommender systems

Jean-Benoît Griesner

▶ **To cite this version:**

Jean-Benoît Griesner. Scalable models for points-of-interest recommender systems. Information Retrieval [cs.IR]. Télécom ParisTech, 2018. English. NNT : 2018ENST0037 . tel-02085091v2

HAL Id: tel-02085091
https://theses.hal.science/tel-02085091v2

Submitted on 27 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0037

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TÉLÉCOM ParisTech

Spécialité « Informatique »

présentée et soutenue publiquement par

Jean-Benoît Griesner

le 3 Juillet 2018

Systèmes de recommandation de POI à large échelle

Directeurs de thèse: **Talel Abdessalem** et **Hubert Naacke**

Jury

M. Talel Abdessalem, Professeur, Télécom ParisTech	Directeur de thèse
Mme. Florence d'Alché-Buc, Professeur, Télécom ParisTech	Examinatrice
M. Stéphane Bressan, Professeur associé, National University of Singapore	Rapporteur
M. Amin Mantrach, Directeur de Recherches, Criteo Labs	Rapporteur
M. Hubert Naacke, Maître de Conférences, Université Pierre et Marie Curie	Directeur de thèse
M. Fabrice Rossi, Professeur, Université Paris 1 Panthéon-Sorbonne	Examinateur

TÉLÉCOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Abstract

The task of points-of-interest (POI) recommendations has become an essential feature in location-based social networks. However it remains a challenging problem because of specific constraints of these networks. In this thesis I investigate new approaches to solve the personalized POI recommendation problem. Three main contributions are proposed in this work.

The first contribution is a new matrix factorization model that integrates geographical and temporal influences. This model is based on a specific processing of geographical data. The second contribution is an innovative solution against the implicit feedback problem. This problem corresponds to the difficulty to distinguish among unvisited POI the actual "unknown" from the "negative" ones. Finally the third contribution of this thesis is a new method to generate recommendations with large-scale datasets. In this approach I propose to combine a new geographical clustering algorithm with users' implicit social influences in order to define local and global mobility scales.

La recommandation de points d'intérêts (POI) est une composante essentielle des réseaux sociaux géolocalisés. Cette tâche pose de nouveaux défis dûs aux contraintes spécifiques de ces réseaux. Cette thèse étudie de nouvelles solutions au problème de la recommandation personnalisée de POI. Trois contributions sont proposées dans ce travail.

La première contribution est un nouveau modèle de factorisation de matrices qui intègre les influences géographique et temporelle. Ce modèle s'appuie sur un traitement spécifique des données. La deuxième contribution est une nouvelle solution au problème dit du feedback implicite. Ce problème correspond à la difficulté à distinguer parmi les POI non visités, les POI dont l'utilisateur ignore l'existence des POI qui ne l'intéressent pas. Enfin la troisième contribution de cette thèse est une méthode pour générer des recommandations à large échelle. Cette approche combine un algorithme de clustering géographique avec l'influence sociale des utilisateurs à différentes échelles de mobilité.

Remerciements

L'épilogue de mon doctorat approche. Cette perspective me conduit à remercier chaleureusement toutes les personnes qui ont rendu possible cette aventure, certes trop rapide, et sans lesquelles ces trois dernières années n'auraient pas eu la même saveur. Qu'elles reçoivent ici l'assurance de ma sincère reconnaissance.

Je voudrais en première intention témoigner toute ma gratitude à mes deux directeurs de thèse, Messieurs Talel Abdessalem et Hubert Naacke, pour leur soutien inconditionnel. Ils ont su me guider avec pertinence et bienveillance et orienter à bon escient mes recherches. Sans leur aide et leurs conseils éclairés l'écriture de cette thèse n'aurait sans doute pas été achevée.

J'exprime également mes remerciements à Messieurs Stéphane Bressan et Amin Mantrach, rapporteurs de ce manuscrit, ainsi qu'aux autres membres de mon jury de soutenance, Madame Florence d'Alché-Buc et Monsieur Fabrice Rossi, pour leur disponibilité et leurs nombreux avis.

J'ai bien conscience d'avoir eu l'opportunité de bénéficier à Télécom ParisTech d'un environnement propice et stimulant. Je remercie à cet égard chaque membre de l'équipe DBWeb pour les incessantes réflexions partagées et les amicales discussions.

Enfin j'adresse à mes parents Patrick et Annie et à mon frère Jean-Baptiste mes pensées les plus aimantes et affectionnées.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We propose in this thesis new efficient methods in order to recommend personalized and relevant points-of-interest to users. To this end, this first chapter is aimed at proposing a global overview of the work carried out throughout this thesis. In particular we describe our research motivation in Section 1.1. Some first definitions come in Section 1.2. Then we describe our general objective in Section 1.3 and our research goals in Section 1.4. We present the contributions that we achieved in Section 1.5. Finally we present briefly the structure of the thesis in Section 1.7.

1.1 Research Motivation

The development of the Web 2.0 [Lewis 2006] these last years has promoted the emergence of a large number of Location-Based Social Networks (**LBSNs**) or online networks with location-based features such as Twitter, Facebook, Google+, Foursquare, Flickr and so on, which have changed deeply our vision of our environment and how we interact with it. A LBSN is a special category of *online social network* [Klein, Ahlf, and Sharma 2015] whose its content is directly associated to our geographical world. As a consequence geography and locations have both a crucial impact first on LBSNs structure and also on the quality of the services provided to the users and on the way their personal data are processed. Indeed LBSNs propose many different location-based services to their users ranging from transport to weather and news or recommendation services for instance. These services are especially interesting for the user facing a new or unknown environment. As a result these networks have developed more and more technologies, services and supports to help their users who want to explore and discover this unknown environment. Moreover users are now so used to interact with these online services that a new *information need* has emerged. Due to this information need and thanks

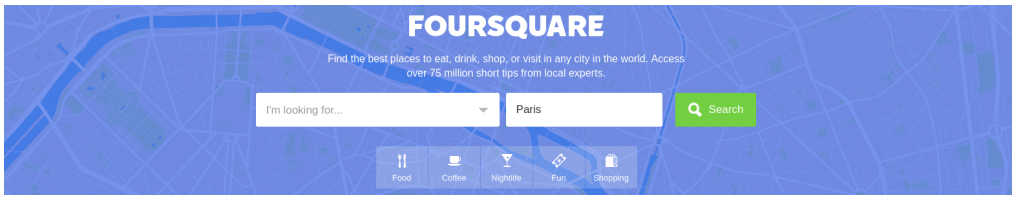


Figure 1.1: The POI search engine of https://foursquare.com/.

to these new services LBSNs constitute nowadays the most abundant sources of available information related to the global users' preferences, habits and activities [Chorley, Whitaker, and Allen 2015].

Indeed the amount of personal information and resources shared on these LBSNs has risen exponentially these last years [Cui, Hero, Luo, and Moura 2016]. For instance on the LBSN Flickr[1] there are more than 110 millions users who produce more than one million images per day. Because of this *information overload* [Toffler 1970], it has become increasingly difficult for the users to find what they are looking for in their surroundings. For instance it is common for a user who is looking online for a restaurant abroad to be overwhelmed by the intractable quantity of information to consider. As a consequence different POI search engines have been developed these last years to meet this need. An example of such a POI search engine can be seen on Figure 1.1.

To address this information overload problem, *Recommender Systems* (**RS**) have become an essential technology. The recommender systems most general purpose is to provide a personalized assistance to the users who require help for searching, ranking or filtering the large amount of information available on LBSNs. More precisely the main goal of recommender systems [Adomavicius and Tuzhilin 2005] is to propose to users personalized recommendations that are useful to discover *interesting and new* items that users would have probably not discovered on their own. These systems are now widely adopted by online business platforms in varied contexts ranging from books (Amazon[2]), movies (Netflix[3]), music (Spotify[4]) or Points-Of-Interest (**POIs**) with applications such as Foursquare[5]. As said previously, these platforms are usually characterized by the large spaces of shared data they manage: 500 million messages exchanged everyday on Twitter by 248 million

[1] https://www.flickr.com/
[2] https://www.amazon.com/
[3] https://www.netflix.com/fr/
[4] https://www.spotify.com/fr/
[5] https://fr.foursquare.com/

users, 200 million products in Amazon, 30 million songs in Spotify, 10,000 movies in Netflix... Such volumes of candidate items to explore impose harsh practical limitations for the user who wants to filter, search or select the interesting online information. Indeed without an efficient online assistant it becomes impossible for the user to navigate in these large spaces. As a consequence the importance to propose highly accurate recommendation lists and efficient filtering tools has become a major issue in this context.

Many recommendation problems have been investigated these last years in a large number of topics and domains, ranging from music recommendation [Cheng, Shen, and Mei 2014] to news recommendation [Hsieh et al. 2016] and movies recommendation [Gantner, Rendle, and Schmidt-Thieme 2010]. As a result one might expect that it exists now a large number of models and approaches to solve most of recommendation difficulties. However the problem of points-of-interest recommendation involves several specific challenges that distinguish it from traditional recommendation tasks (such as books, music or movies...) especially because of geographical influence, side information, user mobility and implicit user feedback. In this thesis we address such recommendation-specific challenges. More precisely we investigate the impact of these challenges on the recommendation quality and we propose new approaches to solve them.

point-of-interest name	Eiffel Tower
#Checkins	7.097.302
Location	Latitude: 48.858° Longitude: N, 2.294° E
Categories	scenic views, monument, entertainment, leisure

Table 1.1: An example of a points-of-interest and its associated information

Motivated by these specific constraints and challenges, the problem of geographical items recommendation and specifically the problem of **POI recommendation** has received an increasing level of interest in the academic world in the last years [Jing, Xin, and Lejian 2017]. Therefore a large number of works have been proposed to address this problem these last years. Existing works span from the industry to the academic world, especially in top tier conferences in computer science such as ACM RecSys [Baral and Li 2016], KDD [Li, Ge, Hong, and Zhu 2016], WWW [Ying, Chen, Xiong, and Wu 2016], CIKM [Xie et al. 2016], IJCAI [Jing, Xin, and Lejian 2017], SIGIR [Yuan, Cong, Ma, et al. 2013] and many more. All these approaches aim at combining different existing layers of information into one recommendation model. However the information layers contain complex objects

with geograhpical and temporal dimensions that are not easy to integrate into any existing model.

1.2 Points-Of-Interest and Social Networks

A *point-of-interest*[6] is a uniquely identified specific site generally associated to a specific category of activities (*e.g.* museum, restaurant, university etc.). Similarly to checkins (defined below) in LBSNs a point-of-interest is generally also associated with some content which corresponds to the set of all comments, pictures, opinions that users have uploaded during the checkins they made. For instance in table 1.1 the point-of-interest *Eiffel Tower* is associated with some of the categories it belongs to. However in many practical cases the categories or other point-of-interest descriptions are not disclosed for various reasons (*e.g.* privacy, confidentiality etc.). This is why in our approaches we have assumed that we only know the locations, *i.e.* the pairs $(latitude, longitude)$ for all points-of-interest.

On the other hand the *checkins* correspond to the visits made by users in points-of-interest. Therefore checkins are always associated at least with a POI and a location (*i.e.* a pair $\{longitude, latitude\}$) and a date (*e.g.* a timestamp). These information are required to deal with geographical and temporal dimensions. Checkins can also be associated to different content. However taking into account of these metadata requires more complex input models, which eventually increases the training duration and the computational complexity of the recommendation model.

We propose on figure 1.2 an overview of the standard structure of a LBSN. On this figure we distinguish three main layers: first the map and the points-of-interest (*i.e.* the **geographical layer** or the **physical layer**), then the users (*i.e.* the **social layer**) and finally the content shared online by the users (*i.e.* the **content layer**). The information contained by these layers come with specific constraints and characteristics that are likely to influence the final quality of the models and that require to be taken into account. For instance it has been demonstrated that the geographical layer content has the most significant impact on the recommendation final quality [Lian et al. 2014]. As a result it is necessary to consider how to exploit all the layers in order to increase the efficiency of our recommendation models. Unfortunately the information contained by the content layer are often not disclosed for privacy purpose. Moreover it does not exist any universal method to manage these data. For these reasons we will not exploit the content layer

[6] All necessary definitions are given in Section 1.6

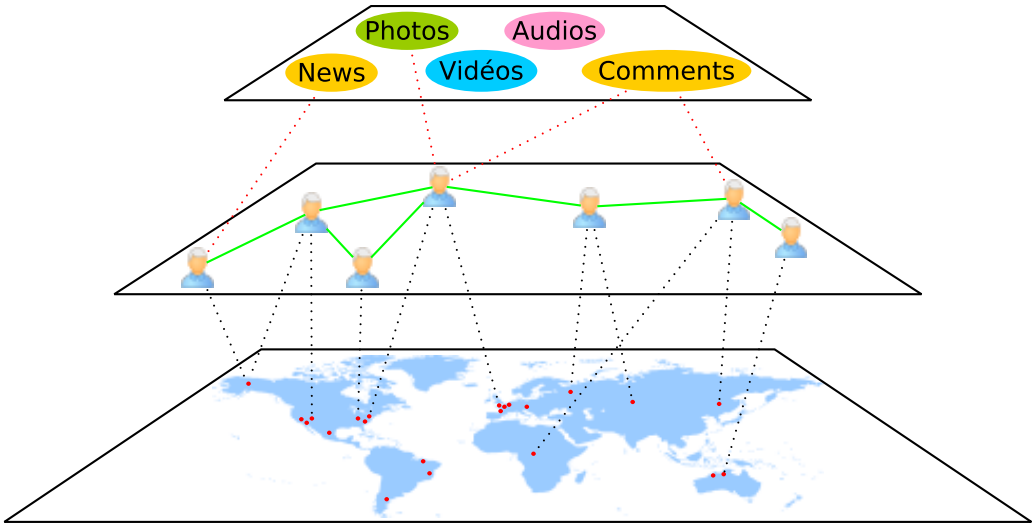


Figure 1.2: Three layers of the information layout in LBSNs.

information directly in this thesis.

1.3 General Objectives

The general objectives of this thesis are twofold. First we investigate and propose new matrix factorization methods to address the POI recommendation problem based on LBSN data. Actually existing matrix factorization methods are not designed to integrate the items locations directly into their models, which results in a poor recommendation quality. Especially we consider that the improvement in terms of "quality" of recommendations has to take into account of how users explore their surroundings and what are their specific mobility patterns. The *quality of the recommendation*[7] can be defined either in terms of ranking or in terms of prediction. The prediction quality aims at generating a recommendation score that reflects directly the preference of a given user for a given POI. In this case the score is an estimation of to which extent the user appreciates a given POI. On the other hand the ranking quality aims at optimizing the ranking of points-of-interest for a given target user. In the case of ranking the computed score gives only a possibility to sort the top-k list recommended.

Our second objective is to address the problem of scalability of such methods. Indeed, as said briefly in the previous section, most of recommendation algorithms

[7]We detail more precisely how the recommendation quality can be measured in Chapter 2

generally suffer from large volume of data. This large volume of data involves a large number of points-of-interest which makes the recommendation models inefficient. For instance most of factorization methods have a quadratic complexity [Hu, Koren, and Volinsky 2008]. The other issue of scalability is due to the large geographical target area that increases significantly the complexity to explore all possible points-of-interest in this area. As a result our goal is to adapt these models and also to investigate new techniques to compute efficient models on very large volume of data such as the YFCC dataset proposed by [Thomee et al. 2016].

1.4 Research Goals

The most general objective of this thesis, as described above, is to improve the quality and the scalability of POI recommendation approaches based on LBSN data. To do this we have been led to pursue the following *research goals* (or **RG**):

RG n°1: Survey of existing methods regarding point-of-interest recommendation. As described in the previous sections, it exists in the related work a large number of methods and models for POI recommendation. However a clear understanding of the effective advantages and shortcomings between these models is still missing. Therefore we inspect and propose a complete panorama of the most efficient techniques and approaches (c.f. Section 2).

RG n°2: Explore and improve matrix factorization approaches. Since the Netflix Prize [Bennett, Lanning, and Netflix 2007] we know that factorization approaches are the most efficient among collaborative filtering methods to face sparsity issues. So we investigate factorization methods in order to connect them with specificities of point-of-interest recommendation. In particular the goal here is to integrate geographical and temporal influences.

RG n°3: Investigate the probabilistic framework for factorization models. Probabilistic rules and assumptions allow to build more flexible and efficient methods that allow to enhance the quality of the model. Therefore we propose new probabilistic approaches to better take into account of sparsity issue and contextual information as well.

RG n°4: Enhance scalability of factorization approaches. Most point-of-interest recommendation techniques fail to handle large volumes of users and POI. As a result most of existing experimental datasets used to test in literature

are several order of magnitude smaller than real-world datasets. This is why we aim at exploring possible solutions to alleviate this issue. We investigate especially geographical clustering methods to tackle the sparsity and the scalability problems.

RG n°5: Explore geographical users' mobility patterns. LBSN provide a rich and precise source of information regarding users habits. This source of information can be exploited to enhance our understanding of the geographical mobility patterns of users. More precisely we exploit the observation of different scales of mobility: some users tend to travel through the whole world, on long distance, while others will concentrate their checkins in local areas.

1.5 Contributions

The work conducted throughout this thesis has resulted in several achievements in the area of POI recommendation. This section describes briefly our main contributions.

Contribution n°1: A geographical matrix factorization model with time dependencies: GeoMF-TD. We have proposed a factorization model that takes into account of the spatio-temporal distributions of checkins in the data. GeoMF-TD divides the given target region in a grid of even cells. This grid is exploited then to model the geographical and temporal latent influences of POI and users' activity through the cells of the grid. These latent influences are then combined linearly with latent vectors to compute the recommendation score. We have proposed such a new model in [Griesner, Abdessalem, and Naacke 2015]

Contribution n°2: GeoSPF, an approach to the implicit feedback problem based on a Poisson factorization model. Poisson factorization has been exploited successfully for various recommendation problems. Based on a Poisson factorization model, we have enhanced the contextual information influence by building an implicit social network. This implicit social network is based on geographical preferences of users. We have investigated this line of research in [Griesner, Abdesssalem, and Naacke 2017]

Contribution n°3: A new model for users' personal geographical mobility patterns. We usually observe in LBSN datasets different user profiles: some users tend to make long distance trips whereas other users only do checkins limited to a local area. We have exploited this observation to face the scalability issue by

applying a hierarchical geographical clustring method.

Contribution n°4: ALGeoSPF, a large-scale extension of GeoSPF. Based on previous contributions we have proposed a new approach that can build personalized POI recommendations on large-scale datasets. This work has resulted in our ALGeoSPF model. We have presented this work in [Griesner, Abdesssalem, Naacke, and Dosne 2018].

1.6 General Definitions

In this section we give the main definitions of terms and expressions used throughout this thesis. A list of notations used in the following is also proposed in Table 1.2 below. We define first what a point-of-interest is:

Definition 1.6.1. *(**Point-of-Interest**) A point-of-interest is a uniquely identified specific site associated to a specific activity (e.g. museum, restaurant, university...).*

In LBSN a POI is generally also associated with some content. For instance in table 1.1 the POI is associated with some of the categories it belongs to. However in many practical cases the categories or other POI descriptions are not disclosed for various reasons (*e.g.* privacy). In our thesis we assume that we only know the location, *i.e.* the pair $(latitude, longitude)$ of every POI. Because of this reason in this thesis the terms *point-of-interest*, *location*, *spot*, *place*, *site* can designate indifferently the same thing. In the same way we can define the check-ins made by users into points-of-interest as follows:

Definition 1.6.2. *(**Check-in**) The check-in activity of a user u visiting a POI p at a time t is a tuple $< u, p, t >$.*

To compute the visit frequency of u in p we count how many corresponding check-ins have been made. Given that in our approach each POI is associated with at least one super-POI, each check-in of a POI increments the corresponding super-POI visit frequency. We draw on figure 1.3 a standard representation of an LBSN with points-of-interest and check-ins. We observe that most of LBSNs have in common the following data attributes: a user set $\mathcal{U}$, a POIs set $\mathcal{P}$, some temporal information $\mathcal{T}$ and a social network $\mathcal{S}$.

In this framework, basically, a user u can make a check-in at some POI p at a time t. These check-ins constitute the user profile of the user as defined below:

Definition 1.6.3. *(**User Profile**) A user profile is the set of all the check-ins that the user made in the past:* $\mathcal{P}^u = \{< u, p_i, t_j > / < u, p_i, t_j > \in \mathcal{D}\}$. *Each user is associated to her profile. The aggregation of all user profiles constitutes the full dataset* $\mathcal{D} = \{\mathcal{P}^u / u \in \mathcal{U}\}$. *The user profile can also be defined as a set of check-ins sequences. A sequence of check-ins that the users made between a* K *number of consecutive points-of-interest can be noted as follows:* $\{p^1 \rightarrow p^2 \rightarrow ... \rightarrow p^k / p \in \mathcal{P}\}$.

We propose now to define more the specific recommendation problem that this thesis aims at solving. For a given set of points-of-interest: $\mathcal{P} = \{p^1, ..., p^m\}$ and for a given set of users: $\mathcal{U} = \{u^1, ..., u^n\}$, each user is associated with a chronologically ordered set of points-of-interest $\mathcal{L}^u$ visited by the user u such that $\mathcal{L}^u = \{p_u^1 \rightarrow p_u^2 \rightarrow ... \rightarrow p_u^k\}$ where $k = |\mathcal{L}^u|$, we define the problem of *points-of-interest recommendation*:

Definition 1.6.4. *(**Points-of-interest recommendation problem**) is the problem of recommending for any user* $u \in \mathcal{U}$ *a top-k list* $\hat{\mathcal{L}}^u$ *of new unvisited points-of-interest, that is to say points-of-interest that belong to the set* $\mathcal{P} \setminus \mathcal{L}^u$, *that are the most likely to match the user preferences and thus to be visited by* u.

We can distinguish two main cases[8]. Indeed we say that the recommendation can be either *generic* if the system proposes points-of-interest without considering the given user, *i.e.* recommends the same list for any user as follows: $\forall u, \hat{\mathcal{L}}^u = \hat{\mathcal{L}}^{Gen}$ On the other hand the recommendation is said *personalized* if the recommendation result depends effectively on the user, as follows: *i.e.* $\forall u, \hat{\mathcal{L}}^u = \hat{\mathcal{L}}^{Pers}[u]$. Generic techniques allow the recommender system to produce recommendation for a large variety of different items: they are called *domain independent*. On the counterpart they perform usually poorly because of the lack of contextual information. On

[8]These cases are described in Chapter 2

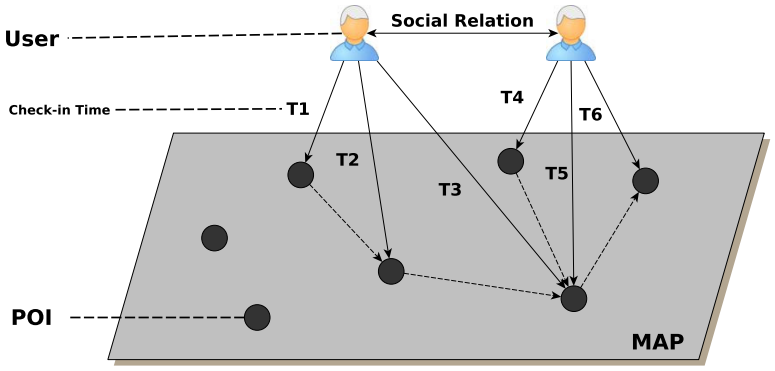


Figure 1.3: Standard Location-based Social Network Components.

Table 1.2: Table of Notations

$\mathcal{U}$	Set of all users $\{u_1, u_2, ..., u_{\lvert\mathcal{U}\rvert}\}$
$\mathcal{P}$	Set of all POIs $\{p_1, p_2, ..., p_{\lvert\mathcal{P}\rvert}\}$
$\mathfrak{P}$	Family of all layers of super-POIs: $\mathfrak{P} = \bigcup_k \mathfrak{P}^k$
$\mathcal{P}^u$, $\mathcal{E}^u$	Sets of POIs and edges visited by user u
$< u, p, t >$	Check-in of user u visiting POI p at time t
$\mathcal{D}$	Collection of all check-ins of all users visiting all POIs: $\mathcal{D} = \{< u_i, p_i, t_i >\}_{i=1}^{\lvert\mathcal{D}\rvert}$
$\mathcal{T}_{j,j+1}$	Transition probability betwen POIs j and $j+1$
$\mathcal{A}_{j,j+1}$	Accessibility between POIs j and $j+1$
$\mathbf{G}$	Geographical accessibility digraph $\mathbf{G} = (V, E, \rho)$
$\Gamma(X)$	Set of POIs accessible in one hop from X $\Gamma(X) = \{p' \in V \lvert p \in X \wedge \mathcal{A}_{p,p'} > 0\}$
$\mathbf{X} = [x_{up}]$	The $\lvert\mathcal{U}\rvert \times \lvert\mathcal{P}\rvert$ user-POI check-in matrix
x_{up}	Visit frequency of user u in POI p
y_{up}	Recommendation score of POI p for user u
$\mathbf{u}_i$	Vector of $\mathbb{R}^{m \times k}$ user latent factors
$\mathbf{v}_j$	Vector of $\mathbb{R}^{n \times k}$ POI latent factors

the other hand, personalized techniques result in a better final quality, but require more complex models.

Based on this definition we could consider two specific instances of the point-of-interest recommendation problem. The first one is a user that has made all her checkins in a given city $\mathcal{C}$. Consequently her associated history list $\mathcal{L}^u$ contains only points-of-interest close from each other. For this user the recommendation model has to deduce what is the maximum distance that the user could accept for visiting a relevant POI: if the relevant POI is too far away, the user will probably not visit it.

Reciprocally the second instance is a user who made checkins in a wide geographical area (e.g. on all continents). We claim that for this user the distance will not be a serious constraint. However we could face a scalability issue given that more points-of-interest will have to be considered. We observe here that an efficient recommendation model has to detect these two patterns and adapt its parameters to the target user.

1.7 Structure of the Thesis

We have structured our thesis as follows.

- Chapter 1 introduces our thesis. That is to say it presents the research motivation of our problem, the general objectives we want to complete, the contributions that we achieved and general definitions used throughout this thesis.
- Chapter 2 proposes first a global overview of the recommendation concept and process. Then the second part of this chapter is dedicated to a specific introduction to points-of-interest recommendation. Finally it concludes by describing the position of the techniques proposed in our work among the related works.
- Chapter 3 dives into our geographical matrix factorization model named GeoMF-TD. We present its structure and main ideas.
- Chapter 4 introduces geoSPF, our Poisson-based factorization approach that integrates geographical and social influences.
- Chapter 5 presents ALGeoSPF, an extension of GeoSPF that takes into account of the specific mobility patterns of the users in a personalized way.
- Chapter 6 contains the conclusion that we reached throughout this thesis and proposes an insight of the possible future works.

Chapter 2

A Survey on Points-Of-Interest Recommender Systems

*Investigating the related work in areas regarding the problem of Points-Of-Interest (**POI**[1]) Recommender Systems (**RS**) represents an essential part of the work that has been conducted in this thesis. Our purpose is to present a comprehensive and systematic exploration of algorithms and ideas from this field. Thus in this chapter we propose a review of main existing approaches and technologies. Specifically we have investigated works linked to geographical, social and temporal influences for POI RS. In particular we propose a global classification of main existing POI recommendation algorithms. We start by defining the general recommendation process in Section 2.1. Then we delve into the specificities of POI recommendation in Section 2.2. Finally we propose an overview of the models we propose in this thesis in Section 2.3.*

2.1 A Recommender Systems Overview

The most general goal of a recommender system (also noted RS) is to suggest to online users items to consume or to select [Adomavicius and Tuzhilin 2005]. Most of the time these items are expected to be new or at least that she could not find on her own. Furthermore these items are also expected to match the user preferences and so to contribute positively to the user experience. This is why these systems propose to the user a personalized exploration of a large space of possible choices. Differently from pure information retrieval systems where the user navigates into this possible space of choices by expressing an explicit query, the RS is not aware "a priori" of what the user really wants or prefers. In other words in a standard

[1] In the following we will name indistinctively: "POI", "destination", "location" or "item".

recommendation scenario the RS have to infer the implicit personal information needs of the users. As a consequence, because there is no explicit query, the RS can only exploit all past interactions of the user with the system to generate recommendations. This is why recommender systems collect and analyse all past user's preferences in order to predict future preferences.

Moreover, from a business point of view the RS goal is to transform a standard user into a consumer. The business value here is to increase the conversion rate of the POIs owner [Ricci, Rokach, Shapira, and Kantor 2010]. This task is completed especially by enhancing the loyalty of the user to the system, which is done by improving her browsing experience. Thus the recommender systems business purpose is to improve the quality of the users' interaction with the system, and to help the user through a large space of possible relevant items to select and consume.

In this Chapter we provide an introduction to the recommendation scenario from a technical point of view. The first part of this introduction is mainly based on many exhaustive state-of-the-art studies and surveys in the field of recommendation, such as [Borris, Moreno, and Valls 2014; Barbieri, Manco, and Ritacco 2014] and many others. The second part of the introduction is mainly based on POI recommendation surveys. First, we provide in subsection 2.1.1 a formal background that introduces some notations used in the sequel. We present a brief classification of existing recommendation algorithms in subsection 2.1.2. Then we will discuss the challenges and evaluation involved by RS in subsections 2.1.3 and 2.1.4 respectively.

2.1.1 Background

As said in the introduction, recommender systems are tools that are used to lead the users to the items they prefer without asking any question [Adomavicius and Tuzhilin 2005]. More formally a recommender system generates to users personalized lists of K items as close as possible to the users' preferences. Modelling this recommendation scenario requires obviously, at least, three entities: *users, items* and *preferences*. Any other contextual entities or information related to the users or items can then be integrated a posteriori into the model. In the following we present some of the notations that we use in the sequel of this thesis to model these entities.

As said above, any recommendation scenario involves at least users and items. So let $\mathcal{U} = \{u_1, ..., u_M\}$ be a set of M users and let $\mathcal{I} = \{i_1, ..., i_N\}$ be a set of N

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	3		1						4	
u_2		5				2				2
u_3	1		2	4			2	1		4
u_4	3	4								2
u_5				1			5			
u_6	2	1				5		1		
u_7						4	2			1

Table 2.1: Illustration of a users' rating matrix $\mathbf{X}$.

items. For sake of clarity we traditionnally represent the users' preferences in a $M \times N$ matrix $\mathbf{X} = [x_{ui}] \in \mathbb{S}^{M \times N}$. This is why we define $\mathbf{X}$ the *user-item rating matrix*. The set $\mathbb{S}$ is the set that contains all possible values for elements x_{ui} of $\mathbf{X}$. In $\mathbf{X}$ the element x_{ui} represents the *preference value* of the user u for the item i. The set $\mathbb{S}$ is called the *domain of scores* and it can contain different types of scores: either bounded ratings (*e.g.* $\mathbb{S} = [1, 2, 3, 4, 5]$), or only two elements (*e.g.* $\mathbb{S} = \{like, dislike\}$), or in the case of behavior data simply frequencies of interactions: $\mathbb{S} = \mathbb{N}^+$ for instance. We propose an illustration of a possible rating matrix on table 2.1. In this example the scenario involves 7 users, 10 items and explicit preferences ordered from degree 1 to 5. Depending on the meaning associated to values in $\mathbf{X}$, this preference can be classified either as *explicit*, either as *implicit*.

- **Explicit data** corresponds to explicit ratings expressed by the users about the corresponding items. Most of the time these explicit ratings are gathered by asking directly to users their feedback regarding items they have consumed or interacted with in the past [Towle and Quinn 2000]. Then this feedback is converted into explicit preferences. This kind of preferences are more difficult to collect since it requires the user availability. On the counterpart they are easier to interpret than implicit data.

- On the other hand **implicit data** corresponds to raw observations of the dyads (u, i) in the dataset. We can view these implicit data as a *behavioral* information that has been recorded only by observing and collecting all past interactions between the users and the system, without asking directly the user's feedback [Oard and Kim 1998]. For example these behavioral information can be clicks in a browser, music listened, web sessions, events or check-in in POI. In this case it implies that the entry x_{ui} is a nothing more than binary value: $x_{ui} = 0$ means that u has not yet consumed i while $x_{ui} = 1$

denotes that u has effectively purchased i.

Behavioral - implicit - data is usually gathered in a silent and passive way, implicit feedback is usually easier to collect than explicit feedback, but it is often **unreliable**, given that the real effective users' evaluations remain hidden. On the other hand explicit feedback is usually less abundant but more accurate. Let's observe that explicit feedback can be either positive or negative, while implicit feedback is always positive. Implicit feedback corresponds here to the *One-Class Collaborative Filtering* problem [Pan et al. 2008].

