Measurement and Analysis of Tips in Foursquare

Yang Chen[1,2,3], Yuxi Yang[1,2], Jiyao Hu[1,2], Chenfan Zhuang[4]
[1]School of Computer Science, Fudan University, Shanghai, China
[2] Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China
[3] The State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China
[4]Malong Technologies, Co. Ltd., Shenzhen, China
Email: {chenyang, yangyx13, hujy13}@fudan.edu.cn, fan@malongtech.cn

Abstract—**Being a leading online service providing both local search and social networking functions, Foursquare has attracted tens of millions of users all over the world. As a location-centric platform, Foursquare maintains the information of numerous venues, and it has recorded a tremendous amount of users' tips for these venues. Tips (micro-reviews) play a critical role in helping users find a good venue and providing comments for the venue owners. A lot of studies have been done in investigating social connections and check-in patterns among Foursquare users. However, there is a lack of a thorough study on tips, the primary type of user-generated contents (UGCs) in Foursquare. In this paper, by crawling and analyzing all tips published by more than 6 million users, we study Foursquare's tips from various aspects. We start from counting the number of tips published by different users and conduct a group-based analysis of the average number of published tips per user. Moreover, we look into the important fields of tips. We study the venue category distribution, and the evolution of the number and diversity of tips. Finally, we conduct a series of sentiment analysis by referring to the texts of all tips and introduce the concept of *happiness index* to evaluate the overall level of satisfaction of a set of tips. To the best of our knowledge, our study presents the first comprehensive and unbiased picture of tips in Foursquare.**

I. Introduction

The rapid development of mobile computing technologies and social networking services drives a significant growth of location-based social networks (LBSNs). Founded in early 2009, Foursquare has been a leading LBSN for several years. As in Aug. 2014, it has attracted more than 50 million users around the world [1]. In Foursquare, users are allowed to leave tips for different venues. Tips can help Foursquare users get information of a venue before visiting it and provide useful comments for the venue owners. Earlier papers [5], [6] have done a lot in studying the social connections and check-in patterns of Foursquare users. However, a comprehensive and unbiased study of tips, the primary type of UGCs in Foursquare, is still needed. Getting in-depth knowledge of tips is useful for different relevant entities, including users, venue owners and the Foursquare platform operators.

Our study is based on a massive set of tips posted by 6.52 million Foursquare users. To the best of our knowledge, our work is the first to provide a comprehensive and unbiased view of Foursquare's tips. We have the following key findings.

First of all, we get a statistical overview of the number of tips per user in Foursquare. We can see that the distribution of the number of tips can be approximated by a two-term exponential model. We also group the users according to their profile photos, gender and country information, respectively. We conduct a comparative study for different user groups.

In addition, we scan the key information fields of all tips. We investigate the venue category distribution, and the evolution of the number and diversity of tips published in each quarter.

Last, we dive into the texts of all tips, and use sentiment analysis to understand the publisher's opinion for each tip. We introduce the new "happiness index" metric to see what type of venues can make people happier and which group of users post more positive tips.

II. Background and Data Collection

Since 2009, Foursquare has been a leading site for the combination of location-based services (LBS) and mobile social networking. Different from traditional online social networks (OSNs) such as Facebook, all activities on Foursquare are related to different venues. When Foursquare was founded in 2009, the two key functions were leaving tips and conducting check-ins. However, since Aug. 2014, the check-in function has been removed from Foursquare and integrated to a new app so-called Swarm. Unfortunately, most of the existing work, such as [5], [6], [8], have not demonstrated an unbiased and comprehensive picture of tips in Foursquare. Therefore, in this paper, we choose tips in Foursquare for a thorough study.

