Finding suitable places for live campaigns using location-based services

Md. Khaledur Rahman
United International University
khaledur@cse.uiu.ac.bd

Muhammad Ali Nayeem
Bangladesh University of Engineering and Technology
ali_nayeem@cse.buet.ac.bd

ABSTRACT

In the recent years, the idea of reaching customers through human experience has triggered a new marketing strategy known as live campaigns. We can expect that live campaigns will become more pervasive and profitable, but not before addressing key business challenges. It can be easily ruined if the campaign agency fails to identify the optimal location and time. In this paper, we address the challenge of finding a suitable location from online location based services for arranging a live campaign according to given schedule among a set of candidate locations. We study the predictive power of various spatio-temporal mining features on the capability of gathering audience through the use of a dataset collected from Foursquare of New York City. Finally, we develop models which will predict the expected audience at a location based on these features. We achieve 50.46% accuracy in individual feature based approach and an accuracy of 72.6% in Support Vector Machine (SVM) regression model.

KEYWORDS

Location-based Service, Recommender System, Social Media, Live Advertisement

ACM Reference format:

Md. Khaledur Rahman and Muhammad Ali Nayeem. 2017. Finding suitable places for live campaigns using location-based services. In *Proceedings of GeoRich'17 , Chicago, IL, USA, May 14, 2017,* 6 pages.
DOI: http://dx.doi.org/10.1145/3080546.3080630

1 INTRODUCTION

The numbers of location based social network (Foursquare, Yelp, etc.) users as well as social network (Facebook, Twitter, etc.) users are increasing day by day. With the growth of users, we see an increasing demand of various location based service applications. Users go to a place, share their location via location based social networks and express their feelings and thoughts (status, likes, comments, etc.) via social networks. These enormous amount of social, temporal and spatial data give us a great opportunity to design application models for numerous fields including urban planning, lifestyle, business, politics, tourism, etc. To get optimal output in any fields we need to take the maximum benefit from these data. Suppose a coffee shop, placed in 100 meters down the road, is not getting enough customers which may force to close the shop within a month. However, mining the location based information of customers; we can discover that if the same coffee shop is placed in the street corner, it may attract lots of customers [13].

In the recent years, Location Based Service (LBS) has become popular due to the enormous use of smartphones. The smartphones, equipped with Global Positioning System (GPS) modules, can process holders' location information. This has brought the flood of LBS applications in the smartphone ecosystem. It has attracted considerable attention due to its potential for a range of highly personalized and context-aware services. A good example can be the smartphone camera with Instagram LBS. If one takes a photo with a smartphone camera, the location where the photo is taken is embedded in the picture automatically, which helps one to remind about the photo. Furthermore, the aggressive expansion of Social Network Services (SNS) has also assisted its growth by constructing connections between location information and social network. Location has become a crucial facet of many online services. People are more willing to share information about their geographic position with friends. As a consequence, service providers have access to a valuable source of data on the geographic location of users as well as online friendship connections among them. The combination of these two factors offers not only a groundbreaking opportunity to understand and exploit the spatial properties of the social networks arising among online users, but also a potential window on real human socio-spatial behavior.

In this paper, we would like introduce a novel problem which is about arranging live campaign through location-based data. To the best of our knowledge, this is the first attempt to study live campaign problem using location based social network. Numerous studies have been conducted to investigate consumers' acceptance of advertising, including mobile and Internet advertising; however, public attitude toward advertising is getting negative [1], because consumers usually find advertising information useless or even annoying. As a result, traditional and previously proven forms of advertisement are no longer as effective as they once were. Now-a-days, human experience is one of the most effective media to reach customers and to utilize it we need to incorporate human interaction into the marketing scenario. This demands a new medium of advertising where producers and consumers both get optimum opportunities to spread information and gather information correspondingly. Live campaign is the new form of advertising growing throughout the world. It offers short business promotional activity in public spaces surprising people through entertainment and refreshment. But a big deal is to find the suitable place and time slot to reach the maximum spectators. So, we effectively incorporate the solution for the problem of finding suitable place with given

GeoRich'17 , Chicago, IL, USA

DOI: http://dx.doi.org/10.1145/3080546.3080630

time slot. Our solution is also applicable to those areas where maximum spectators are needed such as raising social awareness, live billboard, asking vote from young people and live interview for public opinion.

Our overall contributions are summarized as follows:

- We formulate the problem of finding suitable places for live campaigns using location-based data. (Section 2)
- Then, we construct several features from available dataset for training and testing purposes. (Section 3)
- We also show collective impact of several features applying Support Vector Machine (SVM) method. (Section 4)
- Finally, the results from extensive set of experiments show the effectiveness of our approach. (Section 4)

2 PROBLEM FORMULATION

In this section we formalize the problem of finding live campaign hotspot in the context of location-based social networks. Our goal is to identify the best area amongst a candidate set of potential areas for arranging a live campaign on a particular day and time. As a particular place may have very sparse check-ins, we will aggregate the check-ins to all the places within a circular area and find the campaign quality of that area instead of a place. In this case we give an area for the live campaign, the exact place can be selected by the marketing people.

We take a candidate set of areas L in which a commercial enterprise is interested in arranging an advertising campaign for Δt hours of duration on the d^{th} day of the week at h^{th} hour of the day as input. We wish to identify the optimal area $l \in L$, such that the arranged campaign will potentially attract the largest number of audience. An area l is represented by its latitude and longitude coordinates and a radius of 400 meters. The ranking of places according to their advertising quality is then estimated using the features mined by incorporating the characteristics of the area nearby. Our main assumption in the formulation of this task is that the number of empirically observed check-ins by social media users can be used as a proxy for the measure the audience of the live advertisement at a place.

3 SOLUTION OVERVIEW

In this section we present a brief sketch of the steps for solving our problem. At first we introduce the dataset we want to use for our analysis. Then we define the features we mine for ranking geographic areas according to the predicted advertising quality.

3.1 Dataset

The dataset used in this analysis is collected from a widely used location-based social media called Foursquare. We collect this large-scale check-in data[1] available from [2]. This dataset contains check-ins in New York City collected for about 10 month (from 12 April 2012 to 16 February 2013). It contains 227,428 check-ins in New York city for 38,312 unique venues. Each check-in is associated with its time stamp, its GPS coordinates and its semantic meaning (represented by fine-grained venue-categories). This dataset is originally used for studying the spatial-temporal regularity of user activity in LBSNs.

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

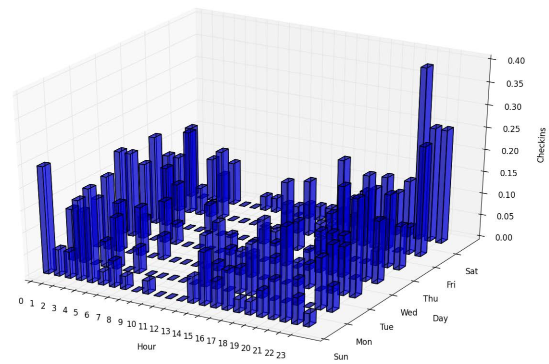

Figure 1: Check-in pattern for Time Square, New York with $\Delta t = 1$

3.2 Prediction features

We now introduce the features that we have mined from the dataset in the city of New York. Each feature returns a numeric score $\hat{\chi}_l$ that corresponds to a quality assessment of the area for live advertising. We employ check-in data analysis for extracting the features. As we are working with user contributed data, a major concern is the effect of noise. So we devise two noise sensitive features which can give low score to the areas affected by noisy check-in pattern.

3.2.1 *Density.*

It measures the number of Foursquare venues in an area $l \in L$. Formally:

$$\hat{\chi}_l = |\{p \in P | dist(p, l) < 400\}|$$

Here, the function *dist* denotes the geographic distance between two places in meters and P denotes the set of venues in New York. We denote the number of neighboring venues within the area l with $N(l)$. Intuitively, a denser area could imply higher likelihood for an opportunistic visit. Thus, we may expect higher audience in that area.

3.2.2 *Neighbors Entropy.*

It measures the spatial heterogeneity of an area $l \in L$. It shows how diverse the area is in terms of types of venues such as food, shop, work, etc. We apply entropy measure from information theory [21] to the frequency of venue types in the area. We denote the number of neighboring venues of type γ with $N_\gamma(l)$. The entropy defines how many bits are required to encode the corresponding vector of type counters $N_\gamma(l) | \gamma \in \Gamma$, where Γ is the set of all types. Formally:

$$\hat{\chi}_l = -\sum_{\gamma \in \Gamma} \frac{N_\gamma(l)}{N(l)} \times log_2 \frac{N_\gamma(l)}{N(l)}$$

The higher the entropy, the more heterogeneous the area is. In general, an area with higher entropy can serve more people.

3.2.3 *Check-in count.*

It measures the estimated check-in count of an area $l \in L$ for Δt hours of duration on the d^{th} day of the week at h^{th} hour of the day. We find the check-in pattern of an area by

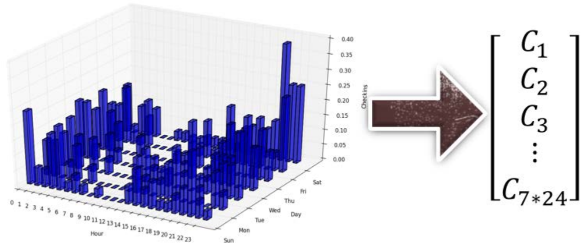

Figure 2: Representing a check-in pattern as a vector

plotting the check-in data with respect to three dimensions: weekday, hours (grouped according to the duration Δt) and normalized check-in count. We show an example check-in pattern in Figure 1. Check-in count represents the normalized check-in count at d^{th} day and h^{th} hour in the check-in pattern.

3.2.4 Check-in consistency. This is our first noise sensitive feature. It measures the degree of changes in check-in pattern of an area with the passage of time. It gives higher score to those areas whose check-in patterns are more stable throughout the year. It may happen that, the check-in count feature gives high score to an area because of a special event (e.g. fair, concert etc.) on some days. In that case check-in count gives us the wrong estimation. In case of live campaign we want to minimize the risk of losing audience. So we prefer areas with stable check-in pattern.

We represent our check-in pattern as a vector of 7×24 cells (7 days each having 24 hours) as in Figure 2. Thus we measure the correlation between two check-in patterns. We calculate the overall check-in pattern considering all check-in data and monthly check-in patterns considering month wise check-in data for each area. Then we measure correlation between the overall check-in pattern and each monthly check-in pattern. Check-in consistency represents the average of all these correlation measures.

3.2.5 Check-in identity. This is our second noise sensitive feature. It makes sure that the check-ins found at the area is not just the noisy effect of the area of interests. For measuring this feature we consider unique check-ins of a place, i.e., we do not consider multiple check-ins of a person for the same location. The reason behind this is that if a person alway create a check-in from his/her home or office that will make noisy effect. We normalize check-in identity of a location in the following way:

$$C_{identity} = \frac{\#uniquecheckins}{\#totalcheckins}$$

3.2.6 Openness. This one is our static feature. It measures the amount of open space within an area. We measure this feature based on the inverse of density. We assume that a place is more open if its density is less. An area with open spaces is suitable for community gathering.

3.2.7 Temporal Signal. We also measure the time of day when people are willing to visit a place. To construct this feature, we divide 24 hours into four time slots, namely, morning (m), noon (n), afternoon (a) and evening (e). The reason behind this feature is that people are interested to visit a certain place in particular time of the day. This feature assigns high value to popular time of the day for a specific location. For example, visitors are generally interested to visit a shopping center in the afternoon or evening and so these two time slots get higher values than morning and noon. We use this feature in SVM regression model. We normalize the temporal signal in the following way:

$$T_{signal} = \sum_{i \in \{m,n,a,e\}} \frac{\#checkins_i}{\#totalcheckins}$$

3.2.8 Distance from nearest subway station. The New York City subway is a popular public transportation systems in New York. This feature measures the distance between the center of the area and the nearest subway station. Mass people have greater access to those areas which are closer to subway stations. We consider spatio-temporal data of the subway stations listed in Table 1. We calculate the distance between a subway station and a venue based on latitude and longitude.

Table 1: List of subway stations in the dataset

Category ID	Category Name
4bf58dd8d48988d1fe931735	Bus Station
4bf58dd8d48988d12b951735	Bus Line
4bf58dd8d48988d1fc931735	Light Rail Station
4bf58dd8d48988d1fd931735	Metro Station
4bf58dd8d48988d129951735	Train Station
4f4531504b9074f6e4fb0102	Platform
4bf58dd8d48988d12a951735	Train
52f2ab2ebcbc57f1066b8b51	Tram Station

4 EXPERIMENTAL RESULTS

In this section, we present experimental study to evaluate the performance of our proposed algorithm. We used the dataset described in section 3.1 to train and test our approach. We divide our experiments into two steps. In the first step, we conduct experiments incorporating each of the features directly and measure performance. In the second step, we train and test across SVM model to see the combined effect of all features. We conduct each experiment 30 times and present the average results. To run all experiments, we use a laptop PC of Core i3 2.2 GHz CPU and 4 GB RAM. We implement all our methods in C++ and R programming language.

We compute the actual rank list R of the candidate areas based on the ground truth i.e. observed check-in and we also compute the predicted rank list R' based on the score predicted by our feature-based model. Given these two ranked lists, we formally define the metrics, we use to assess the quality of predictions achieved by our model.

We measure the fraction of times that the optimal location in the predicted list R' is at the top-X% of the the actual rank list R which represents our ground truth. We refer to this metric as Accuracy@X%. In Figure 3, we report Accuracy based on some individual features for various data size. List size on this Figure represents X% of predicted top ranked locations are on the X% of

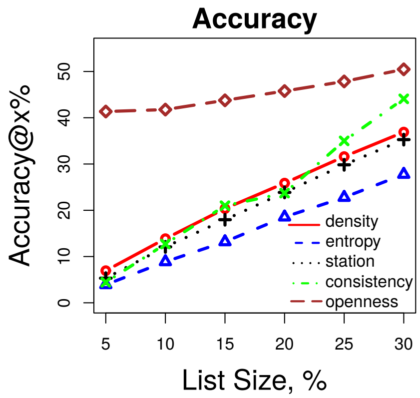

Figure 3: Accuracy@X% for various Sizes of dataset

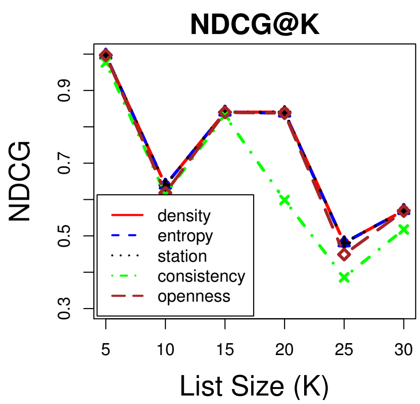

Figure 4: $NDCG@K$ **for various Sizes of dataset**

actual ranked locations. In Figure 3, we also see that the ranking method based on openness feature outperforms other features.

We have not reported some features in Figure 3 because of multi-dimensional or integrity property. For example, the temporal signal feature has four different properties (morning, noon, afternoon, evening) which lacks integrity. However, we use all features together while training SVM model.

We also measure the extent to which the top-k locations in actual rank list R are highly ranked in the predicted list R'. For this we adopt the $NDCG@k$ (Normalized Discounted Cumulative Gain) metric frequently used in the performance evaluation of information retrieval systems [22]. The metric assesses the cumulative gain achieved by placing the most relevant instances in the top-k of the prediction list as formally defined by the Discounted Cumulative Gain measure:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel(l_i)} - 1}{log_2(i+1)}$$

where $rel(l_i)$ is the score relevance of an instance at position i in R'. The result is then normalized by the DCG of the ideal prediction, when the instances are sorted by the relevance with the most relevant in the first position. Thus the resulting scores lie in the range from $0 < NDCG@k \leq 1$. As a relevance score for an

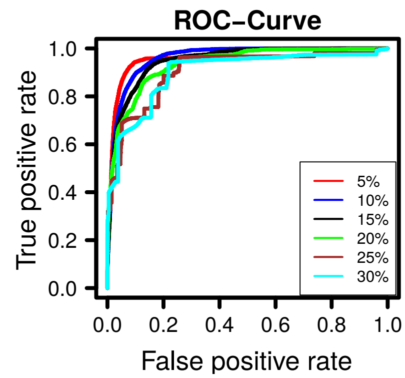

Figure 5: ROC-Curve for various true-valued datasets

instance l_i we will use its relative position in actual ranking R, i.e., $rel(l_i) = \frac{|L|-rank(l_i)+1}{|L|}$. The $rel(l_i)$ score is equal to 1 when the area is ranked first in terms of check-ins and it linearly decreases to 0 as the rank goes down the list. As a baseline for comparison, we use the expected value of $NDCG@k$ for a random ranker which is achieved by randomly permuting the instances in the testing set. In Figure 4, we report $NDCG@k$ based on individual features for various values of k. In this Figure, we see that almost all features show similar NDCG values except consistency feature. When the value of k is 5 then our approach retrieves all relevant locations. For $k = 10$, the value of $NDCG@10$ decreases drastically which means, it retrieves some irrelevant locations. For other values of k, we also see fluctuations in the values of $NDCGs$. We can conclude that all individual features achieve high relevant venues when $k = 5$.

We also design a second kind of experiment where we apply SVM to all of our features altogether. For this experiment, we create a binary classifier using top ranked $X\%$ locations of data. For example, when we take top ranked 5% locations, we assign 1 to all these locations (P) and we assign 0 to rest of the 95% locations (N). Then, we train and test SVM across this dataset using 10-fold cross-validation. We use linear kernel in SVM which represents linear combination of all features. There are four possible scenarios in a binary classifier: when a predicted ranked location is on the actual ranked list we treat this as true positive (TP), but when a predicted place is not on the actual ranked list we treat this as false positive (FP). When a location is not on the predicted ranked locations but present on the actual ranked list, we treat this as true negative (TN); however, when a location is not on the predicted ranked locations and also absent on the actual ranked list, we treat this as false negative (FN). Considering all these scenarios, we estimate accuracy, true positive rate (TPR) and false positive rate (FPR) based on the following formulas:

$$Accuracy = \frac{TP + TN}{38,312}, TPR = \frac{TP}{P}, FPR = \frac{FP}{N}$$

We report Receiver Operating Characteristics (ROC) curves in Figure 5 for various sizes of true valued dataset. When a ROC-curve gets more closer to left upper corner of the graph, it achieves better results [23] and it also gets higher value for area under that ROC-Curve. In this Figure, 5% indicate that there are only 5% true valued data (i.e., assigned to 1) and rest of the data are false valued (i.e., assigned to 0). In the dataset, there are 38,312 distinct locations in total. We assign 5% locations to 1 and other 95% locations to 0 i.e., 1,916 distinct locations are assigned as top 5% ranked based on number of checkin counts and other locations are not on top 5%. We achieve 72.6% and 95.84% accuracy for true valued data size of 30% and 5%, respectively and accuracy for other data size lies between these two values. We get Area Under ROC-curve (AUROC) as 0.90 and 0.96 for true valued data size of 30% and 5%, respectively. From these two analyses, it seems that when we consider top 5% true valued data, SVM performs better. There also exists Precision-Recall curve analysis for binary classification which shows better comparison in case of imbalanced dataset. When a Precision-Recall (PR) curve gets more closer to right upper corner of the graph, it achieves better results and it also gets higher value for Area Under PR (AUPR) curve. To investigate more, we estimate AUPR curve as 0.83 and 0.64 for true valued data size of 30% and 5%, respectively (see Figure 6). In this case, we see that 30% true valued data size performs better. The reason behind is that when we consider only 5% true valued data, there is higher probability to get high TPR value. So, we need to analyze both AUROC and AUPR to reveal the true scenario. Finally, we calculate spearman correlation for 30% and 5% true valued locations against predicted locations and find the results as 0.64 and 0.33, respectively. So, SVM method performs better in the case of 30% true valued data which achieves an accuracy of 72.6%. Finally, we report accuracy as percentage for best individual feature (openness) and SVM regression model in Table 2.

Table 2: Comparison of accuracy for best model of individual feature (openness) and SVM regression model

Size of True Valued Data	Individual Feature (Openness)	SVM Regression
5%	41.34	95.84
10%	41.76	91.67
15%	43.76	87.22
20%	45.77	82.5
25%	47.85	77.6
30%	50.46	72.6

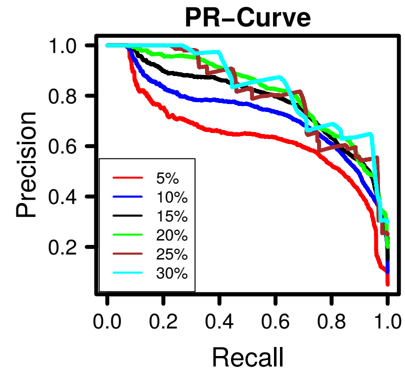

Figure 6: Precision Recall Curve for various true-valued datasets

5 RELATED WORKS

To the best of our knowledge, we are the first to explore live campaign problem using location based social network. However, many related works exist in the field of recommender systems which are discussed as follows.

Georgiev, Noulas and Mascolo propose a recommender system to predict whether a person would attend a future event based on check-in data collected from Foursquare [3]. They mine many popular events along with several social, temporal and spatial features from the data. Finally, they formulate an event prediction task where they try to rank events for each user based on the prediction features. In [3], authors quantify social and spatial factors from users' check-in data successfully. To materialize place type preference of the user into a vector of real numbers they treat user as document and place type as term. To measure the influence of places visited by friends, they design a directional weighted graph (called socio-spatial graph) that combine social and spatial influences. Then, they apply random walks with restart on this graph to calculate probability of participation for a user-event pair. Sklar, Shaw and Hogue propose a real-time event detection engine for Foursquare to measure how unusually busy a place becomes [4]. They focus on probabilistic model that yields a negative binomial distribution over the number of people to check in at any given time. Quercia et al. study event prediction problem where they provide cold-start event recommendations for users whose home location is known [5]. However, in their work, they do not focus on personalization. Three other prominent examples of event recommender systems have been built in the domains of on-going cultural events, scientific and conference talks. Lee exploits trust relations together with explicit user feedback to recommend cultural events [6], while Minkov et al. combine content-based with collaborative filtering approaches to capture user preferences towards latent topics hidden in scientific talk announcements [7]. After that Liao et al. build latent models based on offline spontaneous interactions and co-attendance information to recommend related events in offline ephemeral social networks formed around conference talks [8].

Jensen studies the identification of the appropriate geographic positioning of retail stores [9][10]. Jensen's approach uses a spatial network based formulation of the problem, where nodes are 55 different types of retail stores and weighted signed links are defined to model the attraction and repulsion of entities in the network. Porta et al. propose an approach which is based on the analysis of the spatial distribution of commercial activities [11][12]. In these works, the authors investigate the relationship between street centrality and retail store density in the cities of Bologna and Barcelona respectively, verifying how the former acquires a significant role in the formation of urban structure and land usage. Karamshuk et al. further extend the results of these works by adding to the analysis features mined from the human mobility traces and effectively show that the combination of the geographic and mobility features provides better insights on the quality of an area as a potential spot to open a new retail facility [13]. Specifically, they study the problem of identifying the optimal location for a new retail store placement. They evaluate a diverse set of data mining features, modeling spatial and semantic information about places and patterns of user movements in the surrounding area.

In the recent years, many efforts focus on the temporal dynamics of online social networks and mobile networks. For example, Song et al. study the periodicity of people's activities with respect to their most visited location [14]. Cheng et al. and Ye et al. show that the daily and weekly check-in patterns for specific locations can reveal semantic information (e.g., that two locations are similar), and can facilitate location based search and location recommendation [15][16]. Liang et al. use check-in data and event related tweets to model crowd-based population [18]. They estimate the duration of time a user might spend in a crowd, the number of user leaving a crowd at any time and the number of posts generated from a crowd. Finally, they validate their model by predicting traffic volume for Manhattan and by predicting number of posts for some events. They focus on statistical and probabilistic analysis, and consider both event-driven and location-driven crowds. Hasan and Ukkusuri propose a model to analyze large-scale geo-location data from social media to infer individual activity patterns [19]. Combined with the data from traditional surveys, their model provides an activity generation mechanism which is potentially a useful component of an activity-travel simulator. Lee investigates how microblogging social networks can be used as a reliable information source of emerging events by extracting their spatio-temporal features from the messages to enhance event awareness [20]. In this work, author applies a density-based online clustering method for mining microblogging text streams, in order to obtain temporal and geospatial features of real-world events.

In this paper, we have discussed features that are extracted from analyzing all the business perspectives discussed above. We have also shown the influence of these features on location based social services.

6 CONCLUSION

In this paper, we tackle a novel problem, namely, finding optimal places for live campaigns in the context of location-based social networks. We have analyzed the aggregated effect of check-in from the leading location-based service, Foursquare. Further, we have introduced some static geographic factors that may influence the human mobility pattern. We conduct two kinds of experiments, namely, direct individual feature based experiment and SVM regression using linear kernel with 10-fold cross-validation. For 30% true valued data, we have achieved a maximum of 50.46% accuracy in individual feature based approach and an accuracy of 72.6% running SVM regression model using all features altogether.

We will further extend the work using textual information from twitter [17] and wide range of datasets. This work has a very important business implication in the field of mobile computing. So, we also have a plan to develop a recommender system for Mobile Application.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Mohammed Eunus Ali for advising and encouraging us to work in this field.

REFERENCES

[1] E. J. Zanot:*Public Attitudes Toward Advertising: The American Experience.* International Journal of Advertising, vol. 3, pages 3-15, 1984.
[2] D. Yang, D. Zhang, V. W. Zheng, Z. Yu: *Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs.* IEEE Trans. on Systems, Man, and Cybernetics: Systems, (TSMC), 45(1), 129-142, 2015.
[3] P. Georgiev, A. Noulas, C. Mascolo: *The Call of the Crowd: Event Participation in Location-Based Social Services.* In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014). Ann Arbour, Michigan, USA, June 2014.
[4] M. Sklar, B. Shaw, and A. Hogue: *Recommending interesting events in real-time with foursquare check-ins.* In Proceedings of the 6th ACM Conference on Recommender Systems, RecSys, Dublin, Ireland, September, 2012.
[5] D. Quercia, N. Lathia, F. Calabrese, G. D. Lorenzo and J. Crowcroft: *Recommending social events from mobile phone location data.* In Proceedings of the 10th IEEE International Conference on Data Mining, Sydney, Australia, December 2010.
[6] D. H. Lee: *Pittcult: trust-based cultural event recommender.* In Proceedings of the 2nd ACM Conference on Recommender Systems, RecSys, Lausanne, Switzerland, October, 2008.
[7] E. Minkov, B. Charrow, J. Ledlie, S. Teller and T. Jaakkola: *Collaborative future event recommendation.* In Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM, Toronto, Ontario, Canada, October, 2010.
[8] G. Liao, Y. Zhao, S. Xie and P. S. Yu: *Latent networks fusion based model for event recommendation in offline ephemeral social networks.* In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM, San Francisco, CA, USA, October, 2013.
[9] Pablo Jensen: *Network-based predictions of retail store commercial categories and optimal locations.* Physical Review E, 74:035101, September, 2006.
[10] Pablo Jensen: *Analyzing the localization of retail stores with complex systems tools.* In Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII, IDA, pages 10âĂŞ20, Berlin, Heidelberg, 2009.
[11] S. Porta, V. Latora, F. Wang, S. Rueda, E. Strano, S. Scellato, A. Cardillo, E. Belli, F. CÃărdenas, B. Cormenzana and L. Latora: *Street centrality and location of economic activities in barcelona.* Urban Studies, vol. 49, no. 7, pages 1471-1488, November, 2011.
[12] S. Porta, V. Latora, F. Wang, E. Strano, A. Cardillo, S. Scellato, V. Iacoviello and R. Messora: *Street centrality and densities of retail and services in bologna, italy.* Environment and Planning B: Planning and Design, vol. 36, no. 3, pages 450âĂŞ465, December, 2009.
[13] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia and C. Mascolo: *Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement.* In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, Chicago, IL, USA, August, 2013.
[14] C. Song, Z. Qu, N. Blumm and A.-L. Barabasi: *Limits of predictability in human mobility.* Science, vol. 327, issue 5968, pages 1018âĂŞ1021, February, 2010.
[15] Z. Cheng, J. Caverlee, K. Kamath and K. Lee: *Toward Traffic-Driven Location-Based Web Search.* In Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM, Glasgow, United Kingdom, October, 2011.
[16] M. Ye, D. Shou, W. C. Lee, P. Yin and K. Janowicz: *On the semantic annotation of places in location-based social networks.* In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August, 2011.
[17] Z. Cheng, J. Caverlee, and K. Lee: *Exploring millions of footprints in location sharing services.* In Proceedings of the 5th International Conference on Weblogs and Social Media, ICWSM, Barcelona, Catalonia, Spain, July, 2011.
[18] Y. Liang, J. Caverlee, Z. Cheng and K. Y. Kamath: *How Big is the Crowd? Event and Location Based Population Modeling in Social Media.* In Proceedings of the 24th ACM Conference on Hypertext and Social Media (part of ECRC), HT, Paris, France, May, 2013.
[19] S. Hasan and S. Ukkusuri: *Urban activity pattern classification using topic models from online geo-location data.* Transportation Research Part C, vol. 44, pages 363-381, April, 2014.
[20] C. -H. Lee: *Mining spatio-temporal information on microblogging streams using a density-based online clustering method.* Expert Systems with Applications, Elsevier, vol. 39, issue 10, pages 9623-9641, February, 2012.
[21] T. M. Cover, J. A. Thomas and J. Wiley: *Elements of information theory.* Wiley Online Library, vol. 6, 1991.
[22] K. Jarvelin and J. Kekalainen: *Cumulated gain-based evaluation of ir techniques.* ACM Transactions on Information Systems (TOIS), vol. 20, issue 4, pages 422-446, 2002.
[23] T. Fawcett: *An introduction to ROC analysis.* Pattern recognition letters, vol. 27, issue 8, pages 861-874, 2006.