Regarding **the dimensions**, usually the number of users M as well as the number of items N are very large [Ricci, Rokach, Shapira, and Kantor 2010] with typically: $M >> N$. This is why we say that in traditional real-world recommendation scenarios, the user-item rating matrix $\mathbf{X}$ is characterized by an extreme *sparseness*, given that users give their feedback for a (very) limited amount of available items. In the following we will note $\langle u, i \rangle$ the list of all dyads in $\mathbf{X}$ such that $x_{ui} > 0$.

More formally we usually note $\mathcal{I}_{\mathbf{X}}(u) = \{i \in \mathcal{I} | \langle u, i \rangle \in \mathbf{X}\}$ the set of items rated by user u. On the other hand the set $\mathcal{U}_{\mathbf{X}}(i) = \{u \in \mathcal{U} | \langle u, i \rangle \in \mathbf{X}\}$ will be the set of users that have selected/consumed the item i. If the context does not allow any ambiguity regarding the matrix $\mathbf{X}$ involved, we can simply note $\mathcal{I}(u) = \mathcal{I}_{\mathbf{X}}(u)$ and $\mathcal{U}(i) = \mathcal{U}_{\mathbf{X}}(i)$. Some users have not done any selection yet, so we note: $\mathcal{I}(u) = \varnothing$. Reciprocally we will say that any user u with a rating history, that is to say such that $\mathcal{I}(u) \neq \varnothing$, is an *active user*.

A classical problem appears when either $\mathcal{I}(u)$ or $\mathcal{U}(i)$ is empty (which means that a new user or a new item has been added to the dataset). We call this situation the *cold start* problem [Saveski and Mantrach 2014]. Cold start is generally problematic in RS, since these cannot provide suggestions for users or items if there is not a sufficient amount of information. For instance in table 2.1 the item i_5 has not been rated yet by any user. So this item will never be recommended by any RS given that we have not enough information concerning it.

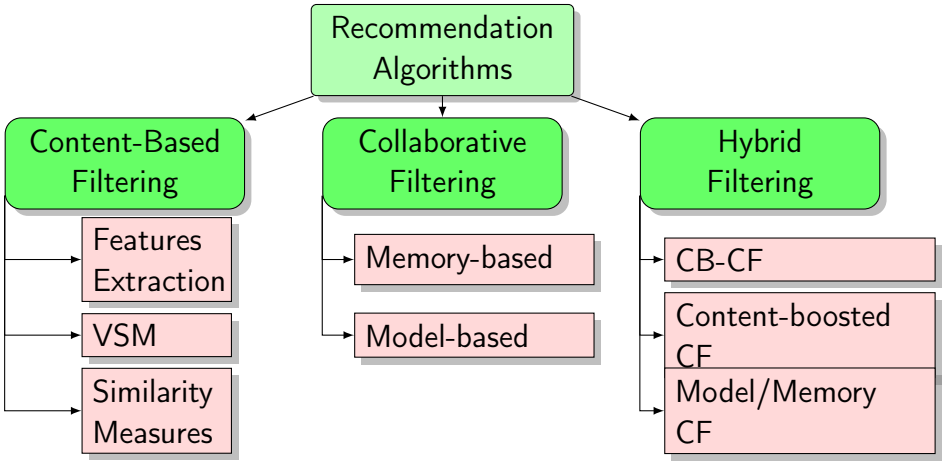
2.1.2 Algorithms Classification

Because of historical reasons [Ricci, Rokach, Shapira, and Kantor 2010], recommendations are generally generated by means of *filtering* or *retrieval* techniques. The main idea of these classes of approaches is to remove unwanted information from an information stream in the case of *information filtering* (**IF**, online), or

from an information database in the case of *information retrieval* (**IR**, offline).

In the context of recommendation, "unwanted" information corresponds to the least relevant items for the target user. To this end, IF/IR based recommendation methods exploit the assumption that human interests and preferences are correlated. According to this assumption a user is likely to prefer what other *similar* users have selected in the past. Thus the most intuitive technique is to collect information about user preferences and to explore similarities between users' profiles, and then to exploit the known preferences of similar users to build a prediction for the target user.

Filtering algorithms can be classified according to multiple criteria [Adomavicius and Tuzhilin 2005] such as the recommendation domain, the type of feedback, the contextual issues... However the most used classification focuses on the exploitation of the interaction between users, items and the system, and distinguishes three classes of algorithms: *content-based*, *collaborative* and *hybrid* that we present in details in the sequel.



2.1.2.1 Content-Based Filtering

Content-based algorithms (**CB**) try to find a matching between a *user's profile* and item *attributes*. This works in two steps. First the model learns users' preferences based on what they purchased in the past. This leads to a *user's profile representation*. Then the CB model ranks the items that are the most *similar* to those the user liked in the past [Pazzani and Billsus 2007]. This requires to have a common *item representation* for all items. Finally the model provides recommendation of unexperienced items based on this ranked list. Usually in CB filtering, item profiles

and user profiles are represented with a description such as a set of keywords or attributes. In the case where items are textual documents for instance, the keywords are simply the regular words of the language contained by the documents. As a result the user profile corresponds to the most *relevant keywords* of the items she purchased in the past. Once the model has learned each profile, the items are ranked according to a *similarity function*.

Vector Space Model (VSM). Specifically we denote $\mathcal{F} = \{f_1, ..., f_q\}$ a set of descriptive features (or attributes) for the items $\mathcal{I}$. As said previously, these attributes are usually keywords or scalars extracted from the side information associated to each item. Then the features are exploited to associate each item with a *features vector* representation. These vectors are then projected into an Euclidean space such as $\mathbb{R}^q$ which is called the *vector space model* (VSM) which is a traditional model in information indexing [Salton, Wong, and Yang 1975]. The recommendation score is then derived from a list of candidate items ranked thanks to the similarity function between vectors in the VSM. Let $\mathbf{w}_i \in \mathbb{R}^q$ be the feature vector associated with item $i \in \mathcal{I}$. Each component $\mathbf{w}_i^f$ represents the contribution weight of feature f for the item. The values $\mathbf{w}_i^f$ can be either binary, categorical or numerical depending on the data specifications.

TF-IDF. One of the most used method [Ricci, Rokach, Shapira, and Kantor 2010] to build these weights is the TF-IDF method, that gives for a given document $\mathcal{D}$ more importance to terms that appear frequently (TF≡*Term Frequency*), but also that penalizes terms that occur frequently in other documents (IDF≡*Inverse Document Frequency*). The TF-IDF function is defined as follows:

$$\text{TF-IDF}(t_k, d_j) = \underbrace{\frac{\text{freq}_{k,j}}{\max_z \text{freq}_{z,j}}}_{TF} \cdot \underbrace{\log(\frac{N}{n_k})}_{IDF} \tag{2.1}$$

where t_k is the target term, d_j the target document, N is the number of documents in the corpus, n_k the number of documents containing the term t_k and $\text{freq}_{k,j}$ refers to the frequency of term t_k in document d_j.

Once the items have a representation in the VSM, one can either apply machine learning techniques, or directly similarity function. For a given user's profile, the main idea is to classify items in two classes: $\mathcal{C} = \{c+, c-\}$ of *positive* and *negative* class depending on if the item is relevant or not for the user. Machine learning techniques have been widely exploited by CB approaches. However this is outside of our scope, so we will not present more this topic.

Similarity functions. Given a target user's profile, similarity functions are necessary to determine how relevant two candidate items are. It exists a large choice of possible ways to measure how close two vectors are. The most commonly used similarity functions are the following:

- **Minkowski distance**: This is a generalization of the notion of distance between two points in an Euclidean space. The distance between items i and j is defined as follows:

$$d_p^{Mink}(i,j) = \left(\sum_{l=1}^{q} |w_{i,l} - w_{j,l}|^p\right)^{\frac{1}{p}} \tag{2.2}$$

- **Cosine similarity**: It measures the similarity of two items with the angle between their corresponding vectors. This similarity function is especially used with sparse feature vectors. It is defined as follows:

$$sim^{Cos}(i,j) = \frac{\mathbf{w}_i^T \cdot \mathbf{w}_j}{\|\mathbf{w}_i\|_2 \cdot \|\mathbf{w}_j\|_2} \tag{2.3}$$

- **Jaccard similarity**: This well-known similarity measures how common features tend to be predominant in the set of features. It is defined as follows:

$$sim^{Jac}(i,j) = \frac{\mathbf{w}_i^T \cdot \mathbf{w}_j}{\mathbf{w}_i^T \cdot \mathbf{w}_i + \mathbf{w}_j^T \cdot \mathbf{w}_j - \mathbf{w}_i^T \cdot \mathbf{w}_j} \tag{2.4}$$

Advantages & Shortcomings. Because they are based only on content information and because they don't require any history of past interactions. This provides *two advantages* [Ricci, Rokach, Shapira, and Kantor 2010]. First CB methods guarantees user independence: the recommendation does not require other users information. Also thanks to this, CB methods don't suffer from the cold-start problem: even for new user or new item it will be possible to make recommendations. Another advantage is the system is more transparent: it is easy to explain the recommendation result based only on other items and not on other users.

On the other hand CB methods suffer from *three important limitations*. First they can only recommend items similar to those already purchased by the user. That is to say these methods don't explore the non-similar items, and hence tend to always recommend the same kind of items, without any diversity but with, possibly, a lot of redundancy. Another problem concerns the features extraction. In

the case of textual data, the features are easy and natural to extract. However for complex data this process is not solved easily yet, because of privacy issues for instance. Moreover it is impossible for CB methods to distinguish between distinct items with the same features, while in reality they could have different value for the user. Finally since the Netflix Prize [Bennett, Lanning, and Netflix 2007], it has been demonstrated that CB methods are globally less accurate than collaborative methods [Koren, Bell, and Volinsky 2009].

2.1.2.2 Collaborative Filtering

Differently from CB approaches, collaborative filtering (**CF**) does not require any description of items. The term *collaborative* here means [Schafer, Frankowski, Herlocker, and Sen 2007] that CF methods exploit *only* all users' past interactions with the system to make recommendations of items selected by the most *similar* users of the target user. The central assumption is that *users who adopted the same behavior in the past will tend to agree also in the future*. As a result CF models are naturally much simpler than CB models,because no side information, such as items description, is required. This makes CF methods more general and especially *domain independent*. Moreover CF methods allow a higher level of privacy, since no personal information is required. Another advantage of CF is that the more feedback the model receives, the more the recommendation will be accurate.

As presented in Table 2.2, collaborative approaches are generally classified in two classes in existing literature [Breese, Heckerman, and Kadie 1998] namely *memory-based* and *model-based*. Memory-based approaches exploit directly all the data of the user-rating matrix $\mathbf{X}$, while model-based approaches use only a compact representation of the matrix $\mathbf{X}$. Neighborhood-based methods are the most widely used approaches among memory-based methods: these methods exploit user/item similarities. Model-based are more personalized, since they work with a compact model for each user and each item. This compact model is then used to predict a recommendation score for each given pair (*user,item*). Globally memory-based approaches are more intuitive given that the recommendation scores are directly computed with the input data. On the other hand this requires a constant access to the whole dataset to produce the recommendations, which can imply serious issues when the data volumes increase. Model-based approaches don't suffer from this problem, as they only require a compact data model. Another difference is that neighborhood models are efficient to model local similarities while model-based methods are more efficient on global relationships. In the sequel of this section we

CF Class	Techniques	Advantages	Shortcomings
Memory-based CF	• Neighborood-based. • Item/User based Top-N.	• Implementation fast and intuitive. • New data does not require to build any new model. • Provide fast recommendations on small datasets.	• Provide poor quality recommendation when sparsity increases. • Cold-start is a problem because no user/item content model is built. • Problem of scalability.
Model-based CF	• Bayesian Networks. • Clustering Methods. • Latent factors Models. • Probabilistic Modeling	• Adress efficiently the sparsity and scalability problems. • Provide better prediction performance. • Make recommendations more intuitive and natural.	• Building the model is generally expensive. • A tradeoff has to be found between prediction quality and scalability. • Can loose some valuable information.

Table 2.2: Collaborative filtering classes: advantages & shortcomings.

present an overview of neighborhood-based and latent factor methods.

Neighborhood-Based Approaches. Based on the idea that users often ask to their friends advices regarding items, neighborhood-based approaches have naturally emerged [Desrosiers and Karypis 2011]. For a given pair (*user,item*) these methods exploit the intuition that the most similar users will tend to share the same preferences. Following this intuition, neighborhood-based approaches will use the preferences of the users the most similar to the target user in order to produce the recommendation score. The set of the most similar users constitutes the *neighborhood* of the user. The most used method is the k-nearest neighbors algorithm (or **KNN**). In KNN a similarity function denoted $\mathcal{S}(u, v)$ is used to estimate the degree of similarity of any users u and v. This function $\mathcal{S}(u, v)$ is then used to build for a target user u the set $\mathcal{N}(u)$ of the K most similar users. Then the recommendation score of user u for the item i is simply the average of the ratings that the neighbors have given to item i, as follows:

$$\hat{x}_{u,i} = \frac{\sum_{v \in \mathcal{N}(u)} \mathcal{S}(u, v) \cdot x_{u,i}}{\sum_{v \in \mathcal{N}(u)} \mathcal{S}(u, v)} \tag{2.5}$$

This *user-based* approach can be considered also as *item-based* by directly ag-

gregating the ratings that the target user has given to the K most similar items. As a result the Equation 2.5 becomes:

$$\hat{x}_{u,i} = \frac{\sum_{j \in \mathcal{N}(i)} \mathcal{S}(i,j) \cdot x_{u,j}}{\sum_{j \in \mathcal{N}(i)} \mathcal{S}(i,j)} \tag{2.6}$$

In equation 2.5 and 2.6, one of the most important term is the similarity function $\mathcal{S}(*,*)$. Indeed this function is used to select the neighborhood first, but to weight the prediction score as well. The *cosine similarity* and *Jaccard similarity* (presented in Section 2.1.2.1) are common choice. Another possibility is the *Pearson Correlation* defined as follows:

$$\mathcal{S}_{Pearson}(i,j) = \frac{\sum_{u \in \mathcal{U}(i) \cap \mathcal{U}(j)} (x_{ui} - \bar{x}_i) \cdot (x_{uj} - \bar{x}_j)}{\sqrt{\sum_{u \in \mathcal{U}(i) \cap \mathcal{U}(j)} (x_{ui} - \bar{x}_i)^2} \sqrt{\sum_{u \in \mathcal{U}(i) \cap \mathcal{U}(j)} (x_{ui} - \bar{x}_j)^2}} \tag{2.7}$$

The KNN model requires just a few number of events to compute similarities and thus offers a good solution to the sparsity issue. However the computation cost of the pairwise similarities for all user/item put a severe limitation to its exploitation. As a result, KNN will be efficient only for relatively small datasets. The authors of [Bell and Koren 2007] have proposed a scalable neighborhood-based approach. Their idea is based on a formal *neighborhood relationship model* to compute the similarity weights as a least square problem.

Latent Factor Approaches. Observed check-ins in the data can always be associated to several motivations: any user has a reason (personal, professional etc.) for having visited a place. In other words any check-ins can be explained by some *factors*. Based on this idea, *latent factor models* have emerged as a solution to decompose the overall user's preferences on a set of latent factors. These factors allow to represent the quality of the interaction between user's preferences and item attributes. These models have a long history. They have been widely exploited by dimensionality reduction methods [Maaten, Postma, and Herik 2008] and by *latent semantic indexing* [Deerwester et al. 1990]

Hybrid Filtering. Finally this class gathers a combination of algorithms of the two other classes. We will not investigate more this class here. We provide in section 2.2.3 some examples of these models.

2.1.3 Challenges

Usually RS have to face common issues relative to the quality, the quantity, the privacy or the security of the data. We propose in the sequel a short review of these issues.

Sparsity. Generally the order of magnitude of the number of distinct items proposed by RS to its users is about 50 millions. As a consequence even the most active users will not be able to consume more than a very limited part of this number of choices. It implies that the *density* of the user-rating matrix will be extremly low: usually the density is between 0.005% and 1.5%. This is a serious issue for the RS that is expected to produce accurate recommendations with such a poor input. We call this phenomenon the *reduced coverage* or *sparsity*. The problem of the *cold-start* is a similar shortcoming: we have defined this phenomenon in subsection 2.1.1.

Scalability. In a world where we are more and more used to real-time communications and an instantaneous access to the information, RS are expected to deliver suggestions as fast as possible. Given that RS are usually associated to large user-item databases (such as Amazon, Google News...), they require large computational resources in order to perform in time (or at max a few of milli-seconds). This demand requires to use scalable methods and an efficient data management. Latent factors approaches are a powerful solution to separate the learning phase (which is done *offline*) and a recommendation phase (*online*).

Obsolescence. The user-item database is not a static or closed system: new users and new items come every day, increasing the data volume. As a consequence a standard RS can become obsolete fast and, thus, unaccurate. To prevent this obsolescence issue, the RS have to update their model frequently through incremental techniques.

Privacy. RS are based on the exploitation of past interactions of the user with existing systems. As a result, RS collect personal data that can represent a serious threat to the individual privacy. Even with anonymous databases, the aggregation of several data sources can lead to the identification of any particular user. This issue can be serious when sensitive information is involved. Recent researchs in this domain have shown that we can keep good recommendation quality even with anonymous data.

Security. The security issue appears in this context when malicious users want to influence system's suggestions about items. Usually these malicious users use *fake*

profiles or *attacker profiles*, that is to say fictitious user identities.

2.1.4 Evaluation

The goal of RS evaluation is to measure the impact of the RS on the user's experience with the system [Ricci, Rokach, Shapira, and Kantor 2010]. An efficient RS is expected to generate a significant positive impact on the user's decision process. Generally this evaluation follows a protocol that provides a good idea of the RS quality. Then these evaluations are used to compare different recommendation algorithms and approaches. Most of the time the quality evaluation is *offline*, that is to say the user-rating matrix $\mathbf{X}$ is split into two matrices $\mathbf{T}$ and $\mathbf{S}$ used for training and test, respectively. Many metrics can be used to evaluate the accuracy of a RS [Karypis 2001]. As detailed previously, the RS purpose is to build a list L of items that the user is most probable to like. To estimate how efficiently the RS performs, we could compare for each given tuple $< user, item, rating >$ the recommendation score computed by the system and the effective rating value. Then the average mean of these comparisons can lead to a conclusion on the RS efficiency. Another possibility could be to estimate directly the quality of the recommended list. For instance we could sort the items choosen by each user in the test set and compare the recommended items order with the effective user's preferences order. Hence the quality of the RS accuracy can be evaluated either on the predicted scores, either on the predicted list L. In this part we review the three well-known classes of existing metrics.

Prediction Evaluation. In this category we measure on average how each recommendation score is far from the effective user rating. To do so we aim at minimizing the average error between the infered and effective scores.

- Mean Absolute Error (**MAE**):

$$MAE = \frac{1}{|\mathbf{S}|} \sum_{\langle u,i \rangle \in \mathbf{S}} |x_{u,i} - \tilde{x}_{u,i}| \tag{2.8}$$

- Mean Squared Error (**MSE**):

$$MSE = \frac{1}{|\mathbf{S}|} \sum_{\langle u,i \rangle \in \mathbf{S}} (x_{u,i} - \tilde{x}_{u,i})^2 \tag{2.9}$$

  and

$$RMSE = \sqrt{MSE} \tag{2.10}$$

- Mean Prediction Error (**MPE**):

$$MPE = \frac{1}{|\mathbf{S}|} \sum_{\langle u,i \rangle \in \mathbf{S}} \mathbb{1}(x_{u,i} - \tilde{x}_{u,i}) \tag{2.11}$$

Recommendation Evaluation. Here we estimate if the recommended set of items is close of the items effectively chosen by the target user. We aim at maximizing the recall, the precision or the F-measure.

- Recall:

$$Recall@N = \frac{1}{M} \sum_{u \in \mathcal{U}} \frac{|\mathcal{L}_u \cap \mathcal{T}_u|}{|\mathcal{T}_u|} \tag{2.12}$$

- Precision:

$$Precision@N = \frac{1}{M} \sum_{u \in \mathcal{U}} \frac{|\mathcal{L}_u \cap \mathcal{T}_u|}{|\mathcal{L}_u|} \tag{2.13}$$

- Hybrid:

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.14}$$

Rank Accuracy. Finally we could compare the recommended items sets and moreover compare the inner rank of each recommended item in the set.

- Kendall's coefficients

$$\mathcal{K}(\tau_u, \tilde{\tau}_u) = \frac{2 \cdot (\sum_{i,j \in \mathcal{I}} \mathbf{S}(\tau_u(i) \prec \tau_u(j)) \wedge \tilde{\tau}_u(j) \prec \tilde{\tau}_u(i))}{N(N-1)} \tag{2.15}$$

- Spearman's coefficients

$$\rho(\tau_u, \tilde{\tau}_u) = \frac{\sum_{i \in \mathcal{I}} (\tau_u(i) - \bar{\tau}_u)(\tilde{\tau}_u(i) - \bar{\tilde{\tau}}_u)}{\sqrt{\sum_{i \in \mathcal{I}} (\tau_u(i) - \bar{\tau}_u)^2 \sum_{i \in \mathcal{I}} (\tilde{\tau}_u(i) - \bar{\tilde{\tau}}_u)^2}} \tag{2.16}$$

Other Evaluation Metrics. Many other evaluation metrics exist, such as the novely, the serendipity, the diversity, the coverage... among many others.

2.2 POI Recommendation

Many recommendation services are provided together in most of LBSN, such as user recommendation, activity recommendation, or POI recommendation. POI recommendation is one of the most challenging problems that received attention both in the academic community (with international conferences dedicated specifically to this problem such as ACM RecSys[2]) and the industry community as well due to its business exploitation. Our aim in this section is to present a general overview of existing models and approaches proposed in literature. We start to define our problem in subsection 2.2.1. Then we provide details about distinct POI recommendation problems in subsection 2.2.2. Subsections 2.2.3 details hybrid methods. Then we explore graph-based approaches in subsection 2.2.4. Finally we investigate matrix factorization approaches in subsection 2.2.5. Notice that we present a comprehensive summary of all existing approaches proposed for POI recommendation in subsection 2.3.1.

2.2.1 Problem Definition.

Let $\mathcal{L} = \{l_1, ..., l_N\}$ be a set of locations. The set $\mathcal{L}$ corresponds to the set $\mathcal{I}$ defined in section 2.1. An element $l_j \in \mathcal{L}$ is called a *location* or a *POI*. Each POI l_j is associated to geographical coordinates (lat_j, lon_j). Each user $u \in \mathcal{U}$ is associated to a history of visited locations denoted $\mathcal{L}^u$. We use these sets $\mathcal{L}^u$ to populate the user-checkin matrix $\mathbf{X}$. Given the matrix $\mathbf{X}$, the problem of *POI recommendation* is to recommend for each user $u \in \mathcal{U}$ a *top-k* list of new POI, that is to say POI in the set $\mathcal{L} \setminus \mathcal{L}^u$, that are the most likely to match the user preferences, and thus to be visited by u. This recommendation can be either *generic* if the system proposes POI without considering the given user, or *personalized* if the recommendation result depends on the user. Unlike traditional recommendation challenges[3], POI recommendation comes with other specific challenges due to geographical, temporal and social influences. Existing approaches usually exploit one or two of these influences either in traditional collaborative method, or in a graph-based approach or in a matrix factorization approach. In the following we briefly present successively the specifications of these influences.

Geographical influence. According to Tobler's first law of geography [Miller 2004] *everything is related to everything else, but near things are more related than distant things.* It means that the user's willingness to check-in a POI is inversely

[2]RecSys: https://recsys.acm.org/
[3]detailed in subsection 2.1.3

proportional to her distance to this POI. In other words the more the POI is far, the less likely the user will visit it. This phenomenon is called the *spatial clustering phenomenon* (**SCP**) and has been widely exploited through a *power-law* assumption in most of existing works [Zhang and Chow 2013; Ye, Yin, Lee, and Lee 2011].

Social influence. Most related work have established that most friends have a small overlapping on their check-in POI [Zhang and Wang 2015; Cheng, Yang, King, and Lyu 2012]. However the overlap is larger than non-friends, and so still interesting to exploit.

Temporal influence. Usually users check-in restaurants during lunch time, while bars are checked-in around midnight. So different users can behave similarly or differently with respect to time. Reciprocally different POI are expected to have different opening hours and a non-uniform distribution of check-ins through time. These two information have been taken into account by few related work yet, including [Gao, Tang, Hu, and Liu 2013; Zhang and Wang 2015].

2.2.2 Different POI Recommendation Problems

The POI recommendation problem described above is the most general case: the RS receives a request $\mathcal{Q}(\mathcal{L}^u)$ that depends only on the user history $\mathcal{L}^u$. However it exists several sub-problems more specific that usually exploit side information to perform a similar task. We present these similar tasks in the sequel.

Next POI Recommendation. In this case the user's request depends also on the current location of the user: $\mathcal{Q}(\mathcal{L}^u, l^u_{current})$. The goal of this problem is to make recommendations for a given location and the current user's location. That is to say the system will take into account of the *visit sequences* [Feng et al. 2015; Cheng, Yang, Lyu, and King 2013]. However most of existing works facing this problem exploit techniques and methods used in traditional POI recommendation.

Time-aware POI Recommendation. In this problem the request that is received by the RS is $\mathcal{Q}(\mathcal{L}^u, t^u_{current})$. As the previous problem, here the recommendation has to take into account of the evolution of user preferences through time. The authors of [Yuan, Cong, Ma, et al. 2013] propose a *user-time-POI* cube to model the temporal influence.

POI Itinerary Recommendation. Many approaches have been proposed to recommend a list of POI subject to a budget in time and/or money. This is what

the authors of [Zhang, Liang, Wang, and Sun 2015] have investigated: they add two strong constraints on the NP-hard *optimal route problem* in order to propose a personalized solution. The authors of [Lucchese et al. 2012] propose a random walk approach to maximize the touristic experiences of users between POI.

In-town/Out-of-town POI Recommendation. This problem separates the problem depending on the location of the geographical area with a city. Some works have been conducted [Ference, Ye, and Lee 2013] to show that POI recommendation out-of-town gives worse results than in-of-town POI recommendation. As a result the authors propose to use different parameter settings for these two different situations. The authors of [Wang, Yin, et al. 2015] have proposed a sparse additive generative model for spatial item recommendation that exploits latent topic distribution to face this problem.

2.2.3 Hybrid Collaborative Filtering Models

Based on the observation that each model family has advantages and shortcomings, many approaches aim at combining the advantages of distinct methods, while minimizing their shortcomings. So *hybrid models* combine several recommendation methods. In this part we present briefly some famous hybrid models.

2.2.3.1 iGSLR: Geo-Social Location Recommendation

This model presented in [Zhang and Chow 2013] integrates geographical and social influences. The social influence is computed with an approach inspired by *friend-based collaborative filtering* proposed by [Ma, King, and Lyu 2009] and by [Ye, Yin, Lee, and Lee 2011]. In iGSLR the social similarity between users u_i and u_j is computed as follows:

$$SGSim(u_i, u_j) = 1 - \frac{distance(u_i, u_j)}{\max_{u_f \in F(u_i)} distance(u_i, u_f)} \tag{2.17}$$

where $F(u_i)$ corresponds to the set of friends of user u_i. The geographical influence is computed with a classic *kernel density estimation* (KDE) done for each user check-in history such as:

$$\tilde{f}(d_{i,j}) = \frac{1}{|D|\,h} \sum_{d' \in \mathcal{L}^u} K\left(\frac{d_{i,j} - d'}{h}\right) \tag{2.18}$$

Based on the resulting distribution, the approach gets then a probability that a user i visit a POI j computing the distances between j and all POI visited by i as

follows:

$$p(l_j|\mathcal{L}^u) = \frac{1}{n}\sum_{i=1}^{n}\tilde{f}(d_{i,j}) \tag{2.19}$$

This approach has two main limitations. First the KDE requires to know where is the home location of each user, while this information is usually not displayed in most of LBSN. The solution to this problem is to assume the home location considering the locations of the most frequent check-in. However this assumption creates a significant bias in the model, and furthermore is not relevant in the context of foreign trips. The second limitation of this model is due to its complexity. Indeed the KDE requires to compute the distance between each pair of visited POI for each user, which is impossible for real-world datasets.

2.2.3.2 GeoSoCa: Geographical, Social and Categorical Correlations

The authors of [Zhang and Chow 2015] have proposed to exploit geographical, social and categorical correlations existing in the data to improve the accuracy of the RS. The geographical correlations are computed in a similar way of iGSLR, with an adaptive kernel density estimation of the geographical relevance score as follows:

$$f_{Geo}(l|u) = \frac{1}{N}\sum_{i=1}^{n}(\mathbf{X}_{u,l_i}\cdot K_{Hh_i}(l-l_i)) \tag{2.20}$$

where differently from iGSLR the kernel $K_{Hh_i}(l-l_i)$ is here a geographically adapted kernel. Then the following social correlation term is computed:

$$F_{So}(x_{u,l}) = 1-(1+x_{u,l})^{1-\beta} \tag{2.21}$$

The third term measures the categorical relevance score between the user and the location. It is computed as follows:

$$F_{Ca}(y_{u,l}) = 1-(1+y_{u,l})^{1-\gamma} \tag{2.22}$$

Finally, the final recommendation score for the pair (u,l) is computed with the three previous terms $f_{Geo}(l|u)$, $F_{So}(x_{u,l})$ and $F_{Ca}(y_{u,l})$ simply as follows:

$$s(u,l) = f_{Geo}(l|u)\cdot F_{So}(x_{u,l})\cdot F_{Ca}(y_{u,l}) \tag{2.23}$$

2.2.4 Graph Based Approaches

Few works have explored graph-based approaches for POI recommendation. However these solutions are interesting for embedding geographical and temporal influences in a natural way. Usually the main limitation of these models is the limited amount of side information they can include. In this part we present GTAG, which is representative of these methods.

2.2.4.1 GTAG: Geo-Temporal Aware Graph

This approach [Yuan, Cong, and Sun 2014] has proposed a time-aware graph-based approach to integrate geographical and temporal influences. GTAG is inspired by STG model proposed in [Xiang et al. 2010]. In this model there are three types of nodes: *POI node*, *session node* and *user node*. This graph structure allows to integrate these influences naturally. The model GTAG is based on four intuitions: (**i**) *user's interests vary with time, and her temporal interest in a time is reflected by the POI she visited at that time*, (**ii**) *check-ins which are closer in time from the target time are more important*, (**iii**) *when two users have similar temporal patterns, they tend to visit the same POI* and (**iv**) *users tend to visit nearby POI*. Based on these intuitions, GTAG proposes a graph traversal approach that exploits different weights based on *breadth-first preference* propagation strategy. The main idea is to inject an initial preference for the target user node u and then to propagate this value towards all candidate POI. The preference propagated through each path p is the production of the initial preference r_u and the weights of all edges on the path as follows:

$$r_u^{(p)} = \prod_{edge_{i,j} \in p} w_{i,j} \cdot r_u \tag{2.24}$$

2.2.5 Matrix Factorization Models

The goal of *Matrix Factorization* (MF) methods is to find a decomposition of the user-checkin matrix $\mathbf{X}$ with 2 matrices: a matrix $\mathbf{U} \in \mathbb{R}^{M \times K}$ of users' latent features and a matrix $\mathbf{L} \in \mathbb{R}^{N \times K}$ of POI latent features where $K \in \mathbb{N}$ is the number of *latent factors*. Usually K is defined between 30 and 100. The idea is to represent in a "small" hidden space (that is to say K is small in comparison to M and N: $K << M, N$) the user profiles and the POI descripions. Specifically each user i is represented by a row $\mathbf{u}_i$ from $\mathbf{U}$ and each POI j is represented by a row $\mathbf{l}_j$ from $\mathbf{L}$. Then we compute the recommendation score of user i for POI j with the inner product of their corresponding latent vectors: $\tilde{\mathbf{X}}_{ij} = \mathbf{u}_i \mathbf{l}_j^T$. After empirical risk minimization, we obtain the most general objective function to minimize: $\min_{\mathbf{U},\mathbf{L}} \|\mathbf{X} - \mathbf{U}\mathbf{L}^T\|_{\mathbf{F}}^2$. However this function can take into account the *overfitting* problem depending on the category of the problem by adding *regularization* terms as follows:

$$\min_{\mathbf{U},\mathbf{L}} \left\|\mathbf{X} - \mathbf{U}\mathbf{L}^{\mathbf{T}}\right\|_{\mathbf{F}}^2 + \underbrace{\lambda_1 \|\mathbf{U}\|_{\mathbf{F}}^2 + \lambda_2 \|\mathbf{L}\|_{\mathbf{F}}^2}_{\text{Regularization terms.}} \tag{2.25}$$

where λ_1, $\lambda_2 \in \mathbb{R}$ are regularization scalars, and $\|\cdot\|_{\mathbf{F}}$ is the Frobenius norm of a matrix. We propose in the following of this section to introduce the most famous MF-based models used for POI recommendation.