Each user in Foursquare is assigned a numeric ID, and the IDs are assigned sequentially. If a user has registered earlier, she will have an ID with a smaller number. Therefore, we can get the maximum ID number by registering a new account, and further generate a random subset of user IDs between 1 and the maximum ID number. We did the data collection from Nov. 3 till Nov. 8 in 2015 by using a crawler implemented by us [1]. Similar to other OSN sites, Foursquare also employs an IP-based rate limiting policy. To crawl the data quickly, we introduce the crowd crawling framework [4], which allows us to use a pool of IP addresses to improve the crawling speed. We launch 60 virtual machines on the East US data center of the Microsoft Azure platform, and each of them has an independent IP address. In total, we have successfully fetched the data of 6.52 million users using this collaborative way. For each user, we get her profile and all published tips. We can further see her firstname, number of published tips, number of followings and number of followers. Also, a user

[1] **https://github.com/chenyang03/Foursquare_Crawler**

can optionally add a profile picture, the gender information, the current location, a Facebook ID, a Twitter ID and a biography to her profile [1]. For each tip, we can get the time when it was published, the text of the tip, the country of the venue and the category information of the venue. For example, a sample tip can be represented as {user_id:12345, tip_text: "Super nice restaurant! It provides nice food, and it has a great location!", venue_country: US, venue_category: Food, date:"Oct. 21, 2015"}.

Among all 6.52 million users, 67.22% of them have uploaded a profile picture, while the rest 32.78% have not. For the gender information, we can see that 51.31% of them are male; 42.16% are female; and the rest 6.53% do not want to disclose their gender. For the users' home countries, 87.75% have added some information to the "location" field. We use the Google Maps Geocoding API to infer each user's home country. According to the number of users, the top four countries are USA (30.36%), Turkey (13.06%), Indonesia (9.76%) and Brazil (6.26%), respectively.

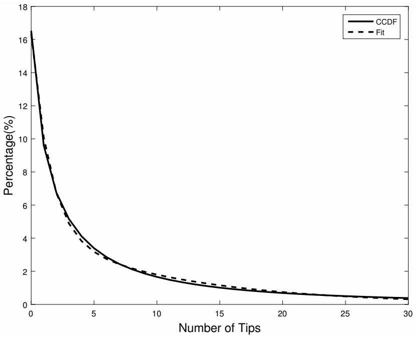


Fig. 1: CCDF of the Number of Tips per User

III. Data Analysis

In this section, we conduct a series of studies to understand tips in Foursquare from different aspects. We first examine the number of tips published by each user, and conduct a series of group-based analysis (§ III-A). Moreover, we check the important fields of tips, and show the venue category distribution, and the evolution of the number and country diversity of tips (§ III-B). Furthermore, we conduct sentiment analysis to see what kind of venues can make people happier and which group of users publish more positive tips (§ III-C).

A. Counting the Number of Tips Published by Each User

In this subsection, we first count the number of tips each user has published and see the distribution of each user's total number of published tips. We can observe the difference among users according to the number of published tips. In Fig. 1, we show the Complementary Cumulative Distribution Function (CCDF) of the number of tips published by each user. We can see that 83.47% users have never published any tip. In other words, most of the Foursquare users are tip readers rather than publishers. Therefore, the average number of a user's published tips is only 0.89. We also rank the users according to the number of published tips. We find that the top 1% users published 47.54% of tips, and the top 10% users published 92.67% of tips.

To better understand the distribution of the number of tips per user, we use four classic distributions, i.e., power law ($P(k) \propto k^{-\alpha}$) [2], power law with exponential cutoff ($P(k) \propto k^{-\alpha}e^{-\lambda k}$) [2], lognormal ($P(k) \propto e^{\frac{(lnx-\mu)^2}{2\tau^2}}$) and two-term exponential ($P(k) \propto ae^{bx} + ce^{dx}$), to see whether it can be approximated by any of them. The fitting accuracy and parameters are calculated using the *cftool* function (Curve Fitting Tool) in MATLAB. To see how well a model fits the data, we use the coefficient of determination, i.e., the R^2 value. The range of a R^2 value is between 0 and 1, and a value of 1 indicates the fitting is perfect. According to our comparison, we can see that the two-term exponential model performs the best ($a = 12.09, b = -0.6727, c = 4.265, d = -0.08713$), and the corresponding R^2 value is 0.9971. Fig. 1 indicates that the model approximates the distribution very well.

