



Information Systems Department
Faculty of Computer & Information Sciences
Ain Shams University

Enhancing Tracking Techniques in Social Networks

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree
of PhD in Computer and Information Sciences

to

Department of Information Systems
Faculty of Computer and Information Sciences
Ain Shams University

by

Ola AlSayed Mostafa Omar AlSenosy

Assistant Lecturer

Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

Under Supervision of

Prof. Dr. Mohamed Hashem Abdalaziz

Professor, Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

Prof. Dr. Hossam Al-Deen Mostafa Faheem

Professor, Computer Systems Department
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

Prof. Dr. Nagwa Lotfy Badr

Professor, Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

Abstract

Understanding business behaviors requires acquiring huge amounts of data from diverse field studies. The additive growing use of mobile devices in social media, especially in recent years, provides large amounts of data transactions that can help in understanding business behaviors replacing the data acquired by the exhaustive field studies. Location Based Social Networks (LBSN) are considered as a solution providing such data used in urban analysis for economic reasons.

Towards more insight for business behavior in this dissertation, a suggestion of global perspective exploiting data collected from LBSNs is introduced in order to predict business behavior according to the business geographical location. Moreover, business behavior prediction in LBSNs is studied in this research for big data application. Prediction of customers' presence rates for business venues is introduced to be implemented using machine learning techniques. Machine learning techniques are investigated for both static and dynamic business predictions in LBSNs. Spatial regression models are thoroughly presented as static machine learning techniques. A comparative study is attained in this dissertation for suitability to model the relationships in LBSNs in order to be used for prediction. Geographically Weighted Regression (GWR) model proved to be the appropriate model in handling the sparse geographical distribution imposed by the LBSNs data. A proposed enhancement over the GWR model is introduced through a distributed training process that is integrated into a partitioned-GWR architecture. The proposed architecture includes a three blocks processes that are designed to deal with LBSNs data heterogeneity pursuing more enhanced predictions for business behavior.

For dynamic business predictions, spatial interpolation techniques are unprecedentedly proposed to be used in LBSNs. Enhancements over the spatial interpolation techniques are proposed in this study. A Local Filtered (LF) spatial interpolation is proposed to handle the data instabilities occurrences in LBSNs, to increase business prediction accuracy. As a second enhancement, a Similarity Embedded (SE) spatial interpolation is introduced to consider the diversity of features provided by the LBSNs' data in the prediction process. The proposed SE spatial interpolation suggested a hybrid feature similarity and distance

measurements involvement in the prediction process. Moreover, a design of a Filtered Similarity Embedded (FSE) spatial interpolation is proposed. The FSE spatial interpolation pursue additive accurate prediction results through a fusion of the two previously proposed spatial interpolation enhancements; LF spatial interpolation and SE spatial interpolation.

Since big data analytics are one of the most important topics of social media related research nowadays, an Iterative Nearest Neighbors First (INNF) search method is designed for timely efficient implementation of the proposed interpolation techniques over big datasets. The proposed INNF search method design the timely efficient solution using the geo indexing property provided by Not Only Structured Query Language (NOSQL) big databases.

For assessing both prediction accuracy and implementation time efficiency of the previously proposed techniques, extensive experiments are implemented over data extracted from Foursquare. Data is collected and analyzed for observation about venues registered in Foursquare residing in Texas State in the United States of America. The experiments results show noticeable improvements for prediction accuracies of the spatial interpolation enhancements proposed in this study. The LF spatial interpolation increases the prediction accuracies with 61% and 54% compared to the classical K Nearest Neighbors (KNN) and Inverse Distance Weighting (IDW) spatial interpolation techniques, respectively. While the SE spatial interpolation increases the prediction accuracies with 44% and 37% compared to the KNN and IDW spatial interpolation techniques, respectively. Moreover, the FSE spatial interpolation increases the prediction accuracies with 67% and 63% compared to the KNN and IDW spatial interpolation techniques, respectively.

The proposed INNF search method shows remarkable improvements in experiments. The INNF search method reduces the implementation runtime of the SE spatial interpolation in terms of milliseconds compared to the tens of seconds classical implementations in the experiments attained over the Foursquare dataset. Moreover the INNF search method shows steady runtimes for successive implementations over increased sized several synthetic datasets.

List of Publications

1. Ola Al Sonosy , Sherine Rady , Nagwa Lotfy Badr , Mohammed Hashem, “Exploiting location based social networks in business predictions”, Proceedings of the 11th International Conference on Innovations in Information Technology (IIT '15), Cairo, Egypt, p.40-50, Nov 2015, DOI:10.1109/INNOVATIONS.2015.7381512.
2. Ola Al Sonosy , Sherine Rady , Nagwa Lotfy Badr , Mohammed Hashem, "Machine Learning Techniques for Mining Location-based Social Networks for Business Predictions", The 10th International Conference on Informatics and Systems (INFOS '16), Cairo, Egypt, May 2016, DOI: [10.1145/2908446.2908475](https://doi.org/10.1145/2908446.2908475).
3. Ola Al Sonosy , Sherine Rady , Nagwa Lotfy Badr , Mohammed Hashem, “Business Behavior Predictions using Location Based Social Networks in Smart Cities”, accepted for publication in ebook of Information Innovation Technology in Smart Cities, Springer Singapore, 2017, ISBN 978-981-10-1741-4.
4. Ola Al Sonosy , Sherine Rady , Nagwa Lotfy Badr , Mohammed Hashem, “Towards Efficient Business Behavior Prediction using Location Based Social Networks”, submitted to Wires Data Mining and Knowledge Discovery Journal. Impact factor: 1.759.
5. Ola Al Sonosy , Sherine Rady , Nagwa Lotfy Badr , Mohammed Hashem, “A Study of Spatial Machine Learning for Business Behavior Prediction in Location Based Social Networks”, 11thIEEE International Conference on Computer Engineering and Systems (ICCES 2016), Cairo, Egypt, p. 266 - 272, December 2016, DOI: 10.1109/ICCES.2016.7822012.

Table of Contents

Abstract.....	I
List of Publications	III
Table of Contents	IV
List of Figures.....	VII
List of Tables	IX
List of Abbreviations	X
Chapter 1: Introduction	1
1.1.Motivation.....	1
1.2.Objectives	2
1.3.Research Contributions.....	3
1.4.Dissertation Organization	5
Chapter 2: Location Based Social Networks Background	8
2.1.Introduction.....	8
2.2.Online Location Based Social Networks.....	8
2.3.Historical background.....	10
2.4.Location-Based Social Network Data Characteristics	11
2.5.LBSNs Prediction Metrics.....	14
2.5.1. Message related metrics	14
2.5.2. Nodes related metrics	16
2.6.Social Data Mining Prediction Techniques	18
2.6.1. Regression methods.....	18
2.6.2. Bayes classifier	19
2.6.3. K-Nearest Neighbor classifier	20
2.6.4. Artificial Neural Network.....	21
2.6.5. Decision Tree.....	21
2.6.6. Collaborative Filtering prediction methods.....	22
2.6.7. Model-based prediction	22
2.7.Dynamic versus Static Data Analysis.....	23

2.8.Summary	24
Chapter 3: Location Based Social Network State of the Art.....	26
3.1.Introduction.....	26
3.2.Location Based Social Networks Studies.....	26
3.2.1. Location-Aware Recommendation systems:.....	27
3.2.2. Location-Aware Business Enhancement systems	39
3.2.3. Location-Aware Community Clustering	40
3.2.4. Human Mobility Analysis Research.....	46
3.3.Summary	49
Chapter 4: Mining LBSNs for Business Behavior Prediction	51
4.1.Introduction.....	51
4.2.Machine Learning.....	52
4.3.Static Business Prediction in LBSNs	54
4.3.1. Spatial Durbin Model	56
4.3.2. Spatial Durbin Error model	57
4.3.3. Spatial Lag Model	58
4.3.4. Spatial Lag X model.....	58
4.3.5. Spatial Error model.....	59
4.3.6 Geographically Weighted Regression	60
4.3.7. The Proposed Partitioned-GWR Architecture.....	62
4.4.Dynamic Business Prediction In LBSNs.....	67
4.4.1. Spatial Interpolation	67
4.4.2. Local Filtered Spatial Interpolation:.....	73
4.4.3. Similarity Embedded Spatial Interpolation	78
4.4.4. Filtered Similarity Embedded Spatial Interpolation.....	82
4.5.Proposed Application of Similarity Embedded Spatial Interpolation In big data	86
4.5.1. Big data analytics	86
4.5.2. The Proposed Iterative Nearest Neighbors First Search	90
4.6.Summary	96
Chapter 5: Performance Evaluation and Experimental Case Studies	98
5.1.Introduction.....	98

5.2.Data Observations:.....	99
5.2.1. Data Extraction.....	99
5.2.2. Data Findings.....	103
5.3.Data Preprocessing	108
5.4.Data Prediction	110
5.4.1. Spatial Regression models Comparison:.....	110
5.4.2. Spatial Interpolation Techniques Comparison	116
5.4.3. Spatial Regression versus Spatial Interpolation	127
5.5.Performance Time	129
5.5.1. Effect of search window size.....	130
5.5.2. Effect of neighborhood size.....	131
5.5.3. Effect of Database Size	132
5.6.Summary	133
Chapter 6 : Conclusions and Future Work.....	137
6.1.Conclusions.....	137
6.2.Future Work suggestions	141
References	143

List of Figures

Figure 2.1 Location Based Social Network.....	9
Figure 2.2 Node representation of social network.....	16
Figure 3.1 LBSNs research studies domains	27
Figure 4.1 Partitioned-GWR Architecture.....	63
Figure 4.2. Business Prediction Spatial Interpolation	71
Figure 4.3 local extreme example.....	75
Figure 4.4 Local Filtered Spatial Interpolation	76
Figure 4.5 Box and Whiskers outlier filter	77
Figure 4.6 Similarity Embedded Spatial Interpolation.....	80
Figure 4.7 Filtered Similarity Embedded Spatial Interpolation Process ..	84
Figure 4.8 Iterative Nearest Neighbor First search.....	93
Figure 4.9 Iterative Nearest Neighbor First Search Algorithm	95
Figure 5.1 Venues Extraction Process	102
Figure 5.2 The Number of venues in each state in the United States of America.....	104
Figure 5.3: (a) Venues densities heat map (b) Venues check-ins heat map	105
Figure 5.4 Number of venues versus number of customers' presence rates per category.....	108
Figure 5.5 Spatial Regression Prediction Error Comparison	112
Figure 5.6 Prediction Error using GWR against partitioned-GWR prediction in each data partition	114
Figure 5.7 Prediction Error using GWR against partitioned-GWR prediction	115
Figure 5.8 Prediction Error for KKN Spatial Interpolation vs LF-KNN Spatial Interpolation for $k=[5:200]$	118

Figure 5.9 Prediction Error for KKN Spatial Interpolation vs SE-KNN Spatial Interpolation for k=[5:200].....	118
Figure 5.10 Prediction Error for KKN Spatial Interpolation vs FSE-KNN Spatial Interpolation for k=[5:200].....	119
Figure 5.11 Prediction Error for KKN Spatial Interpolation vs FSE-KNN Spatial Interpolation for k=[5:200].....	120
Figure 5.12 Prediction Error for IDW Spatial Interpolation vs LF-IDW Spatial Interpolation for k=[5:200].....	121
Figure 5.13 Prediction Error for IDW Spatial Interpolation vs SE-IDW Spatial Interpolation for k=[5:200].....	122
Figure 5.14 Prediction Error for IDW Spatial Interpolation vs FSE-IDW Spatial Interpolation for k=[5:200].....	122
Figure 5.15 Prediction Error for SE Spatial Interpolation and FSE Spatial Interpolation for k=[5:200].....	123
Figure 5.16 Comparison of Prediction Error of IDW vs KNN applied for Spatial Interpolations for k=[5:200]	126
Figure 5.17 Comparison of GWR and Spatial Interpolations Prediction Error	127
Figure 5.18 Implementation Time of INNF search method with w=[500:10000].....	131
Figure 5.19 Implementation Time of INNF search method with neighborhood sizes k =[10:60]	132
Figure 15.20 (a) Classical Implementation Time (b) INNF Search Implementation Time of Similarity neighborhood determination with dataset sizes [1:5] millions of records	133

List of Tables

Table 4.1 Comparison of Spatial Regression Models Applications in LBSNs business customers' presence rates predictions	60
Table 5.1 Foursquare Venue Features	99
Table 5.2 Number of venues versus number of customers' presence rates per category	106
Table 5.3 Partitioned-GWR versus GWR Prediction Error	115
Table 5.4 Comparison between prediction error reduction between KKN Spatial Interpolation, against LF Spatial Interpolation, SE Spatial Interpolation and FSE Spatial Interpolation	120
Table 5.5 Comparison between prediction error reduction between IDW Spatial Interpolation, against LF Spatial Interpolation, SE Spatial Interpolation and FSE Spatial Interpolation	123
Table 5.6 Prediction error reductions of the modified spatial interpolation techniques against the classical KNN and IDW spatial interpolation techniques	126
Table 5.7 Relative reduction of GWR prediction error than Spatial Interpolations techniques	128
Table 5.8 Implementation time for similarity based neighborhood determination in INNf method vs classical implementation	130

List of Abbreviations

CLAF	Collaborative Location and Activity Filtering
CSE-Tree	Compressed Start End Tree
FSE	Filtered Similarity Embedded
GIS	Geographic Information Systems
GLS	Generalized Least Squared
GPS	Geographical Positioning Systems
GWR	Geographically Weighted Regression
HGBSM	Hierarchical Graph Based Similarity Measurement
IDF	Inverse Document Frequency
IDW	Inverse Distance Weight
INNF	Iterative Nearest Neighbor First
k-BCT	k Best-Connected Trajectory
KNN	K Nearest Neighbors
LARS	Location Aware Recommender System
LBS	Location Based Services
LBSN	Location Based Social Network
LDA	Latent Dirichlet Allocation
LF	Local Filtered
MTM	Maximal Travel Match
NOSQL	Not Only Structured Query Language
OLS	Ordinary Least Squared
PCLAF	Personalized Collaborative Location and Activity Filtering
POI	Points Of Interests
RAE	Relative Absolute Error

RPCLAF	Ranking-based Personalized Collaborative Location and Activity Filtering
SAR	Spatial Auto Regression
SDM	Spatial Durbin Model
SE	Similarity Embedded
SEM	Spatial Error Model
SErr	Standard Error
SLH	Semantic Location History
SLX	Spatial Lag X
TBHG	Tree Based Hierarchical Graph
WCH	Weighted Category Hierarchy

Chapter 1

Introduction

With the recent improvements of mobile devices manufacturing and location acquisition technologies, a huge number of location-based data transactions has been available for data scientists. These amounts of location related data is forming a very rich soil for data analytics and realizing new location aware services. Various large scale location based data, including user generated data in location based social networks or geographical positioning systems data reported from vehicles or sensors, have led to research challenges and opportunists to provide location based services, intelligent transportation systems, geographic information systems, urban analysis reports, and smart cities services.

Different computer science techniques, like data mining, machine learning, artificial intelligence, and spatial or spatio-temporal databases, are used or combined to address such challenges.

1.1. Motivation

The rapid technological advances of the mobile networks, social media and the fact that mobile devices with Geographical Positioning Systems (GPS) modules are now been used everywhere, motivate social networks users to share information about their real movements as online transactions, by checking in places or venues especially in Location Based Social Networks (LBSNs) such as Foursquare and Facebook Places. LBSNs have been attracting more and more users by providing services that integrate social activities with geographic information. In

LBSNs, a user can discover places of interests around his current location, check-in at these places and also share his/her check-in information and opinions about places with his/her online friends.

Over years of heavy use, LBSNs have accumulated very large amounts of information related to business activities along with their associated geographical information. Unprecedentedly, urban analysis can be used for mining LBSN's originated datasets to infer information that enables more understanding about business behavior. Business behavior understanding can be beneficial to business owners for making business related decisions like opening new business places, initiating advertisements campaigns, improving provided services qualities or even closing business decisions.

1.2. Objectives

The main objective of this study is to introduce a global comprehension of LBSNs' data to serve business owners. As, LBSNs accumulated with large rates through time, a global analysis of these huge amounts of data can give a deep insight about business behavior in geographical terrains. This dissertation includes a suggestion of quantitative analytical study to provide conclusions of the business behavior and patterns for business owners and decision makers. The study aims to exploit the huge amounts of information available about users' check-ins for business venues on the cumulative LBSNs datasets in predicting users' number of check-ins for venues in order to understand the business behavior.

As data originated by social media are accumulated in millions of records, therefore, methods and algorithms that will be used in analyzing LBSNs data has to be assessed for big data efficient implementations.

1.3. Research Contributions

The research study of this dissertation can be summarized in the following contributions:

- Investigation of spatial regression models as static prediction methods to be applied in LBSNs' data. Several presented models are discussed in a scientific comparative study for suitability for LBSNs' data. The study proved the Geographically Weighted Regression (GWR) as the model the suitable for presenting the relationships of data originated by LBSNs. This is presented in chapter 4, sections 4.3.1- 4.3.6.
- Proposing a distributed training GWR model to deal with the data heterogeneity imposed by the LBSNs' diverse features characterizing their entities. The proposed model is integrated into a design of a proposed partitioned-Geographically Weighted Regression architecture to be used for predictions in LBSNs. This is presented in chapter 4, section 4.3.7.
- Unprecedented introduction of Spatial Interpolation techniques to be used for predictions in LBSNs. The spatial interpolation techniques are used as dynamic prediction techniques which fits the changeable nature of features describing entities in LBSNs. This is presented in chapter 4, section 4.4.1.
- A proposal of a Local Filtered spatial interpolation enhancement is introduced in this study. The proposed method provides enhanced predictions in LBSNs by decreasing the negative effect of the local extremes that are detected in the LBSNs' data. This is presented in chapter 4, section 4.4.2.