2.2.5.1 LRT: Location Recommendation with Temporal Effects

Location Recommendation with Temporal effects (**LRT**) is a time-aware matrix factorization model proposed recently in [Gao, Tang, Hu, and Liu 2013] to address the temporal influences of users' mobility patterns in LBSN data. Indeed LRT proposes (i) to model *temporal non-uniformness* by defining distinct time slots and (ii) to model *temporal consecutiveness* by introducing temporal regularization terms. The authors observe that users' check-in behavior varies with time, and they propose to define latent features vectors for different time slots through the day. Then all the latent vectors are exploited to compute the final recommendation score. Hence a check-in matrix $\mathbf{X}^{(t)}$ is factorized for each time slot t separately, where $t \in \{0, 1, ..., 23\}$ corresponds to an hour of the day. Then regularization terms are added to the objective function 3.1 as follows:

$$\min_{\mathbf{U},\mathbf{L}} \left\|\mathbf{X} - \mathbf{U}\mathbf{L}^{\mathbf{T}}\right\|_{\mathbf{F}}^2 + \lambda_1 \left\|\mathbf{U}\right\|_{\mathbf{F}}^2 + \lambda_2 \left\|\mathbf{L}\right\|_{\mathbf{F}}^2 + \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{m} \psi_i(t, t-1) \left\|\mathbf{u}_i^{(t)} - \mathbf{u}_i^{(t-1)}\right\|_F^2}_{\text{Temporal regularization terms.}} \tag{2.26}$$

where $\psi_i(t, t-1)$ is the similarity between $\mathbf{X}_i^{(t)}$ and $\mathbf{X}_i^{(t-1)}$. Then the recommendation score is computed as the sum of all time slots: $\tilde{x}_{ij} = \sum_t \mathbf{u}_i^t \mathbf{l}_j^T$

2.2.5.2 GeoMF: Geographical Matrix Factorization

The authors of [Lian et al. 2014] have proposed a geographical matrix factorization approach. GeoMF is based on WRMF [Hu, Koren, and Volinsky 2008] which is a regularized matrix factorization approach that give good quality results on implicit feedback datasets. The idea of GeoMF is to capture the *spatial clustering phenomenon*[4]. To do so, GeoMF first divides the whole geographical space into R grids, each of which representing a *geographical region*. A region is a geographical square of more or less 500 meters. Then GeoMF assumes that each POI propagates its influence to surrounding regions. GeoMF models users' activity regions with *user activity vectors* and region with *influence propagation vectors*. Users' geographical latent vectors constitute matrix $\mathbf{G}$ and region propagation vectors constitute matrix $\mathbb{Y}$. Then the GeoMF model computes the estimated recommendation score as follows:

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{L}^{\mathbf{T}} + \mathbf{G}\mathbf{Y}^{\mathbf{T}} \tag{2.27}$$

[4] defined in Section 2.2.1

In a similar form of equation 3.1, GeoMF optimizes finally the following problem:

$$\min_{\mathbf{U},\mathbf{L},\mathbf{G}} \left\| \mathbf{W} \odot \left(\mathbf{X} - \mathbf{U}\mathbf{L}^{\mathbf{T}} - \underbrace{\mathbf{G}\mathbf{Y}^{\mathbf{T}}}_{\text{Geographical terms}} \right) \right\|_{\mathbf{F}}^{2} + \gamma \left(\|\mathbf{U}\|_{\mathbf{F}}^{2} + \|\mathbf{L}\|_{\mathbf{F}}^{2} \right) + \lambda \|\mathbf{G}\|_{1} \quad (2.28)$$

2.3 Overview of Important Models

In this section we present briefly some recent techniques that belong to the family of models that have inspired our research. We present first a classification of the most recent existing methods. Then we describe the conductive line of the models presented in this thesis.

2.3.1 Existing Methods

We present in this part a comprehensive overview of existing points-of-interest recommendation approaches. Table 2.3 contains a classification of some of the most recent works related to points-of-interest recommendation. Some of these methods have been presented in Section 2.2. We classified them depending on the problem they propose to solve, the category of the model they apply and the contextual influences they exploit. The existing current problems in the POI recommendation state-of-the-art are described in Section 2.2.2. We can observe that most of them are factorization methods and use the geographical influences. Also few of them are time-aware methods.

2.3.2 Models of this Thesis

In this thesis our work focuses on the most general problem of points-of-interest recommendation. The approaches that we propose in the following of this thesis are factorization-based methods that exploit both geographical and temporal influences.

We first investigate a weighted factorization-based model, namely GeoMF-TD, that aims at solving the implicit feedback problem (*c.f.* Chapter 4). We present this model in details in Chapter 3. GeoMF-TD gives good results, but suffer from its important complexity. Because of this complexity we have investigated other factorization-based model that could be trained faster. For this reason we have then proposed a Poisson-based factorization model (a geographical social poisson factorization or GeoSPF) that deals efficiently with large datasets. This model is

		Problem				Model					Influences					
Reference	**Name**	POI Reco.	Time Aware	Next-POI	In/Out Town	Collab. Filt.	Graph-based	Mat. Fact.	Probabilistic	Others	Geographical	Social	Temporal	Sequential	Categorical	Textual
Lian et al. 2014	GeoMF	✓						✓			✓					
Gao, Tang, Hu, and Liu 2013	LRT		✓					✓					✓			
Liu and Xiong 2013	TL-PMF	✓						✓	✓		✓					✓
Cheng, Yang, King, and Lyu 2012	FMFMGM	✓						✓			✓	✓				
Ye, Yin, Lee, and Lee 2011	USG	✓				✓					✓					
Zhang and Chow 2013	iGSLR	✓				✓			✓		✓	✓			✓	
Liu, Fu, Yao, and Xiong 2013	GT-BNMF	✓						✓	✓		✓					
Li, Cong, et al. 2015	Rank-GeoFM	✓						✓	✓		✓		✓	✓		
Feng et al. 2015	PRME-G			✓					✓	✓	✓					
Berjani and Strufe 2011	EWI	✓							✓			✓				
Zhang and Wang 2015	LTSCR		✓	✓			✓		✓		✓	✓				
Ference, Ye, and Lee 2013	UPS-CF				✓	✓								✓	✓	
Zhang and Chow 2015	GeoSoCa	✓				✓			✓		✓	✓			✓	

Table 2.3: Overview of some recent points-of-interest recommendation techniques.

built from an implicit social graph that is used to learn social influences. GeoSPF is presented in Chapter 4. However GeoSPF requires a lot of data preprocessing to build the social graphs which prevent to use it on large geographical zones. This is why we propose to augment this model thanks to geographical clustering methods to handle more realistic datasets.

Chapter 3

An Efficient Matrix Factorization Model for POI Recommendation

In the context of POI recommendation the spatial dimension has a significant impact on the recommendation quality, beyond the personal preferences of the users. A varied range of methods have been proposed to deal with the geographical information. Most of these methods are based on the common assumption that the inter check-ins distances are generally small. This corresponds to the Tobler's first law of geography [Miller 2004] which claims that close things are more related than remote ones. However the real world is not an isotropic space: transport networks (roads, bridges, metro...) and geographical characteristics (rivers, mountains, urbanism...) have to be taken into account as well. In this chapter we present extensively a new geographical matrix factorization approach, namely GeoMF-TD, that aims at integrating these characteristics into the model. This work has resulted in a publication Griesner, Abdessalem, and Naacke 2015 in the ACM Conference on Recommender Systems[1]*.*

After a short introduction in Section 3.1 and a brief related work in Section 3.2, we present the GeoMF model in Section 3.3. Then we introduce our approach in Section 3.4. The experiments that we have conducted on a real-world dataset are presented in Section 3.5. Finally we conclude this chapter in Section 3.6.

3.1 Introduction

The rapid emergence of location-based social networks (or LBSNs such as Foursquare, Flickr, Facebook places and so on) has promoted the advent of new forms of online services, such as recommendation services. Many different recom-

[1] http://recsys.acm.org/

mendation services have been experimented in LBSNs (recommendation of places, of activities, of users, of events...). One of the main goals of these services is to offer to users the possibility to interact with each other, to know better their environment and to explore new sets of points-of-interest (or POIs) by sharing their personal experiences and feelings regarding POIs they have visited in the past. Personalized POI recommendation is the task of making recommendations of the POIs matching the best the user preferences. Today this task has become an essential component of the LBSN activities since it allows not only users to have better user experiences but POI owners to get more targeted customers as well.

By collecting the mobility records of users LBSNs constitute a rich and large-scale check-in data source. These data considered as an abundant implicit feedback of the travel experiences of the user give a significant opportunity to improve POI recommendation performances. The traditional way to realize this task is to use classical collaborative filtering (CF) approaches such as matrix factorization. Matrix factorization approaches have demonstrated to be the most accurate recommendation methods some years ago thanks to the NetFlix Prize Koren, Bell, and Volinsky 2009. Many methods exist to include geographical dimension into matrix factorization models. However these methods assume usually that the environment is an isotropic homogeneous space (*i.e.* without any geographical constraint anywhere) whereas different observation from the data could be explained by natural of physical causes. Here our idea is to integrate naturally these causes into the factorization model.

Some years ago the authors Hu, Koren, and Volinsky 2008 have demonstrated that weighted matrix factorization was the most adapted method to CF problems with implicit feedback. This method has been exploited and augmented by Lian et al. 2014 to include the geographical influence of POI by the modeling of the spatial clustering phenomenon Ye, Yin, Lee, and Lee 2011; Zhang and Chow 2013 directly into the factorization process. However LBSN data comes with much more than only geographical information. Notably we have also access to the recorded timestamp of each check-in. In the approach that we present here the main idea is to distinguish among unvisited POIs the "negative" ones from the "unknown" ones. To this aim we take into account of the geographical dimension of the problem considering that an unvisited POI geographically close from a frequently visited POI is more likely to be "negative" than a POI geographically distant from visited POI.

The approach presented in this chapter aims at integrating time dependen-

cies into geographical matrix factorization. Specifically we investigate the idea of augmenting matrix factorization model with both geographical and temporal influences. This leads to the GeoMF-TD algorithm we present in the following. GeoMF-TD has been published in Griesner, Abdessalem, and Naacke 2015.

3.2 Related Matrix Factorization Models

Matrix factorization models belong to a successful class of methods that many previous works have used Cheng, Yang, King, and Lyu 2012; Feng et al. 2015. A lot of different factorization methods have been proposed to solve POI recommendation. However most of these approaches try only to adapt traditional recommendation algorithms to the specific problem of POI recommendation. Some years ago Zheng, Zheng, Xie, and Yang 2010 have proposed the Collaborative Location Activity Filtering (CLAF) algorithm for generic recommendation. CLAF is a collective matrix factorization close to the method presented by Singh and Gordon 2008 which is based on the exploitation of the inferred correlations existing between the features of the locations and the POIs. Differently Regularized Matrix Factorization presented by Berjani and Strufe 2011 apply CF personalized methods on dimensionally reduced user-POI matrices aiming at minimizing squared regularized errors. The authors Sattari et al. 2012 have proposed Improve Feature Combination (IFC), which is based on an extended matrix factorization model that integrates additional data resources before applying the standard singular value decomposition technique to the extended model. It has been proven in several studies that IFC performs better than CLAF in terms of prediction accuracy.

Since each POI comes with a significant geographical dimension, many works have tried to integrate this geographical information into the recommendation model. Some years ago Ye, Yin, Lee, and Lee 2011 have proposed a new technique to integrate geographical influence with classical CF approaches. More precisely the authors have studied the geographical influence of POI assuming a power-law distribution of the visited POIs. On another hand Cheng, Yang, King, and Lyu 2012 have proposed a multi-center gaussian model that exploits the natural spatial clustering phenomenon. Differently Zhang and Chow 2013 have proposed a personalized fusion framework based on kernel density estimation of the distances distribution between POIs of each user.

In addition to the geographical dimension, the temporal dimension is another important factor leveraging the accuracy of the model. Exploring temporal dimen-

sion into matrix factorization is not a new idea. Some years ago Gao, Tang, Hu, and Liu 2013 have proposed a location recommendation framework with temporal effects (LRT). Specifically they showed how to model two main temporal properties of data (i.e. non uniformness and consecutiveness) with matrix factorization. The experiments conducted showed that LRT outperforms traditional recommendation algorithms, but with a high complexity cost.

3.3 Geographical Influence for Factorization Models

In this part we present the GeoMF-TD base model. Here let $\mathbf{u} = \{u_1, u_2, ..., u_m\} \subset U^m$ be the subset of users and $\mathbf{p} = \{p_1, p_2, ..., p_n\} \subset P^n$ the subset of POIs. Then let $\mathbf{C} \in \mathbb{R}^{m \times n}$ be the user-POI matrix containing the m users and the n POIs. The value $c_{u,j}$ in $\mathbf{C}$ refers to the visit frequency of user u to the POI i.

3.3.1 Weighted Matrix Factorization

Basically the goal of matrix factorization is to approximate matrix $\mathbf{C}$ by the product of two matrices $\mathbf{P} \in \mathbb{R}^{m \times k}$, and $\mathbf{Q} \in \mathbb{R}^{n \times k}$ of latent factors with dimension $k \ll \min(m, n)$ by solving the following classical optimization problem:

$$\min_{\mathbf{P},\mathbf{Q}} \left\| \mathbf{C} - \mathbf{P}\mathbf{Q}^T \right\|^2 + \gamma \left(\|\mathbf{P}\|^2 + \|\mathbf{Q}\|^2 \right) \tag{3.1}$$

with γ a non-negative parameter to avoid overfitting by controlling the capability of $\mathbf{P}$ and $\mathbf{Q}$. Then it becomes possible to approximate the missing value $\widetilde{c_{u,j}}$ in $\mathbf{C}$ by computing the inner product between corresponding latent factors $\widetilde{c_{u,j}} = \mathbf{P}_u \mathbf{Q}_j^T$.

However, the application domain of LBSNs is different from traditional recommendation domains. Indeed the check-in datasets in LBSNs provide only indication of *confidence* but no information about *preferences* of users. This property refers to the recommendation problems with *implicit feedback*. Specifically Hu, Koren, and Volinsky 2008 have proven that weighted matrix factorization (WMF) gives the best results with implicit feedback datasets. Weighted matrix factorization takes into account of the asymmetry existing between *confidence* and *preference* and creates two new variables for formalizing this asymmetry. Then WMF turns the problem of Eq(3.1) into the following new optimization problem,

$$\min_{\mathbf{P},\mathbf{Q}} \left\| \mathbf{W} \odot \left(\mathbf{R} - \mathbf{P}\mathbf{Q}^T \right) \right\|^2 + \gamma \left(\|\mathbf{P}\|^2 + \|\mathbf{Q}\|^2 \right) \tag{3.2}$$

where $\odot$ is the element-wise matrices multiplication (i.e. the Hadamard product) and where the only differences with Eq (3.1) is the presence of the matrix $\mathbf{W}$, and the binary 0/1 matrix $\mathbf{R}$ whose each entry $r_{u,i}$ indicates if user u has visited POI i. The idea of WMF is to assume a minimum confidence for all POI, visited or not. This minimum confidence is encoded within the $\mathbf{W}$ matrix, setted as follows:

$$w_{u,i} = \begin{cases} 1 + \alpha(c_{u,i}) & \text{if } c_{u,i} > 0 \\ 1 & \text{otherwise} \end{cases} \tag{3.3}$$

where $\alpha()$ is a monotonically increasing function.

3.3.2 Modelling Geographical Influence

The modeling of geographical influence for POI recommendation in LBSNs has been widely studied in previous works such as Ye, Yin, Lee, and Lee 2011; Cheng, Yang, King, and Lyu 2012; Zhang and Chow 2013 or by Liu and Xiong 2013. Recently Lian et al. 2014 have proposed a geographical matrix factorization (GeoMF) to integrate this influence directly into the factorization model of WMF. The idea of the authors was to distinguish for each user the unvisited but interesting POIs among the negative ones. The intuition is that if a user visits a POI without visiting the other closely located POIs then these "ignored" POIs may not be interesting enough for the user. Consequently these POIs become *negative* for the factorization model.

This approach divides the space into L even grids $\mathbb{L} = \{g_1, g_2, ..., g_L\}$ and computes for each POI its influence area onto each one of these L grids based on the normal distribution of distances. Specifically they augmented the traditional matrix of latent factors $\mathbf{P}$ and $\mathbf{Q}$ with two matrices of latent geographical factors $\mathbf{X} \in \mathbb{R}^{m \times L}$ and $\mathbf{Y} \in \mathbb{R}^{n \times L}$. With these new latent factors, the equation 3.2 is modified as follows:

$$\min_{\mathbf{P},\mathbf{Q},\mathbf{X}} \left\| \mathbf{W} \odot \left(\mathbf{R} - \mathbf{P}\mathbf{Q}^T - \mathbf{X}\mathbf{Y}^T \right) \right\|^2 + \gamma \left(\|\mathbf{P}\|^2 + \|\mathbf{Q}\|^2 \right) + \lambda \|\mathbf{X}\|^2 \tag{3.4}$$

where λ controls the sparsity constraint over the mobility behavior of each user through the L grids. A row $\mathbf{x}_u$ of $\mathbf{X}$ refers to the activities areas of user u i.e. the distribution of his visit frequencies in each grid g_l of the map, while a row $\mathbf{y}_i$ of $\mathbf{Y}$ refers to the influence area of POI i. More precisely we compute for each POI i and for each grid g_l the gaussian geographical influence i has on g_l:

$$\mathbf{y}_i^l = \frac{1}{\sigma} K\left(\frac{d(i,l)}{\sigma}\right) \tag{3.5}$$

where $K()$ is the standard normal distribution and σ the standard deviation. With this augmented geographical model we get the recommendation ranking score for user u and POI i as follows:

$$\widetilde{c_{u,i}} = \mathbf{P}_u \mathbf{Q}_i^T + \mathbf{X}_u \mathbf{Y}_i^T \tag{3.6}$$

One of the most significant advantage of this approach is that it encompasses both preferences of user from latent factors and preferences from geographical factors. We can observe that the augmented latent features vectors have a universal form and can be extended easily with any other external data.

3.4 GeoMF with Temporal Dependencies: GeoMF-TD

The GeoMF model assumes that the space is an isotropic homogeneous space (*i.e.* without any geographical constraint anywhere) . Especially this model assumes that the influence area of each POI follows a normal distribution fixed in advance and only based on distances over space. However the influence areas of two distinct POIs can be very different in reality by considering different parameters other than the distances. Notably the temporal effects in POIs visit sequences play also a significant role as Gao, Tang, Hu, and Liu 2013 have demonstrated. Particularly these effects can reflect that a POI j can be in the influence area of another POI i but not being really negative.

Following the GeoMF approach, our basic idea is to integrate these temporal influences into the GeoMF model. Actually we propose to modify the values of the influence area of each POI i through the grid $g_{l \in \mathbb{N}^L}$ to take into account the time spent by a user to go from the POI i to the other POIs collocated in g_l. More precisely for each POI i, we compute the average time that each user spend to reach j (j is in g_l) from i. We compute this for every user that has at least one check-in at i and another (more recent) check-in at j into g_l. Then we average the per-user values to get a single value related to POI i. Let $t_i^{g_l}$ be the average time computed between i and collocated POIs existing in g_l. We introduce temporal coefficients $\theta_l(t_i^{g_l})$ as follows:

$$\theta_l(t_i^{g_l}) = \begin{cases} \alpha * \mathbf{y}_i^l & \text{if } t_i^{g_l} > \sigma^i \text{ and } \mathbf{y}_i^l < 0.1 \\ \mathbf{y}_i^l & \text{otherwise} \end{cases} \tag{3.7}$$

Number of users	196,591
Number of check-in	6,442,890
Number of social links	950,327
Matrix density	2.9×10^{-5}
Average No. of visited POIs per user	37.18
Average No. of check-ins per POI	3.11

Table 3.1: Statistics of the Gowalla data set

where σ^i refers to the standard variation of time intervals for POI i and $\mathbf{y}_i^l$ has been computed from Eq(3.5). Then we fuse these coefficients with influence vector $\mathbf{y}_i$ for POI i:

$$\mathbf{y}_i = [\theta_1(t_i^{g_1}), ..., \theta_L(t_i^{g_L})] \tag{3.8}$$

The idea of these temporal coefficients is to decrease the *negativeness* of potential negative POIs when no user has checked-in them during a certain time. That is why these coefficients let unchanged the influence area value when this value is low. We use these coefficients as a fusion output between geographical gaussian influence over space, and temporal dependencies existing into the dataset.

3.5 Experiments

In our experiments, we compared the accuracy of our approach with GeoMF. This section describes the dataset we used, the evaluation metrics we chose, and the results we obtained.

3.5.1 Dataset and Experimental Setup

We evaluated the algorithms on check-ins crawled from Gowalla[2] and publicly available thanks to Cho, Myers, and Leskovec 2011. Gowalla was a famous LBSN closed in 2012. Gowalla dataset has already been used in several works on POI recommendation such as the works proposed by Cheng, Yang, King, and Lyu 2012; Cho, Myers, and Leskovec 2011; Zhang and Chow 2013. Table 3.1 presents the main statistics concerning this dataset.

[2]The dataset can be downloaded here: http://snap.stanford.edu/data/loc-gowalla.html

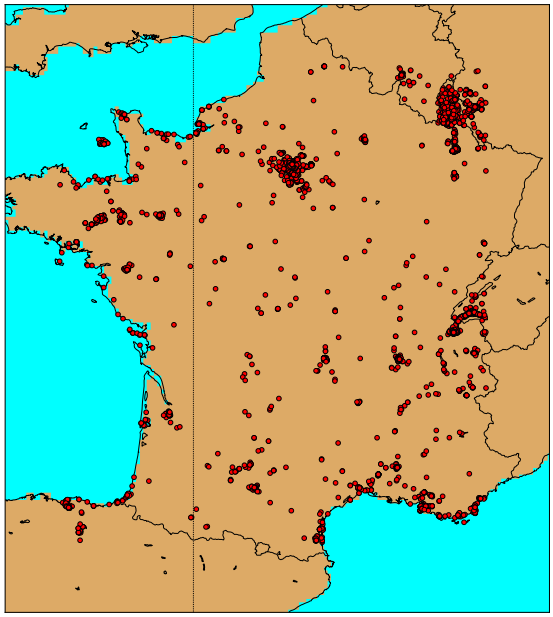


Figure 3.1: Check-in distribution from Gowalla users during 21 months of the most visited POIs in France.

In order to reduce matrix sparsity in the dataset we keep only users with at least 50 check-ins and for practical purposes we use only check-ins localized in France. Figure 3.1 presents the spatial distribution of check-ins over the France area. Finally it remains 161 users, 7697 distincts POIs for 12418 distinct check-ins, which is very few but enough for an initial evaluation. Then we organize this dataset as a user-POI matrix.

3.5.2 Evaluation Metrics

It is traditional for each user $u_i \in U$ to mark off between 20% and 40% of all POIs he has checked-in in the past for testing, while the rest remains for training the model. Basically a recommendation algorithm estimates a ranking score for each candidate POI $i_{cand} \in P$ and returns the top-k highest ranked POIs $p_1, p_2, ..., p_k \in P^k$ as recommendation results for the targeted user. Then we evaluate the recommendation accuracy by finding out how many recommended POIs are effectively present into the test set of this targeted user. More precisely we compute $precision@N$ and

$recall@N$. The former refers to the ratio of recovered POIs to the N recommended POIs, while the latter refers to the ratio of recovered POIs to the set of previously visited POIs as follows:

$$precision@N = \frac{\sum_{u_i \in U} |TopN(u_i) \cap L(u_i)|}{\sum_{u_i \in U} |TopN(u_i)|} \tag{3.9}$$

$$recall@N = \frac{\sum_{u_i \in U} |TopN(u_i) \cap L(u_i)|}{\sum_{u_i \in U} |L(u_i)|} \tag{3.10}$$

where $TopN(u_i)$ represents the set of top-N POIs recommended to user u_i and $L(u_i)$ represents the set of POIs from the test set checked-ins by u_i. We have evaluated $precision@N$ and $recall@N$ with N ranging from 1 to 20 for precision, and from 1 to 100 for recall. We provide the results we obtained on the average after cross-validation with 5 folds in the next section.

3.5.3 Results and Discussions

For comparison purpose, we implemented the GeoMF approach using the LibRec Java library[3]. Figure 3.2 and Figure 3.3 depict a comparative analysis of respectively the $precision@N$ and the $recall@N$ results of GeoMF and our approach (GeoMF-TD) with N ranging from 1 to 20 for the precision, and with N ranging from 1 to 100 for the recall. As expected the temporal coefficients we introduce allowed to take into account the temporal dependencies existing between POIs and thus improve the global accuracy. Figures 3.2 and 3.3 show an average benefit of 60% for recall and 20% for precision. This overall performance comparison does not integrate the study of the influence of the threshold parameter, but gives promising results for the future.

3.6 Conclusions

In this chapter we have proposed a new matrix factorization model for the problem of POI recommendation in LBSNs. Specifically we have investigated matrix factorization algorithms based on geographical influence. Our goal was to try to leverage the factorization model of GeoMF by considering the temporal influences of POIs checked-ins. To this end we have provided GeoMF-TD algorithm as an efficient

[3]LibRec can be downloaded here: http://www.librec.net/

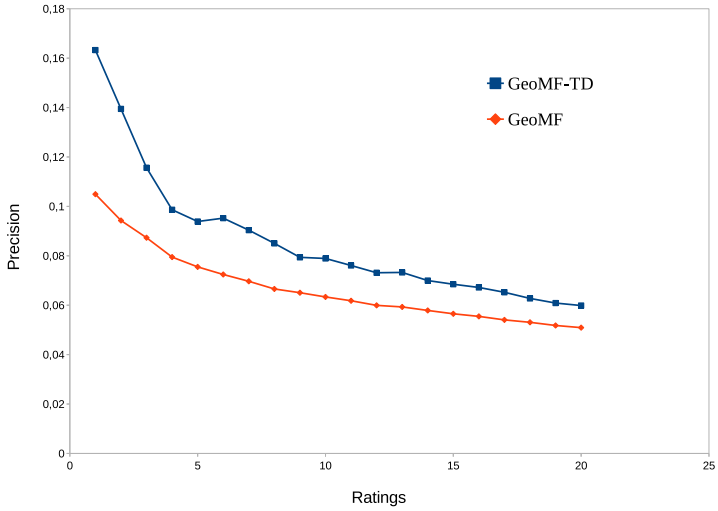


Figure 3.2: Precision comparison between GeoMF and GeoMF-TD

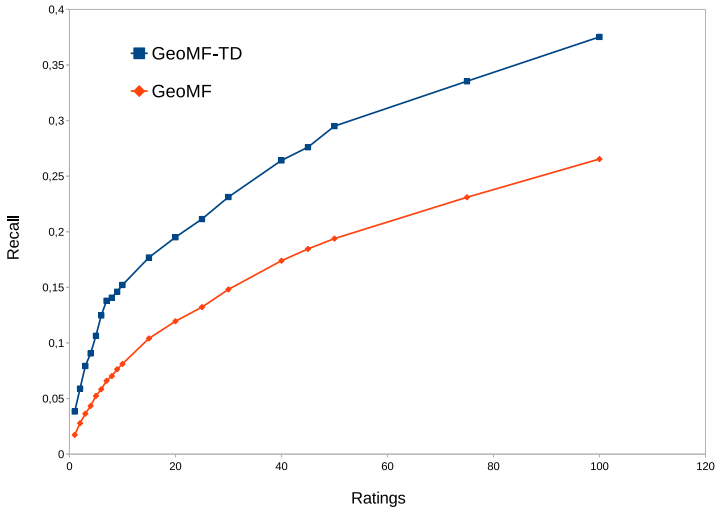


Figure 3.3: Recall comparison between GeoMF and GeoMF-TD

proposal of an extension of GeoMF and we have presented accuracy comparisons. Our experimental evaluation shows that GeoMF-TD presents better accuracy performances than GeoMF.

Considering that the preferences of the user will change over time, a future line of work we should investigate will be to take into account of the online integration of user's preferences changes and to capture this evolution into our model. This problem refers to the recommendation *dynamicity* challenge widely studied in recent studies Gueye, Abdessalem, and Naacke 2013; Gueye, Abdessalem, and Naacke 2015. Additionally, one of our future goal will be to include POI categories into the model, and to cope with the scalability issues.

Chapter 4

A Factorization Based Solution to the Implicit Feedback Problem

One of the main challenges that POI recommendation has to tackle is the implicit feedback problem. This problem corresponds to the difficulty to distinguish among the unvisited POI the "negative" ones from the unknown ones. In this chapter we investigate specifically this problem and we present with details a new geographical matrix factorization approach with implicit social influences, namely GeoSPF. Our central idea is to infer an implicit social network based on geographical patterns observed in the data. Then we use this network to compute a social influence bias that is integrated into a Poisson factorization framework. This work has resulted in a publication Griesner, Abdesssalem, and Naacke 2017 in EGC international Conference[1].

We present a short introduction in Section 4.1 and some related works on implicit feedback problem in Section 4.2. Then we introduce some specific notations in Section 4.3. We describe our process to infer an implicit social network and we present our GeoSPF model in Section 4.4. Finally we propose our experimental results and our conclusion in Sections 4.5 and 4.6 respectively.

4.1 Introduction

The large number of Location-Based Social Networks (LBSNs) such as Foursquare, Flickr, Twitter etc. which have emerged these last years have changed deeply our vision of our environment and how we interact with it. For instance on the LBSN Flickr[2] there are more than 110 millions users who produce more than one million

[1] http://egc2017.imag.fr/
[2] www.flickr.com

images per day. Such large volumes of data provide a rich and precise information on the preferences and interests of the users, making possible new kind of online services, such as POIs recommendation.

Personalized POIs recommendation is the task of proposing to a user a list of relevant POIs the user could be interested to visit. This task has become an essential component of LBSNs, allowing the users to discover new POIs and reciprocally the POIs to increase their attractiveness. Through these networks millions of users can share their experiences and their comments concerning the locations, also known as *points-of-interest* (POIs), e.g., restaurants, museums, buildings etc. that they have visited in the past. These visits are also known as *check-in* activities that correspond to users' preferences on POIs.

Dealing with LBSNs check-ins involves to take into account of several challenging characteristics of POIs recommendation:

- A high level of **sparsity** which means that the density of the user-POI check-in matrix is very low in LBSNs in comparison to other applications such as the appoaches proposed by Zhang and Chow 2013; Cheng, Yang, King, and Lyu 2012; Gao, Tang, Hu, and Liu 2013; Lian et al. 2014. Sparsity puts severe limitations on the accuracy of most of the recommendation approaches, since it makes the extraction of users' preferences very difficult.

- **Frequency data**: that is to say we know only how many times a user has been located to a place. Most of existing works use Gaussian assumption to model the geographical users' mobility as the works from Liu, Fu, Yao, and Xiong 2013; Ye, Yin, Lee, and Lee 2011; Cheng, Yang, King, and Lyu 2012. However Poisson models proposed by Charlin, Ranganath, McInerney, and Blei 2015 for recommender systems are much more efficient.

- **Contextual information**: in existing works for POIs recommendation, social influence is the most exploited contextual information Zhang and Wang 2015; Cheng, Yang, King, and Lyu 2012. However adding these information into the model requires to divide the user-POI check-in matrix into a tensor, which increases even more the sparsity. As a consequence these approaches will fail to deal with large-scale datasets.

- **Implicit feedback**: indeed check-in data only provide positive samples: we know which POIs have been checked in but we cannot know if the user's experiences have been positive or negative. In this case, there is no straightforward way to distinguish between unattractive POIs for the user and those

undiscovered by the user but potentially attractive for her.

The implicit feedback problem has the most significant impact on the recommendation accuracy given that historically most of recommendation data models exploit explicit ratings from users. In the case of POI recommendation, we cannot infer the reason why user has been located at a given place: it could be for professional reasons, or social reasons, and not because he decided it. Because the explicit feedback of the user is not available we have to find other methods to compensate this lack of information.

Our main objective remains to recommend a list of POIs to a given user based on her past check-ins and other available external side information. These side information are here the locations associated to each POI. We use these locations to build an implicit geographical accessibility graph (AGRA). We define below what this graph is exactly, and then we define our problem.

