# Business Location Recommendation Using Check-in Data

Thesis submitted in partial fulfilment of the requirement for the degree of

Bachelor of Computer Science and Engineering

Under the Supervision of

Dr. Jia Uddin

By


Nasim Uddin Ahmed (13101285)

Shafayet Mahmud (14201009)

Md. Tawabul Islam (12301043)

Shadman Shoumik (13101276)

Muhaimin Habib(16101322)

**School of Engineering and Computer Science**


August 2017


BRAC University, Dhaka, Bangladesh

# Declaration

We hereby declare that this thesis is based on results obtained from our own work. Due acknowledgement has been made in the text to all other material used. This thesis, neither in whole nor in part, has been previously submitted to any other University or Institute for the award of any degree or diploma.

**Signature of Supervisor**

Dr. Jia Uddin

**Signature of the Authors**

Nasim Uddin Ahmed (13101285)

Shafayet Mahmud (14201009)

Shadman Shoumik(13101276)

Muhaimin Habib (16101322)

Tawabul Islam (12301043)

# Acknowledgement

First and foremost, we would like to thank Almighty Allah for enabling us to initiate the research, to put our best efforts and successfully conclude it.

Secondly, we submit our heartiest gratitude to our respected Supervisor Dr. Jia Uddin for his contribution, guidance and support in conducting the research and preparation of the report. Every last involvement of his, starting from instilling in us the deadliest of fears to the kindest words of inspiration has permitted us to effectively complete the paper. We are truly grateful to him.

We revere the patronage and moral support extended with love, by our parents as well as our friends. They helped us with their direct or indirect suggestions which aided in achieving our goal. We would also like to acknowledge the assistance we received from numerous resources over the Internet especially from fellow researchers' work.

Last but not the least, we thank BRAC University for providing us the opportunity of conducting this research and for giving us the chance to complete our Bachelor degree.

# Table of Contents

# List of Figures

# List of Abbreviations

| | |
|---|---|
| LBSN | Location Based Social Network |
| CNF | Collaborative Neighborhood Filtering |
| KNN | K Nearest Neighbor |

# Abstract

In this era of Social Networks, large amount of data is generated from the users of social media and mobile applications on day to day basis. This data can be useful for the companies as they provide insight into the location oriented decisions of the businesses and on user behavior patterns in their regular activities. In this thesis work we are interested in the LBSN (Location Based Social network) data which is generated when the Users of social network Interact in the Online Social Networking Platforms and mobile applications by sharing their location data through "check ins" in the various Business locations. This spatial aspect of the LSBN data almost represent an online model of the physical world which can be analyzed to find key insights regarding the business locations. We have used the Geographical and Social distances to partition the city into neighborhoods as place for a new business opportunity. In technique we have used the collaborative neighborhood filtering based on similarity of neighborhoods by establishing correlation between business venues and check in patterns. We have used the New York foursquare data for our experimentation, this experimentation shows promising results for prediction of future business location.

# CHAPTER 01

# Introduction

Modern world is witnessing the rapid growth in business sector. As a result of that, finding an appropriate place to start a business has become a major issue. Factors like population, distance, personal choices are massively involved in the whole process. Check in data from the social media is a workable source from which we can propose a feasible location to start a business. Our goal is to use the check in data of the general people from social media to observe and analyze the common check ins of a particular person and find a relation to predict the best possible business place. Popular Online social networks have incorporated location based service, e.g. Facebook place, Foursquare/ Swarm, Tinder, Sports tracker etc. Foursquare has gained popularity as Location based social networking platform, which was launched in March 2009[1].

Decisions on starting up a new business seems crucial nowadays as the market is getting more competitive day by day. Investors always looks for business opportunities in a new area where the probability of success is high. Emergence of LBSN data and machine learning algorithms made this process simpler for investor as it enables them to make well informed decisions on opening a new business in a hotspot location. Finding the best location is a classical problem with numerous solutions being proposed including a geo-business prediction technique which is an adaption of the k nearest neighbor (KNN) spatial interpolation technique [3]. It implemented training data based on spatial and categorical criteria to make spatially related predications. Spatial regression models are investigated through comparative study for predicting business behavior [4]. Geographically weighted regression compared to Durbin, Durbin error, spatial lag, spatial error and spatial lag X regression models were implemented for business behavior prediction.

## 1.1 Motivation

Social media is significant in the context that it holds huge pool of information, containing geo-located data of users when they visit a location and check-in using their smartphone and location based social network [2]. For instance, users post about recent news in twitter, upload photos in Instagram, and give check in to their favorite places in Foursquare.

In September 2011, Foursquare had more than 10 million registered users who interacted with the platform by giving a total of 1 million check-ins. By April 2012, the number of check-ins doubled [3]. This large amounts of LBSN data served as a motivation for implementing machine learning algorithm and investigate business behavior through online check-ins of foursquare users.

## 1.2 Contribution Summary

Ever since then, a good number of models have been proposed by numerous researchers trying to recommend a good business location leveraging the LBSN data. However, most of them do not identify the set of related categories for each line of business. Related categories of business are of prime importance in order to find a suitable business venue. For e.g. the probability of existence of coffee shops near restaurants are high. This gave us the incentive to conduct a research on finding the set of correlated categories of business in a specific neighborhood. Our aim was to develop two models, Bayes and CNF and evaluate the performance of these individual algorithm.

## 1.3 Thesis Outline

- Chapter 2 provides the Background study in details including the algorithms and techniques used in the system
- Chapter 3 discusses the Literature Review of related works in this field
- Chapter 4 describes the proposed model along with implementation details
- Chapter 5 presents the results of the experiment along with performance analysis and comparisons
- Chapter 6 concludes the paper specifying the limitations and challenges while planning future development of the project

## 1.4 Methodology

### 1.4.1 Data collection

Our dataset contains check-ins in NYC collected for about 10 month (from 12 April 2012 to 16 February 2013). It contains 227,428 check-ins in New York city. 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[14].

### 1.4.2 Tools used

We used Python Language and Pycharm IDE for data analysis.

