

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST-0026

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « INFORMATIQUE et RESEAUX »

présentée et soutenue publiquement par

Xiaolan SHA

le 6 mai 2013

Personnalisation du Contenu et Tendances dans les Médias Sociaux

Directeur de thèse : **Pietro Michiardi**

Jury

M. Ernst BIERSACK, Département Réseaux et Sécurité, EURECOM	Président
M. Guillaume URVOY-KELLER, Laboratoire I3S, Université Nice Sophia Antipolis	Rapporteur
Mme. Elena BARALIS, Dipartimento di Automatica e Informatica, Politecnico di Torino	Rapporteur
M. Daniele QUERCIA, Yahoo ! Lab, Barcelona	Examinateur

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

THÈSE

Abstract

Social media continuously draws the interest of researchers from a variety of perspectives - networks, sociology, marketing etc. In this networked age, the role of mass media at spreading information is increasingly opening itself to individual contributions. Researchers have therefore focused on how information is disseminated by individuals through social networks. Fluctuating along user connections, some content succeeds at capturing the attention of a large amount of users and suddenly becomes trending. Understanding trending content and its dynamics is crucial to the explanation of opinion spreading, and to the design of social marketing strategies. While previous research has mostly focused on trending content and on the network structure of individuals in social media, this work complements these studies by exploring in depth the human factors behind the generation of this content. We build upon this analysis to investigate new personalization tools helping individuals to discover interesting social media content. This work contributes to the literature on the following aspects:

- An in depth analysis on individuals who create trending content in social media that uncovers their distinguishing characteristics;
- A novel means to identify trending content by relying on the ability of special individuals who create them;
- A mechanism to build a recommender system to personalize trending content;
- Techniques to improve the quality of recommendations beyond the core theme of accuracy.

Our studies underline the vital role of special users in the creation of trending content in social media. Thanks to such special users and their "wisdom", individuals may discover the trending content distilled to their tastes. Our work brings insights in two main research directions - trending content in social media and recommender systems.

Résumé

Actuellement, les médias sociaux retiennent continûment l'attention des chercheurs dans des domaines varies comme par exemple les réseaux, la sociologie, le marketing, etc. À notre époque où tout devient interconnecté, les médias de masse accordent de plus en plus d'importance aux contributions des individus dans la diffusion de l'information. Les chercheurs se sont donc intéressés à la façon dont l'information se propage dans les réseaux sociaux. En fonction des connexions entre utilisateurs de ces réseaux, certains contenus peuvent bénéficier d'une large audience et tout d'un coup se transformer en tendance. Comprendre comment du contenu peut se transformer en tendance est donc crucial pour pouvoir expliquer la propagation des opinions ainsi que pour établir des stratégies de marketing sociale. Les précédentes études se sont concentrées sur les caractéristiques du contenu pouvant se transformer en tendance et sur la structure du réseau d'individus dans les médias sociaux. Ce travail complète ces études en explorant les facteurs humains derrières la génération du contenu tendance. Nous nous appuyons sur cette analyse pour définir de nouveaux outils de personnalisation permettant aux individus de repérer le contenu qui les intéresse dans les médias sociaux. Les contributions de ce travail sont les suivantes :

- Une analyse approfondie des individus créant du contenu tendance dans les médias sociaux ce qui permet de découvrir leurs caractéristiques distinctives ;
- Un nouveau moyen d'identifier le contenu tendance en s'appuyant sur la capacité des individus spéciaux qui le créent ;
- Un mécanisme d'élaboration de système de recommandation afin de personnaliser le contenu tendance ;
- Des techniques d'amélioration de la qualité des recommandations allant au-delà de la seule évaluation de la précision.

Nos études montrent le rôle vital de certains utilisateurs spéciaux dans la création de contenu tendance dans les médias sociaux. Ces utilisateurs avec leur sagesse permettent aux autres individus de découvrir du contenu tendance à leur goût. Notre travail contribue aux deux principales orientations de recherche : le contenu tendance dans les médias sociaux et les systèmes de recommandation.

Acknowledgements

This manuscript covers my research work at EURECOM as a PhD candidate. However, what it does not spell out are the good days and the great people happened to be there, without whom this thesis would not have been accomplished.

First, I would like to send a big and sincere thanks to my advisor Prof. Pietro Michiardi, who has opened me the door to research, and has supported me throughout the entire thesis with his passion of research, giving me the freedom to proceed in projects the way I like (and learn from my mistakes). Also, a special, enormous, thanks to Dr. Daniele Quercia and Dr. Matteo Dell'Amico, from whom I have received interesting discussions, great advices, big encouragement, and amazing inspirations. I am grateful to my jury, Prof. Ernst Biersack, Prof. Guillaume Urvoy-Keller, and Prof. Elena Baralis, as well as to all the anonymous reviewers who have been kind and patient to give valuable feedback.

These years would not have been the same without the members (or ex members) of our group, Francesco Albanese, Mario Pastorelli, Antonio Barbuzzi, Daniele Venzano and Duy-Hung Phan, with whom I have shared the offices, the cluster, as well as the sour of fails and the sweetness of our successes. Beyond our group was the corridor, and my EURECOM friends, Leyla Bilge, Davide Balzarotti, Pierre-Antoine Vervier, Jako Fritz, Olivier Thonnard, Hadrien Hours, Andrea Lanzi, as well as the chinese community: Heng Cui, Lei Xiao, Kaijie Zhou, Juan Hao, Xuran Zhao, Rui Min, Xinping Yi, Xueliang Liu, Jinbang Chen, Jingjing Zhang, Qianrui Li, Shengyun Liu. Thanks to them, the journey towards the completion of this thesis was full of joy. Special thanks to KuangTing Liu, Huiyi Chen Haiying Zhou, and Jianyi Huang, with whom, I have enjoyed the leisure time in shopping, dining out, some skiing, golf and the exercise called swimming that I still don't know how to do.

A final thought to my husband Corrado Leita for his unconditional support and caring. I thank my family for their understanding. Last, I would also like to remember Alessandro Duminuco and Marco Paleari for having shared with me their wonderful journey to a PhD, and that inspired me into starting this challenge.

Contents

List of Figures xiii

List of Tables xv

1 Introduction 1
- 1.1 Motivations . . . 1
- 1.2 Research Problems and Contributions . . . 5
- 1.3 Organization of the thesis . . . 6

2 Background and Related Work 7
- 2.1 Trends in Social Media . . . 8
  - 2.1.1 Identification . . . 8
  - 2.1.2 Characterization . . . 10
  - 2.1.3 Influentials . . . 12
- 2.2 Recommender Systems . . . 12
  - 2.2.1 User Preferences . . . 13
  - 2.2.2 Collaborative Filtering . . . 14
  - 2.2.3 Evaluation . . . 17
- 2.3 Summary . . . 20

3 Who Creates Trends 23
- 3.1 Background on the Mobile Social Application . . . 23
- 3.2 Dataset . . . 25
  - 3.2.1 Uploads and Votes . . . 25
  - 3.2.2 Geography . . . 26
  - 3.2.3 Following . . . 26
- 3.3 Identifying Trend Makers and Spotters . . . 28
  - 3.3.1 Defining Trend Spotters and Trend Makers . . . 28
  - 3.3.2 Characterizing Trend Spotters and Trend Makers . . . 30
  - 3.3.3 Who trend spotters and trend makers are . . . 33
- 3.4 Predicting Trend Makers and Spotters . . . 35
  - 3.4.1 Regression Models . . . 35
  - 3.4.2 Support Vector Machines (SVM) . . . 38

3.5 Discussion . . . 39
3.5.1 Theoretical Implications . . . 39
3.5.2 Practical Implications . . . 40
3.6 Summary . . . 40

4 Personalizing Trends **43**
4.1 Background . . . 43
4.2 Trend-aware Recommendation . . . 44
4.2.1 Identify trend makers and trend spotters . . . 45
4.2.2 Identify Trends . . . 48
4.2.3 Recommend Trends . . . 48
4.3 Evaluation . . . 49
4.3.1 Classifying users into trend spotter(maker) classes . . . 50
4.3.2 Determining whether an item is a trend or not . . . 50
4.3.3 Recommending trends . . . 51
4.4 Discussion . . . 54
4.5 Summary . . . 56

5 Serendipitous Recommendations **57**
5.1 Background . . . 58
5.2 Our Proposals . . . 58
5.2.1 Basic Algorithms . . . 59
5.2.2 Beyond User History . . . 61
5.2.3 Beyond User Routine . . . 62
5.3 Evaluation . . . 63
5.3.1 Evaluation Metrics . . . 64
5.3.2 Data . . . 65
5.3.3 Validating Modeling Assumptions . . . 65
5.3.4 Accuracy vs. Serendipity . . . 68
5.4 Discussion . . . 72
5.5 Summary . . . 73

6 Conclusion and Future Work **75**
6.1 Thesis Contributions . . . 76
6.2 Future Work . . . 77

Bibliography **81**

Appendix **90**

A Synthèse en Français **91**
A.1 Introduction . . . 91
A.1.1 Motivation . . . 92

A.1.2 Objectifs de la these et contribuations 95
A.1.3 Structure de la these 97
A.2 Qui Crée Les Tendances 97
A.2.1 Identification de Trend Makers et Spotters 97
A.2.2 Caractérisations des Trend Spotters et Trend Makers 98
A.2.3 Prédiction de Trend Makers et Spotters 100
A.2.4 Résumé 102
A.3 Personaliser Les Tendances 102
A.3.1 Trend-aware Recommendation 103
A.3.2 Notre Expérimentes 104
A.3.3 Résumé 107
A.4 Recommendations à La Sérendipité 107
A.4.1 Notre Proposition 108
A.4.2 Evaluation 111
A.4.3 Résumé 113
A.5 Conclusion 114
A.5.1 Contributions 115
A.5.2 Future Work 117

List of Figures

3.1 Screenshot of the mobile application. 24
3.2 (a) Number of uploads per user; (b) Number of votes per user; (c) Distribution comparison on uploads and votes (log-scale on x-axis) 26
3.3 Empirical CDF of the number of votes, likes, and dislikes. 26
3.4 Number of uploads from each country (Top 6) 27
3.5 Empirical CDF of number of countries (cities) from where each user has uploaded, with y-axis representing the cumulative number of users. . . . 27
3.6 Empirical CDF of the number of followers and followees. 27
3.7 Number of followers(followees) and number of uploads per user 28
3.8 Number of followers(followees) and number of votes per user 28
3.9 Empirical CDF of spotter (maker) scores (log-transformed) versus top-n ranked items. 29
3.10 Distributions of features (a-j), trend maker (k) and trend spotter (l) scores with log-transformed values (except for the age feature). The x-axis represents the range of log-transformed features, and the y-axis represents the number of users. 32
3.11 ROC curve of logistic regression and SVM model (S: trend spotters; M: trend makers). 39

4.1 Trend-aware Recommender System 44
4.2 Trend spotter score (log). We split trend spotters into three classes using a proportional 3-interval discretization, as the two vertical lines show. . . . 47
4.3 ROC curve for the logistic regression that predicts whether an item is a trend or not. 51
4.4 Precision and Recall. Results for trend-aware recommender vs. item-based recommender. The size of the recommended list is N. 52
4.5 Precision and Recall. Results for two trend-aware recommenders (item-based and Implicit SVD) and for recommendations of most popular trends. 53
4.6 Number of days an item (*a*) vs. a trend (*b*) receives votes for. 54

5.1 Probability of a user moving at a certain distance. 66
5.2 The distribution of a venue's mixing (i.e., ability to attract all sorts of "user tribes" among the k tribes) by category. Grey bars reflect the overall *average* distribution of a venue's mixing (without any distinction by category). . . 67

5.3 Probability of visiting venues by mixing value. 68
5.4 Percentage of Checkins per Hour of a Day. 69
5.5 Accuracy of the Bayesian model based on LDA 70
5.6 Accuracy and Serendipity of (top 10) Recommendations. This considers all users (i.e., users who visited at least two venues). 71
5.7 Accuracy and Serendipity of (top 10) Recommendations. This considers users who visited at least 5 venues. 72
5.8 Accuracy and Serendipity of (top 10) Recommendations. This considers users who visited at least 10 venues. 72
5.9 Impact of users with different tendencies for social mixing. 72

A.1 La courbe ROC de régression logistique et le modèle de SVM (S: trend spotters; M: trend makers). 102
A.2 Systeme de Recommendation Trend-aware 103
A.3 Précision et Recall. Les resultats de recommandation trend-aware vs. recommandation item-based. Le N signifie le nombre de recommandations. 106
A.4 La précision et le recall. Les resultats de deux recommandations de trend-aware (item-based et Implicit SVD) et celui de recommander des tendances de le plus populaires. 107
A.5 La précision et la sérendipité de (top 10) recommendations. Ils considèrent tous les utilisateurs (i.e., des personnes qui ont visite au moins deux venues). 112
A.6 La précision et la sérendipité de (top 10) recommandations. Ils considèrent des utilisateurs qui ont visité au moins 5 venues. 112
A.7 La précision et la sérendipité de (top 10) recommandations. Ils considèrent des utilisateurs qui ont visité au moins 10 venues. 113
A.8 Des impacts des tendances de social mixing aux niveaux differents. 113

List of Tables

2.1 Four possible outcomes of recommending an item to a user. 18

3.1 Our Hypotheses ($\surd$: accept hypothesis; $\times$: accept the alternative hypothesis; $*$: unknown) 34
3.2 Summary of Kolmogorov-Smirnov test results of our hypotheses. D-values with significance level < 0.05 are highlighted and come with $*$. M, S and T stand for trend makers, trend spotters and typical users. We test a pair of distributions at a time - e.g., for **S** $>$ **T**, we test whether the daily upload distribution for trend spotters is greater than that of typical users, and report the corresponding D-value. 34
3.3 Pearson Correlation coefficients between each pair of predictors. Coefficients greater than ± 0.25 with statistical significant level < 0.05 are marked with a $*$. 36
3.4 Coefficients of the logistic regression. A correlation coefficient within 2 standard errors is considered statistically significant. We highlight and mark them with *. 37
3.5 Coefficients of the linear regression. A correlation coefficient within 2 standard errors is considered statistically significant. We highlight and mark them with *. 37
3.6 AUC and best accuracy of each predictive model. 38

4.1 Coefficients of the logistic regression (a correlation coefficient within 2 standard errors is statistically significant. The significance levels are marked with $*$'s: $p < 0.001(***)$, $p < 0.01(**)$, $p < 0.05(*)$)) 51

5.1 One's unwillingness of traveling far to visit venues of each category. The higher α, the shorter the trip to a venue for a given category. 66
5.2 Mixing of venues per category. 67
5.3 Accuracy and Serendipity of our three basic algorithms. For LDA in the last model, the number of tribes k is set to 100 because of its best accuracy compared to other k's. 70

A.1 Nos Hypothèses ($\surd$: hypothèse accepté; $\times$:hypothèse alternative accepté; $*$: inconnu) 99

A.2 Résumé de les résultats de Kolmogorov-Smirnov test. Les values D avec ses niveaux significatifs < 0.05 sont mis en évidence et sont livrés avec $*$. M, S et T représent les trend makers, spotters et utilisateurs typiques. Nous testons un pair de distributions a la foi - e.g., pour **S** $>$ **T**, nous testons si la distribution de *daily upload* de spotters est plus grande que cela des utilisateurs typiques, et nous rapportons le valeur D correspondant. . . . 99
A.3 Coefficients de régression logistique. Un coefficient de corrélation dans les 2 erreurs standard est considéré comme statistiquement significatif. Nous les soulignons et marquons avec *. 100
A.4 Coefficients de régression linéaire. Un coefficient de corrélation dans les 2 erreurs standard est considéré comme statistiquement significatif. Nous les soulignons et marquons avec *. 101
A.5 AUC et le meilleure précision de chacun modèle de prédiction. 102
A.6 La précision et la sérendipité de nos trois algorithmes de base. Pour LDA dans le dernière modèle, le nombre de user tribes (k) est fixé a 100, en raison de le meilleur précision. 111

Chapter 1

Introduction

That the majority of people are building an online society is an obvious fact.

Social networks are growing beyond being the playground where *connecting* is the only purpose. It is more a phenomenon that we are enthusiastic about. As a phenomenon, we are ready to share all sorts of information (e.g., photos, parties, alimentation, gossips, political discussion, etc.) along our connections. Through social networks, we also organize social events, e.g. meetings, parties, club activities and even political campaigns. In short, we continue to migrate offline activities to social networks.

This "social" phenomenon with its rich collection of our online behavior has certainly attracted lots of research interests. Many different questions were raised, but solutions were sought particularly by understanding aggregated user behavior and information diffusion along network connections. People noticed that when information is propagated along connections, some is more widely adopted than others, and some is spread faster. Moreover, there is a moment that these information reach the critical mass all of a sudden. Such a moment is the so-called "tipping point", defined by Gladwell in [39].

In this Thesis, we present our studies about digital content that triggers the "tipping point" in online social networks. We call such content trending content - *trends* for short - in the following chapters. Specifically, we study the human factors behind the creation of trends, and, we design a system to provide people early discovery of trends that they might be interested in.

1.1 Motivations

Formally, a *trend* is defined as *any form of behavior that develops among a large population and is collectively followed with enthusiasm for some period, generally as a result of the behavior's being perceived as novel in some way* [62]. Two primary properties of trends emerge from the

definition: *broad adoptions* and *temporal effectiveness*. These properties are the key factors that social marketers and researchers have shown great interests in identifying trends.

It is vital to identify trends in social networks, because knowledge about trends can be translated into *event identification*, *opinion spreading*, *brand management*, etc. These concepts are similar to each other in that they are widely adopted and spread fast. To clarify the role that the term "trend" plays in different scenarios, we elaborate with some examples.

- **Event Identification.** *A trend is an event.* An example comes from Twitter.[1] People tweet about what they see or what they encounter. It can be big global events such as the Olympic games, or small local ones such as neighborhood gatherings. In 2009, an Editor in Chief at Mashable - Adam Ostrow, observed that "*earthquakes are one thing you can bet on being covered on Twitter first, because, quite frankly, if the ground is shaking, you're going to tweet about it before it even registers with the USGS and long before it gets reported by the media* [2]". Indeed, that happened for the 2009 Japan earthquake: tweets about the earthquake have travelled around the world much quicker than the official media reports of its occurrence. This has inspired and motivated many researchers to identify real-world events by keeping track of information diffusion in social networks [13, 14, 63, 103].

- **Opinion Spreading.** *A trend is a piece of opinion.* Examples are comments to news threads, reviews to shopping items, and political or societal discussions are all different types of opinions that spread in the Internet. Borge-Holthoefer *et al.* [18] have studied a particular case - the discussion of 15-M movement invoked by economic crisis in Spain in 2011. By collecting Twitter messages (tweets) for one month (approximately from two weeks before the movement till a week after it took place), they noticed that the *"movement-in-the-making had been brewing for a while in the social media"* [18]. Hashtags related to the discussions of camping the Puerta del Sol square in Madrid have been mentioned by tons of tweets, and the underlying "following" and "followers" structure in social networks have pushed their reach of receivers much further. Understanding how opinions are spread is undoubtedly of great importance. In addition to its impact on societal and political opinions, Wu *et al.* [127] have also studied how public opinions form in online voting and review systems.

- **Brand Management.** *A trend is a fashion fad.* Jansen *et al.* proposed to use *microblogging as online word-of-mouth branding* [54]. They stated that in commercial situations, a positive word-of-mouth branding has strong effect on consumers, since it is based on the *trust* built upon social relationships. Coupled with sentiment analysis, their studies of tweets collected for 13 weeks show that user satisfactions with the brands change with time: these changes are correlated with the word-of-mouth spreading. Motivated by such word-of-mouth effect, researchers also proposed to

[1] http://www.twitter.com
[2] http://mashable.com/2009/08/12/japan-earthquake/

leverage social blogs in new product development [36], as well as the early prediction of customers' reactions [17].

Having discussed that various types of "trends", which all receive a burst of attention at a certain time and require people to react fast, a natural question to ask is the following:

How to Detect Trends?

To react fast, we have to be aware of trends sufficiently early. However, capturing a trend before its "tipping point" is hard. Great research efforts have also been spent on *characterizing trends*. Most of the devised solutions are built upon the fact that trends result from *aggregated* user behavior. In other words, the main signal of the birth of a trend is the intensive responses from people. Therefore, many studies have been carried out to study trends from different user reactions, namely:

- **Clicks.** A click, on the Web, is a basic reaction that indicates implicitly our interests. Learning from the aggregated clicks in the video sharing site YouTube,[3] Crane *et al.* [28] identified different patterns of aggregated clicks associated to different types of trends. Relying on these clicks patterns, one can tell exogenous trends (those triggered by factors external to the site, e.g. reporting of a piece of news on TV) from endogenous ones (those triggered by internal factors, e.g. spreading of a piece of news within the site).

- **Posts/Retweets.** Posts or retweets activities are more explicit and proactive than clicks. They show explicitly one's willing to spread the content. Crawling trending topics from Twitter and associating them with related tweets, Nikolov [81] proposed a statistical nonparametric classification method to capture trending topics by learning their time series pattern of tweeting rate.

- **Content.** Direct analyses about the content of trends are also studied with a certain granularity. An example of exogenous trends as news disseminating in Twitter, is that these tweets often contains an url pointing to the external media site where the news was reported [77].

- **Social Connections.** "Following" and "followers" are the fundamental functionality provided in any social networks. Thanks to these social links, information flows from one to another. Intuitively, people who establish many links have better chances to propagate content to others; this is the case that generates endogenous trends such as celebrities' gossips [77, 130].

Knowing the characteristics of trends is of great help to *detect* trends. But, unveiling how do trends evolve with human dynamics can provide people (e.g. social marketers) the

[3] http://www.youtube.com

knowledge of who *create* trends. To this end, searching "influentials" within the network becomes the central theme.

Who are Influentials?

The fundamental theory of "influentials" goes back to the *two-step flow* paradigm proposed by Katz and Lazarsfeld in 1955 [64], which was originally formulated to understand how public opinions form. It says that the information diffusion cascade is *"a process of the moving of information from the media to opinion leaders, and influence moving from opinion leaders to their followers"* [20].

For decades, the two-step flow theory had been dominant in the research of information diffusion processes. Their definition of *opinion leaders* has been well accepted and later on also adopted as the definition of *influentials* [73]. That is, influentials are those *individuals who are likely to influence other persons in their immediate environment* [64].

Modern studies about influentials (especially with the easy access to information diffusion traces in online social networks) have developed two different opinions.

- **Influentials are special individuals.** Adhering to the two-step flow theory, researchers in this group believe that influential individuals are different from the crowd to some extent. Gladwell in his book *The Tipping Point* states that *"the success of any kind of social epidemic is heavily dependent on the involvement of people with a particular and rare set of social gifts"* [39]. He identifies three actors as special individuals who created social epidemics (trends), that is *connectors* (those who know many people in the community), *mavens* (information specialities) and *salesman* (persuaders). By modeling and analyzing information diffusion in social networks, researchers have confirmed the existences of these different types of special individuals who are able to spot trends early on [64, 101]. These special individuals are socially well connected (connectors) [57]; are able to easily influence others (salesman) [45]; are considered to be experts (mavens) [113, 125]; or are celebrities [130].

- **Influentials can be anyone.** Duncan Watts claims that being influential is *mostly an accident of location and timing* [10, 121]. It is a matter of adopting correct opinions at a correct moment, regardless of who you are. Also, he stated that the *influentials* are not necessarily *"head of formal organizations, nor public figures such as news paper columnists, critics, or media personalities, whose influence is exerted indirectly via organized media or authority structures"* [122]. To highlight the concepts of *unexpectedness* and *unplanned*, the individuals who are involved in the diffusion of trends are then called as "accidental influentials" [121].

We have seen that identifying trends can be translated to event identification, opinion spreading, brand management, etc. A variety of studies of trends are performed with two themes. That is, 1) what are trends; 2) who creates trends. We have given a brief overview about how people try to detect trends by their characteristics, as well as the debate on

whether the individuals who create trends are special (more detailed background and related work please refer to Chapter 2). Next, we define our research scope of this Thesis and position our contributions with respect to the literature on trend exploration and the associated tools at our disposal.

1.2 Research Problems and Contributions

We have stated that trends result from aggregated user behavior. They are pieces of information that are disseminated in the network and obtained a wide coverage of adopters. Unsurprisingly, the notion of *diffusion* within the network is the focus. However, the complete process of the birth of a trend should also include the creation of the information itself. It is an undeniable fact that people who create the information (who initially bring it into the network) are an important filter to the information from the outside of the network.

Considering both the creation and the diffusion process of trends, there are some questions that still need a clear answer. These questions are raised from three main aspects - human factors behind trends, identifying trends and exploring them.

- **Human Factors.** In spite of the debate on influentials, the human dynamics of individuals who create trends (no matter whether they are special) are still unclear. Considering both the individuals that originally bring the information to the network and the ones that spread them, what are their characteristics? Do they share any common traits and what are their differences?

- **Identification.** Knowing the characteristics of people who create trends, is it possible to identify trends by leveraging their knowledge? To which extent trends could be identified accurately as such?

- **Exploration.** Suppose that we are able to precisely identify trends. How can we build upon this ability to help users discover the trends of their interests? To the end of providing such personalized content exploration, how to guarantee the quality of the personalizations?

In this Thesis, we are going to tackle these questions in steps. In the course of seeking the answers, we make the following contributions:

- We approach the analysis of who creates trends by defining two distinct classes of individuals: trend spotters (those who rate items before they become trends) and trend makers (those who upload items that become trends). We characterize them by combining multiple characteristics including their activity, content, network and geographical features. We find that trend spotters and trend makers differ from typical users, in that, they are more active, show interest in a variety of items, and attract social connections. We then study what differentiates trend spotters from

trend makers. We learn that successful trend spotters are early adopters who hold interests in very diverse items, while successful trend makers are individuals of any age who focus on specific types of items (Chapter 3).

- Using linear regression, we predict the extent to which one is a trend spotter or trend maker. Then, with an existing machine learning algorithm (SVM) and with a logistic regression, we perform a binary classification of whether one is likely to be a trend spotter (trend maker) or not. While linear regression has produced informative results, SVM and logistic regression have returned accurate predictions (Chapter 3).

- We propose a method that detects trends by relying on the activities of two types of users: trend makers and trend spotters. We then construct a preference matrix based on the identified trends, and test the extent to which a state-of-the-art matrix factorization algorithm (*Implicit SVD* [51]) effectively recommends trends (Chapter 4).

- Going beyond the goal of making accurate recommendations, we explore the possibility to enrich serendipity in final recommendations by leveraging network analysis techniques, and validate our proposals in the context of a location-based mobile recommender system. To be precise, we tackle the possibility of introducing serendipity by promoting places that go beyond those that would be recommended based on past visited places and on one's typical routine. We quantitatively evaluate to which extent we are able to introduce serendipity without compromising the accuracy of the recommendations upon the real-world dataset (Chapter 5).

1.3 Organization of the thesis

Chapter 1 has spelled out our research problems.

Chapter 2 gives the background of our work from two principal related research directions, i.e., trends in social medias and recommender systems.

Chapter 3 differentiates trend makers and trend spotters from typical users, characterizes them and experimentally shows that they can be accurately predicted with a variety of features.

Chapter 4 proposes a recommender system to satisfy people with personalized trending contents.

Chapter 5 leverages network analysis techniques to introduce serendipity into recommendations.

Chapter 6 concludes our research work and summarizes our contributions to the state-of-the-art.

CHAPTER 2

Background and Related Work

The concept of *trend* is tied to abrupt spikes in the attention toward a specific item or concept. Such a property underlines how the term of *trend* can be generalized to a variety of related concepts in the literature, e.g. events, opinions, topics, social memes, etc. This Chapter aims at providing an overview on the current state of the art on trend modeling and identification.

We have motivated our research work in the previous Chapter through its impact on event identification, opinion spreading and brand management. In this Chapter, We first provide a broad overview of the literature related to the concept of trends (Section 2.1). We identify two main approaches to the problem of trend identification and analysis: approaches to identify spikes of interest inherent of the nature of a trend, and approaches to study their overall characteristics. Finally, trends are produced by aggregated user activities, and the human role in the generation of trends has been the subject of debate in the scientific community. In Section 2.1.3 we will review the debate on the role of "influential" users in the generation of trends.

One of our main research objectives of this work is to help users discover and consume trends. After having studied and understood the mechanism behind the generation of trends, we will therefore focus on the problem of identifying trends of one's interests. This type of problem is analogous to that of making personalized recommendations. In Chapter 4, we study how to recommend personalized trends by leveraging the power of the crowds by using collaborative filtering techniques. We review in this Chapter (Section 2.2) the current state of the art in collaborative filtering techniques developed in recommender systems and their applications, tools that will be later used as building blocks to our work.

2.1 Trends in Social Media

The general topic of studying trends in social media has received considerable interests by the research community, motivated by the demand of understanding viral marketing, opinion spreading, and event and topic identifications. We review relevant studies on trends from two aspects: 1) how to identify trends; 2) what are the characteristics of trends.

2.1.1 Identification

The sharp increase in user interest is generally regarded as a signal for identifying trends. In the literature, capturing such bursts of interests is one of the main research approaches to the problem of trend identification, and most related work is built upon time series analysis and modeling. Depending on the types of content, different approaches are proposed to identify bursts in social media. For instance, text mining techniques have been widely explored when the content is text, e.g. news streams. When dealing with other types of content such as pictures and videos (whose content is more complex and costly to mine), the identification of trends relies instead more often on the dynamics of the user activities or interactions (e.g. posting, replying, forwarding, viewing, commenting, etc.).

Text mining

Text mining techniques are commonly applied in detecting and tracking emerging topics in news streams. Mostly, they are built upon word segmentation and topic modeling. To identify trends in general topics, Kleinberg developed a framework to describe the time stream of the frequency of words with a finite state automaton, in which the bursts could be signaled at the state transitions [58]. His further analyses on the burst patterns of the terms reveal that the mixture of these "trending" terms form a latent hierarchical structure that has a meaning - that is, a topic. Therefore, the detection of trending terms could contribute to the identification of trending topics. Kleinberg's model describes the temporal change of co-occurrence of words, which can be viewed as the topic change over time. Instead, the approach proposed by Wang and McCallum assumes that topic itself does not change (i.e., the term mixture of a topic), while the topic co-occurrence patterns of documents change over time [119]. With a different granularity of pattern mining over time, their approach exhibits a better performance on the trending topic identification.

Both of the previous approaches focus on mining trending topic from a single text stream. But in some cases, different data streams may cover the same topics. Such situation is commonly seen in the applications of news media. When a major event happens, the same news could be reported by multiple news agencies, and thus are disseminated through multiple news streams. Wang *et al.* discover that the bursts of related topics from different media triggered by the same events are correlated from the temporal as-

pect [120]. By mining such *correlated bursty pattern*, the authors show that it is possible to identify global trending topics across multiple news streams regardless the language used in the stream. Moreover, such bursty pattern could tell the local trends apart from the global ones.

While these models have been successful at identifying trending topics in news media, in a larger scale, Leskovec *et al.* studied the information dissemination of news cycling, in which information does not stay "locked" within its news media, but gets propagated to social medias like blogs through the user interactions among different web services [66]. By quantitatively analyzing millions of articles collected from over one million media sites and blogs, they found that there exist competition among individual "memes" (i.e., trending topics in the news streams) to become the trends. The fact that such competition occurs is likely associated with another findings from their studies, which states that different news agencies are very close to each other on what to report and when to publish. On the level of local news trends, the authors observed that the volume of attention decreases exponentially in both direction from the peak of its bursts (i.e., both prior to and after the peak). Additionally, another notable phenomenon they found is that for the same story which becomes a trend, the time it gets trendy in blogosphere is in average 2.5 hours later than in news media stream.

Overall, trends and topic identification in news and blogosphere has been widely studied by text mining. Diverse probabilistic models have been built above the mixture of terms to successfully capture the temporal dynamics of the trends [3, 15, 40, 47, 58, 119, 120]. When applying similar analysis to social media content, because of the limited size of the content (e.g., Twitter), the approach to detect bursts usually can be simplified and applied to the identification of terms that appear in a certain time period much more frequently than expected [14, 77, 78, 90].

User activity and interactions

The approaches to identify trends based on text mining work well when the entity of trends are textual content, e.g., news and blogs. However, the Internet enables increasingly rich approaches for sharing information that go beyond simple text. Mining information out of an image, or a video, can be extremely expensive and this renders the application of similar mining approaches to these formats more costly. Especially when analyzing the heavily used social medias, the identification of trends may need to build upon different types of information such as the way in which the mass of users respond to them.

When looking at the dynamics of the aggregated user activities and at the response time, researchers have documented two opposing behaviors. The reaction to trends has in fact been shown to be either completely random, or highly correlated with the activities of others [12, 46, 118]. When looking at the latter case, researchers have distinguished between trends generated by highly correlated user activities *within* the user community

and those that are resulted from other factors *outside* the community. This has led to the respective definition of endogenous and exogenous trends [28, 108].

To sort out the bursts of user activities of different types (i.e., endogenous or exogenous trends), Crane and Sornette have analyzed the time series of daily viewing patterns in YouTube [1] (an online video sharing service) [28]. Their studies reveal that the distribution of the waiting time before the user's response to the videos is sufficient to describe the different burst patterns for endogenous and exogenous trends. More precisely, in the case of endogenous trends the burst of attention is preceded by a smooth increase, associated to the gradual spreading along the social connections within the community. In the case of exogenous trends, instead, the burst happens shortly after the upload of the videos. These immediate peaks of attention are triggered by factors that are external to the social media, e.g., reporting of a piece of news on TV, and they thus bypass the social interactions. Similarly, to capture the trending hashtag/topic in Twitter, Nikolov *et al.* proposed a non-parametric classification algorithm to learn the temporal pattern for the tweeting rate of trending hashtags/topics, and succeeded to capture the trends approximately half an hour before the topics were shown as trends on Twitter [81].

In addition to the work on identifying trends by catching the bursts of user activities, researchers have attempted to understand the descriptive and comparative characteristics of trends. Such studies were widely done in the context of micro-blogging social media like Twitter.

2.1.2 Characterization

Rather than focusing on the identification of trends, a parallel branch of research has focused on leveraging trends to extract insights from social media (e.g. predicting large-scale events) [14, 77, 130]. In general, researchers have attempted to gather a better understanding of trends by means of descriptive and comparative analyses, often focusing on trends or content related to real-life events in Twitter.

Generic studies on trends

People tend to tweet about real-world events prior to the traditional news media [63, 89, 103]; this fact led several research efforts to attempt to analyze the trends and events identified in Twitter. To estimate the location of an earthquake or the trajectory of typhoon, Sakaki *et al.* have studied social, spatial and temporal characteristics of earthquake/typhoon related tweets [103]. By analyzing the early messages associated with an event, Petrovic *et al.* [89] discovered that the number of users who tweet about the event is more indicative than the volume of the tweets written about the event.

[1] www.youtube.com

Looking at local news events, Yardi *et al.* [129] studied the characteristics of messages related to them, and those of the users who posted them. They found that active users who are in the center of the online network are more likely physically centered around the local events. They also found that the local networks are denser than the global one, so that local news sources and the people who witnessed the local events are more efficient at spreading the events.