Definition 4.1.1. *(**AGRA**) An **Accessibility Graph**, AGRA for short, denoted* $\mathbf{G} = (V, E, \rho)$ *is a directed graph where each node* $v \in V$ *represents a POI associated to its geographical location, each edge* $e = (p_i, p_j) \in E$ *exists if the transition* $p_i \rightarrow p_j$ *exists, and* ρ *is a function that associates to each edge* $e = (p_i, p_j)$ *the corresponding accessibility measure* $\mathcal{A}_{i,j}$ *(as defined in equation A.14). An edge exists only if a transition from* p_i *to* p_j *is observed in at least one user's itinerary. This is an efficient structure to explore users itineraries.*

Problem 4.1.1. ***Implicit Feedback:*** *The problem is to distinguish among unvisited POI the negative ones from the unknown ones. This is an instance of the positive unlabeled (i.e. PU) classification problem. This problem has been investigated some years ago by Elkan and Noto 2008.*

This problem is also called the one-class collaborative problem, or simply the one class problem. It is frequent in many applications and real-world datasets where the users' feedback has not been tracked. A way to deal with it is usually to use cost sensitive optimization functions. In the following of this section we propose to tackle this problem by extracting an implicit social network from the users' behavioral patterns.

4.2 Existing Implicit Feedback Approaches

POIs recommendation corresponds to a wide category of problems that counts many related sub-problems, such as *next-POI recommendation*, *in/out town POI*

recommendation, *time-aware POI recommendation* etc. In this section we present some related works regarding the most general POIs recommendation problem with a specific focus on the implicit feedback issue.

A promising way to solve the implicit feedback problem in POIs recommendation problem was to exploit both geographical influence and social influence into a collaborative filtering (CF) framework. Many memory-based approaches have been proposed such as Zheng, Zheng, Xie, and Yang 2010; Ye, Yin, Lee, and Lee 2011 in order to exploit weights directly computed from source data and integrate contextual information to their model. Some years ago Ye, Yin, Lee, and Lee 2011 have proposed a memory-based CF method that integrates both the social and geographical influence by linear interpolation. In particular they put in evidence that the geographical factor played an essential role into the recommendation quality. The main limitation of their approach is its important complexity which makes it practically unusable to deal with sparse large-scale datasets.

Model-based CF methods have been widely used as well Lian et al. 2014; Zhang and Wang 2015; Hu, Koren, and Volinsky 2008. Model-based approaches are usually much scalable than memory-based methods given that they separate the training phase (offline) from the recommendation phase (online): this allows a fast computation of the recommendation score. The idea of these approaches is to build a predictive model of the data, based on statistical geographical assumption. These methods work efficiently on datasets where the users' feedback is explicit, but the results are globally disappointing when the users' feedback is implicit.

Others classes of methods have been proposed to solve the implicit feedback problem such as factorization based models such as proposed by Cheng, Yang, King, and Lyu 2012; Zhang and Wang 2015. As described in the previous chapter, factorization models aim at catching the user-item interaction assuming that both can be expressed as a vector of latent attributes. In a probabilistic framework, Probabilistic Matrix Factorization as Salakhutdinov and Mnih 2007 proposed PMF which is a successful approach which aims at minimizing a sum-of-squared-errors objective function with quadratic regularization terms. Some years ago Zhang and Wang 2015 have proposed a model which includes both the geographical influence and the temporal influence. However this work differs a bit from ours, since it belongs to the next-POI recommendation problem Feng et al. 2015. Another recent work Gao, Tang, Hu, and Liu 2013 proposed to use a temporal regularization between sessions, in order to put constraints on latent factors to minimize too high contrasts from a session to another. In the same way the authors of Koren, Bell,

and Volinsky 2009 have proposed to used temporal bias to deal with the temporal influence. However this should not be appropriate for frequency data and this creates problems in the recommendations computed.

To deal with implicit behavioral data records, a model has been proposed some years ago by Hu, Koren, and Volinsky 2008. Their model introduces a distinction between the user's preference estimation, and the confidence we can have into this estimation. The authors have demonstrated that their approach was more efficient on small datasets than alternative existing methods. This method has been exploited and augmented by the authors of Lian et al. 2014 to include the geographical influence of POI by modeling the spatial clustering phenomenon Ye, Yin, Lee, and Lee 2011; Zhang and Chow 2013 directly into the factorization process. However the complexity is far too high to use it on real-world datasets.

Poisson Factorization has emerged recently Charlin, Ranganath, McInerney, and Blei 2015 as a successful alternative solution. It is a scalable probabilistic factorization model that outperforms state-of-the-art models subject to sparsity and diversity constraints. Many recent works have proposed to increase the recommendation quality with social influence such as Cho, Myers, and Leskovec 2011; Zhang and Chow 2013; Zhang and Wang 2015; Cheng, Yang, King, and Lyu 2012. The idea of these methods is to exploit the knowledge a user's friends have on unvisited POIs. Some years ago Zhang and Wang 2015 have proposed a model called LTSCR which uses the social similarities of users and integrate them into the factorization model. Unfortunately the social networks are usually not associated in LBSNs real-world datasets. The extraction of an implicit social network has been widely investigated last years by Losup et al. 2014; Hu, Koren, and Volinsky 2008 but remains still unexploited for POIs recommendation.

4.3 A Factorization Model for Implicit Feedback

Let's consider the $m \times n$ user-POI frequency matrix $\mathbf{X}$ representing m users' visit frequencies for n POIs. The goal of traditional matrix factorization approaches is to approximate the matrix $\mathbf{X}$ by the inner product of k-rank factors such that: $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{m \times k}$ and $\mathbf{V} \in \mathbb{R}^{n \times k}$ with $k \ll \min(m, n)$. Since users visit only a few number of POIs, the matrix $\mathbf{X}$ is usually very sparse. Poisson factorization, denoted PF in the following, is a generative probabilistic latent factors based approach that exploits a Poisson law assumption to model the observations. PF is based on the GaP topic model from Canny 2004. Gopalan *et al.* have demonstrated that this model is adapted to behavioral and sparse data as proposed by Charlin,

Ranganath, McInerney, and Blei 2015.

Moreover, the posterior inference of PF is much faster than other approaches because the likelihood of the data depends only of the observed values. If we denote by $x_{i,j}$ the number of times user i has visited POI j, the PF model assumes that $x_{i,j}$ comes from a Poisson distribution, parameterized by the inner product of the user's preferences and the POI attributes. Thus PF estimates for each user i and each item j the following sample rating:

$$y_{i,j} \sim \text{Poisson}(\mathbf{u}_i^T \cdot \mathbf{v}_j) \tag{4.1}$$

Once the posterior distribution has been fitted, PF ranks each user's target items by their recommendation score, that is to say the expected posterior, as follows:

$$\hat{r}_{i,j} = \text{E}\left[\mathbf{u}_i^T \cdot \mathbf{v}_j | y\right] \tag{4.2}$$

where $\mathbf{u}_i$ et $\mathbf{v}_j$ are the k-vectors the user's latent preferences, and the POI latent attributes, respectively. User's latent preferences, and POI latent attributes are considered as hidden variables. Furthermore $\mathbf{u}_i$ and $\mathbf{v}_j$ have assigned empirical Gamma priors.

4.4 GeoSPF: Modeling Geographical and Social Influences

This section presents GeoSPF, that is to say our method to extract implicit social influences from users' mobility behaviors. We introduce our accessibility aware graph (AGRA) that we use to model the geographical influences. This graph is used to build the implicit social network that we exploit then for POIs recommendation. Then we describe how we fuse these influences with a Poisson matrix factorization model. A list of notations used in the following is proposed in Table 1.2.

4.4.1 General Idea

We argue that similar POIs, similar cities and similar regions are expected to share similar characteristics and similar user preferences. These characteristics can explain in part why users have visited such place. As a consequence similar places are expected to share similar latent features. GeoSPF is based on the assumption that it exists a combination of personal preferences, and geographical and social influences, behind the decision process of the user. We integrate these elements into our model following the intuition that a user will prefer, among the POIs that

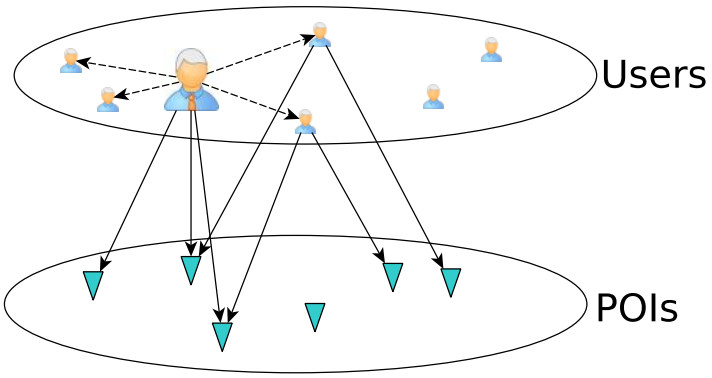


Figure 4.1: An illustration of a user's social network and the check-ins of her friends. GeoSPF is based on the central idea that the target user should benefit from the visit experiences of her friends. Her social network is extracted through the geographical mobility patterns observed in the data. Then our approach integrates her friends' existing check-ins into a factorization model.

match her preferences, the most accessible ones. We model this accessibility as a transition probability from one POI to another. If we denote by $\alpha(u,p)$ the degree of interest of a user u has for a POI p, $\mathcal{S}(u,p)$ the social influence user u got on p, and $\mathcal{G}(u,p)$ the geographical preference of user u regarding p, the probability to observe the pair (u,p) in the dataset should be directly proportional to the interest of u for p, and decreases monotonically with the accessibility:

$$\mathrm{P}(u,p) \propto \mathbb{F}[\alpha(u,p), \mathcal{G}(u,p), \mathcal{S}(u,p)] \tag{4.3}$$

where $\mathbb{F}[\cdot]$ is a function which combines the personal interests, the social influence and the geographical influence. Existing approaches such as the ones proposed by Lian et al. 2014; Yuan, Cong, Ma, et al. 2013; Griesner, Abdessalem, and Naacke 2015 have verified that geographical influence has a significant impact on the recommendation quality. They usually deal with an uniform isotropic space and use only distances between check-ins. We plot in Figure 4.2 the normalized distribution of the inter check-ins distances in two real-worls datasets. The observations seem to confirm the Tobler's law: the willingness to check-in a place decreases with the distance.

However such approaches do not take into account the constraints (ex. road and transportation networks, natural obstacles, country borders etc.) that could make the mobility between two POIs difficult, even if they are close to each other. We can observe a high number of transitions between two POIs that are far from

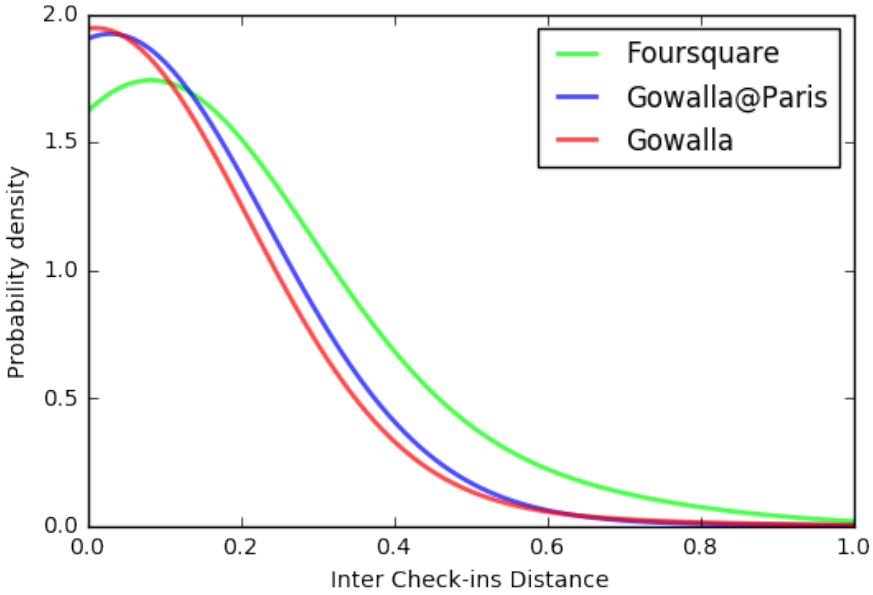


Figure 4.2: Density of inter check-ins distances distribution on 4 datasets.

each other and, conversely, we can observe a very low number of transitions between two popular POIs that are close to each other. Figure 4.2 shows that some transitions observed in real-world datasets have been made between far POIs.

Beyond the distances, we introduce the concept of accessibility, as we will see in Section 4.4.2, to better capture the geographical influence in the users' choices. Figure 4.3 plots the density distribution of accessibility in two datasets that we have gathered. We observe that a similar pattern emerges clearly on the four distributions. More precisely this pattern allows us to select a subset This pattern appears to be very biased and to have a long tail representing the subset of accessible paths between pairs of POIs. We aim to favor these paths in our model.

To sum up, the **main steps** of GeoSPF are as follows:

1. We build an accessibility-aware graph (AGRA) based on the observed transitions (from a POI to another) and their probabilities.

2. We infer an implicit social network (ISN) from AGRA and the similarities between the check-in history of the users.

3. We integrate the ISN into a social Poisson factorization recommendation model to obtain GeoSPF model.

4.4.2 Geographical Accessibility

The idea of accessibility is to model the probability that a user will move to a POI p_{j+1} after visiting POI p_j. To do so, we apply first-order Markov models, which were used successfully for handling sequential data, to our context.

A transition is observed in the itinerary of a user u if it exists in the dataset two successive check-ins $< u, p_i, t_1 >$ and $< u, p_j, t2 >$, done in two different POIs p_i and p_j at two timestamps t_1 and t_2, such that $t_1 < t_2$ and no other intermediary check-in $< u, p_k, t' >$ $(t_1 < t' < t_2)$ exits in the dataset. We will note this transition as follows: $p_i \rightarrow p_j$ in the rest of the chapter. Thus, for a given user, the probability to visit p_{j+1} will be inferred from the last visited one. Formally, we have:

$$P(p_{j+1}|p_j, p_{j-1}, ..., p_1) = P(p_{j+1}|p_j) \tag{4.4}$$

where we define $P(p_{j+1}|p_j)$ as the transition probability $\mathcal{T}_{j,j+1}$ from p_j to p_{j+1}. We can compute this probability using the empirical maximum likelihood estimation as follows :

$$\mathcal{T}_{j,j+1} = P(p_{j+1}|p_j) = \frac{N(p_j, p_{j+1})}{N(p_j)} \tag{4.5}$$

where $N(p_j, p_{j+1})$ is the number of users having the sequence $p_j \rightarrow p_{j+1}$ in their past check-ins, and $N(p_j)$ is the number of users having visited p_j. Since we have $N(p_j, p_{j+1}) \leqslant N(p_j)$ we know that $\mathcal{T}_{j,j+1}$ is naturally bounded: $\mathcal{T}_{j,j+1} \in [0, 1]$. Note that to compute this probability, the check-ins have to follow the temporal order of their occurrence. Then, we can combine this probability with the geographical information in order to estimate the accessibility $\mathcal{A}_{j,j+1}$ between POIs p_j and p_{j+1}. We define this accessibility as follows:

$$\mathcal{A}_{j,j+1} = \frac{1}{0.5 + d(p_j, p_{j+1})} \cdot \mathcal{T}_{j,j+1} \tag{4.6}$$

where $\mathcal{T}_{j,j+1}$ refers to equation 4.5 and $d(p_j, p_{j+1})$ is the euclidean distance between POIs p_j and p_{j+1}. If p_{j+1} is far form p_j, the accessibility will tend to be low. But, if a lot of transitions have been observed from p_j to p_{j+1}, then the accessibility will increase accordingly. Equation A.14 is inspired from the geographical weights used by Liu and Xiong 2013. Figure 4.3 compare the distribution density of the inter check-ins on four real-world datasets. The obtained curves share a common pattern, which highlights a uniformity in the behavior of the users while considering the accessibility.

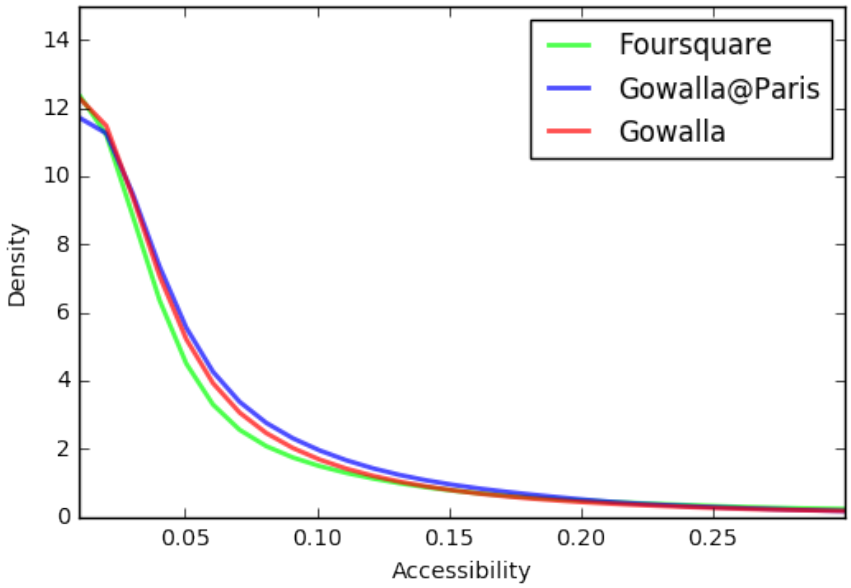


Figure 4.3: Density of inter check-ins accessibilities distribution on 4 datasets.

4.4.3 AGRA: Accessibility Graph

Previous works such as Cheng, Yang, King, and Lyu 2012; Zhang and Chow 2013; Ma, Zhou, Lyu, and King 2011; A., J., and Tauro 2014 have shown that social networks data can play a significant role in POI recommendation quality. They exploit the fact that the connections close to a given user (friends, colleagues ...) have an influence on her choices. However, in LBSNs data we do not have an explicit social network: we only have the history of the check-ins without explicit social links between the users.

Thus, our approach builds an Implicit Social Network (ISN) based on the similarity between the check-in history of the users and their transitions in the AGRA graph. We propose the four possible similarity measures below, chosen for their scalability and the quality of their results, as we will show later in the experiments.

- **Adamic/Adar**: This measure gives a high importance to the rare transitions (i.e., with a low accessibility). Intuitively, the more two users share POIs involved in rare transitions, the more they are supposed to be close to each other. Formally, the Adamic/Adar similarity $S_{\mathrm{AA}}(u_1, u_2)$ is defined as follows:

$$S_{\mathrm{AA}}(u_1, u_2) = \sum_{v \in \mathcal{P}^{u_1} \cap \mathcal{P}^{u_2}} \frac{1}{\log(D(v))} \tag{4.7}$$

Where $D(\cdot)$ is a function returning the degree of a node. From equation 4.7 we see that if u_1 and u_2 have in common POIs involved in a small number of transitions, they tend to be similar.

- **Standard Jaccard on POIs**: This is the standard Jaccard measure. Given the sets of the POIs visited by two users u_1 and u_2, we define their Jaccard similarity $S_J(u_1, u_2)$ as follows:

$$S_J(u_1, u_2) = |\mathcal{P}^{u_1} \cap \mathcal{P}^{u_2}| \,/\, |\mathcal{P}^{u_1} \cup \mathcal{P}^{u_2}| \tag{4.8}$$

- **Accessibility Weighted Symmetric Jaccard:** With this measure, we extend the standard Jaccard measure by considering the accessibility between the visited POIs. Intuitively, the more the POIs visited by two users are accessible to each other, the more these two users are similar. To do so, we add to the set of visited POIs those ($\Gamma(\mathcal{P}^u)$) that are accessible in one hop throw the AGRA graph. Let $G = \Gamma(\mathcal{P}^{u_1}) \cup \Gamma(\mathcal{P}^{u_2})$ be the set of visited POIs, by either user u_1 or u_2. Let $N = |G|$. Let $\rho^{u_1} \in \mathbb{R}_+^N$ and $\rho^{u_2} \in \mathbb{R}_+^N$ be two vectors of accessibility weights. Vector ρ^{u_1} is constructed as follows: $\forall i \in [0, N]$ **if** $p_i \in \mathcal{P}^{u_1}$ **then** $\rho_i^{u_1} = 1$ **else if** $p_i \in \Gamma(\mathcal{P}^{u_1})$ **then** $\rho_i^{u_1} = \sum_{v \in \mathcal{P}^{u_1}} \mathcal{A}_{v,p}$ **otherwise** $\rho_i^{u_1} = 0$. Similarly, we construct Vector ρ^{u_2}. This is a symmetric metric. Then, we define the accessibility weighted Jaccard similarity $S_{AWS}(u_1, u_2)$ as follows:

$$S_{AWS}(u_1, u_2) = \frac{\sum_{i \in [0,N]} \min(\rho_i^{u_1}, \rho_i^{u_2})}{\sum_{i \in [0,N]} \max(\rho_i^{u_1}, \rho_i^{u_2})} \tag{4.9}$$

- **Accessibility Weighted Antisymmetric Jaccard:** In this metric, we try to take into account the asymmetry that could exist in terms of influence between two users. To do so, we change the definition of G as follows: $G = \Gamma(\mathcal{P}^{u_1}) \cup \mathcal{P}^{u_2}$ Instead of extending both sets $\mathcal{P}^{u_1}$ and $\mathcal{P}^{u_2}$, we only extend the set of POIs visited by user u_1. Then, we compute the Accessibility Weighted Antisymmetric Jaccard $S_{AWA}(u_1, u_2)$ using equation 4.9. Note that $S_{AWA}(u_1, u_2) \neq S_{AWA}(u_2, u_1)$.

4.4.4 GeoSPF: An Implicit Social Factorization

Poisson factorization has been widely used to deal with numerous recommendation problems such as investigated by Charlin, Ranganath, McInerney, and Blei 2015; Ma, Liu, King, and Lyu 2011; Gopalan, Hofman, and Blei 2015; Chaney, Blei,

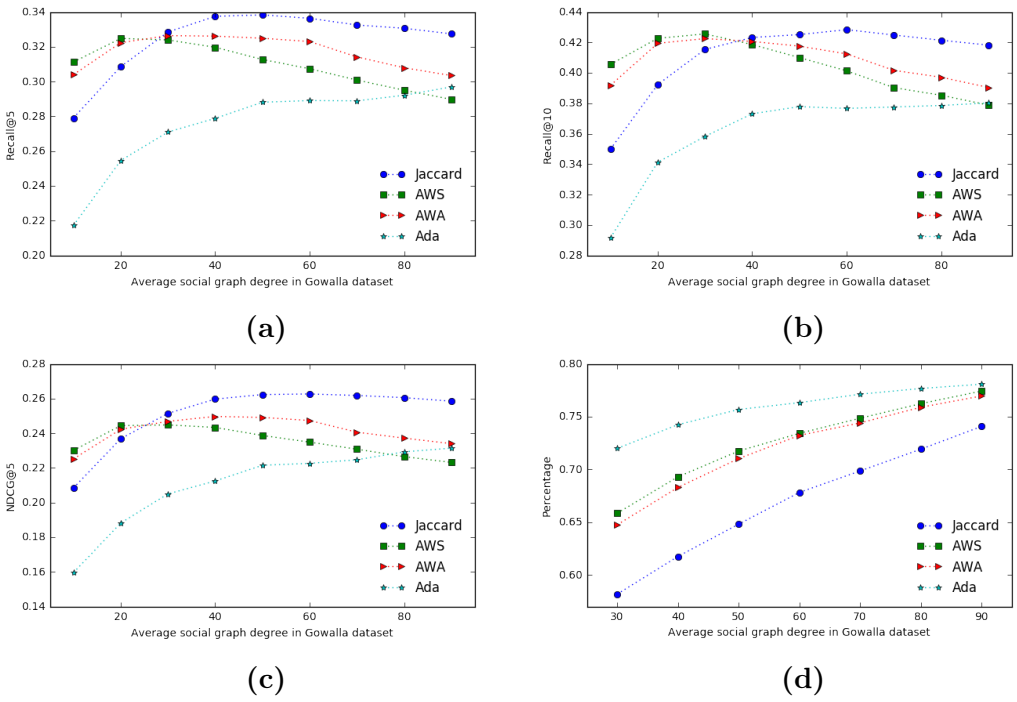


Figure 4.4: Performance results of 4 Methods on Gowalla. Each figure represents the performance results of the four metrics described in section 4.4 for a different number of edges in the graph. This number of edges is controlled by the average social graph degree.

and Eliassi-Rad 2015; Liu, Fu, Yao, and Xiong 2013. All these approaches propose extensions of the PF model where some weights are used to influence the latent factors distribution. These weights are usually related to the specificities of the considered problem.

Recently Chaney *et al.* proposed an extension to the Poisson factorization framework in Chaney, Blei, and Eliassi-Rad 2015 called SPF. We argue that extending the SPF recommendation model leads to a good approach for POI recommendation, because SPF has some interesting properties that match our needs:

- SPF is based on probabilistic matrix factorization which is known to perform well, both in terms of **quality** and **scalability**, in the context of sparse data containing only positive sample.*c.f.* the case of implicit feedback mentioned in the introduction of this paper (see Section 4.1).

- SPF allows integrating **social information** which is meaningful in our context. SPF takes as input the influence circle of each user, *i.e.* the set of

neighbors that may influence a user. In the context of POI recommendation, we consider that users influence each other mainly when they share experiences about POIs. Although we are constrained to deal with limited contextual information (only GPS locations and check-ins date are disclosed), we claim that such information is sufficient to characterize who is influencing who in a trip planning scenario.

- SPF separates the questions: *who is a member of the circle ?* from *how much influence does that member actually transmit ?* SPF assumes that the circle membership is known in advance, whereas the influence level is learned. This **separation** is essential in our case because the level of influence of a user does not depend on the POIs he/she shares with the other users, but rather on hidden (undisclosed) interactions that the users may have.

The idea of **GeoSPF** is to integrate the influence of the possible friends of the target user into the recommendation process, by taking into account the ratings of his/her neighbors. Differently from equation 4.1, GeoSPF considers the following distribution:

$$y_{i,j} \sim \text{Poisson}\left[\mathbf{u}_i^T \cdot \mathbf{v}_j + \sum_{k \in V(i)} \mathbf{s}_{i,k} \cdot x_{k,j}\right] \tag{4.10}$$

where $V(i)$ refers to the set of neighbors of user i in the ISN, and $\mathbf{s}_{i,k}$ refers to the latent social influence factor. This latent random variable models the influence that neighbor k has on user i. In equation 4.10 we still have the dot product of users and POIs latent vectors like in equation 4.1, but we introduce an additional social influence term which is the sum of the influences of each user in the neighborhood. The choice for the neighborhood $V(i)$ is important: $V(i)$ will contain all the neighbors who are the most similar to user i. Our central intuition is that we consider that the more two areas that two users are used to visit are accessible, the more they are likely to benefit from the influence from each other.

A probabilistic graphical model of GeoSPF is proposed on Figure 4.5. In contrast with SPF, here we are not using an explicit social networks. Furthermore, we can tune the recommendation quality based on the used similarity metrics and on the graph filtering. Indeed, for each constructed graph we can choose different selection criteria for the neighborhood of the users. We can also apply the filters either on the accessibility graph or on the implicit social graph. This adds some flexibility to GeoSPF.

4.4.5 Inference

GeoSPF is based on a Bayesian generative process. The first purpose of such a process is to model some underlying unobserved data assuming we have already computed the latent vectors.The parameters estimation, or inference, of the model requires to reverse this process, *i.e.* estimate the parameters thanks to the observed data. The goal is to compute the posterior on hidden variables. Unfortunately, because it has not a closed analytic form, exact posterior is impossible to compute. Consequently it is necessary to use an approximation. Many methods exist to approximate this posterior: Expectation propagation, Gibbs sampling, message passing and MCMC are the most widely used ones. However, they require to be tailored to the generative process. Variational methods have emerged as the most efficient and scalable alternative to fit a Bayesian model. This is why we have chose such inference method in GeoSPF.

The idea behind variational inference is to set a distribution family on latent factors indexed by variational parameters. Then, the method tries to find these variational parameters which make the true posterior as close as possible to this family. As in Gopalan, Hofman, and Blei 2015 Kullback-Liebler distance is used to measure the distance between the indexed distribution family and the true posterior. Then, an alternative minimization method is used to find the optimum parameters. Specifically, we chose a mean-field variational family where each latent variable has its own variational parameter.

Finally the stochastic variational Bayes inference used to learn the parameters has a complexity of $\Theta(N(K+V))$ where N is the number of non-zero entry in matrix $\mathbf{X}$, K the number of latent factors, and V the maximum user degree in the social graph. This model allows to catch the influence of users' friends on their check-ins. SPF is the approach we will apply in one of the variants of our model but

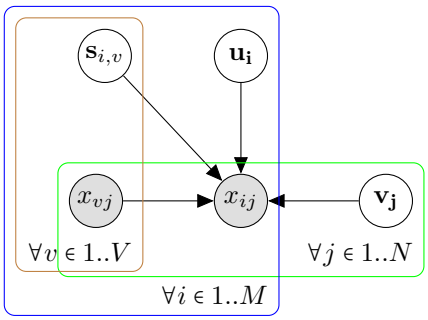


Figure 4.5: Graphical model of our approach.

with a social networks built thanks to a similarity network based on accessibility. We chose it because its complexity is much lower than alternative methods.

4.5 Experimental Evaluation

In this section we evaluate the performance of our method for POIs recommendation. We evaluate how our method fares in comparison with state-of-the-art approaches. We also study how our proposed social similarity metrics perform, and how they improve the recommendation quality. Before we describe the experiments, we first present the LBSN datasets that we crawled and the evaluation metrics we used.

4.5.1 Data Sets and Metrics Description

We conducted experiments on three real-world datasets containing check-ins from widely-used `YFCC`, `Gowalla` and `Foursquare` LBSNs. To assess the behavior of our solution at various geographical scales, we filtered the datasets such that they cover a small, medium and large area respectively. Namely, Gowalla@Paris covers a city, Foursquare covers a region (around Paris), Gowalla covers a country (France) and YFCC covers Europe.

Figure 4.6 depict how the check-ins are geographically distributed in the YFCC datasets. The YFCC dataset has been proposed recently by Thomee et al. 2016. It is the largest dataset existing for POI recommendation: the full dataset contains 50 millions geo-located check-ins. As a consequence, most of existing approaches for POI recommendation fail to cope with such large volume of data. The Foursquare

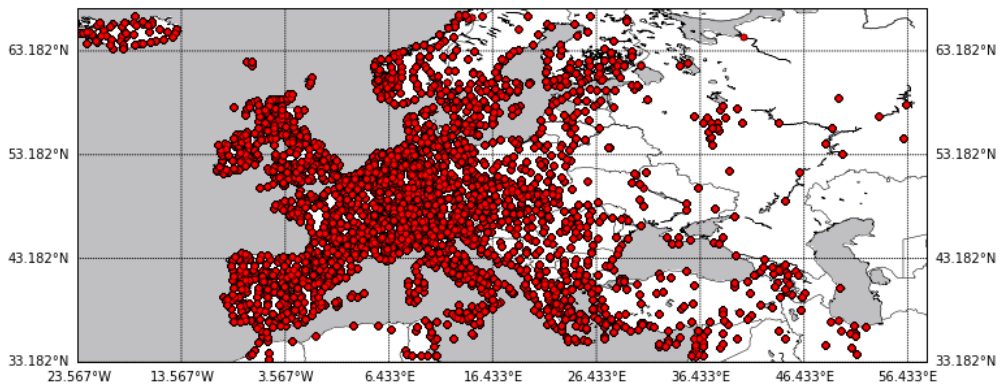


Figure 4.6: European YFCC dataset.

Table 4.1: Statistics on the datasets

Dataset	#Check-ins	#Users	#POIs	avg #POIs	Density
`Gowalla@Paris`	42323	2384	4895	5.6	0.362 %
`Foursquare` [1]	109077	4825	19645	3.1	0.115 %
`Gowalla` [2]	191365	6749	24353	4.1	0.116 %
`YFCC` [3]	48453357	214328	12758657	61.2	0.0017 %

dataset has been used in Yuan, Cong, Ma, et al. 2013. It contains check-ins made between Apr. 2012 to Sep. 2013. The Gowalla dataset has been used in Cho, Myers, and Leskovec 2011. It contains check-ins gathered over the period of Feb. 2009 to Oct. 2010. Table A.2 presents the basic statistics regarding the datasets we used.