# CHAPTER 02

# Background Analysis and Literature Review

## 2.1 Business Location Recommendation

Due to the increment of the various kind of gadgets and online services in the modern world, users can benefit by getting wide range of access to location-based services from anywhere via mobile devices. Adding to that, users can share location-related information with each other by using LBSNs[10].

According to Wikipedia, LBSNs are a type of social networking in which geographic services and capabilities such as geo-coding and geo-tagging are used to enable additional social dynamics. As shown in Figure 1, users can visit locations in the real world and can provide geo-tagged information content (i.e.: comments, photos, videos). There are three layers: User layer, Location layer, Content layer.

- User Layer: We can calculate the similarity among users based on social network that exists in the User layer.
- Location Layer: We can compute geographical distance between each pair of places in the Location layer.
- Content Layer: From the content layer, we can compute the similarity among the information (i.e.: videos, tags etc.) based on their metadata.

Acquiring these contextual information, LBSN can improve the quality of services.

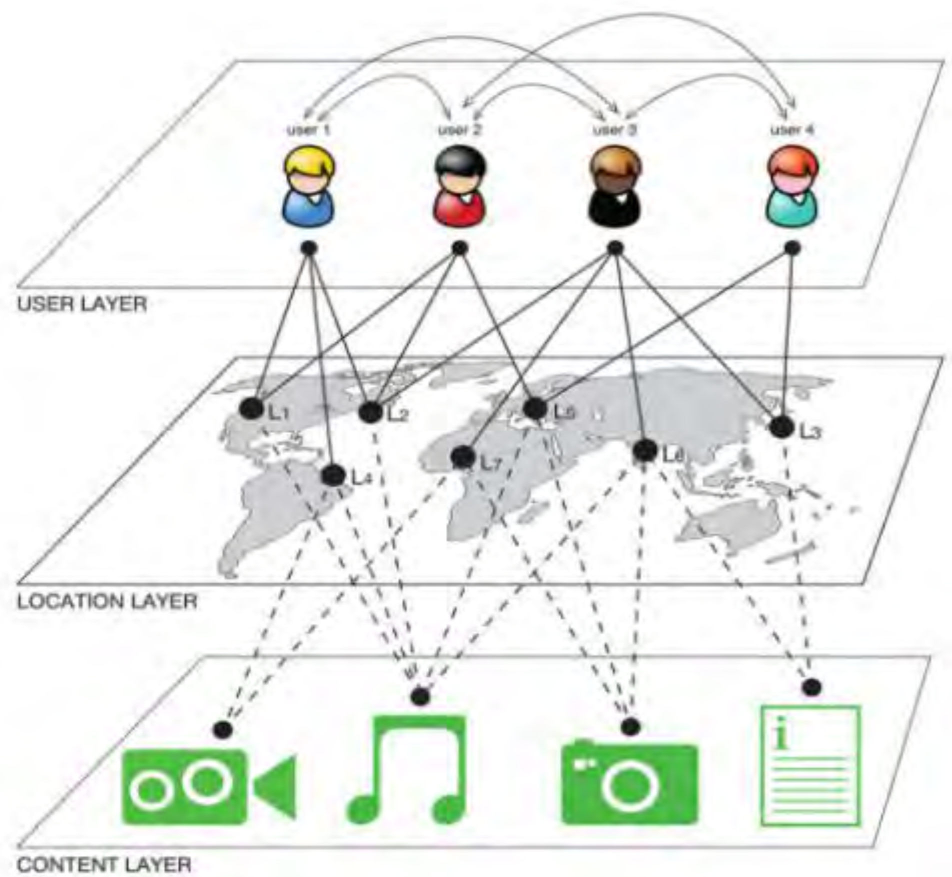


Fig. 2.1.1. Representation of user, location, and contents

**2.1.1 Generic Recommendations:** Generic Recommendations compute the same recommendation list (location, activity, events etc.) for all users, regardless the personalized preferences of each individual user.

**2.1.2 Personalized Recommendations:** The personalized recommender systems rely on past "check-in" history of users. Then, they correlate them with other users that have similar preferences and suggest to them new locations, activities and events.

## 2.2 K-means Clustering

K means clustering organizes the objects of a set into several groups or clusters [5]. Given a data set D, of n objects in Euclidean space. The objects in D, are partitioned into k clusters, $C_1$ ,……., Ck which is similar to $C_i \subset D$ and $C_i \cap C_j = \emptyset$. An objective function ensures that objects within a cluster are similar to another and dissimilar to objects in other clusters.

K means clustering is a centroid based partitioning technique which uses the centroid of the cluster, $C_i$. The centre point of the cluster is known as centroid. The difference between an object $p \in C_i$ and $c_i$, is measured by dist(p, $c_i$) where dis(x, y) is the Euclidian distance between two points x and y. The quality of the cluster $C_i$ can be measured by the within cluster variation, which is the sum of squared error between all objects in $C_i$ and the centroid $c_i$,

$$E = \sum_{i=1}^{k} \sum_{p \in \mathrm{Ci}} dist(p, ci)^2 \tag{1}$$

Here, E - the sum of the squared for all objects in the data set

p - the point in space representing a given object

$c_i$ - the centroid of the cluster $C_i$

Therefore, in equation, for each object in each cluster, the distance from the object to the cluster is squared and the distances are summed.

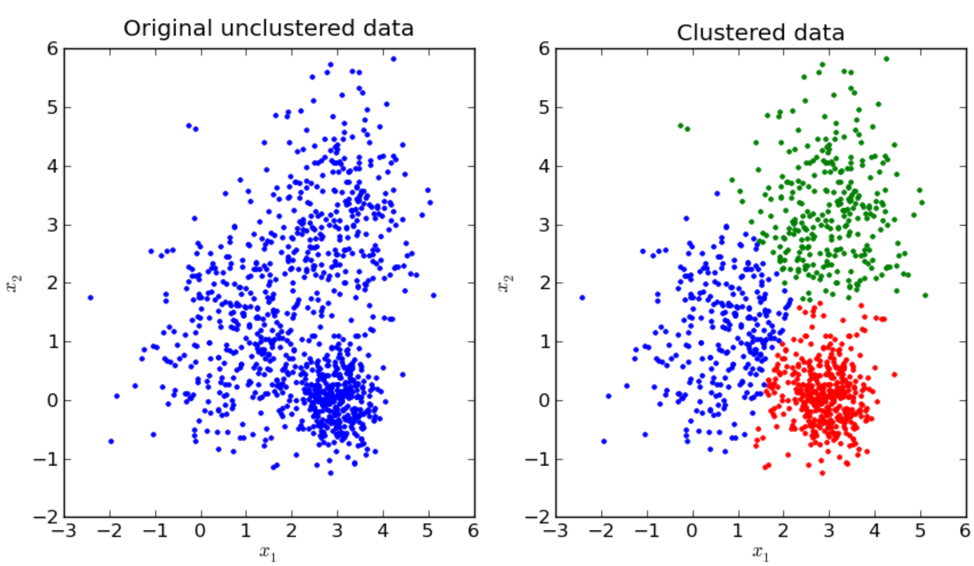


Fig. 2.2.1. Clustering [6]

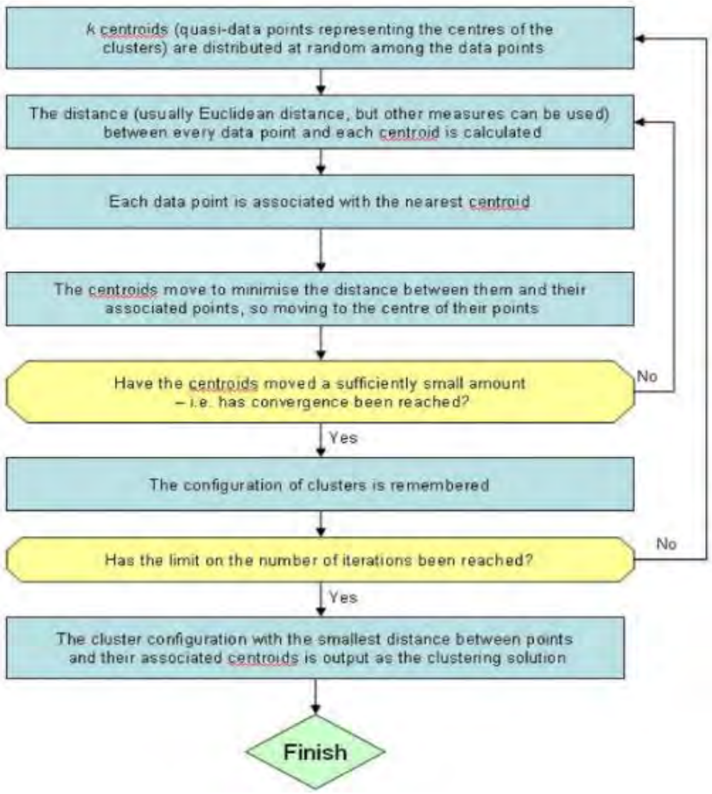


Fig. 2.2.2. Flowchart of K Means Clustering [7]

### 2.2.1 Steps

The following steps are used for K-means Clustering [5]:

Inputs taken as:

- K: the number of clusters.
- D: a data containing n objects.

Output should be: A set of clusters.

Work process:

(1) Arbitrarily choose k objects from D as the initial cluster centres.
(2) Repeat step 1.
(3) (re) assign each object to the cluster to which the object is the most similar, based on the mean value of objects in the cluster.
(4) update the cluster means, by calculating the mean value of the objects for each Cluster until no change

## 2.3 Bayesian Inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference is an important technique in statistics, and especially in mathematical statistics. Bayesian updating is particularly important in the dynamic analysis of a sequence of data. Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, sport, and law. In the philosophy of decision theory, Bayesian inference is closely related to subjective probability, often called "Bayesian probability"[9].

$$P(Cj = 1|Cj) = \frac{P(Cj|Cj = 1)P(Cj = 1)}{\Sigma_{l=0}^{1} \mathrm{P(Cj|Cj = l)\,P(Cj = l)}} \qquad (2)$$

## 2.4 Collaborative Neighborhood Filtering (CNF)

The growth of the Internet has made it much more difficult to effectively extract useful information from all the available online information. The overwhelming amount of data necessitates mechanisms for efficient information filtering. Collaborative filtering is one of the techniques used for dealing with this problem[11].

The motivation for collaborative filtering comes from the idea that people often get the best recommendations from someone with tastes similar to themselves. Collaborative filtering encompasses techniques for matching people with similar interests and making recommendations on this basis.

Collaborative filtering algorithms often require (1) users' active participation, (2) an easy way to represent users' interests, and (3) algorithms that are able to match people with similar interests.

Typically, the workflow of a collaborative filtering system is:

1. A user expresses his or her preferences by rating items (e.g. books, movies or CDs) of the system. These ratings can be viewed as an approximate representation of the user's interest in the corresponding domain.

2. The system matches this user's ratings against other users' and finds the people with most "similar" tastes.
3. With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user (presumably the absence of rating is often considered as the unfamiliarity of an item).

A key problem of collaborative filtering is how to combine and weight the preferences of user neighbors. Sometimes, users can immediately rate the recommended items. As a result, the system gains an increasingly accurate representation of user preferences over time.

## 2.5 Previous Works and Technical Overview

By using check in data, it is possible to generate a good solution for someone who is willing to start a new business. There has been some works which are done before with this.

Wang, Terrovitis, Mamoulis[12] used Bookmark-coloring algorithm, check-in data, geo filtering to recommend business friendly location.

Noulas, Scellato, Lathia, Mascolo[13] also worked on venue recommendation. Their approach includes visiting popular venues, attending venues by category, following friends, staying close to home, like-mindedness and similarity.

If we use K-means clustering with Bayesian Inference and K-means clustering with Collaborative Neighborhood Filtering, it is possible to get more accurate results which will be able to recommend the best possible location.

### 2.5.1 Defining Problem

At first area is defined as a neighborhood in a specific city where different types and number of venues exist. This can be found in different ways, such concept is given by Cranshaw[8].

Then the set of users is recommended for which a venue in a specific business category is estimated to be an appropriate business location. This input is taken in the proposed solution in the following way:

For a specific type of business, a set of neighborhoods is recommended, where the existence of highly probable business location is found. This will also ask for a specific number of neighborhood areas.

### 2.5.2 Neighborhood searching

The locations and their check-ins are used into the neighborhoods. Then the distance is approximated as the Euclidian distance between coordinates of the venues. After setting the definition, the data is clustered. For this purpose, K-means with k number of clusters is used. α is the distance weight parameter[8].

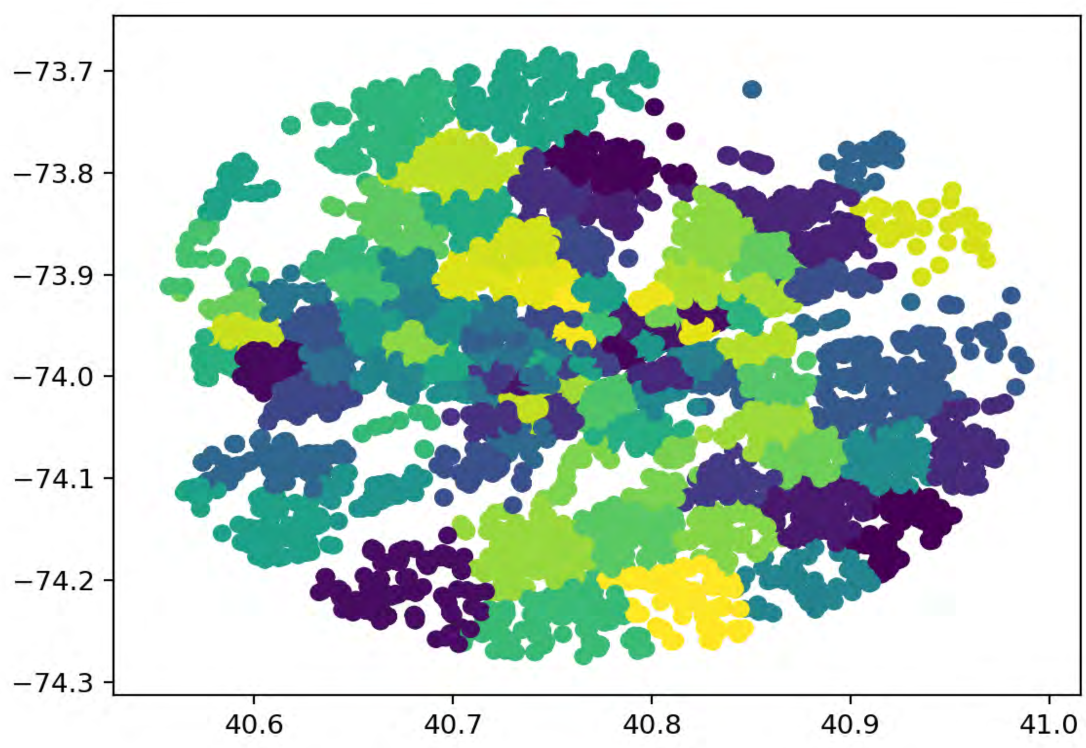


Fig. 2.5.1 : K means clustering (k=100)

### 2.5.3 Implementation of Bayesian Inference

There are different categories of business which are correlated in neighbourhoods. As mentioned in the paper[8] there are few drawbacks of the approach they have used. The PNS algorithm proposed in the paper is quite straightforward, hence it is hard to find something to improve. The drawback of this algorithm is the fact that it is based on an assumptions that every category is independent of each other when the category Cj is considered, which is not always true in reality.

If we rewrite this algorithm, I will avoid using 0 and 1 as the value in the formula. It may lead to many 0 and 1 value in the final posterior probability (because of multiplication), therefore we cannot consider which neighborhood is better when both of them have probability for a given category = 1.

### 2.5.4 Implementation of Collaborative Neighborhood Filtering

We tried to solve the same problem using the Collaborative Filtering approaches. The main idea in this approach is to find similar entities (in this case, neighborhood) to the one which was asked and find the common things as a recommendation.

$$J\left(N_i, N_m. C_j\right) = \frac{N_i \cap N_m}{N_i \cup N_m} \tag{3}$$

$$= \frac{\sum_{n \in C_j'} \min\left\{ N_{CAT(i,n)}, N_{CAT(m,n)}\right\}}{\sum_{n \in C_j'} \max\left\{ N_{CAT(i,n)}, N_{CAT(m,n)}\right\}}$$

$$where\ C_j' = (1,2, \dots, j-1, j+1, \dots, |C|-1, |C|)$$

$$L(N_i, C_j = \frac{\sum_{N_m \in SimN_i} J\left(N_i, N_m, C_j\right). GA_{MAT(m,j)}}{\sum_{N_m \in SimN_i} J\left(N_i, N_m, C_j\right)} \tag{4}$$

# CHAPTER 03

# Proposed Model

## 3.1 System Design

- *Block 1:* Dataset from foursquare is given as input to the system.
    - The dataset is clustered into different neighbourhoods.
- *Block 2:* The different venues from the dataset are mapped to cluster
- *Block 3:* Category- neighbourhood existence matrix is built from the clusters
    - If venue of a category exists in Neighbourhood, 1 is assigned as a value and 0 for otherwise.
- *Block 4:* Prior probability when $(P(C_j = 1))$ and $(P(C_m = 0))$ is calculated without any prior knowledge of the neighbourhood.
- *Block 5:* Posterior probability of success of each category in every neighbourhood is calculated.
- *Block 6:* Posterior probability matrix is built for recommendation to the user.

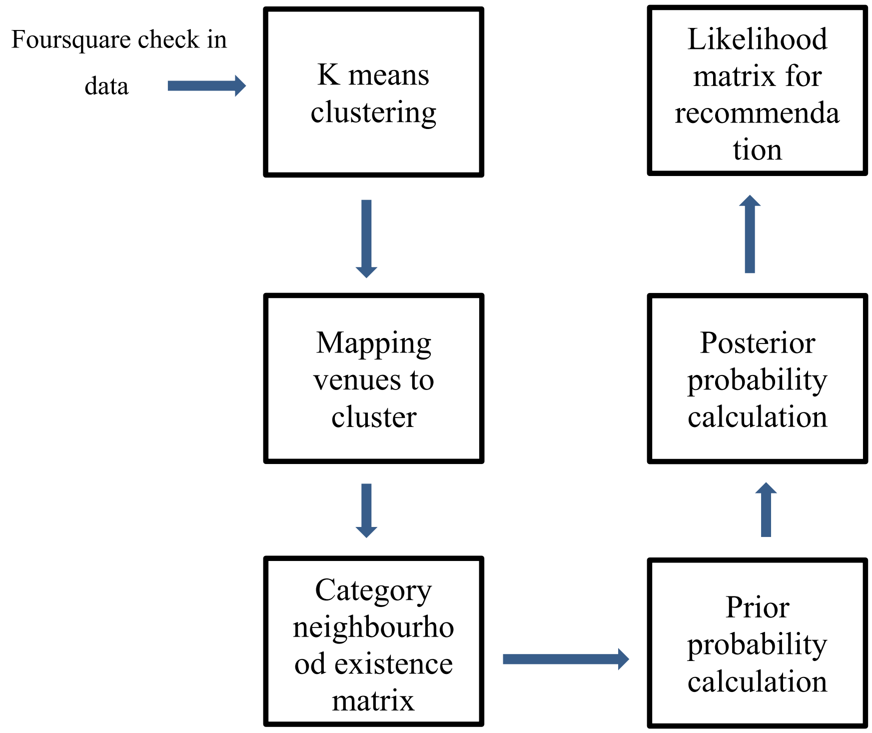


Fig. 3.1.1. Block diagram of the proposed model

K means clustering was initially applied to the dataset for clustering the business venues across the city based on geographical distance and social distance. Geographical distance is the Euclidian distance between the coordinates of venues. This distance is measured along the surface of the Earth. In addition, social distance was calculated using the Jaccard distance between the users of venues which significantly measures the distance between any two users at different venues.

In this paper, we proposed the Bayesian inference which is modified to an algorithm called Probabilistic Neighborhood Selection(PNS) [8]. We found correlation of categories $C_j$ and $C_m$ by using the posterior probability of the different categories. However, the drawbacks of this algorithm is the fact that it is based on an assumption that every category is independent of each other when category $C_j$ is considered. We calculated the prior probability using the following equation:

$$P(C_l, C_m|C_j) = P(C_l|C_j).P(C_m|C_j) \qquad (5)$$

Then we calculated the posterior probability to find the most probable neighbourhood.

$$P(Cj = 1|C1, C2, \ldots, Cj - 1, Cj + 1, \ldots, CJ \quad (6)$$

The posterior probability matrix is built and saved for efficient recommendation.

## 3.2 Workflow


Fig. 3.2.1. Workflow

# CHAPTER 04

# Experimental Results

We have used the location based data from foursquare check-in. For performance evaluation we have implemented the two algorithm for the recommendation system: Bayesian Inference(BAYES) and Collaborative Neighborhood Filtering(CNF) to evaluate performance of these algorithms, we define two metrics:

In this case, we feed the system with a category Cj and the test neighborhoods and retrieve n neighborhoods which the system recommends. We have used four different values n =1, 3,5,10 for our testing purposes which encompasses the practical scenarios where and investor is not expected to request more than 10 neighborhoods for investment recommendation.

First we import the necessary libraries and load the dataset for testing. Then we load the cleaned data to get info about venues. Afterwards we get data frame of unique venues from check-ins. Then we load the cluster member matrix.

We define a method to check if a category C_j exists in a neighborhood N_i (1 for exist, 0 for not):

The accuracy for each category C_j is defined as the percentage of neighborhoods in the recommendation list N_rec that actually have venues in this category. Now we choose a set of 25 random categories from top 150 popular categories (at least 0.1% of venues in this category).

We retrieve the top-n neighbourhood for category C_j with Bayes and CNF algorithm.

When n=1, we get:

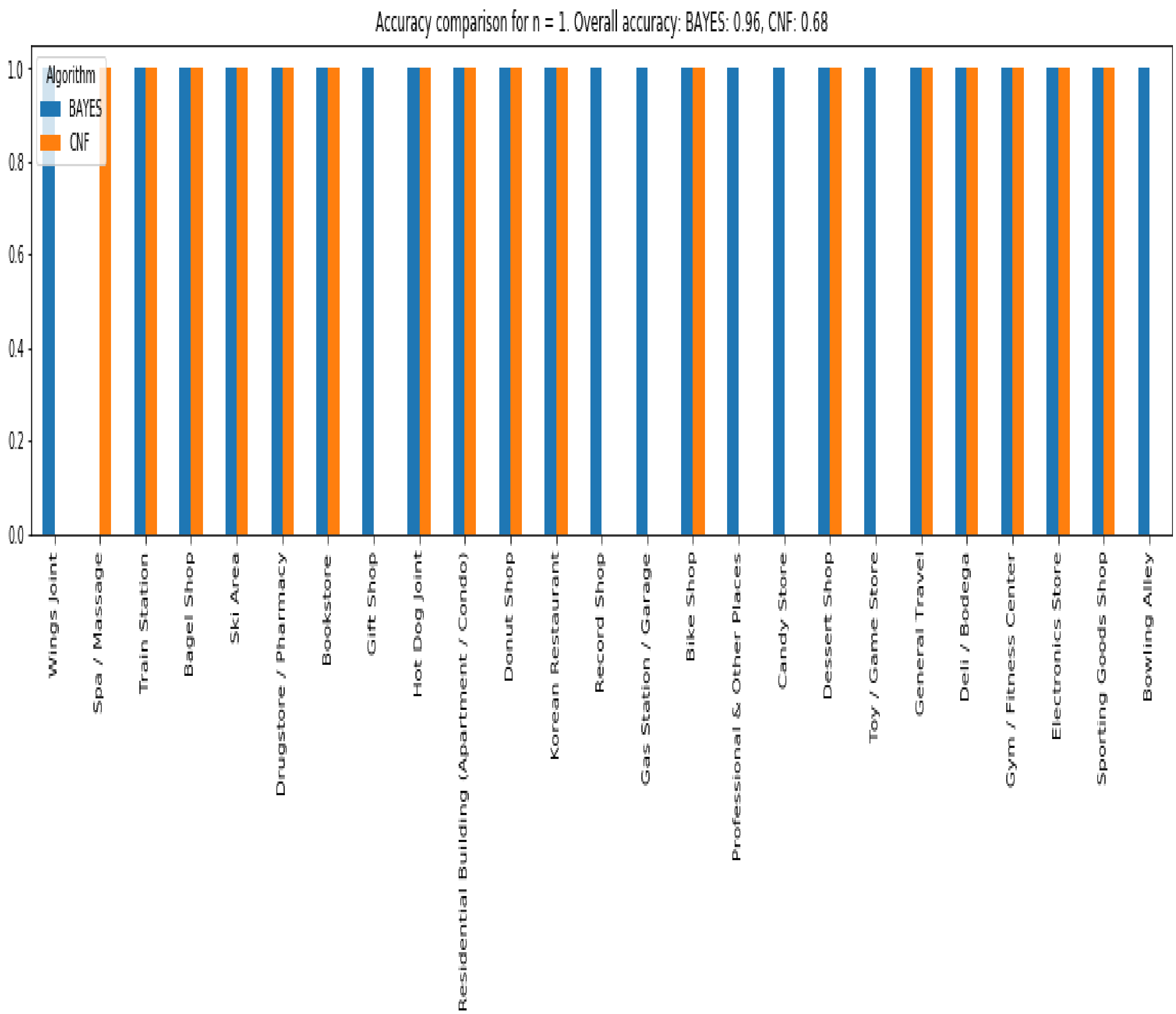


Fig. 4.1 (a). Bayesian accuracy 96%, CNF accuracy 68%

When n=3, we get:

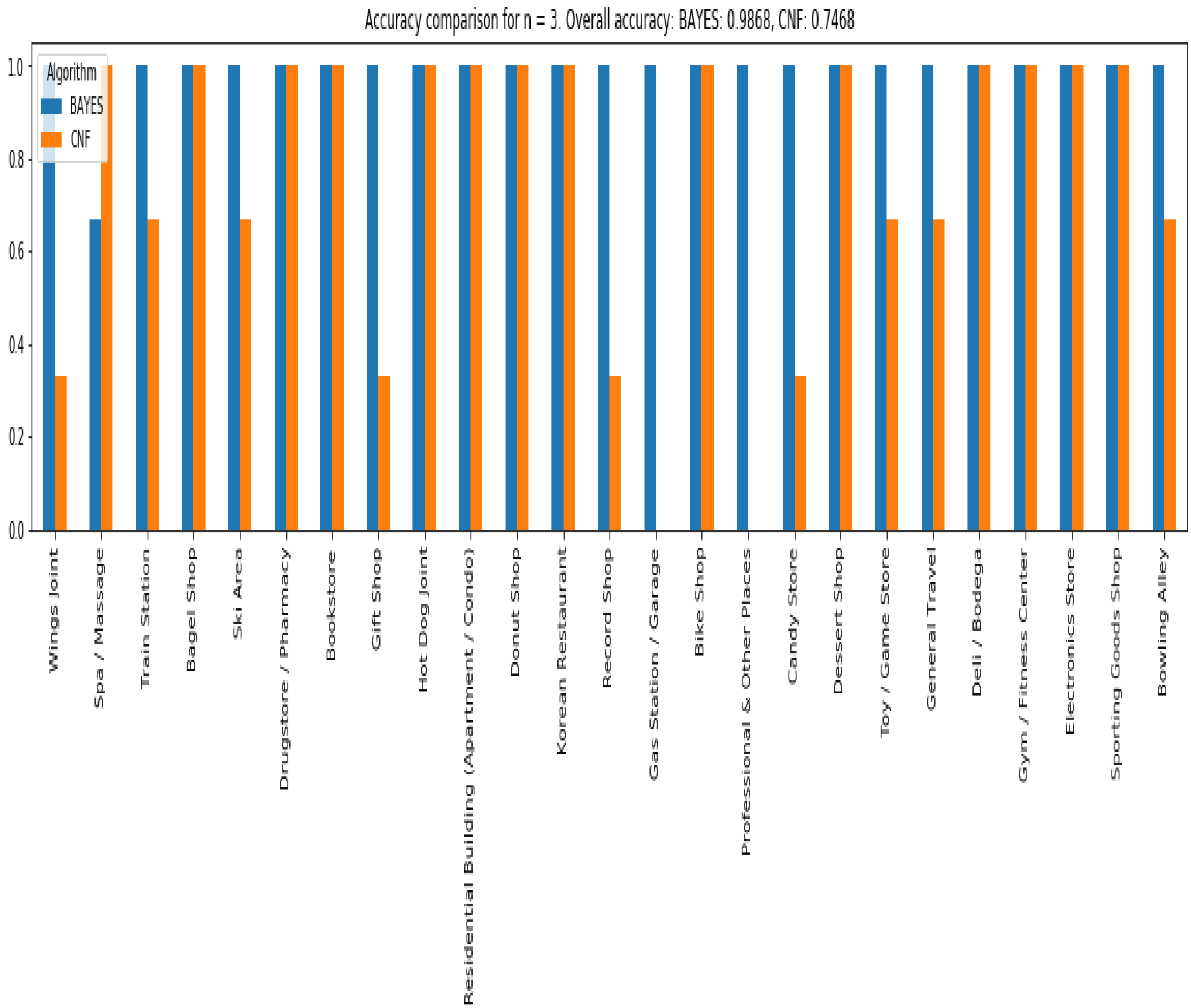


Fig. 4.1 (b). Bayesian accuracy 98%, CNF accuracy 74%

When n=5, we get:

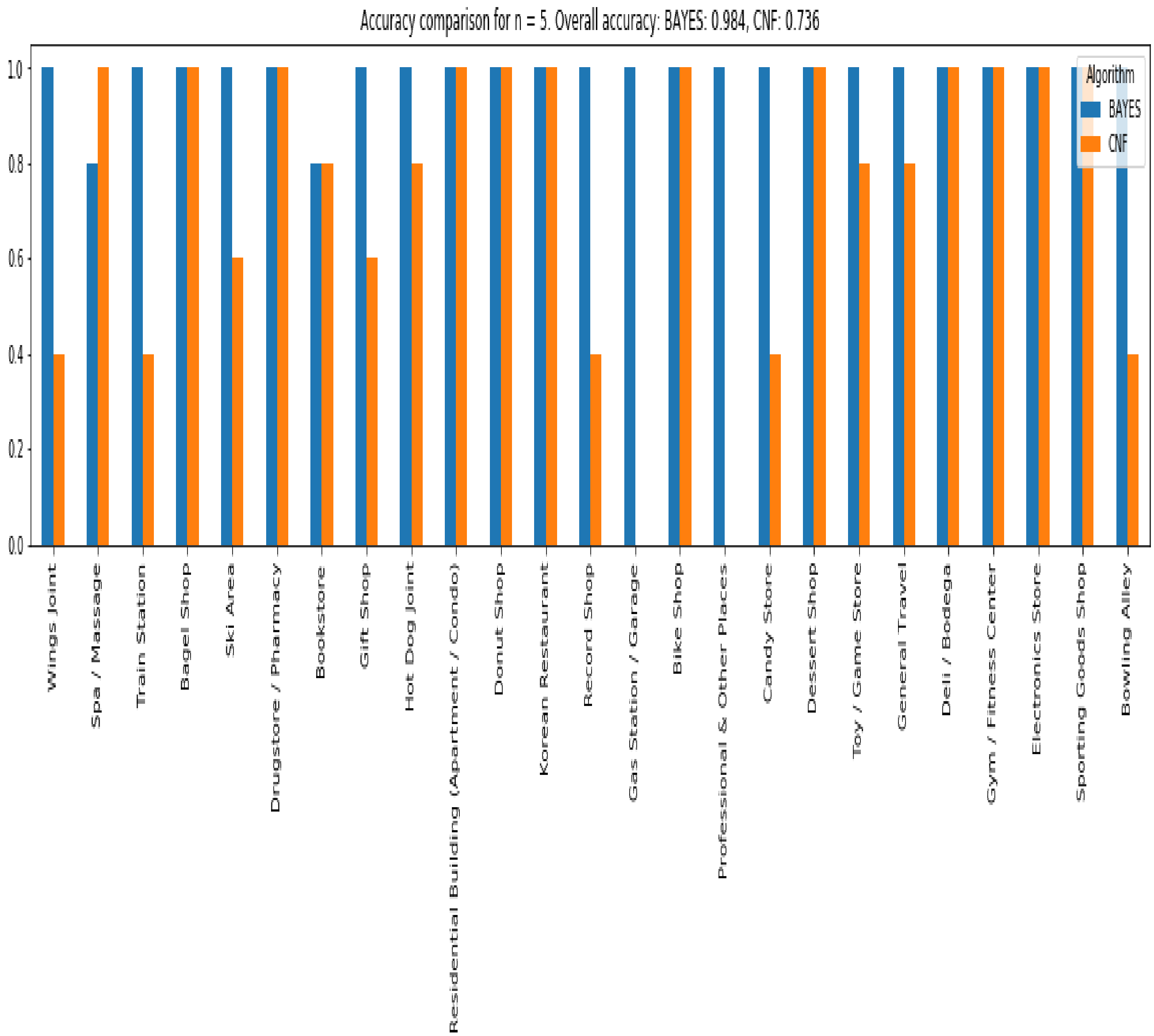


Fig. 4.1 (c) Bayesian accuracy 98%, CNF accuracy 73%

When n=10, we get:

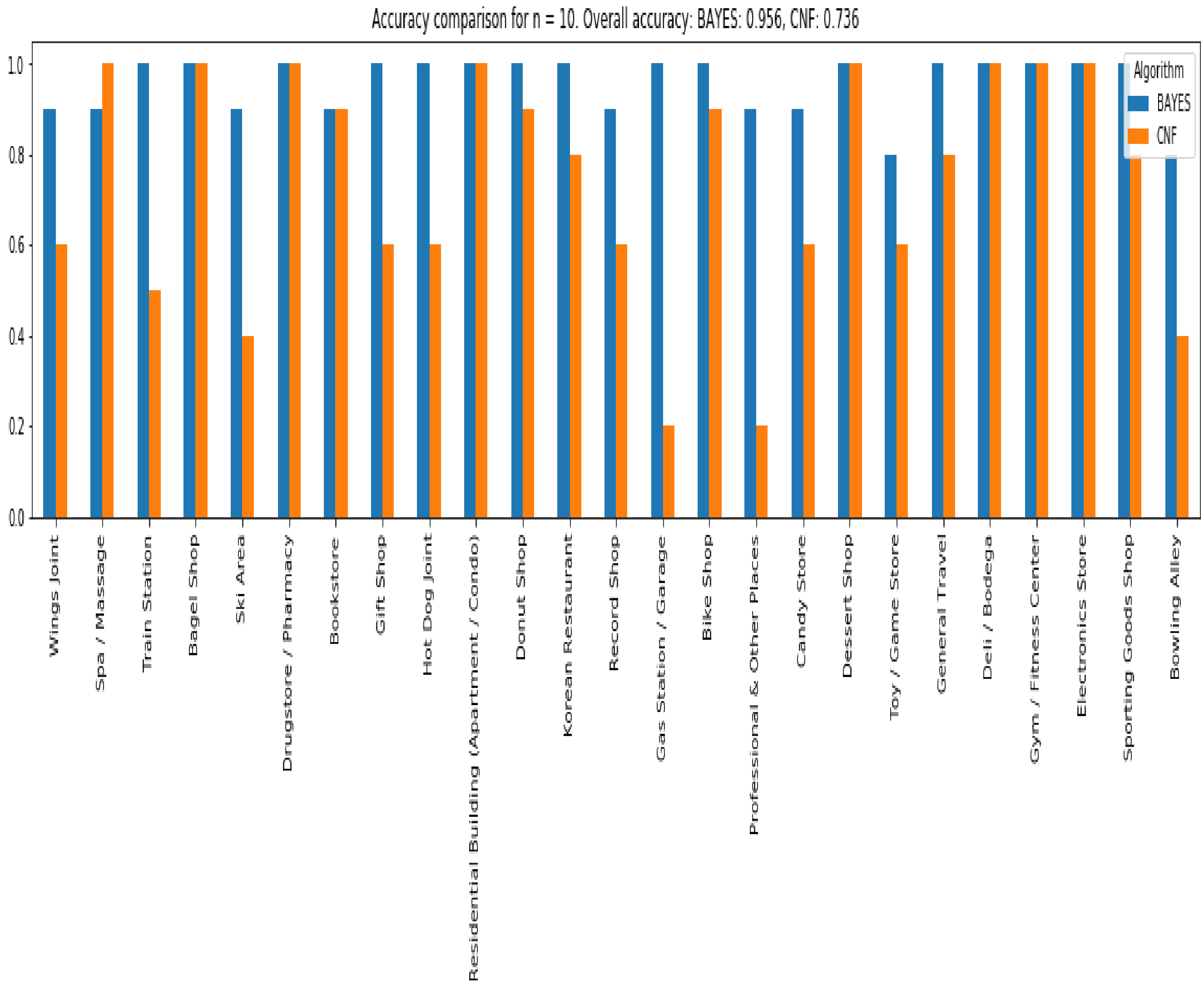


Fig. 4.1 (d) Bayesian accuracy 95%, CNF accuracy 73%

# CHAPTER 05

# Conclusion and Future Work

## 5.1 Conclusion

We have proposed a business recommendation solution by analyzing similarities of geographic neighborhoods using check-in data sets. Our solution generates a new neighborhood where a specific type of business location can be presented. The result is used to generate an appropriate neighborhood for a new venue. We have approached with two solutions: one on Bayesian inference and its approximation, and another on collaborative filtering. We have shown with experiments by using data from foursqare that the proposed solution can give us 96% accuracy with Bayesian Inference and 72% accuracy with CNF .

## 5.2 Future Work

We plan to work on an improved recommendation system where the system can estimate the appropriate location with high accuracy level. The amount of dataset we used were quite enough for the approximation, it needs a larger data-set and more inputs with different categories to estimate the result with more accuracy.

# REFERENCE


[1] “Foursquare.”[Online].Available:https://foursquare.com/.

[2]J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, “A Survey on Recommendations in Location-based Social Networks,” ACM Trans. vol. V, no. 212, 2013

[3] O. A. Sanosy, S. Rady, N. Badr, and M. Hashem, “Exploiting Location Based Social Networks in Business Predictions,” thesis, 11th International Conference on Innovations in Information Technology (IIT), Cairo, 2015.

[4] O. A. Sonosy, S. Rady, N. L. Badr, and M. Hashem, “A Study of Spatial Machine Learning for Business Behavior Prediction in Location Based Social Networks,” thesis, IEEE, 2016.

[5] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, 3rd ed. Waltham: Morgan Kaufmann, 2012.

[6] “K-means Clustering, major understanding issue,” matlab - K-means Clustering, major understanding issue - Stack Overflow. [Online]. Available: https://stackoverflow.com/questions/24645068/k-means-clustering-major-understanding-issue. [Accessed: 19-Aug-2017].

[7] “Flowchart summarising the method used by k-means cluste | Open-i,” U.S. National Library of Medicine. [Online]. Available: https://openi.nlm.nih.gov/detailedresult.php?img=PMC2655095_1471-2105-9-490-1&req=4. [Accessed: 19-Aug-2017].

[8] B. Eravci, N. Bulut, C. Etemoglu, and H. Ferhatosmanoglu, “Location Recommendations for New Businesses using Check-in Data,” thesis, IEEE, 2016.

[9] “Bayesian inference,” Wikipedia, 19-Aug-2017. [Online]. Available: https://en.wikipedia.org/wiki/Bayesian_inference. [Accessed: 20-Aug-2017].

[10] Symeonidis, P., Manolopoulos, Y. and Ntempos, D. (n.d.). Recommender Systems for Location-based Social Networks. [online] Google Books. Available at: https://books.google.com.bd/books?id=CUG9BAAAQBAJ&pg=PA35&lpg=PA35&dq=LBSN+wikipedia&source=bl&ots=BwMoHP5zdF&sig=5MEuial0ftacLfhcyioFzOjMpQE&hl=en&sa=X&redir_esc=y#v=onepage&q=LBSN%20wikipedia&f=false.

[11] “Collaborative filtering,” Wikipedia, 11-Aug-2017. [Online]. Available: https://en.wikipedia.org/wiki/Collaborative_filtering. [Accessed: 20-Aug-2017].

[12] H. Wang, M. Terrovitis, and N. Mamoulis, “Location Recommendation in Location-based Social Networks using User Check-in Data.” [Online]. Available: http://delivery.acm.org/10.1145/2530000/2525357/p364-wang.pdf?ip=123.49.46.144&id=2525357&acc=ACTIVE%20SERVICE&key=CB7B71C8A2C31385%2E231EC91298109147%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=799599138&CFTOKEN=53054139&__acm__=1503230525_af3e7cfb55c6f9052e67b76ec7078eb1.

[13] A. Anastasios, S. Scellato, N. Lathia, and C. Mascolo, “A Random Walk Around the City: New Venue Recommendation in Location-Based Social Networks.” [Online]. Available: https://www.cl.cam.ac.uk/~cm542/papers/socialcom12_noulas.pdf.

[14] Dingqi Yang, Daqing Zhang, Zhiyong Yu and Zhiwen Yu, Fine-Grained Preference-Aware Location Search Leveraging Crowdsourced Digital Footprints from LBSNs. In Proceeding of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013), September 8-12, 2013, in Zurich, Switzerland.