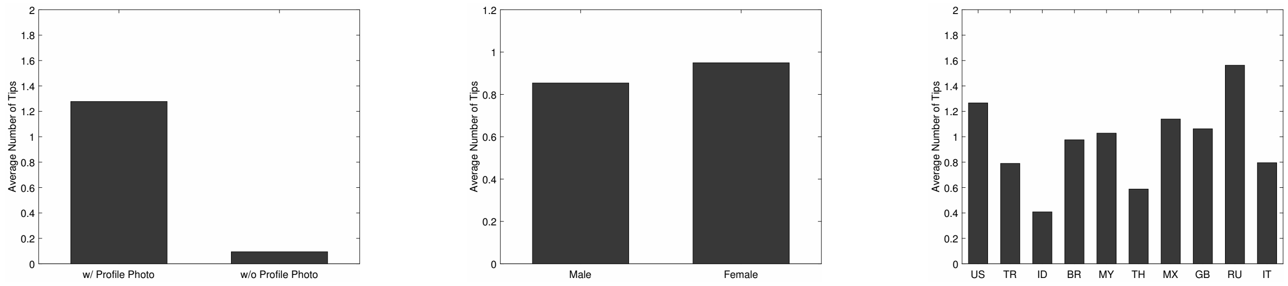
Besides considering all users as an integrated whole, we also divide users into groups based on their profile photos, gender information and home country information, respectively. According to Fig. 2(a), among the users who have uploaded profile photos, the average number of published tips per user is 1.28. In contrast, this number is only 0.09 for the users who have not uploaded profile photos. Therefore, whether or not uploading a profile photo is an indicator for the number of published tips. The gender difference is shown in Fig. 2(b) and we find that the average number of tips published by male users is 0.85 while this number is 0.95 for female users. In other words, in average female users published about 11.76% more than male users in terms of the number of tips, while the difference between male and female is not very significant. In Fig. 2(c), we look at the users' home countries and focus on the users from the 10 countries with the highest Foursquare population. Among them, users in Russia and the United States are more active in publishing tips, while users from Indonesia are the least active.

B. Scanning the Important Fields of Tips

In addition to the user-centric study of the number of tips, we classify all tips according to different key information fields of a tip, such as the venue's country, the venue's category and the published date of a tip.

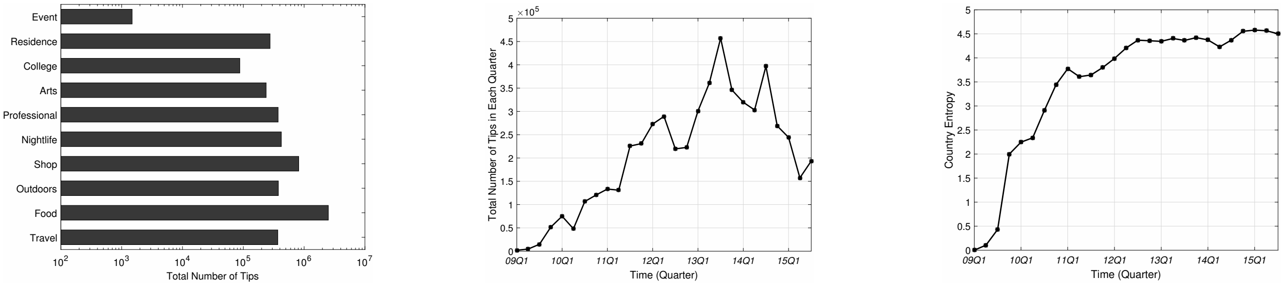
In Fig. 3(a), we demonstrate the venue category distribution of all published tips. We can see that the most popular venue category is "food", which has attracted 45.06% of all tips. The second most popular venue category is "shop" and 14.68% of all tips are posted in this category of venues. For the rest of the categories, none of them has received more than 10% of all tips. Therefore, restaurants are the most attractive venues for tip publishers in Foursquare.