We can observe that the datasets are very sparse (*i.e.* very low sparsity value). Indeed, since we expect to demonstrate that our approach is viable for sparse datasets, we did not filter out users nor POIs that have few check-ins. Thus, the average number of visited POIs by user is small (less than 6) in the used datasets.

Each dataset have been split into a training and a test set through a random process: approximately 20% of the data are used for post-inference testing while the rest is used for training. We are interested in evaluating the ranking accuracy of our approach. For this reason we use the usual recall (denoted $recall@N$) and the normalized discounted cumulative gain (denoted $nDCG_p$) as main evaluation metrics.

4.5.2 Comparison with competitor models

In order to estimate the effective benefits of our approach with respect to well-known state-of-the-art solutions[4]. We have chosen efficient recommendation models Specifically we have compared $GeoSPF$ along with the following recommendation techniques:

- **NMF:** Non Negative Matrix Factorization proposed by Lee and Seung 2000 is

[1] data available at https://sites.google.com/site/yangdingqi/home/foursquare-dataset

[2] data available at: http://www.yongliu.org/datasets

[3] data available at: https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67

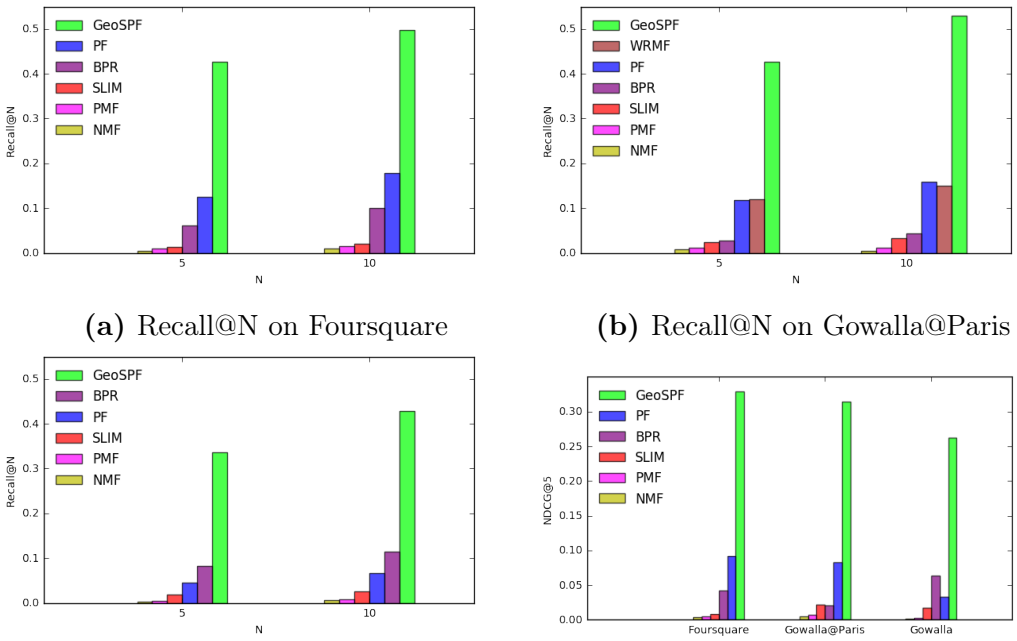
[4] The code to reproduce the experiments described below is available at https://gitlab.telecom-paristech.fr/griesner/geopfModeles

one of the most popular factorization model.This model factorizes the original data matrix thanks to multiplicative iterative update rules

- **PMF:** Probabilistic Matrix Factorization proposed by Salakhutdinov and Mnih 2007 is an effective probabilistic factorization model based on Gaussian priors on dat
- **SLIM:** Sparse Linear Methods proposed by Ning and Karypis 2012 are adapted to sparse datasets. They are based on a linear model that exploits a sparse aggregation of coefficients
- **BPR:** Bayesian Personalized Ranking proposed by Rendle, Freudenthaler, Gantner, and Schmidt-Thieme 2009 has been designed to tackle implicit feedback problems. This is a scalable probabilistic approach that basically optimizes a ranking criterion. BPR is a strong competitor among the state-of-the-art approaches.
- **WRMF:** Weighted Regularized Matrix Factorization proposed by Hu, Koren, and Volinsky 2008 has been designed precisely for implicit feedback datasets, which perfectly fits the requirements of POI recommendation.
- **PoissonMF:** This is a recent probabilistic Poisson based model proposed by Gopalan, Hofman, and Blei 2015 that we used as a building block for our approach.
- **GeoSPF:** This is our approach. It uses equation 4.10 presented in the previous section.

Figure A.4 shows the overall performances of all the above baseline methods. On Figure A.4a the Recall@5 and Recall@10 are reported for the Foursquare dataset. The same metrics are reported on Figures A.4b (resp. A.4c) for Gowalla@Paris (resp. Gowalla) dataset. Finally, Figure A.4d reports the NDCG@5 metric for the three datasets.

As a first observation, on the first three datasets (Foursquare, Gowalla@Paris and Gowalla) we notice that our approach (GeoSPF) significantly outperforms all the other ones. As expected, NMF and PMF do not yield a good quality since they were not designed to cope with implicit feedback datasets. This is consistent with the results in Liu and Xiong 2013. Although SLIM is known to perform well on sparse datasets, it fails to achieve a good quality in our context because it assumes explicit feedback (instead of implicit one). Unfortunately the complexity of WRMF makes it practically useless on large datasets: the WRMF computation



(a) Recall@N on Foursquare

(b) Recall@N on Gowalla@Paris

(c) Recall@N on Gowalla

(d) NDCG@5 on three datasets

Figure 4.7: Performance comparison *w.r.t.* state-of-the-art approaches for three datasets: Foursquare, Gowalla@Paris and Gowlla. We plot Recall@N for N=5 and N=10 on Figures A.4a, A.4b and A.4c. We plot NDCG@5 on figure A.4d. We observe that GeoSPF outperforms significantly baselines on the three datasets for the three performance measures.

time is prohibitively long beyond 5000 POIs. This is why we only get the recall@N for Gowalla@Paris, but not the two other datasets where the number of POIs is too high. This is due to the fact that the complexity of WRMF depends highly on the number of POIs, and according to table 5.1, Gowalla@Paris has less POIs than the others. Among all the state-of-the-art competitors, PF achieves the best quality. Thus, we focus our analysis on comparing PF vs. GeoSPF. As a major result, the relative benefit of GeoSPF on all the datasets is around 200%. This impressive gain makes GeoSPF suitable for POI recommendation over wide geographical areas. It confirms that exploiting restricted contextual information (only GPS and check-in date) through a combined geographical/social solution yields a high quality for POI recommendation.

4.6 Conclusion

In this chapter we have proposed a new scalable approach for the POIs recommendation task in LBSNs called GeoSPF. The main goal of GeoSPF was to build an implicit social network which does not suffer from the lack of explicit users' feedback regarding their check-ins. Based on the new concepts of *accessibility* and *social* similarity metrics that we have introduced in this work, our GeoSPF approach succeeded **(i)** to build efficiently an implicit scalable factorization model and **(ii)** to capture the user's social similarity and finally **(iii)** to present significant better results than baselines on large-scale datasets. We have demonstrated with extensive experiments that GeoSPF significantly outperforms all the alternative approaches in terms of *recall* and *NDCG*.

Chapter 5

ALGeoSPF: A Clustering Based Factorization Model for Large Scale POI Recommendation

The implicit feedback problem - described in the previous chapter - is usually solved by adding more contextual parameters (geographical, content...) into the model which tends to increase its associated complexity. As a result, most of the existing methods suffer from scalability and are unusable on real-world data. Actually scalability is a common practical limitation for most of existing approaches in POI recommendation. This chapter presents a new geographical matrix factorization based approach, namely ALGeoSPF, that is scalable. ALGeoSPF exploits a geographical clustering algorithm that allows both to increase the density of the dataset and to generate more personalized recommendations thanks to the definition of user mobility behavior classes. We present experiments conducted on a global real-world dataset containing more than 50 millions check-ins. This work has resulted in two publications: one Griesner, Abdesssalem, Naacke, and Dosne 2018 in EGC international conference[1] and another in BDA 2017 [2]

After a short introduction in Section 5.1 and a brief related work of existing clustering methods in Section 5.2, we present the ALGeoSPF model in Section 5.3. Our experiments on a real-world dataset are presented in Section 5.5. Finally the Section 5.6 concludes this chapter.

[1] https://egc18.sciencesconf.org/
[2] https://project.inria.fr/bda2017/

5.1 Introduction

The recommendation models that we discussed in the previous chapters are efficient on small volumes of data. However they cannot handle realistic datasets where the number of users and POI exceeds some thousands. We consider in this chapter the specific problem of POIs recommendation where the check-ins data are covering a large geographical area which can be as large as the entire world. As a direct consequence the volume of data collected is much more important than smaller datasets. However in the same time the sparsity of the dataset increases. Indeed when the area covered by a dataset is becoming wider and wider, the dataset is becoming sparser and sparser because the number of POIs is increasing whereas the average number of visited POIs per user remains invariant. In other words, the density defined as the average fraction of POI that a user has visited, is decreasing.

For instance, we analyzed the YFCC dataset[3], and compared the density at the level of a country (France), a region (Europe), and the world. The density values are respectively 0.50%, 0.21% and 0.11%, *i.e.* the world wide level is 5 times less dense than the country level. To tackle that low-density challenge, recent recommendation solutions have been proposed among those Poisson Factorization (noted PF in the following) is known to be one of the most efficient ones. However PF still fails to handle very low-density dataset such as the YFCC. To figure out the effect of low density data on a recommendation task we applied the PF algorithm on the YFCC: the recommendation restricted to France yields a quality of 2.4% in terms of recall@10, whereas the same recommendation task covering the larger Europe continent performs almost 67% times less. Clearly, the low-density of wide area datasets is still an open challenge in the context of POI recommendation. We propose below a definition of the problem that this chapter aims at solving.

Problem 5.1.1. ***Large-Scale POIs Recommendation:*** *Given a large check-ins collection $\mathcal{D}$ of low density, the goal of large-scale POIs recommendation is to provide to a given user u a top-k list of POIs he could be interested to visit with the highest probability.*

5.1.1 Contributions

In this chapter we deal with the challenges involved by large-scale geographical datasets for POIs recommendation. One of our main assumptions is that users'

[3]We present in details the datasets in Section 5.5.

mobility behaviors should be analyzed at different spatial scales and that social influence can rarely spread throw users from different categories (globetrotters versus city dwellers). In this work we propose an efficient factorization model for POIs recommendation (namely ALGeoSPF) that takes into account of the contextual information as well as the social influence based on specific users' mobility behaviors. Moreover we exploit a flexible hierarchical clustering structure of **super-POIs** to detect these behaviors.

The idea of super-POIs is to catch the different scales existing in the data and to recommend to each class of users the corresponding super-POI she needs. In order to classify users, a naive solution could be for instance to compute for each user the diameter of the bounding box of all her check-ins, and then to define a threshold: if the user's diameter is above, she is a globetrotter, otherwise she is an urban user. The problem with this solution is that it does not take into account of the existing density fluctuations of check-ins in the datasets. That is to say, regions with few check-ins tend to have long distance trips although the users will not necessarily be globe-trotters.

This is why a more efficient approach should be proposed first to cluster the check-ins based on the density, through a spatial clustering preprocessing step. Here we propose to use the STatistical INformation Grid-based clustering (STING) method proposed by Wang, Yang, and Muntz 1997. STING algorithm offers an ideal solution for our problem given that it will create a hierarchical structure based on the density of check-ins, minimizing the density fluctuations in the leafs of the tree. Then we propose to apply a geographical social Poisson factorization method exploiting only the social influences of the users with the most similar mobility behaviors.

To sum up, in this chapter we propose a hierarchical geographical clustering-based approach using an implicit social Poisson factorization model which takes into account geographical influence and the implicit social influence for the POIs recommendation problem.

We can summarize the contributions that we achieve in this chapter as follows:

- We propose a **scalable** probabilistic factorization approach for POIs recommendation problem.
- We propose a **hierarchical structure** to define several levels of **superPOIs** thanks to a flexible clustering algorithm.

- We build more **personalized recommendations** based on users specific mobility behaviors.
- Finally we conduct exhaustive experiments on a large-scale dataset which confirm the **efficiency** of our approach.

5.1.2 Road Map

The rest of this chapter is organized as follows. Section 5.2 presents the related work. Section 5.3 gives an exhaustive formulation of the problem we are dealing with and introduces our hierarchical clustering algorithm. Section 5.5 presents some experiments and the results that we conducted on 3 real-world datasets. Finally Section 5.6 concludes this chapter.

5.2 POI Recommendation at Large Scale

To handle the problem of implicit behavioral data records, a model has been proposed some years ago by Hu, Koren, and Volinsky 2008. Their model introduces a distinction between the user's preference estimation, and the confidence we can have into this estimation. The authors have demonstrated that their approach was more efficient on small datasets than alternative existing methods. This method has been exploited and augmented by the authors of [Lian et al. 2014] to include the geographical influence of POI by modeling the spatial clustering phenomenon [Ye, Yin, Lee, and Lee 2011; Zhang and Chow 2013] directly into the factorization process. However the complexity is far too high to use it on large-scale datasets.

Few works have been proposed to deal specifically with the large scale POI recommendation problem. Recently [Zong et al. 2016] have proposed a cascading bandits model that can deal with large scale datasets. However their approach is not efficient with sparse data. Recently [Lee and Abu-El-Haija 2017] have proposed a new method to recommend videos on large-scale datasets. Their approach is based on a content-based model that exploits deep video embeddings. Moreover very few datasets are available to test the models scalability for POI recommendation. Recently [Sidana et al. 2017] have proposed an approach to build a dataset with large dimensions that can be exploited for scalability testing.

Poisson Factorization has emerged as a successful solution to handle large-scale volumes of data given that this models is generated only with the observed values. It is a scalable probabilistic factorization model applied for factorization by

[Charlin, Ranganath, McInerney, and Blei 2015] that outperforms state-of-the-art models subject to sparsity and diversity constraints. Many recent works have proposed to increase the recommendation quality with social influence such as [Cho, Myers, and Leskovec 2011; Zhang and Chow 2013; Zhang and Wang 2015; Cheng, Yang, King, and Lyu 2012]. The idea of these methods is to exploit the knowledge a user's friends have on unvisited POIs. Some years ago [Zhang and Wang 2015] have proposed a model called LTSCR which uses the social similarities of users and integrate them into the factorization model. Unfortunately the social networks are usually not associated in LBSNs real-world datasets. The extraction of an implicit social network has been widely investigated last years by [Losup et al. 2014; Hu, Koren, and Volinsky 2008] but remains still unexploited for POIs recommendation.

The impact of the contextual influences on the scalability has been widely studied these last years in related works [Liu, Fu, Yao, and Xiong 2013; Cho, Myers, and Leskovec 2011; Zhang and Chow 2013; Griesner, Abdessalem, and Naacke 2015; Cheng, Yang, King, and Lyu 2012; Gao, Tang, Hu, and Liu 2013; Lian et al. 2014]. These methods are based before all on the assumption that geographical proximity of POIs significantly influences the users' decision process. For instance considering Tobler's first law of geography [Miller 2004] we can imagine that generally the user next check-in will tend to be close to the last visited one. Indeed, a spatial clustering phenomenon can be easily observed (for instance, around the main cities) in LBSN datasets, since users tend to visit nearby POI.

5.3 ALGeoSPF: Local-Global Spatial Influence Modeling

In this section we present ALGeoSPF and we describe our geographical clustering algorithm. Then we propose a new approach for generating personalized class of mobility behavior.

5.3.1 General Idea

Existing works fail to model the users' mobility behaviors both at the microscopic scale and at the macroscopic scale. The definitions of these scales are flexible: for instance the microscopic scale could correspond to cities and the macroscopic scale could correspond to countries. Indeed this is a frequent problem with massive datasets: working on large-scale problems does not allow for interpretation on a

microscopic scale.

Furthermore, in reality different classes of travelers exist simultaneously in the data: some users are more likely to make long distance trips while others will be limited to restricted areas such as cities or regions. We will call in the following the former the *globetrotters* and the latter the *urban* users as defined below. We argue that the urban users should not benefit from the social influence of the globetrotters, and reciprocally.

Definition 5.3.1. ***(Globetrotter/Urban)*** *Users who travel frequently to faraway POIs are called globetrotters. Reciprocally users who travel only to geographically close POIs are called urban users.*

Existing POIs recommendation approaches [Liu and Xiong 2013; Gao, Tang, Hu, and Liu 2013; Hu, Koren, and Volinsky 2008; Lian et al. 2014] cannot deal with these problems at a large-scale, trying only to adapt traditional collaborative filtering solutions to POIs recommendation. Our work differs from existing approaches because it targets the usual case where the only available contextual information is the GPS location and the date of checkins. This opens the challenge to infer some kind of contextual social knowledge from the raw checkins data.

5.3.2 Super-POIs

To address the **low-density challenge** involved by the large-scale POI recommendation problem we first investigate solutions to increase the density of a dataset without reducing the geographical area covered by the dataset. To do so we select a fraction of the POIs while ensuring that the selected fraction covers the entire area. We propose to filter the POIs based on their geographical position and the number of visits they received.

The main idea is to define a set of **superPOIs**, each one being representative of a group of POIs. The set of superPOIs constitutes itself a hierarchical structure including other superPOIs. For instance, if we consider one superPOI per city, the YFCC dataset would have 15,886 POIs and a density of 0.23%, which matches the density requirement for the recommendation task. While segmenting the space into regular cells to make groups of POIs is rather straightforward, the basic fixed-size grid segmentation does not apply in its own because it ends up with superPOIs representing many POIs (*e.g.* cities) and some other superPOIs representing very

few POIs (*e.g.* deserts). Here is below a definition of superPOIs:

Definition 5.3.2. ***(SuperPOI)*** *The aggregation of several distinct POIs or superPOIs constitutes a superPOI. A superPOI corresponds to a unique specific geographical area. It is a set of existing POIs or superPOIs.*

The areas defining the superPOIs are per-wise disjoints. Intuitively a hierarchical structure is required to fit the density scale of the dataset: some dense geographical areas will be divided into a lot of superPOIs while less dense areas will be divided in less superPOIs. We define formally $\mathfrak{P}$ recursively: let $\mathfrak{P}^O = \mathcal{P}$ and let $\mathfrak{P}^{k+1} = \{\{p_1, p_2, ...\} | \{p_1, p_2, ...\} \in \mathfrak{P}^k\}$. Then we get that: $\mathfrak{P} = \bigcup_k \mathfrak{P}^k$. Different clustering techniques can be used to build this family. We discuss with more details the advantages of this structure in Section 5.4. The same observation applies to users: some superPOIs are visited by many users (*e.g.* famous districts) whereas some other superPOIs are visited by very few people (off-road areas). This amplifies the skew on the distribution of the number of user per (super)POIs and effects in degrading the recommendation quality.

Consequently, we need a space segmentation approach that guarantees that each superPOI has been visited by at most n users. Moreover, this bound on the maximum number of users per POI is meaningful for the recommendation task because this allows for computing closer similarity between users. User similarity is an essential building block of the recommendation task as detailed in Section 4.4. For instance we can imagine two users having visited Hollywood and Venice beach (part of LA city) and no other place in common. If the superPOI is LA, then they only have one common point, namely LA. If no, LA is divided into four districts with one superPOI per district, then the users have two common points and become closer neighbors during the neighborhood discovery phase of the recommendation task.

5.3.3 Mobility Behaviors

We observe that distinct mobility behaviors emerge from the data. However in most of existing works, no distinction is made between these distinct users' mobility behaviors. As a consequence the social influence of globetrotters could be spread, for instance, to urban users, decreasing the recommendation quality. One of the aims of our approach is to exploit the social influence of users who belong to the same class.

As a result the second challenge to be addressed while dealing with wide area datasets is to capture the **mobility behaviors** of users both at several geographical scales from the local one (city) to the global one (world). Inded considering a very wide geographical area dataset such as the YFCC, we observe that some users still move rather locally, only visiting POIs in a small area, while some other users really move further across a large area. One can distinguish several classes of users. For instance the urban users visiting only a very restricted area (*e.g.* a city) and the globetrotters who have check-ins all over the world.

The main motivation for considering user mobility classes is to ensure that for every user, enough information is available to the recommendation model. Our approach is not limited to only 2 classes and can handle any number of classes depending on the datasets specificities. In practice, the recommendation task requires to know at least 5 distinct POIs per user to train the model and then make prediction of acceptable quality. Thus to recommend an urban users one has to know the various visited locations. This somehow contradicts the first objective about wide area recommendation. For example, in a wide area scenario where a superPOI may represent a cell as big as a city, an urban user may have all her check-ins fused into a single superPOI. That means that on the one hand the density requirement requires to aggregate POIs into larger superPOIs, on the other hand the low mobility pattern of urban users requires to keep thinly located POIs.

5.3.4 Final Objective

To handle this **tradeoff** between low-density , we propose a unified solution that consistently relies on the superPOI definition in order to:

1. Aggregate POIs in a flexible way: the geographical aggregation level being controlled by choosing an upper bound density measure (c.f. Section 5.3).

2. Classify users such that they belong to the aggregation level where they satisfy the recommendation learning requirement, *i.e.* the minimal number of visits per user.

By doing this we are able to make recommendation for every user at the adequate aggregation level. In the following we instantiate this framework for two classes: the globetrotters are recommended at the superPOI ("country") level. The urban users are recommended at the original POI level where the area has been restricted to a continent to match the density requirement. An assumption we make

is that urbans should not be influenced by globetrotters. That is to say the globetrotters are excluded from the urban user recommendation scenario. Empirical experiments (presented in Section 5.5) confirm the benefit of excluding globetrotters from the urban recommendation case.

Problem 5.3.1. ***Spatial Clustering Impact:*** *Given a large check-ins collection $\mathcal{D}$ of low density and a recommendation model, the problem is to to improve the recommendation quality by leveraging on a spatial clustering method to characterize various user mobility classes. Specifically the clustering method is used to characterize globetrotters and urban users based on their geographical mobility.*

Our Augmented Local-Global GeoSPF model (denoted ALGeoSPF) consists in defining local and global layers of superPOIs in order to increase the dataset density, select a class of users - based on a personalized optimal parameter N_{max} - and therefore enable the GeoSPF recommendation task (see Section 4.4) while targeting a dedicated set of superPOIs and users. The first advantage of ALGeoSPF is its ability to detect users' mobility behaviors either at a local scale or at more global scales. ALGeoSPF captures the existing mobility behaviors of urban users and globetrotters that we observed in our datasets.

The second advantage of our multi-scale solution is that it allows isolating all the steps of the recommendation process (social network inference, learning, prediction) within each class of user, avoiding thereby any noise propagation through users' classes. This reflects the fact that, for instance, the mobility behavior of urban users can not by influenced by the mobility of the globetrotters (users of a higher spatial scale).

5.4 Hierachical SuperPOIs Layers

The ALGeoSPF approach exploits a hierarchical structure of super-POI. The idea of hierarchical structures for POIs recommendation has been investigated recently by [Zhang, Wang, et al. 2017]. The authors investigated the idea to exploit a hierarchical structure of categories and a geographical influence distributed between different regions. However their problem is different than ours, given that they propose an approach to predict the category of the next POIs visited.

SuperPOIs aggregation enables to deal with large-scale low-density datasets by aggregating parts of the original datasets. We note $\mathfrak{P}$ the set of all the superPOIs

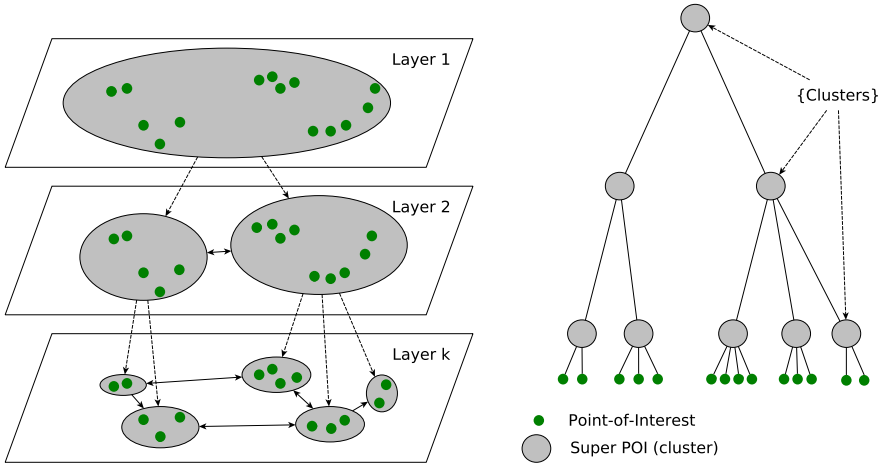


Figure 5.1: An illustration of three different hierarchical layers containing some check-ins and the superPOIs.

layers. We define a multi layer structure to gradually aggregate the POIs into superPOIs visited by an increasing number of users. Figure 5.1 represents an illustration of this structure. This allows to meet the density requirements of our recommendation task for different classes of users. Let k be the number of layers, $\mathfrak{P}_k$ be the set of superPOIs defined at level k, and N^k_{max} be the maximum number of users visiting a superPOI in $\mathfrak{P}_k$. The following condition defines the maximum area that a superPOI represents, noting p a superPOI: $\forall p \in \mathfrak{P}_k, N(p) < N^k_{max}$ with $N(p)$ being the total number of distinct users visiting the POIs in p. Moreover, each level aims to aggregate the POIs as much as possible to ensure that every p is not "too small", *i.e.* there is no p' in the upper layer $\mathfrak{P}_{k+1}$ such that p' aggregates p and p' satisfies $N(p') < N^k_{max}$.

5.4.1 Geographical Clustering Algorithm

We adopt a clustering approach that consists in dividing the initial geographical space (*e.g.* the entire world for the YFCC large scale dataset) into even rectangular cells. As proposed by [Wang, Yang, and Muntz 1997] and usual quad-tree based approaches, a cell can be recursively divided into 4 cells. Thus we construct a tree where the root is the whole world map and each node is a quarter of its parent region. The principle of the algorithm is to recursively divide a cell c until it satisfies the condition concerning the number of different users who made check-ins in that cell: $N(c) < N_{max}$. The cells satisfying that condition are chosen to be the superPOIs. The result of the clustering algorithm is a set of superPOI cells denoted S. The N_{max} parameter allows controlling the aggregation level. We specify the

condition of a superPOI cell based on $N(\cdot)$ instead of the number of POIs because it better detects users' mobility behaviors when many popular POIs are close to each other within an area which has a few number of POIs (*e.g.* two close theaters, two close museums...), which happens to be a frequent case. The algorithm is presented in algorithm 1 below.

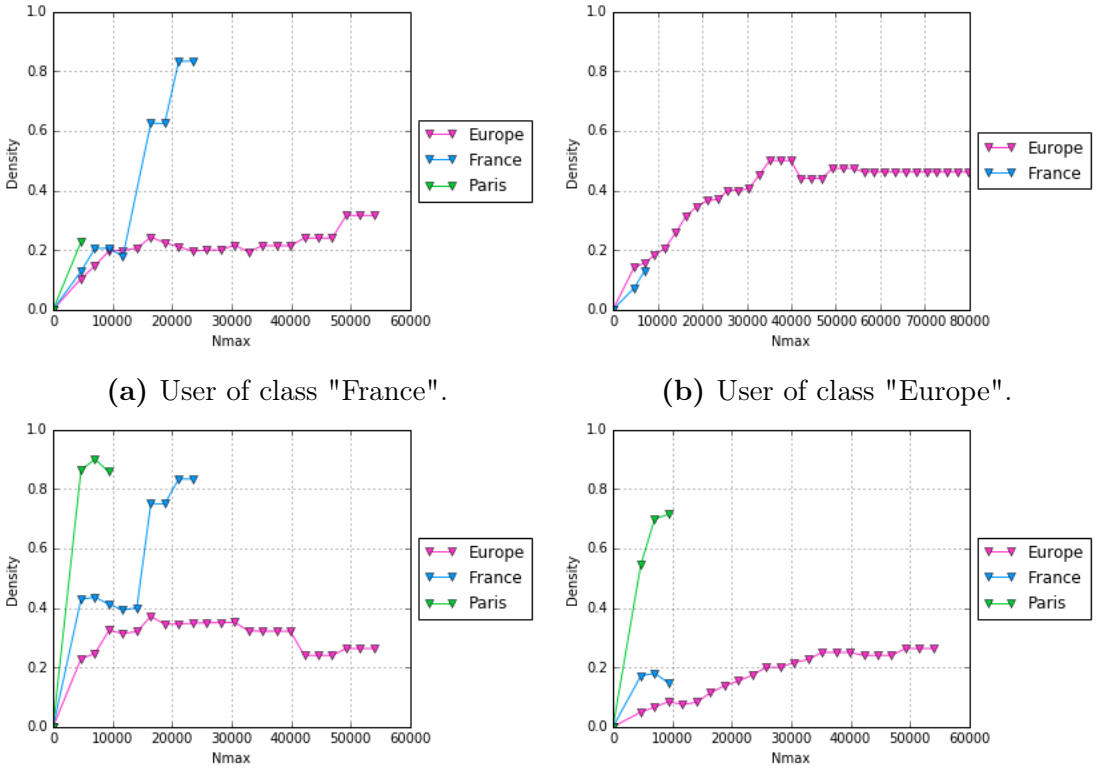
Algorithm 1 Top-down Clustering Method for ALGeoSPF

1: **Input**:
   - N_{max}: maximum number of users having visited a cell.
2: **Global Output**:
   - S: the set of superPOIs cells.
3: **Initialize**: $S \leftarrow \varnothing$
4: **function** WORLDTOSUPERPOIS (C: a cell)
5: Split C into 4 even rectangular cells $C_1, \cdots, C_4$
6: **for each** C_i **do**
7: **if** $N(C_i) > N_{max}$ and $\#POIs(C_i) \geqslant 2$ **then**
8: worldToSuperPOIs (C_i)
9: **else** Put C_i into S

5.4.2 Personalized Class Selection

Since the recommendation process is known to perform better with higher density datasets, we explain how the clustering method allows for improving the density in a more personalized way. Given a user asking for a recommendation, we rely on the clustering algorithm to tune (*i.e.* to optimize) the two parameters that effect in changing the dataset density: the initial cell and the cluster size.

- The **initial cell** on which to apply the clustering. Considering the entire world as the initial cell suits well for users with many checkins spread all over the world. But for most users, the area covering all the user's checkins is smaller (*e.g.* Europe, France, Paris). Considering a smaller initial cell effects in increasing the dataset density.
- The **cluster size** is defined by N_{max}. Increasing N_{max} will result in fewer and larger superPOI therefore will increase the dataset density. However N_{max} is bounded: for each user, there exists a maximal N_{max} (denoted N_{max}^{user}) beyond which the recommendation is no more possible because the user will not have



(a) User of class "France". **(b)** User of class "Europe".

(c) User of class "Paris". **(d)** User of class "Paris".

Figure 5.2: We represent the evolution of the density of the dataset for different users and for different values of N_{max} and three different geographical areas: Europe, France and Paris. We observe that each user has a peak of density depending on an optimal N_{max} which characterizes the class of the user.

visited enough distinct superPOIs (we require at least 5 distinct visits per user).

While tuning these parameters, we found that their impact on the density varies a lot depending on the user, which justifies our attempt to propose a personalized approach. For some users a rather small initial cell (*e.g. Paris*) yields the highest density. For some other users, although *Paris* could be the initial cell (because the *Paris* cell contains all the user's chekins), the *France* initial cell allows for higher N_{max} and eventually brings a higher density. More generally, we observed that such optimization method helps to detect several classes of users that share the same near-optimal couple of (initial cell, N_{max}) parameters. Here we define the class of a user as the initial geographical area that matches the best her needs. In the approach described above we have introduced distinct classes of users (urbans,

globetrotters...) defined a priori at the beginning. These classes are associated to specific mobility patterns. However we observe that each user belongs most significantly to one or another of these classes. Thus, as we said above, we associate each user with her personalized optimal N_{max}^{user} which is the N_{max} that maximizes the density of distinct check-ins in the clusters.

To put this into evidence, we have selected three geographical areas (that is to say, three distinct root cells: Europe, France and Paris city) and for each of them we have selected some users to observe how their density evolves for a varying N_{max} inside each area. For instance in figure 5.2a the N_{max} is around 24,000. Then, once the N_{max}^{user} has been computed for each user, we can select the social influences the most adapted to the user. We observe on figures 5.2 that the densities are very high (close to 1). This is due to our filtering process. Indeed we removed from the dataset users and POIs with less than 5 check-ins. This allows us to detect more the distribution of optimal N_{max} among users. For instance on figure 5.2c we observe that the user seems to belong to class "Paris". However if we observe in the data a lot of users in class above ("France") we can include this user into "France" as well, giving more flexibility to the model.