Focusing instead on a more global scenario, that of the trending topics in Twitter, Kwak *et al.* [63] have analyzed tweets of top trending topics from the temporal behavior and user participation. These trending topics were mostly news headlines and user response to fresh news, and were found to be active for durations of a week or shorter. Long-lasting trending topics did not always have new users joining the discussion. Similar temporal characteristics were found in [9] as well. Asur *et al.* discovered that trends in Twitter were determined by the retweets from other users instead of users who posted them originally, and were more related to their content instead of the characteristics of users [9].

Studies on categorized trends

As we have reviewed in the previous section, in a social media site, there are two types of trends - exogenous (if trends are caused by factors external to the social media) and endogenous (if trends are created because of factors within the social media). These two types of trends have been shown to exhibit different temporal patterns of waiting time (i.e., the duration of time before user respond to the content) [28]. Exogenous and endogenous trends with their distinguishable temporal patterns were also observed in Twitter [63]. Among all the identified trends, the authors found larger percentage of exogenous trends (e.g., headline news) than endogenous ones. Focusing on Twitter data from a metropolitan area (New York City), Naaman *et al.* separated trends into different groups by refining the original exogenous and endogenous categories [77]. Their work suggested that even within the same category of trends, different types of trends existed and could be distinguished from a rich set of features - content, user interactions and social networks [77]. Based on the differences lying in these features for different types of trends, Becker *et al.* then used clustering techniques to distinguish real world events from non-events messages in Twitter [14].

While a large percentage of trends in Twitter were found to be exogenous, and were news stories in particular, Yu *et al.* have explored the trends in weibo (Twitter-alike service in China), [2] and found trends in China were mostly created due to the retweet of content such as jokes, images and videos, and were thus mostly endogenous [130].

[2] http://www.weibo.com/

2.1.3 Influentials

In addition to trend analysis and identification, researchers have also looked at the individuals behind them and have tried to investigate the process of the generation of a trend. Research efforts to answer this question intersected with the studies on opinion formation and information dissemination in social networks. Mainly, there are three different views on the trend creation process.

Special individuals

The first vision sees trends as *generated by influentials*. In his popular book "The Tipping Point", Malcolm Gladwell argued that the creators belong to the "special few", and are often called "influentials" [39]. These influentials are found to be special kinds of individuals who: are able to spot trends early on [39, 64, 101]; are socially well connected [57]; are able to easily influence others [45]; are considered to be experts [113, 125]; or are celebrities [130].

Accidental influential

The second view on trend creation sees trends as *generated by coincidences*: anyone can be influential. As a result, what becomes popular in a network does not depend on the initiators and is thus an accidental process. Duncan Watts uses the terms "accidental influentials" as he considers social epidemics to be "mostly an accident of location and timing" [121], and ideas spread and ultimately become popular only if there is societal willingness to accept them.

Combined process

Lately, researchers have found that there are different classes of individuals who contribute to two parallel processes: early participants start contributing and thus create random seeding, and that contribution spreads then through low threshold individuals [8, 44]. Based on this recent literature which has focused on the two parallel processes, our work in this Thesis will take a close look at the individuals who contribute to those processes.

2.2 Recommender Systems

Good user experience is what makes online services enticing. To be outstanding, many services not only try to provide easy access to the content of what users are looking for,

but also attempt to help them discover new information which they might be interested in. Mostly, these services use recommender systems to give personalized suggestions to each user [99, 100]. Depending on the context of the applications, recommended items are of all kinds - e.g., books, CDs [69], movies [5, 135], videos [32], news [31], music [133], events [95], places [134], search keywords [67], social connections [23] etc. However, the fundamental idea of all these recommender systems is the same, that is, to seek relevant items into one's preferences. In general, two essential components are needed to construct a recommender system: 1) user preferences; and 2) algorithms.

2.2.1 User Preferences

To make personalized recommendations that are highly likely to be accepted by the end user, understanding one's preferences is evidently of great importance. Depending on the applications, user preferences can be collected in two ways: through explicit or implicit ratings.

Explicit Ratings

To obtain explicit feedbacks, users are asked to rate items on a Likert scale (e.g., on a scale of 1 to 5 points) depending on the degree of their preferences. For example, the online e-commerce platform Amazon [3] collects customer reviews about products on the scale of one to five stars [69], and similar Likert scaled feedbacks are also collected in movie recommender system MovieLens [74]. But, the user self-expressed ratings were found not as robust to quantify their real preferences as expected [6]. Users might report inconsistent ratings because of the impacts of environment, and thus introduce noises into the user preferences [6, 55, 83]. To reduce such inconsistencies, some services simplify the explicit rating to a single "thumb up" if a user likes an item, e.g., YouTube [32].

Implicit Feedbacks

Explicit ratings are not always available in all the applications. When they are missing, an alternative to infer one's preference is to extract his/her implicit feedbacks. Implicit ratings could be obtained by measuring different user behaviors [82, 114], depending on the items to recommend. Binary implicit ratings only formulate whether one likes an item, and the user behavior to indicate such "likes" include a purchase [69, 97], a click [114], the fact of joining a community/group [24], etc. The numerical values of explicit feedbacks requires finer grained information about user behaviors, which often tend to describe the frequency of actions [51]. For instance, it could be how much time the user watched a certain show [51], how often a user listens to an album [87], etc.

[3] www.amazon.com

Implicit preferences inferred from user behavior are inevitably noisy, but they provide the confidence of the fact the users like an item [51]. For instance, by performing user studies in the context of music consumption, the time one listens to an album is shown to be clearly correlated with the explicit ratings from the user [87].

2.2.2 Collaborative Filtering

There are a variety of algorithms designed for recommender systems of all kinds. In the literature, Adomavicius *et al.* [2] provided a comprehensive overview about the state of the art of the algorithms by grouping them into content-based, collaborative filtering and hybrid approaches, as well as the pros and cons of each group, while Su *et al.* [110] conducted a survey dedicated to collaborative filtering techniques in particular. Instead of giving yet another overall review of these diverse algorithms, in this section, we will focus on two notable collaborative filtering algorithms (i.e., item-based and SVD) that are often used as baseline in the field, which will be also applied later on in our work (will be presented in Chapter 4 and Chapter 5).

As classified in [2], there are three groups of algorithms designed for recommender systems - content-based methods, collaborative filtering techniques and hybrid approaches. Among them, collaborative filtering techniques are the most successful ones [110]. Collaborative filtering is a term coined from the first recommender system Tapestry [42], and is meant to "helping people help each other" [115]. Its main idea is to leverage the "the wisdom of crowds" [111] and recommend items that people with similar tastes and preferences liked in the past [11]. A typical collaborative filtering based recommender system takes the user preferences ratings (explicit ratings or implicit ones inferred from user behaviors) as input, and outputs/predicts the missing preferences (i.e., ratings on not yet consumed items). Items are then sorted according to the predicted ratings, and top N ranked items are returned as recommendations (known as Top-N Recommendations). Depending on the way to predict the ratings, two types of collaborative filtering techniques are spotted - memory-based and model-based approaches.

Memory-based

Memory-based approaches [19, 34, 35, 65, 98] predict ratings based on the entire collection of previously rated items by the users [2]. The most common memory-based approaches are *k-nearest neighbor* models, and the original model is user-based [48] approach. Such user oriented approach tries to predict missing ratings based on the ratings from like-minded users. Due to its better scalability and more accurate predictions, an analogous approach but item oriented (known as item-based approach [104]) are better adopted in practice than user-based. The item-based approach makes recommendations in two steps: 1) computing similarity between items; 2) predicting missing ratings by aggregating user preferences on similar items. Items are then sorted by their predicted

ratings in the descending order. The top N ranked items are commonly output as final recommendations.

Computing Similarity. Similarity computation is the critical step to find the *k-nearest neighbors* of the items that one likes. The similarity between each pair of the items should be examined, and it could be determined by user preference ratings. There are a number of ways to compute the similarity $s_{i,j}$ between item i and item j. Popular ones include cosine-based, correlation-based and adjusted cosine similarity.

User preferences ratings are commonly used to construct a preference matrix, in which each row corresponds to a user, and each column is an item. Each element of the matrix describes one's preference towards the corresponding item. Under such formulation, item i and j can be thought as two vectors in the user space. Therefore, the cosine of the angle between these two vectors can be viewed as their similarity $s_{i,j}$:

$$s_{i,j} = cosine(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{||\vec{i}||_2 \times ||\vec{j}||_2} \tag{2.1}$$

Another widely adopted method to measure similarity is the *Pearson Correlation Coefficient*, which measures the degree of linearity on the intersection of the pair of item profiles. To make the correlation computation accurate, it's better to focus on ratings from the set of users ($\mathcal{U}$) who co-rated both of the item i and j [104]. Then, the correlation-based similarity could be computed as:

$$s_{i,j} = \frac{\sum_{u \in \mathcal{U}} (r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in \mathcal{U}} (r_{u,i} - \bar{r}_i)^2} \sqrt{\sum_{u \in \mathcal{U}} (r_{u,j} - \bar{r}_j)^2}} \tag{2.2}$$

where $r_{u,i}$ is the rating user u gave to item i, and $\bar{r}_i$ is the average rating item i received.

Counting that users may have different rating scales - some may prefer to give neutral feedbacks then extreme "likes" or "dislikes", an adjusted cosine similarity measurement is also proposed to offset such situation by subtracting the corresponding user average rating from each co-rated pair [104]:

$$s_{i,j} = \frac{\sum_{u \in \mathcal{U}} (r_{u,i} - \bar{r}_u)(r_{u,j} - \bar{r}_u)}{\sqrt{\sum_{u \in \mathcal{U}} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{u \in \mathcal{U}} (r_{u,j} - \bar{r}_u)^2}} \tag{2.3}$$

in which $r_{u,i}$ is the rating user u gave to item i, and $\bar{r}_u$ is the average rating that user u used to give.

There are also other ways to compute similarities (e.g., Jaccard similarity, Euclidean distance, etc). Different similarity measurement metrics may lead to different effects in recommendations, and one may outperform another in different datasets.

Predicting Ratings. Once the most similar items are identified, we need to predict ratings of unrated items for each user. To predict the rating $r_{u,k}$ that user u might give to unrated

item k, the most common way is to aggregate u's ratings towards the set of similar items ($\mathcal{N}$) of k. And, the rating $\hat{r}_{u,k}$ could be predicted as the weighted average of ratings on nearest neighbors:

$$\hat{r}_{u,k} = \frac{\sum_{i \in \mathcal{N}}(s_{i,k} \times r_{u,i})}{\sum_{i \in \mathcal{N}}(|s_{i,k}|)} \tag{2.4}$$

Items are then sorted according to their predicted ratings, and the Top-N are returned as recommendations. With item-based approach, what user receives as recommendations are the items that are similar to what they ever liked. The easy explanation of why one receives the list of recommendations as such, has also contributed to its success.

Model-based

In contrast to memory-based models which are based on the entire user ratings, model-based approaches use a sample dataset constructed from a subset collection of ratings. Typically, for these algorithms, a learning phase is dedicated to learn user preference/rating models, which then are used to make predictions about user preferences on unrated items. In the literature, research efforts tried to model the user preferences using data mining or machine learning algorithms [110].

One of the problems that recommender systems face is data sparsity. That is, the number of items rated by each user is always a small portion with respect to the full set of items in the application. The unrated user-item combination results is a sparse preference matrix. A prominent algorithm to address such problem is based on matrix factorization, which is often built upon dimensional reduction techniques - such as Singular Value Decomposition (SVD) [105].

As addressed in [4], the key of an SVD decomposition of a matrix is to find a new lower dimensional feature space, in which each feature represents a "concept" and the importance of each of the concept is eligible to be computed. Applying SVD decomposition on a preference matrix ($\mathcal{R}$) with n users and m items in a recommender system, is to find two descriptive matrices $\mathcal{U}$ ($n \times r$) and $\mathcal{V}$ ($r \times m$) for a given number of new features r, that can be used to approximate the original preference matrix in a lower dimensional feature space:

$$\mathcal{R} = \mathcal{U}\lambda\mathcal{V}^T \tag{2.5}$$

in which λ is a diagonal matrix that contains singular values (which represent the semi-axes of the r-dimensional ellipsoid of the "concept" space). The $\mathcal{U}$ matrix can be interpreted as the "user to concept" similarity matrix, while the $\mathcal{V}$ matrix is the "item to concept" similarity matrix.

By uncovering the user-item latent relationships with "concepts", there are two different ways to use the decomposed matrices in making recommendations [105].

- In the low dimensional feature space, one could measure similarity between each pair of users (or items) to identify the k-nearest neighbors in that reduced space. Therefore, memory-based recommender systems can be further applied.

- Relying on the decomposed matrices, the ratings for a user to an item can be described as the dot product between the user's feature vector ($\mathcal{U}$) and the item's feature vector ($\mathcal{V}$). In other words, for a user u and item i, the predicted rating $\hat{r}_{u,i}$ is:

$$\hat{r}_{u,i} = \sum_{f=0}^{r} \mathcal{U}_{u,f} \times \mathcal{V}_{f,i} \tag{2.6}$$

There are other matrix factorization techniques (e.g., Principle Component Analysis - PCA) [43], and their different variants such as the Non-negative Matrix Factorization have also been used in the literature [128]. These algorithms are similar to SVD in the sense that they all aim to decompose the ratings matrix into two matrices, one of which contains features that describe the users and the other contains features describing the items.

We have seen that the basic process to build a recommender system requires two steps: 1) extracting user preferences from their explicit ratings or by inferred from user behaviors; 2) choosing an algorithm to predict user preferences on unrated items. The decision about which algorithm to use is difficult to make without any performance metrics to optimize for. Next, we take a brief look at how recommender systems are evaluated.

2.2.3 Evaluation

The performance of recommender systems is difficult to evaluate, because 1) an algorithm may perform differently depending on the datasets; 2) the goal that a recommender system is expected to achieve differs from one application to another [49].

The main stream of algorithms handle the task of making personalized recommendations as solving the problem of predicting ratings. To evaluate how well an algorithm is able to make such predictions, an experimental dataset is often required and it gets split into two parts - a training and a testing subset. While the algorithm learns user preferences from the training set, it tries to predict the withheld preferences in the test set. The most common way to quantify its power of prediction is that of accuracy metrics.

Accuracy

Prediction accuracy is the most commonly measured quality of a recommender system. Based on the different focuses, there are three classes of accuracy metrics: 1) measuring the accuracy of predicted numeric ratings; 2) measuring the accuracy of binary preferences (e.g., whether one performed an activity); 3) measuring the ranking of the items

in the recommendation list [107]. Now, we present three popular accuracy metrics from each class of the accuracy measurement respectively.

MAE and RMSE. As we have discussed in Section 2.2.1, some applications require users to rate items on a Likert scale (e.g., from 1 to 5 points). In such cases, to evaluate the predicted numeric ratings, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are probably the most popular metrics. In a test set $\mathcal{T}$, if the withheld rating $r_{u,i}$ from user u to item i is predicted as $\hat{r}_{u,i}$, the accuracy of the algorithm under evaluation could be computed by MAE and RMSE respectively as below:

$$MAE = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} |\hat{r}_{u,i} - r_{u,i}| \tag{2.7}$$

$$RMSE = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} (\hat{r}_{u,i} - r_{u,i})^2} \tag{2.8}$$

Both of these two metrics focus on the numeric difference between the real rating and its predicted value. However, if using these two metrics to compare different algorithms, RMSE favors the algorithm whose predictions are all associated to small errors, while MAE metric gives preference to algorithms with minimal errors in most cases tolerating a certain number of predictions with large errors [49]. A recommender system with accurate predictions evaluated with RMSE would make general acceptable recommendations to all the users. The one with accurate predictions evaluated with MAE would give perfect relevant recommendations in most cases, but it is likely to make extremely incorrect recommendations sometimes.

The consequence is that an accurate recommender system yielded from the comparison on RMSE would make generally acceptable recommendations to most of users, while the one outperform in the terms of MAE would give perfect relevant recommendations in most cases, but might also makes extremely incorrect recommendations.

Precision and Recall. There are many applications, where user preferences are not explicit ratings on a numeric scale, but are inferred as binary from user behaviors (e.g., 1 if the user made a purchase, 0 otherwise). In these applications, the objective of a recommender algorithm could be thought of as to predict whether one might perform an activity. On top of the withheld facts (i.e., hidden information on the fact that one has purchased a certain item) in the test set, the outcome of a recommendation from such binary predictions would fall into four cases as shown in Table 2.1.

	Recommended	**Not Recommended**
Adopted	True Positive (TP)	False Negative (FN)
Not Adopted	False Positive (FP)	True Negative (TN)

Table 2.1: Four possible outcomes of recommending an item to a user.

Counting these possible outcomes, the overall accuracy of the algorithm could be quantified by the precision and recall metrics defined as following:

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{2.9}$$

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{2.10}$$

It should be noticed that precision and recall metrics are dependent from the number of recommendations the user receives, that is, they depend on the size of the recommendation list. So, while comparing different algorithms using these two metrics, it's often preferred to control the size of recommendations, and to measure the accuracy using precision/recall at N (i.e., top-N recommendations).

Percentile Ranking. Predicted ratings are not the only property to reflect the accuracy of recommendations. The position of the relevant items (i.e., the recommended item that was adopted by the user) in the recommendation list also matters. An accurate algorithm ranks the most relevant items in the top tier of the recommendation list. One of the metric to address such fact is percentile ranking [51].

Before the final output of top-N recommendations, all the unrated items are sorted according to their predicted ratings. In such ordered list of items, each item i receives its percentile ranking $\mathrm{rank}_{i,u}$ as:

$$\mathrm{rank}_{i,u} = \frac{index\ of\ item\ i\ in\ the\ ordered\ list}{size\ of\ the\ ordered\ list} \tag{2.11}$$

In this way, $\mathrm{rank}_{i,u} = 0$ would mean that item i is predicted to be the most desirable for user u. Taking consideration of the withheld rating $r_{u,i}$ from user u to item i in the test set $\mathcal{T}$, the overall accuracy of the algorithm could be quantified using such percentile ranking as following:

$$\overline{\mathrm{rank}} = \frac{\sum_{i,u \in \mathcal{T}} r_{i,u} \times \mathrm{rank}_{i,u}}{\sum_{i,u \in \mathcal{T}} r_{i,u}} \tag{2.12}$$

Invariably from the size of the recommendation list, a lower $\overline{\mathrm{rank}}$ in such metric tells a better accurate recommender system.

Depending on the applications of recommendations, different metrics should be chosen carefully to measure the accuracy. In the literature, the prediction accuracy has been regarded as the major property of a recommender system. Such importance is built upon the assumption that an accurate prediction is a good recommendation. However, prediction accuracy itself is insufficient to explain one's decision to adopt a recommendation [107], thus is not enough to conclude about the quality of recommendations. Complementary evaluations on other properties of recommender system should be explored and investigated.

Beyond Accuracy

In [72], McNee *et al.* pointed out that accuracy is not enough to describe the quality of a recommender system, and that the focus of accuracy may start to hurt user experiences. They underlined how recommendations shown to be accurate according to the algorithmic metric sometimes are not useful to users. To tackle this problem, they proposed to shift the attention to user-centric recommendations.

Various properties of recommendations then have been taken into consideration [107], e.g., *novelty* [60], *serendipity* [76, 133], *diversity* [131, 137], etc.

Novelty. *Novel recommendations are the recommended items that users did not know about* [60, 107]. Based on such definition, a direct way to evaluate whether the recommended item is new to the user is to perform a user study [21, 56]. In the offline experiment setting, the evaluation could be done by splitting the dataset into training and testing subsets along the time, simulating one's knowledge about items [107].

Serendipity. *A serendipitous recommendation is an unexpected (or surprising) recommendations that users do enjoy*. Quantifying such "surprises" is challenging. An early attempt from Murakami *et al.* thought serendipity of recommendations as deviation from a "nature" prediction [76]. And Zhang *et al.* has tried to quantify the serendipity as the amount of information relevant but new to the user in recommendations [132].

Diversity. *A diverse recommendation list contains items that are very different from one to another*. It could be measured as the new item's diversity from the items already in the list, which is commonly quantified using distance metrics (i.e., the opposite of similarity computation in item-based approach) [137].

While the concepts of these new evaluation aspects of recommendations remain the same across various applications, their measurement metrics are often tailored according to the context. Moreover, the exploration of techniques to improve the recommendations from these new perspectives are still in its infancy.

2.3 Summary

In this Chapter, in the social media setting, we have introduced various approaches developed in the community to detect trends based on their nature of bursts of interests, as well as the characteristics of identified trends of different kinds. From the perspective of human factors, we also discussed the debate on the role of "influentials" in the creation of trends.

In addition to the background on trends in social media, we have also presented recommender systems - the tools to personalize user content. We see that to build a recommender system, the common practice includes: extracting user preferences on items;

choosing a suitable algorithm to predict one's preferences about unrated items; and evaluating the quality of recommendations. The development of recommender system faces a number of challenges, and the quality of recommendations beyond the accuracy is in particular at demand.

Based on the background and related work in these two different fields, we notice that the studies about "influentials" in the creation of trends have been focused on their power of influencing others to adopt an idea or an item, while the dynamics of their diverse online behaviors in a social media are still unclear. Moreover, it is yet unclear to which extent the knowledge of these people could benefit the identification of trends, and the user consumption of trending content in a social media. In this Thesis, our work aims at tackling these questions by: 1) exploring the characteristics of the people who create trends in a social media; 2) designing a recommender system dedicated to facilitate users discover trending content of their interests; 3) proposing a few approaches to improve the quality of recommendation by introducing serendipity.

CHAPTER 3

Who Creates Trends

We have seen that in the literature (Chapter 2), media marketers and researchers have shown great interests in what becomes a trend within social media sites. Their interests have focused on analyzing the items that become trends. In this Chapter, We will focus on people rather than items. Research efforts about people who create trends in social media sites have been focused on their power of influencing others to adopt an idea or an item. We go beyond the ability of influence, and refine the roles of these individuals with two classes of users - trend makers (those who generate trends) and trend spotters (those who spread them).

First, we introduce the mobile social-network application used in our study in Section 3.1. Then, we provide an overview about the dataset collected from the application in Section 3.2. Our analysis on trend makers and trend spotters unfolds in two steps: in Section 3.3, we perform a comparative hypotheses analysis to characterize trend makers and trend spotters from a variety of features; and in Section 3.4, we build statistical models based on a set of selected features to predict which users, through their activities, will become trend makers and trend spotters. This work reveals the different types of users involved in the creation of trends. This brings some insights in the literature of opinion spreading and practical implications in the design of recommender systems, which will be discussed in Section 3.5.

3.1 Background on the Mobile Social Application

The application under study (i.e., *iCoolhunt*[1]) is a social application with a mobile phone client in which users can share pictures of "cool items". Users upload photos of items that they encounter in the real world and that they consider "cool" (Figure 3.1). Upon submission, each photo must be tagged with a specific category selected among the five prede-

[1] http://www.icoolhunt.com



Figure 3.1: Screenshot of the mobile application.

fined categories (technology, lifestyle, music, design and fashion), and must be textually described. If one enables geolocation, one's photos are automatically geo-referenced with the locations (latitude and longitude of the location is registered) in which they were uploaded. Users can vote others' photos with either a *like* or a *dislike* button. Every photo can receive limited-size comments from users, including the uploader. Every user is directly allowed to retrieve the list of popular pictures and latest uploads from any other user. In a way similar to what happens in Twitter, users of the application can follow each other. At the time of data collection, iCoolhunt didn't provide any "news-feed" of pictures uploaded by followed users: activities from social connections were only accessible by manually browsing their profile pages.

We consider the *complete* dataset from February 2010 (the application's launch) to August 2010. The iCoolhunt web application was launched after the end of our study period and, as such, our dataset includes only mobile application users. The format of this dataset could not be acquired via crawling but directly from the service providers. We did acquire data only until the end of August 2010. Within those first six months, 9,316 users uploaded 6,395 photos and submitted 13,893 votes.

The unique characteristics of this dataset fit particularly well our interest in characterizing trend spotters and trend makers. The dataset contains user demographic information, the follower-followee graph, votes, comments, and geographical location of the place where items were uploaded.

To better interpret the results of our data analysis and compare them to the findings in the literature (which are mainly about Twitter and, to a lesser degree, Foursquare), we spell out the similarities and differences between the mobile application and Twitter/Foursquare:

Similarities. In a way similar to Twitter, the mobile application's users can follow each other and, in a way similar to Foursquare, they can receive "honors" depending on how

active they are (these honors are called "guru", "observer", "rookie", and "spotter" based on the number of followers one has and sum of votes his uploads have received).

Differences. There are two main differences with Twitter. The first is about social interactions. Twitter users can *reply*, *retweet* others' tweets, *mention* specific users, but cannot vote explicitly tweets (although similar information can be inferred from "favorites"). Instead, users of the mobile application can vote items of others with a "like" or "dislike", comment items, but can neither forward ("retweet") items nor *mention* any other user. The second difference is about the user interface. Twitter users exchange status updates with each other, while the mobile application's users have no transparent way of being aware of what others are up to. We will discuss how these differences impact the effectiveness of our mobile application later on.

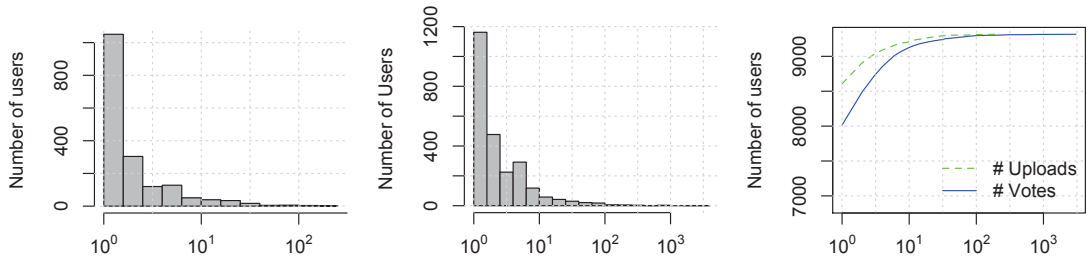
3.2 Dataset

The dataset includes 5,092 males and 4,224 females: 19% of males and 18% of females uploaded items and, out of a total of 13,893 votes, 69% of those were produced by males and 31% by females. This suggests that males are more active in voting than females. To understand how many users are actually using the application, we initially conduct a preliminary analysis of the behavior related to uploading, voting, and managing social relations. The size of the dataset is relatively small with respect to other popular social media sites as Twitter. However, the number of active users and their activities registered in the dataset are sufficient to draw concrete conclusions with the statistical analysis methods we use in the following sections.

3.2.1 Uploads and Votes

Uploading and voting pictures are the main activities in iCoolhunt, but, as one sees from the distributions in Figure 3.2(a) and Figure 3.2(b), only a small portion of the users are active in either uploading or voting. Out of the total 9,316 users, 1,761 (2,463) uploaded (voted) at least once and 710 (1,301) of them uploaded (voted) more than once. However, the minority who have uploaded more than once have contributed to 83% of the pictures, while those who voted more than once have contributed to 94% of the votes. That means that users prefer to vote pictures rather than to upload and, more importantly, that a minority of the users have contributed most of the content. This observation is coherent with the typical power law in social media sites.

Users are allowed to vote explicitly with a "like" or "dislike". Thus, to understand users' voting behavior, we separate votes into likes and dislikes and also consider any type of vote on aggregate. Figure 3.3 shows the distribution of the total number of votes per user and distinguishes "likes" from "dislikes": 584 registered users have submitted "dislike" votes, 2,349 have submitted "like" votes, and 2,463 have simply voted with



(a) #Uploads (b) #Votes (c) Empirical CDF

Figure 3.2: (a) Number of uploads per user; (b) Number of votes per user; (c) Distribution comparison on uploads and votes (log-scale on x-axis)

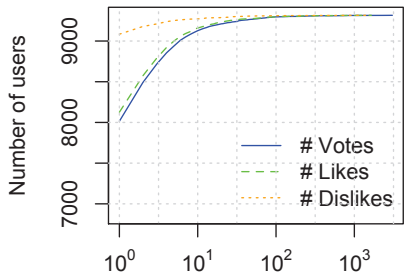


Figure 3.3: Empirical CDF of the number of votes, likes, and dislikes.

either a "like" or "dislike". This suggests that iCoolhunt users are comfortable to express far more positive votes than negative ones.

3.2.2 Geography

Pictures are geographically tagged while they are uploaded. By tracing the locations of pictures, we are able to infer the number of places each user has been to. Using Google Maps, we are able to classify the coordinates into countries and cities/towns. The pictures have been uploaded from 57 different countries and regions. Among those there are only six countries with more than 100 uploads (Figure 3.4): United Kingdom (UK), Italy (IT), United States (US), Germany (DE), Ireland (IR) and France (FR). Also, each user could upload from multiple countries: among those who uploaded pictures, 89 users did so from more than one country (Figure 3.5(a)) and 249 users from at least two different cities/towns (Figure 3.5(b)).

3.2.3 Following

To cope with information overload, iCoolhunt users define lists of people they know or whose content they like. Each user then preferentially receives pictures from his/her own list of following, and eventually leaves comments and messages on those pictures. Figure 3.6 shows the number of followers/followees for each user - only few users follow

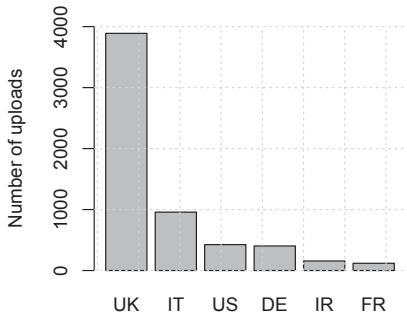


Figure 3.4: Number of uploads from each country (Top 6)

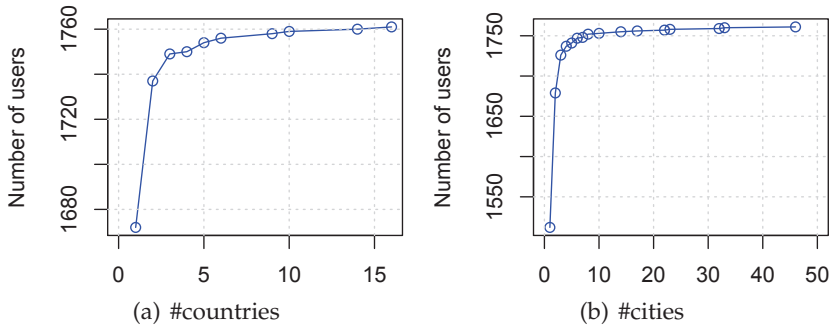


Figure 3.5: Empirical CDF of number of countries (cities) from where each user has uploaded, with y-axis representing the cumulative number of users.

other users, and even fewer users are followed. To then check whether users who upload more also have more followers/followees, we graph the scatter-plots of the number of followers/followees (y-axis) as a function of the number of uploads and votes for each user. In a way similar to [63], we bin the number of uploads/votes on a log scale and show both of the mean and median of each bin. The relationships are clear: the number of followers increases with the number of uploads (Figure 3.7(a)) and number of votes (Figure 3.8(a)); so does the number of followees(Figures 3.7(b) and 3.8(b)). In short, people who contribute to the community get followed, those who lurk do not. This makes sense as lurkers are essentially invisible.

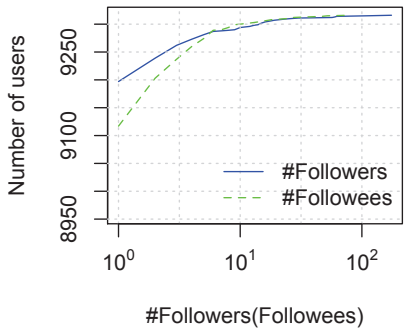
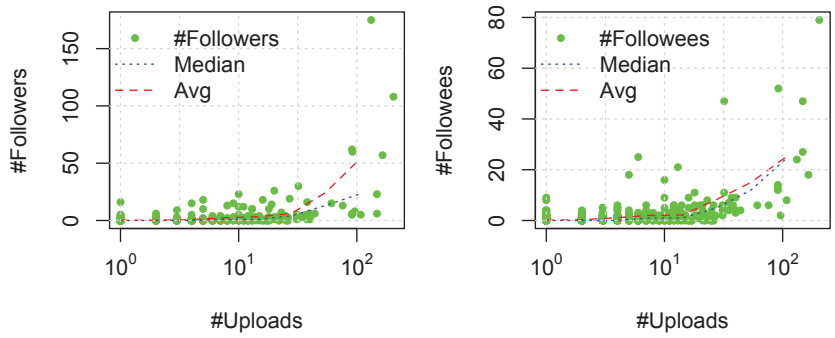
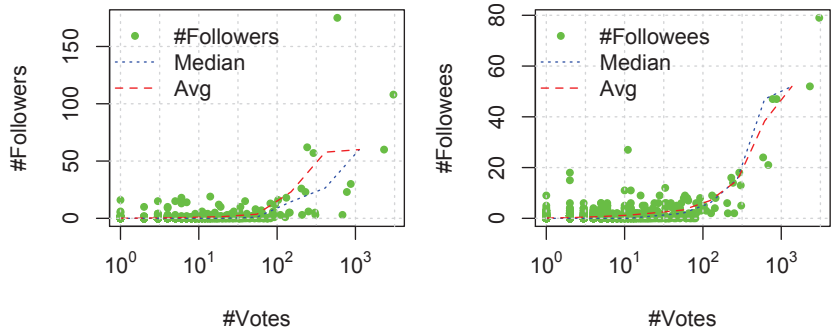


Figure 3.6: Empirical CDF of the number of followers and followees.



(a) Uploads vs. Followers (b) Uploads vs. Followees

Figure 3.7: Number of followers(followees) and number of uploads per user



(a) Votes vs. Followers (b) Votes vs. Followees

Figure 3.8: Number of followers(followees) and number of votes per user

To sum up, as one expects, a minority of users have uploaded and voted most of the pictures. Since we cannot get hold of access logs, we are not able to identify lurkers (those who simply browse) among inactive users. What we are able to differentiate though is trend spotters (users who spread trends) from trend makers (those who upload trends), and we do so next.

3.3 Identifying Trend Makers and Spotters

Our analysis unfolds three steps: identify trend spotters and trend makers; characterize them by conducting a quantitative analysis based on selected features; and build a statistical model that identifies who is a trend spotter and who is a trend maker.

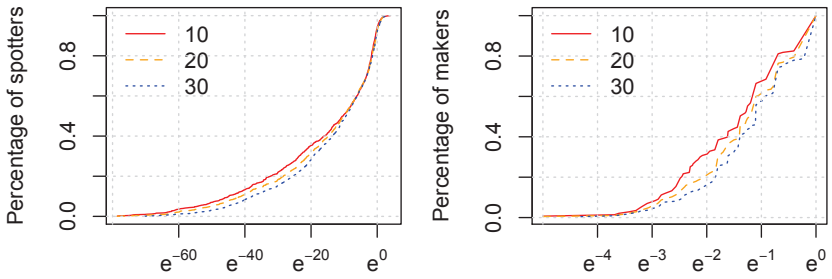
3.3.1 Defining Trend Spotters and Trend Makers

To identify trend spotters and trend makers, we need to define what a trend is first.

Trends. Trends differ from popular items, in that, they are not necessarily popular but they receive abrupt attention within a short period of time. To identify trends in the dataset, we define a "trend score" metric, which is derived from a simple burst detection algorithm proposed in [77]. At each time unit t (one-week window that incrementally slides every day), we assign to item i a $trendScore(i,t)$ that increases with the number of votes it receives:

$$trendScore(i,t) = \frac{|v_{i,t}| - \mu_i}{\sigma_i} \tag{3.1}$$

where $|v_{i,t}|$ is the number of votes item i has received within time unit t, μ_i is the mean number of votes it has received per time unit[2], and σ_i is the corresponding standard deviation. The higher the item's *trendScore*, the more votes it has received. For each time unit (each week), we sort items by their trend scores in descending order and select the top-n items to be trends. We have experimented with different $n \in \{10, 20, 30\}$ and found that spotter (maker) scores (defined later on) do not change very much (Figure 3.9) and, for $n = 10$, the resulting numbers of trend makers (140) and trend spotters (671) were sensible compared to the total number of users who voted (1,301) or uploaded (710) more than once. We thus report the results for $n = 10$.



(a) Spotter Scores (b) Maker Scores

Figure 3.9: Empirical CDF of spotter (maker) scores (log-transformed) versus top-n ranked items.

Trend Spotters. Trend spotters are those who tend to vote items that, after a while, end up being trends. Not all trend spotters are equally good at voting trends. Considering a set of trending items, one's ability of voting trends depends on three factors: *how many* trending items one has voted, *how early* one has voted them, and *how popular* the voted items turned out to be.

We incorporate these three factors in a *spotterScore* for each user u by dividing the number of trends user u has voted ($\sum_{i \in \mathcal{I}_u} g_{u,i}$) by u's total number of votes (v_u):

$$spotterScore\,(u) = \frac{\sum_{i \in \mathcal{I}_u} g_{u,i}}{v_u} \tag{3.2}$$

[2]The last time unit we consider is that in which the item received the last vote.

In the numerator, $g_{u,i}$ is the gain user u acquires when voting on trending item i and incorporates the three factors of *how many*, *how early*, and *how popular*:

$$g_{u,i} = v_i \times \alpha^{-p_{u,i}} \tag{3.3}$$

where $\mathcal{I}_u$ is set of trends that u has voted ($\sum_{i \in \mathcal{I}_u}$ reflects the *how many*); v_i is total number of votes item i has received (which reflects the *how popular*), and α is a decay factor (which reflects the *how early*, $\alpha = 2$ in our experiments) whose exponent is the order with which u has voted item i (i.e., $p_{u,i}$ means that u is the p^{th} user who voted i). A trend spotter is then anyone with trend spotter score greater than zero.

Trend Makers. Trend makers are those who tend to (not simply vote but) *upload* trending items. So the trend maker score of user u increases with the number of trends u has uploaded. The numerator of the score is $\sum_{i \in \mathcal{I}_u} I(i\ is\ a\ trend)$, where $\mathcal{I}_u$ is the set of items u has uploaded, and I is the indicator function, which is 1, if the enclosed expression "*i is a trend*" is true; 0, otherwise. This numerator is then normalized by u's total number of uploads ($|\mathcal{I}_u|$) to account for those users who indiscriminately upload a large number of items without any quality control. A trend maker is then anyone with trend maker score greater than zero.

$$makerScore\,(u) = \frac{\sum_{i \in \mathcal{I}_u} I(i\ is\ a\ trend)}{|\mathcal{I}_u|} \tag{3.4}$$

Typical users. If an active user (i.e., one who uploaded or voted more than once) is not a trend spotter or a trend maker, then he/she is considered to be a typical user. We discover that in our application, there were 1,705 typical users.

3.3.2 Characterizing Trend Spotters and Trend Makers

To characterize trend spotters and trend makers, we conduct a quantitative analysis that considers four types of features: activity, content, network, and geographical features.

Activity Features

The first activity feature we consider reflects how long an individual has been actively using the application. We call this feature "lifetime", and previous studies have identified it to be important as it conveniently identifies "early adopters" [36]. The literature recognizes that early adopters are a special interest group that heavily shapes usage of the application and ultimately determines the social norms within the application [29]. Once social norms are formed, changing them becomes very difficult and might backfire at times [30]. In addition to early adopters, we consider typical users as well. Their activ-

ity mainly consists of producing content (uploading pictures) and consuming it (voting pictures). Thus, we add two activity features to "lifetime": how frequently a user has been uploading pictures (*daily uploads*) and how many pictures the user has voted (*daily votes*).

Content Features

When users upload pictures, they are able to categorize them by selecting a proper category from the five predefined ones (technology, lifestyle, music, design and fashion). Previous studies on Twitter have linked category diversity to influence. According to [22] and [125], to become influential, one should "stay focused" – one tweet content in a specific category and become the "guru" in it. One may thus wonder whether our trend spotters and trend makers focus on specific categories of pictures, or, rather, whether they diversify consumption and production of content. To answer this question, we adopt a measure of categorical diversity from information theory called Shannon Index [71]:

$$s = -\sum_{i \in \mathcal{C}} (f_i \ln f_i) \tag{3.5}$$

where $\mathcal{C}$ is the above set of five categories (echnology, lifestyle, music, design and fashion) and f_i is the fraction of items (out of the total number of items) that belong to i^{th} category. Using this expression, we measure three types of diversities for each user – *upload diversity*, *vote diversity*, *consumption diversity* (consumption translates into either voting or commenting pictures).

Network Features.

Users of our application follow each other in a way similar to what happens in Twitter. Thus, the simplest network measures we could consider are in-degree (*number of followers*) and out-degree (*number of followees*). To then account for local network properties, we also consider the *clustering coefficient* [123] of a user, which is computed from the undirected graph whose nodes are users, and edges link users between whom there is at least one following relation. Clustering coefficient reflects the extent to which one's network is densely connected.

Geographical Features.

Information propagation faces geographical constraints, caused by the decrease of the probability of a social tie between a pair of individuals with the increase of the geographical distance between the pair [25, 84]. In our application, when users upload pictures of items, these pictures are automatically geo-referenced – they report the longitude-latitude

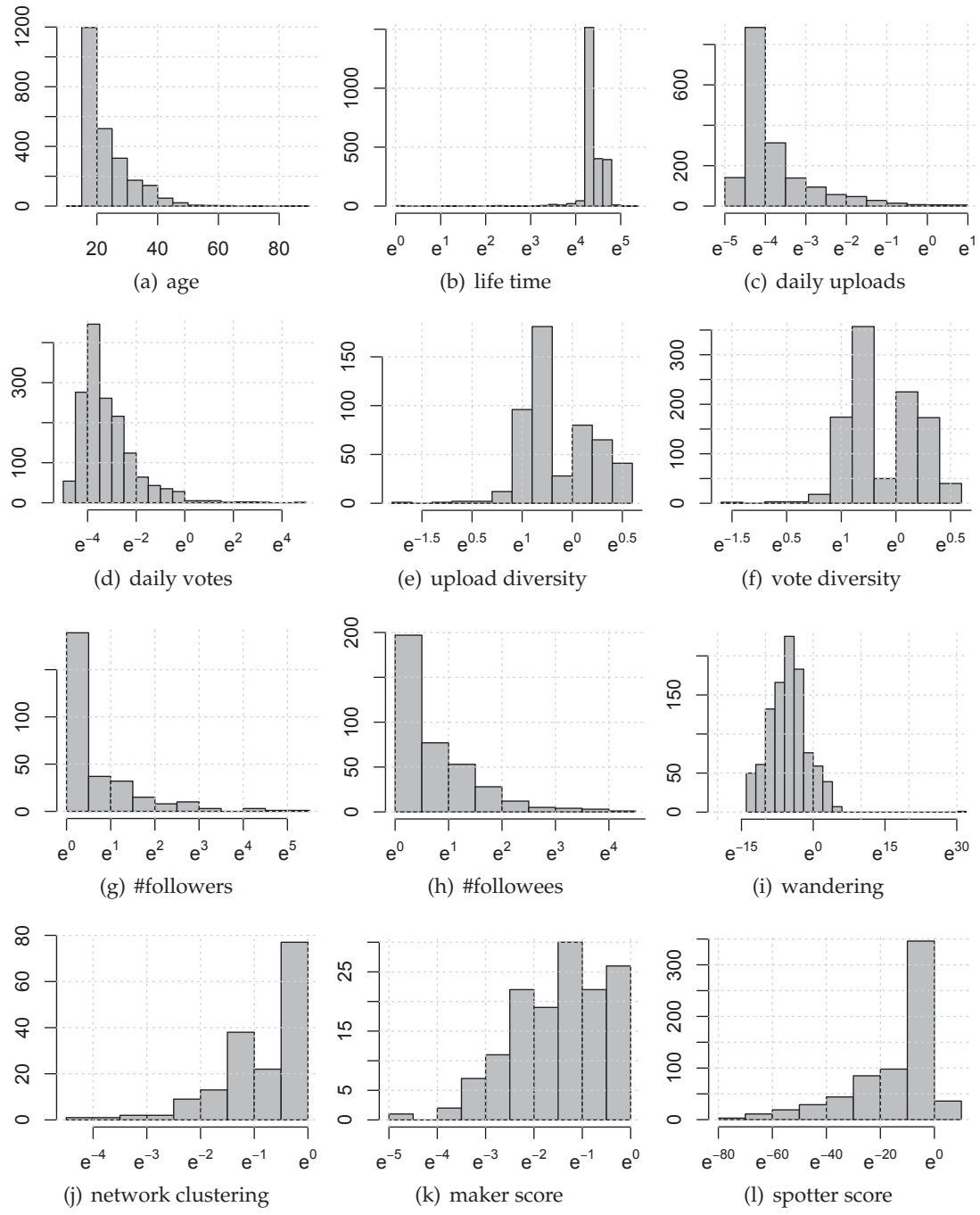


Figure 3.10: Distributions of features (a-j), trend maker (k) and trend spotter (l) scores with log-transformed values (except for the age feature). The x-axis represents the range of log-transformed features, and the y-axis represents the number of users.

pairs of the items' positions. Thus, we can compute how often and how far users physically move (*wandering*), and we do so using the radius of gyration [26]:

$$r_u = \sqrt{\frac{1}{n} \sum_{i \in I_u} d^2_{l_i, c_u}} \tag{3.6}$$

where n is number of u's uploaded items, I_u is the set of u's uploaded items, c_u is user u's center of mass (which is the "average point" of all geographical locations of u's items), l_i is the location where item i has been uploaded, and d_{l_i, c_u} is the Euclidean distance between user u's center of mass and the location of each item i.

Since locations are not only associated with pictures but also with users, we also compute the geographic span of a user's network [84]:

$$s_u = \frac{1}{m} \sum_{j \in \mathcal{F}_u} d_{u,j} \tag{3.7}$$

where m is number of u's followers, $\mathcal{F}_u$ is the list of u's followers, and $d_{u,j}$ is the distance between user u's center of mass and each follower j's center of mass.

We display the distribution of each feature in Figure 3.10. Since the distributions of the features are skewed, we show their log-transformed distributions. We see that in general, trend makers and trend spotters are young (Figure 3.10(a)), have been using the application for a considerable period (Figure 3.10(b)). A few of them upload (vote) actively daily (Figure 3.10(c) and 3.10(d)), and are followed by many other users (Figure 3.10(g)).

3.3.3 Who trend spotters and trend makers are

Having all the features at hand, we are now able to run a comparative analysis. We compare trend spotters, trend makers, and typical users by testing hypotheses drawn from the literature, which Table A.1 collates for convenience. We will now explain these hypotheses one-by-one.

Trend spotters (makers) vs. Typical users.

Previous studies have shown that, compared to typical Twitter users, influentials tend to be more active, more specialized in specific categories, and be more popular (i.e., attract more followers) [22, 109, 113, 125, 130]. To draw parallels between Twitter influentials and trend spotters (makers), we hypothesize that, compared to typical users, trend spotters (makers) are more active (H1.1 and H2.1 in Table A.1), specialized (H1.2 and H2.2), and popular (H1.3 and H2.3).

To test these hypotheses, we run Kolmogorov-Smirnov tests (K-S tests [85]), and we report the results in Table A.2. The idea is that we consider a pair of distributions, say, those of "daily uploads" for trend spotters (S) and for typical users (T) and compare them - we compare whether the mean of the distribution of trend spotters is greater than that of

	Content	Result
Spotters vs. Typical	H1.1 Trend spotters are more active than typical users.	$\surd$
	H1.2 Trend spotters tend to be more specialized than typical users in certain category of items.	$\times$
	H1.3 Trend spotters attract more followers than typical users.	$\surd$
Makers vs. Typical	H2.1 Trend makers are more active than typical users.	$\surd$
	H2.2 Trend makers are more specialized than typical users in certain category of items.	$\times$
	H2.3 Trend makers attract more followers than typical users.	$\surd$
Spotters vs. Makers	H3.1 Trend makers upload content more often than trend spotters.	$\surd$
	H3.2 Trend makers vote less often than spotters.	$\surd$
	H3.3 Trend spotters upload more diverse content than trend makers.	$*$
	H3.4 Trend spotters vote less diverse content than trend makers.	$\times$
	H3.5 Trend makers have more followers than trend spotters.	$\surd$

Table 3.1: Our Hypotheses ($\surd$: accept hypothesis; $\times$: accept the alternative hypothesis; $*$: unknown)

typical users (i.e., we test $S > T$). We find that, compared to typical users, both trend spotters and trend makers are more active (they upload and vote more) and are more popular (attract more followers). These results are statistically significant, that is, the corresponding p-values are below 0.05. Hence the four hypotheses H1.1, H1.3, H2.1 and H2.3 are confirmed. By contrast, hypotheses H1.2 and H2.2 are not confirmed. When consuming and producing content, trend spotters and trend makers neither focus on specific content categories nor diversify themselves more than what typical users do.

However, by separating what users vote and what they upload, we find that the items voted by trend spotters are more diverse than those uploaded. This preliminary difference between trend spotters and trend makers opens up the way for dwelling on the similarities and differences between these two types of users.

Feature (log-transformed)	**S > T**	**M > T**	**M > S** (if not shown otherwise)
Daily Uploads	**0.07** $*$	**0.45** $*$	**0.58** $*$
Daily Votes	**0.66** $*$	**0.18** $*$	**0.57** $*$ (M < S)
Upload Diversity	**0.31** $*$	**0.35** $*$	0.02 (M < S)
Vote Diversity	**0.31** $*$	**0.23** $*$	**0.27** $*$ (M < S)
#Followers	**0.06** $*$	**0.32** $*$	**0.26** $*$

Table 3.2: Summary of Kolmogorov-Smirnov test results of our hypotheses. D-values with significance level < 0.05 are highlighted and come with $*$. M, S and T stand for trend makers, trend spotters and typical users. We test a pair of distributions at a time - e.g., for **S > T**, we test whether the daily upload distribution for trend spotters is greater than that of typical users, and report the corresponding D-value.

Trend spotters vs. Trend makers.

Since no previous study has compared the characteristics of trend spotters and trend makers, we need to start with some initial hypotheses based on our intuition. So we initially consider that trend makers tend to upload items, while trend spotters tend to vote items. More specifically, we hypothesize that, compared to trend spotters, trend makers upload more content (H3.1), vote less (H3.2), upload less diverse content (H3.3), vote more diverse content (H3.4), and are more popular (H3.5).

After running Kolmogorov-Smirnov tests (Table A.2), we find that trend makers *upload* more frequently than trend spotters who, by contrast, *vote* more frequently. That confirms both H3.1 and H3.2. By then considering what users upload/vote, we find that trend makers "stay focused" (i.e., they upload and vote items in specific categories), while trend spotters vote items belonging to a variety of categories. So trend makers act in a way similar to the content contributors discussed in [75, 79] who tended to care specially about producing quality content. In a similar way, our trend spotters tend to upload items in the few categories they are more familiar with, while they vote on items of different categories, suggesting a wide spectrum of interests. Finally, trend makers tend to be more popular (are followed more) than trend spotters.

To recap, trend spotters preferentially engage in voting and do so across a broad range of categories, trend makers engage uploading within a limited number of categories. Both of them are popular, but trend makers are followed more than trend spotters.

3.4 Predicting Trend Makers and Spotters

By considering four types of features, we have been able to find statistically significant similarities and differences among trend spotters, trend makers, and typical users. Now we study to which extent these features are potential predictors of whether users are trend spotters (makers), and do so in two steps:

1. We model trend spotter (maker) score as a linear combination of the features.
2. We predict trend spotter (maker) using a logistic regression and a machine learning model: Support Vector Machines (SVM).

Upon the set of 140 trend makers, 671 trend spotters and 1,705 typical users (identified in the previous section), we now run our predictions.

3.4.1 Regression Models

Before running the regression, we compute the (Pearson) correlation coefficients between each pair of predictors (Table 3.3). As one expects, we find that different types of activities

are correlated (i.e., high positive correlation between the *number of followees*, *daily uploads*, *daily votes*, and content diversity). Attracting followers is correlated more with uploading content (i.e., positive correlation between the *number of followers* and *daily uploads*) rather than voting content (i.e., no significant high correlation between the *number of followers* and *daily votes*).

	Age	Life Time	Daily Uploads	Daily Votes	Upload Diversity	Vote Diversity	Wandering	Follower Geo Span	#Followers	#Followees	
Life Time	0.21										
Daily Uploads	0.02	-0.12									
Daily Votes	0.05	-0.09	**0.47** $*$								
Upload Diversity	0.02	0.09	**0.40** $*$	0.08							
Vote Diversity	0.04	0.08	0.22	0.08	**0.42** $*$						
Wandering	0.004	0.13	0.16	0.11	0.06	0.05					
Follower Geo Span	0.05	0.12	0.16	0.10	0.12	0.11	0.23				
#Followers	0.03	0.23	**0.37** $*$	0.14	0.22	0.16	0.44	0.16			
#Followees	0.05	0.17	**0.52** $*$	**0.31** $*$	**0.29** $*$	0.22	**0.56** $*$	0.21	**0.64** $*$		
Network Clustering	0.03	0.13	0.22	0.04	0.24	0.23	-0.001	**0.27** $*$	0.08	0.22	
Spotter Score	0.07	0.18	0.03	0.01	0.05	0.10	0.04	0.07	0.13	0.11	0.15
Maker Score	0.07	0.10	0.06	0.01	0.07	0.06	0.02	0.12	0.12	0.09	0.10

Table 3.3: Pearson Correlation coefficients between each pair of predictors. Coefficients greater than ± 0.25 with statistical significant level < 0.05 are marked with a $*$.

Next, we perform both logistic and linear regressions on input of the following predictors that tend not to be strongly correlated with each other: *age*, *life time*, *daily votes*, *daily uploads*, *votes diversity*, *upload diversity*, *wandering*, *number of followers* and *network clustering*.

We model trend spotter (maker) score as a combination of the features in two steps, as it is commonly done [38]. In the first, we use a logistic regression to model whether a user has trend spotter (maker) score greater than zero or not:

$$Pr(score_u > 0) = logit^{-1}(\alpha + \sum_{i \in \mathcal{V}} \beta_i \mathcal{U}_{u,i}). \tag{3.8}$$

In the second step, we take only those users with trend spotter (maker) scores greater than zero, and predict their scores with a linear regression of the form:

$$log(score_u) = \alpha' + \sum_{i \in \mathcal{V}} \beta_i' \mathcal{U}_{u,i}, \tag{3.9}$$

In Equation 3.8 and 3.9 , $\mathcal{V}$ is a set of predictors, $\mathcal{U}_{u,i}$ refers to user u's value of predictor i, and coefficients β_i and β_i' reveals the importances of each predictor i in each model respectively.

Features	$\mathbf{I(Score > 0)}$	
	Spotters	**Makers**
Age	2e-04	0.001
Life Time	**0.006***	**0.001***
Daily Votes (Daily Uploads)	**0.007***	**0.16***
Vote Diversity (Upload Diversity)	**0.38***	**0.14***
Wandering	-6e-15	-7e-15
#Followers	2e-05	**0.009***
Network Clustering	0.08	**0.28***

Table 3.4: Coefficients of the logistic regression. A correlation coefficient within 2 standard errors is considered statistically significant. We highlight and mark them with *.

The results of the logistic regression (coefficients in Table A.3) show that the significant predictors for trend spotters are *life time*, *daily votes* and *vote diversity*. For trend makers, significant predictors include *life time*, *daily votes*, *vote diversity*,*number of followers* and *network clustering*. These statistical significant predictors suggest that trend spotters tend to be early adopters who vote often and are interested in diverse items, and trend makers tend to be early adopters who upload often and also upload items from different categories, moreover, they tend to attract followers and have a dense connected network.

Considering then only the users who have trend spotter (maker) scores greater than zero, we focus on the features that can potentially predict how successful a trend spotter (maker) is. The results of the linear regression (β coefficients in Table A.4) shows that the significant predictors for successful trend spotters are *age*, *life time* and *vote diversity*, while the significant predictors for successful trend makers contain *daily uploads*, *upload diversity*, *number of followers* and *network clustering*. The sign of the coefficients of these significant predictors suggest that successful trend spotters are adult early adopters who vote items from various categories. By contrast, successful trend makers are users of any age who upload items belonging to specific categories (they "stay focused") and tend to attract social followers from different communities.

Features	$\mathbf{log(Score)}$	
	Spotters	**Makers**
Age	**0.36***	0.01
Life Time	**0.19***	0.0001
Daily Votes (Daily Uploads)	0.16	**-1.03***
Vote Diversity (Upload Diversity)	**7.28***	**-1.09***
Wandering	-2.1e-13	-1.4e-15
#Followers	-0.06	**0.01***
Network Clustering	2.75	**-0.64***
R^2	0.15	0.65
Adjusted R^2	0.14	0.64

Table 3.5: Coefficients of the linear regression. A correlation coefficient within 2 standard errors is considered statistically significant. We highlight and mark them with *.

The goodness of fit of a linear regression model is indicated by R^2. In our case, the adjusted R^2 is very similar to R^2, which is 0.15 for trend spotter score and 0.65 for trend

maker. So one is able to explain 15% variability in trend spotter score and 65% in trend maker score. The difference in these two results might be explained by either: 1) the idea that trend spotters might well be "accidental influentials" [121] and, as such, trend spotters are harder to identify than trend makers; or 2) the fact that our predictors simply encapsulate complex phenomena and, as such, their explanatory power is limited. Next, we test whether trend makers and trend spotters can be predicted by a machine learning model that has shown good performance in similar learning settings – that is, we use SVM.

3.4.2 Support Vector Machines (SVM)

We formulate the task of predicting trend spotters (makers) as a binary classification problem, where the response variable is whether a user's trend spotter (maker) score is greater than or equal to zero. To our sample of 671 trend spotters and 140 trend makers, we add an equal number of typical users (those 1,705 users have been identified in the previous section). By construction, the resulting sample is balanced (the response variable is split 50-50), and interpreting the results becomes now easy as the accuracy of a random prediction model would be 50%.

We split randomly each set of samples into two subsets, 80% of them are used for training and 20% for testing. We apply SVM [3] on the input of the same seven features previously used in the regressions to predict trend spotter and trend maker scores. We compare the results with those obtained by the previous logistic regression model, and we show their prediction performance by ROC (Receiver Operating Characteristic) curve (Figure A.1), AUC (area under the ROC curve), and accuracy (Table A.5). In a ROC curve plot, an ideal prediction model is expected to achieve a high true positive rate but with a low false positive rate. In Figure A.1, the point (0,1) corresponds to the perfect prediction while (1,0) corresponds to the worst, and the diagonal line reflects the baseline of a random guess. Points above the baseline indicate good prediction results [71].

	Spotters		**Makers**	
	AUC	**Accuracy**	**AUC**	**Accuracy**
Logistic	0.77	71.52%	0.85	82.09%
SVM	0.77	71.85%	0.90	88.06%

Table 3.6: AUC and best accuracy of each predictive model.

We see from the results in Figure A.1 and Table A.5 that SVM and logistic regression show comparable performance (for both, $AUC = 0.77$; accuracy is 71.52% for the regression, and 71.85% for SVM). SVM only slightly outperforms the logistic regression in identifying trend makers. This suggests that one is able to effectively identify trend spotters and trend makers even with a simple logistic regression. Also, SVM might not have shown considerable prediction gain simply because of our (limited) dataset's size.

[3] In our experiment, we apply SVM from the package of *e1071* in R programming language.

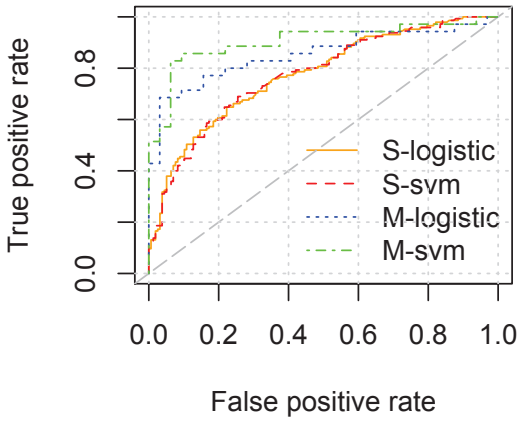


Figure 3.11: ROC curve of logistic regression and SVM model (S: trend spotters; M: trend makers).

3.5 Discussion

We have characterized trend spotters and trend makers based on four types of features (i.e., activity, content, network, and geographical features) and proposed a statistical model to accurately identify them. This work has both theoretical and practical implications.

3.5.1 Theoretical Implications

We show that trend spotters and trend makers are similar only to a certain extent. Compared to typical users, both of them: are more active in uploading/voting; attract more social connections; and upload/consume more diverse content. Yet, when they are compared not with typical users but with each other, differences emerge:

1. Trend spotters prefer voting more than uploading, and when they vote, they do so in very diverse categories. By contrast, trend makers act in a way similar to the content contributors in [75] who have special care in producing quality content and "stay focused" – they upload and vote items in very specific categories.

2. Successful trend spotters are early adopters who are attracted by diverse *items*, while successful trend makers attract diverse social *relations* (they tend to be followed by users from different social clusters).

These notable differences between trend spotters and trend makers would call for a rethink of current studies in opinion spreading. In studies of opinion spreading, social networks are traditionally modeled as graphs in which people are look-alike nodes characterized only by their graph properties (e.g., in-degree, out-degree). More recently, studies on the relationship between users' personality traits and use of social media have shown that graph properties are not sufficient to explain influence in social networks [41, 93, 94].

In iCoolhunt, geographical features seem not to matter. That might suggest that, to be successful, trend spotters or trend makers do not necessarily need to move often or travel around. However, based on further analysis, we have learned that, while the application was originally designed to let users share items on the move, some users have started to assume unexpected behaviors – for example, some have started to post content (e.g., images from the Web) that was not explicitly related with the location from which it was uploaded. Given such behaviors, to make more grounded claims, a longitudinal analysis would be required.

3.5.2 Practical Implications

The ability of identifying trend spotters and trend makers has implications in designing recommender systems, new products and user interfaces.

Recommender Systems. Every user has different interests and tastes and, as such, might well benefit from personalized suggestions of content. These suggestions are automatically produced by so-called "recommender systems". Typically, these systems produce recommendations people might like by *equally* weighting all user ratings. Given that trend spotters are effective social filters, one could imagine to weight their ratings more than those from typical users to construct a new recommender system.

New Products. Some web services (*e.g.*, 99designs[4]) provide a platform to crowd source design work, where clients submit their requests and designers try to fulfill them. Since trend spotters and trend makers are "fashion leaders", soliciting their early feedbacks might result into avoiding mistakes when designing new products. Often, at design stage, costs of correcting minor mistakes are negligible, while, at production stage, they become prohibitive.

User Interfaces. Trend spotters and trend makers do not connect to as many users as one would expect. That is likely because it is hard for iCoolhunt users to be aware of what others are up to. The user interface does not come with clear-cut "social features" that create a sense of connection and awareness among users as much as Facebook or Twitter sharing features do (as we have detailed in the Application section).

3.6 Summary

A community is an emergent system. It forms from the actions of its members who are reacting to each other's behavior. We have studied a specific community of individuals who are passionate about sharing pictures of items (mainly fashion and design items) using a mobile phone application. This community has a specific culture in which a set of habits, attitudes and beliefs guide how its members behave. In it, we have seen and

[4]http://www.99designs.com

quantified the importance of early adopters. In general, these individuals are those who initially set the unwritten rules that other community members learn (from observing those around them), internalize, and follow. In our case, early adopters tend to be successful trend spotters who like very diverse items. Trend makers, by contrast, tend to be highly organized individuals who focus on specific items. Understanding the characteristics of "the many" – of regular individuals with specific interests (trend makers) connected to early adopters with very diverse interests (trend spotters) – turned out to be more important than trying to find the "special few". At least, it has been so for the social application in our study, and for a variety of (more) complex networks [8, 44, 124].

Having fully understood the process of the creation of the trends, and the individuals participating in the process, we then ask whether these information could be used for the actual identification and exploration of trends.

CHAPTER 4

Personalizing Trends

In a community where users are design-conscious individuals, temporal dynamics matter, and users would greatly profit from ways of identifying the latest design *trends*. In this Chapter, based on the same dataset used in Chapter 3, we study the potential of providing "the crowds" a customized way to explore trends by leveraging the wisdom of these "special" users.

To begin with, in Section 4.1, we introduce recommender systems - the tools that we will use to help users explore trends. In Section 4.2, we propose a new way of recommending personalized trends to users, which includes three steps. Based on our findings from Chapter 3, where we saw that trends are created in a combined process by two types of the "special" users - trend makers and trend spotters - we first identify these two types of users from the "crowds" (Section 4.2.1). Second, based on what those "special" users have uploaded and rated, trends are identified early on (Section 4.2.2). Third, trends are recommended using existing algorithms (Section 4.2.3). Using the complete longitudinal dataset of the mobile application, we compare the performance of our approach to a traditional recommender system (Section 4.3).

4.1 Background

Recommender systems are used in different online services. Traditionally, studies focused on recommending books, CDs [69], movies [5, 135], songs [59, 133], news [31], and videos [32]. With the advent of mobile services, many applications are able to be aware of where users are, and some services have thus started to recommend location-based events [95], activities, and *POIs* (Points of Interests) [134]. Adding the users' social connections to their geographic information has been found to improve the quality of recommendations [52, 61]. Also, new social connections have themselves become "items to recommend" [23, 92]. There has been a lot of work on algorithms over the last few years

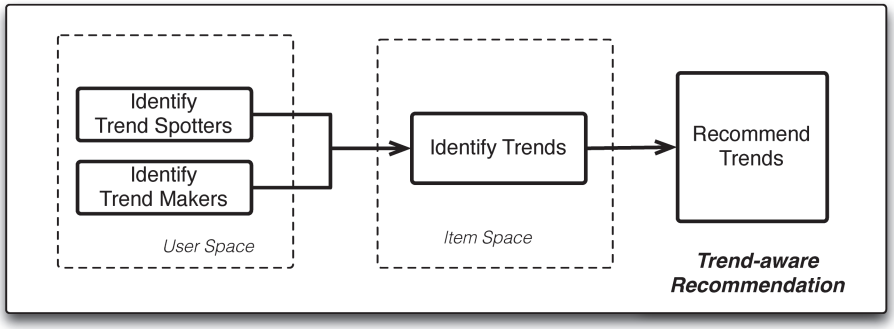


Figure 4.1: Trend-aware Recommender System

(a useful categorization of them can be found in Adamic *et al.*'s work [2]), and effective techniques such as matrix factorization have emerged [51, 52, 134].

Here, we bring the research line of recommender systems together with trend analysis, and study whether it is possible to build simple mechanisms to facilitate discovery and recommendation of trends.

4.2 Trend-aware Recommendation

To recommend personalized trends, we could build a trend-aware recommender system by following the common practice introduced in Chapter 2. Precisely, the process to set up a recommender system includes: 1) extracting implicit/explicit user preferences about items; and 2) implementing an algorithm to predict users' preferences about unrated items.

To learn user preferences on trends, we first need to identify them across the entire set of items. To obtain this prerequisite knowledge, we build upon the findings in Chapter 3, which suggest that trends in a social application are generated by a combined process in which two different users types are involved - regular individuals with specific interests (trend makers as defined in our work) and early adopters with very diverse interests (trend spotters). In Section 4.4, we will explain why it is not a good idea to identify trends directly and, instead, it is beneficial to identify trend spotters (makers) first and through them, then the trends themselves.

We go beyond the common practice and build our trend-ware recommender system by performing the following three steps (as shown in Figure A.2):

1. **Identify trend makers and trend spotters.** In this step, our goal is to identify two types of "special" users who are the origins of trends. We extend the prediction models (in Section 3.4) that we have presented in the previous Chapter, to identify trend makers and trend spotters of different levels of success (Section 4.2.1).

2. **Identify trends.** We then propose a method that identifies trends through the knowledge of trend makers and trend spotters (Section 4.2.2).

3. **Recommend the previously identified trends.** We extract implicit user preferences on trends, and construct a preference matrix based on the ones that we identified. Using a state-of-the-art matrix factorization algorithm (*Implicit SVD* [51]), we finally recommend trends (Section 4.2.3).

We now describe each step in detail.

4.2.1 Identify trend makers and trend spotters

In every social application, there are large behavioral differences among users [75]: some are able to identify trends early on, and some are leisure laggards. To identify the former type, we focus on two user categories – trend makers and trend spotters.

In Chapter 3, we have defined trend makers as those who tend to *upload* items that then become trends, and trend spotters as those who tend to *vote* items that then become trends early on. And, we have conducted a comprehensive statistical analyses about trend makers and trend spotters. The results have shown that they can be quantitatively identified using the following features:

Activity. The main activities on the application we consider in our work are two - voting and uploading. From them, we compute three activity features: *daily votes*, *daily uploads* and *lifetime*. This last feature reflects whether users are early adopters (i.e., are those who shape social norms [29]) or not [36].

Content. Users vary in how diverse their interests are: one could have a wide variety of interests, while another one could "focus" on very specific and limited set of interests. In Twitter, for example, it has been shown that influential users focus on very specific topics [22, 125]. To differentiate users based on their interest diversity, we consider two measures of content diversity. Both use the Shannon Index [71] and are called *upload diversity* and *vote diversity*.

Social Network. Since information might partly propagate along social connections, we also account for how well a user is connected by considering the *number of followers*, the *number of followees*, and the user's *clustering coefficient* (computed on the social graph in which each node represents a user, and an edge links two users with at least one following relationship [123]).

Geography. We finally consider: 1) how much and how often a user is *wandering* in the real world by using the radius of gyration [26]; and 2) a user's *geographical span of followers* computed as the average distance between where the user is and where his/her followers are.

As we have seen in Chapter 3, among trend makers and trend spotters, not every one is equally good at uploading or spotting trends. And there are always some more successful than others. Intuitively, trend makers and trend spotters of high level of success are those who are good at creating trends. Here, we go beyond the identification of whether one is a trend maker or trend spotter, and attempt to identify trend makers and trend spotters of different levels of success. To do so, we build a statistical model to predict at which level of success a user is a trend maker or trend spotter by means of three steps. For each user, we:

Step 1 Compute the user's spotter score and maker score (as to be defined below).

Step 2 Discretize the user's scores into k intervals. By doing so, users are able to be clustered into k classes, each of which corresponds to a level of success in creating trends.

Step 3 Predict the user's discretized scores on the input of previously defined features of activity, content, social network, and geography.

Next, we describe each of the steps.

Step 1. To begin with, we use two metrics defined in our previous work (Chapter 3), that reflect the extent to which a trend maker (spotter) u is successfully uploading (spotting) trends. We have defined in Section 3.3.1 a user u's $makerScore(u)$ as:

$$makerScore\,(u) = \frac{\sum_{i \in \mathcal{I}_u} I(i\ is\ a\ trend)}{|\mathcal{I}_u|}, \tag{4.1}$$

where $\mathcal{I}_u$ is the set of trends that u has uploaded, and $I(i\ is\ a\ trend)$ is an indication function which equals to 1, if i is a trend; otherwise, it is 0. To establish whether an item is a trend or not, we use a metric similar to the one proposed in [77]. That is, for each time unit t, each item i is assigned with a $trendScore(i,t)$ computed as:

$$trendScore(i,t) = \frac{|v_{i,t}| - \mu_i}{\sigma_i}, \tag{4.2}$$

where $|v_{i,t}|$ is the number of votes item i has received within time unit t, μ_i is the mean number of votes it received per time unit, and σ_i is its standard deviation. A high trend score tells that the item have received more attention than expected within the time unit. In each time unit, items are sorted according to their trend scores, and top-N items are extracted and identified as trends. From our analysis, we found that the temporal resolution (one week or two weeks) and the length of the recommended list do not significantly change the scores. In Figure 4.2, one observes that the trend spotter score does not change as the list length (top-10 vs. top-50) changes.

To add the spotter score to the maker score, the ability of spotting trends is largely determined by three factors – how many, how early, and how popular one's spotted trends

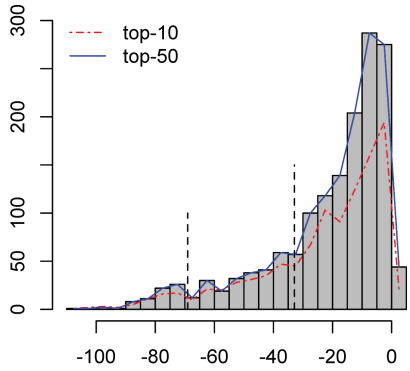


Figure 4.2: Trend spotter score (log). We split trend spotters into three classes using a proportional 3-interval discretization, as the two vertical lines show.

become. To incorporate the factor of how early and how popular trends become, for each trend i that u has spotted (voted), we compute the following gain $g_{u,i}$ score:

$$g_{u,i} = v_i \times \alpha^{-p_{u,i}}, \tag{4.3}$$

in which v_i is the total votes i received, $p_{u,i}$ captures that u is the p^{th} user who spotted i (the lower p, the better), and α is a decay factor. Combining a user's gains all together, we obtain a cumulative $spotterScore$ for user u (which is normalized by user u's total number of votes) (as in Section 3.3.1):

$$spotterScore\,(u) = \frac{\sum_{i \in \mathcal{I}_u} g_{u,i}}{v_u}, \tag{4.4}$$

Step 2. Based on users' maker scores and trend scores, we are able to cluster them into k classes, which indicate their ability of uploading (spotting) trends. To do so, we apply a proportional k-interval discretization [126] over the whole range of maker (spotter) scores in log-scale and assign each user to one of the three classes (with $k = 3$). We chose log-scale because the distribution of trend spotter(maker) scores in Figure 4.2 are shown to be skewed. And the increments in score is shown not linear. We partially identify three different speeds of increment and thus we cluster users into three classes.

Step 3. For each user, we compute the values of all his/her (activity, content, social network, and geographic) features defined above (and have been described in Section 3.3.2 in detail). Based on these features, we have shown in Section 3.4.2 that a machine learning technique - Support Vector Machine (SVM) is able to identify accurately whether one is a trend maker or a trend spotter. Similarly in this Chapter, we use SVM based on the same set of features, but now we predict to which class of trend maker or trend spotter one belongs to. We evaluate to which extent the SVM model could be used to identify accurately these different classes of trend makers and trend spotters in Section 4.3.

4.2.2 Identify Trends

with a variety of features to describe user behavior, we have described the means of three steps to learn a SVM model to predict to which extent a user is a successful trend maker or a trend spotter. We then explore the possibility to identify trends by relying on these identified special users of different classes.

Trend makers and trend spotters are the source of trends, but not all items uploaded and voted by those users become trends - there is a certain probability that they will be so. More generally, an item is likely to become a trend depending on:

- the extent to which the item's uploader is a trend maker;
- the extent to which the item's voters are trend spotters.

We model these insights in the following logistic regression [38]:

$$Pr(y_i = 1) = logit^{-1}(X_i\beta) \tag{4.5}$$

where $Pr(y_i = 1)$ is the probability of an item y_i is a trend ($Pr(y_i = 1)$), and X_i are a set of predictors, which are the uploader's trend maker class and the number of voters from each of the trend spotter classes. Coefficients β indicate the contributions of each predictor in X_i.

We validate the contribution of trend makers and trend spotters of different levels of success in trends identification in Section 4.3.

4.2.3 Recommend Trends

Up to now, we have designed a mechanism to identify a trend relying on: the extent to which its uploader is a successful trend makers; and the extent to which its voters are successful trend spotters. To recommend trends, we then need two major components: user preferences on trends; and an algorithm to predict user preferences on un-voted trends.

User preferences on trends. Having identified items that are likely to be trends, we are now able to build a trend-aware preference matrix $\mathcal{P}'$, in which $p'_{u,t}$ is 1 or 0 depending on whether u liked item t that has then become a trend. Since the application that we used in this study does not ask users to submit implicit ratings over Likert scale, we infer users' implicit ratings from their votes. When a user votes an item, we consider it as the signal of interest, and register his preference on the item as a rating 1.

Collaborative Filtering. On the trend-aware preference matrix, we apply two popular recommender systems algorithms: *Implicit SVD* [51] and *item-based* collaborative filtering [69].

As we have seen in Section 2.2, *Implicit SVD* aims at finding two descriptive matrices $\mathcal{U}$ ($n \times r$) and $\mathcal{V}$ ($r \times m$) for a given number of new features r, that can be used to approximate the original preference matrix $\mathcal{P}$ in a lower dimensional feature space. The inferred rating that user u would grant to item i then could be predicted as:

$$\hat{r}_{u,i} = \sum_{f=0}^{r} \mathcal{U}_{u,f} \times \mathcal{V}_{f,i} \tag{4.6}$$

Instead, *item-based* predicts the preference $\hat{r}_{u,i}$ of user u on item i by aggregating u's preferences toward the set of similar items ($\mathcal{N}$) of i as the weighted average of ratings on these nearest neighbors:

$$\hat{r}_{u,i} = \frac{\sum_{k \in \mathcal{N}} (s_{i,k} \times r_{u,k})}{\sum_{k \in \mathcal{N}} (|s_{i,k}|)} \tag{4.7}$$

in which $s_{i,k}$ is the similarity between item i and one of its nearest neighbor k. In both cases of *implicit SVD* and *item-based*, candidate items are then sorted by descending order of their predicted ratings, and the top-N ranked ones are commonly returned as the final recommendations.

We compare how these algorithms perform by comparing the trend-aware matrix as input with a traditional preference matrix $\mathcal{P}$ (in which $p_{u,t}$ is 1 or 0 depending, again, on whether u voted on item t that has then become a trend). The difference between the two preference matrices is that the trend-aware one is less sparse because, at the columns, it does not have all items but only those that we have predicted to be trends.

4.3 Evaluation

In Section 4.2, we proposed to construct a trend-aware recommender system by: 1) using a SVM model to identify trend makers and trend spotters of different level of success, with a variety of features to describe user behavior; 2) using a logistic regression to identify trends by relying on the level of success to which the uploader is a trend maker and level of success to which their voters are trend spotters; and 3) recommending trends by applying *Implicit SVD* and *item-based* collaborative filtering techniques to operate a trend-aware preference matrix [1].

In this Section, we evaluate the effectiveness of each of the three steps using the same dataset that we have introduced in Chapter 3. By effectiveness, we refer to: 1) the accuracy of statistical predict models; and 2) the accuracy of the trend-aware recommender system.

[1] We use Mahout (`https://mahout.apache.org/`) implementation of the algorithms.

4.3.1 Classifying users into trend spotter(maker) classes

We first evaluate the extent to which *SVM* is able to classify each user into one of the three maker/spotter classes on input of the user's features (introduced in Section 4.2.1). To this end, we use the dataset that introduced in Chapter 3 and sampled 209 unique trends, 50 trend makers, and 531 trend spotters, we run a 10-fold cross validation. We randomly split our experimental dataset into 10 subsets with the same size. The validation includes 10 repeated processes, in which each subset is used once as the test set to validate, while the remaining subsets are used for training. With 10-fold cross validation, the extended *SVM* model are shown to be able to identify accurately 83.80% of trend spotters and 60.7% of trend makers of different classes.

4.3.2 Determining whether an item is a trend or not

After ascertaining that *SVM* is able to identify trend spotters and trend makers with acceptable accuracy, we now need to test whether the logistic regression in Section 4.2.2 is able to identify trends based on information about uploaders and voters.

Since we divide trend makers and trend spotters into three classes according to their different levels of successfulness (as described in Section 4.2.1). The regression predicts whether an item is a trend or not based on four features:

- the uploader's trend maker class
- the number of votes from users who belong to the *low* spotter class
- the number of votes from users who belong to the *medium* spotter class
- the number of votes from users who belong to the *high* spotter class

To test the logistic regression, we build a balanced dataset that contains our 209 trends plus 209 (randomly extracted) non-trends and obtain the results in Table 4.1. The statistically significant coefficients suggest that an item is more likely to become a trend, if its uploader is a good trend maker and its voters are in the upper (trend spotter) class.

To avoid overfitting in Equation 4.5 (considering the size of the dataset used in this study might be limited), we add a commonly used regularization term - Tikhonov regularization term [91]. The problem of learning β then translates into the following optimization problem:

$$\beta' = \arg\min_{\beta} \sum_{i} log(1 + exp(-y_i \beta X_i)) + \lambda ||\beta||_2^2 \tag{4.8}$$

We split the dataset of trends into two subsets: the first subset consists of 80% of the entire dataset and is used to *train* the model, while the remaining 20% is used to *test* the model. Again, with a 10-fold cross validation, we first fix the value of λ and then fit the model with the training set.

To measure the accuracy of the regularized logistical regression model, we apply the trained model to the test set. We obtain the ROC curve plot that reflects both the model's TPR (true positive rate) and FPR (false positive rate) [71]. As we have introduced in Section 3.4.2 that an ideal classification model is expected to achieve a high TPR but with a low FPR, and in a ROC curve plot, the best classification performance is at the coordinate (0,1) while the worse is at the coordinate (1,0), and the diagonal line represents a random guess. In Figure 4.3, We see that the ROC curve of the regularized logistical regression model in Figure 4.3 are above the baseline (the diagonal line), which indicates that the regularized regression model is able to classify whether an item is a trend a not.

Predictors	**Coefficient**
Uploader's trend maker class	$6.21***$
#Voters from low trend spotter class	-1.30
#Voters from medium trend spotter class	$-1.17*$
#Voters from high trend spotter class	$0.64***$

Table 4.1: Coefficients of the logistic regression (a correlation coefficient within 2 standard errors is statistically significant. The significance levels are marked with $*$'s: $p < 0.001(***)$, $p < 0.01(**)$, $p < 0.05(*)$))

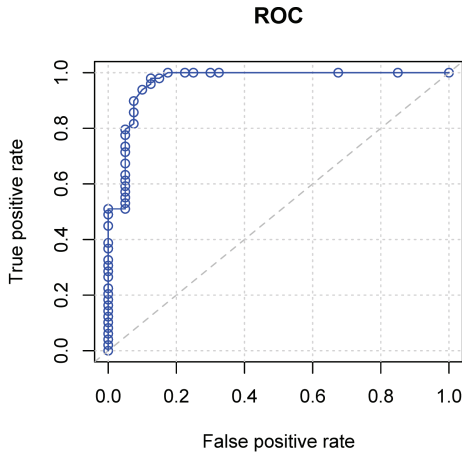
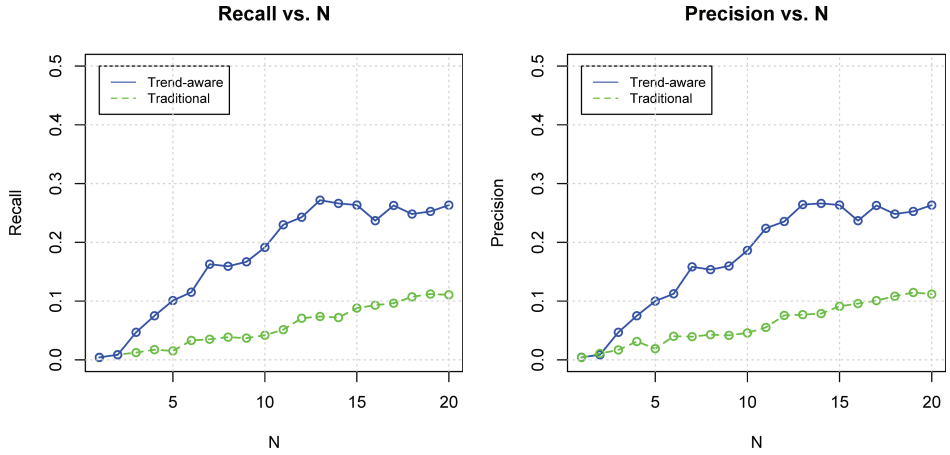


Figure 4.3: ROC curve for the logistic regression that predicts whether an item is a trend or not.

4.3.3 Recommending trends

We have validated our ability to identify trends by relying on the successfulness of trend makers and trend spotters. Now the question is: if we were to build a *user-by-trend* matrix out of the predicted trends, what would be the performance of an existing collaborative filtering algorithm?

To answer the question, we need to determine three components - a recommender system algorithm, the evaluation metrics, and a baseline with which we could compare the performance of our recommender system:



(a) Recall vs. top-N (b) Precision vs. top-N

Figure 4.4: Precision and Recall. Results for trend-aware recommender vs. item-based recommender. The size of the recommended list is N.

1. **A recommender system algorithm.** First, we choose the most popular memory-based approach, that is, a simple item-based collaborative filtering algorithm [104]. Later, we will see whether we can improve performance with *Implicit SVD* [51].

2. **The metrics that reflect recommendation performance.** To be in line with the literature, we compute precision and recall [51] defined as following:

$$recall(N) = \frac{\#hits}{|T|} \tag{4.9}$$

$$precision(N) = \frac{\#hits}{N * |T|} \tag{4.10}$$

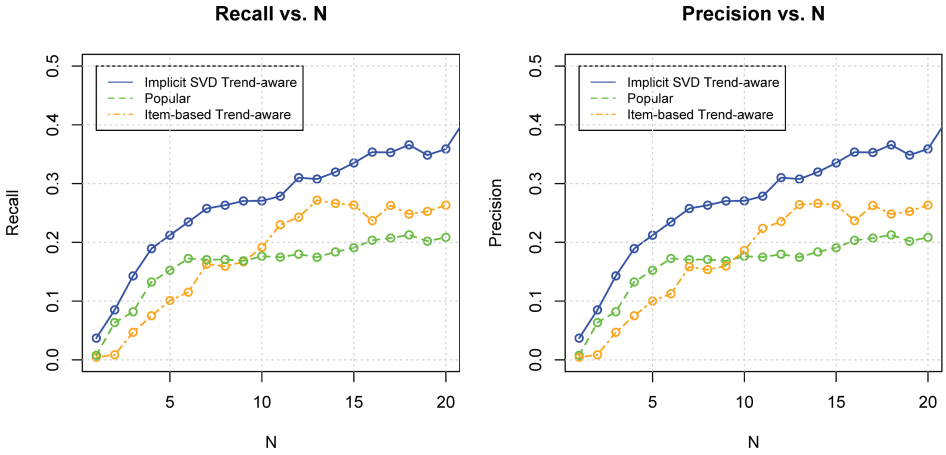
   where T is the test set, and N is number of items to recommend.

3. **The baseline against which our trend-aware approach will be compared.** To ease interpretability of the results, we again select item-based collaborative filtering but, this time, the algorithm would take in input the original user-item preference matrix $\mathcal{P}$, in which:

$$p_{u,i} = \begin{cases} 1 & \text{if } u \text{ likes } i \text{ \& } i \text{ is a trend} \\ 0 & \text{otherwise} \end{cases}$$

Having the three components (i.e.,the algorithms, the evaluation metrics and the baseline) defined, we now examine to which extent our trend-aware recommender system is able to profit users with the discoveries of personalized trends.

Traditional item-based vs. Trend-aware item-based. Figure A.3 shows precision and recall for the traditional and trend-aware item-based collaborative filtering as a function of the recommended list size (top-N recommendations). For both systems, precision and recall improve as N increases. However, at increasing value of N, both precision and recall



(a) Recall vs. top-N (b) Precision vs. top-N

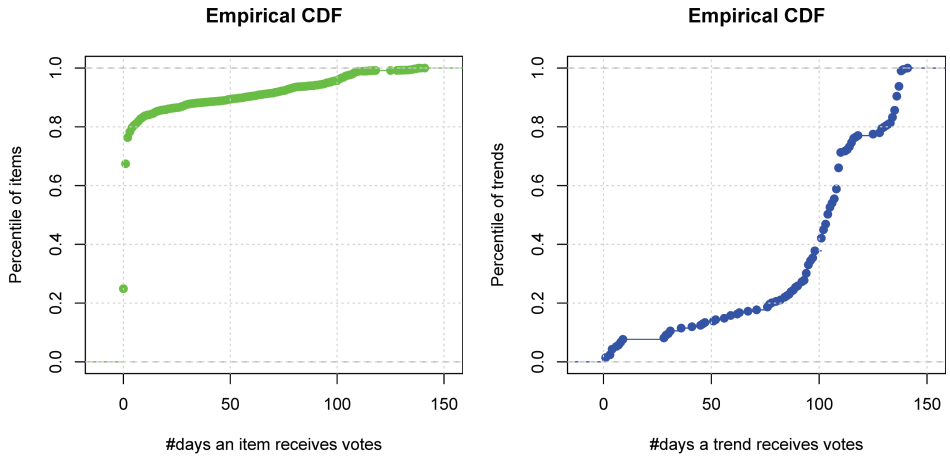
Figure 4.5: Precision and Recall. Results for two trend-aware recommenders (item-based and Implicit SVD) and for recommendations of most popular trends.

increase at a faster rate in the trend-aware case. For instance, at top-10 recommendations, precision/recall for the traditional item-based recommender system is 0.05, while the trend-aware item-based approach achieves 0.2 (as shown in Figure A.3). This significant difference in performance indicates that a traditional item-based recommender system would not be able to recommend trends, while a trend-aware system would. These results also suggest that, in the presence of data sparsity, relying on few expert ratings is an effective way of recommending trends.

Trend-aware item-based vs. Trend-aware Implicit SVD. So far, we have analyzed how an item-based collaborative filtering algorithm would perform to recommend trends. Considering the sparsity of the dataset, next, we test whether a papular matrix factorization approach - *Implicit SVD* - would improve the performance. Figure A.4 shows this to be the case. We could see that at any given top-N recommendation, the precision and recall from *Implicit SVD* are consistently better than the *item-based*. Additionally, as the size of the recommended list increases, *Implicit SVD* trend-aware recommender system improves precision and recall faster than item-based trend-aware one.

We have shown the ability of a trend-aware item-based in recommending trends. And using *Implicit SVD*, the performance of our trend-aware recommender system could be further improved.

Popularity. In a recommender system, popular items (those receive most number of votes) are often easier to recommend, because it is highly likely that similar users have already voted them [96]. Trends are similar to popular items in the sense that both of them receive a considerable number of votes. Differently, trends are the content that receives *abrupt* increase of votes, while the voting rate to popular items does not necessarily increases. Inevitably, among trends, some have better adoptions from users (i.e., more popular) than others. If the popularity is the only reason that people consumes



(a) Item Lifetime (b) Trend Lifetime

Figure 4.6: Number of days an item (*a*) vs. a trend (*b*) receives votes for.

trends, then recommending popular trends would yield the best accuracy. We examine this assumption by comparing the performance of our trend-aware recommender system and a simple strategy to recommend only popular trends.

Interestingly, as we can see from Figure A.4, if the recommender system recommends only popular trends, precision and recall would be worst. The precision and recall do not improve much while recommending more than 5 top popular items. This indicates that even for trends - items that one expects to be non-long tail - personalization makes sense. But up to a point, precision and recall results are limited, and that is largely because of the very nature of trends.

To sum up, in this Section, we have validated that: 1) based on a variety of (activity, content, social network and geography) features to describe user behavior, trend makers and trend spotters of different levels of success could be identified using SVM; and 2) trends could be identified by relying on the activities of trend makers and trend spotters. We have confirmed that for non-long tail items like trends, personalized recommendation makes sense. And we have examined the performance of our trend-aware recommender system in terms of accuracy, and *Implicit SVD* trend-aware is turned out to perform the best. In addition, we have answered a fundamental question - whether trend detection helps the recommendation process; and the answer is a definite 'Yes'.

4.4 Discussion

We have proposed a novel mechanism to recommend personalized trends, in which a key component is to identify trends by leveraging the "wisdom" of trend makers and trend spotters. We have validated each step of the trend-aware recommender system and have

examined its recommendation quality in the terms of accuracy. In this Section, we discuss the implications of our work related to the literature of trends and recommender system.

Why not detecting activity bursts directly? Since a burst detection algorithm could be applied to identify trends (in a way similar to expression (A.1)), one might wonder why we add the intermediate step of identifying trend spotters (makers) and not, instead, identifying trends directly. The main reasons for this choice are efficiency and time:

- **Efficiency.** We have seen in Chapter 3 that trends are created by some "special" users (trend makers and trend spotters). These trend makers and trend spotters behave differently from the typical users, and thus are able to be identified with a variety of features. While a typical burst detection algorithm requires the complete and up-to-date view of the system - to monitor *all* votes on *all* items, monitoring a limited number of users who are the source of trends requires much less computation resources.

- **Time.** This is the most important reason and comes from the temporal dynamics typical of trends. In Figure 4.6, we display the empirical distribution of number of days items receive their votes by the means of the cumulative distribution function (CDF). Given a number of days that an item receives votes as x, the y-axis of CDF plots shows the percentage of items that have received votes not more than x days. We could see that the average item is generally short-lived and dies off the first or second day (Figure 4.6(a)); by contrast, a trend persists for a longer period of time (as one would expect based on preferential attachment[2]), yet it also takes off after two weeks or so (Figure 4.6(b)). This observation is coherent with the findings from Crane *et al.* [28] that endogenous trends receive a smooth increase of user interests before the spike. As such, burst detection would miss trends for a long period of time, while monitoring key individuals - trend spotters (makers) - can be done quickly and efficiently. One contribution of this work has been to show that monitoring key individuals not only is quick and efficient but is also an accurate way of identifying trends.

Online Updating. In our analysis, we have not registered the frequent emergence of new trend spotters and trend makers. However, in a system with a larger user base, that might be the case, and ways of updating the pool of key users - trend spotters (makers) - would be needed. To decide when and how to run such updates, one of the idea is to explore the use of controllers that automatically and accurately estimate frequency of updates. These techniques have been recently introduced with the idea of ensuring stable and high-quality recommendations in dynamically evolving environments [53].

[2] http://en.wikipedia.org/wiki/Preferential_attachment

4.5 Summary

Recommender systems are a powerful tool to personalize content. They are of great use in providing people with information close to their interests. In this Chapter, we have designed a trend-aware recommender system to help people explore and discover trends of their preferences. We have shown that, upon activity, network, and geographic attributes, a machine learning approach (SVM) can identify key users with different levels of skills in creating trends - trend spotters and trend makers. Whether an item will be a trend or not can then be reliably identified based on whether the item has been uploaded by a successful trend maker and voted by successful trend spotters. We have then seen that existing recommender systems can profit from this ability of identifying such "special" users, and we have evaluated the effectiveness of the system in recommending trends from the perspective of recommendation accuracy. The results have confirmed that trends - as non-long tail items - are worth being personalized as well.

While recommender systems are meant to help people to explore the items of their interests, it learns continuously user preferences based on their behavior. Every acceptance of recommended item reinforces the beliefs of the system about user preferences, and the system thus continues to recommend items of the same type. These accurate recommendations then gradually narrow down the diversity of items that user would get recommended next. To tackle this general problem in recommender systems, in next Chapter, we then shift our focus from accuracy to another demanding quality measurement of recommender system - serendipity, and will do in the context of a location-based recommender system.

CHAPTER 5

Serendipitous Recommendations

In the previous Chapter, we have presented the design and evaluation of a recommender system to make personalized suggestions of trends. However, recommender systems suffer from a major drawback. That is, they incrementally learn our preferences and suggest items we might like. With accuracy as their core objective, these systems run into the following problem: they "trap" users in their own "filter bubbles" (i.e., recommended items tend to be liked only by users with similar preferences). The consequence is that the recommendations get more and more focused on one's central interests, thus fail to provide interesting discoveries. In this Chapter, we tackle the problem by improving the quality of recommendations in terms of serendipity. That is, we explore the potential to provide unexpected recommendations that people do enjoy. Due to the limitation of its size and the sparsity, we chose to move away from the dataset introduced in Chapter 3. We focus on a new dataset that was collected from a well established location based mobile service (i.e., Foursquare [1]). However, it should be noticed that the methodology and strategy we propose in this Chapter are also applicable to other recommender systems in general (e.g., our trend-aware recommender system).

We start by addressing the demand of serendipitous recommendations in Section 5.1. To lead users out of the bounded personalization in mobile social network services, we first design a basic location recommender system that combines a traditional item-based collaborative filtering algorithm with geographical information taken into consideration (Section 5.2.1). Then, we explore different strategies that introduce serendipity into the recommendation process by leveraging network analysis techniques. This allow us to tackle the possibility of introducing serendipity by promoting places that go beyond those that would be recommended based on past visited places (Section 5.2.2) and on one's typical routine (Section 5.2.3). We quantitatively evaluate to which extent we are able to introduce serendipity without compromising the accuracy of the recommendations upon the real-world dataset in Section 5.3.

[1] https://foursquare.com/

5.1 Background

Based on our online behavior, a recommender system suggests items (e.g., books, songs) a user might like. Until recently, considerable research efforts have focused on improving the accuracy of recommendations, including the trend-aware recommender system described in previous Chapter. However, accuracy-focused recommenders may not necessarily translate into enjoyable user experience. They might well produce ineffective or "expectable" recommendations, harming a user's personal growth and experience by pandering to the user's existing taste. That is because a traditional recommender system incrementally learns one's preferences and its accuracy increases over time. Critics says that what users consequently end up with is a narrow, biased, subjective "filter bubble", in which they are left with recommendations of limited scope [86].

To fix that, researchers have recently started to consider factors other than accuracy that contribute towards the quality of recommendation [72]. They defined concepts such as novelty and serendipity, and proposed ways to quantify them [49, 117, 136, 137]. These approaches have been mostly tailored to Web-based systems (e.g., music recommenders [133]) but have not been explored in the context of mobile recommendations. In mobile settings where location counts, most of research work has been focused on studying the spatial dynamics of people [26, 106, 112], and making accurate recommendations about events [95], bars and restaurants [102] that people could be physically presented. The consequence of these accuracy-focused recommendations is that users are gradually "trapped" into places where only like-minded users go, potentially contributing to geographic segregation [95].

To balance accuracy and serendipity of mobile recommendations simultaneously, in this Chapter, we extend the related work and propose a variety of techniques for recommending venues that are accurate and serendipitous, and whose recommendations are easy to explain. Here, our study focuses on users and places within the city of London, and explore the possibility to bring more "surprises" into ordinary recommendations.

5.2 Our Proposals

Our goal is to develop a recommender system that aims at finding the right balance between accuracy (i.e., the ability to recommend venues that a user likes) and serendipity (i.e., the ability to recommend venues that a user finds novel and surprising). To this end, we propose a set of basic algorithms (Section 5.2.1) that considers various factors which might impact one's decision to adopt an item (in our context, to visit a location), and we then extend those algorithms to introduce serendipity in the recommendation list (Sections 5.2.2 and 5.2.3).

5.2.1 Basic Algorithms

As basic algorithms, we aim to provide "accurate" location recommendations. To do so, we first design a basic location recommender that takes into consideration several factors that might impact one's decision to visit a location, and it does so using a Bayesian model.

Whether a user goes to a place might depend on two main factors: whether the user likes the place (taste) and how far the place is (distance) [102]. We thus first model user taste by introducing a concept of "user tribes" (i.e., clusters of like-minded people) and then include further attributes such as physical distance later on. Additionally, we introduce venues' social *mixing* - a feature to reflect how attractive a venue is to users with diverse tastes. Finally, we describe how these features could be integrated to predict the probability for a user to visit a place using Bayesian model.

User Taste. People visit a place because they like the place. Predicting the extent to which one might like an item (in our case, a venue) is the main goal of a recommender system. Therefore, we could model every user's taste on each place by incorporating a popular item-based collaborative filtering algorithm. As we have seen in both Chapter 2.2 and Chapter 4, one of the most important tasks in item-based approach is to find k nearest neighbors to each item, which is commonly accomplished by computing similarities between each pair of items. To compute similarities, in the traditional means, each venue can be described as a vector of users who have visited it (which is used as our baseline in Section 5.3.4). However in our algorithm, we introduce a concept of "user tribes" and then describe how it could be used to compute the similarities between each pair of places. The reason that we introduce the concept of "user tribes" is because it could be of great helpful for us to understand the the attractiveness of venues later on.

- *User Tribes.* The input of our system is, for each venue, the set of users who have visited it. We first cluster users into a set of *tribes* - a tribe consists of users who tend to visit the same/similar places. To do so, we use Latent Dirichlet Allocation (*LDA*) [16], which is typically used to learn topics out of a collection of textual documents.

  In topic modeling, a document is considered as a mixture of topics, and each word in the document attributes to one of these topics. Given a collection of documents, LDA was originally designed to learn the word compositions of the topics, and the different importance of each topic to every document [15, 16].

  To paraphrase *LDA* in our case, we consider our venues as documents, and users who visited those venues as words in the corresponding documents. In such a context, *LDA* learns "topics" that are groups of users who have visited similar venues - we call those topics "user tribes". The distribution of user tribes for each venue indicates the extent to which the venue is visited by like-minded users (if only few tribes visit it) or not (if, by contrast, a variety of tribes visit it). We have used *LDA*

because it has been shown to counter data sparsity in recommender systems, which is a major problem in our context of mobile recommendations.

- *Venue similarity.* Once we have identified user tribes, we can compare the similarity between each pair of venues (i,j). Let $\mathcal{T}$ be the complete set of user tribes, and $w_{i,t}$ be the ratio of users from tribe t within the entire set of visitors to venue i; the similarity between venue i and j is measured by the cosine similarity

$$\mathrm{sim}(i,j) = \frac{\sum_{t\in\mathcal{T}} w_{i,t} \times w_{j,t}}{\sqrt{\sum_{t\in\mathcal{T}}(w_{i,t})^2} \times \sqrt{\sum_{t\in\mathcal{T}}(w_{j,t})^2}}. \tag{5.1}$$

  Having the similarity computed for each pair of venues, we now combine the process with the traditional item-based collaborative filtering algorithm. The algorithm now outputs a score $l_{u,i}$ for each venue i, personalized for each user u based on how similar the user's past visited places ($\mathcal{H}_u$) are to venue i:

$$l_{u,i} = \frac{1}{|\mathcal{H}_u|} \sum_{h\in\mathcal{H}_u} \mathrm{sim}(i,h). \tag{5.2}$$

As the item-based algorithm, this score $l_{u,i}$ is commonly treated as the predicted taste of every user u to each item i.

Physical Distance. Taste could be one of the key reasons for a user to visit a place. But, physical distance to the place also matters. A user goes to a place might because 1) it is nearby, for instance, a cafeteria next to his/her working place; or 2) it fulfills his/her interests. - e.g., visiting a special exhibition organized in a museum on the other side of the city. Depending on the extent to which one likes a place or a place fulfills one's interests, his/her willingness to displace changes. Therefore, for each venue i and user u, we consider the distance $d_{u,i}$ (in meters) between venue i and the centroid of all coordinates (latitude and longitude) of places visited by u.

Social Mixing. Different venues may attract different people: some venues are tailored to niche crowds, while others are open to anyone. The latter type of venues encourages forms of social mixing more than what the former do. It is also reasonable to consider that users are more likely to travel far to reach a place that appeals to their "niche" tastes. We define a venue's social *mixing* as the extent to which the venue attracts diverse sets of users. We compute i's social mixing score s_i as the Shannon diversity [71] of the vector w_i (which has as many k elements as there are tribes) - the more different tribes there are, the higher the venue's social mixing.

$$s_i = -\sum_{t\in\mathcal{T}} w_{i,t} \log w_{i,t}. \tag{5.3}$$

Having defined the three factors (i.e., user taste, physical distance, social mixing) that are likely to impact one's decision to visit a place, next, we describe how to recommend personalized venues by combining these factors in a Bayesian model [71].

Bayesian Modeling. Using Bayesian model, we could translate the task of generating recommendations as to predict whether a user will go to a given venue. To the purpose of providing accurate recommendations at this stage, venues with high probability to be visited are worthy to be recommended.

We then consider the random variables L (related to predicted Likes), D (geographic Distance) and S (venue's Social Mixing) obtained by discretizing respectively $l_{u,i}$, $d_{u,i}$ and s_i; we want to compute the probability of G (Go) event - that is, whether user u visits venue i.

To find out how much the "social mixing" feature is important, we use two Bayesian models. The first ignores the value of S and is thus about computing $p(G|L, D)$. The second is about computing the full $p(G|L, D, S)$. We obtain these two values as

$$p(G|L, D) = \frac{p(L|G, D) \times p(G|D)}{p(L|D)} \tag{5.4}$$

and

$$p(G|L, D, S) = \frac{p(S, L|G, D) \times p(G|D)}{p(S, L|D)}, \tag{5.5}$$

where

$$p(L|G, D) = \frac{\text{\# of venues } u \text{ visited with scores } L \text{ at distance } D}{\text{\# of venues } u \text{ visited at distance } D},$$

$$p(L|D) = \frac{\text{\# of venues with scores } L \text{ at distance } D}{\text{\# of venues at distance } D},$$

$$p(S, L|G, D) = \frac{\text{\# of venues } u \text{ visited with scores } S, L \text{ at distance } D}{\text{\# of venues } u \text{ visited at distance } D},$$

$$p(S, L|D) = \frac{\text{\# of venues with scores } S, L \text{ at distance } D}{\text{\# of venues at distance } D},$$

$$p(G|D) = \frac{\text{\# of venue } u \text{ visited at distance } D}{\text{\# of venues at distance } D}.$$

The venues with highest $p(G|L, D, S)$ (or $p(G|L, D)$) values (those having the best chances to be visited will be our recommended venues - i.e., we rank venues by either of those two probabilities.

5.2.2 Beyond User History

The previous equations model why a user would go to a place depending on user taste, distance and the extent to which the place act as a social mixer. We now try to consider

the additional feature of serendipity by introducing two personalized ways of integrating it in the recommendation process.

The first way (which we call "Beyond User History") constructs a "local preference" graph for each user u, in which vertices are venues that u has visited (H_u), and an edge $e_{i,j}$ between venues i and j exists if the similarity $\mathrm{sim}(i,j)$ from Equation (5.1) is greater than the average similarity among venues user u visited. Well-connected venues form clusters. Venues in the same cluster can belong to a given type, while those connected across clusters are "brokering venues" (i.e., venues that do not necessarily belong to a given type).

As a next step, we add each *candidate* venue x (venue to be potentially recommended) temporarily to u's local preference graph. Again, we create edges $e_{x,i}$ only if $\mathrm{sim}(x,i)$ is larger than the average similarity between venues in u's history. To introduce serendipity, we reward venues that lie on the edge of different venue clusters in u's preference graph, by ranking them by the clustering coefficient $c_{x,u}$ of node x in u's local preference graph.This ranking has in the first positions venues with a lower $c_{x,u}$ value, which are further away from u's central interests; we control the influence of this customized ranking by interpolating it with the basic ranking:

$$\mathrm{ranking}_{x,u} = (1-\alpha)\cdot r_{\mathrm{basic},x,u} + \alpha \cdot r_{\mathrm{history},x}. \tag{5.6}$$

where $r_{\mathrm{basic},x,u}$ is the percentile ranking of user u for venue x produced by the basic algorithm, $r_{\mathrm{history},x}$ is the percentile ranking of venue x sorted by its clustering coefficient in the local preference graph, and α is the interpolation factor that balances the influence of the clustering coefficient over the basic algorithm.

5.2.3 Beyond User Routine

Ordinary human mobility is often a repeated behavior among "home", "office" and "elsewhere" (e.g., gym). Such routine reflects the itinerary triangle described by French sociologist Paul-Henry Chombart de Lauwe in 1952 [33]. Based on this repeated behavior, Eagle and Pentland [37] have shown that it is easy to predict one's location.

Our second way of improving serendipity (which we call "Beyond User Routines") breaks users' itinerary triangles. To do so, we transform the local preference graph seen above into a local *routine* graph. To fix that, we consider temporal aspects that typically characterize movements. We shall see in Section 5.3.4 that venues of different categories have different daily checkin patterns (Figure 5.4). This implies that in u's routine, home, work and "elsewhere" venues will have different temporal checkin patterns. For each venue, we compute a vector with the fraction of checkins that happen in each of the 24 hours of the day. Subsequently, we construct u's routine graph in which vertices are venues that user u visited (H_u); for each pair (i,j) of venues, an edge $e_{(i,j)}$ is added if the cosine similarity of the checkin-per-hour vectors of i and j is *less* than the average.

In other words, the routine graph connects places that have *different* temporal checkin patterns. Venues with similar temporal checkin patterns are likely of the same category, and thus can replace each other (one can visit one today and visit another place at the same hour tomorrow) and cannot be part of one's routine. As a result, connected venues in the routine graph form a set of places in which distinct activities take place.

Similarly to what we have done for the previous algorithm, we then tentatively add each candidate venue x to u's routine graph. We add an edge between x and a venue i if $\mathrm{sim}(x, i)$ is larger than the average similarity between venues in u's history. Finally, we rank candidate venues by their clustering coefficient in the resulting graphs.

Now, a high clustering coefficient for node x means that, for each pair of edges $e_{x,i}$ and $e_{x,j}$, an edge $e_{i,j}$ is likely to exist. Edges $e_{x,i}$ and $e_{x,j}$ exist if $\mathrm{sim}(x, i)$ and $\mathrm{sim}(x, j)$ are high, that is if i and j are visited by similar "tribes" of users; an edge $e_{i,j}$ exists if i and j are generally visited at different times of the day, meaning that they belong to a different category of venues. This implies that a node x has high clustering coefficient if it is visited by the same groups of people that u meets *at different points in time in the day*. By recommending venues with lower clustering coefficient, we bias recommendations towards places different from those where u spends most of her/his time.

Similar to Equation (5.6), here, we also rank candidate venues by their clustering coefficient in u's routine graph, and we interpolate this ranking with the basic algorithm's:

$$\mathrm{ranking}_{x,u} = (1 - \alpha) \cdot r_{\mathrm{basic},x,u} + \alpha \cdot r_{\mathrm{routine},x}. \tag{5.7}$$

where $r_{\mathrm{routine},x}$ is the clustering-based customized ranking.

We have designed so far two basic algorithms to provide accurate recommendations by taking consideration of user taste, physical distance and venue's social mixing. In addition, we have introduced two methods to tackle the possibility of introducing serendipity by promoting places that go beyond those that would be recommended based on past visited places and on one's typical routine.

5.3 Evaluation

We now evaluate our proposals. To do so, we first define two main metrics for recommendation quality (accuracy and serendipity in Section 5.3.1). We then introduce a dataset of checkins from Foursquare (Section 5.3.2). Upon this dataset, we evaluate the extent to which the different proposed techniques are able to introduce serendipity without compromising recommendation accuracy (Sections 5.3.3 and 5.3.4).

5.3.1 Evaluation Metrics

The goal of this work is to introduce serendipity into mobile recommendations, whilst ensuring high accuracy. To validate the final recommendation quality, we evaluate the recommendation output against two aspects: *accuracy* and *serendipity*.

Accuracy. To evaluate to which extent our system can recommend a venue that one might visit, we adopt a recall-oriented accuracy metric - *percentile ranking*, which has been applied in similar situations [51, 95, 102, 133]. Percentile ranking evaluates the position of a hit (a recommended item that the user indeed visited) in the recommended list, regardless of the size of recommendation list or the number of users. We measure the accuracy of a recommender system by the overall average percentile ranking as:

$$\overline{\text{rank}} = \frac{\sum_{i,u \in \mathcal{T}} \text{gone}_{i,u} \times r_{i,u}}{\sum_{i,u \in \mathcal{T}} \text{gone}_{i,u}} \tag{5.8}$$

where $r_{i,u}$ is the percentile-ranking of a venue i in the ordered list of recommendations for user u, and $\text{gone}_{i,u}$ is an indication of whether user u visited venue i. A perfect accurate recommendation yields 0.0 as $\overline{\text{rank}}$, while random guess is 0.5. We then define our accuracy metric as:

$$\text{accuracy} = (0.5 - \overline{\text{rank}}) \times 2 \tag{5.9}$$

So, a random guess will receive 0.0 as accuracy, and 1.0 indicates a perfectly accurate recommender.

Serendipity. The serendipity of a recommendation list reflects the extent to which "surprises" and "unexpectedness" are brought into a user's list, compared to his/her past visits. Some researchers argue that such "unexpectedness" can be measured as deviation from mature predictions [76]. That is, a high serendipitous recommendation list contains items that are difficult to predict. Others hold that a serendipitous recommender system can reward items that are contextually distant from one's past preferences [107, 131]. A similar idea was already implemented to measure the overall serendipity of a recommendation list by looking at the average distance between the profile of recommended items and one's previously preferred items [133]. We quantify the total "surprise" the new recommendations bring using Kullback-Leibler divergence (*KL* divergence) [71], which is often used to quantify one's information gain. Therefore, the "surprise" a recommended venue i carries with respected to user's past visited venue h could be computed as:

$$dvg(i,h) = \sum_{t \in \mathcal{T}} w_{i,t} log \frac{w_{i,t}}{w_{h,t}} \tag{5.10}$$

in which $\mathcal{T}$ is the full set of user tribes, $w_{i,t}$ is the importance of user tribe t to the venue i and $w_{h,t}$ is the importance of user tribe t to the venue h. Finally, the serendipity of a recommender system is the overall average serendipity of the recommendation lists

generated for all the test users ($\mathcal{U}$):

$$\text{serendipity} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{H}_u|} \sum_{h \in \mathcal{H}_u} \sum_{i \in \mathcal{S}_{u,n}} \frac{\text{dvg}(i,h)}{n} \tag{5.11}$$

where $\mathcal{H}_u$ is user u's past visited venues, $\mathcal{S}_{u,n}$ represents a list of n recommendations to user u, and $\text{dvg}(i,h)$ is the *KL* divergence of the recommended venue i with respect to the past visited venue h.

5.3.2 Data

Having defined the metrics, we now assess the extent to which our algorithms are able to make serendipitous recommendations. We do so upon a dataset containing Foursquare checkins in London. We use Foursquare checkins (e.g., announcement of arrival/presence [70]) of users who published them on Twitter. This dataset was collected by Cheng *et al.* [26]. To it, we add venues' geographic locations and categories by crawling the Foursquare open API. The dataset is very sparse, in that, it does not contain a user's entire set of checkins, due to the crawling constraints from Twitter.

We code whether a user visited a venue into one of the binary cells of the preference matrix, and that matrix (once completed) will be the recommender system's input. We consider users who visited at least two venues, and venues that have at least two users. This results into 28,791 (user, venue) pairs (user u visiting venue i) for a total of 3,293 users and 3,137 venues. The sparsity of the corresponding preference matrix is as severe as 0.003 (i.e.. the ratio of (user, venue) pairs for which a checkin appears in the dataset). Only few users visited multiple venues, and a few venues have been visited by multiple visitors. Upon this data, we answer two questions:

1. How do geographic distance, venues' social mixing propensities, temporal patterns impact one's decision to visit a place (Section 5.3.3). We are interested in those three aspects because they are the building blocks of our modeling (Section 5.2.1).

2. What is the trade-off between accuracy and serendipity of our proposals? (Section 5.3.4).

5.3.3 Validating Modeling Assumptions

Distance. Distance is one of the main factors that impact one's decision to visit a venue. People might be willing to travel far to visit certain venues (e.g., Michelin star restaurants) more than others (e.g., petrol stations). To see to which extent that is true, we compute the probability of an individual to visit a venue at a certain distance, and do so across all our venue categories (Figure 5.1). The probability distributions approximately

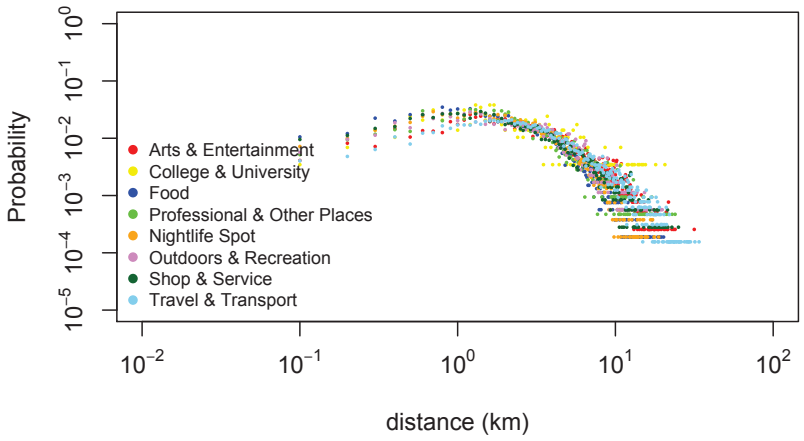


Figure 5.1: Probability of a user moving at a certain distance.

follow the distribution (as also shown in [84, 102]):

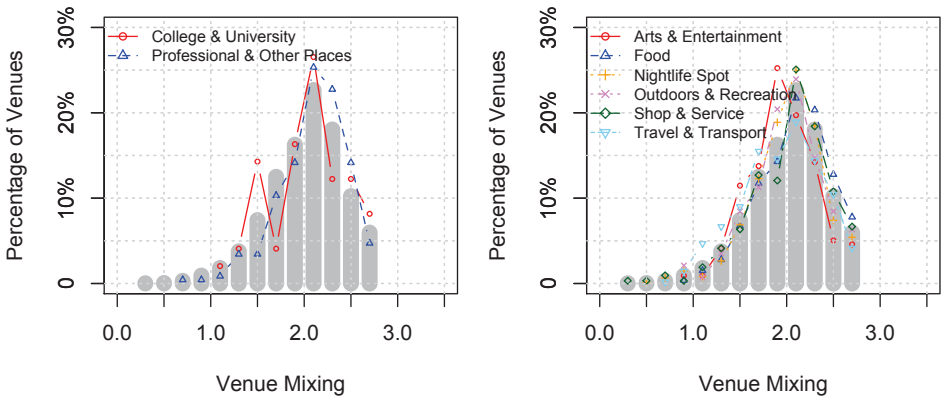
$$p_{\text{close}} = k\frac{1}{d_{u,i}^{\gamma}} \tag{5.12}$$

where $d_{u,i}$ is the distance between venue i and user u's center of interest. What changes across categories is the decay factor α - high values for it are associated with short-range trips (e.g., walk to a bar, dining at a local restaurant), while low values are associated with long-range trips (e.g., going to university) . Table 5.1 shows that across the different categories: high values of decay (i.e., short-range trips) are associated with categories such as "Food" (1.59) and "Nightlife Spot"(1.49), and the lowest value (i.e., long-range trips) with"College & University" (0.49).

Category	α
Arts & Entertainment	1.19
College & University	0.49
Food	1.59
Professional & Other Places	1.17
Nightlife Spot	1.49
Outdoors & Recreation	1.08
Shop & Service	1.30
Travel & Transport	1.33

Table 5.1: One's unwillingness of traveling far to visit venues of each category. The higher α, the shorter the trip to a venue for a given category.

Venue's Mixing. The frequency distribution of mixing values as defined in Equation 5.3 for our venues is skewed (Figure 5.2): few venues target specific tribes (e.g., dance school), while many attract a mix of them (e.g., shopping mall). This holds across categories (Table 5.2). Despite mixing values being similar across categories, there are still differences within each category: for example, for nightlife spots, the highest mixing value is 2.72, while the lowest is 0.47. Finally, to test whether users are attracted to more (less) mixing venues, we plot the probability of a user going to a venue with a given



(a) Work/College Venues (b) Remaining Venues

Figure 5.2: The distribution of a venue's mixing (i.e., ability to attract all sorts of "user tribes" among the k tribes) by category. Grey bars reflect the overall *average* distribution of a venue's mixing (without any distinction by category).

mixing value (Figure 5.3):

$$p(\text{mixing} \mid \text{go}) = \frac{\#\text{venues } u \text{ visited with mixing } m}{\#\text{venues visited by } u} \tag{5.13}$$

The two peaks in Figure 5.3 suggest that, on average (for both mean and median), people tend to visit either niche places (mixing value of 0.6) or high-mixing (likely popular) places (mixing value of 2.4). This observation motivates us later on to separate users according to their different tendencies for social mixing in the experiment of introducing serendipitous recommendations.

Category	Median	Mean	Min	Max
Arts & Entertainment	1.95	1.95	0.88	2.73
College & University	2.05	2.03	1.03	2.69
Food	2.08	2.08	0.93	2.72
Professional & Other Places	2.01	2.09	0.72	2.72
Nightlife Spot	2.04	2.01	**0.47**	**2.72**
Outdoors & Recreation	2.02	2.01	0.94	2.68
Shop & Service	2.04	2.02	**0.40**	**2.71**
Travel & Transport	1.97	1.92	0.78	2.71

Table 5.2: Mixing of venues per category.

Routine. We aggregate the checkins from the 3,293 users to 3,137 venues (including repeated checkins at a same place) and show their daily patterns in Figure 5.4. There are three main peak hours: around 7am in the morning, 12am at lunch time, and 5-6pm in the afternoon (Figure 5.4(a)). This overall daily pattern is coherent with what has been already found in other major cities like New York City and Los Angeles [26].

Considering one's routine is constituted by "home", "work" and "elsewhere" locations, we then break down the daily checkin patterns into each category of venues. For the

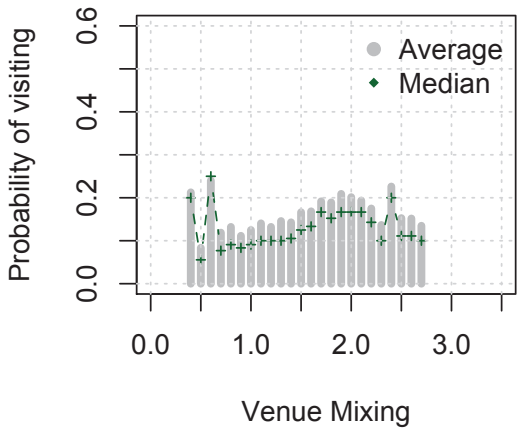


Figure 5.3: Probability of visiting venues by mixing value.

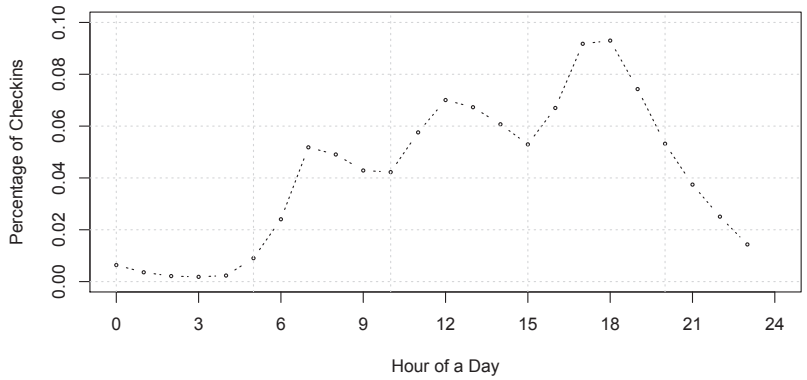
"invariant" routines of venues (home/work/college), it shows that people tend to start their day by checking at their residence places at around 6am, followed by arrivals at work (school) at 8am (9am). Popular checkins are also observed after work (school) at 6pm/8pm (Figure 5.4(b)).

Different daily checkin patterns are observed in other categories of venues (Figure 5.4(c)). Peak hours for nightlife hotspots are registered from 6pm; and restaurants are visited at lunch (peak at 12am) and at dinner (at around 6-7pm). Art and entertainment venues also receive most checkins at 12am and 6pm, but the volume of checkins during the afternoon does not drop as much as restaurants'. Shops and recreation places share similar patterns -they are mostly visited during opening hours (from 11am to 5pm), and that stays constant throughout the day. Places dedicated to public transportation shows checkin peaks corresponding to peak commuting hours (7am and 5pm). All of this suggests that the temporal patterns upon which we based our proposal "Beyond User Routine" (Section 5.2.3) do exist and tend to be consistent.

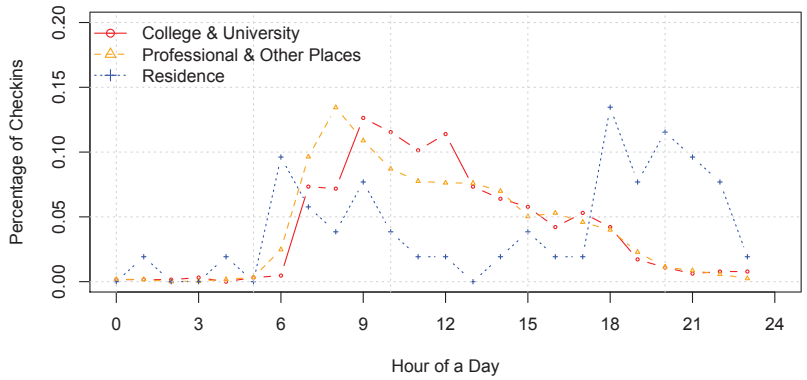
From the Foursquare checkin dataset, we have seen that people do consider distance when they decide to visit a place, and the degree to which it is concerned varies among different categories of venues. In addition, people exhibit different tendencies for social mixing. Some prefer to visit niche places, while some like high-mixing places better. These two observations have empirically confirmed that in a location recommender system, user taste is not the only reason why a user adopts an item (visiting a place), physical distance and venues' social mixing might also matter.

5.3.4 Accuracy vs. Serendipity

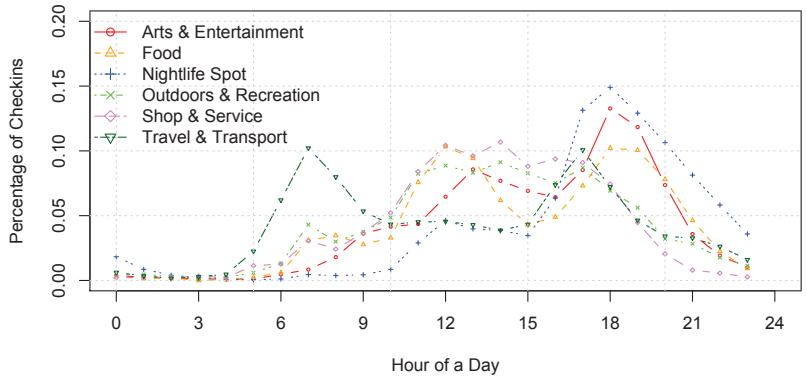
Three basic Bayesian models. We now evaluate our basic Bayesian models described in Section 5.2.1. We consider the three variants:



(a) All



(b) Home/Work/College Venues



(c) Other Venues

Figure 5.4: Percentage of Checkins per Hour of a Day.

i+d (item-based + distance). In this algorithm, we compute recommendations without taking into account the "social mixing" feature, that is by adopting Equation A.6. In addition, we don't take advantage of the LDA-based user clustering: each user counts as a different tribe, resulting in a set of $l_{u,i}$ scores that reflect a traditional item-based recommender algorithm. We consider this as our baseline algorithm.

i+d+s (item-based + distance + social mixing). Here, we improve the previous algorithm by taking into account the social mixing feature: rankings are based on the score of Equation A.7.

L+d+s (LDA + distance + social mixing). This last algorithm combines all the features that we have proposed: here, we introduce the LDA modeling user in tribes [2].

For the three algorithms, we obtain the values of L, D and S so that each discrete class empirically obtains a sufficient number of samples, as follows: $L = \lfloor 100 \times l_{u,i} \rfloor$, $D = \lfloor \log_{10} d_{u,i} \rfloor$, $S = \lfloor s_{u,i} \rfloor$.

Model	Features	Accuracy	Serendipity
Baseline (i+d)	item-based + distance	0.195 $\pm$0.048	3.288 $\pm$0.017
(i+d+s)	item-based + distance + social mixing	0.226 $\pm$0.061	**3.359** $\pm$0.019
Full (L+d+s)	LDA + distance + social mixing	**0.478** $\pm$0.034	3.175 $\pm$0.020

Table 5.3: Accuracy and Serendipity of our three basic algorithms. For LDA in the last model, the number of tribes k is set to 100 because of its best accuracy compared to other k's.

Accuracy vs. Serendipity of the three basic models. To avoid easily-predictable venues, we consider all venues other than those in the category residences/work/education places. Given the severe sparsity of our data (Section 5.3.2), we cannot execute our evaluation using cross-validation. We thus resort to leave-one-out [35], in that, we randomly withhold one checkin venue for each user, and leave the rest as the training set. We measure accuracy and serendipity defined in Section 5.3.1, and Table A.6 shows the results. Compared to the baseline *(i+d)*, the *(i+d+s)* method (which considers the "social mixing" feature) increases accuracy. The *LDA*-based model performs best in terms of accuracy and also strikes the right balance between accuracy (which is twice that of the item-based model *(i+d+s)*) and serendipity (which is comparable to item-based model's). These results refer to the case in which the number k of user tribes in *LDA* is set to 100. We show the results corresponding to that value because 100 happens to return the most accurate recommendations (Figure 5.5).

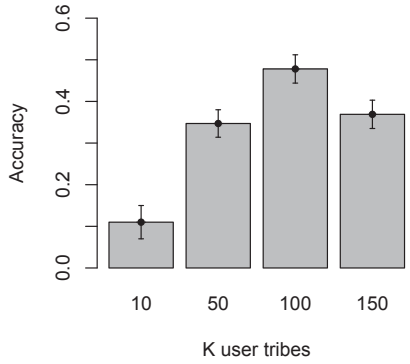
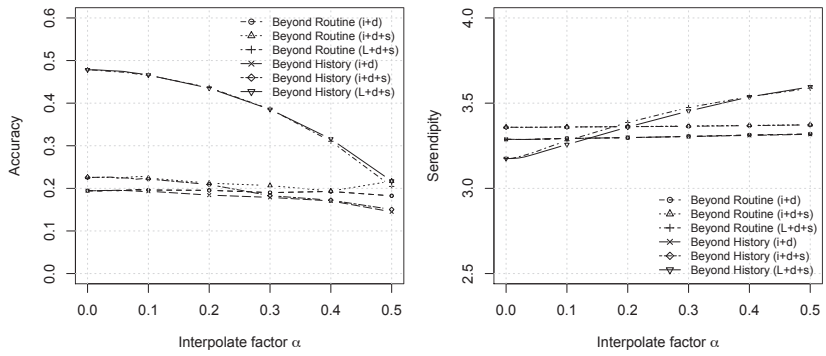


Figure 5.5: Accuracy of the Bayesian model based on LDA

[2] In our empirical study, we use the LDA implementation from Mallet (`http://mallet.cs.umass.edu/`) framework.



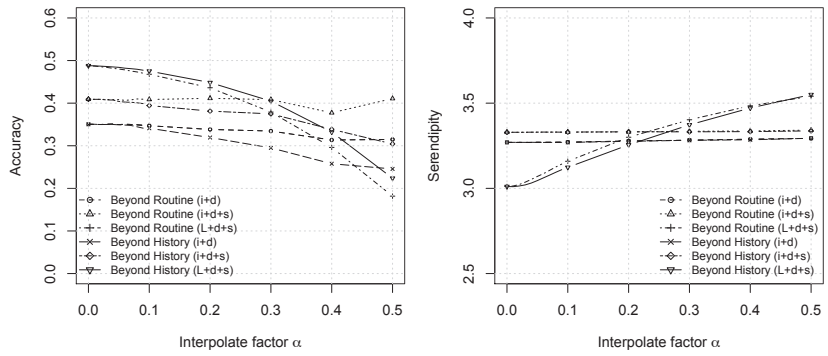
(a) Accuracy (b) Serendipity

Figure 5.6: Accuracy and Serendipity of (top 10) Recommendations. This considers all users (i.e., users who visited at least two venues).

Accuracy vs. Serendipity of the two serendipity-enhanced models. The *LDA*-based model returns better accuracy than the item-based one (Figure A.5(a)). The interpolation factor α does not impact the item-based's serendipity but does impact the *LDA*-based model's; on the other hand, it is possible to enhance the serendipity of the recommended items by raising the α parameter, as it can be seen in Figure A.5(b). The drawback is that, by raising α, accuracy decreses. A good tradeoff appears to be a value of $\alpha \in [0.2, 0.3]$, where the *LDA*-based model offers relatively high accuracy (well above 0.4) as well as high serendipity (above 3.4).

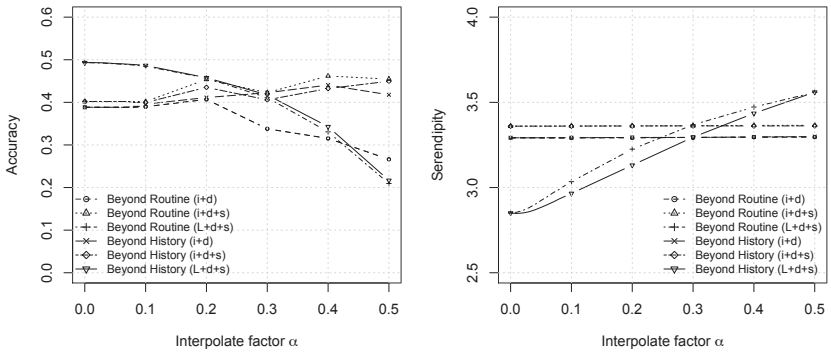
Impact of user activity. To test how users' activity levels impact the results, we consider users who visited at least 5 and those who visited at least 10 venues. In those situations of lower data sparsity, the item-based model increases its accuracy (Figures A.6 and A.7), but does not reach that of the *LDA*-based model in high data sparsity situation (Figure A.5) - accuracy is below 0.40. More generally, this suggests that *LDA*-based model's accuracy is indeed more robust to data sparsity. User activity does not impact the metric of serendipity.

Impact of users with different tendencies for social mixing. We consider three types of users: those whose average social mixing values of their visited venues is in the first quartile (niche users), those whose average is in the last quartile (social mixers), and those remaining (average mixers). For these three types, we do not register any change for the serendipity metric across all algorithms. By contrast, the accuracy metric shows some differences. For niche users, the item-based model shows higher accuracy than the LDA-based model (Figure A.8(a)). This may be because grouping users in the compact representation of "tribes" may lead to information loss, in that, it may dilute information about their specificities. The opposite holds for social mixers (Figure A.8(c)): considering the fine-grained variety of other users mixers meet boosts recommendation accuracy.



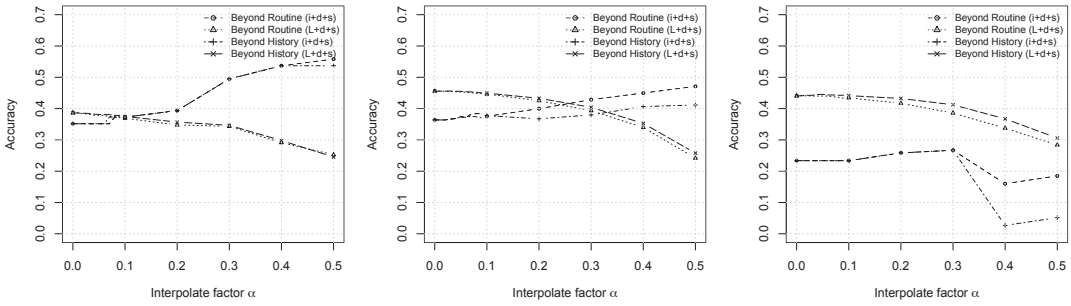
(a) Accuracy (b) Serendipity

Figure 5.7: Accuracy and Serendipity of (top 10) Recommendations. This considers users who visited at least 5 venues.



(a) Accuracy (b) Serendipity

Figure 5.8: Accuracy and Serendipity of (top 10) Recommendations. This considers users who visited at least 10 venues.



(a) Niche (b) Average Mixers (c) Social Mixers

Figure 5.9: Impact of users with different tendencies for social mixing.

5.4 Discussion

Active Learning. Recommendations not only need to capture user tastes (i.e., be accurate) but also need to be diverse and suggest things users are not familiar with. However,

different users enjoy different levels of serendipitous encounters: forcing the same level upon all users could be perceived as a patronizing act by some [133]. This work has proposed to make the trade-off between serendipity and accuracy tunable. In the future, based on how a user reacts to certain recommendations (e.g., whether the user tends to follow more or less serendipitous recommendations), the proposed algorithm could be extended, in that, it could tune itself automatically based on its user's reactions.

Whole Past vs. Routine Triangle. We have proposed two ways to increase serendipity in mobile recommendations. The first avoids recommending places similar to those visited in the past. The second way avoids recommending places that are similar to a user's routine (triangle). The two show very similar accuracy and serendipity values, and that might be because of data sparsity - one's limited past (e.g., 3 or 4 distinct visits) could look like one's routine triangle. In the future, it would be interesting to understand in which situations these two ways tend to perform similarly and in which they tend to be complementary.

Data Sparsity. Data sparsity is an open problem in recommender systems. We have shown that the technique based on *LDA* effectively counters sparsity (accuracy is far better than an item-based model's) and allows for novel recommendations (with an interpolation parameter between 0.2 and 0.3, the technique makes not only accurate but also serendipitous recommendations).

Scalability and Explanability. The proposed Bayesian models are scalable as they depend on coefficients (e.g., distance decay factor γ) that can be set offline. Therefore, the online computation can scale reasonably well with a growth of the number of users. The resulting recommendations are also easy to explain as one can easily tease apart which recommendations depend on similar past visits and which on one's geographic center of interest.

5.5 Summary

In this Chapter, we have designed a mobile recommender system that produces not only accurate but also serendipitous recommendations. Through experimental analysis, we have observed that there are two classes of users in a location sharing social media: those who go to niche venues (i.e., places visited by like-minded users) and those who go to popular venues. Different users are comfortable with different levels of serendipity. That is why we made the trade-off α between accuracy and serendipity tunable. To counter data sparsity, we have proposed an approach based on *LDA*, which groups like-minded users in the same "user tribes", increasing recommendation accuracy, especially for users who have not been very active. Also, characterizing venues by a "social mixing" fea-

ture (i.e., its tendency to be visited by different user tribes) increases recommendation accuracy.

Chapter 6

Conclusion and Future Work

This Thesis addresses an important problem in the analysis of social media, that of understanding who are creating trending content in the networked world. Prior research efforts on trending content have focused on two parallel perspectives: trends - the entity itself, and the source of trends - people who create trends. While the first research direction has mainly focused on burst detection techniques to capture trends by their special property of the "tipping point", the work in second research direction has been centered around the core concept of being "influential". And the discussion about whether the people who create trends are "influential" exhibits different views.

However, influence - as the power to persuade others to accept one's idea - is a function of people, content and environment (e.g., activity, locations etc). Building upon previous findings about influentials, we redefined in this Thesis two types of people who contribute to the creation of trends - trend makers (those who generate trends) and trend spotters (those who spread trends). Through an in depth analysis of a variety of features of trend makers and trend spotters - activity, content, social connections and geographical ones (Chapter 3), we have shown that trends are indeed created by "special" users, and in social media sites, they seem to be *many* rather than *a few*.

The appealing characteristics of trend makers and trend spotters make them distinguishable from the remaining typical users, which gives us the insights on the underlying causes of trends in social media. Moreover, these notable special users provide an opportunity to identify trends in a novel way - to rely on the wisdom of the origins (i.e., the people who create trends). We have shown how this idea can be integrated and used as one of the major components to a trend-aware recommender system (Chapter 4), which is shown to be able to serve personalized trends effectively to individuals with different interests.

An accuracy focused recommender system is often expected to learn perfectly one's preferences and to output the most "close" recommendations - those that are in the center of one's tastes. Actively adopting these "accurate" recommendations gradually narrows

down the range of one's exploration. To expand one's scope of recommendations, various network analysis strategies are investigated in Chapter 5. Specifically in a location recommender system, we have shown that by leveraging network analysis techniques, accuracy and serendipity could be balanced.

6.1 Thesis Contributions

In this thesis, we have: 1) analyzed the human factors in the creation of trends; 2) studied to which extent the special users could be used to identify trends; and 3) deployed the tools to make personalizations in the mobile social media. Our overall contributions are relevant to two main research topics - trends and recommender systems in social media.

Trends

In the search of the answers to our research problems related to trends, we make the following three main contributions.

Human Factors. Our analysis have uncovered that the creation of trends is a combined process, in which two types of users are involved - regular individuals with specific interests who are connected with different user clusters (defined as trend makers in this thesis) and early adopters with diverse interests (defined as trend spotters). Both of them are able to be identified from the remaining typical users with a handful features (i.e., activities, content consumptions, social network connections and geographical features, etc.) using standard machine learning tools such as SVM or a logistic regression model.

Accounting the fact that not every one is equally good at generating trends, we have grouped trend makers/spotters into three classes (i.e., high, medium and low) based on their different levels, and have extended the statistical models to successfully identify which class of trend maker/spotter one belongs to.

Identifications. Based on these identified trend makers and trend spotters of different levels, we have shown trends could be identified by using a logistic regression model. At the same time, the coefficients of the model indicated that an item has a good chance to become a trend if its uploader is a trend maker of high level, and if it receives great attentions from successful trend spotters.

Explorations. Finally, incorporating the statistical models to identify trends using a collaborative filter technique (i.e., implicit SVD), we have designed a trend-aware recommender system to effectively help users discover trending content close to their preferences.

Recommender Systems

In parallel to the research on trends in social media sites, our work in this Thesis also contributes to the state of the art in recommender systems. Specifically, we have examined the design of recommender systems tailored to two different *mobile* social networks.

Accuracy Focused. First, we have designed a trend-aware recommender system to serve users with trending content of their interest. On top of the common practice of building a recommender system, we propose a novel means to enrich the traditional user item preferences matrix by converting it to "trend-aware". Aiming at making *accurate* recommendations about trending content, the system is shown to be able to recommend trends effectively. Moreover, we show that recommending trends outperforms recommending popular content in general.

Serendipity Enhanced. Second, we have explored the potential improvement introduced by serendipity to the quality of recommendations. Specifically, we have performed the experiment in the context of a a mobile recommender system where *location* is the primary type of information. We designed the location recommender system by incorporating one's preferences and physical distances using a Bayesian model. While our strategies to promote locations beyond the recommendations from like-minded users and one's routine are shown to be effective in enhancing serendipity, our analyses on users show that people prefer different levels of serendipity. Such differences obviously should to be taken into consideration in the process of balancing the accuracy and serendipity.

While recommender systems are designed as the tools to make personalizations of "long tail" items (a large number of items that are have relatively small quantity of adoptions of each [7]), our work in this Thesis shows that it also makes sense to personalize the non-long tail items like trending content in social media. In addition, the design of a recommender system should definitely be application specific, that is, to be tailored according to the context (such as the type of items, the reasons for one to adopt an item, and the level of one's adoptions etc).

The work in this Thesis has immediate impact on researchers that are interested in understanding information dissemination in social media in general, especially for those who are interested in: 1) identifying events (e.g., trending news); 2) understanding opinion spreading; and 3) designing viral marketing strategy etc. It is also relevant to practitioners who are trying to boost user experiences in personalized information consumption.

6.2 Future Work

This work builds upon two datasets from two real mobile social networks. While we have good confidence in the generality of our observations, it would be interesting to

conduct a comparative study across different social networks to understand the impact of dealing with different usage patterns and different types of individuals. Moreover, this thesis could be extended from the following directions:

Geography. As online social networks becomes an important part of our daily life, a social media report [80] claims that *"when it comes to accessing social content, it's all about mobile"*. The most appealing feature that mobile applications carry is that they are location-aware.

While the popular theory of "six degrees of separation" [116] says that we live in a small world where everyone could be connected to another person within six steps, research work revealed that people's social connections and mobility are still constrained by geographical distances [27, 68, 84]. Since content dissemination along social connections is vital to the creation of trends, such geographical constraints might effect trends as well, and as such bound the diffusion process. A typical example is related to the concept of event identification. In such setting, a trend could be a global event, or it could be a local event as well [129]. Limited by the size of dataset, our work on trends has been focused on global trending content in the entire mobile social application. However, in a larger mobile social application, it would be interesting to separate the local trends from the global ones. Controlling these two classes of trends, one could study: 1) what are the major underlying differences between the creation of global trends and local trends? 2) whether there exist global/local trend makers and trend spotters? 3) to which extent global/local trend makers and trend spotters contribute to the creation of trends of different levels?

Temporal Dynamics. Temporal dynamics is another important factor that to be considered. Related to our work, there are two types of temporal dynamics could studied: temporal dynamics in trends and temporal dynamics in user preferences.

- **Trends.** In Chapter 4, we have seen that trends tend to persist for a longer time than normal content, but the volume of attention they gain stops to increase after a period. However, in some cases (e.g., when a trend is a fashion fad), the life of a trend could be cyclic [1]. Taking the temporal complexity into consideration, the studies about individuals in the creation of trends could be further extended, as well as the model to identify trends by leverage the knowledge of their creators.

- **User preferences.** User preferences shift with time, that is a well-known problem in the research studies on recommender systems. Our trend-aware recommender system could be improved by integrating it with an extra component to model user preferences shift. Moreover, shift in user preferences might also lead to the change in the level of serendipity one accepts. An in-depth study could be performed to understand such impact, and then be counted in the practice of generating serendipitous recommendations.

Online Updating. A recommender system learns one's preferences. However, it impacts one's choice in adopting the items. As we have discussed in Chapter 4, in our

trend-aware recommender system, individuals are recommended with trending content of their preferences. The consequence is that in the course of accepting the recommended trending content, users are "trained" to be trend spotters. As the system relies on identifying trend spotters of different levels of success, the identification models then need to be updated periodically. But up to which point the models are insufficient to serve the purpose of accurate identifications, it requires further analysis. A recent proposal to address this problem that could be employed in our context, is to integrate controllers to automatically estimate the frequency of updating models [53].

Sentiment Analysis. In this thesis we have observed that people behave differently. For instance, significantly different behaviors were observed among trend makers, trend spotters, and typical users. And in the reaction to recommendations, people are also found to accept different levels of serendipity. One might ask what are the fundamental causes to all the diverse human behaviors?

Similar questions have been asked in the research of individuals' different power of social influences, and explanations were sought by looking into the divergence of personality traits [41, 93, 94]. However, whether personality traits could be used to explain the different behaviors among trend makers, trend spotters and typical users still needs to be examined.

Moreover, some researchers started to make recommendations based on user personalities [50, 88]. While our work focused on the modeling one's personality from his/her past ratings or preferences, it would be also interesting to investigate whether personality could explain individuals' different acceptance levels of serendipitous recommendations, and thus could be also put into practice while tuning the balance between accuracy and serendipity in personalized recommendations for people with different personality traits.

Bibliography

[1] A. Acerbi, S. Ghirlanda, and M. Enquist. The Logic of Fashion Cycles. *PloS one*, 7(3), 2012.

[2] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 2005.

[3] J. Allan. Introduction to Topic Detection and Tracking. In *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers, 2002.

[4] X. Amatriain, A. Jaimes, N. Oliver, and J. M. Pujol. Data Mining Methods for Recommender Systems. *Recommender Systems Handbook*, pages 39–71, 2011.

[5] X. Amatriain, N. Lathia, J. Pujol, H. Kwak, and N. Oliver. The Wisdom of the Few: A Collaborative Filtering Approach Based on Expert Opinions from the Web. In *Proceedings of the* 32^{nd} *ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 2009.

[6] X. Amatriain, J. Pujol, and N. Oliver. I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. *User Modeling, Adaptation, and Personalization*, pages 247–258, 2009.

[7] C. Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, 2008.

[8] Aral, S. and Walker, D. Identifying Influential and Susceptible Members of Social Networks. *Science*, 337, 2012.

[9] S. Asur, B. A. Huberman, G. Szabo, and C. Wang. Trends in Social Media: Persistence and Decay. In *Proceedings of the* 5^{th} *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[10] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the* 4^{th} *ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2011.

[11] M. Balabanović and Y. Shoham. Fab: Content-Based, Collaborative Recommendation. *Communications of the ACM*, 1997.

[12] A.-L. Barabasi. The Origin of Bursts and Heavy Tails in Human Dynamics. *Nature*, 2005.

[13] H. Becker, M. Naaman, and L. Gravano. Learning Similarity Metrics for Event Identification in Social Media. In *Proceedings of the 3^{rd} International Conference on Web Search and Data Mining (WSDM)*, 2010.

[14] H. Becker, M. Naaman, and L. Gravano. Beyond Trending Topics: Real-world Event Identification on Twitter. In *Proceedings of the 5^{th} International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[15] D. M. Blei and J. D. Lafferty. Dynamic Topic Models. In *Proceedings of the 23^{rd} International Conference on Machine learning (ICML)*, New York, NY, USA, 2006. ACM.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[17] F. Bodendorf and C. Kaiser. Detecting Opinion Leaders and Trends in Online Social Networks. In *Proceedings of the 2^{nd} ACM workshop on Social Web Search and Mining (SWSM)*, 2009.

[18] J. Borge-Holthoefer, A. Rivero, I. García, E. Cauhé, A. Ferrer, D. Ferrer, D. Francos, D. Iñiguez, M. Pérez, G. Ruiz, et al. Structural and Dynamical Patterns on Online Social Networks: the Spanish May 15th Movement as a Case Study. *PLoS One*, 6(8), 2011.

[19] J. S. Breese, D. Heckerman, and C. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14^{th} ACM Conference on Uncertainty in Artificial Intelligence (UAI)*, 1998.

[20] R. Burt. The Social Capital of Opinion Leaders. *The Annals of the American Academy of Political and Social Science*, 566(1), 1999.

[21] Ò. Celma and P. Herrera. A New Approach to Evaluating Novel Recommendations. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys)*, pages 179–186. ACM, 2008.

[22] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4^{th} International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[23] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make New Friends, But Keep the Old: Recommending People on Social Networking Sites. In *Proceedings of the 27^{th} ACM International Conference on Human Factors in Computing Systems (CHI)*. ACM, 2009.

[24] W.-Y. Chen, J.-C. Chu, J. Luan, H. Bai, Y. Wang, and E. Y. Chang. Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior. In *Proceedings of the 18^{th} International Conference on World Wide Web (WWW)*. ACM, 2009.

[25] Z. Cheng, J. Caverlee, and K. Lee. You are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In *Proceedings of the 19^{th} ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2010.

[26] Z. Cheng, J. Caverlee, K. Lee, and D. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the 5^{th} International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[27] E. Cho, S. A. Myers, and J. Leskovec. Friendship and Mobility: User Movement In Location-Based Social Networks. In *Proceedings of the* 17^{th} *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 2011.

[28] R. Crane and D. Sornette. Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System. In *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, October 2008.

[29] danah boyd. The Future of Privacy: How Privacy Norms Can Inform Regulation. Invited Talk at the 32^{nd} International Conference of Data Protection and Privacy Commissioners, October 2010.

[30] danah boyd. Designing for Social Norms (or How Not to Create Angry Mobs). http://www.zephoria.org/ thoughts/archives/2011/08/05/design-social- norms.html, August 2011.

[31] A. Das, M. Datar, A. Garg, and S. Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of the* 16^{th} *ACM International Conference on World Wide Web (WWW)*, 2007.

[32] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, et al. The youtube video recommendation system. In *Proceedings of the* 4^{th} *ACM Conference on Recommender Systems (RecSys)*, 2010.

[33] P. de Lauwe. *Paris et l'agglomération parisienne*. Presses Universitaires de France, 1952.

[34] J. Delgado and N. Ishii. Memory-Based Weighted Majority Prediction. In *Proceedings of the* 22^{nd} *ACM International Conference on Research and Development in Information Retrieval (SIGIR) - workshop on recommender systems*. ACM, 1999.

[35] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22, 2004.

[36] C. Droge, M. Stanko, and W. Pollitte. Lead Users and Early Adopters on the Web: The Role of New Technology Product Blogs. *Journal of Product Innovation Management*, 27(1), 2010.

[37] N. Eagle and A. Pentland. Eigenbehaviors: Identifying Structure in Routine. *Behavioral Ecology and Sociobiology*, 63, 2009.

[38] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[39] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Little, Brown and Company, 2000.

[40] N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated Trend Discovery for Weblogs. In *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics*, 2004.

[41] J. Golbeck, C. Robles, and K. Turner. Predicting Personality with Social Media. In *Proceedings of the* 29^{th} *ACM Conference on Human Factors in Computing Systems (CHI)*, May 2011.

[42] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using Collaborative Filtering to Weave An Information Tapestry. *Communications of the ACM - Special Issue on Information Filtering*, 1992.

[43] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval*, 4(2), 2001.

[44] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. The Dynamics of Protest Recruitment through an Online Network. *Scientific reports*, 1, 2011.

[45] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning Influence Probabilities in Social Networks. In *Proceedings of the* 3^{rd} *International Conference on Web Search and Data Mining (WSDM)*, 2010.

[46] F. A. Haight. *Handbook of the Poisson Distribution*. Wiley, 1967.

[47] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: Visualizing Theme Changes over Time. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, 2000.

[48] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An Algorithmic Framework for Performing Collaborative Filtering. In *Proceedings of the* 22^{nd} *International ACM Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 1999.

[49] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions of Information Systems*, 2004.

[50] R. Hu and P. Pu. Acceptance Issues of Personality-Based Recommender Systems. In *Proceedings of the* 3^{rd} *ACM Conference on Recommender Systems (RecSys)*. ACM, 2009.

[51] Y. Hu, Y. Koren, and C. Volinsky. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the* 8^{th} *IEEE International Conference on Data Mining (ICDM)*, 2008.

[52] M. Jamali and M. Ester. A Matrix Factorization Technique with Trust Propagation for Recommendation in Social Networks. In *Proceedings of the* 4^{th} *ACM Conference on Recommender Systems (RecSys)*, 2010.

[53] T. Jambor, J. Wang, and N. Lathia. Using Control Theory for Stable and Efficient Recommender Systems. In *Proceedings of the* 21^{st} *International Conference on World Wide Web (WWW)*, 2012.

[54] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as Online Word of Mouth Branding. In *Proceedings of the* 27^{th} *International Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2009.

[55] G. Jawaheer, M. Szomszor, and P. Kostkova. Characterisation of Explicit Feedback in An Online Music Recommendation Service. In *Proceedings of the* 4^{th} *ACM Conference on Recommender Systems (RecSys)*. ACM, 2010.

[56] N. Jones and P. Pu. User Technology Adoption Issues in Recommender Systems. *Proceedings of Networking and Electronic Commerce Research Conference (NAEC)*, pages 379–39, 2007.

[57] D. Kempe, J. Kleinberg, and E. Tardos. Influential Nodes in a Diffusion Model for Social Networks. In *Proceedings of the* 32^{nd} *International Conference on Automata, Languages and Programming (ICALP)*, 2005.

[58] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the* 8^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2002.

[59] N. Koenigstein, G. Dror, and Y. Koren. Yahoo! Music Recommendations: Modeling Music Ratings with Temporal Dynamics and Item Taxonomy. In *Proceedings of the* 5^{th} *ACM conference on Recommender systems (RecSys)*. ACM, 2011.

[60] J. A. Konstan, S. M. McNee, C.-N. Ziegler, R. Torres, N. Kapoor, and J. T. Riedl. Lessons on Applying Automated Recommender Systems to Information-Seeking Tasks. In *Proceedings of the National Conference on Artificial Intelligent*. AAAI Press, 2006.

[61] I. Konstas, V. Stathopoulos, and J. Jose. On Social Networks and Collaborative Recommendation. In *Proceedings of the* 32^{nd} *International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2009.

[62] W. Kornblum and C. Smith. *Sociology in a Changing World*. Wadsworth Publishing Company, 2007.

[63] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proceedings of the* 19^{th} *ACM Conference on World Wide Web (WWW)*, 2010.

[64] P. Lazarsfeld and E. Katz. *Personal Influence: the Part Played by People in the Flow of Mass Communications*. Glencoe, Illinois, 1955.

[65] D. Lemire and A. Maclachlan. Slope One Predictors for Online Rating-Based Collaborative Filtering. In *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM)*, 2005.

[66] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the* 15^{th} *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 2009.

[67] H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Chang. Pfp: Parallel FP-growth for Query Recommendation. In *Proceedings of the* 2^{nd} *ACM Conference on Recommender Systems*, 2008.

[68] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic Routing in Social Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33), 2005.

[69] G. Linden, B. Smith, and J. York. Amazon.com Recommendations: Item-to-item Collaborative Filtering. *IEEE Internet Computing*, 2003.

[70] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the Mayor of My House: Examining Why People Use Foursquare - a Social-Driven Location Sharing Application. In *Proceedings of the* 29^{th} *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2011.

[71] D. J. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[72] S. M. McNee, J. Riedl, and J. A. Konstan. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *Proceedings of the* 24^{th} *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) - Extended Abstracts on Human Factors in Computing Systems*. ACM, 2006.

[73] R. Merton. Patterns of Influence: Local and Cosmopolitan Influentials. *Social Theory and Social Structure*, 1957.

[74] B. N. Miller, I. Albert, S. K. Lam, J. A. Konstan, and J. Riedl. MovieLens Unplugged: Experiences with An Occasionally Connected Recommender System. In *Proceedings of the* 8^{th} *International Conference on Intelligent User Interfaces (IUI)*. ACM, 2003.

[75] M. Muller, N. S. Shami, D. R. Millen, and J. Feinberg. We are all Lurkers: Consuming Behaviors among Authors and Readers in an Enterprise File-Sharing Service. In *Proceedings of the* 16^{th} *ACM International Conference on Supporting Group Work (GROUP)*, 2010.

[76] T. Murakami, K. Mori, and R. Orihara. Metrics for Evaluating the Serendipity of Recommendation Lists. *New Frontiers in Artificial Intelligence*, 2008.

[77] M. Naaman, H. Becker, and L. Gravano. Hip and Trendy: Characterizing Emerging Trends on Twitter. *Journal of the American Society for Information Science and Technology*, 65, May 2011.

[78] M. Nagarajan, K. Gomadam, A. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences. *Web Information Systems Engineering-WISE*, 2009.

[79] C. Neustaedter, A. Tang, and J. K. Tejinder. The Role of Community and Groupware in Geocache Creation and Maintenance. In *Proceedings of the* 28^{th} *International Conference on Human Factors in Computing Systems (CHI)*, 2010.

[80] Nielsen. State of the Media: the Social Media Report 2012. http://blog.nielsen.com/nielsenwire/social/2012/, 2012.

[81] S. Nikolov. Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series. Master's thesis, Massachusetts Institute of Technology, 2012.

[82] D. W. Oard, J. Kim, et al. Implicit Feedback for Recommender Systems. In *Proceedings of the AAAI Workshop on Recommender Systems*, 1998.

[83] M. P. O'Mahony, N. J. Hurley, and G. Silvestre. Detecting Noise in Recommender System Databases. In *Proceedings of the* 11^{th} *International Conference on Intelligent User Interfaces (IUI)*. ACM, 2006.

[84] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis. Geographic Constraints on Social Network Groups. *PloS One*, 6, 2011.

[85] M. Panik. *Advanced Statistics From an Elementary Point of View*. Academic Press, 2005.

[86] E. Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin Press HC, 2011.

[87] D. Parra and X. Amatriain. Walk the Talk: Analyzing the Relation between Implicit and Explicit Feedback for Preference Elicitation. *User Modeling, Adaption and Personalization*, pages 255–268, 2011.

[88] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles. Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- And Model-Based Approach. In *Proceedings of the* 16^{th} *ACM Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

[89] S. Petrović, M. Osborne, and V. Lavrenko. Streaming First Story Detection with Application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics, 2010.

[90] O. Phelan, K. McCarthy, and B. Smyth. Using Twitter to Recommend Real-Time Topical News. In *Proceedings of the* 3^{rd} *ACM conference on Recommender systems (RecSys)*. ACM, 2009.

[91] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge university press, 2007.

[92] D. Quercia and L. Capra. FriendSensing: Recommending Friends using Mobile Phones. In *Proceedings of the* 3^{th} *ACM Conference on Recommender Systems (RecSys)*. ACM, 2009.

[93] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. In the Mood for Being Influential on Twitter. In *Proceedings of the* 3^{rd} *IEEE Conference on Social Computing (SocialCom)*, 2011.

[94] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *Proceedings of the* 3^{rd} *IEEE Conference on Social Computing (SocialCom)*, 2011.

[95] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending Social Events from Mobile Phone Location Data. In *Proceedings of the* 10^{th} *IEEE International Conference on Data Mining (ICDM)*, 2010.

[96] A. M. Rashid, I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan, and J. Riedl. Getting to Know You: Learning New User Preferences in Recommender Systems. In *Proceedings of the* 7^{th} *International Conference on Intelligent User Interfaces (IUI)*. ACM, 2002.

[97] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. Factorizing Personalized Markov Chains for Next-Basket Recommendation. In *Proceedings of the* 19^{th} *International Conference on World Wide Web (WWW)*. ACM, 2010.

[98] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW)*. ACM, 1994.

[99] P. Resnick and H. R. Varian. Recommender Systems. *Communications of the ACM*, 40(3), 1997.

[100] F. Ricci, L. Rokach, and B. Shapira. Introduction to Recommender Systems Handbook. *Recommender Systems Handbook*, pages 1–35, 2011.

[101] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto. Finding Trendsetters in Information Networks. In *Proceedings of the* 18^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2012.

[102] D. Saez-Trumper, D. Quercia, and J. Crowcroft. Ads and the City: Considering Geographic Distance Goes a Long Way. In *Proceedings of the* 6^{th} *ACM Conference on Recommender Systems (RecSys)*, 2012.

[103] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the* 19^{th} *International Conference on World Wide Web (WWW)*, pages 851–860. ACM, 2010.

[104] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-Based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the* 10^{th} *International ACM Conference on World Wide Web (WWW)*, 2001.

[105] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the* 2^{nd} *ACM Conference on Electronic Commerce*. ACM, 2000.

[106] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial Properties of Online Location-based Social Networks. In *Proceedings of the* 5^{th} *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[107] G. Shani and A. Gunawardana. Evaluating Recommender Systems. *Recommender Systems Handbook*, pages 257–298, 2009.

[108] D. Sornette and A. Helmstetter. Endogenous versus Exogenous Shocks in Systems with Memory. *Physica A: Statistical Mechanics and its Applications*, 318(3), 2003.

[109] C. Steinfield, N. B. Ellison, and C. Lampe. Social Capital, Self-esteem, and Use of Online Social Network Sites: A Longitudinal Analysis. *Journal of Applied Developmental Psychology*, 29(6), 2008.

[110] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009.

[111] J. Surowiecki. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business. *Economies, Societies and Nations*, 2004.

[112] Y. Takhteyev, A. Gruzd, and B. Wellman. Geography of Twitter Networks. *Social Networks*, 2011.

[113] J. Tang, J. Sun, C. Wang, and Z. Yang. Social Influence Analysis in Large-scale Networks. In *Proceedings of the* 15^{th} *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

[114] J. Teevan, S. T. Dumais, and E. Horvitz. Potential for Personalization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2010.

[115] L. Terveen and W. Hill. Beyond Recommender Systems: Helping People Help Each Other. *HCI in the New Millennium. Addison Wesley*, pages 487–509, 2001.

[116] J. Travers and S. Milgram. An Experimental Study of the Small World Problem. *Sociometry*, 1969.

[117] S. Vargas and P. Castells. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the* 5^{th} *ACM conference on Recommender systems (RecSys)*, 2011.

[118] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling Bursts and Heavy Tails in Human Dynamics. *Physical Review E*, 73, 2006.

[119] X. Wang and A. McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the* 12^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2006.

[120] X. Wang, C. Zhai, X. Hu, and R. Sproat. Mining Correlated Bursty Topic Patterns from Coordinated Text Streams. In *Proceedings of the* 13^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2007.

[121] D. Watts. Challenging the Influentials Hypothesis. *Measuring Word of Mouth*, 3, 2007.

[122] D. Watts and P. Dodds. Influentials, Networks, and Public Opinion Formation. *Journal of consumer research*, 34(4), 2007.

[123] D. Watts and S. Strogatz. Collective Dynamics of 'Small-world' Networks. *Nature*, 393(6684), 1998.

[124] D. J. Watts. *Everything Is Obvious: *Once You Know the Answer*. Crown Business, March 2011.

[125] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the* 3^{rd} *ACM International Conference on Web Search and Data Mining (WSDM)*, 2010.

[126] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition* . Morgan Kaufmann Publishers Inc., 2005.

[127] F. Wu and B. A. Huberman. How Public Opinion Forms. In *Proceedings of the* 4^{th} *International Workshop on Internet and Network Economics (WINE)*. Springer-Verlag, 2008.

[128] M. Wu. Collaborative Filtering via Ensembles of Matrix Factorizations. In *Proceedings of KDD Cup and Workshop*, 2007.

[129] S. Yardi and D. Boyd. Tweeting from the Town Square: Measuring Geographic Local Networks. In *Proceedings of the* 4^{th} *AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[130] L. Yu, S. Asur, and B. A. Huberman. What Trends in Chinese Social Media. *The* 5^{th} *Workshop on Social Network Mining and Analysis (SNA-KDD)*, 2011.

[131] M. Zhang and N. Hurley. Avoiding Monotony: Improving the Diversity of Recommendation Lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys)*. ACM, 2008.

[132] Y. Zhang, J. Callan, and T. Minka. Novelty and Redundancy Detection in Adaptive Filtering. In *Proceedings of the* 25^{th} *ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 81–88. ACM, 2002.

[133] Y. Zhang, D. Séaghdha, D. Quercia, and T. Jambor. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the* 5^{th} *ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2012.

[134] V. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative Location and Activity Recommendations with GPS History Data. In *Proceedings of the* 19^{th} *ACM International Conference on World Wide Web (WWW)*, 2010.

[135] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale Parallel Collaborative Filtering for the Netflix Prize. *Algorithmic Aspects in Information and Management*, 2008.

[136] T. Zhuo, Z. Kuscik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the Apparent Diversity-Accuracy Dilemma of Recommender Systems. *PNAS*, 2010.

[137] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the* 14^{th} *International Conference on World Wide Web (WWW)*. ACM, 2005.

ANNEXE A

Synthèse en Français

A.1 Introduction

C'est évident que la majorité sont en train de construire d'une grande société on ligne.

Les réseaux sociaux sont de plus en plus d'être le terrain de jeu où *connexion* et le seul objectif. Il est plus un phénomène que nous sommes enthousiastes. En tant que phénomène, nous sommes prêts à partager toutes sortes d'informations (par exemple, les photos, les soirées, l'alimentation, les commères, la politique, etc) le long de nos connexions. Grâce aux réseaux sociaux, nous organisons aussi des événements sociaux, par exemple, les réunions, les fêtes, les activités du club et même les campagnes politiques. En bref, nous continuons de migrer nos activités offline à les réseaux sociaux.

Ce phénomène "social" avec sa riche collection de notre comportement en ligne a certainement attiré beaucoup d'intérêts de recherche. Beaucoup de questions ont été posées, mais des solutions ont été recherchées notamment par la compréhension des activités de l'utilisateur agrégées et diffusion de l'information au long de la connexion dans les réseaux sociaux. Les gens ont remarqué que lorsque l'information se propage le long des connexions, une partie est plus largement adopté que d'autres, et certains sont réparties plus rapide. En outre, il ya un moment que ces informations atteindre la masse critique tout d'un coup. Ce moment est la soi-disant "tipping point", défini par Gladwell dans [39].

Dans cette thèse, nous présentons nos études sur le contenu numérique qui déclenche le "tipping point" dans les réseaux sociaux. Nous appelons un tel contenu comme contenu de tendance - *tendances* - dans les chapitres suivants. Plus précisément, nous étudions les facteurs humains dans la création de tendances, et, nous créons un système pour aider des gens à découvrir des tendances qu'ils pourraient être intéressés.

A.1.1 Motivation

Il est essentiel d'identifier les tendances dans les réseaux sociaux, parce que la connaissance des tendances peut être une signal de *l'identification d'événement*, *la propagation des opinions*, *la gestion de la marque*, etc. Ces concepts sont semblables les uns aux autres par le fait que ils sont bien adoptées par beaucoup de gens et ils se propagent très rapidement. Pour faire la clarification du terme "tendance" dans les scénarios diverses, nous les élaborons avec quelques exemples.

- **Identification de l'événement.** *Une tendance est un événement.* Un exemple est fourni par Twitter.[1] Les gens publient sur Twitter au sujet de ce qu'ils voient ou ce qu'ils rencontrent. Il peut être de grands événements mondiaux tels que les Jeux olympiques ou les petits locaux tels que les réunions de quartier. En 2009, un rédacteur en chef à Mashable - Adam Ostrow, a observé que "*des tremblements de terre sont une chose que vous pouvez parier sur étant couverte sur Twitter abord, parce que, franchement, si le sol tremble, vous allez* tweet *ce sujet avant qu'il enregistre même avec l'USGS et avant qu'il ne soit rapporté par les médias* [2]". En effet, ce qui s'est passé pour le tremblement de terre au Japon 2009 : tweets sur le tremblement de terre ont voyagé autour du monde beaucoup plus rapide que les rapports officiels des médias. Cela a inspiré et motivé de nombreux chercheurs à identifier les événements du monde réel en gardant la trace de la diffusion de l'information dans les réseaux sociaux [13, 14, 63, 103].

- **Propagation de l'opinion.** *Une tendance est une pièce d'opinion.* Par exemples, des commentaires à de nouvelles, des avis aux points d'achat, et des discussions politiques ou de la société sont tous les types différents d'opinions qui se propagent dans l'Internet. Borge-Holthoefer *et al.* [18] ont étudié un cas particulier - la discussion du mouvement 15-M invoqué par la crise économique en Espagne en 2011. En recueillant des messages Twitter (tweets) pour un mois (environ de deux semaines avant le mouvement jusqu'à une semaine après il a eu lieu), ils ont remarqué que la *mouvement-dans-le-décision couvait depuis un certain moment dans les médias sociaux* [18]. Hashtags liés aux discussions du camping à la place Puerta del Sol à Madrid ont été mentionnés par des tonnes de tweets, et le sous-jacent structure de "following" et "followers" dans les réseaux sociaux ont poussé des récepteurs de cette décision beaucoup plus loin. Comment les opinions sont répartis est sans doute une grande question très important à comprendre. En plus de son impact sur les opinions sociales et politiques, Wu *et al.* [127] ont étudié également la façon dont les opinions publiques se forment dans les systèmes de vote et d'examen en ligne.

- **Gestion de la marque.** *Une tendance est un phénomène de mode.* Jansen *et al.* a proposé d'utiliser *micro-blogging comme word-of-mouth branding en ligne* [54]. Ils ont déclaré que dans des situations commerciales, une marque avec l'effet positive de word-

[1] http://www.twitter.com
[2] http://mashable.com/2009/08/12/japan-earthquake/

of-mouth a une forte influence sur les consommateurs, car elle est fondée sur la *confiance* construit sur des relations sociales. Avec l'analyse des sentiments, leurs études de tweets collectés pendant 13 semaines montrent que les satisfactions des utilisateurs avec les marques changent avec le temps : ces changements sont corrélés avec la propagation d'opinion par word-of-mouth. Motivé par effet comme le word-of-mouth, les chercheurs ont également proposé de tirer des blogs sociaux dans le développement de nouveaux produits [36], ainsi que la prédiction de la réaction des clients [17].

Après avoir discuté des types différents de "tendances", qui reçoivent tous une salve d'attention à un certain moment et obliger les gens à réagir rapidement, une question on pose naturellement est la suivante :

Comment Détecter Les Tendances ?
Pour réagir rapidement, nous devons être conscients des tendances suffisamment tôt. Toutefois, il est très difficile de capturer une tendance avant son "tipping point". Il y a des grands efforts de recherche qui ont été consacrés à *caractériser les tendances*. La plupart des solutions élaborées sont construites sur le fait que les tendances sont les résultats de comportement *agrégé* de les utilisateurs. En d'autres termes, le signal principal d'une tendance est les réponses intensifs d'individus. Par conséquent, de nombreuses études ont été menées pour étudier les tendances depuis des réactions diverses des utilisateurs, à savoir :

- **Des clics.** Un clic, sur le Web, est une réaction de base qui indique implicitement nos intérêts. Apprendre des clics agrégées dans le site YouTube [3] où on partage de vidéos, Crane *et al.* [28] ont identifié des modèles différents de clics agrégés associés à diverses types de tendances. S'appuyant sur ces modèles de clics, on peut dire tendances exogènes (celles déclenchées par des facteurs externes au site, par exemple, la notification d'un morceau de nouvelles à la télévision) de ceux endogènes (celles déclenchées par des facteurs internes, par exemple, la partage d'un morceau de nouvelles entre les utilisateurs de le site).

- **Des Posts/Retweets.** Ils sont plus explicites et proactifs des activités de posts ou retweets activités que des clics. Ils montrent explicitement que l'un est prêt à répandre le contenu. En collectant des contenues tendance en Twitter et des tweets associés, Nikolov [81] a proposé une méthode statistique de classification non paramétrique pour capturer des sujets tendances par l'apprentissage de leur modèle d'allures de "tweeting".

- **Des Contenues.** Des contenue de tendances sont bien analysés avec certaine granularité. Un exemple de tendances exogènes comme la diffusion de nouvelles sur Twitter, c'est que souvent, ces tweets contient une URL lier à un site externe où les nouvelles a été crées [77].

[3] http://www.youtube.com

- **Des Connexions Sociaux.** Des fonctionnalités de "Following" et "followers" sont fondamental dans réseaux sociaux. Grâce à ces liens sociaux, l'information s'écoule de l'un à l'autre. Intuitivement, les gens qui établissent de nombreux liens sociaux ont de meilleures chances de se propager contenu à d'autres, c'est bien le cas que comment tendances endogènes (comme les commères de célébrités) sont générés [77, 130].

La connaissance des caractéristiques des tendances est d'une grande aide pour *détecter* des tendences. Mais, en dévoilant comment les tendances évoluent avec la dynamique humaine peut fournir aux gens (par exemple, marketing social) la connaissance de qui *créer* les tendances. À cette fin, la recherche d'"influentes" au sein du réseau sociaux devient le thème central.

Qui Sont Influentes ?
La théorie fondamentale d'"influentes" remonte à la paradigme de *two-step flow*, proposé par Katz et Lazarsfeld en 1955 [64], qui a été initialement conçu pour comprendre comment les opinions publiques se forment. Il dit que la cascade de la diffusion de l'information est *"un processus de déplacement de l'information par les médias aux des leaders d'opinion, et l'influence se déplace des leaders d'opinion à leurs followers"* [20].

Pendant des décennies, la théorie de l'écoulement en deux étapes a été dominante dans la recherche de diffusion l'information. Leur définition de *leaders d'opinion* a été bien acceptée, et plus tard adopté comme la définition d'*influentes* [73]. C'est, des influentes sont *les personnes qui sont susceptibles d'influencer d'autres personnes dans leur environnement immédiat* [64].

Les études modernes au sujet d'influentes (notamment avec l'accès facile à des traces de diffusion de l'information dans les réseaux sociaux) ont mis au point deux opinions différentes.

- **Des influentes sont des personnes spéciales.** Adhérant à la théorie de l'écoulement en deux étapes, les chercheurs de ce groupe croient que les influentes sont différents de la majorité dans une certaine mesure. Dans son livre *The Tipping Point*, Gladwell indique que *"le succès de n'importe quel type de épidémie sociale est fortement tributaire de la participation des personnes avec des talents sociaux particuliers et rares"* [39]. Il a identifié trois acteurs comme des individus particuliers qui ont créé les épidémies sociales (tendances). Ils sont des *connecteurs* (ceux qui savent beaucoup de gens dans la communauté), des *mavens* (des spécialités de l'information) et des *vendeurs* (qui savent comment persuader des autres). En modélisant et analyse de diffusion du l'information dans les réseaux sociaux, les chercheurs ont confirmé les existences de types différents de personnes spéciales qui peuvent repérer les tendances plus tôt possible [64, 101]. Ces personnes spéciales sont bien connectés avec des autres (connecteurs) [57] ; savent influencer les autres facilement (vendeurs) [45] ; sont considérés comme des experts (mavens) [113, 125] ; ou sont les célébrités [130].

- **Des influentes peuvent être n'importe qui.** Duncan Watts affirme que d'être influente est *surtout un accident de localisation et de synchronisation* [10, 121]. Il s'agit de l'adoption de bonnes opinions à un bon moment, peu importe qui vous êtes. En outre, il a déclaré que l'*influentes* ne sont pas nécessairement *"les chefs des organisations formelles, ni des personnalités publiques telles que les columnists des journaux, des critiques, ou des personnalités des médias, dont l'influence est exercée indirectement par des médias organisés ou des structures d'autorité"* [122]. Pour mettre en évidence les concepts de *inattendu* et *imprévu*, les individus qui sont impliqués dans la diffusion des tendances sont ensuite appelé comme "influentes accidentelles" [121].

Nous avons vu que l'identification des tendances peut être traduit à l'identification d'événements, la propagation d'opinion, la gestion de marque, etc. Une variété d'études de tendances sont effectuées avec deux thèmes. Ce sont, 1) quelles sont les tendances ; 2) qui crée les tendances. Nous avons donné un bref aperçu sur la façon dont les gens essaient de détecter les tendances par leurs caractéristiques, ainsi que le débat de savoir si les personnes qui créent les tendances sont spéciaux (des informations plus détaillées et les related works s'il vous plaît se référer au chapitre 2). Ensuite, nous définissons notre champ de cette thèse de recherche et positionnons nos contributions à l'égard de la littérature sur l'exploration des tendances et les outils associés à notre disposition.

A.1.2 Objectifs de la these et contribuations

Nous avons spécifié que les tendances sont les résultats de comportement agrégées des utilisateurs. Ils sont des informations qui sont diffusées dans le réseau et ont obtenu une large couverture des adoptants. Sans surprise, la notion de *diffusion* au sein du réseau est la concentration. Cependant, le processus complet de la naissance d'une tendance devrait également inclure la création de l'information elle-même. Il est indéniable que les personnes qui créent l'information (qui ont d'abord le mis en réseau) sont un filtre important de l'information de l'extérieur du réseau.

Considérant à la fois la création et la diffusion des tendances, il y a des questions qui ont encore besoin des réponses claires. Ces questions sont posées à partir de trois aspects principaux - les facteurs humains derrière les tendances, identification des tendances et leur explorations.

- **Facteurs humains.** Malgré du débat au sujet de les influentes, la dynamique humaine d'individus qui créent les tendances (peu importe si elles sont spéciales) sont encore mal connues. Considérant les deux individus qui apportent à l'origine des informations sur le réseau et ceux qui les propagent, quelles sont leurs caractéristiques ? Partagent-ils des traits communs et quelles sont leurs différences ?

- **Identification.** En sachant des caractéristiques des personnes qui créent les tendances, est-il possible d'identifier les tendances en s'appuyant sur leurs connais-

sances ? Dans quelle mesure les tendances ont pu être identifiées précisément en tant que tel ?

- **Exploration.** Supposons que nous sommes capable d'identifier précisé les tendances. Comment pouvons-nous construire sur cette capacité à aider les utilisateurs à découvrir les tendances de leurs intérêts ? Pour fournir une telle exploration de contenu personnalisé, comment on peut garantir la qualité des personnalisations ?

Dans cette thèse, nous allons aborder ces questions dans les étapes. A la recherche des réponses, nous faisons les contributions suivantes :

- Nous abordons l'analyse de qui crée des tendances en définissant deux classes distinctes de personnes : trend spotters (ceux qui évaluent les articles avant qu'ils ne deviennent des tendances) et trend makers (ceux qui chargent des objets qui deviennent tendances). Nous les caractérisons par la combinaison de plusieurs caractéristiques, notamment leur activité, le contenu, le réseau et les caractéristiques géographiques. Nous constatons que trend spotters et makers sont diffèrents des utilisateurs typiques, en ce que, ils sont plus actifs, sont intérêts très variétés, et attirent des liens sociaux. Nous étudions ensuite ce qui différencie les trend spotters des makers. Nous apprenons que les trend spotter réussis sont des adopteurs précoces qui aiment très divers articles, tandis que les trend makers succès sont des personnes de tout âge qui se concentrent sur des types spécifiques (Chapter 3).

- En utilisant la régression linéaire, nous prévoyons la mesure de trend spotters et makers. Puis, avec un algorithme machine learning (SVM) et une régression logistique, nous procédons à une méthode classification binaire de savoir si l'on est susceptible d'être un trend spotter (maker). Bien que la régression linéaire a donné des résultats intéressants, SVM et régression logistique ont retourné des prédictions précises (Chapter 3).

- Nous proposons une méthode qui détecte les tendances en s'appuyant sur les activités de deux types d'utilisateurs : les trend makers et spotters. On construit alors une matrice de préférence fondée sur les tendances identifiées, et de tester la moteur de recommandation avec un algorithme de matrice factorisation (*Implicit SVD* [51]) (Chapter 4).

- Au-delà de l'objectif de faire des recommandations précises, nous explorons la possibilité d'enrichir la sérendipité dans les recommandations finales en s'appuyant sur des techniques d'analyse de réseau, et de valider nos propositions dans le cadre d'un système de recommandation mobiles. Nous nous attaquons à la possibilité d'introduire la sérendipité par la promotion de lieux qui vont au-delà ceux qui seraient recommandées en base de lieux déjà visité dans le passé ou sur sa routine quotidien. Nous évaluons quantitativement dans quelle mesure nous pourrons pousser la sérendipité, sans compromettre la précision des recommandations sur un dataset real (Chapter 5).

A.1.3 Structure de la these

Chapter 1 a énoncé nos problèmes de recherche.

Chapter 2 donne le background de notre recherche depuis deux directions principaux, i.e., les tendances dans les médias sociaux et les systèmes de recommandation.

Chapter 3 différencie les trend spotters et makers des utilisateurs typiques, étudie leur caractéristiques et montre expérimentalement qu'ils peuvent être prédits précisément avec une variété de caractéristiques.

Chapter 4 propose un système de recommandation pour satisfaire les gens avec des contenus tendances personnalisés.

Chapter 5 tire des techniques d'analyse de réseau pour introduire la sérendipité dans les recommandations.

Chapter 6 conclut notre recherche et résume nos contributions à la state-of-the-art.

A.2 Qui Crée Les Tendances

Des commerciaux du média et des chercheurs ont de grands intérêts à ce qui devient une tendance au sein de sites de médias sociaux. Leurs intérêts ont porté des analyses des contenues qui deviennent tendances. Dans ce travail, nous allons nous concentrer sur les personnes plutôt que les contenues. Les recherches sur les personnes qui créent les tendances dans les réseaux sociaux ont été concentrés sur leur pouvoir d'influencer les autres à adopter une idée ou un produit. Nous allons au-delà de la capacité d'influence, et d'affiner les rôles de ces personnes avec deux catégories d'utilisateurs - trend makers (ceux qui génèrent les tendances) et trend spotters (ceux qui les propagent). Et nous menons nos analyses sur un database réelles collecté auprès d'un média social mobile. Cet database contient les activités des utilisateurs depuis Février 2010 (son lancement) à Août 2010 (avant le moment où cet application lancé son application web).

A.2.1 Identification de Trend Makers et Spotters

Pour indentifier les trend spotters et makers, d'abord, nous définissons une métrique "trend score" pour identifier les **tendances** dans les dataset. A chaque unité de temps t (une fenêtre d'une semaine qui glisse progressivement tous les jours), nous attribuons à l'objet i un $trendScore(i,t)$ qui augmente avec le nombre de votes qu'il reçoit :

$$trendScore(i,t) = \frac{|v_{i,t}| - \mu_i}{\sigma_i} \tag{A.1}$$

où $|v_{i,t}|$ est le nombre de votes l'objet i a reçu dans les délais unité t, le μ_i est le nombre moyen de votes qu'il a reçu par unité de temps. Pour chaque unité de temps (chaque semaine), nous trions les objets par leurs "trend scores" dans l'ordre décroissant et sélectionnons les top-n objets comme les tendances.

Trend Spotters. Trend spotters sont ceux qui vont voter des objets qui, après un certain moment, deviennent les tendances. Compte tenu d'une groupe d'objets de tendance, la capacité d'une à voter des objets tendances dépend de trois facteurs : *combien* de tendances, *comment les premiers* on les a voté, et *la popularité* des contenues votés se sont avérés être. Pour chaque utilisateur u, nous intégrons ces trois facteurs dans un *spotterScore* en divisant le nombre de tendances de u a voté ($\sum_{i \in \mathcal{I}_u} g_{u,i}$) par le nombre total de son votes (v_u) :

$$spotterScore\,(u) = \frac{\sum_{i \in \mathcal{I}_u} g_{u,i}}{v_u} \tag{A.2}$$

Dans le numérateur, $g_{u,i}$ est le gain que l'utilisateur u acquiert lors du son vote sur le tendance i et il intègre les trois facteurs de *combien*, *comment les premiers* et *la popularité*($g_{u,i} = v_i \times \alpha^{-p_{u,i}}$) dont $\mathcal{I}_u$ est un groupe de tendances que u a voté ($\sum_{i \in \mathcal{I}_u}$ reflet le *combien*) ; v_i est le total de votes que l'objet i a reçu (qui reflet *la popularité*), et puis, α est une facteur de décroissance (which reflet le *comment les premiers*, $\alpha = 2$ dans nos expérimentes) dont exposant est dans l'ordre chronique où u a voté l'objet i (i.e., $p_{u,i}$ signifie que u est le p^{th} qui a voté le i). Un trend spotter est donc n'importe qui avec spotter score plus que zéro.

Trend Makers. Trend makers sont ceux qui (pas seulement voter, mais) *uploader* les tendances. Donc, le score trend maker d'utilisateur u augmente avec le nombre de tendances que u a uploadé. Le numerateur du score est $\sum_{i \in \mathcal{I}_u} I(i\ is\ a\ trend)$, dont $\mathcal{I}_u$ est le groupe d'objets que u a uploadé, et le I est la fonction indicateur, qui est 1, si l'expression "*i est une tendance*" est vrai ; 0, dans l'autre cas. A compté des utilisateurs qui uploadent indistinctement un grand nombre d'objets sans aucun contrôle de la qualité, ce numerateur et puis est normalise par le nombre total de upload de u ($|\mathcal{I}_u|$). Un trend maker donc est n'importe qui reçoit le maker score plus que zéro.

$$makerScore\,(u) = \frac{\sum_{i \in \mathcal{I}_u} I(i\ is\ a\ trend)}{|\mathcal{I}_u|} \tag{A.3}$$

Utilisateurs Typiques. Si un utilisateur actif (i.e, qui a uploadé ou voté plus d'une fois) n'est pas un trend spotter ou maker, alors il/elle est considérée comme un utilisateur typique.

A.2.2 Caractérisations des Trend Spotters et Trend Makers

Pour caractériser les trend spotters et makers, nous effectuons une analyse quantitative qui considère quatre types de traits : l'activité, le contenu, le réseau et les caractéristiques

géographiques. Et nous comparons les spottes et les makers avec les utilisateurs typiques par tester des hypothèses tirées de la littérature, que table A.1 rassemble pour plus de commodité. Pour tester ces hypothèses, nous appliquons Kolmogorov-Smirnov teste (K-S tests [85]), et nous rapportons les résultats dans le Tableau A.2.

	Content	Result
Spotters vs. Typiques	H1.1 Les trend spotters sont plus actifs que les utilisateurs typiques.	$\surd$
	H1.2 Les trend spotters sont plus spécialisés que les utilisateurs typiques dans certain catégories d'objets.	$\times$
	H1.3 Les trend spotters attirent plus followers que les utilisateurs typiques.	$\surd$
Makers vs. Typiques	H2.1 Les trend makers sont plus actifs que les utilisateurs typiques.	$\surd$
	H2.2 Les trend makers sont plus spécialisés que les typiques dans certain catégories d'objets.	$\times$
	H2.3 Les trend makers attirent plus followers que les typiques.	$\surd$
Spotters vs. Makers	H3.1 Les trend makers upload des contenues plus souvent que les spotters.	$\surd$
	H3.2 Les trend makers votent moins souvent que les spotters.	$\surd$
	H3.3 Les trend spotters upload des contenues plus diverse que les makers.	$*$
	H3.4 Les trend spotters votent des contenues moins diverse que les makers.	$\times$
	H3.5 Les trend makers ont plus followers que celui de spotters.	$\surd$

TABLE A.1: Nos Hypothèses ($\surd$: hypothèse accepté ; $\times$:hypothèse alternative accepté ; $*$: inconnu)

Trend spotters (makers) vs. Utilisateurs typiques. L'idée d'utiliser de K-S test est que nous considérons comme une paire de distributions, par exemple, celles de "daily uploads" de spotters (S) et des utilisateurs typiques(T), et nous les comparons - nous comparons si la moyenne de la distribution des spotters est plus grande que celle des typiques (i.e., nous testons $S > T$). Nous constatons que, par rapport aux utilisateurs typiques, les spotters et les makers sont plus actifs (qu'ils upload et votent plus) et sont plus populaires (attirent plus de followers). En revanche, les hypothèses H1.2 et H2.2 ne sont pas confirmées. Lors de la consommation et la production de contenu, les spotters et makers ni l'accent uniquement sur les catégories de contenu spécifiques ni se diversifient de plus de ce que les utilisateurs typiques font. Cependant, en séparant ce que les utilisateurs votent et ce qu'ils téléchargent, nous constatons que les onjets votés par les spotters sont plus diversifiées que celles uploadés.

Traits (log-transformé)	**S > T**	**M > T**	**M > S** (si pas démontré le contraire)
Daily Uploads	**0.07** $*$	**0.45** $*$	**0.58** $*$
Daily Votes	**0.66** $*$	**0.18** $*$	**0.57** $*$ (M < S)
Upload Diversity	**0.31** $*$	**0.35** $*$	0.02 (M < S)
Vote Diversity	**0.31** $*$	**0.23** $*$	**0.27** $*$ (M < S)
#Followers	**0.06** $*$	**0.32** $*$	**0.26** $*$

TABLE A.2: Résumé de les résultats de Kolmogorov-Smirnov test. Les values D avec ses niveaux significatifs < 0.05 sont mis en évidence et sont livrés avec $*$. M, S et T représent les trend makers, spotters et utilisateurs typiques. Nous testons un pair de distributions a la foi - e.g., pour **S > T**, nous testons si la distribution de *daily upload* de spotters est plus grande que cela des utilisateurs typiques, et nous rapportons le valeur D correspondant.

Trend spotters vs. Trend makers. Selon des tests Kolmogorov-Smirnov (Tableau A.2), nous observons que les makers *upload* plus souvent que les spotters qui, en contre, *votent* plus souvent. En considérant ce que des gens upload/votent, on trouve que les makers "restent concentrer" (i.e., ils uploadent et votent les objets dans les catégories spécifiques), mai les spotters votent des objets appartiennent à des catégories variées. Donc, les makers se soucient particulièrement de produire un contenu de haute qualité. De même, les spotters uploadent des objets dans les catégories limites que ils sont familiers avec, mais ils votent des objets plus variées, en suggérant un large éventail d'intérêts. Finalement, les makers sont plus populaire (ont plus followers) que les spotters.

A.2.3 Prédiction de Trend Makers et Spotters

Ayant compris les caractéristiques des trend makers et spotters, nous étudions maintenant dans quelle mesure les traits des utilisateurs sont des facteurs prédictifs potentiels de savoir si quelqu'un est un trend spotter (maker), et nous le faisons en deux étapes : 1) nous modélisons les scores de trend spotter (maker) comme une combinaison linéaire des traits ; et 2) nous prévoyons le score de trend spotter (maker) avec une régression logistique et un modèle machine learning : Support Vector Machines (SVM). (Nous gérons nos prévisions sur le groupe des 140 makers, 671 spotters et 1705 utilisateurs typiques identifiés dans le dataset expérimental).

Modèles de Régression. Nous effectuons deux méthodes de régression - logistique et linéaire, avec des inputs de prédicateurs (des traits) ne pas être fortement corrélée avec l'autre, et nous modélisons le score de spotter (maker) dans deux étapes comme il est fait dans le littérature [38]. Première, nous modélisons si un utilisateur a le score de spotter (maker) supérieur à zéro par une régression logistique. Et puis, nous prenons seuls les utilisateurs avec les scores spotter (maker) supérieurs à zéro, et de prédirons ses scores avec une régression linéaire.

Traits	$\mathbf{I(Score > 0)}$	
	Spotters	**Makers**
Age	2e-04	0.001
Life Time	**0.006** *	**0.001** *
Daily Votes (Daily Uploads)	**0.007** *	**0.16** *
Vote Diversity (Upload Diversity)	**0.38** *	**0.14** *
Wandering	-6e-15	-7e-15
#Followers	2e-05	**0.009** *
Network Clustering	0.08	**0.28** *

TABLE A.3: Coefficients de régression logistique. Un coefficient de corrélation dans les 2 erreurs standard est considéré comme statistiquement significatif. Nous les soulignons et marquons avec *.

Les résultats de régression logistique (des coefficients dans le Tableau A.3) montrent que les prédicateurs signifiants pour spotters sont *life time*, *daily votes* and *vote diversity*. Pour les makers, ce sont *life time*, *daily votes*, *vote diversity*,*number of followers* et *network clustering*. Ces predicateurs signifiants statistiquement suggèrent que les spotters sont des

"early adopteurs" qui votent souvent et sont intéressés à des objets diverses. Et les makers sont des "early adopteurs" qui téléchargent souvent et aussi téléchargent des objets très diverse, de plus, ils bien attirent des followers et ont beaucoup connexions sociaux.

Maintenant, nous considérons uniquement des gens qui ont les score spotter (maker) supérieur à zéro. Les résultats de régression linéaire (coefficients β dans le Tableau A.4) montent que les prédicateurs signifiants pour les spotters qui réussissent sont *age*, *life time* et *vote diversity*, mais les prédicateurs signifiants pour les makers qui réussissent sont *daily uploads*, *upload diversity*, *number of followers* et *network clustering*. Le signe de coefficient de prédicateur suggère que les spotters qui réussissent sont les early-adpteurs adultes qui votent des objets dans les catégories variées. En contre, les makers qui réussissent sont les utilisateur de n'importe quel âge qui upload des objets dans les catégories spécifiques (ils "restent concentrer") et ils attirent beaucoup followers de communautés différentes.

Traits	$\log(\mathbf{Score})$	
	Spotters	**Makers**
Age	**0.36** *	0.01
Life Time	**0.19** *	0.0001
Daily Votes (Daily Uploads)	0.16	**-1.03** *
Vote Diversity (Upload Diversity)	**7.28** *	**-1.09** *
Wandering	-2.1e-13	-1.4e-15
#Followers	-0.06	**0.01** *
Network Clustering	2.75	**-0.64** *
R^2	0.15	0.65
Adjusted R^2	0.14	0.64

TABLE A.4: Coefficients de régression linéaire. Un coefficient de corrélation dans les 2 erreurs standard est considéré comme statistiquement significatif. Nous les soulignons et marquons avec *.

Support Vector Machines (SVM). Nous formulons la tâche de prévoir les spotters (makers) comme un problème de classification binaire, où la variable de réponse est de savoir si le score de spotter (maker) est supérieur ou égal à zéro. Sur notre dataset de 671 spotters et de 140 makers, nous ajoutons un nombre égal d'utilisateurs typiques. Par construction, le sample est équilibré (la variable de réponse est 50-50), et l'interprétation des résultats devient maintenant facile, et la précision d'un modèle de prédiction aléatoire serait de 50 %. Nous séparons au hasard chaque série de sample en deux subsets , 80% d'entre eux sont utilisés a la formation et 20% pour les tests. Nous appliquons SVM sur des sept mêmes traits utilisées dans les modèles de régressions. On compare les performances de prédiction avec ceux obtenus par le modèle de régression logistique. Les résultats en montrent avec la forme ROC (Receiver Operating Characteristic) (Figure A.1) , AUC (aire sous la courbe ROC), et la précision (tableau A.5) disent que SVM et régression logistique ont des performances comparables. SVM surpasse légèrement la régression logistique a la identification de makers. Cela nous donne à penser que les spotters et makers sont peut être identifier efficacement même avec une régression logistique simple. Aussi, SVM peut-être n'a pas montré gain de prédiction considérable simplement en raison de taille limité de notre dataset.

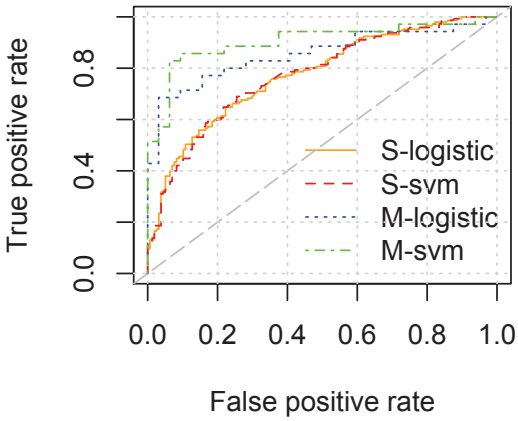


FIGURE A.1: La courbe ROC de régression logistique et le modèle de SVM (S : trend spotters ; M : trend makers).

	Spotters		Makers	
	AUC	**Accuracy**	**AUC**	**Accuracy**
Logistic	0.77	71.52%	0.85	82.09%
SVM	0.77	71.85%	0.90	88.06%

TABLE A.5: AUC et le meilleure précision de chacun modèle de prédiction.

A.2.4 Résumé

Nous avons étudié une communauté spécifique de personnes qui se passionnent pour le partage des photos d'objets (principalement des objets de mode et de design) à l'aide d'une application mobile. Dans cette communauté, nous avons vu et quantifié l'importance de early-adopteurs. Dans notre cas, les early-adopteurs peuvent être les trend spotters succès qui aiment des objets très divers. Les trend makers, en revanche, peuvent être des individus très organisés qui concentrent aux objets spécifiques. Comprendre les caractéristiques des "nombreux" - des personnes ordinaires ayant des intérêts spécifiques (trend makers) connectés à des early-adopteurs qui ont des intérêts très divers (trend spotters) - s'est avéré être plus important que d'essayer de trouver la "special few".

A.3 Personaliser Les Tendances

Dans cette partie de recherche, nous étudions le potentiel de fournir "les crowds" avec un nouvel façon d'explorer des tendances par l'utilisation de la sagesse de les individus "spéciaux". Nous proposons un nouveau système de recommandations pour personnaliser des tendances, qui comprend trois étapes. Basé sur nos résultats présenté dans la Section A.2, que nous avons trouve que les tendances ont crées dans un processus combine par deux types d'utilisateurs "spéciaux" - trend makers et spotters. Nous d'abord identifions ces deux types entre les "crowds". Et puis, fonde sur ceux que les "spéciaux" ont téléchargé et voté, nous identifions les tendances. Dans la troisième étape, nous personnalisons et recommandons des tendances avec un algorithme de state-of-the-art. Sur

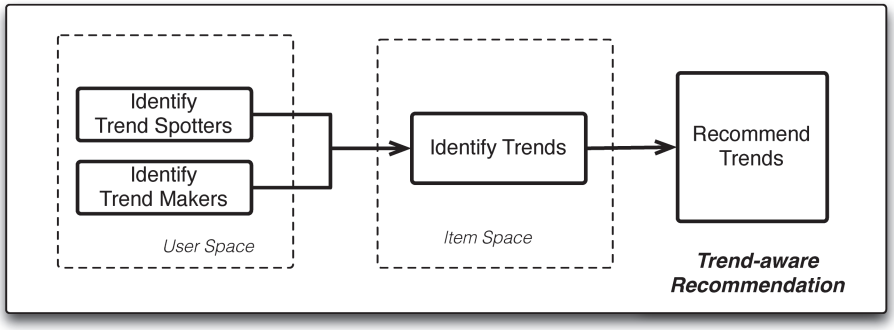


FIGURE A.2: Systeme de Recommendation Trend-aware

un dataset d'une application mobile, nous comparons le performance de notre moyen avec un système de recommandation traditionnel.

A.3.1 Trend-aware Recommendation

Pour recommander des tendances personnalisés, nous pourrions construire un système de recommandation trend-aware par la suite de la pratique courante. Autrement dit, le processus à mettre en place un système de recommandation consiste à extraire les préférences des utilisateurs implicites/explicites, et la mise en oeuvre d'un algorithme pour prédire les préférences des utilisateurs sur les objets ne sont pas encore notés/votés/consumés.

Pour en savoir des préférences de l'utilisateur sur les tendances, nous devons d'abord identifier les tendances. Pour obtenir cette connaissance préalable, nous construisons en fondant des conclusions à la Section A.2, ce qui suggère que les tendances dans les réseaux sociaux sont générés par un processus combiné dans lequel deux types d'utilisateurs différents sont engagés - des individus ordinaires qui ont des intérêts spécifiques (trend makers) et des early-adopteurs qui ont des intérêts très divers (trend spotters). Précisément, nous allons au-delà de la pratique courante et construisons notre système de recommandation trend-aware en effectuant les trois étapes suivantes(Figure A.2) : 1) identifier les trend makers et spotters ; 2) identifier les tendances par les sagesse de trend makers et spotters ; et 3) recommender les tendances identifies avec un algorithme de factorisation de matrice à la state-of-the-art (*Implicit SVD* [51]).

a. Identification de trend makers et spotters. Nous avons constaté que les trend makers et spotters peuvent être quantitativement identifiés en utilisant les caractéristiques des leur activités, leur téléchargement de contenu et les consommations, la structure de leur réseaux sociaux et de couverture géographique. Ici, nous allons au-delà de l'identification de savoir si l'on est un trend maker ou spotter, et tentons d'identifier makers et spotters des différents niveaux de succès. Pour cela, nous construisons un modèle statistique pour prédire à quel niveau de succès un utilisateur est un trend maker ou spotter, au moyen de trois étapes. Pour chaque utilisateur, nous : 1) calculons son score de spotter et de ma-

ker (defini dans la Section A.2) ; 2) discrétisons ses scores (log-scale) avec k intervalles (comme ça, les utilisateurs sont regroupés dans k classes) ; 3) prédirons les scores discréditées avec une technique de machine learning - Support Vector Machine (SVM) sur des traits de l'activité, le contenu, la structure de réseau social, et de la géographie.

b. Identification des Tendances. Après avoir construit un modèle de SVM pour prédire dans quelle mesure un utilisateur est un trend spotter ou un maker réussie. Nous explorons ensuite la possibilité d'identifier les tendances par ces utilisateurs spécifiques des classes différentes. Plus généralement, un objet est susceptible de devenir une tendance en fonction de : 1) la mesure dans laquelle son uploader est un trend maker, et 2) la mesure dans laquelle ses voteurs sont les trend spotters. Et nous en modélisons avec une régression logistique.

c. Recommender des Tendances. Pour recommander des tendances, nous avons besoin de deux composants majors : des préférences de tendance de chaque utilisateur ; et une algorithme a prédiquer le préférence d'un sur un tendance q'il ne savais pas.

Des préférences personnel de tendances. Avec des objets identifie comme des tendances potentiels, nous allons construire une matrice trend-aware de préférence personnel $\mathcal{P}'$, dont $p'_{u,t}$ est 1 or 0 dépendant si u aime le tendance t. Dans le contexte de notre dataset, nous inférons le rating implicite depuis des votes. Si un a voté un tendance, nous en considérons comme le signal de intérêt, et le enregistrons son préférence personnel sur cet objet comme 1.

Collaborative Filtering. Dans cette matrice trend-aware de préférences personnels, nous appliquons deux algorithmes populaire de système recommandation : *Implicit SVD* [51] et *item-based* collaborative filtering [69].

Nous comparonsla façon dont ces algorithmes effectuent en comparant la matrice trend-aware comme input avec la matrice de préférence traditionnel $\mathcal{P}$ comme input (où $p_{u,t}$ est 1 ou 0 selon, encore une fois, si u a voté sur l'objet t qui a alors devenu une tendance). La différence entre les deux matrices de préférence est que l'une des trend-aware est moins rare, car, dans les colonnes, il n'a pas tous les objets mais seulement ceux que nous avons prévu comme tendance.

A.3.2 Notre Expérimentes

Nous évaluons l'efficacité de chacune étape en utilisant le même dataset que dans la Section A.2. Par efficacité, nous nous référons à la précision des modèles statistiques de prédictions et la précision du système de recommandation trend-aware.

Classifier de trend spotter(maker) de classes déférents. Nous évaluons la mesure dans laquelle *SVM* est capable de classifier chaque utilisateur dans l'une des trois classes de trend makers/spotters, par ses traits humains. À cette fin, nous faisons une 10-fold cross-

validation, et le modèle *SVM* présentés arrive la précision de 83,80% pour des trend spotters et 60,7% pour des makers de classes différentes.

Déterminer si un objet est une tendance ou pas. Nous testons maintenant de savoir si un modèle de régression logistique est suffisant d'identifier les tendances par les informations de uploaders et voteurs. En particulier, les facteurs prédictifs que nous prenons sont : 1) la classe de trend maker de le uploader de cet objet, 2) le nombre de votes depuis des spotters de la classe *bas* ; 3) le nombre de votes depuis des spotters de la classe *moyen* ; et 4) le nombre de votes depuis des spotters de la classe *haute*. Pour éviter un surajustement (compte tenu de la taille limite de notre dataset expérimente), nous ajoutons un terme de régularisation - Tikhonov régularisation [91] a le modèle logistique. Et notre expérience montre que le modèle régression après régularisation est capable de classer précisé les tendances.

Recommander des tendances. Nous essayons maintenant de recommander des tendances personnalisés par construire d'un *m*atrice user-by-trend sur les tendances prévues, et nous en évaluons ses performances (en termes de précision et de recall) avec les algorithmes de collaborative filtering de state-of-the-art (l'algorithme simple de item-based [104], et puis, nous allons voir si nous pouvons améliorer les performances par *implicite SVD* [51]). Pour être en accord avec la littérature dans l'évaluation de la performance, nous calculons la précision et le recall [51] définis comme ensuit :

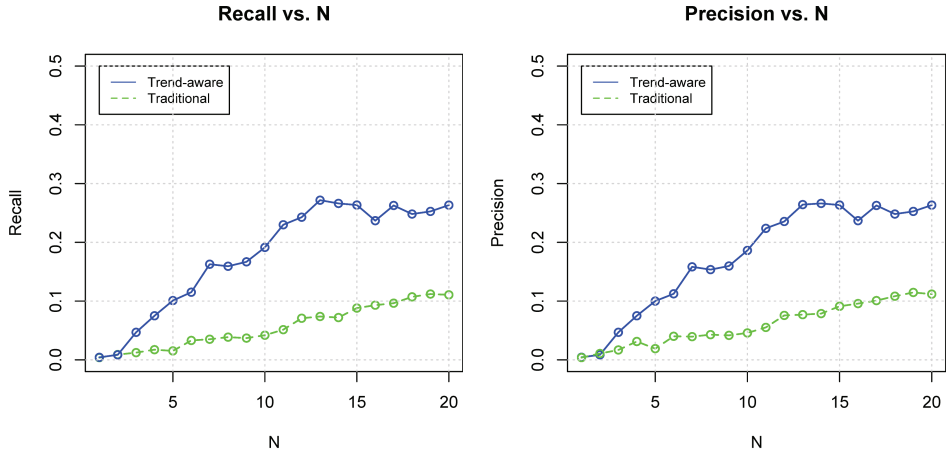
$$recall(N) = \frac{\#hits}{|T|} \tag{A.4}$$

$$precision(N) = \frac{\#hits}{N * |T|} \tag{A.5}$$

dont T est le test set, et N est le nombre d'objets à recommander. Enfin, pour faciliter l'intelligibilité des résultats, nous avons besoin d'un baseline. Et nous avons encore sélectionnez le item-based collaborative filtering, mais, cette fois, l'algorithme prendrait la matrice d'original de préférences personnels $\mathcal{P}$, où $p_{u,i}$ est 1 si u aime l'objet i et i est une tendance, sinon 0. Nous examinons dans quelle mesure notre système de recommandation trend-aware est capable de profiter des utilisateurs à la découverte des tendances personnalisés.

Item-based traditionnel vs. Item-based trend-aware. Figure A.3 montre la précision et le recall de collaborative filtering de item-based traditionnel et trend-aware, en fonction de le nombre de recommandations par personne (recommandations de top-N). Pour les deux systèmes, la précision et le recall améliorent quand le N augmente. Cependant, à accroître la valeur de N, la précision et le recall augmentent plus rapide dans le cas de trend-aware. Par exemple, au top-10 recommandations, la précision/rappel pour le système de recommandation de item-based traditionnelle est de 0.05, tandis que l'approche d'item-based de trend-aware atteint 0.2 (Figure A.3).

Cette différence significative des performances indique qu'un système de recommandation d'item-based traditionnelle ne serait pas capable de recommander les tendances,



(a) Recall vs. top-N (b) Precision vs. top-N

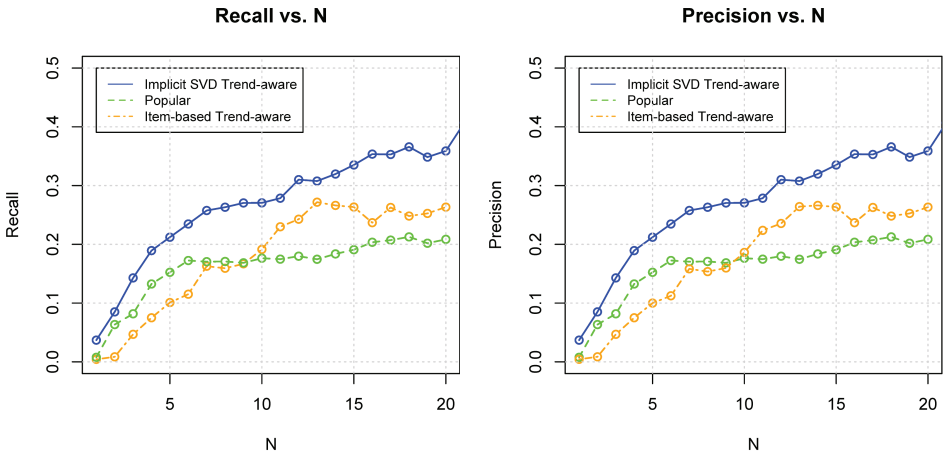
FIGURE A.3: Précision et Recall. Les resultats de recommandation trend-aware vs. recommandation item-based. Le N signifie le nombre de recommandations.

mais un système de trend-aware le serait. Ces résultats suggèrent également que, en présence de "sparsity" de data, en s'appuyant sur quelques experts est un moyen efficace de recommander tendances.

Item-based trend-aware vs. Implicit SVD Trend-aware. Nous testons ensuite si une approche populaire de factorisation de matrice - *implicite SVD* - permettrait d'améliorer la performance. Figure A.4 montre que ce soit le cas. Nous avons pu voir que, à tout top-N recommandation donnée, la précision et le recall de*implicite SVD* sont toujours mieux que le *item-based*. En outre, comme la taille de la liste des recommandées augmente, le système de recommandation de *implicite SVD* trend-aware améliore la précision et le recall plus rapide que l'item-based trend-aware.

Nos expériences ont montré la capacité d'item-based trend-aware à recommander des tendances. Et avec *implicite SVD*, la performance de notre système de recommandation trend-aware pourrait être améliorée encore plus loin.

Popularité. Dans un système de recommandation, des objets populaires (ceux qui reçoivent le plus grand nombre de votes) sont souvent le plus faciles à recommander, car il est très probable que leurs utilisateurs similaires ont déjà les voté [96]. Les tendances sont similaires à des objets populaires dans le sens que tous les deux reçoivent un nombre considérable de votes. Mais, les tendances sont le contenu qui reçoit des votes qui augmentent *brusquement*, tandis que le vélocité de participation à des objets populaires ne augmente pas nécessairement. Si la popularité est la seule raison que les gens consomment les tendances , puis recommander tendances populaires donneraient la meilleure précision. Nous examinons cette hypothèse en comparant la performance de notre système de recommandation trend-aware avec une stratégie simple de recommander seulement les tendances populaires. Ce qui est intéressant, comme nous pouvons le voir sur la Figure A.4, si le système de recommandation recommande seulement les



(a) Recall vs. top-N (b) Precision vs. top-N

FIGURE A.4: La précision et le recall. Les resultats de deux recommandations de trend-aware (item-based et Implicit SVD) et celui de recommander des tendances de le plus populaires.

tendances populaires, la précision et le recall serait le pire. La précision et le recall ne s'améliorent pas beaucoup après le top-5. Cela indique que, même pour les tendances - objets que l'on attend d'être non-long-tail - personnalisation a du sens. Mais jusqu'à un certain point, la précision et le recall des résultats sont limitées, et c'est parse que de la nature des tendances .

A.3.3 Résumé

Dans cette étude, nous avons montré que, sur l'activité, le réseau et les attributs géographiques, une approche de machine learning (SVM) permet d'identifier les utilisateurs principaux qui ont des compétences de niveaux différents de créer des tendances - trend spotters et makers. Si un objet sera une tendance ou pas peut alors être identifiée de façon fiable, selon que l'objet a été transféré par un trend maker réussie et voté par les spotters réussies. Nous avons alors vu que les systèmes de recommandation existants peuvent tirer profit de cette capacité d'identifier ces "utilisateurs spéciaux", et nous avons évalué l'efficacité du système en recommandant les tendances personnalisés en terme de la précision. Les résultats ont confirmé que les tendances - comme les objets non-long-tail - méritent d'être personnalisé ainsi.

A.4 Recommendations à La Sérendipité

Dans les études précédentes, nous avons présenté la conception et l'évaluation d'un système de recommandation de faire des suggestions personnalisées de tendances. Cependant, les systèmes de recommandation souffrent d'un inconvénient majeur. C'est, ils apprennent progressivement nos préférences et suggèrent les objets que nous pourrions ai-

mer certainement. Avec la précision comme l'objectif principal, ces systèmes rencontrent le problème suivant : ils "trap" des utilisateurs dans leur propre "filter bubble" (i.e., les objets recommandés peuvent être aimé que par les utilisateurs ayant des préférences similaires). La conséquence est que les recommandations deviennent de plus en plus concentrer autour des intérêts centraux, ainsi, ne pas fournir des découvertures diverses. Dans ce travail, nous nous attaquons au problème par l'amélioration de la qualité des recommandations en termes de la sérendipité. Nous explorons le potentiel de fournir des recommandations inattendues mais les gens peut-être aimer. Nous le faisons dans le cadre d'un média social de location-based (i.e., Foursquare [4]).

A.4.1 Notre Proposition

Nous proposons d'abord un algorithme de base pour générer des recommandations de localisation, puis de concevoir des techniques supplémentaires pour augmenter la sérendipité.

a. Algorithmes de Base
Comme les algorithmes de base, nous visons à fournir des recommandations "précises". Nous concevons d'abord un recommandation de base qui considère les plusieurs facteurs qui pourraient avoir une incidence sur sa décision de visiter un endroit, et nous le faisons en utilisant un modèle Bayésien. Plus précisément, nous considérons trois facteurs principaux : si l'utilisateur aime l'endroit (le taste), dans quelle mesure le lieu est (distance) et comment *mixing* d'une venue (qui reflète si le lieu est attirant pour les utilisateurs des goûts variés) .

Pour modéliser le goût de l'utilisateur et la social mixing d'une venue, nous proposons tout d'abord la notion de "user tribes" - une tribu se compose des utilisateurs visitent les mêmes venues ou les venues similaires. Pour cela, nous appliquons Latent Dirichlet Allocation (*LDA*) [16], tout ce qui a été initialement conçu pour apprendre les compositions de mots de chaque sujet, et l'importance de chaque sujet lier à chaque document [15, 16]. Pour paraphraser *LDA* dans notre cas, nous avons considéré une venues comme un document, et les personnes qui ont visité la venue comme les mots dans le document. Dans ce cas, les "topics" qui sont appris par *LDA* sont des groupes d'utilisateurs qui ont visité des venues similaires - nous les appelons "user tribes". Pour chaque venue, sa distribution de visiteurs de chaque tribus indique la mesure dans laquelle cette venue est visités par des personnes like-minded.

User Taste. Les gens visitent un venue parce qu'ils l'aiment. Pour prédire la mesure dans laquelle on voudrait un objet (dans notre cas, une venue) est le objectif principal d'un système de recommandation. Par conséquent, nous pourrions modéliser le goût de chaque utilisateur sur chaque venue en incorporant un algorithme populaire de item-based collaborative filtering. Une fois que nous avons identifié les user tribes, nous pouvons com-

[4] https ://foursquare.com/

parer la similarité entre chaque paire de venues (i, j) mesurés comme la distance cosine entre leur distributions de user tribes. Après, pour chaque venue i, l'algorithme de item-based maintenant émet un score $l_{u,i}$ personnalisé pour chaque utilisateur u basée sur la similarité de cette venue et les venues qui sont visités dans la passé($\mathcal{H}_u$). Et ce score $l_{u,i}$ reflète le goût prévu de chaque utilisateur u pour chaque venue i.

Physical Distance. Un utilisateur visite une venue pourrait parce que 1) il est juste dans le coin, ou 2) il remplit ses intérêts. Selon la mesure dans laquelle on aime une venue ou la venue remplit ses intérêts, sa volonté de se déplacer changements. Par conséquent, pour chaque venue i et utilisateur u, nous considérons la distance $d_{u,i}$ (en mètres) entre le lieu i et le centroid de toutes les coordonnées (latitude et longitude) de lieux que u a visité.

Social Mixing. Des venues peuvent attirer des gens différents : certains venues sont adaptés à la crowd de niche, tandis que d'autres sont accessible pour tous. Le dernier type de venues encourage plus la vie sociale. Nous définissons le *mixing* sociale d'une venue comme la mesure dans laquelle le venue attire des groupes d'utilisateurs diverses. Nous calculons un score de mixing s_i pour chaque venue i comme la diversité de Shannon [71] du vecteur w_i (qui a autant de k éléments comme des nombre de user tribes) - le plus différent il y a des user tribes d'une venue, la plus mixité sociale est la venue.

Bayesian Modeling. Avec le modèle Bayésien, nous pourrions traduire la tâche de générer des recommandations comme prédire si un utilisateur visitera une venue. Pour formuler des recommandations précises, les venues avec une grande probabilité d'être visités sont des endroits d'être recommandé. Nous examinons ensuite les variables L (liée à Likes prévus), D (Distance géographique) et S (de la social mixing de cette venue) obtenus par discrétiser respectivement $l_{u,i}$, $d_{u,i}$ et s_i, nous voulons calculer la probabilité de événement G (Go) - qui est, si l'utilisateur u visites la venue i. Pour savoir comment importante la social mixing de venues, nous appliquons deux modèles Bayésiens. La première ne tient pas compte de la valeur de S et le modèle est donc sur le calcul de $p(G|L, D)$. Le second est sur du plein $p(G|L, D, S)$. Nous obtenons ces deux valeurs comme :

$$p(G|L, D) = \frac{p(L|G, D) \times p(G|D)}{p(L|D)} \tag{A.6}$$

and

$$p(G|L, D, S) = \frac{p(S, L|G, D) \times p(G|D)}{p(S, L|D)}, \tag{A.7}$$

Les venues obtenues les plus grands valeurs de $p(G|L, D, S)$ (or $p(G|L, D)$)(qui ont les plus grands possibilités d'être visités.) vont être des top recommandations.

b. Beyond User History

Il première manière que nous nous proposons de renforcer la sérendipité construit un graphe "préférence locale" pour chaque utilisateur u, dont les noeuds sont des venues que u a visités (H_u), et un bord $e_{i,j}$ entre les noeuds i et j existe si leur similarité est supérieure à la similarité moyenne entre touts les venues que u a visité. Les venues qui sont bien connectés forment des clusters. Des venues dans le même groupe peuvent apparte-

nir à un même type, mais, ceux qui sont liés entre les clusters sont des venues "brokerage" (i.e., des venues qui n'appartiennent pas nécessairement à un même type). Dans la prochaine étape, nous ajoutons chaque venue *candidat* x (une venue à recommander potentiellement) temporairement à la graphique de "préférence locale" de u. Encore une fois, nous créons des bords $e_{x,i}$ seulement si $\mathrm{sim}(x, i)$ est supérieure que la similarité moyenne entre les venues de l'histoire du u. Pour introduire la sérendipité, nous récompensons des venues qui se trouvent sur le bord de clusters différents dans graphique de préférence du u, nous les rangeons par les clustering coefficient $c_{x,u}$ de noeud x dans le graphique local du u.

c. Beyond User Routine
Notre deuxième moyen d'améliorer la sérendipité (que nous appelons "Beyond User Routines") rompt des triangles itinéraire des utilisateurs. Pour ce but, nous transformons la graphique de la préférence locale vu ci-dessus dans une graphique *routine*. Pour résoudre ce problème, nous considérons les aspects temporels des venues qui caractérisent généralement les activités. Des venues de catégories différentes ont différents modèles de checkin quotidien. Cela implique que dans la routine de u, des venues de la maison, du travail et ailleurs auront différents modèles de checkin temporelles. Nous construisons une graphe $routine$ de u dans lequel les noeuds sont des venues que u a visité (H_u), pour chaque paire de venues (i, j), un bord $e_{(i,j)}$ est ajouté si la similarité cosine de le modèle temporel de checkin de i et j est *moins* que le moyenne.

En conséquence, les venues connectés dans le graphe de routine constituent un ensemble de lieux où les activités différents peut être effectuées. De même façon que nous avons fait pour l'algorithme précédent, nous ajoutons alors provisoirement chaque venue de candidat x à le graphe routine de u. Nous ajoutons un bord entre les venues x et i si leur similarité est supérieure à la moyenne entre les venues que u a visité. Enfin, nous nous classons des venues de candidats par leur clustering coefficient dans les graphes routine. En recommandant des venues avec un clustering coefficient inférieur, nous biaise des recommandations vers des endroits qui pourraient avoir des activités différentes de celles où u a passé souvent.

Dans les deux algorithmes de "beyond user history" et "beyond user routine", des venues de candidats se sont rangés par leurs clustering coefficients. Pour chaque utilisateur u, nous combinons ensuite ses recommandations de base avec ces deux techniques par interpoler le percentil ranking ($r_{\mathrm{base},x,u}$) de chaque venue de candidat (x) dans les résultats de recommandation de base (modèle Bayésien) avec ses percentile ranking personnalisés ($r_{\mathrm{algorithme},x}$) sortie des techniques de la sérendipité augmentée :

$$\mathrm{ranking}_{x,u} = (1 - \alpha) \cdot r_{\mathrm{basic},x,u} + \alpha \cdot r_{\mathrm{algorithm},x,u}. \tag{A.8}$$

dont $r_{\mathrm{algorithm},x,u}$ est le percentile ranking sortie des algorithmes de clustering-based - "beyond user history" et "beyond user routine".

A.4.2 Evaluation

Nous évaluons la performance de nos modèles Bayésiens de base décrites ci-dessus (avec le percentile ranking comme l'evaluation de la précision et KL divergence [71] entre les recommandations et les venues visités dans la passé comme la mesure de la sérendipité), et nous considérons les trois variantes :

i+d (item-based + distance). Dans cet algorithme, nous calculons des recommandations sans tenir compte de l'option de "social mixing", en adoptant L'équation A.6. De plus, nous ne profitons pas de la user tribes du LDA : chaque utilisateur est considéré comme un user tribe différent, résultant en un ensemble de scores $l_{u,i}$ qui reflètent une recommandation traditionnelle de item-based. Nous considérons cela comme notre algorithme de base.

i+d+s (item-based + distance + social mixing). Ici, nous améliorons l'algorithme précédent en tenant compte du trait de social mixing : des venues de candidats se sont range par leur scores de L'équation A.7.

L+d+s (LDA + distance + social mixing). Ce dernier algorithme combine toutes les traits que nous avons proposées : ici, nous introduisons les user tribes sorties de l'algorithme de LDA.

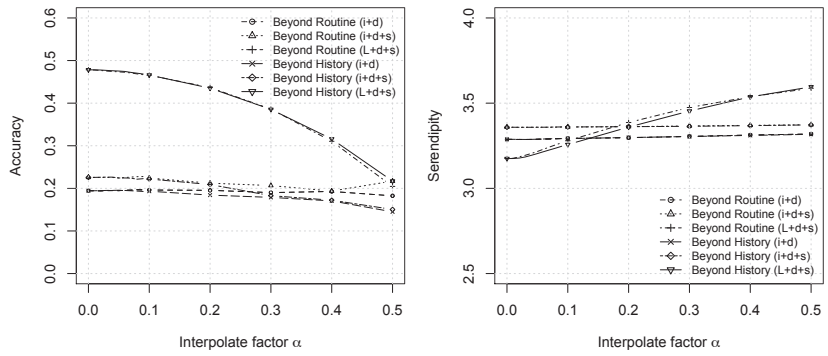
Pour les trois algorithmes, nous obtenions les valeurs de L, D and S afin que chaque classe discrète obtient des nombre de samples suffisants, comme suit : $L = \lfloor 100 \times l_{u,i} \rfloor$, $D = \lfloor \log_{10} d_{u,i} \rfloor$, $S = \lfloor s_{u,i} \rfloor$.

Modèle	Traits	Précision	Sérendipité
Baseline (i+d)	item-based + distance	0.195 $\pm$0.048	3.288 $\pm$0.017
(i+d+s)	item-based + distance + social mixing	0.226 $\pm$0.061	**3.359** $\pm$0.019
Full (L+d+s)	LDA + distance + social mixing	**0.478** $\pm$0.034	3.175 $\pm$0.020

TABLE A.6: La précision et la sérendipité de nos trois algorithmes de base. Pour LDA dans le dernière modèle, le nombre de user tribes (k) est fixé a 100, en raison de le meilleur précision.

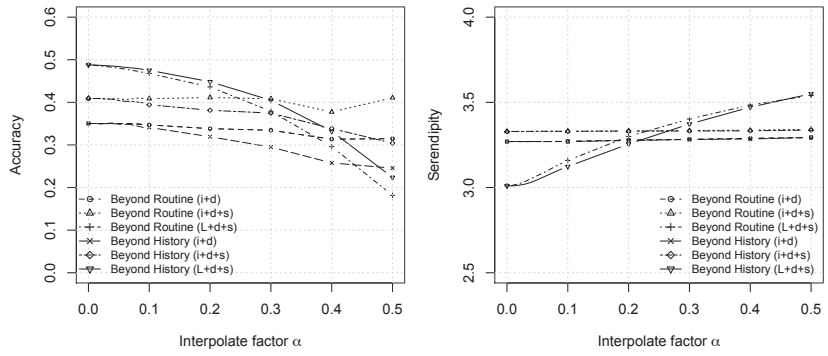
Précision vs. Sérendipité des trois modèles de base. Pour éviter des venues facile a prévoir, nous considérons tous les venues que ne sont pas dans les catégories de résidences/travail/éducation. Par rapport à la baseline *(I+D)*, la méthode de *(i+d+s)* (qui considère l'option de "social mixing") augmente la précision. Le modèle de *LDA*-based effectue le meilleure en terme de précision et il aussi équilibre juste entre la précision (ce qui est le double de celle du modèle item-based *(i+d+s)*) et sérendipité (qui est comparable à modèle item-based). Ces résultats se réfèrent au cas où le nombre de user tribes (k) dans *LDA* est fixé à 100.

Précision vs. Sérendipité des deux modeles de sérendipité augmenté. Le modèle *LDA*-based renvoie la précision mieux que la de item-based (Figure A.5(a)). Le facteur d'interpolation α n'affecte pas la sérendipité de item-based, mais incidence sur le modèle *LDA*-based ; d'autre part, il est possible d'améliorer la sérendipité des venues recommandés



(a) Précision (b) Sérendipité

FIGURE A.5: La précision et la sérendipité de (top 10) recommendations. Ils considèrent tous les utilisateurs (i.e., des personnes qui ont visite au moins deux venues).



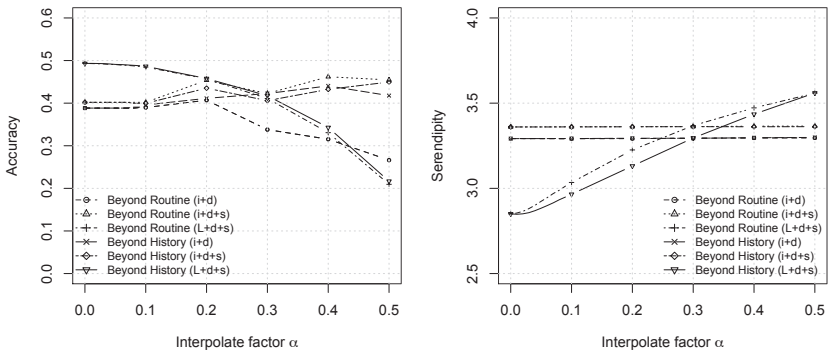
(a) Précision (b) Sérendipité

FIGURE A.6: La précision et la sérendipité de (top 10) recommandations. Ils considèrent des utilisateurs qui ont visité au moins 5 venues.

par incrémenter la paramètre α, comme on peut le voir sur la Figure A.5(b). L'inconvénient est que, en augmentant α, la précision decreses. Un bon compromis semble être une valeur de $\alpha \in [0.2, 0.3]$, où le modèle *LDA*-based offre une précision relativement élevée (bien au-dessus de 0.4) ainsi que de haute sérendipité (au-dessus de 3.4).

Des impacts des activités. Pour tester comment les niveaux d'activité des utilisateurs influe les résultats, nous considérons les utilisateurs qui ont visité au moins 5 et ceux qui ont visité au moins 10 venues séparément. Dans ces situations de inférieure rareté de dataset, le modèle item-based augmente sa précision (Figures A.6 et A.7), mais n'atteint pas celle du modèle *LDA*-based en situation de haute rareté de dataset (Figure A.5) - la précision est inférieure à 0.40. Plus généralement, cela suggère que la précision de modèle *LDA*-based est en effet plus robuste à faible densité de dataset. Les activité de l'utilisateurs n'a pas d'impact de la métrique de sérendipité.

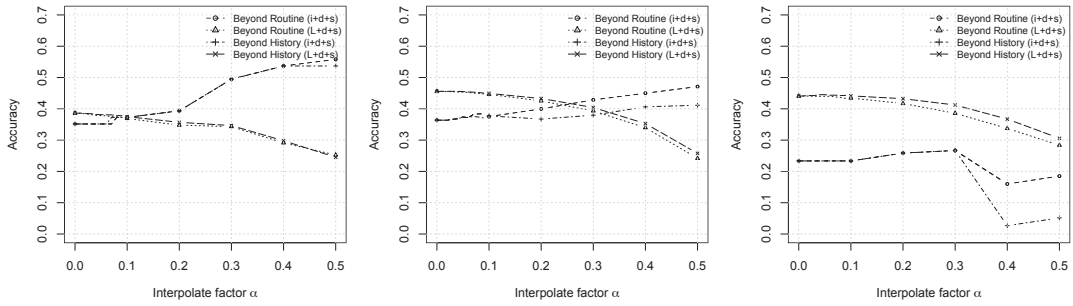
Des impacts des tendances de social mixing aux niveaux differents. Nous considérons trois types d'utilisateurs : ceux dont la moyenne des values social mixing de venues visités est dans le premier quartile (niche users), ceux dont la moyenne est dans le der-



(a) Précision (b) Sérendipité

FIGURE A.7: La précision et la sérendipité de (top 10) recommandations. Ils considèrent des utilisateurs qui ont visité au moins 10 venues.

nier quartile (social mixers), et ceux qui restent (average mixers). Pour ces trois types, nous n'enregistrons pas de changement de la sérendipité dans tous les algorithmes. En revanche, la précision montre certaines différences. Pour les niche users, le modèle item-based montre la précision mieux que le modèle LDA-based (Figure A.8(a)). C'est peut-être parce que regroupant les utilisateurs dans la représentation compacte de "user tribes" peut conduire à la perte de l'information, en ce que, il peut diluer l'information spécificités à la personne. Les résultats sont inverse pour social mixers (Figure A.8(c)) : compte tenu de la variété à fine-grained d'autres utilisateurs que mixers rencontrent, la précision augmente.



(a) Niche (b) Average Mixers (c) Social Mixers

FIGURE A.8: Des impacts des tendances de social mixing aux niveaux differents.

A.4.3 Résumé

Dans ce chapitre, nous avons conçu un système de recommandation mobile qui produit non seulement des recommandations précises, mais aussi à la sérendipité. Grâce à l'analyse expérimentale, nous avons observé qu'il existe deux catégories d'utilisateurs dans un média social de location-based : ceux qui vont à la venue de niche (par exemple, les venues visités par les personnes partageant les mêmes idées) et ceux qui vont à la venue

populaire. Chaque personne est à l'aise avec du niveau différent de sérendipité. C'est pourquoi nous avons fait le compromis α entre la précision et la sérendipité accordable. Pour contrer la "sparsity" de dataset, nous avons proposé une approche basée sur *LDA*, qui regroupe les personnes partageant les mêmes idées dans les mêmes "user tribes". Cette approche augmente la précision de recommandation, surtout pour les utilisateurs qui n'ont pas été très actifs. En outre, la caractérisation des venues par le trait de "social mixing" (c'est à dire, sa tendance à être visité par les user tribes différents) augmente recommandation précision aussi.

A.5 Conclusion

Cette thèse aborde un problème important dans l'analyse des médias sociaux, que de la compréhension qui créent le contenu de tendance dans les réseaux sociaux. Des efforts de recherche antérieurs sur des tendances ont porté sur deux perspectives parallèles : les tendances - l'entité elle-même, et la source des tendances - des gens qui créent les tendances. Alors que la première direction de recherche est principalement sur les techniques de saisir les tendances par les propriétés particulières de leur "tipping point", le travail dans la seconde direction de recherche a été centrée autour de la notion d'être "influent". Et la discussion pour savoir si les gens qui créent les tendances sont "influent" présente des points de vue différents.

Cependant, l'influence - comme le pouvoir de persuader les autres d'accepter son idée - est une fonction de la population, le contenu et l'environnement (par exemple, activité, lieux, etc.) S'appuyant sur les conclusions précédentes sur les personnes influentes, nous avons redéfini dans cette thèse deux types de personnes qui contribuent à la création de tendances - trend makers (ceux qui génèrent les tendances) et trend spotters (ceux qui propagent les tendances). Grâce à une analyse approfondie d'un large éventail de fonctions de trend makers et trend spotters - l'activité, le contenu, les liens sociaux et ceux géographiques (Chapitre 3), nous avons montré que les tendances sont en effet créés par "special users", et dans les sites de médias sociaux, ils semblent être *beaucoup* plutôt que *a few*.

Les caractéristiques attrayantes des trend makers et les trend spotters rendent distinguer des autres utilisateurs typiques, cela qui nous donne les connaissances sur les causes sous-jacentes des tendances dans les médias sociaux. En outre, ces utilisateurs particuliers notables apportent l'occasion d'identifier les tendances d'une manière nouvelle - de s'appuyer sur la sagesse des origines (i.e., les gens qui créent les tendances). Nous avons montré comment cette idée peut être intégré et utilisé comme l'un des éléments principaux dans un système de recommandation trend-aware (Chapitre 4), qui est représenté ses capacités de servir les tendances personnalisés efficacement aux personnes avec les intérêts différents.

Un système de recommandation concentré de précision est souvent censé apprendre parfaitement des préférences d'une personne et de sortie les recommandations de le plus en plus "proches" - ceux qui sont au centre de ses goûts. Adopter activement ces recommandations précises rétrécit progressivement la gamme de l'exploration pour des utilisateurs. Pour élargir son champ des recommandations, diverses stratégies d'analyse de réseau sont étudiées dans le Chapitre 5. En particulier dans un système de recommandation location-based, nous avons montré que en tirant parti des techniques d'analyse de réseau, la précision et la sérendipité pourraient être équilibrées.

A.5.1 Contributions

Dans cette thèse, nous avons : 1) analysé les facteurs humains dans la création de tendances ; 2) ont étudié dans quelle mesure les utilisateurs spéciaux pourraient peuvent contribuer a l'identification des tendances, et 3) les outils déployés pour faire personnalisations dans les médias sociaux mobiles. Nos contributions globales sont pertinentes pour les deux thèmes de recherche principaux - les tendances et les systèmes de recommandation dans les médias sociaux.

Tendances
A la recherche des réponses à nos questions de tendances, nous faisons les trois contributions principales suivantes.

Facteurs Humains. Notre analyse a révélé que la création de tendances est un processus combiné, dans lequel deux types d'utilisateurs sont engagés - des individus ordinaires qui ont des intérêts spécifiques et sont liées à des utilisateurs de différents groupes (définis comme trend makers dans cette thèse) et des adopteurs précoces avec des intérêts divers (défini comme trend spotters). Tout les deux types peuvent être identifiés à partir des autres utilisateurs typiques avec des caractéristiques (activités, consommation de contenu, les connexions de réseau social et caractéristiques géographiques, etc) à l'aide des outils de machine learning standard tels que SVM ou un modèle de régression logistique.

En compte du fait que pas tout le monde est aussi capable à la création des tendances, nous avons regroupé les trend makers/spotters en trois classes (i.e., haute, moyenne et basse) en fonction de leurs niveaux différents, et nous avons étendu les modèles statistiques pour identifier à quele classe de succès un trend maker/spoter appartient.

Identifications. Sur la base de trend makers/soitters de niveaux differentes identifiées, nous avons montré que les tendances pourraient être identifiés avec un modèle de régression logistique. Dans le même temps, les coefficients du modèle indiqué qu'un objet a une bonne chance de devenir une tendance si sa uploader est trend maker de haut niveau, et si elle reçoit de grandes attentions de trend spotters succès.

Explorations. Enfin, l'intégration des modèles statistiques pour identifier les tendances en utilisant une technique de collaborative filter (i.e., implicite SVD), nous avons conçu

un système de recommandation trend-aware d'aider efficacement les utilisateurs à découvrir des tendances contenu proche de leurs préférences.

Systèmes de Recommandation
En parallèle à la recherche sur les tendances dans les sites de médias sociaux, notre travail dans cette thèse contribue également à l'état de l'art dans les systèmes de recommandation. Plus précisément, nous avons examiné la conception de systèmes de recommandation adaptés aux deux réseaux sociaux *mobiles* différents.

Précision Concentré. Tout d'abord, nous avons conçu un système de recommandation trend-aware pour servir les utilisateurs avec une tendance de leur intérêt. En plus de la pratique traditionnel de construire d'un système de recommandation, nous proposons un nouveau moyen d'enrichir la matrice des préférences traditionnelles par la convertir en "trend-aware". Visant à faire des recommandations "précises" sur les tendances, le système s'est montré sa capacité de recommander les tendances efficacement. En outre, nous montrons que recommander les tendances surpasse recommandant les contenu populaire en général.

Sérendipité Augmentée. Deuxième, nous avons exploré le potentiel d'amélioration la sérendipité des recommandations. Plus précisément, nous avons réalisé l'expérience dans le contexte d'un système de recommandation mobile où *location* est l'informations principaux. Nous avons conçu le système de recommandation location-based en intégrant les préférences et les distances physiques à l'aide d'un modèle Bayésien. Alors que nos stratégies visant à promouvoir les vneues au-delà des recommandations des utilisateurs qui partageant les mêmes idées et de la routine sont capable d'améliorer efficacement la sérendipité, nos analyses sur les utilisateurs montrent que les gens préfèrent da la sérendipité de niveaux différents. Ces différences doivent évidemment être pris en considération dans le processus d'équilibrage de la précision et de la sérendipité.

Alors que les systèmes de recommandation sont conçus comme des outils pour faire des personnalisations des objets "long tail" (un grand nombre d'objets qui ont quantité relativement faible d'adoptions [7]), notre travail dans cette thèse montre que il est également logique de personnaliser les objets non-long-tail comme des tendances dans les médias sociaux. En outre, la conception d'un système de recommandation devrait être application spécifique, peut être adaptée au contexte (tels que le type d'objets, les raisons pour quelqu'un à adopter un objet, et le niveau de ses adoptions, etc).

Le recherche présenté dans cette thèse a un impact immédiat sur les chercheurs qui s'intéressent à la compréhension de la diffusion d'informations dans les médias sociaux en général, en particulier pour ceux qui sont intéressés à : 1) identifier les événements (par exemple, de nouvelles tendances) ; 2) comprendre la propagation des opinions et 3) conception d'une stratégie de marketing viral, etc. Il est également pertinent pour les praticiens qui cherchent à renforcer l'expérience des utilisateurs de la consommation personnalisée.

A.5.2 Future Work

Ce travail s'appuie sur deux datasets de deux réseaux sociaux mobiles réels. Alors que nous avons une bonne confiance dans la généralité de nos observations, il serait intéressant de mener une étude comparative sur les réseaux sociaux différents pour comprendre l'impact de traiter avec différents modes d'utilisation et les différents types de utilisateurs. En outre, cette thèse pourrait être étendu aux directions suivantes :

Géographie. Les réseaux sociaux en ligne devient un élément important de notre vie quotidienne, un rapport sur les médias sociaux [80] affirme que *"quand il s'agit d'accéder à un contenu social, il s'agit de mobiles"*. La caractéristique la plus attrayante que les applications mobiles portent, c'est qu'ils sont géolocalisation.

Alors que la théorie populaire de "six degrés de séparation" [116] dit que nous vivons dans un petit monde où tout le monde pourrait être reliée à une autre personne dans les six étapes, le travail de recherche a révélé que les connexions et la mobilité sociale des gens sont encore limitée par les distances géographiques [27, 68, 84]. Depuis la diffusion de contenu sur les liens sociaux est essentiel à la création de tendances, ces contraintes géographiques pourraient effectuer tendances ainsi, et en tant que telle liés au processus de diffusion. Un exemple typique est liée à la notion de l'identification de l'événement. Dans ce cadre, une tendance pourrait être un événement mondial, ou ce pourrait être un événement local ainsi [129]. Limité par la taille du dataset, notre travail sur les tendances a été axée sur les tendances globals de le dataset entier de l'application sociale mobile. Cependant, dans une application mobile social plus large, il serait intéressant de séparer les tendances locales de celles mondiaux. Le contrôle de ces deux classes de tendances, on pourrait étudier : 1) quelles sont les différences principales entre les trend makers des tendances mondiales et des tendances locales ? 2) s'il ya des trend makers/spotters mondiales/locales ? 3) la mesure dans laquelle les trend makers et spotters mondiaux/locaux contribuent à la création de tendances de niveaux différents ?

Dynamics Temporelles. Dynamique temporelle est un autre facteur important qui doit être pris en considération. Liées à notre travail, il existe deux types de dynamique temporelle a étudier : la dynamique temporelle des tendances et la dynamique temporelle des préférences de l'utilisateur.

- **Tendances.** Dans le Chapitre 4, nous avons vu que les tendances persistent plus longtemps que le contenu normal, mais le volume de l'attention qu'ils acquièrent s'arrête à augmenter après une période. Cependant, dans certains cas (par exemple, quand une tendance est un phénomène de mode), la durée de vie d'une tendance peut être cyclique [1]. Prenant la complexité temporelle en considération, les études sur les individus dans la création de tendances pourraient être encore étendues, ainsi que le modèle pour identifier les tendances par effet de levier la connaissance de leurs créateurs.

- **Preferences de Utilisateurs.** Les préférences des utilisateurs changent avec le temps, c'est un problème bien connu dans les études de recherche sur les systèmes de recommandation. Notre système de recommandation trend-aware pourrait être améliorée par l'intégration avec un composant supplémentaire pour modéliser le déplacement des préférences des utilisateurs. En outre, l'évolution des préférences des utilisateurs peut également conduire à la modification du niveau de sérendipité on accepte. Une étude approfondie pourrait être effectuée pour comprendre cet impact, et ensuite être prise en compte dans la pratique de générer des recommandations à la sérendipité.

Online Updating. Un système de recommandation apprend ses préférences. Cependant, il impacts son choix dans l'adoption des articles. Comme nous l'avons vu au Chapitre 4, dans notre système de recommandation trend-aware, les individus sont recommandés avec les tendances de leurs préférences. La conséquence est que dans le cadre de l'acceptation du tendance recommandé, les utilisateurs sont "formés" pour être trend spotters. Comme le système repose sur l'identification des trend spotters des niveaux différents de succès, les modèles d'identification doivent ensuite être mis à jour périodiquement. Mais jusqu'à quel point les modèles sont insuffisantes à faire prédiction précisé, elle nécessite une analyse plus approfondie. Une récente proposition pour résoudre ce problème qui pourrait être utilisé dans notre contexte, est d'intégrer les contrôleurs pour estimer automatiquement la fréquence de mise à jour des modèles [53].

Analyses Sentimentaux. Dans cette thèse, nous avons observé que les gens se comportent différemment. Par exemple, beaucoup des comportements différents ont été observés parmi les trend makers, trend spotters, et les utilisateurs typiques. Et dans la réaction aux recommandations, les gens se retrouvent également à accepter les différents niveaux de la sérendipité. On peut se demander quelles sont les causes fondamentales de la diverse dans les comportements humains ?

Des questions similaires ont été posées à la recherche de la puissance différente des influents sociales, et des explications ont été demandées par la recherche dans la divergence des traits de personnalité [41, 93, 94].Toutefois, si les traits de personnalité pourraient être utilisés pour expliquer les différents comportements des trend makers, trend spotters et les utilisateurs typiques doit encore être examiné.

De plus, certains chercheurs ont commencé à faire des recommandations fondées sur des personnalités de l'utilisateur [50, 88]. Bien que notre travail a porté sur la modélisation de la personnalité par celui qu'ils notent ou préfèrent dans la passée, il serait également intéressant d'étudier si la personnalité pourrait expliquer les différents niveaux de sérendipité que chaque personne peut accepter, et pourrait donc être également mis en pratique lors de l'accord de l'équilibre entre la précision et la sérendipité dans nos recommandations personnalisés aux individu de personnalités différentes.