Moreover, we conduct a temporal analysis by counting the number of tips published at different time periods. In Fig. 3(b), we look at the total number of tips from an evolutionary view. We divide the whole span of time into a number of time periods, and each period covers three continuous months,



(a) Profile Photo (b) Gender (c) Home Country

Fig. 2: Group-Based Analysis for the Number of Tips per User



(a) Category Distribution (b) Evolution: # of Tips (c) Evolution: Venue Country Entropy

Fig. 3: Investigating Key Fields of Tips

i.e., a quarter of a year. In total, we have 27 continuous time periods, from the first quarter of 2009 all the way up to the third quarter of 2015. As we crawled the data in Nov. 2015, we do not consider the tips published in the fourth quarter of 2015. We calculate the number of published tips in each quarter, and we can thus analyze the evolution of the total number of published tips on a quarterly base. From the very beginning, the total number of tips increased steadily. After reaching a significant peak in the third quarter of 2013, this number started to decrease. In the middle of 2014, Foursquare started to release the dedicated check-in app Swarm, and in Aug. 2014, the check-in interface was completely removed from Foursquare. This made Foursquare a dedicated tip-sharing app and could at least boost the tip posting for a while. As a result, we can observe another peak in the third quarter of 2014.

Besides evaluating the evolution of the total number of tips, we are also interested in how the tips are distributed among different countries. For a certain quarter, we use k to represent the number of countries which have been visited and use p_i to denote the probability of the visited venues belonging to the $i-th$ country. We introduce the concept of the *venue country entropy* E, using the formula $E = -\sum_{i=1}^{k} p_i \log_2 p_i$. In Fig. 3(c), we show the evolution of the venue country entropy. It increases quickly in the first three years since 2009. This means the country distribution of visited venues has become more and more diverse. Since the second quarter of 2012, this number has become stable, as the country distribution of tip venues have reached a relatively stable status.

TABLE I: Sentiment Analysis - Category

Category	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
Travel	63.37	21.25	15.38	0.74
Food	68.77	17.68	13.55	0.78
College	53.15	33.41	13.44	0.70
Nightlife	66.96	20.82	12.22	0.77
Event	65.66	24.15	10.19	0.78
Shop	65.49	22.19	12.32	0.77
Residence	52.43	35.75	11.82	0.70
Professional	58.29	30.34	11.37	0.73
Outdoors	65.80	23.71	10.48	0.78
Arts	64.27	23.46	12.27	0.76

TABLE II: Sentiment Analysis - Home Country

Country	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
BR	70.18	19.90	9.92	0.80
US	64.37	22.48	13.15	0.76
TR	65.78	25.91	8.31	0.79
ID	60.44	29.92	9.64	0.75

C. *Sentiment Analysis*

In § III-A and § III-B, we count the number of tips published by every user and classify all tips according to the key fields. In this subsection, we dive into the main component of a tip, i.e., the tip text. This component records the tip publisher's

TABLE III: Sentiment Analysis - Gender

Gender	Positive (%)	Neutral (%)	Negative (%)	H_{idx}
Male	65.83	23.23	10.94	0.77
Female	63.73	22.90	13.37	0.75

detailed comment for a venue. To understand the publishers' opinion of the tips, a series of sentiment analysis are expected.

In our study, we calculate a "sentiment score" for each tip. We use a Python-based natural language processing (NLP) library, so-called TextBlob [2], to extract the publisher's attitude from the text. Using this tool, we can obtain a sentiment score for each tip within the range of [-1, 1]. A score of -1 means the tip is surely negative, and a score of 1 means the tip is certainly positive. In our study, if a score is within the range of [-1, -0), we conclude the tip as a negative tip. If a score is within the range of (0, 1], we regard the tip as a positive tip. For the rest of the tips, with a sentiment score of zero for each of them, are defined as neutral tips. As TextBlob can only process English-based tips, we filter out all tips published in other languages. Among all English-based tips, 65.72% of all tips are positive, 21.34% are neutral, and the rest 12.94% are negative. In other words, there are much more positive tips in Foursquare than the sum of the neutral and negative tips.

We use P_{pos} to denote the fraction of positive tips and P_{neu} to represent the fraction of neutral tips. To study the overall sentiment of a set of tips, we introduce a new metric, so-called happiness index (H_{idx}), using the formula $H_{idx} = P_{pos}+P_{neu}/2$. The value of H_{idx} is within the range of [0, 1]. A higher value of H_{idx} indicates higher level of happiness. For all tips in our study, the overall H_{idx} is 0.76.

From Table I, we can see that the tips published in "food", "event" and "outdoors" categories have the highest values of H_{idx}, while the tips published in "college" and "residence" categories have the lowest values. Moreover, we also group the tips according to the tip publishers' home country. According to Table II, tips published by Brazil users have the highest H_{idx}, while the tips published by users from Indonesia have the lowest H_{idx}. Finally, we classify the tips according to the gender of the publisher. From Table III, we find that the male users have a slightly higher H_{idx} than female users.

IV. Related Work

There are a number of literatures related to measurement and analysis of Foursquare [5], [6]. Scellato et al. have explored the socio-spatial properties of Foursquare, by examining the geographic distances among Foursquare friends [6]. Preoţiuc-Pietro et al. have investigated the check-in patterns of Foursquare users, by considering a number of temporal and venue category issues [5]. These studies have contributed a lot to the understanding of the social structure and the check-in feature of Foursquare. However, Foursquare has become a tip-dominated platform, and thus this paper presents a dedicated study on tips in Foursquare.

There are also some work on tips in LBSNs. Vasconcelos et al. [8] have crawled 1.6 million Foursquare venues and extracted 527K user IDs from the obtained venue data. Based on the profiles of these users, they have also studied the distribution of the number of tips per user. However, the set of studied user IDs are obtained in a biased way, for example, all users with zero tips will not be covered. In contrast, we fetch the user IDs through an unbiased sampling, and we have covered more than 10 times more users. Costa et al. have studied the spam tips in Apontador, a popular Brazilian LBSN system [3]. They propose a spam detection mechanism, which is able to identify most of the spam tips. Vasconcelos et al. [7] have explored the prediction of the future popularity of tips.

V. Conclusion and Future Work

In this paper, we present a comprehensive and unbiased analysis of tips in Foursquare. We conduct a data-driven study by crawling and analyzing all tips published by randomly selected 6.52 million users. We count the number of tips published by different users, scan the important fields of tips and finally perform sentiment analysis for all tips. Our study shows a thorough view of Foursquare's tips.

For the next step, we wish to explore the following issues in order to enhance the experience of Foursquare users from different aspects.

First, we aim to look at the spammers, while we have not considered the existence of them in this paper. These accounts might publish some malicious tips to mislead legitimate users. We aim to find out the outstanding features of these users and introduce a machine learning-based approach to discover them. Identifying these accounts can provide a better environment for legitimate users.

Second, we want to investigate the relationship between the tip publishing and the social structure of Foursquare. We will consider a user's followers, followings and links to the external websites such as Facebook and Twitter to see whether these issues will influence the user's tip posting behavior.

Last but not least, we plan to study some applications based on the tips data, such as venue recommendation, online advertisement, social games and cross-OSN data aggregation.

Acknowledgement

This work has been sponsored by Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment.

References

[1] Y. Chen, C. Zhuang, Q. Cao, and P. Hui. Understanding cross-site linking in online social networks. In *Proc. of the 8th ACM Workshop on Social Network Mining and Analysis (SNAKDD)*, 2014.

[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

[3] H. Costa, F. Benevenuto, and L. H. C. Merschmann. Detecting Tip Spam in Location-based Social Networks. In *Proc. of SAC*, 2013.

[4] C. Ding, Y. Chen, and X. Fu. Crowd Crawling: Towards Collaborative Data Collection for Large-scale Online Social Networks. In *Proc. of ACM COSN*, 2013.

[5] D. Preoţiuc-Pietro and T. Cohn. Mining User Behaviours: A Study of Check-in Patterns in Location Based Social Networks. In *Proc. of ACM WebSci*, 2013.

[6] S. Scellato, A. Noulas, and et al. Socio-spatial Properties of Online Location-based Social Networks. In *Proc. of AAAI ICWSM*, 2011.

[7] M. Vasconcelos, J. Almeida, and M. Goncalves. What Makes your Opinion Popular? Predicting the Popularity of Micro-Reviews in Foursquare. In *Proc. of SAC*, 2014.

[8] M. Vasconcelos, S. Ricci, and et al. Tips, Dones and To-Dos: Uncovering User Profiles in FourSquare. In *Proc. of ACM WSDM*, 2012.

[2] **https://textblob.readthedocs.org/**