5.5 Experimental Evaluation

In this section we evaluate the performances of our method for POIs recommendation. We evaluate how our method fares in comparison with state-of-the-art approaches. We also study how our proposed social similarity metrics perform, and how they improve the recommendation quality. Before we describe the experiments, we first present the LBSN datasets that we crawled and the evaluation metrics we used.

5.5.1 Datasets and Metrics Description

We conducted experiments on three real-world datasets containing check-ins from widely-used `YFCC`, `Gowalla` and `Foursquare` LBSNs. To assess the behavior of our solution at various geographical scales, we filtered the datasets such that they cover a small, medium and large area respectively. Namely, Gowalla@Paris covers a city, Foursquare covers a region (around Paris), Gowalla covers a country (France) and YFCC covers Europe. Figure 4.6 depict how the check-ins are geographically distributed in the YFCC datasets. The YFCC dataset has been proposed recently by [Thomee et al. 2016]. It is the largest dataset existing for POI recommendation: the full dataset contains 50 millions geo-located check-ins.

Table 5.1: Statistics on the datasets

Dataset	#Check-ins	#Users	#POIs	avg #POIs	Density
`Gowalla@Paris`	42323	2384	4895	5.6	0.362 %
`Foursquare` [1]	109077	4825	19645	3.1	0.115 %
`Gowalla` [2]	191365	6749	24353	4.1	0.116 %
`YFCC` [3]	48453357	214328	12758657	61.2	0.0017 %

As a consequence, most of existing approaches for POI recommendation fail to cope with such large volume of data. The Foursquare dataset has been used in [Yuan, Cong, Ma, et al. 2013]. It contains check-ins made between Apr. 2012 to Sep. 2013. The Gowalla dataset has been used in [Cho, Myers, and Leskovec 2011]. It contains check-ins gathered over the period of Feb. 2009 to Oct. 2010. Table 5.1 presents the basic statistics regarding the datasets we used. We can observe that the datasets are very sparse (*i.e.* very low sparsity value). Indeed, since we expect to demonstrate that our approach is viable for sparse datasets, we did not filter out users nor POIs that have few check-ins. Thus, the average number of visited POIs by user is small (less than 6) in the used datasets.

Each dataset have been split into a training and a test set through a random process: approximately 20% of the data are used for post-inference testing while the rest is used for training. We are interested in evaluating the ranking accuracy of our approach. For this reason we use the usual recall (denoted $recall@N$) and the normalized discounted cumulative gain (denoted $nDCG_p$) as main evaluation metrics.

5.5.2 Comparison with competitor models

To measure the benefit of our work with respect to well-known state-of-the-art solutions[4], we have compared ALGeoSPF with the following recommendation models:

- **NMF:** Non Negative Matrix Factorization [Lee and Seung 2000] is one of

[1] data available at https://sites.google.com/site/yangdingqi/home/foursquare-dataset
[2] data available at: http://www.yongliu.org/datasets
[3] data available at: https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67
[4] The code to reproduce the experiments described below is available at https://gitlab.telecom-paristech.fr/griesner/geopfModeles

the most popular factorization model.This model factorizes the original data matrix thanks to multiplicative iterative update rules

- **PMF:** Probabilistic Matrix Factorization [Salakhutdinov and Mnih 2007] is an effective probabilistic factorization model based on Gaussian priors on data
- **SLIM:** Sparse Linear Methods [Ning and Karypis 2012] are adapted to sparse datasets. They are based on a linear model that exploits a sparse aggregation of coefficients
- **BPR:** Bayesian Personalized Ranking [Rendle, Freudenthaler, Gantner, and Schmidt-Thieme 2009] has been designed to tackle implicit feedback problems. This is a scalable probabilistic approach that basically optimizes a ranking criterion. BPR is a strong competitor among the state-of-the-art approaches.
- **WRMF:** Weighted Regularized Matrix Factorization [Hu, Koren, and Volinsky 2008] has been designed precisely for implicit feedback datasets, which perfectly fits the requirements of POI recommendation.
- **PoissonMF:** This is a recent [Gopalan, Hofman, and Blei 2015] probabilistic Poisson based model that we used as a building block for our approach.
- **GeoSPF:** This is our approach. It uses equation 4.10 presented in the previous section.
- **ALGeoSPF:** Finally this corresponds to our augmented local-global GeoSPF.

Figure A.4 shows the overall performances of all the above baseline methods. On Figure A.4a the Recall@5 and Recall@10 are reported for the Foursquare dataset. The same metrics are reported on Figures A.4b (resp. A.4c) for Gowalla@Paris (resp. Gowalla) dataset. Finally, Figure A.4d reports the NDCG@5 metric for the three datasets.

As a first observation, on the first three datasets (Foursquare, Gowalla@Paris and Gowalla) we notice that our approach (GeoSPF) significantly outperforms all the other ones. As expected, NMF and PMF do not yield a good quality since they were not designed to cope with implicit feedback datasets. This is consistent with the results in [Liu and Xiong 2013]. Although SLIM is known to perform well on sparse datasets, it fails to achieve a good quality in our context because it assumes explicit feedback (instead of implicit one).

Unfortunately the complexity of WRMF is $O(f^2n)$. This complexity makes it practically useless on large datasets: the WRMF computation time is prohibitively long beyond 5000 POIs. This is why we only get the recall@N for Gowalla@Paris, but not the two other datasets where the number of POIs is too high.

This is due to the fact that the complexity of WRMF depends highly on the number of POIs, and according to table 5.1, Gowalla@Paris has less POIs than the others. Among all the state-of-the-art competitors, PF achieves the best quality. Thus, we focus our analysis on comparing PF vs. GeoSPF. As a major result, the relative benefit of GeoSPF on all the datasets is around 200%. This impressive gain makes GeoSPF suitable for POI recommendation over wide geographical areas. It confirms that exploiting restricted contextual information (only GPS and check-in date) through a combined geographical/social solution yields a high quality for POI recommendation.

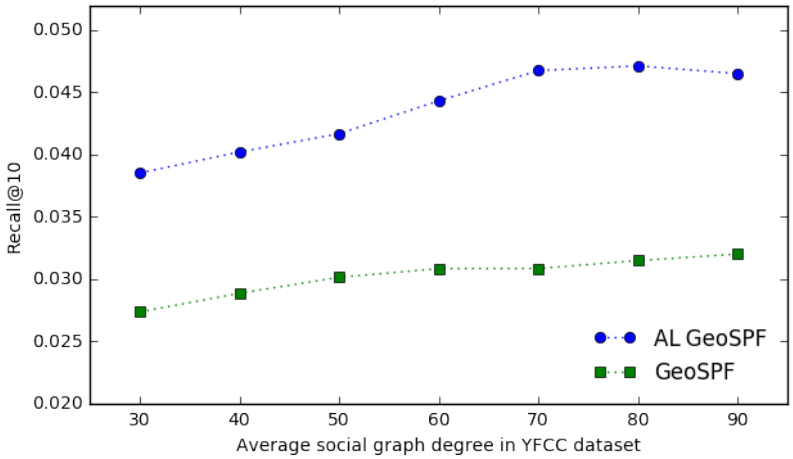
Our last experiment aims at assessing the benefits of our geographical clustering approach for user-class aware recommendation. Figure A.5 reports the recommendation quality (recall) ALGeoSPF applied on the YFCC dataset considering the urban users isolated from the globetrotters. More precisely, Figure A.5a reports the recall@10 of GeoSPF and ALGeoSPF for different average sizes of the implicit social network. For every size of the network, we observe that ALGeoSPF always improves significantly the recall of 50%. Figure A.5b reports the recall@5 and recall@10 of all the competitors as well as ALGeoSPF, on the YFCC dataset (using a fixed social network size of 80). We can see that ALGeoSPF outperforms the other methods, although BPR yields close quality. We observe also that globally the recall measures of tested models for the YFCC dataset are much lower than other datasets: this is due to the low density because of the geographical area covered is large (see Section 5.1).

To sum up, the experiments demonstrate the interest of integrating user classes based on mobility into a social POI recommendation task. We observe that

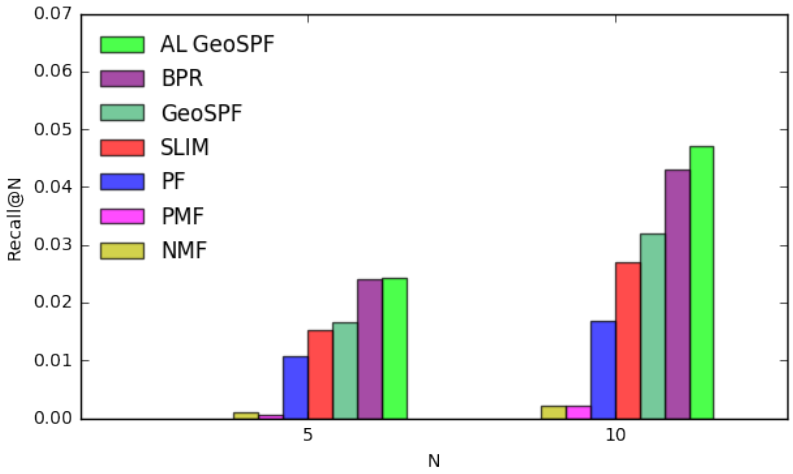
5.6 Conclusion

In this chapter we have proposed a new approach for the POIs recommendation task in LBSNs called ALGeoSPF. The specificity of this approach is that it can handle large volume of check-ins. The main goal of ALGeoSPF is to build a model which does not suffer from the low density of large-scale geographical datasets, and which takes into account of the specific users' mobility behaviors. Based on the new

concept of *superPOI* and our clustering algorithm that we have introduced in this work, our approach succeeded **(i)** to build efficiently an implicit scalable factorization model and **(ii)** to capture the user's mobility preferences into a hierarchical structure and finally **(iii)** to present significant better results than baselines on large-scale datasets. We have demonstrated with extensive experiments that AL-GeoSPF significantly outperforms all the alternative approaches in terms of *recall* and *NDCG* on a large dataset. To the best of our knowledge, we are among the first to test a POIs recommendation approach on the YFCC dataset for our experiments.



(a) Recall@10 on YFCC



(b) Recall@5 and Recall@10

Figure 5.3: Performance comparison of ALGeoSPF *wrt.* state-of-the-art approaches for 2 levels of the YFCC dataset. We plot on figure A.5a the recall@10 results of GeoSPF and ALGeoSPF for different size of the average social graph degree. Figure A.5b presents the results of AMGeoSPF in terms of recall@5 and recall@10.

Chapter 6

Conclusion

In this chapter we propose first a summary of the contributions and achievements reached throughout this thesis in Section 6.1. Then we describe a brief outlook on future works and research perspectives in Section 6.2.

6.1 Summary

In this PhD thesis we have investigated the problem of POIs recommendation. Specifically we have explored three major lines: (i) how to integrate efficiently geographical and temporal influences into a matrix factorization model, (ii) how to deal with the implicit feedback problem, and (iii) how to build recommendation models for large real-world volumes of data.

Geographical Matrix Factorization. First we have proposed a new matrix factorization approach that improves the final recommendation quality. Matrix factorization models became widely used after the Netflix Prize [Bennett, Lanning, and Netflix 2007]. However existing approaches proposed for POI recommendation fail to manage the geographical and temporal influences in an efficient way. To handle these influences we have proposed a method that explores the distribution of the geographical and temporal influences between POI. These influences are expressed with geographical latent features. Then these features are fused by linear combination into a single model. This model is named GeoMF-TD (for *Geographical Matrix Factorization with Temporal Dependencies* and has resulted in a publication Griesner, Abdessalem, and Naacke 2015 in the ACM Conference on Recommender Systems[1] 2015.

[1] http://recsys.acm.org/

Implicit Feedback Problem. Then we have proposed a new approach to tackle the implicit feedback problem. This problem corresponds to the difficulty to generate personalized recommendations without any explicit feedback regarding the user's past choices of POIs. Only positive labels are available, but we cannot distinguish among the unvisited POIs the unknown ones from the "disliked" ones. In other words, this problem corresponds to the question to which extent a check-in of a user means that this user has appreciated the corresponding POI. To tackle this problem we proposed to extract the social influences of users between them. Then we integrate these implicit social influences into a factorization model based on Poisson factorization. This approach corresponds to GeoSPF. This work has resulted in a publication Griesner, Abdesssalem, and Naacke 2017 in EGC 2017 international Conference[2].

Large-Scale POI Recommendation. Finally we have investigated algorithms and methods to deal efficiently with large-scale POI recommendation problems. Because volumes of data gathered everyday on LBSN are nowadays massive, traditional recommendation techniques generally cannot scale. As a result most of existing works on POI recommendation limit their experiments to small datasets, which are far from real-world use cases. This is why we have explored solutions to apply efficient recommendation methods to large datasets. To this end we have exploited a geographical clustering based method that classifies the different user mobility patterns that exist in the data. Our primary objective was to apply GeoSPF to a large-scale dataset released by Yahoo[3] containing 100 millions check-ins. This work has resulted in two publications: one Griesner, Abdesssalem, Naacke, and Dosne 2018 in EGC 2018 international conference[4] and another in BDA 2017 [5]

6.2 Outlook

In this section we propose a brief outlook of future interesting lines of research in the domain of POI recommendation. As we did in Section 6.1 we divide this outlook within the lines of (i) cascade models, (ii) implicit feedback problem, and (iii) large-scale recommendations.

Cascade Models. An interesting line of future research could be to integrate the influence of each recommended POI to the user in a sequential order. Indeed

[2] http://egc2017.imag.fr/
[3] http://yfcc100m.appspot.com/
[4] https://egc18.sciencesconf.org/
[5] https://project.inria.fr/bda2017/

the user decision process to choose which POI is worth to visit is a sequential process. That is to say the user considers the POIs of the recommended top-K list $\mathcal{L} = (p_1, ..., p_K)$ in a sequential order from the first one p_1 to the last one p_K. A convenient model to formulate this specific user behavior is the cascade model proposed some years ago by Craswell, Zoeter, Taylor, and Ramsey 2008. This model depends on attraction probabilities $\omega \in [0, 1]^K$ that reflects the attraction influence of the POI. Consequently the POI p_j has a probability $\omega(p_j)$ to have a positive influence on the user. The main assumption of cascade model is that each considered POI has an influence on the user independently of the other POIs. Once the user has selected a POI p_j, he stops to examine the remaining POIs. Otherwise he continues examining the next POI p_{j+1}. The resulting cascade model optimizes the probability to find at least an attractive POI in the recommended list. Recently Zong et al. 2016 have proposed an efficient cascade-based bandits model for large-scale recommendation use cases.

Implicit Feedback Problem. The available check-ins data are said *positive only*. As a result there is an important difficulty for the recommender systems to distinguish relevant POIs from uninteresting ones. The GeoSPF model that we proposed generates implicit social influences between users. These influences depend on different similarity metrics proposed in Section 4.4. The problem is that the complexity to compute all the similarities has an important cost. A futur line of research to consider to avoid this complexity cost would be to apply directly existing efficient PU-learning approaches such as investigated by Niu et al. 2016; Elkan and Noto 2008 into an augmented recommendation model.

Large-Scale POI Recommendation. Our solution ALGeoSPF to deal with large volumes of data is based on a geographical clustering based approach. Different lines of research could be investigated to improve ALGeoSPF. First regarding the personalized user class definitions. We have constraint the number of mobility classes to a limited number. However we plan to explore solutions with more classes in order to integrate different mobility patterns. These classes could be based on continuous mobility densities. Another line of investigation would be to improve the processing of the

Appendix A

Résumé en français

Summary. *In this appendix we propose a detailed summary in French of the main contributions achieved during this PhD. It is for the most part a translation of the Introduction chapter with a large overview of the main components of this thesis. Note that this summary does not include any additional scientific content relative to the rest of the thesis and so may be skipped.*

Résumé. *Dans cette annexe nous proposons un résumé en français des principales contributions auxquelles ma thèse a abouti. Il s'agit essentiellement d'une traduction du chapitre d'introduction et d'une présentation synthétique des grandes étapes qui ont structuré ma réflexion. Ce résumé n'inclut aucun contenu scientifique supplémentaire par rapport à la thèse en elle-même et n'est donc pas nécessaire en première lecture.*

A.1 Introduction

Le développement du Web 2.0 ces dernières années a favorisé l'émergence d'un grand nombre de réseaux sociaux basés sur la localisation (noté **LBSN** dans la suite de ce chapitre) tels que Twitter, Facebook, Google+, etc., qui ont profondément modifié le regard que nous portons sur notre environnement et la manière dont nous interagissons avec lui. Par définition un LBSN appartient à une catégorie particulière de *réseau social en ligne* [Youssef 2014] dont le contenu est directement associé à notre monde physique. En effet les LBSN proposent à leurs utilisateurs de nombreux services qui s'appuient directement sur leurs localisations. Ces services sont particulièrement utiles à l'utilisateur qui se retrouve dans un environnement nouveau ou inconnu. Pour proposer ces nouveaux services, tous ces réseaux ont développé de plus en plus de technologies et de supports pour aider leurs utilisateurs, jusqu'à devenir aujourd'hui l'une des sources d'informations les plus riches sur les préférences, les habitudes et les activités des utilisateurs dans le monde [Chorley, Whitaker, and Allen 2015].

La quantité d'informations personnelles et de ressources partagées sur ces LBSN a augmenté de manière exponentielle ces dernières années [Cui, Hero, Luo, and Moura 2016]. Par exemple sur le réseau social Flickr[1] plus de 110 millions d'utilisateurs produisent plus d'un million d'images par jour. En raison de cette *surcharge d'informations* [Toffler 1970] il est devenu de plus en plus difficile pour les utilisateurs de trouver ce qu'ils recherchent dans leur environnement. Pour résoudre ce problème de surcharge d'informations, les *systèmes de recommandation* (noté **RS** dans ce qui suit) sont devenus incontournables. Le but principal des RS [Adomavicius and Tuzhilin 2005] est de fournir une assistance personnalisée aux utilisateurs qui ont besoin d'aide pour rechercher, classer ou filtrer la grande quantité d'informations disponibles sur les LBSN.

Ces systèmes sont maintenant largement adoptés par la plupart des plates-formes commerciales en ligne dans des contextes variés allant des livres (Amazon[2]) aux films (Netflix[3]), la musique (Spotify [4]) ou aux points d'intérêt (**POIs**) avec des applications telles que Foursquare [5]. Motivés par ces enjeux commerciaux, de nombreux problèmes de recommandation ont été étudiés ces dernières années dans un grand nombre de domaines allant de la recommandation musicale [Cheng, Shen,

[1] www.flickr.com
[2] https://www.amazon.com/
[3] https://www.netflix.com
[4] https://www.spotify.com
[5] https://fr.foursquare.com/

and Mei 2014] à la recommandation d'actualités [Hsieh et al. 2016] ou encore de la recommandation de films [Gantner, Rendle, and Schmidt-Thieme 2010]. On aurait donc pu s'attendre à ce qu'il existât aujourd'hui différentes approches efficaces permettant de répondre à la plupart des problèmes de recommandation. Or le problème de la recommandation de points d'intérêt implique plusieurs problèmes spécifiques qui le distinguent des tâches de recommandation traditionnelles. Dans cette thèse nous explorons les difficultés spécifiques à la recommandation de points d'intérêt (noté **POI** par la suite) et nous proposons de nouvelles solutions pour les contourner.

Notons que le problème de la recommandation de la recommandation de POI a soulevé un intérêt croissant dans le monde universitaire. En effet un grand nombre de travaux ont été proposés pour résoudre ce problème ces dernières années, notamment dans des conférences internationales en informatique telles que *ACM RecSys* [Baral and Li 2016], *KDD* [Li, Ge, Hong, and Zhu 2016], *WWW* [Ying, Chen, Xiong, and Wu 2016], *CIKM* [Xie et al. 2016], *IJCAI* [Jing, Xin, and Lejian 2017], *SIGIR* [Yuan, Cong, Ma, et al. 2013] et bien d'autres.

A.2 Axes de recherche

Les objectifs principaux de cette thèse sont doubles. Il s'agit tout d'abord d'étudier et de proposer de nouvelles méthodes de factorisation de matrices. Les méthodes de factorisation de matrices ne sont pas prévues pour intégrer les emplacements géographiques des produits recommandés dans leurs modèles, ce qui entraîne souvent une mauvaise qualité de recommandation. Notre deuxième objectif est de résoudre le problème du passage à l'échelle de telles méthodes. En effet la plupart des algorithmes de recommandation souffrent généralement d'un volume trop important de données. Ce volume important de données implique un grand nombre de points d'intérêt, ce qui rend les modèles de recommandation inefficaces car incapables de distinguer lesquels sont les plus pertinents. Pour atteindre ces objectivs, nous avons été amenés à explorer les axes de recherche suivants :

Axe n°1 : Étudier les méthodes existantes dans le domaine de la recommandation de POI. Il existe dans l'état-de-l'art un grand nombre de méthodes et de modèles pour la recommandation de points d'intérêt. Cependant une compréhension claire des avantages et des inconvénients effectifs entre ces modèles fait toujours défaut. Par conséquent nous proposons un panorama complet des techniques et des approches les plus efficaces (*c.f.* Chapitre 2).

Axe n°2 : Explorer et améliorer les approches de factorisation de matrices. Depuis le *Netflix Prize* [Bennett, Lanning, and Netflix 2007] nous savons que les approches de factorisation sont les plus efficaces parmi les méthodes de filtrage collaboratif. Nous avons donc étudié les méthodes de factorisation afin de les associer aux spécificités de la recommandation de POI. Il s'agit en particulier d'intégrer les influences géographiques et temporelles.

Axe n°3 : Étudier le cadre probabiliste pour les modèles de factorisation. Les règles et hypothèses probabilistes permettent de mettre en oeuvre des méthodes plus flexibles et plus sophistiquées. Nous proposons donc de nouvelles approches probabilistes pour mieux prendre en compte les problèmes de densité et d'informations contextuelles.

Axe n°4 : Améliorer le passage à l'échelle des approches de factorisation. La plupart des techniques de recommandation de POI ne parviennent pas à s'appliquer sur de gros volumes d'utilisateurs et/ou de POI. En conséquence de quoi la plupart des jeux de données expérimentaux utilisés pour les tests dans la littérature sont de plusieurs ordres de grandeur plus petits que les jeux de données du monde réel. C'est la raison pour laquelle nous cherchons à explorer les solutions possibles pour remédier à ce problème.

Axe n°5 : Explorer les schémas de mobilité des utilisateurs. Chaque LBSN représente une source d'informations riches et précises sur les habitudes de ses utilisateurs. Cette source d'informations peut être exploitée pour améliorer notre compréhension des schémas de mobilité géographique des utilisateurs. Plus précisément nous cherchons à à exploiter l'observation de différentes échelles de mobilité : certains utilisateurs ont tendance à visiter des POI dans le monde entier, sur de longues distances, tandis que d'autres concentrent leurs visites dans des zones locales.

A.3 Contributions

Les travaux accomplis tout au long de cette thèse ont abouti à plusieurs innovations dans le domaine de la recommandation de points d'intérêt. Cette section décrit brièvement les principales contributions issues de ce travail.

Contribution n°1 : Un modèle de factorisation de matrices géographique avec dépendances temporelles: GeoMF-TD. Nous avons proposé

un modèle de factorisation qui prend en compte les distributions spatio-temporelles des checkins dans les données. GeoMF-TD divise la région cible donnée dans une grille de cellules. Cette grille est ensuite exploitée pour modéliser les influences latentes géographiques et temporelles des POI et l'activité des utilisateurs à travers les cellules de la grille. Ces influences latentes sont ensuite combinées linéairement avec des vecteurs latents pour calculer le score de recommandation. Ce travail a été publié dans [Griesner, Abdessalem, and Naacke 2015].

Contribution n°2 : GeoSPF, une solution au problème de feedback implicite basée sur un modèle de factorisation de Poisson. La factorisation de Poisson a été exploitée avec succès pour divers problèmes de recommandation. C'est pourquoi en partant d'un modèle de factorisation de Poisson nous avons essayé de renforcer l'influence des informations contextuelles en exploitant un graphe social implicite. Ce graphe social est basé sur les préférences géographiques des utilisateurs. Cet axe de recherche a fait l'objet d'une publication dans [Griesner, Abdesssalem, and Naacke 2017].

Contribution n°3 : Un nouveau modèle pour les schémas de mobilité géographique des utilisateurs. Nous observons généralement dans les données des LBSN des profils d'utilisateurs différents : certains utilisateurs ont tendance à faire des trajets touristiques sur de longues distances alors que d'autres utilisateurs ne font que des visites limitées à une zone géographiques locale. Nous avons exploité cette observation pour faire face au problème du passage à l'échelle en appliquant une méthode de clustering géographique hiérarchisée.

Contribution n°4 : ALGeoSPF, une extension de GeoSPF qui passe à l'échelle. Sur la base des contributions précédentes, nous avons proposé une nouvelle approche permettant de générer des recommandations personnalisées de POI sur des jeux de données à grande échelle. Ce travail a abouti à notre modèle ALGeoSPF qui a été présenté dans [Griesner, Abdesssalem, Naacke, and Dosne 2018].

A.4 GeoMF-TD : Un modèle de factorisation de matrices pour la recommandation de POI

Nous commençons par présenter GeoMF-TD qui est une première méthode efficace de recommandation de POI. GeoMF-TD est un modèle de factorisation de matrices

qui repose sur un modèle spatio-temporel qui tient compte de la distribution des POI dans l'espace. Nous présentons GeoMF dans la partie A.4.1 et GeoMF-TD dans la partie A.4.2.

A.4.1 Factorisation de matrices géographique

La recommandation suppose l'existence d'un ensemble d'utilisateurs et d'un ensemble de POI que l'on cherche à proposer aux utilisateurs. Ainsi soit $\mathbf{u} = \{u_1, u_2, ..., u_m\} \subset U^m$ un ensemble d'utilisateurs et soit $\mathbf{p} = \{p_1, p_2, ..., p_n\} \subset P^n$ un ensemble de POI. De plus soit $\mathbf{C} \in \mathbb{R}^{m \times n}$ la matrice utilisateurs-POI initiale contenant les m utilisateurs et les n POI. La valeur $c_{u,j}$ de $\mathbf{C}$ correspond à la fréquence de visite de l'utilisateur u au POI i.

A.4.1.1 Factorisation pondérée de matrices

L'objectif de la factorisation de matrices est d'approximer la matrice $\mathbf{C}$ par le produit de deux matrices $\mathbf{P} \in \mathbb{R}^{m \times k}$, et $\mathbf{Q} \in \mathbb{R}^{n \times k}$ dites de facteurs latents de dimension $k \ll \min(m, n)$ en résolvant le problème d'optimisation suivant :

$$\min_{\mathbf{P},\mathbf{Q}} \left\|\mathbf{C} - \mathbf{P}\mathbf{Q}^T\right\|^2 + \gamma\left(\|\mathbf{P}\|^2 + \|\mathbf{Q}\|^2\right) \tag{A.1}$$

avec γ qui est un paramètre positif qui permet d'éviter le sur-apprentissage en contrôlant la capacité de $\mathbf{P}$ et $\mathbf{Q}$. Ainsi il est possible d'approximer la valeur manquante $\widetilde{c_{u,j}}$ de $\mathbf{C}$ en calculant le produit scalaire entre les vecteurs de facteurs latents correspondants :

$$\widetilde{c_{u,j}} = \mathbf{P}_u \mathbf{Q}_j^T. \tag{A.2}$$

Ce modèle suppose que la matrice $\mathbf{C}$ contient directement des valeurs de préférence pour chaque couple utilisateur-poi. Or les jeux de données issus des LBSN ne fournissent qu'une estimation de *confiance* mais ne donne pas d'information sur les *préférences* des utilisateurs. Cette propriété fait référence aux problèmes de recommandation dits avec *feedback implicite*. Plus précisément [Hu, Koren, and Volinsky 2008] ont prouvé que la factorisation de matrices pondérée (notée WMF par la suite) donne les meilleurs résultats avec des jeux de données de feedback implicite. En effet la factorisation de matrices pondérée prend en compte l'asymétrie qui existe entre *confiance* et *préférence* et crée deux nouvelles variables pour formaliser cette asymétrie. Ensuite la WMF transforme le problème de l'équation (A.1) en un nouveau problème d'optimisation comme suit :

$$\min_{\mathbf{P},\mathbf{Q}} \left\|\mathbf{W} \odot \left(\mathbf{R} - \mathbf{P}\mathbf{Q}^T\right)\right\|^2 + \gamma\left(\|\mathbf{P}\|^2 + \|\mathbf{Q}\|^2\right) \tag{A.3}$$

où $\odot$ est l'opération de multiplication de matrices élément par élément (i.e. le produit d'Hadamard) et où la seule différence avec l'équation (A.1) est la présence de la matrice $\mathbf{W}$ et de la matrice binaire $\mathbf{R}$ dont chaque valeur $r_{u,i}$ indique si l'utilisateur u a visité le POI i. L'idée de WMF est de supposer une confiance minimale pour tous les POI, visités ou non. Cette confiance minimale est encodée dans la matrice $\mathbf{W}$ définie tel que :

$$w_{u,i} = \begin{cases} 1 + \alpha(c_{u,i}) & \text{if } c_{u,i} > 0 \\ 1 & \text{otherwise} \end{cases} \tag{A.4}$$

où $\alpha()$ est une fonction monotone strictement croissante.

A.4.1.2 Modéle de l'influence géographique

Récemment les travaux de [Lian et al. 2014] ont proposé une factorisation de matrices géographique (GeoMF) pour intégrer cette influence directement dans le modèle de factorisation de WMF décrit à la section précédente. L'idée des auteurs était de distinguer pour chaque utilisateur les points d'intérêt non visités mais toutefois intéressants parmi les POI négatifs. L'intuition est que si un utilisateur visite un POI sans visiter les autres points d'intérêt proches, ces POI ignorés risquent de ne pas être suffisamment intéressants pour l'utilisateur. Par conséquent ces POI deviennent négatifs pour le modèle.

Ainsi cette approche divise l'espace en L cellules $\mathbb{L} = \{g_1, g_2, ..., g_L\}$ et calcule pour chaque POI son aire d'influence sur chacune des cellules de la grille en se basant sur une distribution normale des distances entre les POI. Cette approche revient à augmenter les facteurs latents de la factorisation de matrices traditionnelle $\mathbf{P}$ et $\mathbf{Q}$ avec deux matrices de facteurs latents géographiques $\mathbf{X} \in \mathbb{R}^{m \times L}$ et $\mathbf{Y} \in \mathbb{R}^{n \times L}$. Avec ces nouveaux facteurs l'équation A.3 est modifiée ainsi :

$$\min_{\mathbf{P},\mathbf{Q},\mathbf{X}} \left\| \mathbf{W} \odot \left(\mathbf{R} - \mathbf{P}\mathbf{Q}^T - \mathbf{X}\mathbf{Y}^T \right) \right\|^2 + \gamma \left(\|\mathbf{P}\|^2 + \|\mathbf{Q}\|^2 \right) + \lambda \|\mathbf{X}\|^2 \tag{A.5}$$

où λ controle la faible densité sur les schémas de mobilité des utilisateurs à travers la grille. Une rangée $\mathbf{x}_u$ de $\mathbf{X}$ correspond aux aires d'activité de l'utilisateur u *i.e.* la distribution des fréquences de visite dans chaque cellule de la grille g_l de la carte, tandis qu'une colonne $\mathbf{y}_i$ de $\mathbf{Y}$ correspond à l'aire d'influence du POI i. Plus précisément nous calculons pour chaque POI i et pour chaque cellule g_l l'influence gaussienne que i a sur la cellule g_l :

$$\mathbf{y}_i^l = \frac{1}{\sigma} K\left(\frac{d(i,l)}{\sigma}\right) \tag{A.6}$$

où $K()$ est la distribution normale standard et σ la déviation standard. Avec the modèle augmenté géographique nous obtenons le score de recommandation pour l'utilisateur u et le POI i ainsi :

$$\widetilde{c_{u,i}} = \mathbf{P}_u \mathbf{Q}_i^T + \mathbf{X}_u \mathbf{Y}_i^T \tag{A.7}$$

A.4.2 GeoMF avec dépendances temporelles

Le modèle GeoMF décrit ci-dessus suppose que l'espace est un espace homogène isotrope sans contraintes physiques. Il suppose en particulier que la zone d'influence de chaque POI suive une distribution normale fixée à l'avance et basée uniquement sur les distances dans l'espace. Cependant les zones d'influence de deux POI distincts peuvent être très différentes dans la réalité si l'on considère des paramètres autres que les distances. Notamment l'effet temporel dans les séquences de visites de POI jouent également un rôle important comme l'ont démontré [Gao, Tang, Hu, and Liu 2013]. En particulier ces effets peuvent refléter le fait qu'un POI j peut être dans la zone d'influence d'un autre POI i sans être réellement négatif.

Pour tenir compte de ce problème nous proposons de modifier les valeurs de la zone d'influence de chaque POI i à travers la grille $g_{l\epsilon\mathbb{N}^L}$ pour prendre en compte le temps nécessaire à un utilisateur pour aller du POI i aux autres POI situés dans la même cellule g_l. Ainsi pour chaque POI i nous calculons le temps moyen que chaque utilisateur a mis pour atteindre j (j étant dans g_l) à partir de i. Nous calculons ce temps moyen pour chaque utilisateur qui ont au moins un checkin à i et un autre checkin plus récent à j dans g_l. Puis nous faisons la moyenne pour chaque utilisateur des valeurs calculées pour chaque cellule pour obtenir un score d'influence pour le POI i. Soit $t_i^{g_l}$ le temps moyen calculé entre i et les POI qui se trouvent dans la même cellule g_l.

Ainsi nous introduisons des coefficients temporels pour diminuer le caractère négatif potentiel des POI négatifs lorsqu'un utilisateur ne les a pas visités pendant un certain temps. C'est pourquoi ces coefficients laissent inchangée la valeur de la zone d'influence lorsque cette valeur est faible. Ces coefficients sont le résultat direct de la fusion de l'influence géographique gaussienne sur l'espace et des dépendances temporelles existant dans l'ensemble des données. Nous définissons des coefficients temporels $\theta_l(t_i^{g_l})$ comme suit :

$$\theta_l(t_i^{g_l}) = \begin{cases} \alpha * \mathbf{y}_i^l & \text{if } t_i^{g_l} > \sigma^i \text{ and } \mathbf{y}_i^l < 0.1 \\ \mathbf{y}_i^l & \text{otherwise} \end{cases} \tag{A.8}$$

où σ^i correspond à la déviation standard des intervalles de temps pour le POI i et $\mathbf{y}_i^l$ qui est obtenu à partie de l'équation A.6. Puis nous intégrons ces coefficients avec le vecteur d'influence $\mathbf{y}_i$ du POI i comme suit :

$$\mathbf{y}_i = [\theta_1(t_i^{g_1}), ..., \theta_L(t_i^{g_L})] \tag{A.9}$$

A.4.3 Résultats expérimentaux

Dans nos expériences nous avons comparé la précision de GeoMF-TD et de GeoMF. Nous avons évalué les algorithmes sur les checkins collectés à partir du LBSN Gowalla [6] et accessibles au public grâce à [Cho, Myers, and Leskovec 2011]. Gowalla était un célèbre LBSN fermé en 2012. Le jeu de données Gowalla a déjà été utilisé dans plusieurs travaux sur la recommandation de points d'intérêt, tels que ceux proposés par [Cheng, Yang, King, and Lyu 2012; Cho, Myers, and Leskovec 2011; Zhang and Chow 2013]. Le tableau A.1 présente les principales statistiques concernant cet ensemble de données.

Afin d'augmenter la densité de la matrice dans le jeu de données, nous ne conservons que les utilisateurs qui comptent au moins 50 checkins. Pour des raisons pratiques nous n'utilisons que des checkins localisés en France. Suite à ce filtrage il reste 161 utilisateurs, 7697 POI distincts pour 12418 checkins distincts, ce qui est très peu, mais suffisant pour une évaluation initiale.

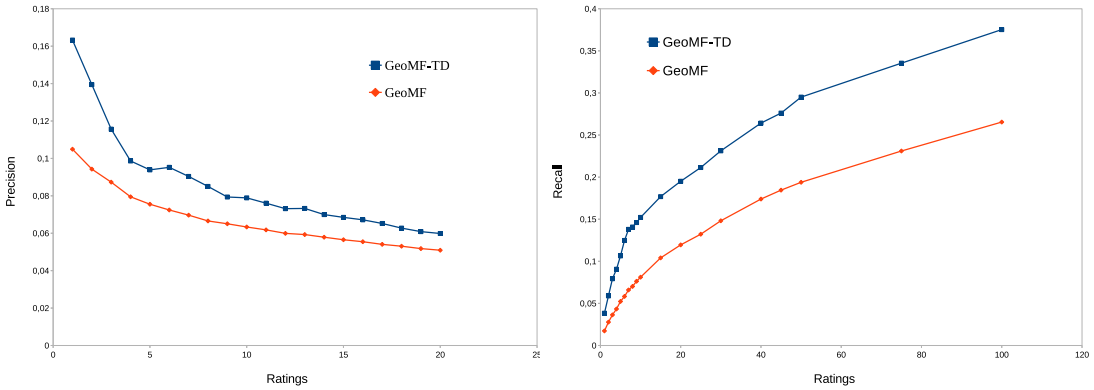
Dans une perspective de comparaison nous avons implémenté le modèle GeoMF en utilisant la librairie Java LibRec[7]. La figure A.1a ainsi que la figure A.1b les

[6] http://snap.stanford.edu/data/loc-gowalla.html
[7] http://www.librec.net/

nombre d'utilisateurs	196,591
nombre de checkins	6,442,890
nombre de liens sociaux	950,327
densité de la matrice	2.9×10^{-5}
nombre moyen de POI par utilisateur	37.18
nombre moyen de checkins par POI	3.11

Table A.1: Statistiques sur le jeu de données issu du LBSN Gowalla utilisé dans les expériences.

résultats comparatifs respectivement de la *prcision@N* et du *rappel@N* entre GeoMF et GeoMF-TD avec N allant de 1 à 20 pour la précision, et de 1 à 100 pour le rappel. Comme prévu les coefficients temporels introduits ont permis de prendre en compte les dépendances temporelles existant dans les données et d'améliorer ainsi la précision globale. Les résultats présentés sur les figures A.1a et A.1b indiquent un bénéfice moyen de 60% pour le rappel et de 20% pour la précision. Cette comparaison de performance globale n'intègre pas l'étude de l'influence du paramètre de seuil mais donne des résultats prometteurs.



(a) Résultats de précision@N entre GeoMF et GeoMF-TD

(b) Résultats de recall@N entre GeoMF et GeoMF-TD

Figure A.1: Résultats comparatifs entre GeoMF et GeoMF-TD

A.4.4 Conclusions

Nous avons présenté GeoMF-TD, un nouveau modèle de factorisation de matrices pour le problème de la recommandation de POI dans les LBSN. Notre objectif était d'essayer de tirer parti d'un modèle de factorisation géographique en prenant en compte les influences temporelles des séquences de visite de POI. C'est dans cette perspective que nous avons proposé l'algorithme GeoMF-TD. Notre évaluation expérimentale montre que GeoMF-TD permet une qualité significativement supérieure à GeoMF en termes de précision et rappel.

A.5 GeoSPF : influences sociales implicites

Nous présentons à présent une méthode de recommandation de POI qui offre une solution efficace au problème dit du feedback implicite. Nous décrivons ce problème dans la partie A.5.1.

A.5.1 Factorisation de Poisson et feedback implicite

Étant donné que les utilisateurs ne visitent qu'un nombre limité de points d'intérêt, la matrice $\mathbf{X}$ est généralement très peu dense. La factorisation de Poisson, notée PF dans la suite, est une solution possible à ce problème de densité. Il s'agit d'une approche générative à base de facteurs latents probabilistes qui exploite une hypothèse de la loi de Poisson pour modéliser les observations. PF est basé sur le modèle de sujet GaP de [Canny 2004]. Gopalan *et al.* ont démontré que ce modèle est adapté aux données de type comportemental et peu dense, comme le proposent [Charlin, Ranganath, McInerney, and Blei 2015]. De plus l'inférence à posteriori de la PF est beaucoup plus rapide que les autres approches car les probabilités inférées dépendent uniquement des valeurs observées. Si nous notons $x_{i,j}$ le nombre de fois où l'utilisateur i a visité le POI j, le modèle PF suppose que $x_{i,j}$ provient d'une distribution de Poisson, paramétrée par le produit scalaire des préférences de l'utilisateur et des caractéristiques du POI. Ainsi la PF estime pour chaque utilisateur i et chaque élément j le modèle suivant:

$$y_{i,j} \sim \text{Poisson}(\mathbf{u}_i^T \cdot \mathbf{v}_j) \tag{A.10}$$

Une fois la distribution postérieure inférée, la PF classe les POI candidats de chaque utilisateur en fonction de leur score de recommandation, c'est-à-dire en fonction du résultat postérieur attendu, comme suit:

$$\hat{r}_{i,j} = \mathrm{E}\left[\mathbf{u}_i^T \cdot \mathbf{v}_j | y\right] \tag{A.11}$$

où $\mathbf{u}_i$ et $\mathbf{v}_j$ sont les k-vecteurs de préférences latentes de l'utilisateur et les caractéristiques latentes des POI, respectivement. Les préférences latentes de l'utilisateur et les attributs latents des points d'intérêt sont ici les variables cachées du modèle.

A.5.2 Modèle d'influence sociale

Dans cette section nous présentons GeoSPF, c'est-à-dire notre méthode pour extraire les influences sociales implicites issues des schémas de mobilité des utilisateurs. Nous présentons notre graphe d'accessibilité (*AGRA* pour Accessibility GRAph) que nous utilisons pour modéliser les influences géographiques. Ce graphe est utilisé pour construire le réseau social implicite que nous exploitons ensuite pour la recommandation. Une liste des notations utilisées dans ce qui suit est proposée dans le tableau 1.2.

A.5.2.1 Accessibilité géographique

L'idée de l'accessibilité est de modéliser la probabilité qu'un utilisateur passe à un POI p_{j+1} après avoir visité le POI p_j. Pour ce faire nous appliquons des modèles de Markov de premier ordre, qui ont été utilisés avec succès pour le traitement de données séquentielles. Une transition est observée dans l'itinéraire d'un utilisateur u si elle existe dans le jeu de données à deux reprises successives $< u, p_i, t_1 >$ et $< u, p_j, t2 >$, effectuées dans deux POI différents p_i et p_j à deux timestamps t_1 et t_2, tels que $t_1 < t_2$ et aucun autre checkin intermédiaire $< u, p_k, t' >$ $(t_1 < t' < t_2)$ n'existant dans les données. Nous noterons cette transition ainsi : $p_i \rightarrow p_j$ par la suite. Ainsi, pour un utilisateur donné, la probabilité de visiter p_{j+1} sera déduite à partir de celle du dernier POI visité. Ainsi nous avons:

$$P(p_{j+1}|p_j, p_{j-1}, ..., p_1) = P(p_{j+1}|p_j) \tag{A.12}$$

où nous définissons $P(p_{j+1}|p_j)$ comme la probabilité de transition $\mathcal{T}_{j,j+1}$ de p_j vers p_{j+1}. Nous pouvons calculer cette probabilité en utilisant l'estimation empirique du maximum de vraisemblance comme suit:

$$\mathcal{T}_{j,j+1} = P(p_{j+1}|p_j) = \frac{N(p_j, p_{j+1})}{N(p_j)} \tag{A.13}$$

où $N(p_j, p_{j+1})$ est le nombre d'utilisateurs ayant la séquence $p_j \rightarrow p_{j+1}$ dans leur profil, et $N(p_j)$ est le nombre d'utilisateurs ayant visité p_j. Comme nous avons $N(p_j, p_{j+1}) \leqslant N(p_j)$ nous savons que $\mathcal{T}_{j,j+1}$ est bornée : $\mathcal{T}_{j,j+1} \in [0, 1]$. Notons que pour calculer cette probabilité les checkins doivent suivre l'ordre chronologique de leur apparition. Ensuite on peut combiner cette probabilité avec l'information géographique pour estimer l'accessibilité $\mathcal{A}_{j,j+1}$ entre les POI p_j et p_{j+1}. Nous définissons cette accessibilité comme suit :

$$\mathcal{A}_{j,j+1} = \frac{1}{0.5 + d(p_j, p_{j+1})} \cdot \mathcal{T}_{j,j+1} \tag{A.14}$$

où $\mathcal{T}_{j,j+1}$ correspond à l'équation A.13 et $d(p_j, p_{j+1})$ est la distance euclidienne entre les POI p_j et p_{j+1}. Si p_{j+1} est loin de p_j, alors l'accessibilité sera faible. Cependant, si de nombreuses transitions ont été observées de p_j à p_{j+1}, l'accessibilité augmentera en conséquence. L'équation A.14 s'inspire des poids géographiques utilisés par [Liu and Xiong 2013].

A.5.2.2 Graphe d'accessibilité

Notre approche construit un réseau social implicite (ISN) basé sur la similarité entre l'historique de visites des utilisateurs et leurs transitions dans le graphe AGRA.

Nous proposons ci-dessous les quatre mesures de similarité possibles, choisies pour leur évolutivité et la qualité de leurs résultats :

- **Adamic/Adar**: Cette mesure accorde une grande importance aux transitions rares (c.-à-d. Avec une faible accessibilité). Intuitivement plus deux utilisateurs partagent des POI impliqués dans des transitions rares plus ils sont supposés être proches l'un de l'autre. Formellement la similarité Adamic/Adar $S_{\mathrm{AA}}(u_1, u_2)$ est définie ainsi :

$$S_{\mathrm{AA}}(u_1, u_2) = \sum_{v \in \mathcal{P}^{u_1} \cap \mathcal{P}^{u_2}} \frac{1}{\log(D(v))} \tag{A.15}$$

  où $D(\cdot)$ est une fonction renvoyant le degré d'un noeud. D'après l'équation A.15, nous voyons que si u_1 et u_2 ont en commun des POI impliqués dans un petit nombre de transitions, ils auront tendance à être similaires.

- **Jaccard standard sur POI**: Il s'agit ici de la similarité de Jaccard standard. Étant donné les ensembles de points d'intérêt visités par deux utilisateurs u_1 et u_2, nous définissons leur similarité Jaccard $S_J(u_1, u_2)$ comme suit:

$$S_{\mathrm{J}}(u_1, u_2) = |\mathcal{P}^{u_1} \cap \mathcal{P}^{u_2}| \,/\, |\mathcal{P}^{u_1} \cup \mathcal{P}^{u_2}| \tag{A.16}$$

- **Accessibility Weighted Symmetric Jaccard:** Avec cette mesure, nous étendons la mesure Jaccard standard en prenant en compte l'accessibilité entre les POI visités. Intuitivement, plus les POI visités par deux utilisateurs sont accessibles, plus ces deux utilisateurs sont similaires. Pour ce faire, nous ajoutons à l'ensemble des POI visités ceux ($\Gamma(\mathcal{P}^u)$) qui sont accessibles en un bond avec le graphe AGRA. Soit $G = \Gamma(\mathcal{P}^{u_1}) \cup \Gamma(\mathcal{P}^{u_2})$ l'ensemble des POI visités par l'utilisateur u_1 ou par u_2. Soit $N = |G|$. Soit $\rho^{u_1} \in \mathbb{R}_+^N$ et $\rho^{u_2} \in \mathbb{R}_+^N$ deux vecteurs pondérés par les accessibilité. Le vecteur ρ^{u_1} est construit ainsi : $\forall i \in [0, N]$ **if** $p_i \in \mathcal{P}^{u_1}$ **then** $\rho_i^{u_1} = 1$ **else if** $p_i \in \Gamma(\mathcal{P}^{u_1})$ **then** $\rho_i^{u_1} = \sum_{v \in \mathcal{P}^{u_1}} \mathcal{A}_{v,p}$ **otherwise** $\rho_i^{u_1} = 0$. De la même façon nous construisons le vecteur ρ^{u_2}. C'est une métrique symétrique. Ensuite, nous définissons la similarité Jaccard pondérée en accessibilité $S_{AWS}(u_1, u_2)$ comme suit:

$$S_{AWS}(u_1, u_2) = \frac{\sum_{i \in [0,N]} \min(\rho_i^{u_1}, \rho_i^{u_2})}{\sum_{i \in [0,N]} \max(\rho_i^{u_1}, \rho_i^{u_2})} \tag{A.17}$$

- **Accessibility Weighted Antisymmetric Jaccard:** Avec cette métrique, nous essayons de prendre en compte l'asymétrie qui pourrait exister en termes d'influence entre deux utilisateurs. Pour ce faire, nous modifions la

définition de G comme suit : $G = \Gamma(\mathcal{P}^{u_1}) \cup \mathcal{P}^{u_2}$ Au lieu d'étendre les deux ensembles $\mathcal{P}^{u_1}$ et $\mathcal{P}^{u_2}$, nous étendons uniquement l'ensemble des POI visités par l'utilisateur u_1. Puis nous calculons la Accessibility Weighted Antisymmetric Jaccard $S_{AWA}(u_1, u_2)$ en utilisant l'équation A.17. Notons que $S_{AWA}(u_1, u_2) \neq S_{AWA}(u_2, u_1)$.

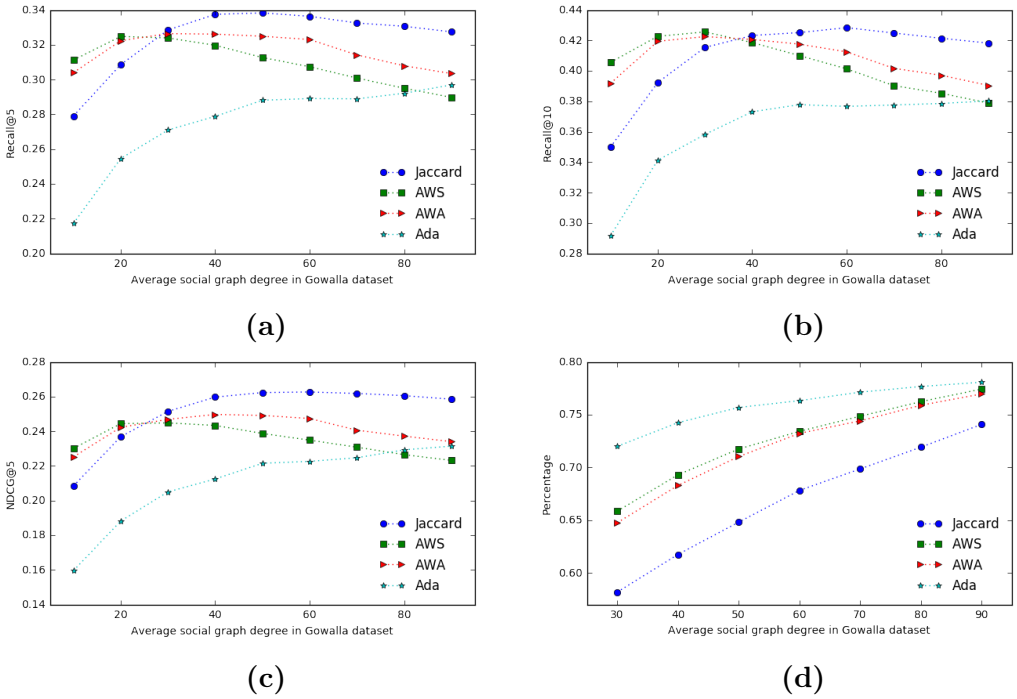


Figure A.2: Résultats de performance des 4 métriques sur Gowalla. Chaque figure représente les résultats de performance des quatre mesures décrites dans la section A.5 pour un nombre différent d'arêtes dans le graphique. Ce nombre d'arêtes est contrôlé par le degré moyen du graphe social.

A.5.2.3 GeoSPF : un modèle de factorisation sociale

La factorisation de Poisson a été largement utilisée pour traiter de nombreux problèmes de recommandation tels que ceux étudiés par [Charlin, Ranganath, McInerney, and Blei 2015; Ma, Liu, King, and Lyu 2011; Gopalan, Hofman, and Blei 2015; Chaney, Blei, and Eliassi-Rad 2015; Liu, Fu, Yao, and Xiong 2013]. Toutes ces approches proposent des extensions du modèle PF où des poids sont utilisés pour influencer la distribution des facteurs latents. Chaney *et al.* a récemment proposé une extension du cadre de la factorisation de Poisson dans [Chaney, Blei, and Eliassi-Rad 2015] appelé SPF. L'extension du modèle de recommandation SPF

semble très adapté à la recommandation de POI car SPF possède un certain nombre de propriétés intéressantes qui répondent à nos besoins, à savoir :

- SPF est basé sur une factorisation de matrices probabiliste connue pour ses performances, à la fois en termes de **qualité** et de **passage à l'échelle** dans le contexte de données éparses ne contenant que des échantillons positifs *c.f.* cas de feedback implicite mentionné dans l'introduction de ce document (voir la section A.5.1).

- SPF permet d'intégrer des **informations sociales**, ce qui est notre objectif dans ce contexte. SPF prend en entrée le cercle d'influence de chaque utilisateur, c'est-à-dire l'ensemble des utilisateurs qui peuvent influencer un utilisateur.

- SPF sépare les questions: *qui est membre du cercle?* de *combien d'influence ce membre transmet-il réellement?* SPF suppose que l'appartenance à un cercle est connue à l'avance, alors que le niveau d'influence est appris. Cette **séparation** est essentielle dans notre cas car le niveau d'influence d'un utilisateur ne dépend pas des POI qu'il/elle partage avec les autres utilisateurs, mais plutôt des interactions cachées (non divulguées) que ces derniers peuvent avoir.

L'idée de GeoSPF est d'intégrer l'influence des éventuels amis de l'utilisateur cible dans le processus de recommandation, en tenant compte des notations de ses voisins. Contrairement à l'équation A.10, GeoSPF exploite la distribution suivante :

$$y_{i,j} \sim \text{Poisson}\left[\mathbf{u}_i^T \cdot \mathbf{v}_j + \sum_{k \in V(i)} \mathbf{s}_{i,k} \cdot x_{k,j}\right] \tag{A.18}$$

où $V(i)$ fait référence à l'ensemble des voisins de l'utilisateur i dans l'ISN et $\mathbf{s}_{i,k}$ fait référence au facteur d'influence sociale latent. Cette variable aléatoire latente modélise l'influence que le voisin k a sur l'utilisateur i. Dans l'équation A.18, nous avons toujours le produit scalaire des utilisateurs et des vecteurs latents des POI, comme dans l'équation A.10, mais nous introduisons un terme d'influence sociale supplémentaire qui correspond à la somme des influences de chaque utilisateur dans le quartier. Le choix du voisinage $V(i)$ est important car $V(i)$ contiendra tous les voisins qui ressemblent le plus à l'utilisateur i. Notre intuition est que nous considérons que plus deux utilisateurs que deux utilisateurs ont l'habitude de visiter sont accessibles, plus ils sont susceptibles de bénéficier de l'influence l'un de l'autre. Le modèle graphique probabiliste de GeoSPF est proposé sur

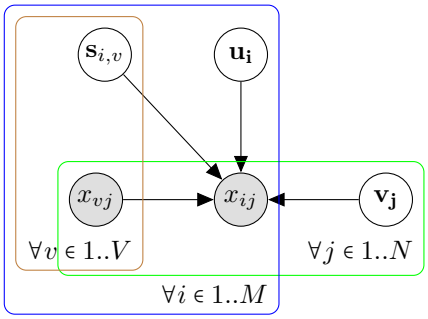


Figure A.3: Modèle graphique de GeoSPF.

la figure A.3. Contrairement à SPF, nous n'utilisons pas ici de réseaux sociaux explicites. De plus, nous pouvons ajuster la qualité de la recommandation en fonction des métriques de similarité utilisées et du filtrage des graphes. En effet, pour chaque graphe construit, nous pouvons choisir différents critères de sélection pour le voisinage des utilisateurs. Nous pouvons également appliquer les filtres sur le graphique d'accessibilité ou sur le graphique social implicite. Cela ajoute une certaine flexibilité à GeoSPF.

A.5.3 Résultats expérimentaux

Nous avons effectué des expériences sur trois ensembles de données du monde réel contenant des checkins issus du `YFCC`, de `Gowalla` et de `Foursquare` qui sont tous largement utilisés. Pour évaluer le comportement de notre solution à différentes échelles géographiques, nous avons filtré les jeux de données afin qu'ils couvrent respectivement une petite, une moyenne et une grande surface géographique, à savoir, Gowalla@Paris couvre une ville, Foursquare couvre une région (autour de Paris), Gowalla couvre un pays (France) et le YFCC couvre l'Europe.

Le jeu de données YFCC a été proposé récemment par [Thomee et al. 2016] qui est le plus grand jeu de données existant pour la recommandation de POI : le jeu de données complet contient 50 millions de check-ins géolocalisés. Par conséquent la plupart des approches existantes en matière de recommandation de POI ne parviennent pas à gérer un volume de données aussi important. Le jeu de données Foursquare a été utilisé dans [Yuan, Cong, Ma, et al. 2013]. Il contient les checkins effectués entre avril 2012 et septembre 2013. Le jeu de données Gowalla a été utilisé dans [Cho, Myers, and Leskovec 2011]. Il contient les checkins effectués entre février 2009 et octobre 2010. Le tableau A.2 présente les statistiques de base concernant les jeux de données que nous avons utilisés.

Table A.2: Statistiques sur les jeux de données utilisés.

Dataset	#Check-ins	#Users	#POIs	avg #POIs	Density
`Gowalla@Paris`	42323	2384	4895	5.6	0.362 %
`Foursquare` [1]	109077	4825	19645	3.1	0.115 %
`Gowalla` [2]	191365	6749	24353	4.1	0.116 %
`YFCC` [3]	48453357	214328	12758657	61.2	0.0017 %

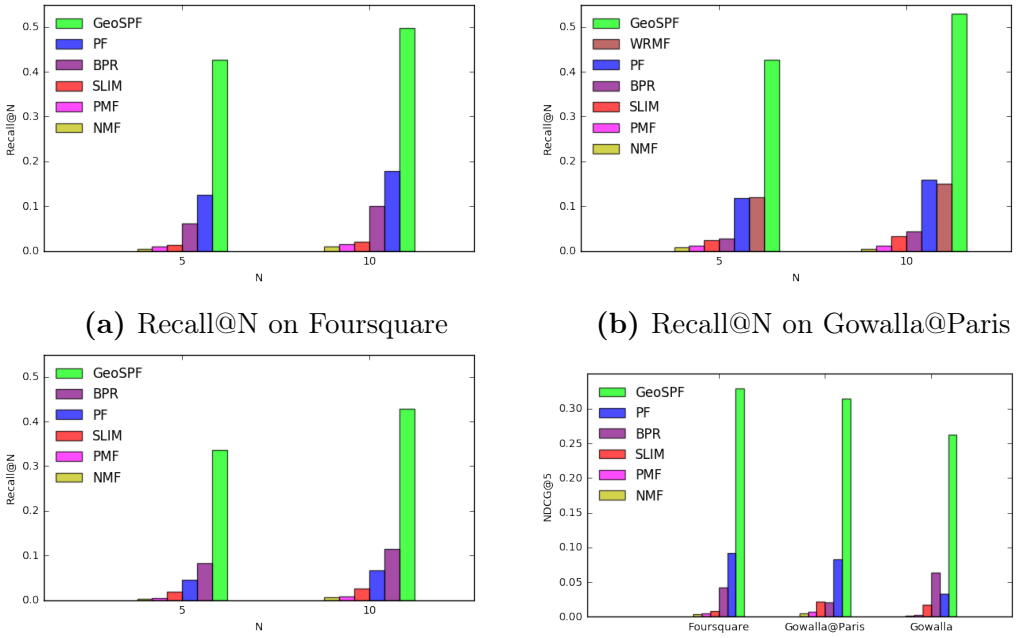
Afin d'estimer les avantages effectifs de notre approche par rapport à des solutions bien connues[4] nous avons choisi des modèles de recommandation efficaces. Plus précisément nous avons comparé $GeoSPF$ aux techniques de recommandation suivantes :

- **NMF:** *Non Negative Matrix Factorization* proposé par [Lee and Seung 2000] est l'un des modèles de factorisation les plus populaires. Ce modèle factorise la matrice de données d'origine grâce aux règles de mise à jour itérative multiplicatives.
- **PMF:** *Probabilistic Matrix Factorization* proposé par [Salakhutdinov and Mnih 2007] est un modèle de factorisation probabiliste efficace basé sur des priors gaussiens sur les données.
- **SLIM:** *Sparse Linear Methods* proposé par [Ning and Karypis 2012] sont adaptés à des ensembles de données très peu denses. Ils sont basés sur un modèle linéaire qui exploite une agrégation clairsemée de coefficients.
- **BPR:** *Bayesian Personalized Ranking* proposé par [Rendle, Freudenthaler, Gantner, and Schmidt-Thieme 2009] a été conçu pour traiter les problèmes de feedback implicite. Il s'agit d'une approche probabiliste évolutive qui optimise un critère de classement. BPR est un concurrent sérieux parmi les approches de pointe.
- **WRMF:** *Weighted Regularized Matrix Factorization* proposé par [Hu, Koren, and Volinsky 2008] a été conçu précisément pour les jeux de données au feedback implicite, ce qui correspond parfaitement aux exigences de la recommandation de POI.

[1] https://sites.google.com/site/yangdingqi/home/foursquare-dataset
[2] http://www.yongliu.org/datasets
[3] https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67
[4] https://gitlab.telecom-paristech.fr/griesner/geopfModeles

- **PoissonMF:** C'est un modèle probabiliste de Poisson récent proposé par [Gopalan, Hofman, and Blei 2015] que nous utilisons comme base de départ de notre approche.

La figure A.4 présente les performances globales de toutes les méthodes comparées listées ci-dessus. Les performances sont évaluées sur deux critères largement utilisés en recommandation de POI : le rappel (ou *recall@N* en fonction de la taille N du sous-ensemble de POI sélectionné) et le NDCG. Sur la figure A.4a le Recall@5 et le Recall@10 sont présentés pour le jeu de données Foursquare. Les mêmes mesures sont reportées sur les figures A.4b (resp. A.4c) pour Gowalla@Paris (resp. Gowalla). Finalement la figure A.4d présente le NDCG@5 pour les trois jeux de données.



(a) Recall@N on Foursquare

(b) Recall@N on Gowalla@Paris

(c) Recall@N on Gowalla

(d) NDCG@5 on three datasets

Figure A.4: Comparaison des performances avec des approches alternatives pour trois jeux de données : Foursquare, Gowalla@Paris et Gowlla. Nous traçons le Recall@N pour N = 5 et N = 10 sur les figures A.4a, A.4b et A.4c. Nous traçons le NDCG@5 sur la figure A.4d. Nous observons que GeoSPF dépasse de manière significative les autres modèles sur les trois jeux de données pour les trois mesures de qualité choisis.

En première observation sur les trois premiers jeux de données (Foursquare, Gowalla@Paris et Gowalla) nous remarquons que notre approche (GeoSPF) sur-

passe de manière significative toutes les autres. Comme prévu, NMF et PMF ne donnent pas une bonne qualité car ils n'ont pas été conçus pour gérer des jeux de données à feedback implicite. Ceci est cohérent avec les résultats présentés dans [Liu and Xiong 2013]. Bien que l'on sache que SLIM fonctionne bien sur des jeux de données épars, il n'atteint pas une qualité satisfaisante dans notre contexte car il suppose un retour d'information explicite. Malheureusement la complexité de WRMF le rend pratiquement inutile sur de grands ensembles de données : le temps de calcul de WRMF est excessivement long au-delà de 5000 POI. C'est pourquoi nous n'obtenons que le rappel@N pour Gowalla@Paris mais pas les deux autres jeux de données où le nombre de points d'intérêt est trop élevé. Cela est dû au fait que la complexité de WRMF dépend fortement du nombre de points d'intérêt, et selon la table 5.1, Gowalla@Paris a moins de points d'intérêt que les autres.

Parmi les meilleurs concurrents en termes de performances PF atteint la meilleure qualité. Nous concentrons donc notre analyse sur la comparaison de PF et de GeoSPF. Le principal avantage de GeoSPF sur l'ensemble des jeux de données est d'environ 200%. Ce gain impressionnant rend GeoSPF apte à la recommandation de POI sur de vastes zones géographiques. Il confirme que l'exploitation d'informations contextuelles restreintes (uniquement le GPS et la date d'enregistrement) au moyen d'une solution combinée géolocalisation/sociale permet d'obtenir une recommandation de grande qualité pour les POI.

A.5.4 Conclusion

Dans ce travail nous avons proposé une nouvelle approche appelée GeoSPF pour le problème de la recommandation de points d'intérêt dans les LBSN. L'objectif principal de GeoSPF est de construire un réseau social implicite qui ne souffre pas du manque de commentaires explicites des utilisateurs concernant leurs visites dans les POI. Sur la base des nouveaux concepts de similarité en termes *d'accessibilité* et d'influence *sociale* que nous avons introduits, notre approche GeoSPF a réussi à **(i)** construire efficacement un modèle de factorisation évolutive implicite et **(ii)** à capturer la similarité sociale de l'utilisateur et enfin **(iii)** à présenter des résultats nettement meilleurs que les niveaux de référence obtenus avec des jeux de données à grande échelle. Nous avons démontré par ailleurs par de nombreuses expériences que GeoSPF surpasse de manière significative toutes les approches alternatives en termes de *rappel* et *NDCG*.

A.6 Passage à l'échelle avec ALGeoSPF

Les approches de recommandation de POI existantes [Gao, Tang, Hu, and Liu 2013; Lian et al. 2014] souffrent du passage à grande échelle et s'appliquent seulement sur de petits volumes de données. Notre travail cible dans cette partie le cas plus réaliste où les jeux de données considérés sont de plusieurs ordres de grandeur (des millions d'utilisateurs et de POI) supérieurs à ceux fréquemment utilisés. Dans cette partie nous présentons notre méthode GeoSPF local/global (noté **ALGeoSPF**) qui correspond à l'approche décrite dans la partie A.5 mais qui passe à l'échelle.

A.6.1 Idée générale

ALGeoSPF définit des couches locales et globales de superPOI afin d'augmenter la densité du jeu de données et de sélectionner une classe d'utilisateurs - en fonction d'un paramètre optimal personnalisé N_{max} - et ainsi d'exploiter le modèle GeoSPF (voir la Section A.5) tout en visant un ensemble dédié d'utilisateurs. Le premier avantage de ALGeoSPF est sa capacité à détecter les comportements de mobilité des utilisateurs, à une échelle locale ou à une échelle plus globale. ALGeoSPF capture les comportements de mobilité existants des utilisateurs citadins et des globetrotters que nous avons observés dans nos jeux de données.

Le deuxième avantage de notre solution multi-échelle est qu'elle permet d'isoler toutes les étapes du processus de recommandation (inférence de réseau social, apprentissage, prédiction) au sein de chaque classe d'utilisateurs, évitant ainsi toute propagation de bruit à travers les classes d'utilisateurs. Cela reflète le fait que, par exemple, le comportement des utilisateurs citadins en matière de mobilité ne peut pas être influencé par la mobilité des globe-trotters.

A.6.1.1 SuperPOI

Pour résoudre le problème de la faible densité des données, nous définissons un ensemble de **superPOI** chacun représentant un groupe de POI. L'ensemble de superPOI constitue lui-même une structure hiérarchique récursive comprenant d'autres superPOI, les zones définissant les superPOI étant disjointes. Bien que la segmentation de l'espace en cellules pour créer des groupes de POI soit relativement simple, la segmentation en grille de taille fixe ne fonctionne pas car elle se aboutit à des superPOI représentant de nombreux POI (par exemple des villes) et d'autres superPOI représentant très peu de POI (par exemple les déserts).

Nous définissons plus formellement cet ensemble de superPOI $\mathfrak{P}$ récursivement : soit $\mathfrak{P}^O = \mathcal{P}$ et soit $\mathfrak{P}^{k+1} = \{\{p_1, p_2, ...\} | \{p_1, p_2, ...\} \in \mathfrak{P}^k\}$. Alors nous avons : $\mathfrak{P} = \bigcup_k \mathfrak{P}^k$. Différentes techniques de clustering peuvent être utilisées pour construire cette famille. Nous avons besoin d'une méthode de segmentation de l'espace garantissant que chaque superPOI a été visité par au plus n utilisateurs. Cette limite sur le nombre maximum d'utilisateurs par superPOI a du sens pour notre problème car elle permet de calculer une similarité plus étroite entre les utilisateurs. La similarité des utilisateurs est un élément essentiel de la recommandation, comme indiqué dans la partie A.5. Par exemple nous pouvons imaginer deux utilisateurs ayant visité Hollywood et la plage de Venice (une partie de la ville de Los Angeles) et aucun autre lieu commun. Si le superPOI est Los Angeles, alors ils n'ont qu'un seul point commun, à savoir Los Angeles. Sinon Los Angeles est divisée en quatre districts avec un superPOI par district : les utilisateurs ont alors deux points communs et deviennent des voisins plus proches lors de la phase de découverte de voisinage.

A.6.1.2 Schémas de mobilité

Nous observons dans les données des comportements de mobilité distincts. Cependant dans la plupart des travaux existants aucune distinction n'est faite entre les comportements de mobilité des utilisateurs car ils ne parviennent pas à modéliser à la fois à l'échelle microscopique et l'échelle macroscopique. Or il existe en réalité différentes classes de voyageurs : certains utilisateurs sont plus susceptibles de faire des voyages à longue distance, tandis que d'autres seront limités à des zones restreintes telles que des villes ou des régions. Nous appellerons par la suite les premiers des *globetrotters* et les seconds des *citadins*. L'objectif de notre approche est d'exploiter l'influence sociale latente des utilisateurs appartenant exclusivement à la même classe.

En pratique la recommandation nécessite de connaître au moins 5 points d'intérêt distincts par utilisateur afin de construire un modèle avec une qualité acceptable. Ainsi pour recommander un utilisateur citadin il faut connaître les différents lieux visités. Cela contredit en quelque sorte le premier objectif décrit à la partie A.6.1.1. Par exemple, dans un scénario étendu où un superPOI peut représenter une cellule aussi grande qu'une ville, un utilisateur citadin peut faire fusionner toutes ses visites en un seul superPOI. Cela signifie que, d'une part, l'exigence de densité nécessite d'agréger des points d'intérêt en de plus grands superPOI et, d'autre part, que la structure de faible mobilité des utilisateurs citadins nécessite de conserver des points d'intérêt peu localisés.

A.6.2 Hiérarchie de superPOI

ALGeoSPF exploite une structure hiérarchique de superPOI. Les auteurs de [Zhang, Wang, et al. 2017] ont récemment étudié l'idée d'exploiter une structure hiérarchique de catégories et une influence géographique répartie entre différentes régions. Cependant leur problème est différent du nôtre, dans la mesure où ils proposent une approche permettant de prédire la catégorie des prochains POI visités. L'agrégation en superPOI permet de traiter des jeux de données à grande densité comme à faible densité en agrégeant des parties des jeux de données d'origine. Nous définissons une structure à plusieurs couches comme détaillée dans la partie A.6.1 pour agréger les POI en superPOI visités par un nombre croissant d'utilisateurs.

A.6.2.1 Clustering géographique

Nous utilisons une technique de clustering qui consiste à diviser l'espace géographique initial en cellules rectangulaires. Comme proposé par [Wang, Yang, and Muntz 1997] une cellule peut être divisée de manière récursive en 4 cellules. Ainsi nous construisons un arbre dont la racine est la carte du monde entier et chaque noeud est le quart de sa région parente. Le principe de l'algorithme est de diviser récursivement une cellule c jusqu'à ce qu'elle réponde à la condition : $N(c) < N_{max}$. Les cellules remplissant cette condition sont choisies pour être les superPOI. Le résultat de l'algorithme de classification est un ensemble de cellules superPOI noté S. Le paramètre N_{max} permet de contrôler le niveau d'agrégation. Cet algorithme est présenté dans l'algorithme 2. Deux paramètres controlent la qualité du clustering, à savoir le choix de la cellule initiale, et la taille du cluster.

Algorithm 2 Top-down Clustering Method for ALGeoSPF

Input:
- N_{max}: maximum number of users having visited a cell.

Global Output:
- S: the set of superPOIs cells.

Initialize: $S \leftarrow \varnothing$
function WORLDTOSUPERPOIS (C: a cell)
  Split C into 4 even rectangular cells $C_1, \cdots, C_4$
  for each C_i **do**
    if $N(C_i) > N_{max}$ and $\#POIs(C_i) \geqslant 2$ **then**
      worldToSuperPOIs (C_i)
    else Put C_i into S

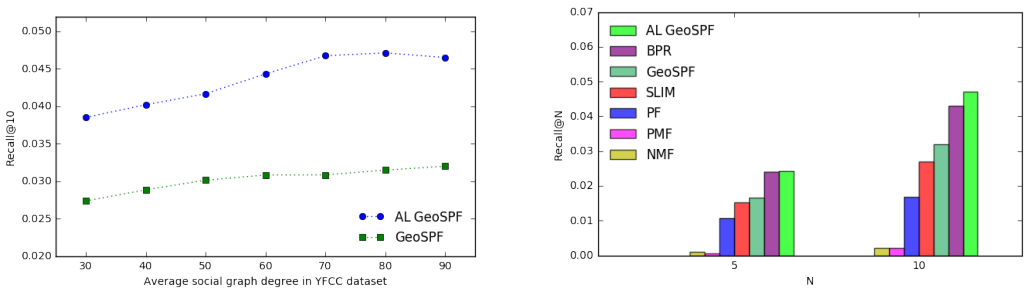
- La **cellule initiale** sur laquelle appliquer le clustering. Considérer le monde entier comme la cellule initiale convient bien aux utilisateurs avec de nombreuses archives réparties dans le monde entier. Mais pour la plupart des utilisateurs, la zone couvrant l'ensemble de leurs enregistrements est (par exemple, Europe, France, Paris). Prendre en compte des effets de cellule initiaux plus petits pour augmenter la densité du jeu de données.

- Le **taille du cluster** est défini par N_{max}. L'augmentation de N_{max} donnera un nombre de superPOI inférieur et supérieur, ce qui augmentera la densité du jeu de données. Cependant, N_{max} est lié: pour chaque utilisateur, il existe un maximum N_{max} (noté $N_{max}^{utilisateur}$) au-delà duquel la recommandation n'est plus possible car l'utilisateur ne ont visité suffisamment de superPOI distincts (nous avons besoin d'au moins 5 visites distinctes par utilisateur).

A.6.3 Résultats expérimentaux

Pour évaluer le comportement de notre solution à différentes échelles géographiques, nous avons filtré les jeux de données afin qu'ils couvrent respectivement une petite, une moyenne et une grande surface. À savoir, Gowalla@Paris couvre une ville, Foursquare couvre une région (autour de Paris), Gowalla couvre un pays (France) et YFCC couvre l'Europe. Le jeu de données YFCC a été proposé récemment par [Thomee et al. 2016]. Il s'agit du plus grand jeu de données existant pour la recommandation de POI. Le jeu de données complet contient plus de 50 millions d'enregistrements géolocalisés.

Notre expérience vise à évaluer les avantages de notre approche de clustering géographique pour les recommandations tenant compte de la classe d'utilisateurs. La figure A.5 indique la qualité de la recommandation (rappel) ALGeoSPF appliquée sur le jeu de données YFCC, en considérant les utilisateurs urbains isolés des globetrotters. Plus précisément, la figure A.5a rapporte le rappel @ 10 de GeoSPF et ALGeoSPF pour différentes tailles moyennes du réseau social implicite. Pour chaque taille du réseau, nous observons qu'ALGeoSPF améliore toujours considérablement le rappel de 50La figure A.5b indique le rappel @ 5 et le rappel @ 10 de tous les concurrents, ainsi que ALGeoSPF, sur le jeu de données YFCC (avec une taille de réseau social fixe de 80). Nous pouvons voir que ALGeoSPF surpasse les autres méthodes, bien que BPR offre une qualité proche. Nous observons également que globalement les mesures de rappel des modèles testés pour le jeu de données YFCC sont beaucoup plus basses que d'autres jeux de données: ceci est dû à la faible densité en raison de la taille de la zone géographique couverte (voir la section 5.1).



(a) Recall@10 on YFCC **(b)** Recall@5 and Recall@10

Figure A.5: Performance comparison of ALGeoSPF *wrt.* state-of-the-art approaches for 2 levels of the YFCC dataset. We plot on figure A.5a the recall@10 results of GeoSPF and ALGeoSPF for different size of the average social graph degree. Figure A.5b presents the results of AMGeoSPF in terms of recall@5 and recall@10.

A.6.4 Conclusion

Nous avons proposé une nouvelle approche pour la recommandation de points d'intérêt appelée ALGeoSPF. Basée sur le concept de *superPOI* et sur un algorithme de clustering géographique, notre approche a réussi à **(i)** intégrer efficacement un modèle de factorisation qui passe à l'échelle et à **(ii)** capturer les préférences de l'utilisateur en termes de mobilité dans une structure hiérarchique et, enfin, **(iii)** présenter des résultats nettement meilleurs que les seuils de référence sur des jeux de données à grande échelle.

A.7 Conclusion générale

Notre travail de recherche a permis d'explorer davantage les méthodes de factorisation de matrices et également d'aboutir à plusieurs solutions efficaces au problème de la recommandation de points d'intérêt. En particulier nous avons décrit un modèle de factorisation qui intègre les distributions géographiques et temporelles des visites faites par les utilisateurs dans les POI. Dans la perspective d'appliquer la factorisation de matrices à des jeux de données plus réalistes, nous avons proposé un modèle basé sur la factorisation de Poisson qui a permis d'obtenir des résultats très prometteurs. Enfin nous avons proposé de segmenter les profils utilisateurs de façon à distinguer, dans un modèle de factorisation, les grandes classes de voyageurs. Ainsi nous avons réussi à limiter les biais d'apprentissage entre les utilisateurs de classes différentes. Nous espérons que ce dernier axe de recherche sera exploré plus en détail dans le futur.

Bibliography

Ramesh A., Anusha J., and Clarence J.M. Tauro (2014). "A Novel, Generalized Recommender System for Social Media Using the Collaborative-filtering Technique". In: *SIGSOFT Softw. Eng. Notes* (cit. on p. 70).

Gediminas Adomavicius and Alexander Tuzhilin (2005). "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions". In: *IEEE Trans. on Knowl. and Data Eng.* (Cit. on pp. 16, 27, 28, 31, 104).

Ramesh Baral and Tao Li (2016). "MAPS: A Multi Aspect Personalized POI Recommender System". In: *RecSys'16* (cit. on pp. 17, 105).

Nicola Barbieri, Giuseppe Manco, and Ettore Ritacco (2014). "Probabilistic Approaches to Recommendations". In: Synthesis Lectures on Data Mining and Knowledge Discovery (cit. on p. 28).

Robert M. Bell and Yehuda Koren (2007). "Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights". In: ICDM '07 (cit. on p. 36).

James Bennett, Stan Lanning, and Netflix Netflix (2007). "The Netflix Prize". In: (cit. on pp. 20, 34, 99, 106).

Betim Berjani and Thorsten Strufe (2011). "A Recommendation System for Spots in Location-based Online Social Networks". In: SNS '11 (cit. on pp. 47, 51).

Joan Borris, Antonio Moreno, and Aida Valls (2014). "Review: Intelligent Tourism Recommender Systems: A Survey". In: *Expert Syst. Appl.* (Cit. on p. 28).

John S. Breese, David Heckerman, and Carl Kadie (1998). "Empirical Analysis of Predictive Algorithms for Collaborative Filtering". In: UAI'98 (cit. on p. 34).

John Canny (2004). "GaP: A Factor Model for Discrete Data". In: SIGIR '04 (cit. on pp. 65, 113).

Allison J.B. Chaney, David M. Blei, and Tina Eliassi-Rad (2015). "A Probabilistic Model for Using Social Networks in Personalized Item Recommendation". In: RecSys '15 (cit. on pp. 71, 72, 116).

Laurent Charlin, Rajesh Ranganath, James McInerney, and David M. Blei (2015). "Dynamic Poisson Factorization". In: RecSys '15 (cit. on pp. 62, 65, 71, 85, 113, 116).

Chen Cheng, Haiqin Yang, Irwin King, and Michael R. Lyu (2012). "Fused Matrix Factorization with Geographical and Social Influence in Location-based Social Networks". In: AAAI'12 (cit. on pp. 41, 47, 51, 53, 55, 62, 64, 65, 70, 85, 111).

Chen Cheng, Haiqin Yang, Michael R. Lyu, and Irwin King (2013). "Where You Like to Go Next: Successive Point-of-interest Recommendation". In: IJCAI '13 (cit. on p. 41).

Zhiyong Cheng, Jialie Shen, and Tao Mei (2014). "Just-for-me: An Adaptive Personalization System for Location-aware Social Music Recommendation". In: SIGIR '14 (cit. on pp. 17, 104).

Eunjoon Cho, Seth A. Myers, and Jure Leskovec (2011). "Friendship and Mobility: User Movement in Location-based Social Networks". In: KDD '11 (cit. on pp. 55, 65, 76, 85, 94, 111, 118).

Martin J. Chorley, Roger M. Whitaker, and Stuart M. Allen (2015). "Personality and location-based social networks". In: *Computers in Human Behavior* (cit. on pp. 16, 104).

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey (2008). "An Experimental Comparison of Click Position-bias Models". In: WSDM '08 (cit. on p. 101).

Shuguang Cui, Alfred O. Hero, Zhi-Quan Luo, and Jos M. F. Moura (2016). *Big Data over Networks* (cit. on pp. 16, 104).

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990). "Indexing by latent semantic analysis". In: *Journal of the American Society for Information Science* (cit. on p. 36).

Christian Desrosiers and George Karypis (2011). "A Comprehensive Survey of Neighborhood-based Recommendation Methods." In: (cit. on p. 35).

Charles Elkan and Keith Noto (2008). "Learning Classifiers from Only Positive and Unlabeled Data". In: KDD 08 (cit. on pp. 63, 101).

Shanshan Feng, Xutao Li, Yifeng Zeng, Gao Cong, Yeow Meng Chee, and Quan Yuan (2015). "Personalized Ranking Metric Embedding for Next New POI Recommendation". In: IJCAI'15 (cit. on pp. 41, 47, 51, 64).

Gregory Ference, Mao Ye, and Wang-Chien Lee (2013). "Location Recommendation for Out-of-town Users in Location-based Social Networks". In: CIKM '13 (cit. on pp. 42, 47).

Zeno Gantner, Steffen Rendle, and Lars Schmidt-Thieme (2010). "Factorization Models for Context-/Time-aware Movie Recommendations". In: CAMRa '10 (cit. on pp. 17, 105).

Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu (2013). "Exploring Temporal Effects for Location Recommendation on Location-based Social Networks". In: RecSys '13 (cit. on pp. 41, 45, 47, 52, 54, 62, 64, 85, 86, 110, 122).

Prem Gopalan, Jake M. Hofman, and David M. Blei (2015). "Scalable Recommendation with Hierarchical Poisson Factorization". In: UAI'15 (cit. on pp. 71, 74, 77, 95, 116, 120).

Jean-Benoît Griesner, Talel Abdessalem, and Hubert Naacke (2015). "POI Recommendation: Towards Fused Matrix Factorization with Geographical and Temporal Influences". In: *RecSys '15* (cit. on pp. 21, 49, 51, 67, 85, 99, 107).

Jean-Benoît Griesner, Talel Abdesssalem, and Hubert Naacke (2017). "Un Modèle de Factorisation de Poisson pour la Recommandation de Points d'Intérêt". In: *Revue des Nouvelles Technologies de l'Information* (cit. on pp. 21, 61, 100, 107).

Jean-Benoît Griesner, Talel Abdesssalem, Hubert Naacke, and Pierre Dosne (2018). "ALGeoSPF: Un modèle de factorisation basé sur du clustering géographique pour la recommandation de POI". In: *Revue des Nouvelles Technologies de l'Information* (cit. on pp. 22, 81, 100, 107).

Modou Gueye, Talel Abdessalem, and Hubert Naacke (2013). "Technique de factorisation multi-biais pour des recommandations dynamiques". In: (cit. on p. 59).

Modou Gueye, Talel Abdessalem, and Hubert Naacke (2015). "Dynamic Recommender System: Using Cluster-Based Biases to Improve the Accuracy of the Predictions". In: Studies in Computational Intelligence (cit. on p. 59).

Cheng-Kang Hsieh, Longqi Yang, Honghao Wei, Mor Naaman, and Deborah Estrin (2016). "Immersive Recommendation: News and Event Recommendations Using Personal Digital Traces". In: WWW '16 (cit. on pp. 17, 105).

Yifan Hu, Yehuda Koren, and Chris Volinsky (2008). "Collaborative Filtering for Implicit Feedback Datasets". In: ICDM '08 (cit. on pp. 20, 45, 50, 52, 64, 65, 77, 84–86, 95, 108, 119).

He Jing, Li Xin, and Liao Lejian (2017). "Category-aware Next Point-of-Interest Recommendation via Listwise Bayesian Personalized Ranking". In: IJCAI'17 (cit. on pp. 17, 105).

George Karypis (2001). "Evaluation of Item-Based Top-N Recommendation Algorithms". In: CIKM '01 (cit. on p. 38).

Andreas Klein, Henning Ahlf, and Varinder Sharma (2015). "Social Activity and Structural Centrality in Online Social Networks". In: *Telemat. Inf.* (Cit. on p. 15).

Yehuda Koren, Robert Bell, and Chris Volinsky (2009). "Matrix Factorization Techniques for Recommender Systems". In: *Computer* (cit. on pp. 34, 50, 64).

Daniel D. Lee and H. Sebastian Seung (2000). "Algorithms for Non-negative Matrix Factorization". In: NIPS'00 (cit. on pp. 76, 94, 119).

Joonseok Lee and Sami Abu-El-Haija (2017). "Large-Scale Content-Only Video Recommendation". In: (cit. on p. 84).

Daniel Lewis (2006). "What is Web 2.0?" In: *XRDS* (cit. on p. 15).

Huayu Li, Yong Ge, Richang Hong, and Hengshu Zhu (2016). "Point-of-Interest Recommendations: Learning Potential Check-ins from Friends". In: KDD '16 (cit. on pp. 17, 105).

Xutao Li, Gao Cong, Xiao-Li Li, Tuan-Anh Nguyen Pham, and Shonali Krishnaswamy (2015). "Rank-GeoFM: A Ranking Based Geographical Factorization Method for Point of Interest Recommendation". In: SIGIR '15 (cit. on p. 47).

Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui (2014). "GeoMF: Joint Geographical Modeling and Matrix Factorization for Point-of-interest Recommendation". In: KDD '14 (cit. on pp. 18, 45, 47, 50, 53, 62, 64, 65, 67, 84–86, 109, 122).

Bin Liu, Yanjie Fu, Zijun Yao, and Hui Xiong (2013). "Learning Geographical Preferences for Point-of-interest Recommendation". In: KDD '13 (cit. on pp. 47, 62, 72, 85, 116).

Bin Liu and Hui Xiong (2013). "Point-of-Interest Recommendation in Location Based Social Networks with Topic and Location Awareness". In: ICDM'13 (cit. on pp. 47, 53, 69, 77, 86, 95, 114, 121).

A. Losup, van de Bovenkamp R., Shen S., Lu Jia A., and Kuipers F. (2014). "Analyzing Implicit Social Networks in Multiplayer Online Games". In: *IEEE Internet Computing* (cit. on pp. 65, 85).

Claudio Lucchese, Raffaele Perego, Fabrizio Silvestri, Hossein Vahabi, and Rossano Venturini (2012). "How Random Walks Can Help Tourism". In: ECIR'12 (cit. on p. 42).

Hao Ma, Irwin King, and Michael R. Lyu (2009). "Learning to Recommend with Social Trust Ensemble". In: SIGIR '09 (cit. on p. 42).

Hao Ma, Chao Liu, Irwin King, and Michael R. Lyu (2011). "Probabilistic Factor Models for Web Site Recommendation". In: SIGIR '11 (cit. on pp. 71, 116).

Hao Ma, Tom Chao Zhou, Michael R. Lyu, and Irwin King (2011). "Improving Recommender Systems by Incorporating Social Contextual Information". In: *ACM Trans. Inf. Syst.* 29.2 (cit. on p. 70).

L.J.P. van der Maaten, E. O. Postma, and H. J. van den Herik (2008). *Dimensionality Reduction: A Comparative Review* (cit. on p. 36).

Harvey J. Miller (2004). "Tobler's first law and spatial analysis". English. In: *Annals of the American Association of Geographers* (cit. on pp. 40, 49, 85).

Xia Ning and George Karypis (2012). "Sparse Linear Methods with Side Information for Top-n Recommendations". In: RecSys '12 (cit. on pp. 77, 95, 119).

Gang Niu, Marthinus C. du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama (2016). "Theoretical Comparisons of Positive-unlabeled Learning Against Positive-negative Learning". In: NIPS'16 (cit. on p. 101).

Douglas Oard and Jinmook Kim (1998). "Implicit Feedback for Recommender Systems". In: (cit. on p. 29).

Rong Pan, Yunhong Zhou, Bin Cao, Nathan N. Liu, Rajan Lukose, Martin Scholz, and Qiang Yang (2008). "One-Class Collaborative Filtering". In: ICDM '08 (cit. on p. 30).

Michael J. Pazzani and Daniel Billsus (2007). "The Adaptive Web". In: (cit. on p. 31).

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme (2009). "BPR: Bayesian Personalized Ranking from Implicit Feedback". In: UAI '09 (cit. on pp. 77, 95, 119).

Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (2010). *Recommender Systems Handbook* (cit. on pp. 28, 30, 32, 33, 38).

Ruslan Salakhutdinov and Andriy Mnih (2007). "Probabilistic Matrix Factorization". In: NIPS'07 (cit. on pp. 64, 77, 95, 119).

G. Salton, A. Wong, and C. S. Yang (1975). "A Vector Space Model for Automatic Indexing". In: *Commun. ACM* (cit. on p. 32).

Masoud Sattari, Murat Manguoglu, Ismail H. Toroslu, Panagiotis Symeonidis, Pinar Senkul, and Yannis Manolopoulos (2012). "Geo-activity Recommendations by Using Improved Feature Combination". In: UbiComp '12 (cit. on p. 51).

Martin Saveski and Amin Mantrach (2014). "Item Cold-start Recommendations: Learning Local Collective Embeddings". In: RecSys '14 (cit. on p. 30).

J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen (2007). "Collaborative Filtering Recommender Systems". In: (cit. on p. 34).

Sumit Sidana, Charlotte Laclau, Massih R. Amini, Gilles Vandelle, and André Bois-Crettez (2017). "KASANDR: A Large-Scale Dataset with Implicit Feedback for Recommendation". In: SIGIR '17 (cit. on p. 84).

Ajit P. Singh and Geoffrey J. Gordon (2008). "Relational Learning via Collective Matrix Factorization". In: KDD '08 (cit. on p. 51).

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li (2016). "YFCC100M: the new data in multimedia research". In: *Commun. ACM* (cit. on pp. 20, 75, 93, 118, 125).

Alvin Toffler (1970). *Future Shock*. Random House (cit. on pp. 16, 104).

B. Towle and C. Quinn (2000). "Knowledge Based Recommender Systems Using Explicit User Models". In: (cit. on p. 29).

Wei Wang, Jiong Yang, and Richard R. Muntz (1997). "STING: A Statistical Information Grid Approach to Spatial Data Mining". In: VLDB '97 (cit. on pp. 83, 90, 124).

Weiqing Wang, Hongzhi Yin, Ling Chen, Yizhou Sun, Shazia Sadiq, and Xiaofang Zhou (2015). "Geo-SAGE: A Geographical Sparse Additive Generative Model for Spatial Item Recommendation". In: KDD '15 (cit. on p. 42).

Liang Xiang, Quan Yuan, Shiwan Zhao, Li Chen, Xiatian Zhang, Qing Yang, and Jimeng Sun (2010). "Temporal Recommendation on Graphs via Long- and Short-term Preference Fusion". In: KDD '10 (cit. on p. 44).

Min Xie, Hongzhi Yin, Hao Wang, Fanjiang Xu, Weitong Chen, and Sen Wang (2016). "Learning Graph-based POI Embedding for Location-based Recommendation". In: CIKM '16 (cit. on pp. 17, 105).

Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee (2011). "Exploiting Geographical Influence for Collaborative Point-of-interest Recommendation". In: SIGIR '11 (cit. on pp. 41, 42, 47, 50, 51, 53, 62, 64, 65, 84).

Haochao Ying, Liang Chen, Yuwen Xiong, and Jian Wu (2016). "PGRank: Personalized Geographical Ranking for Point-of-Interest Recommendation". In: WWW '16 Companion (cit. on pp. 17, 105).

Bassant E. Youssef (2014). "ONLINE SOCIAL NETWORK INTERNETWORKING ANALYSIS". In: *International Journal of Next-Generation Networks (IJNGN)* (cit. on p. 104).

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat- Thalmann (2013). "Time-aware Point-of-interest Recommendation". In: SIGIR '13 (cit. on pp. 17, 41, 67, 76, 94, 105, 118).

Quan Yuan, Gao Cong, and Aixin Sun (2014). "Graph-based Point-of-interest Recommendation with Geographical and Temporal Influences". In: CIKM '14 (cit. on p. 44).

Chenyi Zhang, Hongwei Liang, Ke Wang, and Jianling Sun (2015). "Personalized Trip Recommendation with POI Availability and Uncertain Traveling Time". In: CIKM '15 (cit. on p. 42).

D. Y. Zhang, D. Wang, H. Zheng, X. Mu, Q. Li, and Y. Zhang (2017). "Large-scale point-of-interest category prediction using natural language processing models". In: pp. 1027–1032 (cit. on pp. 89, 124).

Jia-Dong Zhang and Chi-Yin Chow (2013). "iGSLR: Personalized Geo-social Location Recommendation: A Kernel Density Estimation Approach". In: SIGSPATIAL'13 (cit. on pp. 41, 42, 47, 50, 51, 53, 55, 62, 65, 70, 84, 85, 111).

Jia-Dong Zhang and Chi-Yin Chow (2015). "GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations". In: SIGIR '15 (cit. on pp. 43, 47).

Wei Zhang and Jianyong Wang (2015). "Location and Time Aware Social Collaborative Retrieval for New Successive Point-of-Interest Recommendation". In: CIKM '15 (cit. on pp. 41, 47, 62, 64, 65, 85).

Vincent W. Zheng, Yu Zheng, Xing Xie, and Qiang Yang (2010). "Collaborative location and activity recommendations with GPS history data". In: WWW '10 (cit. on pp. 51, 64).

Shi Zong, Hao Ni, Kenny Sung, Nan Rosemary Ke, Zheng Wen, and Branislav Kveton (2016). "Cascading Bandits for Large-scale Recommendation Problems". In: UAI'16 (cit. on pp. 84, 101).

Scalable Models for Points-Of-Interest Recommender Systems

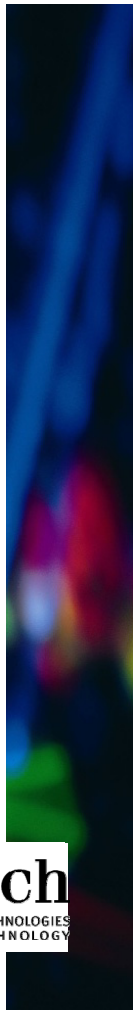
Jean-Benoît Griesner

RÉSUMÉ : La recommandation de points d'intérêts (POI) est une composante essentielle des réseaux sociaux géolocalisés. Cette tâche pose de nouveaux défis dûs aux contraintes spécifiques de ces réseaux. Cette thèse étudie de nouvelles solutions au problème de la recommandation personnalisée de POI. Trois contributions sont proposées dans ce travail.
La première contribution est un nouveau modèle de factorisation de matrices qui intègre les influences géographique et temporelle. Ce modèle s'appuie sur un traitement spécifique des données. La deuxième contribution est une nouvelle solution au problème dit du feedback implicite. Ce problème correspond à la difficulté à distinguer parmi les POI non visités, les POI dont l'utilisateur ignore l'existence des POI qui ne l'intéressent pas. Enfin la troisième contribution de cette thèse est une méthode pour générer des recommandations à large échelle. Cette approche combine un algorithme de clustering géographique avec l'influence sociale des utilisateurs à différentes échelles de mobilité.

MOTS-CLÉS : recommandation, POI, factorisation, matrice, Poisson, géographie

ABSTRACT: The task of points-of-interest (POI) recommendations has become an essential feature in location-based social networks. However it remains a challenging problem because of specific constraints of these networks. In this thesis I investigate new approaches to solve the personalized POI recommendation problem. Three main contributions are proposed in this work.
The first contribution is a new matrix factorization model that integrates geographical and temporal influences. This model is based on a specific processing of geographical data. The second contribution is an innovative solution against the implicit feedback problem. This problem corresponds to the difficulty to distinguish among unvisited POI the actual "unknown" from the "negative" ones. Finally the third contribution of this thesis is a new method to generate recommendations with large-scale datasets. In this approach I propose to combine a new geographical clustering algorithm with users' implicit social influences in order to define local and global mobility scales.

KEYWORDS: recommendation, POI, factorization, matrix, Poisson, geographical



TELECOM ParisTech

EDITE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY