IMPROVED LINK PREDICTION FOR LOCATION BASED SOCIAL NETWORKS WITH NOVEL FEATURES AND CONTEXTUAL FEATURE REDUCTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET ENGIN BAYRAK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

DECEMBER 2019

Approval of the thesis:

IMPROVED LINK PREDICTION FOR LOCATION BASED SOCIAL NETWORKS WITH NOVEL FEATURES AND CONTEXTUAL FEATURE REDUCTION

submitted by **AHMET ENGIN BAYRAK** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Faruk Polat
Supervisor, **Computer Engineering Department, METU** _____

Examining Committee Members:

Prof. Dr. Özgür Ulusoy
Computer Engineering Department, Bilkent University _____

Prof. Dr. Faruk Polat
Computer Engineering Department, METU _____

Prof. Dr. Reda Alhajj
Department of Computer Science, University of Calgary _____

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Tansel Özyer
Computer Engineering Department, TOBB ETU _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ahmet Engin Bayrak

Signature :

ABSTRACT

IMPROVED LINK PREDICTION FOR LOCATION BASED SOCIAL NETWORKS WITH NOVEL FEATURES AND CONTEXTUAL FEATURE REDUCTION

Bayrak, Ahmet Engin
Ph.D., Department of Computer Engineering
Supervisor: Prof. Dr. Faruk Polat

December 2019, 106 pages

High penetration of broadband Internet access has made a revolution on the web usage, where users have become content generators rather than just consuming. People started to communicate, interact, maintain relationship and share data (image, video, note, location, etc.) with their acquaintances through varying online social network sites which are the key factors of that internet usage revolution. Online social networks with location sharing and interaction between people are called Location Based Social Networks (LBSNs). To use and benefit more from social networks, real life social links (friendship, acquaintanceship) should be represented well on them. Link Prediction problem has a motivation of studying social network evolution and trying to predict future possible links for representing the real-life relations better. In this work, we studied a comprehensive feature set which combines topological features with features calculated from temporal interaction data on LBSNs. We proposed novel features which are calculated by using time, category and common friend details of candidates and their social interaction in LBSNs. In addition, we proposed an effective feature reduction mechanism which helps to determine best feature subset in

two steps. Contextual feature clustering is applied to remove redundant features and then a non-monotonic selection of relevant features from the calculated clusters are done by a custom designed genetic algorithm. Results depict that both new features and the proposed feature reduction method improved link prediction performance for LBSNs.

Keywords: link prediction, social networks, location based social networks, feature extraction, feature reduction

ÖZ

ORJİNAL ÖZNİTELİKLER VE BAĞLAMSAL ÖZNİTELİK AZALTMA YÖNTEMİ İLE KONUM TABANLI SOSYAL AĞLAR İÇİN GELİŞTİRİLMİŞ BAĞLANTI TAHMİNİ

Bayrak, Ahmet Engin
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Faruk Polat

Aralık 2019 , 106 sayfa

Geniş bant internet erişiminin yayılışı, kullanıcıların sadece tüketen olmak yerine içerik üretici olduğu bir internet kullanım devrimi yaptı. Bu devrimin önemli faktörlerinden olan çeşitli çevrimiçi sosyal ağ siteleri aracılığıyla; insanlar tanıdıkları ile iletişim kurmakta, etkileşmekte, kurdukları ilişkileri idame ettirmekte ve bilgi (resim, video, not, konum, vb.) paylaşmaktadır. Kişiler arasında konum paylaşımı ve etkileşimi yapılan sosyal ağlara Konum Tabanlı Sosyal Ağlar (KTSA) denir. Sosyal ağların daha fazla kullanılması ve yararlanılması için; gerçek hayat sosyal bağlantılarının (arkadaşlık, tanıdıklık) iyi temsil edilmesi ile mümkündür. Bağlantı tahmini probleminin, sosyal ağ evrimini inceleme ve gerçek yaşam ilişkilerini daha iyi temsil edebilmek için gelecekteki olası bağlantıları tahmin etme hedefi vardır. Araştırmamızda, KTSA'lardaki zamansal etkileşim verilerinden hesaplanan öznitelikler ile topolojik öznitelikleri birleştiren kapsamlı bir öznitelik kümesi çalışıldı. Biz bağlantı adayları için zaman, mekan kategorisi ve ortak arkadaşların detaylı bilgilerini kullanarak hesaplanan orjinal öznitelikler önerdik. Ayrıca, en iyi performanslı öznitelik alt

kümesini belirlemek için iki aşamalı bir öznitelik azaltma mekanizması geliştirdik. Önce gereksiz özniteliklerden kurtulmak için benzer öznitelikleri kümeledik. Daha sonra, hesaplanmış kümelerden ilişkili öznitelikleri özel olarak tasarlanmış bir genetik algoritma yardımı ile monoton olmayan bir şekilde seçtik. Bizim önerdiğimiz bu yeni özniteliklerin ve öznitelik azaltma yönteminin, KTSA'lar için bağlantı tahmini performansını geliştirdiği gözlemlenmiştir.

Anahtar Kelimeler: bağlantı tahmini, sosyal ağlar, konum tabanlı sosyal ağlar, öznitelik çıkarma, öznitelik azaltma

To My Dad

ACKNOWLEDGMENTS

I would like to present my deepest thanks to my thesis supervisor Prof. Dr. Faruk Polat for his valuable guidance, motivation and support throughout this thesis study.

My special thanks go to my wife Ayşe Hilal Bayrak and my son Mete Ediz Bayrak for their help and support to complete this work and to all my friends who gave me support whenever I needed.

I am very grateful to my family for all their patience and tolerance. This thesis is dedicated to the memory of my father, İsmail Bayrak, who always believed in my ability to be successful in the academic arena. You are gone but your belief in me has made this journey possible.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xv

LIST OF FIGURES xviii

LIST OF ABBREVIATIONS xix

CHAPTERS

1 INTRODUCTION 1

2 BACKGROUND AND RELATED WORK 5

2.1 Location Based Social Networks 5

2.2 Link Prediction Problem for Location Based Social Networks 8

2.3 Related Work 9

2.3.1 Link Prediction Researches 9

2.3.2 Link Prediction Researches for Location Based Social Networks 12

2.3.3 Optimal Feature Subset Selection Research 15

3 OUR RESEARCH FUNDAMENTALS 19

3.1 Problem and Solution Formulation 19

3.2 Data 20
3.2.1 Data Enhancement 21
3.2.2 Data Groups 22
3.3 Link Prediction Framework 23
3.3.1 Feature Extraction 23
3.3.2 Supervised Machine Learning 24
3.3.3 Predictions and Performance Evaluation 25
4 CONTEXTUAL FEATURE ANALYSIS 27
4.1 Features from Literature 27
4.1.1 Topological Features 27
4.1.2 Interaction Features 29
4.1.3 Performance Evaluation 32
4.1.3.1 Friend-of-Friends Group (FOF) Performances 32
4.1.3.2 Place-Friends Group (PF) Performances 32
4.1.3.3 Both-Friends Group (BF) Performances 32
4.1.3.4 Whole Group (WG) Performances 32
4.2 Proposed Features 37
4.2.1 Proposed Contextual Features 38
4.2.2 Performance Evaluation 39
4.2.2.1 Friend-of-Friends (FOF) Group Performances 40
4.2.2.2 Place-Friends (PF) Group Performances 40
4.2.2.3 Both-Friends (BF) Group Performances 40
4.2.2.4 Whole Group (WG) Performances 41

4.2.3 Using Proposed Contextual Features with Existing Ones 41
4.2.3.1 Friend-of-Friends (FOF) Group Performances 42
4.2.3.2 Place-Friends (PF) Group Performances 42
4.2.3.3 Both-Friends (BF) Group Performances 43
4.2.3.4 Whole Group (WG) Performances 43
4.2.4 Evaluation 45
5 FEATURE REDUCTION RESEARCH 47
5.1 Greedy Feature Reduction Method for Link Prediction Efficiency . . 48
5.1.1 Cost of Feature Extractions 48
5.1.2 Individual Link Prediction Performance of Features 48
5.1.3 Selecting an Efficient Feature Subset 50
5.1.4 Performance Evaluation 50
5.1.4.1 Naive Bayes Classifier (NBC) Performances 51
5.1.4.2 Bayesian Network (BN) Performances 51
5.1.4.3 Random Forest (RF) Performances 53
5.1.5 Evaluation 53
5.2 Clustering Based Feature Reduction for Link Prediction Effectivity . 55
5.2.1 Clustering Similar Features 55
5.2.1.1 K-means Clustering 59
5.2.1.2 Agglomerative Clustering 60
5.2.2 Optimal Subset from Clustered Features 61
5.2.3 GA Formulation 62
5.2.4 Proposed Fitness Function 62

5.2.5 Performance Evaluation 63
5.2.5.1 Comparison of Clustering Algorithms 64
5.2.5.2 Friend-of-Friends Group (FOF) Performances 64
5.2.5.3 Place-Friends Group (PF) Performances 68
5.2.5.4 Both-Friends Group (BF) Performances 68
5.2.5.5 Whole Group (WG) Performances 70
5.2.5.6 Temporal Proximity Analysis for Positive Link Predictions 77
5.2.6 Evaluation 78
6 FEATURE MINING 81
6.1 Proposed Place Category Based Features 81
6.1.1 Performance Evaluation 83
6.1.1.1 Friend-of-Friends Group (FOF) Performances 85
6.1.1.2 Place-Friends Group (PF) Performances 85
6.1.1.3 Both-Friends Group (BF) Performances 85
6.1.1.4 Whole Group (WG) Performances 85
6.1.2 Evaluation 90
7 CONCLUSION AND FUTURE WORK 93
7.1 Conclusions 93
7.2 Future Work 95
REFERENCES 97
CURRICULUM VITAE 105

LIST OF TABLES

TABLES

Table 3.1 Data Groups' Statistics 23

Table 4.1 FOF - Literature Features' Single Performances 33

Table 4.2 PF - Literature Features' Single Performances 34

Table 4.3 BF - Literature Features' Single Performances 35

Table 4.4 WG - Literature Features' Single Performances 36

Table 4.5 Classification with Topological Features 36

Table 4.6 Classification with Interaction Features 37

Table 4.7 Classification with All Features 37

Table 4.8 FOF - Our Proposed Features' Single Performances 40

Table 4.9 PF - Our Proposed Features' Single Performances 40

Table 4.10 BF - Our Proposed Features' Single Performances 41

Table 4.11 WG - Our Proposed Features' Single Performances 41

Table 4.12 FOF - Proposed Features' Performances with Existing Features . . . 42

Table 4.13 FOF - Prediction Performances Comparison with Other Research . . 42

Table 4.14 PF - Proposed Features' Performances with Existing Features 43

Table 4.15 PF - Prediction Performances Comparison with Other Research . . . 43

Table 4.16 BF - Proposed Features' Performances with Existing Features 44

Table 4.17 BF - Prediction Performances Comparison with Other Research . . . 44

Table 4.18 WG - Proposed Features' Performances with Existing Features . . . 44

Table 4.19 WG - Prediction Performances Comparison with Other Research . . 45

Table 5.1 Feature Extraction Costs for WG in Milliseconds 49

Table 5.2 Measurements with NBC 51

Table 5.3 Picked Feature Subsets for NBC 52

Table 5.4 Measurements with BN 53

Table 5.5 Picked Feature Subsets for BN 54

Table 5.6 Measurements with RF 55

Table 5.7 Picked Feature Subsets for RF 56

Table 5.8 WG - Comparison of Clustering Algorithms 64

Table 5.9 FOF - Compared Feature Subset Performances (ROC-AUC) 68

Table 5.10 FOF - Picked Feature Subsets 69

Table 5.11 FOF - Prediction Performances Comparison with Other Research . . 70

Table 5.12 PF - Compared Feature Subset Performances (ROC-AUC) 70

Table 5.13 PF - Picked Feature Subsets 71

Table 5.14 PF - Prediction Performances Comparison with Other Research . . . 72

Table 5.15 BF - Compared Feature Subset Performances (ROC-AUC) 72

Table 5.16 BF - Picked Feature Subsets 73

Table 5.17 BF - Prediction Performances Comparison with Other Research . . . 74

Table 5.18 WG - Compared Feature Subset Performances (ROC-AUC) 74

Table 5.19 WG - Picked Feature Subsets 75

Table 5.20 WG - Prediction Performances Comparison with Other Research . . 76

Table 5.21 WG - True Positive Rates for Two Proximity Groups 77

Table 5.22 WG - True Positive Rates for Three Proximity Groups 78

Table 6.1 Existing Features 84

Table 6.2 FOF Group Results with Mined Category Features 86

Table 6.3 PF Group Results with Mined Category Features 87

Table 6.4 BF Group Results with Mined Category Features 88

Table 6.5 WG Results with Mined Category Features 89

LIST OF FIGURES

FIGURES

Figure 5.1 Feature Selection Process 57

Figure 5.2 NBC - Agglomerative Hierarchical Clustering Dendrogram with Single Linkage 65

Figure 5.3 NBC - Agglomerative Hierarchical Clustering with Average Linkage Dendrogram 65

Figure 5.4 NBC - Agglomerative Hierarchical Clustering with Complete Linkage Dendrogram 66

Figure 5.5 BN - Agglomerative Hierarchical Clustering with Single Linkage Dendrogram 66

Figure 5.6 BN - Agglomerative Hierarchical Clustering with Average Linkage Dendrogram 67

Figure 5.7 BN - Agglomerative Hierarchical Clustering with Complete Linkage Dendrogram 67

Figure 5.8 WG - ROC Curve of Other Research with NBC Algorithm . . . 72

Figure 5.9 WG - ROC Curve of OUR OPTIMAL SUBSET with NBC Algorithm 74

Figure 5.10 WG - ROC Curve of Other Research with BN Algorithm 76

Figure 5.11 WG - ROC Curve of Our Optimal Subset with BN Algorithm . . 76

LIST OF ABBREVIATIONS

ABBREVIATIONS	
AACF	Adamic Adar Common Friends
AACP	Adamic Adar Common Places
AACPE	Adamic Adar Common Places Entropy
AUC	Area Under Curve
BF	Both Friends
BN	Bayesian Network
BPFS	Best Performing Feature Set
CCC	Common Check-in Count
CCCP	Common Category Check-in Counts Product
CCCPR	Common Category Check-in Counts Product Ratio
CCCSP	Common Category Check-in Count Sum Product
CFC	Common Friend Count
CFR	Common Friend Ratio
CMIM	Conditional Mutual Information Maximization
CPC	Common Place Count
CPCP	Common Place Check-in Counts Product
CPCPR	Common Place Check-in Counts Product Ratio
CPCPS	Common Place Check-in Count Product Sum
CPR	Common Place Ratio
DLH	Distance Of Lat-Lon Homes
DMVH	Distance Of Most Visited Homes
DMVHR	Distance Of Most Visited Homes Ratio
ED	Euclidean Distance

ELF	Existing Literature Features
FCBF	Fast Correlation Based Filter
FF	Fitness Function
FNFS	Filtered New Feature Set
FOF	Friend Of Friends
FP	Single Feature Performance
GA	Genetic Algorithm
GPS	Global Positioning System
LLH	Lat-Lon Home
LP	Link Prediction
MCC	Minimum Check-ins Count of Common Places
MEV	Minimum Entropy Value of Common Places
MIFS	Mutual Information-Based Feature Selection
MKL	Multiple Kernel Learning
MP	Mutual Feature Performance
MPV	Mutual Performance Vector
MRMR	Max-Relevance and Min-Redundancy
MST	Minimum Spanning Tree
MVH	Most Visited Home
NASA	National Aeronautics and Space Administration
NBC	Naïve Bayes Classifier
NP	Non-deterministic Polynomial
OSF	Only Successful Features
OSLCF	Only Successful Low Cost Features
OSN	Social Network
PAC	Preferential Attachment Check-ins
PAF	Preferential Attachment Friends

PAP	Preferential Attachment Places
PPBD	Prediction Period Beginning Date
PF	Place Friends
RF	Random Forest
RLH	Radius Length from Lat-Lon Home
ROC	Receiver Operating Characteristic
ROC-AUC	Area Under Receiver Operating Characteristic Curve
SA	Simulated Annealing
SAGA	Simulated Annealing Genetic Algorithm
SN	Social Network
SNA	Location Based Social Network
SNA	Social Network Analysis
SPD	Shortest Path Distance
SRLH	Sum of Radius Length from Lat-Lon Homes
SVM	Support Vector Machine
TCFC	Total Common Friend Closeness
TCFCC	Total Common Friend Common Check-in
TSP	Total Shortest Paths
WEKA	Waikato Environment for Knowledge Analysis
WG	Whole Group

CHAPTER 1

INTRODUCTION

Social network is a form of network which describes a social structure between actors using relations and interactions. Mostly, actors (nodes) are humans or organizations and interactions between actors (edges) vary according to services provided. Social networks were used by social scientists for analyzing the transmission of information between actors like individuals, groups and organizations.

Researchers from various disciplines have researched on social networks since 1930s. Main motivation of such researches was extracting information of the social structure using some analysis methods. Related research that combines graph theory, algebra, statistics, sociometry and psychometry is called **Social Network Analysis (SNA)**. SNA mainly focuses on analyzing the relationship between people on a society modeled as a network. Analyzing human behavior by location and time is another aspect of SNA for social science researchers [1].

Online Social Networks (OSNs) are internet platforms that such network data is formulated digitally by the help of web technologies. In recent years, smart phones and broader web accessibility technologically boosted content sharing and reaching through evolving OSNs which are believed to be reflection of physical social network.

OSNs have grown rapidly with varying platforms that empowered users to upload miscellaneous digital social content. Users adopted the idea of easily sharing and easily reaching personal digital content (photo, video, location, etc.) with closed network of acquaintance [2].

Human social life is expanding by new people and new actions/interactions everyday. Therefore, new nodes and edges emerge on their corresponding social networks. Such

dynamic structure of social networks makes it harder to represent them for the online ones. OSNs are trying to formulate that graph evolution to keep their graphs up-to-date. Success of an OSN platform is correlated with number of users who are sharing information with the friends through them. Growth in the number of users and growth in the friendship link between users are two key factors for such success [3]. To help on those growth, OSNs generally comes with recommendations about possible friendship links such as:

- people you may know in Facebook,
- who to follow in Twitter and
- friend suggestions of Foursquare

Considering the scale of OSNs and the importance of their growth, smart friendship recommendations should be performed through data-driven analysis and algorithms. SNA researchers studied the defined problem as **Link Prediction (LP)** problem. Main motivation of LP is to formulate evolution of society in fast evolving network with continuously added new users and new relationships. LP problem in a social network can be defined as:

> *"Given a social network at a time t, predict new edges to be added to the network at some future time t'" [4].*

Existing studies on link prediction make use of statistical and topological features in addition to domain related custom ones. There are various applications that benefit from such prediction approaches like citation analysis [5] and online marketing [6].

Smartphones and mobile devices embedded with geographical sensors triggered development of OSNs capable of sharing locations with others. Location Based Social Networks (LBSNs) are platforms where users share location information with friends, named **check-in**. Check-in is performed as a user's location data sharing in places like "SFO Airport" or "Stanford University" with a timestamp. LBSNs such as Facebook, Foursquare, Gowalla and etc. collect temporal location information attached with user social profiles and friendship data between user profiles [7].

LBSNs use two different structural subgraphs [8] which make data richer and encourage SNA researchers to study separate LP solutions for LBSNs:

- *Friendship subgraph*, where friendship edges formed between two user profile vertices

- *Check-In subgraph*, where check-in edges formed between place and user profile vertices

In our research, we focused on improving LP performance to predict new friendships in LBSNs with feature-based approaches. LP solutions for OSNs are mainly based on features calculated from friend topologies (users and friendship) like **Common Friend Count (CFC)** of link candidates. CPC is total number of unique users who are friends with both link candidates.

Link prediction for LBSNs is a similar research problem with some data richness provided by exclusive check-in and place information. In addition to topological features, exclusive information can be used for calculating semantically rich interaction features like **Common Place Count (CPC)** of link candidates. CPC is the total number of places where both link candidates checked-in. Features extracted from check-in information play vital role for predictor performances.

After making a comprehensive literature survey, we collected a core feature set for the problem. In addition to those features, we analyzed LBSN data to design and propose new contextual features based on time, common friend details and place category information in check-in data [9]. Our primary aim was to make use of available information in LBSN data that cannot be utilized by the existing features. Experiments showed that using our proposed features improve the link prediction performance.

During experiments with whole set of combined existing and proposed features, we also observed the requirement of feature reduction to determine optimized feature subset. Feature reduction has a motivation to eliminate features that deteriorates link prediction performance because of the overlapping information.

We proposed two reduction methods:

- A greedy feature reduction method for efficient link prediction using feature extraction time costs and performance ranking [10].
- A clustering based feature reduction methodology for effective link prediction using features' interaction with each others [11].

Clustering based feature reduction method helped to improve LP performance and greedy feature reduction method reduced the time cost of LP solution.

With the motivation of finding features with exclusive information gain, we also performed feature mining focused on place category information. Two proposed groups of features were calculated separately for each category in dataset for a link candidate pair. Our results depict that some place categories are more correlated with new friendship and we could enhance prediction performance by usage of category semantic through features from those categories [12].

This thesis is structured as follows. **Chapter 2** gives a general introduction to the LP problem, LBSNs and extensive combination of related work in SNA literature. In **Chapter 3**, research fundamentals like data, solution formulation and framework are discussed. In the following chapter, **Chapter 4**, a presentation of contextual feature analysis is given, both for the literature existing features and proposed contextual features. In **Chapter 5**, we covered feature reduction studies performed. **Chapter 6** is meant to share details of feature mining efforts and findings. The closing chapter, **Chapter 7**, contains a summary of the study and suggestions for future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

This chapter aims to present an overview of researched problem domain; Location Based Social Networks and Link Prediction Problem. It also includes the combined summaries for related studies in literature together with their approaches for similar problems in similar domain.

2.1 Location Based Social Networks

Graphs are used to represent relationship among entities. Generally, nodes represent entities and edges are for relationships among them. Edges are:

- directed when relationship is unidirectional (one-way) like file transfer, or
- undirected when relationship is bidirectional (two-way) like handshaking.

Social networks are for representing relationships between individuals and organizations. They have edges and nodes, but there is additional social information stored on the graph items in order to represent social structure. Nodes and edges contain various types of content which is specific to the domain of the social network.

At early days, social science researchers use the social networking concept to investigate the underlying rules and mechanism at societies. Moreover, artificial and natural theories for other scientific areas like economy, politics, history and etc. are studied by the help of social networks [13].

Computer technology and its products have become the key determiner of the way people live especially after 2000s. Enormous penetration of broadband internet made

us online and internet connected. Then, wide usage of internet connected mobile phones changed the way people behave and communicate with each other. At the early days of internet, users were mostly consumers of the content generated by others. By the help of 'Web 2.0' concept, users are enabled to produce content for web. Size of user generated mobile content exceeded others; especially after high ownership rates for a smart phone connected to internet. Online social networks are the computer technology products which have leaded such revolution.

Online social networks are growing dramatically with mobile and web software tools for almost all information technology environments [2]. These tools provide a very suitable environment to share multimedia information with other people on the network. People tend to be interactive with each other through such social networks for any step of their lives by shared photo, video, status, message, location information.

In online social networks, neighbor nodes are mostly called friends; such interactivity is established with the edges called friendship between that neighbor friend nodes. For example, we can define:

- Facebook as an online social network of friendship with undirected relationship between users, with various content types are supported for sharing,
- Instagram as an online social network of photo and video sharing of people, with directed relationship (following) between users,
- Twitter as an online social network of message (status) sharing of people, with directed relationship (following) between users.

In addition to well-known global social networks like Facebook, Twitter, Instagram with hundreds of millions of users (nodes), there are many smaller ones which are very popular within their own geographical scope like city, country, etc. Moreover, there are social networks for specific vertical communities based on shared interests and goals. Some networks are only for professionals or for organizations. This variety is based on the social content type stored and shared at nodes and edges of the network.

Location Based Social Networks (LBSNs), is one of the emerged online social net-

working concepts after wide usage of smart phones and location based services. One of the key contributions of smart phones to daily life is the available GPS. When combined with the internet access, GPS on mobile phones have made our life easier at many aspects through location-based services like LBSNs.

Some of the utilities that LBSNs enabled for people are as follows:

- Share their location with friends (usually called check-in),
- Find out local places of their interest,
- Get directions to specific location,
- Discover places with discounts or with special offers,
- Read and write comments, recommendations and ratings about places,
- Connect with their friends and
- Find nearby friends.

Foursquare, Gowalla and Facebook Places are some of fast growing LBSNs. Most of the LBSNs are designed based on following concepts:

1. **User Node:** Main nodes of the network which resembles a person using the LBSN. These nodes can contain various social content according to the LBSN.
2. **Friendship Edge:** Edge between user nodes of the network that resembles a social link between two friends. Generally, time information is stored on these edges.
3. **Place Node:** Secondary nodes of the network which resembles a place which can be visited by users. Descriptive information, multimedia content, user comments and ratings are some of mostly stored together with the spatial information for the place (latitude, longitude).
4. **Check-in Edge:** Edge between user and place nodes of the network that resembles the visit of the user to that specific place. It is generally created with user action and these edges include time information.

Check-in activity is the most important user generated content for LBSNs as it enriches the social interaction over the network by removing the gap with real physical world and online world. By that activity, when and where information of users' mobile trajectory is stored and shared with friends. Moreover, check-in is a way to generate recommendation/rating/multimedia content for places. That content is another useful service of LBSNs while user trying to discover a place of his/her interest. Moreover, services like place ranking, mobile marketing, traffic forecasting and path planning benefit from the check-in related content [14].

Rapid growth of those social networks also created a tremendous amount of user social data with friendships, geographical trajectories. That data is full of challenges and opportunities for researchers that analysis social networks to investigate social, temporal and spatial points of mobile behavior for humans.

2.2 Link Prediction Problem for Location Based Social Networks

In this research, we studied link prediction problem for online location based social networks. As explained on previous sections, LBSNs contains social links between its users (nodes) which is generally called friendship. Those links resemble the friendship between two people in real world.

Social communities are known to be very dynamic in real world by addition and removal of friendship between people. SNA researchers have been trying to formulate a model to predict that additions and removals of the social links with the social network information and historical data. That SNA problem of finding future social link (edge) possibility between two graph node is called **"Link Prediction (LP)"**. LP problem in a social network can be defined as follows:

> *Given a social network at a time t, predict new edges to be added to the network at some future time t' [4].*

Link prediction is an important challenge for online social networks considering the dynamic social link evolution of real world and the motivation of keeping their network edges up-to-date. Amount and quality of user generated content is directly

dependent to the existence of correct social links (friendships) between the users. Therefore, social networking sites are providing link recommendation tools by predicting the possible link as follows:

- people you may know in Facebook,
- who to follow in Twitter and
- friend suggestions of Foursquare.

Existing studies on link prediction defines the problem as two-class classification problem. If there is not any existing friendship edge between two user nodes, those users can be chosen as social link candidates for the classifiers. Classifiers are trained with the features calculated from the social network data. They usually make use of statistical and topological features as measurement parameters, together with domain related custom ones. Recommendation tools designed for OSNs have size limitations as a result of the big data and time cost concerns; therefore, high precision is targeted for the classifiers. There are various applications that benefit from such prediction approach like anti-terrorism and marketing [15].

Link prediction for LBSNs are very similar to pure link prediction problem, only difference is the available usage of **Place nodes** and **Check-in edges**. Those additional data enabled some new interaction features which are calculated from time, location and frequency information retrieved from users' check-ins. Temporal and spatial information of check-in define candidate user's mobile behavior [16]. Geographical distance, mobile behavior similarity and common place information are important for SNA researchers and they can be used as features to train classifier for LP for LBSNs.

2.3 Related Work

2.3.1 Link Prediction Researches

One of the main researches on link prediction problem on social networks was made by Liben-Nowell and Kleinberg. They summarized comprehensive subset of features which were calculated using graph topology [4].

Generally accepted approach is solving the link prediction problem as 2-class classification where you can define 2 classes as: existence of link (friendship), or not. Most of the good performing results in the field were due to supervised learning techniques on this classification problem [17] [18].

Hasan et al. made a comprehensive research for analyzing supervised machine learning algorithm performances on a link prediction problem. In 2 different networks, SVM, Decision Tree and Bagging (multi classifier fusion) algorithms had good performances where SVM overperformed all others. Moreover, they analyzed features to make ranking and elimination. Their results depicted that network distance was most critical feature [18]. Such a feature analysis will be very beneficial especially in decreasing the prediction time without losing much from prediction rate.

Davis et al. studied link prediction in heterogeneous networks and their results depicted how supervised approach over-performed comparing with unsupervised approach. [19]

Fire et al. studied computationally efficient topological features to perform link prediction in social networks. They showed the time cost difference at the LP problem based on the selected configurations. [20]

Lee et al. studied efficiency in link predicted. They used computationally less expensive features but still got high performance by predictor. [21]

As described before, there are lots of usage types for social networks. There is a research on biological gene networks [22]; which also studied link prediction. They used main classification algorithms like naïve bayes classifier and decision trees with a problem specific feature selection. Rather than predicting future links that will be created, they also predicted the existing links that will be deleted in a network.

Alan E. Mislove studied [23] SNA methods from different perspectives to design information systems for special purposes. One of the main contributions of his study is about the growth of OSNs. He used empirical features rather than using only theoretical models like popularity and network distance. He analyzed Flickr, YouTube and Wikipedia data to make better predictions using specific proximity features (like destination indegree) and results overperformed classical methods.

Zhu et al. [24] also studied link prediction problem on different kinds of social networks like SBLP, Wikipedia and IMDB. In order to use related success on specific sub-problem, they proposed a hybrid solution framework for social networks:

- In time based model, they used historical network data with a factor of passed time (bigger factor for sooner),
- In probabilistic relational model, they used Relational Markov Networks to make dynamical prediction model from specific feature data (time or sub-network).

Bliss et al. proposed an evolutionary algorithm approach on the same problem. They used Twitter reply data to predict feature links between users. In their research, 16 well known topological measurements are used to represent a solution [25]. By giving random multiplier to each feature a solution instance is created in that study. By using 100 initial seed instances in a *Covariance Matrix Adaptation Evolutionary Strategy* algorithm, they determined best solution instance by evolutionary population generation and fitness function. Their approach seems to be a generic solution for most of the social networks. However, it lacks lots of problem specific feature and measurement methods because of strict generic 16 features and simple algorithm.

Shin et al. proposed a hierarchical (multi-scale) solution for link prediction, that work with low time requirements in social networks with huge data. In their research [26], data is hierarchically clustered top-down and then their algorithm makes an approximate calculation based on measurements at each level of hierarchy. Prediction time for large scale data could be decreased by that approach.

Temporal information existing in social network as history was used by Soares et al. [27] to propose a solution for predicting future links. As temporal data defines the network evolution model, it would be also a good feature for future links that is also an evolution. They used rewards for temporal events according to their effect on future link association and got a stable prediction model. Rather than using actions of users in future link, they accounted dynamics of common neighborhood when applicable. Usage of higher multiplier for recent activity also performed well.

Usage of timestamps in social network's historical data is crucial for getting better prediction rates. One of the researches emphasizing this is done by Tylenda et al.

from Saarbrücken [28]. Latest and frequent social interactions of an actor in the social networks are used for defining neighborhoods and similarity. When those edges weighted higher, their prediction rates over performed as expected. The main reason of such improvements comes from the dynamic model of societies that is also time aware.

A group from The University of Notre Dame Computer Science Department studied link prediction problem with some new approaches to solve sub problems like variance reduction and sampling issues. That problem is mostly faced when dealing with large social networks where possible feature links (not existing now) are much more than realized ones. They applied oversampling but did not perform well. Then, usage of under sampling by selecting a special group for training (sub-network or neighborhood) performed better. In addition, they advised usage of Hellinger Trees (skew insensitive classifier) within an ensemble [29].

Domain independent strategy for selecting features is modeled by Bao et al. [30] using "Principal Component Analysis". Their research also performs better on sparse datasets where positive and negative links have a strong imbalance where most of the supervised learning algorithms flunk.

2.3.2 Link Prediction Researches for Location Based Social Networks

A group from Cambridge University Computer Laboratory focused on analysis of location based social networks [31]. They used a data set that contains temporal information about created social ties and made place visits (check-ins). That dataset, which includes longitudinal data for 4 months, was collected from Gowalla. Growth of network is modeled by combining social and spatial factors. Their research on defining main factors of social tie creation and social interaction initiation showed that only social relationships like friend-of-friends, common friend count, indegree/outdegree ratio were not enough for modeling the growth of social networks. They propose that place friends (people that has common interest and related check-in on same places) are also good candidate for new social tie even if they do not share any social feature (friend, etc.). In addition, their results depicted that speed of network growth for a node is directly dependent to the time spent on that social network.

Zhiyuan Cheng et al. studied geospatial footprint analysis for LBSNs. For data collection, they performed an analysis on public Twitter data to filter only LBSN created tweets. They calculated a home location for each person from the coordinates of check-ins, using recursive grid search [32]. They also performed an analysis of user behavior during day and during week according to temporal information. Human mobility pattern is researched by making spatial calculations like "radius of gyration", "user displacement" and "returning probability". Their results depicted followings:

- People with higher social influence (follower/following rate) has higher radius of gyration.
- People from denser cities (like metropolitan) have much more often displacement (travel).
- People are more prone to make negative comments on location based social networks.

A group from Arizona State University Computer Science and Engineering Department also made researches on analysis of LBSNs [33]. Their motivation was to model the location sharing action of a person using social and historical ties. Any historical data would be good estimation parameter for future as such social interactions in history are the roots of future social interaction model. They used Twitter public status updates to collect location based network data of Foursquare. They proposed a social historical model for predicting check-ins by the help of two foundlings:

- There is a power-law distribution on location sharing of users. (less places with frequent check-in and more places with few check-ins)
- People tend to share location on places after another sharing another place as combo; like having lunch and going to coffee shop. (short-term effect)

One of the link prediction approaches using location data effectively is proposed by Scellato et al. [34]. They used the Gowalla dataset with 4 months data in their research. Their contribution is on the selection of possible link candidates on social networks. Only 2 people with a common friend or 2 people with a common check-in

place (same location shared by both) are selected as link candidates. Such an effort makes data 15 times smaller while losing only 1/3 of new links. Additionally, they used location based social activity to generate new features for their supervised machine learning algorithms like location entropy, common places and place distances. Their study can be increased by more systematic usage of such user social actions especially with a time-based hierarchy.

Zheng et al. from Microsoft Research Asia studied LBSNs to recommend place and friend [35]. Their study used a hierarchical similarity measurement based on users' longitudinal data history. Some regions are defined dynamically from the historical data and users' movements and users are described by that regions and sequences of those regions. They also consider the importance of a location for creating a friendship by analyzing the historical data for popularity, etc. They proposed a similarity measurement approach from that specifications and calculations.

Sun et al. studied social networks with multi typed relations [36] where some of relationship types can be inferred from others. To handle such variance in types they proposed a meta-path based approach for defining features. Their main contribution on the research is extending the linear prediction problem definition with a new consideration, creation time prediction of that future link.

A group from Carnegie Mellon NASA Ames Research Park studied the link prediction problem for LBSNs. They emphasized on data analysis like feature selection using location information of social interactions and threshold usage on the selection of possible future links. Their results depicted the impact of using such data pre-processing and interaction thresholding (timestamp and frequency based) [37].

Hristova et al. composed a multilayer social network using Twitter and Foursquare to solve link prediction in LBSNs [16]. They used cross network data and made cross network predictions using social and geographic features to enhance individual network prediction scope.

Ye et al. tried to calculate a category distribution for each user to predict his next check-in [38]. Their results depict that category of places helped them to understand user-preference for predicting his next action.

2.3.3 Optimal Feature Subset Selection Research

Optimal feature subset selection is shown to be NP hard problem [39]. Thousands of researches have been conducted to help on that problem by leveraging statistics, math and information theory. Feature selection and feature reduction are two different sub-problems on this area.

Dash and Liu prepared one of first comprehensive survey for feature selection mechanisms. They presented an overview of studies since 1970s and proposed a generic model of feature selection methods with four main steps:

- **Generation:** This step is for producing and determining subset candidates. That production can be performed as a complete, heuristic or random search.
- **Evaluation Function:** This step is for creating a score for subsets based on goals. Distance, information, dependence or consistency metrics can be used as evaluator. Moreover, real classification error rate can be used for high accuracy requirements with huge time usage.
- **Stopping Criterion:** This is the criterion to halt the subset generation and evaluation loop,
- **Validation:** This step is for checking the validness of the selected subset.

They identified all different possible four step combinations from literature and analyzed their probabilities and potential usages. Guidelines and checklists were described for domain specific feature selection method designers [40].

Yu and Liu from Arizona State University analyzed features to create a selection mechanism based on relevance and redundancy [41]. They proposed a framework that uses relevance and redundancy of features in the evaluation function. Firstly, features are divided into three relevance groups (strongly relevant, weakly relevant, and irrelevant) based on the correlation metric relevance. Fast Correlation-Based Filter (FCBF) algorithm is presented that selects pre-dominant features by removing redundant features from relevance groups.

Battiti's research is one of earliest studies that benefit from mutual information of two

features. Information gain based calculations are performed for features. A greedy algorithm selects features one at a time by considering the new feature's mutual information with already selected ones. Most informative new feature is added to the subset [42].

A faster feature selection based on mutual information is offered by Fleuret [43]. This research proposed a iterative feature selection very similar to the Battiti's work. However, a conditional elimination is applied for the new features added based on their similarities with any of previously selected features. Similar features are determined based on the minimal mutual information of that pair.

Wang et al. also used conditional mutual information for text categorization where they benefit from detection of non-informative (redundant) features' removal [44].

Peng et al. also used mutual information estimation as a core of feature selection. [45] Maximizing relevance and minimizing the redundancy, which is called mRMR, is used for the ranking the features. One by one addition of features to subset with mRMR is proven to be equal to select maximum dependent subset. Their first step of feature selection is making a ranked list of features by the help of mRMR and cross validation classification error as evaluation metric. Second step is for deciding the number of features to use from the ranked list to optimize the goal. A wrapper feature selector is defined in two schemes:

- Backward Wrappers removes one redundant feature from subset to lower error rate in each step.
- Forward Wrappers incrementally adds a feature to subset in each step which lowers error rate.

Genetic algorithm has been one of the mostly used methods in feature reduction. Prakash and Murty leveraged genetic algorithm in feature reduction and got good solutions by fast and robust genetic algorithm [46]. They used nearest neighbor classifier accuracy as the fitness function criterion of evolutionary search where each binary gene of a solution resembles a feature. Gene with value 1 is usage of that feature in subset and 0 is the opposite. Single point crossover and random mutation is used on this research.

A hybrid solution for feature selection is proposed by Gheyas and Smiths [47]. They proposed an algorithm SAGA that combines Simulated Annealing (SA), Genetic Algorithm (GA), a generalized regression neural network and a greedy search algorithm. SA is used to guide the search for global solution space. Then, an initial population for GA is created from best solutions find by SA. A greedy algorithm, hill-climbing feature selection, is applied to perform a local search on k-best solutions given by SA and GA. Best neighbor is selected according to the euclidean distance. Much smaller number of features are selected by SAGA in less time.

Research of Xu and Rockmore is very similar to our research as they try to optimize feature subset for link prediction as we do. They proposed an algorithm that ranks the features and then applies weight to them. Information gain is used as a discriminative ability metric is calculated for each feature. Mutual information of any two features are used as the redundancy metric. Then two different approaches are researched for link prediction:

- **Feature Ranking Method:** An iterative greedy search is used to choose features for the optimization problem that maximizes the discriminative power and minimizes the redundancy of feature subset.
- **Feature Weighting Method:** An active set method of Dantzig and Wolfe is used to solve the same optimization problem by assigning weights to the features.

Their methods over-performed and feature ranking method reaches higher classification accuracy for their experiments on Link Prediction [48].

Most of the related researches applied iterative algorithms that select features one by one. A study of Xu et al. proposed a non-monotonic feature selection to avoid possible lacks in iterative ones [49]. A Multiple Kernel Learning (MKL) is used for proposed non-monotonic feature selection's combinatorial optimization. Original problem is relaxed into a convex optimization problem. As it can be solved efficiently by expressing it as a Quadratically Constrained Quadratic Programming, a solution is proposed for feature selection. Their performance is much better than incremental ones for some datasets.

Mitra et al. studied feature similarity based selection. A new metric called *Maximal Information Compression Index* is presented. They calculate this metric from covariance matrix and helps the feature selection on not inserting a new feature which is very similar to any of previously selected features [50].

One of few studies that try to cluster features for feature reduction is performed by Butterworth and et al. [51]. They worked on hierarchical clustering of features by a defined metric. A cluster tree, dendrogram, is generated after the clustering and feature subset is selected from the clusters of that dendrogram. Features which are located at the cluster centers are chosen.

Song et al. proposed a two step algorithm to make subset selection faster. First, features are clustered into clusters. Then, feature with the highest information gain is chosen from each cluster to determine feature subset. A graph is constructed from features where nodes are features and edges between them are correlation value of related node pair. A minimum spanning tree (MST) of that graph is determined using Prim's algorithm. MST is divided into sub-trees using the relevance values. Each sub-tree becomes a cluster of features. Finally, a feature from each sub-tree is selected for the subset by choosing the greatest relevance valued feature [52].

A study of Appice et al. focused on the feature reduction by pairwise analysis of feature pairs and their logical redundancy. Determining useless features and optimizing the classification coverage are two successful approaches they have evaluated [53].

CHAPTER 3

OUR RESEARCH FUNDAMENTALS

This section presents fundamental details of our research about link prediction problem for LBSNs. Problem and solution formulation, research data and developed link prediction framework are described in detail.

3.1 Problem and Solution Formulation

In this study, link prediction problem is solved as classification of the future link status for potential link candidates. Two-class classification is modeled for the research as there are two cases:

- **1 (positive)** corresponds to the existence of future link between candidates and
- **0 (negative)** corresponds to no future link between them.

Supervised machine learning algorithms are used for solving this classification problem by using features extracted from the LBSN. Features and their formulas were defined by social network data analysis literature.

In order to use temporal LBSN data in binary classification solution for link prediction problem, we determined a **Prediction Period Beginning Date (PPBD** from the timestamps covered in the research data. Potential link candidates for classification are decided based on the LBSN data status in the PPBD, pairs with existing social link (friends) are ignored and not considered as a link candidate pair. LBSN data (friendship, check-in) that were created before that date are used to calculate features of potential link candidates. LBSN data that were created later is used as ground truth

of class label which was decided based on whether a link will be created between potential link candidates or not.

As we know that positively classified pairs are used for recommendation tools by LBSNs, we focused on correctness of the positive link prediction in our research. Therefore, we searched for contributions that will improve link prediction performance especially the classification correctness of positive future links.

3.2 Data

In this thesis, we use a dataset provided by Cambridge University Computer Laboratory [34]. It was collected from a well-known LBSN, Gowalla. Existing data contains the following schema:

- Friendships (1508860 unique)
  - Edge or social link between 2 user nodes of the network.
  - Timestamp of creation and ids of users are provided.
- Places (1542003 unique)
  - It includes address as spatial data (lat, lon).
  - It includes category of the place. Category is one of available 283 categories.
- Check-ins (10062916 unique)
  - It is an edge / interaction between a user node and a place node of network.
  - Timestamp of creation and ids of user and place are provided.

In the original dataset, users were not directly given as a separate table for nodes. By mining the friendship table, we constructed one table for 163487 unique users. Moreover, we analyzed the edge direction for friendships. As we see that, over 99 percent of the social links (friendships) have the counterpart link with opposite direction. Therefore, we decided to study the link prediction in LBSNs as an undirected link prediction between any two nodes in any order.

Researched LBSN data was collected from daily snapshots between May, 4th and August, 19th 2010. We decided to use August 1st 2010 as the *'prediction period beginning date'*. Considering all unique users, there were billions of possible link candidate pairs. To be able to be agile and be able to try new things faster, we required to have a reasonable sized data. We decided to have a sampled subset of pairs to use throughout our research.

- After analyzing the ratio of newly established links; realized that near 88 percentage of them was between two candidates who are either friend of friends or between two candidates who have a common place that both made check-in before (called place-friends).
- Decided to consider only candidates which are friend-of-friends or place-friends or both. This also helped for reducing data imbalance effectively.
- We randomly selected a subset (200K) of link candidate pairs from the ones which satisfy this constraint. as training data set to achieve results faster.

3.2.1 Data Enhancement

During literature survey, we observed the importance of place related attributes in the feature set. We enhanced place information with a new property called "entropy". This was used earlier [34] to make use of check-in information for predicting the possibility of two candidates becoming friends after being at the same place.

To evaluate the value of entropy for researched problem, lets consider 2 places: a small coffee shop and an airport. Since an airport is a public place with huge number of visitors, we expect that there would be much more check-in's at the airport than the coffee shop. However, when we analyze individual visitor data we can see that the ratio of that visitor's check-in count to his/her all check-in count is much higher at places like coffee shops.

For each place in the network, we calculate and store the entropy value using **equation 31**:

$$E_k = -\sum_{u_i \in V_k} q_{ik} * \log q_{ik} \tag{31}$$

Where E_k is the entropy for place k, V_k is the set of all visitors of place k, u_i is a visitor of that place and q_{ik} given in **equation 32**:

$$q_{ik} = \frac{c_{ik}}{C_k^P} \tag{32}$$

Where q_{ik} is a user-place metric, c_{ik} is total number of check-ins of user i at place k and C_k^P is the total number of check-ins at place k.

3.2.2 Data Groups

For better analysis of contextual information, we distributed the candidates into groups according candidates' latest condition before prediction period beginning date:

- **Friend-of-Friends Group (FOF):** This sub-group includes all pairs that have at least one common friend.
- **Place-Friends Group (PF):** This sub-group includes all pairs that have at least one common place that both made check-in previously.
- **Both-Friends Group (BF):** This sub-group is the intersection of "FOF" and "PF" groups.
- **Whole Group (WG):** All candidate pairs are included in this group

Those groups are not fully disjoint between each other. **Table 3.1** includes the data details for each group.

Candidate count is the total number of user pairs that we will use through training and test phases of our supervised learners.

Candidate count percentage is the ratio of candidate count for related data group to the candidate count for whole group (all candidates).

Positive link percentage is the ratio of the count for pairs that became friends to the count for all pairs in related data group.

Table 3.1: Data Groups' Statistics

Data Group	**Candidate Count**	**Candidate Count Percentage** (%)	**Positive Link Percentage** (%)
Friend-of-Friends Group (FOF)	78506	39.25	0.40
Place-Friends Group (PF)	128861	64.43	0.19
Both-Friends Group (BF)	7367	3.68	2.51
Whole Group (WG)	200000	100	0.19

3.3 Link Prediction Framework

A Link Prediction Framework is developed for this research to handle complete process of link prediction using Java programming language. We used our framework from initial input level of LBSN research data to the final prediction outputs and evaluation. All features, learning algorithms, feature selection mechanisms to improve link prediction performance are also implemented and analyzed on top of this framework.

Our link prediction process can be presented through following units of the framework:

3.3.1 Feature Extraction

This step is for preparing the required data format for the supervised learning algorithms (classifiers). Using the features formulas from proposed and literature, feature scores/values are calculated for every candidate pairs.

Our dataset is quite large and calculating all features requires a lot of time. To over-

come that problem; we developed a distributed feature extractor that maximizes the parallelization, caching and sharing of the calculated values of features. For example; feature formulas are dependent to metrics of user interactions in LBNS. Those metrics could be used by more than one feature without recalculation in the framework.

In each experiment, different subset of features is needed and framework handles to form the correct feature set for machine learning training and testing.

3.3.2 Supervised Machine Learning

Most of the successful link prediction results were obtained by using supervised learning algorithms. In our framework, core algorithms are used from the WEKA [54]. Training of supervised learners are performed by using the calculated feature scores for labeled candidate pairs. Trained learning models are used as predictor through the classification of test candidates.

Throughout this research, one or more of below classifiers are trained to create link predictors:

1. **Naïve Bayes Classifier (NBC)** [55],
2. **Bayesian Network (BN)** [56] and
3. **Random Forest (RF)** [57].

NBC and BN are methods assumes an underlying probabilistic model for diagnostic and predictive problems. They are named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Those learners are robust to noise in data.

Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms.

Moreover, we applied RF (with 40 trees, 4 features each) to evaluate learner performances. NBC, BN and RF are known to be fast and good performing predictors.

3.3.3 Predictions and Performance Evaluation

Using the trained classifiers, unlabeled link status for candidate pairs are predicted. As given previous sections, LBSNs focus on correctness of positively predicted social links as their goal is to optimize recommendation of friendships. For each data group; different prediction results are experimented and analyzed. Performance improvements with new propositions are also evaluated according to results obtained at this unit of the framework.

We applied a 4 fold cross validation in our prediction evaluations and presented the averaged results for comparison.

To decide on the evaluation metric for LP performance, following domain-specific requirements are considered:

- Highly imbalanced dataset, with much more negative instances for classification
- Similar studies [58] use area under curve (AUC) of receiver operating characteristic (ROC) curve as performance metric **(ROC-AUC)**
- Candidate pairs with higher potential of being friends are used for recommendation tools by LBSNs. Therefore, we focused on correctness of the positive link prediction (true positive) in our study, and we searched for contributions that improve link prediction performance from that perspective.

We chose ROC-AUC as the performance evaluation metric throughout the whole study. ROC-AUC met all requirements and goals.

ROC is drawn using the true positive rate together with the false positive rate at all classification thresholds [58]. Its AUC is equivalent to the probability of a randomly selected positive instance appearing above a randomly selected negative instance [29].

CHAPTER 4

CONTEXTUAL FEATURE ANALYSIS

This chapter provides the details of collected feature set from comprehensive literature search together with our proposed contextual features.

4.1 Features from Literature

We reviewed most of the related studies in literature and identified successful and commonly used features for link prediction in social networks. Some studies focus on the network structure based topological features while others attempt to make use of location based social networks' check-in data.

This section describes the contextual definition and detailed mathematical formulation of all reviewed and used features from literature in linked prediction problem for LBSNs.

Feature information is divided into two subsections that contain the topological features and the interaction features, respectively. we utilized both topological features and interaction features to make use of all available information in data to better estimate links.

4.1.1 Topological Features

Features calculated from social networks by using graph theory concepts are combined as topological features. Below list covers mainly used topological feaures in link prediction problem:

1. **Common Friend Count (CFC)**

   - *Description:* It is total number of people who are friends for both 2 link candidate nodes.
   - *Formula:* $CFC(x, y) = |\alpha_x \bigcap \alpha_y|$
     where α_i is the friend list of user i.

2. **Common Friend Ratio (CFR)**

   - *Description:* It is the Jaccard ratio [59] of common friends' count to total friends' count for 2 link candidate nodes.
   - *Formula:* $CFR(x, y) = \frac{CFC(x,y)}{|\alpha_x \bigcup \alpha_y|}$

3. **Preferential Attachment Friends (PAF)**

   - *Description:* It is the product of total friend counts for 2 link candidate nodes [60].
   - *Formula:* $PAF(x, y) = |\alpha_x| * |\alpha_y|$

4. **Adamic Adar Common Friends (AACF)**

   - *Description:* It is the Adamic Adar score [61], calculated from common friends of candidates; based on the number of friends of that common friend.
   - *Formula:* $AACF(x, y) = \sum_{z \in (\alpha_x \bigcup \alpha_y)} \frac{1}{\log |\alpha_z|}$

5. **Shortest Path Distance (SPD)**

   - *Description:* It is the inverse (1/n) value of the shortest distance between 2 link candidate nodes.
   - *Formula:* $SPD(x, y) = \frac{1}{\forall_p min(distance(p))}$
     where $p \in pathsBetween(x, y)$

6. **Total Shortest Paths (TSP)**

   - *Description:* It is the total sum of inverse values of path distances between 2 link candidate nodes [62]. Only paths with length 2 or 3 are considered.
   - *Formula:* $\sum_{p \in pathsBetween(x,y)} \frac{1}{distance(p)}$
     where $2 \leq distance(p) \leq 3$

4.1.2 Interaction Features

Interaction features, which are calculated from the check-in interaction of LBSN users to predict links, can be listed as follows:

1. **Common Place Count (CPC)**
   - *Description:* It is total number of places which both 2 link candidates have check-in.
   - *Formula:* $CPC(x, y) = |\beta_x \bigcap \beta_y|$
     where β_i is the place list of user i where he/she made check-in.
2. **Common Place Ratio (CPR)**
   - *Description:* It is the Jaccard ratio [59] of common places count to total check-in places count for 2 link candidates.
   - *Formula:* $CPR(x, y) = \frac{CPC(x,y)}{|\beta_x \bigcup \beta_y|}$
3. **Preferential Attachment Places (PAP)**
   - *Description:* It is the product of total different places count for 2 link candidate nodes [60].
   - *Formula:* $PAP(x, y) = |\beta_x| * |\beta_y|$
4. **Adamic Adar Common Places (AACP)**
   - *Description:* It is the Adamic Adar score [61], calculated from common places of link candidates; based on the number of total check-ins (all users) at that common place.
   - *Formula:* $AACP(x, y) = \sum_{t \in (\beta_x \bigcup \beta_y)} \frac{1}{\log |\beta_{C_t}|}$
     where C_i is the check-in list of all users at place i.
5. **Minimum Check-ins Count of Common Places (MCC)**
   - *Description:* It is the minimum value of check-in count for common places of link candidates.
   - *Formula:* $MCC(x, y) = \forall_{t \in (\beta_x \bigcup \beta_y)} min(|C_t|)$

6. **Adamic Adar Common Places Entropy (AACPE)**

   - *Description:* It is the Adamic Adar score [61], calculated from common places of link candidates; based on the entropy value of that common place.
   - *Formula:* $AACPE(x,y) = \sum_{t \in (\beta_x \bigcup \beta_y)} \frac{1}{Entropy(t)}$
     where Entropy(i) is calculated score of place i from all check-ins at that location C_i.

7. **Minimum Entropy Value of Common Places (MEV)**

   - *Description:* It is the minimum value of calculated entropy values for common places for 2 link candidate users.
   - *Formula:* $MEV(x,y) = \forall_{t \in (\beta_x \bigcup \beta_y)} min(Entropy(t))$

8. **Preferential Attachment Check-ins (PAC)**

   - *Description:* It is the product of total check-in counts for 2 link candidate nodes [60].
   - *Formula:* $PAC(x,y) = |\gamma_x| * |\gamma_y|$
     where γ_i is the check-in list of user i.

9. **Common Place Check-in Counts Product (CPCP)**

   - *Description:* It is the summation of products (dot product) where each product is calculated with check-in count for 2 link candidate users at specific location.
   - *Formula:* $CPCP(x,y) = \sum_{t \in (\beta_x \bigcup \beta_y)} |\gamma_{x,t}| * |\gamma_{y,t}|$
     where $\gamma_{i,m}$ is the check-in list of user i at place m.

10. **Common Place Check-in Counts Product Ratio (CPCPR)**

    - *Description:* It is the ratio of common place check-in count product [34] to the product of all check-in counts of users (cosine similarity of two users' common place check-in count vectors).
    - *Formula: CPCPR(x,y) =* $\frac{CPCP(x,y)}{\sqrt{\sum_{t \in \beta_x} |\gamma_{x,t}|^2 * \sum_{t \in \beta_y} |\gamma_{y,t}|^2}}$

11. **Distance Of Most Visited Homes (DMVH)**

- *Description:* A home location is determined from most visited place for each candidate and distance between those home values is calculated for that link candidate pair.
- *Formula:* $DMVH(x, y) = distanceBetween(MVH_x, MVH_y)$ where MVH_x is the home location of user I where he has max check-in.

12. **Distance Of Most Visited Homes Ratio (DMVHR)**

- *Description:* A ratio of 'Distance Of Most Visited Homes' to the product of check-in counts at that value.
- *Formula:* $DMVHR(x, y) = \frac{DMVH(x,y)}{\left|\gamma_{x,MVH_x}\right| * \left|\gamma_{y,MVH_y}\right|}$

13. **Distance Of Lat-Lon Homes (DLH)**

- *Description:* A home location is calculated (average) from lat-lon information of each check-in for each candidate and distance is calculated for that lat-lon coordinates (homes).
- *Formula:* $DLH(x, y) = distanceBetween(LLH_x, LLH_y)$
  where LLH_i is a lat-lon position average home location by $LLH_i = \frac{\sum_{m \in \gamma_i} (m^{lat}, m^{lon})}{|\gamma_i|}$

14. **Sum of Radius Length from Lat-Lon Homes (SRLH)**

- *Description:* A radius is calculated from each candidates' each check-ins' distance to the determined lat-lon home. (average of all distances) Then a feature is defined by summing the radius of both candidates.
- *Formula:* $SRLH(x, y) = rlh_x + rlh_y$
  where rlh_i is the radius of user i calculated by
  $rlh_i = \frac{\sum_{m \in \gamma_i} distanceBetween(m, LLH_i)}{|\gamma_i|}$

4.1.3 Performance Evaluation

To analyze features in an isolated way, we decided to start from using single features for training link predictors. Classifications for each data group are made separately.

4.1.3.1 Friend-of-Friends Group (FOF) Performances

Table 4.1 shows the performance (ROC-AUC value) of existing literature features with all 3 classification algorithms:

4.1.3.2 Place-Friends Group (PF) Performances

Table 4.2 shows the performance (ROC-AUC value) of existing literature features with all 3 classification algorithms:

4.1.3.3 Both-Friends Group (BF) Performances

Table 4.3 shows the performance (ROC-AUC value) of existing literature features with all 3 classification algorithms:

4.1.3.4 Whole Group (WG) Performances

Table 4.4 shows the performance (ROC-AUC value) of existing literature features with all 3 classification algorithms:

Table 4.5 includes link prediction performances (ROC-AUC) when all topological features from literature are used on each data group.

Table 4.6 includes link prediction performances (ROC-AUC) when all interaction features from literature are used on each data group.

Table 4.7 includes link prediction performances (ROC-AUC) when all features from literature (20) are used on each data group.

Table 4.1: FOF - Literature Features' Single Performances

Feature	**NBC**	**BN**	**RF**
CFC	0.737	0.818	0.788
CFR	0.519	0.512	0.546
PAF	0.625	0.718	0.503
AACF	0.930	0.925	0.455
SPD	0.500	0.500	0.500
TSP	0.698	0.816	0.712
CPC	0.617	0.750	0.734
CPR	0.734	0.748	0.443
PAP	0.599	0.689	0.406
AACP	0.759	0.742	0.450
MCC	0.476	0.759	0.567
AACPE	0.760	0.738	0.468
MEV	0.728	0.759	0.505
PAC	0.605	0.686	0.399
CPCP	0.700	0.751	0.713
CPCPR	0.720	0.746	0.365
DMVH	0.684	0.729	0.521
DMVHR	0.695	0.718	0.513
DLH	0.674	0.749	0.511
SRLH	0.626	0.609	0.472

Table 4.2: PF - Literature Features' Single Performances

Feature	**NBC**	**BN**	**RF**
CFC	0.791	0.853	0.836
CFR	0.846	0.840	0.470
PAF	0.733	0.766	0.456
AACF	0.856	0.838	0.508
SPD	0.875	0.841	0.875
TSP	0.826	0.867	0.770
CPC	0.704	0.725	0.707
CPR	0.439	0.521	0.509
PAP	0.584	0.633	0.510
AACP	0.737	0.862	0.526
MCC	0.907	0.878	0.720
AACPE	0.762	0.880	0.468
MEV	0.918	0.898	0.571
PAC	0.600	0.654	0.520
CPCP	0.751	0.756	0.707
CPCPR	0.533	0.500	0.516
DMVH	0.812	0.780	0.537
DMVHR	0.671	0.783	0.530
DLH	0.765	0.716	0.514
SRLH	0.687	0.652	0.519

Table 4.3: BF - Literature Features' Single Performances

Feature	**NBC**	**BN**	**RF**
CFC	0.790	0.790	0.777
CFR	0.643	0.572	0.614
PAF	0.596	0.566	0.546
AACF	0.864	0.841	0.567
SPD	0.500	0.500	0.500
TSP	0.648	0.651	0.610
CPC	0.618	0.588	0.579
CPR	0.633	0.571	0.553
PAP	0.569	0.515	0.524
AACP	0.563	0.746	0.588
MCC	0.804	0.783	0.663
AACPE	0.591	0.732	0.568
MEV	0.825	0.788	0.645
PAC	0.573	0.530	0.551
CPCP	0.607	0.619	0.618
CPCPR	0.600	0.554	0.535
DMVH	0.753	0.686	0.547
DMVHR	0.543	0.707	0.529
DLH	0.737	0.672	0.561
SRLH	0.699	0.636	0.537

Table 4.4: WG - Literature Features' Single Performances

Feature	**NBC**	**BN**	**RF**
CFC	0.795	0.777	0.798
CFR	0.703	0.723	0.529
PAF	0.667	0.767	0.465
AACF	0.862	0.865	0.469
SPD	0.733	0.718	0.733
TSP	0.828	0.811	0.733
CPC	0.621	0.646	0.675
CPR	0.478	0.500	0.509
PAP	0.573	0.575	0.523
AACP	0.793	0.805	0.551
MCC	0.725	0.824	0.700
AACPE	0.803	0.810	0.518
MEV	0.727	0.827	0.578
PAC	0.588	0.584	0.502
CPCP	0.655	0.688	0.682
CPCPR	0.515	0.500	0.506
DMVH	0.670	0.684	0.536
DMVHR	0.677	0.702	0.526
DLH	0.641	0.648	0.517
SRLH	0.628	0.593	0.514

Table 4.5: Classification with Topological Features

Classifier	**FOF**	**PF**	**BF**	**WG**
Naive Bayes Classifier	0.918	0.896	0.822	0.875
Bayesian Network	0.929	0.900	0.841	0.880
Random Forest	0.796	0.759	0.826	0.737

Table 4.6: Classification with Interaction Features

Classifier	**FOF**	**PF**	**BF**	**WG**
Naive Bayes Classifier	0.770	0.901	0.810	0.822
Bayesian Network	0.784	0.919	0.827	0.847
Random Forest	0.722	0.751	0.753	0.700

Table 4.7: Classification with All Features

Classifier	**FOF**	**PF**	**BF**	**WG**
Naive Bayes Classifier	0.881	0.948	0.860	0.915
Bayesian Network	0.926	0.966	0.886	0.950
Random Forest	0.878	0.854	0.865	0.856

Results with the existing features from literature depicts that Bayesian Networks are best supervised learner for all data groups. Best classifier performances for all data groups are achieved when all features are used except the FOF group. For FOF data group, best classifier performance is achieved when only topological features are used.

4.2 Proposed Features

After collecting an extensive feature list from literature; we analyzed LBSN data and designed new features based on time, common friend detail and place category information of check-in data. Our primary aim was to make use of available information in LBSN data that cannot be utilized by the existing features. Experiments showed that using our proposed features improve the link prediction performance. We compared our results with best results in the literature on the same dataset [9].

We combined our proposed features with the well-known topological features and location specific interaction features to have a qualified full set of features to improve link prediction performance. This section covers the proposed contextual features and their impact within the full feature set.

4.2.1 Proposed Contextual Features

Time (temporal) and place (spatial) information of check-in data is combined to calculate **'Common Check-in Count (CCC)'** feature that tries to distinguish the cases that candidates see each other.

CCC is the total number of occurrences that both of the link candidates' check-in at the same place at around same time (within 1 hour) calculated using equation 41:

$$CCC(x,y) = \sum_{t \in (\beta_x \bigcup \beta_y)} \left| \gamma_{x,t} \bigcap \gamma_{y,t} \right| \tag{41}$$

Where $\gamma_{i,n}$ is the check-in list of user i at place n, β_i is the place list of user i (where check-in made) and intersection is based on the timestamp closeness (max 1 hour difference).

We also observed that existing features measuring the common friends of link candidates were not deep enough. Therefore, we also mined the network link data to calculate two new features for digging the level of friendship with candidates for the common friends; **'Total Common Friend Closeness (TCFC)'** and **'Total Common Friend Common Check-in (TCFCC)'**.

Main idea for such calculation is *If common friends are closer for the candidates, its effect on making new friendship is higher.*

TCFC is the summation of products where each product is calculated for a common friend of candidates. It is an *advanced topological feature* which uses the multiplication of the common friends' count of that common friend with each candidate. We can formulate as **equation 42**:

$$TCFC(x,y) = \sum_{z \in (\alpha_x \bigcup \alpha_y)} CFC(x,z) * CFC(y,z) \tag{42}$$

Where **CFC(a,b)** is the common friend count of users a with b and α_i is the set of friends of user i.

TCFCC is the summation of products where each product is the number of common check-ins of that common friend with each candidate (the same place, at around the

same time). *Time (temporal) and place (spatial)* information is used effectively for this feature. We can formulate it as **equation 43**:

$$TCFCC(x,y) = \sum_{z \in (\alpha_x \bigcup \alpha_y)} CCC(x,z) * CCC(y,z) \tag{43}$$

Using the check-in place and category data, we calculated **'Common Category Check-in Counts Product (CCCP)'** and **'Common Category Check-in Counts Product Ratio (CCCPR)'** features that mainly represent similarity of candidates according to their check-in habits (category of places). They are novel features in literature that benefit from the *spatial* information together with *category* type. Main idea for such calculation is *If two people have similar life-styles (similarity), the possibility for them to meet and become friends is higher.*

CCCP is the summation of products, where each product is calculated with check-in count of each user at specific categorized places (dot product) and **CCCPR** is the ratio of common category check-in count to the all check-in counts of users (cosine similarity). Below equations are the mathematical formulations of CCCP 44 and CCCPR 45, where ω_i is the set of categories checked-in previously by user i and $\phi_{i,n}$ is the check-in list of user i at places with category n:

$$CCCP(x,y) = \sum_{m \in (\omega_x \bigcup \omega_y)} |\phi_{x,m}| * |\phi_{y,m}| \tag{44}$$

$$CCCPR(x,y) = \frac{CCCP(x,y)}{\sqrt{\sum_{m \in \omega_x} |\phi_{x,m}|^2 * \sum_{m \in \omega_y} |\phi_{y,m}|^2}} \tag{45}$$

4.2.2 Performance Evaluation

To analyze features in isolated way, we decided to start from using single features for training link predictors. Evaluations for each data group are made separately.

Table 4.8: FOF - Our Proposed Features' Single Performances

Feature	**NBC**	**BN**	**RF**
TCFC	0.705	0.775	0.518
CCC	0.640	0.640	0.515
TCFCC	0.692	0.700	0.515
CCCP	0.649	0.695	0.504
CCCPR	0.637	0.704	0.447

Table 4.9: PF - Our Proposed Features' Single Performances

Feature	**NBC**	**BN**	**RF**
TCFC	0.671	0.795	0.509
CCC	0.668	0.666	0.496
TCFCC	0.718	0.729	0.513
CCCP	0.628	0.696	0.499
CCCPR	0.669	0.582	0.521

4.2.2.1 Friend-of-Friends (FOF) Group Performances

Table 4.8 shows the FOF group performance (ROC-AUC value) of our proposed features with all 3 classification algorithms:

4.2.2.2 Place-Friends (PF) Group Performances

Table 4.9 shows the PF group performance (ROC-AUC value) of our proposed features with all 3 classification algorithms:

4.2.2.3 Both-Friends (BF) Group Performances

Table 4.10 shows the BF group performance (ROC-AUC value) of our proposed features with all 3 classification algorithms:

Table 4.10: BF - Our Proposed Features' Single Performances

Feature	**NBC**	**BN**	**RF**
TCFC	0.656	0.673	0.495
CCC	0.659	0.658	0.509
TCFCC	0.721	0.759	0.568
CCCP	0.592	0.539	0.499
CCCPR	0.616	0.539	0.528

Table 4.11: WG - Our Proposed Features' Single Performances

Feature	**NBC**	**BN**	**RF**
TCFC	0.693	0.785	0.502
CCC	0.645	0.644	0.512
TCFCC	0.673	0.677	0.498
CCCP	0.689	0.604	0.506
CCCPR	0.622	0.649	0.504

4.2.2.4 Whole Group (WG) Performances

Table 4.11 shows the WG performance (ROC-AUC value) of our proposed features with all 3 classification algorithms:

4.2.3 Using Proposed Contextual Features with Existing Ones

To see the basic information provided by newly proposed features, we used them together with all existing (20) features together.

First, a predictor is trained with only existing literature features. Then, adding only one of the proposed new features to existing ones 5 classifiers are trained with 21 features. Finally, all features (25) are used to train and test the prediction performance. Classifications for each data group are made separately.

We also compared our best (all features using proposed contextual features) AUC val-

Table 4.12: FOF - Proposed Features' Performances with Existing Features

Features	**NBC**	**BN**	**RF**
Only ELF	0.881	0.926	0.878
ELF + TCFC	0.896	0.925	0.889
ELF + CCC	0.897	0.928	0.884
ELF + TCFCC	0.898	0.930	0.887
ELF + CCCP	0.897	0.923	0.881
ELF + CCCPR	0.882	0.924	0.880
ALL	0.908	0.930	0.891

Table 4.13: FOF - Prediction Performances Comparison with Other Research

Features	**NBC**	**BN**	**RF**
OTHER RESEARCH	0.882	0.918	0.880
OUR RESEARCH - ALL	0.908	0.930	0.891

ues with the AUC values calculated using the features from most successful research[34] in same problem, same data and same algorithm. Compared research is called as OTHER RESEARCH in related tables.

4.2.3.1 Friend-of-Friends (FOF) Group Performances

Table 4.12 shows the FOF group performance (ROC-AUC value) of our predictors when our proposed new features are used with Existing Literature Features (ELF):

Table 4.13 compares the FOF group performance for our research's all feature set with performance for features from best performing research in same field.

4.2.3.2 Place-Friends (PF) Group Performances

Table 4.14 shows the PF group performance (ROC-AUC value) of our predictors when our proposed new features are used with Existing Literature Features (ELF):

Table 4.14: PF - Proposed Features' Performances with Existing Features

Features	**NBC**	**BN**	**RF**
Only ELF	0.948	0.966	0.854
ELF + TCFC	0.952	0.967	0.873
ELF + CCC	0.951	0.967	0.873
ELF + TCFCC	0.955	0.968	0.875
ELF + CCCP	0.947	0.965	0.872
ELF + CCCPR	0.948	0.966	0.872
ALL	0.959	0.970	0.878

Table 4.15: PF - Prediction Performances Comparison with Other Research

Features	**NBC**	**BN**	**RF**
OTHER RESEARCH	0.948	0.955	0.866
OUR RESEARCH - ALL	0.959	0.970	0.872

Table 4.15 compares the PF group performance for our research's all feature set with performance for features from best performing research in same field.

4.2.3.3 Both-Friends (BF) Group Performances

Table 4.16 shows the BF group performance (ROC-AUC value) of our predictors when our proposed new features are used with Existing Literature Features (ELF):

Table 4.17 compares the BF group performance for our research's all feature set with performance for features from best performing research in same field.

4.2.3.4 Whole Group (WG) Performances

Table 4.18 shows the WG performance (ROC-AUC value) of our predictors when our proposed new features are used with Existing Literature Features (ELF):

Table 4.19 compares the WG performance for our research's all feature set with

Table 4.16: BF - Proposed Features' Performances with Existing Features

Features	**NBC**	**BN**	**RF**
Only ELF	0.860	0.886	0.865
ELF + TCFC	0.862	0.890	0.871
ELF + CCC	0.863	0.887	0.878
ELF + TCFCC	0.879	0.896	0.870
ELF + CCCP	0.862	0.886	0.867
ELF + CCCPR	0.862	0.886	0.869
ALL	0.881	0.898	0.880

Table 4.17: BF - Prediction Performances Comparison with Other Research

Features	**NBC**	**BN**	**RF**
OTHER RESEARCH	0.862	0.879	0.853
OUR RESEARCH - ALL	0.881	0.898	0.880

Table 4.18: WG - Proposed Features' Performances with Existing Features

Features	**NBC**	**BN**	**RF**
Only ELF	0.915	0.950	0.856
ELF + TCFC	0.917	0.950	0.858
ELF + CCC	0.917	0.951	0.859
ELF + TCFCC	0.922	0.954	0.866
ELF + CCCP	0.916	0.950	0.860
ELF + CCCPR	0.916	0.951	0.855
ALL	0.925	0.955	0.872

Table 4.19: WG - Prediction Performances Comparison with Other Research

Features	**NBC**	**BN**	**RF**
OTHER RESEARCH	0.927	0.941	0.839
OUR RESEARCH - ALL	0.925	0.955	0.872

performance for features from best performing research in same field.

4.2.4 Evaluation

Proposed features improve the link prediction especially when they contextually augment the link information required to make prediction. Common Check-In Count (CCC) and Total Common Friend Closeness (TCFC) are the most effective ones as social network information utilized by them is highly required and cannot be replaced with any existing/available features in literature. We can see that using a relevant and effective subset of features will improve link prediction performance of the classifier.

Comparison with the prediction performances calculated from features of most successful research also showed the information gain additional features. Even compared research features already have very high ROC-AUC performance, our features were better in 11 experiments of performed 12 experiments (4 dataset X 3 algorithm). At this band of AUC-ROC values, those constant performance improvements clear the value of proposed features and shows that they would be very helpful for the online location based social networks.

CHAPTER 5

FEATURE REDUCTION RESEARCH

In our research, we focus on improving LP performance to predict new friendships in LBSNs with feature-based approaches. We had proposed new features leveraging Friendship and Check-In subgraphs [9] and their different structure and characteristics to utilize the available information with new features [10]. After an extensive literature survey and evaluation of existing features [31, 34], we ended up with a combined core set of 25 features which are beneficial for LP performance.

Each of the feature in full feature set was relevant and non-redundant witch the classification individually by helping with information gain from friendship subgraph topology or from the check-in subgraph. However, they don't form the optimal feature set from both effectivity and efficiency manner when all used together.

Picking the optimal feature subset became another challenge for us which had a potential to improve the LP prediction performance with improved accuracy and decreased time cost [53]. Feature reduction research has the motivation to help this problem by eliminating the logically redundant features in multiple dataset.

This chapter covers the feature reduction efforts to find optimal feature subset. It presents details of proposed greedy method to enhance the efficiency and proposed clustering based method to enhance the effectivity of link predictors. Performance evaluations for both methods are shown in corresponding sections.

5.1 Greedy Feature Reduction Method for Link Prediction Efficiency

Throughout feature analysis, we had observed that higher number of features means higher time cost for both feature extraction and classification stages. In this section, we will show how we improved the speed of LP solution by optimizing the time cost on those stages by the help of a proposed greedy feature reduction method. We started with analyzing those costs and performances for individual features deeply with detailed measurements to give data driven decisions for speed.

5.1.1 Cost of Feature Extractions

We measured the cost of extracting each feature for 200000 candidate pairs (Whole Group (WG)) and come with metrics (in milliseconds) shown in the **Table 5.1**.

- Average: Average time cost for extracting feature.
- 50th Percentile: Time cost value which is the middle point of all cost calculations (median).
- 90th Percentile: Time cost value which is greater than the 90 percentage of all cost values.

5.1.2 Individual Link Prediction Performance of Features

Our goal is to keep accuracy high while improving the speed. To determine the feature value from accuracy perspective, we re-used the individual link prediction performances of each feature in full feature set for Whole Group (WG) candidate pairs for different algorithms.

Individual performances were given in previous section through **Table 4.4** and **Table 4.11**.

Table 5.1: Feature Extraction Costs for WG in Milliseconds

Feature	**Average**	**50th Percentile**	**90th Percentile**
CFC	101.18	93.30	107.48
CFR	59.71	52.90	72.56
PAF	57.17	47.93	78.33
AACF	654.44	104.93	1421.59
TCFC	1548.46	130.60	2425.29
SPD	13651.23	2049.48	43907.09
TSP	39452.34	19581.30	60809.31
CPC	151.39	56.10	306.26
CPR	46.54	9.68	88.99
PAP	30.82	6.60	80.36
AACP	369.33	174.23	793.43
MCC	396.58	176.60	892.85
AACPE	80.40	11.48	219.72
MEV	25.95	6.18	61.07
CCC	29.45	6.43	77.83
TCFCC	937.96	593.78	1403.09
PAC	87.50	12.05	185.20
CPCP	50.60	9.20	115.28
CPCPR	27.77	7.10	55.00
CCCP	8230.80	5169.58	14018.36
CCPR	97.96	30.80	206.16
DMVH	41.11	8.00	99.09
DMVHR	32.94	7.83	79.41
DLH	22.88	6.88	60.63
SRLH	15.41	6.35	41.47

5.1.3 Selecting an Efficient Feature Subset

Considering the individual feature costs together with the individual prediction performances, we decided to select a feature subset that will help us keep effectivity but will enhance the overall speed for the predictor. This will be achieved by reducing the extraction time cost of the unselected features together with reduced training and classification time costs. Below steps are used for such selection which was performed separately for each classification algorithm:

- First, apply a threshold for the features and pick the features with individual performance higher than median value, to a **picked subset**.
- Then, filter the highest costly (extraction time) features from the picked subset, top three expensive feature and related features are eliminated by this way.

This is a greedy method for selecting a feature subset with the goal of making faster classifiers for link prediction problem in LBSNs, with keeping the effectivity as high as possible. This reduction can be extended to complex algorithms with optimal performance. However, we picked the simple approach to show how we could use individual feature extraction costs and their performance to reduce the total time cost.

5.1.4 Performance Evaluation

Using the selected subsets, both prediction performance (ROC-AUC) and time cost are measured for all algorithms. Evaluations for each algorithm are made separately. To show the efficiency enhancements, we will measure the prediction performance and time cost with following picked feature subsets for all algorithms:

- **All Features (ALL):** Results are measured when all 25 features are extracted, trained and used for classification.
- **Only Successful Features (OSF) Subset:** Results are measured when only successful features are extracted, trained and used for classification. Those features are determined by comparing the individual prediction performance of each feature with the median of all feature performances.

Table 5.2: Measurements with NBC

Subset	**Total Feature Extraction Time (min)**	**Total Training Time (sec)**	**One Candidate Pair Feature Extraction Time (sec)**	**Prediction Performance (ROC-AUC)**
AF	220666.7	30.3	66.2	0.927
OSF	191000.0	22.4	57.3	0.923
OSLCF	11000.0	15.6	3.3	0.922

- **Only Successful Low Cost Features (OSLCF) Subset:** OSF subset is filtered and features which require high extraction cost (time) are eliminated from that subset.

5.1.4.1 Naive Bayes Classifier (NBC) Performances

Table 5.2 presents the Naive Bayes Classifier (NBC) performances and measurements for each subset.

Table 5.3 lists the picked feature list of each subset for the Naive Bayes Classifier (NBC).

5.1.4.2 Bayesian Network (BN) Performances

Table 5.4 presents the Bayesian Network (BN) performances and measurements for each subset.

Table 5.5 lists the picked feature list of each subset for the Bayesian Network (BN).

Table 5.3: Picked Feature Subsets for NBC

Feature	**NBC-OSF**	**NBC-OSFLC**
CFC	X	X
CFR	X	X
PAF		
AACF	X	X
TCFC	X	X
SPD	X	
TSP	X	
CPC		
CPR		
PAP		
AACP	X	X
MCC	X	X
AACPE	X	X
MEV	X	X
CCC		
TCFCC	X	
PAC		
CPCP		
CPCPR		
CCCP		
CCCPR		
DMVH		
DMVHR	X	X
DLH		
SRLH		

Table 5.4: Measurements with BN

Subset	**Total Feature Extraction Time (min)**	**Total Training Time (sec)**	**One Candidate Pair Feature Extraction Time (sec)**	**Prediction Performance (ROC-AUC)**
AF	220666.7	35.2	66.2	0.955
OSF	188333.3	23.0	56.5	0.949
OSLCF	11333.3	18.3	3.4	0.951

5.1.4.3 Random Forest (RF) Performances

Table 5.6 presents the Random Forest performances and measurements for each subset.

Table 5.7 lists the picked feature list of each subset for the Random Forest (RF).

5.1.5 Evaluation

Performance results from feature subsets depict that we could keep high effectivity by eliminating the computationally costly and low performing features. That elimination paid back by the greatly reduced time cost for the link predictor while not losing much from the accuracy.

NBC and BN classifiers responded better to feature reduction as their performance degrade is below 1 percentage. However, for RF predictor there is up to 5 percentage decrease in the performance. Elimination of high performing features like SPD and TSP not caused as high performance degrade. It shows us that there are other features which are giving the similar information gain like them and their elimination became a redundancy removal.

From the time consumption point, total cost of feature extraction for our dataset (200000 potential link pairs) reduced significantly (20 times) from 153 days (assuming non-parallel computation) to around 7 days with our proposed greedy method.

Table 5.5: Picked Feature Subsets for BN

Feature	**BN-OSF**	**BN-OSFLC**
CFC	X	X
CFR	X	X
PAF	X	X
AACF	X	X
TCFC	X	X
SPD	X	
TSP	X	
CPC		
CPR		
PAP		
AACP	X	X
MCC	X	X
AACPE	X	X
MEV	X	X
CCC		
TCFCC		
PAC		
CPCP	X	X
CPCPR		
CCCP		
CCCPR		
DMVH		
DMVHR	X	X
DLH		
SRLH		

Table 5.6: Measurements with RF

Subset	**Total Feature Extraction Time (min)**	**Total Training Time (sec)**	**One Candidate Pair Feature Extraction Time (sec)**	**Prediction Performance (ROC-AUC)**
AF	220666.7	418.2	66.2	0.863
OSF	181666.7	310.2	54.5	0.791
OSLCF	4666.7	197.1	1.4	0.799

5.2 Clustering Based Feature Reduction for Link Prediction Effectivity

Considering the performance issues caused by redundancy and relevance interactions between features, we proposed a custom two-step feature reduction method [11]. Proposed method starts with clustering features based on the interaction related similarity measurement and ends with non-monotonically selecting optimal feature subset from those clusters by the help of a custom formulated genetic algorithm as shown in (**Figure 5.1**).

This section covers extended details of our proposed feature reduction method together with performed empirical analysis to evaluate novelty and verify the contributions for LP problem in LBSNs. Results from multiple data groups depict the effectivity improvements of proposed method by comparing with 3 well-known feature reduction algorithms from literature [42, 43, 45].

5.2.1 Clustering Similar Features

Similarity of features have been studied heavily before [50, 63]. Most of the existing studies kept analyzing only two features and their statistics to come-up with a similarity measurement. There were few studies that considered feature interaction while selecting feature subsets [64]. Towards the goal of eliminating the logically redundant features, we decided to keep focus on the impact of a feature when used with others in a subset. Therefore, we proposed a custom similarity measurement to contribute

Table 5.7: Picked Feature Subsets for RF

Feature	**RF-OSF**	**RF-OSFLC**
CFC	X	X
CFR	X	X
PAF		
AACF		
TCFC		
SPD	X	
TSP	X	
CPC	X	X
CPR		
PAP	X	X
AACP	X	X
MCC	X	X
AACPE	X	X
MEV	X	X
CCC		
TCFCC		
PAC		
CPCP	X	X
CPCPR		
CCCP		
CCCPR	X	
DMVH	X	X
DMVHR	X	X
DLH		
SRLH		

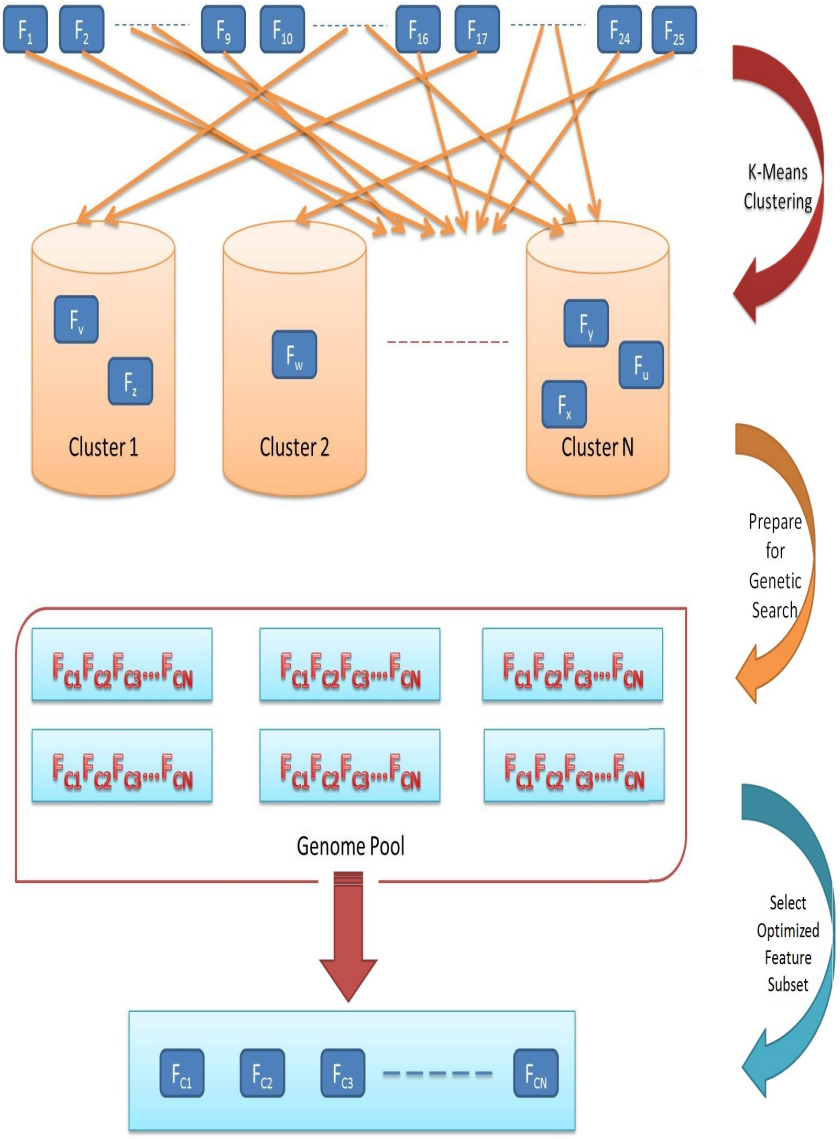

Figure 5.1: Feature Selection Process

with impact of features' interaction.

We proposed an impact-based similarity metric from feature interactions and used that measurement as a basic tool for clustering features. Below formulas are the fundamentals for that metric:

- **Single Feature Performance (FP_i):** LP performance (accuracy) for each feature, which is calculated when a learning algorithm is trained and tested using only that feature **i**. This metric evaluates the correlation of classification with information gain from that feature.
- **Mutual Feature Performance ($MP_{i,j}$):** Impact of a feature to another feature is calculated through performance gain of using two features (**i** and **j**) together rather than using one of them individually. Such calculation is performed for all possible feature pairs (300 pairs are possible for 25 features) in our data. Main idea is to determine the value added by a second feature when they interacted:
  - $FP_{i,j}$**:** LP performance for two features i and j when only those two features are used by the learning algorithm.
  - To analyze the impact of two features interaction, contribution (when both used together) is calculated by a delta from the maximum individual performance using equation 51:

$$MP_{i,j} = FP_{i,j} - max(FP_i, FP_j) \tag{51}$$

  - $MP_{i,j}$ is expected to be low (even negative) for logically redundant features because of overlapping information, etc.

To be able to represent the overall impact of an individual feature when interacting with others in the set, an impact vector for that feature is formed through the Mutual Feature Performance values of that feature with all other features in set. That constructed vector is called **Mutual Performance Vector (MPV)**, equation 52:

$$MPV_i = [MP_{i,1}, MP_{i,2}, MP_{i,3}, \cdots, MP_{i,25}] \tag{52}$$

MPV_i is constructed for defining characteristics of feature **i** and used as similarity measurement for clustering. Similar features are expected to have similar impact characteristics when interacted with other features.

For clustering step, we used K-means and Agglomerative hierarchical clustering algorithms and analyzed their contributions.

5.2.1.1 K-means Clustering

Following the trend with good performance from similar studies, we used k-means clustering algorithm for the clustering step. That simple but effective algorithm distributes N object to K clusters by leveraging vector quantization with distance calculations. **MPV** for each feature in our study is a perfect fit for k-means as it characterize individual features as vectors.

Details of the applied k-means feature clustering are as follows:

- A predefined count for clusters, **k**, is required and that number is determined by the required feature subset size.
- For the changing numbers of k, algorithm determines k centroids and perform distribution of features to those k clusters.
  - Each feature is associated with nearest centroid by a distance metric
  - Locations of centroids are updated step by step until targeted barycentre is achieved.
  - Centroids are placed as much as far away from each other
- K-means clustering help us to be non-monotonic in feature selection because the cluster contents are not fully maintained in changed numbers of k:
  - A feature selected in a subset of size m is not guaranteed to be selected again for a subset of size $m + 1$. We experimented with cluster sizes from 1 to 25 to find best feature subset.
- **Euclidean distance (ED)** is one of the mostly used distance metrics, which suits well to our k-means clustering task as well.
  - Each feature is resembled by a vector of size 25 in previous section, ED between two vectors is used as a distance function for k-means clustering,

as given in equation 53:

$$ED(MPV_i, MPV_j) = \sqrt{\sum_{n=1}^{25} (MP_{i,n} - MP_{j,n})^2} \tag{53}$$

Note that similarity metric we used (i.e. ED of w vectors) is based on two features' relation with the remaining features. As impact of feature **i** and **j** among other feature **n** become similar, $(MP_{i,n} - MP_{j,n})$ value decreases in the ED calculation. By that way, distance of two feature vectors decreases, and they are high-probably placed into the same cluster. Moreover, using two features **i and j** from the same cluster would not improve the relevance value with their small $MP_{i,j}$ value. Otherwise, they would not be placed in the same cluster. If $MP_{i,j}$ value was higher, ED would be also higher and they had been placed into the different cluster high-probably.

5.2.1.2 Agglomerative Clustering

Agglomerative clustering is a bottom-up hierarchical clustering method where initially each object is considered as a separate cluster. Then, pairs of clusters are merged based on the similarity until one big cluster is formed to contain all objects [65]. Generally, whole hierarchical process in this method is represented by an object tree called **dendrogram**.

At any step of clustering, two clusters are merged based on the distance between those clusters using different linkage approaches like single, average and complete. We used single, average and complete linkage approaches which are leveraging the distance matrix between all elements in two clusters while considering merging them:

- **Single** linkage, distance is defined as the minimum value in distance matrix.
- **Average** linkage, median value of values in distance matrix is used as cluster distance.
- **Complete** linkage, maximum value of distance matrix is used as distance between related clusters.

We experimented with agglomerative clustering outputs from cluster counts hierarchically changing from 25 to 1 and evaluated different cluster count impact on LP performance.

5.2.2 Optimal Subset from Clustered Features

Dividing features into clusters prepared the environment for formulating an optimal subset for LP problem. Next challenge was picking correct features from those clusters to form a high relevant feature subset. That challenge was also comprised from two sup problems: finding the feature to pick from the cluster and finding the order of clusters to pick. Those subproblems itself require exhaustive search and most of greedy approaches have high risk of trending local maxima and missing the final optimality.

Most commonly used approach for both problems are ranking similar features based on their correlation with the classification while choosing from a cluster [42, 43, 45]. This has the risk of missing a secondary similar feature which may contribute more when interacted with other features in other clusters. To avoid that problem, we decided to not prefer any greedy approach and keep all similar features in a cluster **"pickable"** while formulating the subset. We would still pick one of those similar features in a cluster, but we also keep chance for other features within that cluster. Another bottleneck for greedy approaches was monotonically choosing the features to form incremental selection of features. This method reduces the search space, and this may undermine the feature interaction impact [49].

To keep features in same cluster pickable and perform a non-monotonic subset formulation we proposed a customized **Genetic Algorithm (GA)**. GA is suited well for a non-monotonic feature selection because the selected features are determined in one step (as bulk, not one by one). Also our gene formulation made it possible for each feature in a cluster pickable for each genome.

5.2.3 GA Formulation

Starting point while formulating a problem for GA, first step is to pick a genome representation which helps to do optimal search in solution space. Our GA solution has the genome representation $G_1G_2 \cdots G_k$, where

- k is the total number of clusters,
- G_i is an **integer gene** specifying a label for a feature in related cluster **i**
- Features are labeled with integers from 1 to n in a cluster. where n is the number of features in that cluster. if G_i is j $(1 \leq j \leq n)$ then feature with label j is picked from cluster i.

GA starts with a random valid population of genomes. Each gene in a genome stores the index of a selected feature from the corresponding cluster. Therefore, we used uniform crossover so that information from each cluster is transferred to child genomes properly. Furthermore, random single gene mutation is used for mutation step.

5.2.4 Proposed Fitness Function

One of the hardest tasks in designing an effective GA for the problem is determining the fitness function. We applied a Heuristic-Based Fitness Function in our work by using the Mutual Feature Performance **(MP)** gain values calculated for clustering:

- A total sum is used as fitness value for each genome. That summation is based on:
  - Best performing feature (gene) within the genome and
  - Gain values for all gene pairs in the genome. Number of selected features (pairs) are fixed for each GA run, averaging is not required.
- Formulation of fitness function is given below in equation 54:

$$FF(G) = max_{i \in G}(FP_i) + \sum_{j \in G} \sum_{k \in G} MP_{j,k} \tag{54}$$

  where G is the evaluated chromosome; i, j and k are genes in G.

The GA is applied on the clusters, and a feature subset is determined. Link prediction performance of that feature subset is the main evaluation criteria for our performance improvements in link prediction problem.

5.2.5 Performance Evaluation

To analyze proposed clustering based feature reduction performance, we used its output feature subset to train Naive Bayes Classifier (NBC) and Bayesian Network (BN) classifiers for link prediction.

Proposed feature selection method aims to pick a smaller subset of features to have more effective and efficient LP solution. A good evaluation of its performance is to compare with performance when all features used. However, to see the contribution and novelty of our proposed solution comparison should be done with performances when other feature reduction and feature selection studies used. To achieve so, we used 3 other algorithms from WEKA [54]:

- **CMIM:** Fleuret's Conditional Mutual Information Maximization [43]
- **MIFS:** Battiti's Mutual Information-Based Feature Selection [42]
- **MRMR:** Peng's Max-Relevance and Min-Redundancy [45]

Evaluations show the LP performance (ROC-AUC values) when selected feature subsets are used for link prediction. Evaluations for each data group are made separately.

Initially feature clustering analysis is performed to decide on the clustering algorithm to use at our proposal. Then, proposed feature reduction method together with 3 well-known methods, are used to create 4 different subsets for each data group and for each classification algorithm. Prediction performances with all 25 features are also given for better comparison for effectivity improvements.

We also compared our best (optimal feature subset used) AUC values with the AUC values calculated using the features from most successful research[34] in same problem, same data and same algorithm. Compared research is called as OTHER RESEARCH in related tables.

Table 5.8: WG - Comparison of Clustering Algorithms

Algorithm	**NBC Perf. (ROC-AUC)**	**NBC Subset Size**	**BN Perf. (ROC-AUC)**	**BN Subset Size**
Agglomerative with Single linkage	0.940	5	0.956	3
Agglomerative with Average linkage	0.946	5	0.957	6
Agglomerative with Complete linkage	0.953	5	0.958	7
K-means	0.956	7	0.965	12

5.2.5.1 Comparison of Clustering Algorithms

We started with comparing the clustering algorithms in our proposed method according to overall LP performances when each used for whole dataset with both classifiers. **Table 5.8** shows the WG performances (ROC-AUC value) and selected feature subset size of predictors when different clustering methods used for feature reduction.

Best results for agglomerative hierarchical clustering were achieved by complete linkage. Dendrograms for each linkage type with both algorithms are shown in **Figure 5.2, 5.3, 5.4, 5.5, 5.6 and 5.7**.

Performances from k-means were better than all them in all experiments therefore we decided to use **k-means** as core clustering algorithm in our evaluations. This was expected as k-means restart whole clustering from scratch for varying cluster count however hierarchical clustering approaches are monotonically merging the clusters and reduces the search space as it is greedily eliminating other possible distributions.

5.2.5.2 Friend-of-Friends Group (FOF) Performances

Table 5.9 shows the FOF group performances (ROC-AUC value) of predictors when all feature set and multiple subsets selected from different feature reduction methods.

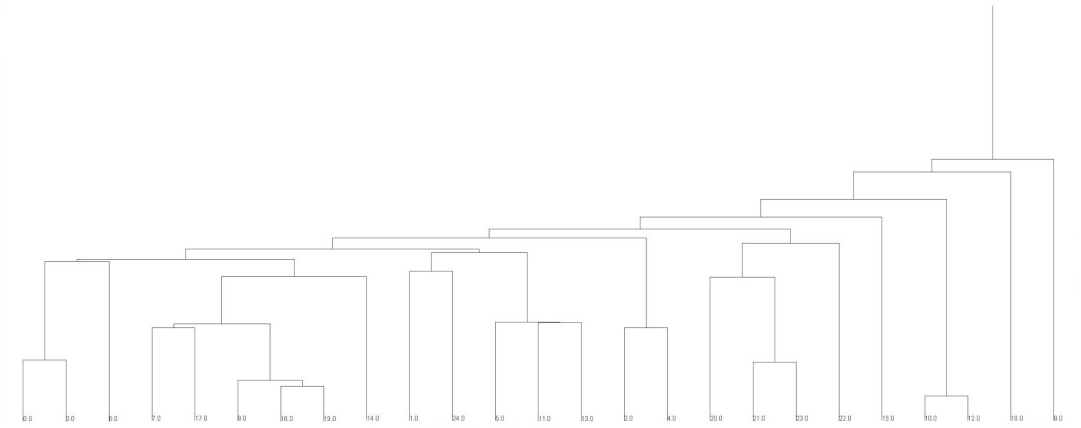

Figure 5.2: NBC - Agglomerative Hierarchical Clustering Dendrogram with Single Linkage

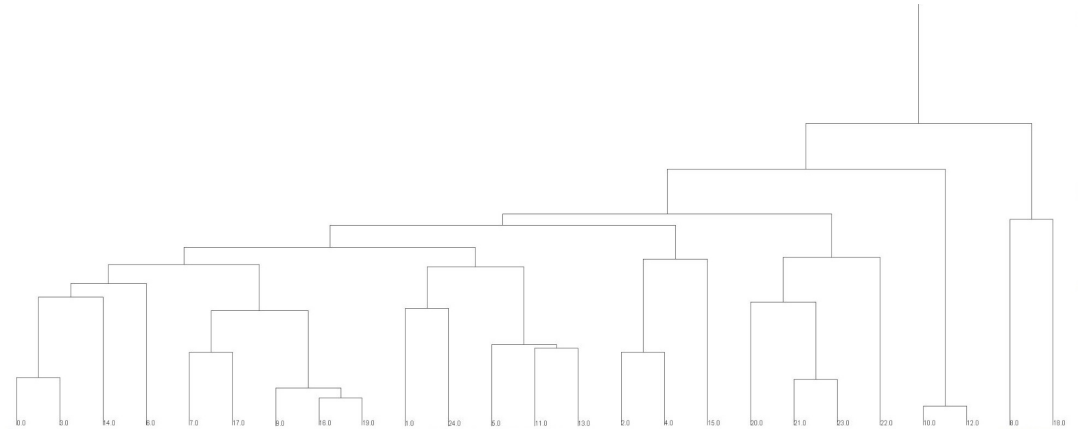

Figure 5.3: NBC - Agglomerative Hierarchical Clustering with Average Linkage Dendrogram

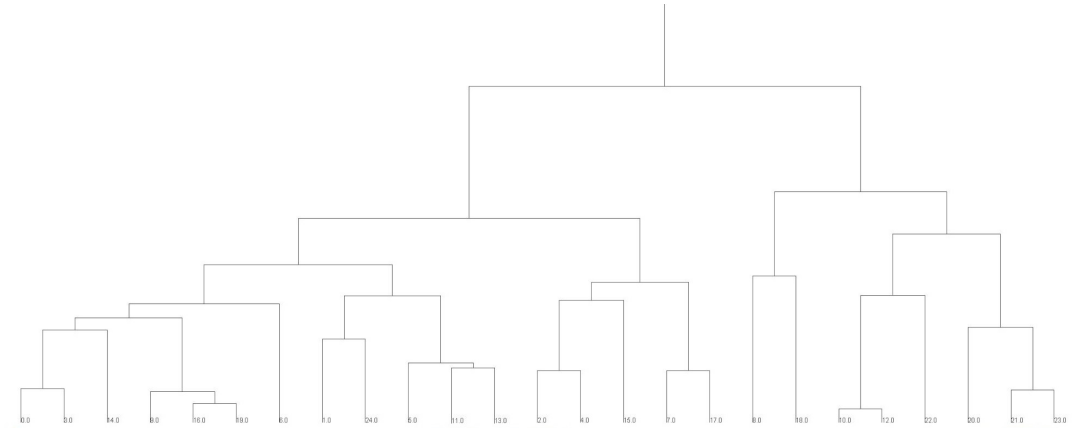

Figure 5.4: NBC - Agglomerative Hierarchical Clustering with Complete Linkage Dendrogram

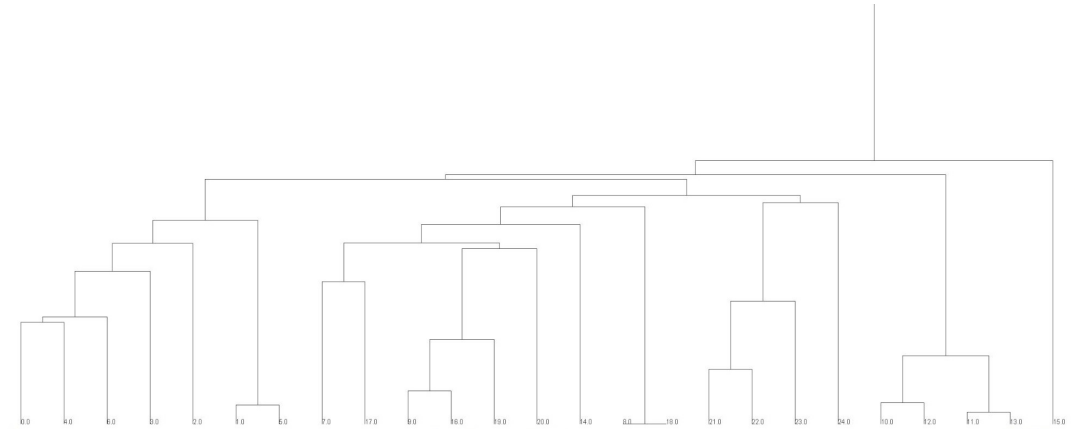

Figure 5.5: BN - Agglomerative Hierarchical Clustering with Single Linkage Dendrogram

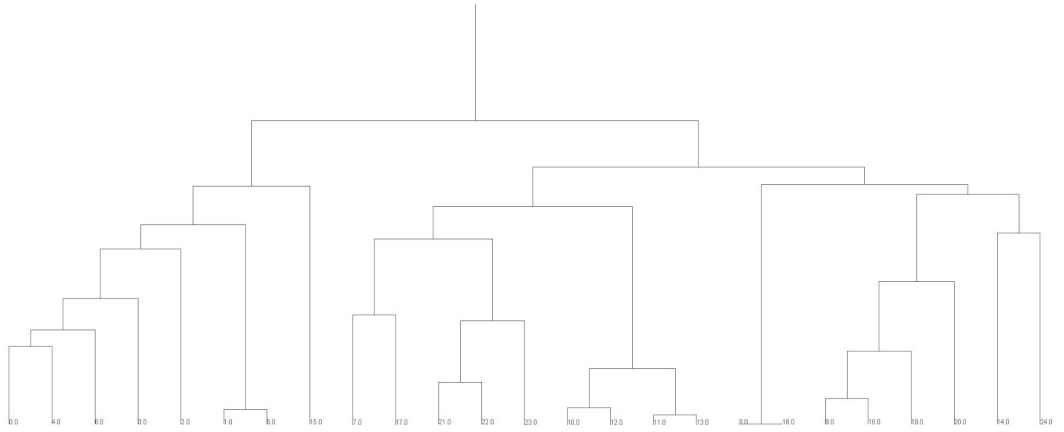

Figure 5.6: BN - Agglomerative Hierarchical Clustering with Average Linkage Dendrogram

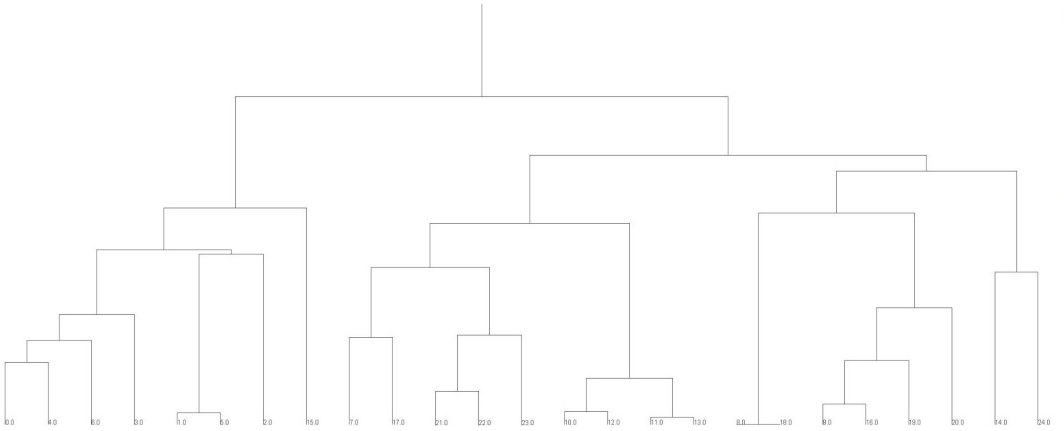

Figure 5.7: BN - Agglomerative Hierarchical Clustering with Complete Linkage Dendrogram

Table 5.9: FOF - Compared Feature Subset Performances (ROC-AUC)

Features	**NBC**	**BN**
All Features	0.908	0.930
Subset Selected by Our Method	0.957	0.960
Subset Selected by CMIM	0.946	0.960
Subset Selected by MIFS	0.948	0.955
Subset Selected by MRMR	0.957	0.956

Table 5.10 lists the picked features of FOF group for each learning algorithm by our feature reduction method.

Table 5.11 compares best link prediction performance for FOF Group from our research (with applied reduction) with best performing research from literature.

5.2.5.3 Place-Friends Group (PF) Performances

Table 5.12 shows the PF group performances (ROC-AUC value) of predictors when all feature set and multiple subsets selected from different feature reduction methods.

Table 5.13 lists the picked features of PF group for each learning algorithm by our feature reduction method.

Table 5.14 compares best link prediction performance for PF Group from our research (with applied reduction) with best performing research from literature.

5.2.5.4 Both-Friends Group (BF) Performances

Table 5.15 shows the BF group performances (ROC-AUC value) of predictors when all feature set and multiple subsets selected from different feature reduction methods.

Table 5.10: FOF - Picked Feature Subsets

Feature	**NBC**	**BN**
CFC		
CFR		X
PAF		
AACF	X	X
TCFC	X	
SPD		
TSP		X
CPC		
CPR		
PAP		
AACP		
MCC		X
AACPE	X	
MEV		
CCC	X	X
TCFCC	X	X
PAC		
CPCP		
CPCPR		
CCCP	X	
CCCPR		
DMVH		
DMVHR		
DLH		
SRLH		

Table 5.11: FOF - Prediction Performances Comparison with Other Research

Features	**NBC**	**BN**
OTHER RESEARCH	0.882	0.918
OUR OPTIMAL SUBSET	0.957	0.960

Table 5.12: PF - Compared Feature Subset Performances (ROC-AUC)

Features	**NBC**	**BN**
All Features	0.959	0.970
Subset Selected by Our Method	0.972	0.976
Subset Selected by CMIM	0.969	0.976
Subset Selected by MIFS	0.961	0.970
Subset Selected by MRMR	0.964	0.970

Table 5.16 lists the picked features of BF group for each learning algorithm by our feature reduction method.

Table 5.17 compares best link prediction performance for BF Group from our research (with applied reduction) with best performing research from literature.

5.2.5.5 Whole Group (WG) Performances

Table 5.18 shows the WG performances (ROC-AUC value) of predictors when all feature set and multiple subsets selected from different feature reduction methods.

Table 5.19 lists the picked features of WG for each learning algorithm by our feature reduction method.

Table 5.20 compares best link prediction performance for WG from our research (with applied reduction) with best performing research from literature. ROC curves for both algorithms are drawn in **Figure 5.8, 5.9 5.10 and 5.11** where y-axis is True-Positive-Rates and x-axis is False-Positive-Rates.

Table 5.13: PF - Picked Feature Subsets

Feature	**NBC**	**BN**
CFC	X	X
CFR		
PAF		
AACF	X	X
TCFC	X	X
SPD	X	
TSP		X
CPC		
CPR		X
PAP		
AACP		
MCC		X
AACPE	X	
MEV	X	X
CCC	X	X
TCFCC	X	X
PAC		
CPCP	X	
CPCPR		
CCCP		
CCCPR		
DMVH		
DMVHR		
DLH		
SRLH	X	X

Table 5.14: PF - Prediction Performances Comparison with Other Research

Features	**NBC**	**BN**
OTHER RESEARCH	0.948	0.955
OUR OPTIMAL SUBSET	0.972	0.976

Table 5.15: BF - Compared Feature Subset Performances (ROC-AUC)

Features	**NBC**	**BN**
All Features	0.881	0.898
Subset Selected by Our Method	0.899	0.912
Subset Selected by CMIM	0.899	0.912
Subset Selected by MIFS	0.891	0.911
Subset Selected by MRMR	0.882	0.908

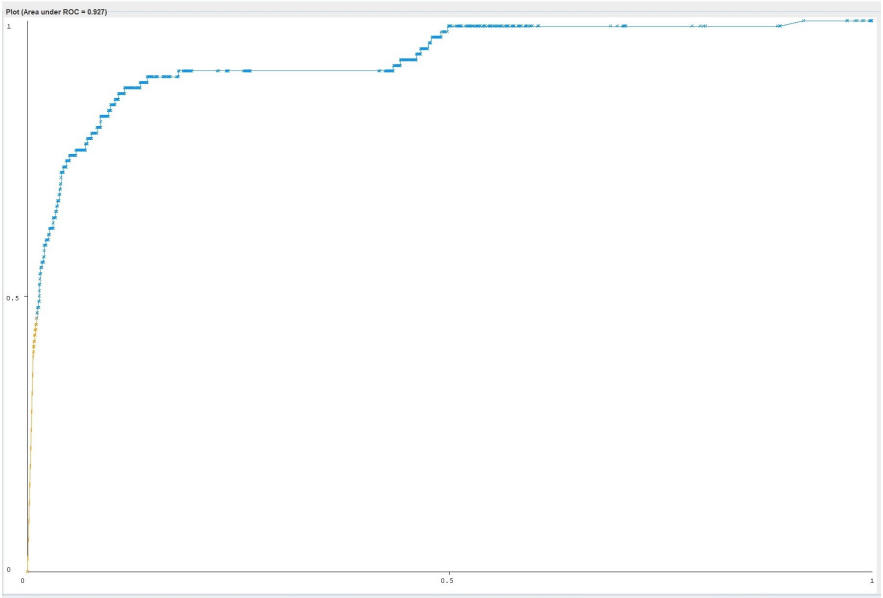

Figure 5.8: WG - ROC Curve of Other Research with NBC Algorithm

Table 5.16: BF - Picked Feature Subsets

Feature	**NBC**	**BN**
CFC	X	X
CFR		X
PAF	X	
AACF	X	X
TCFC	X	X
SPD		
TSP		
CPC		
CPR	X	
PAP		
AACP		
MCC		X
AACPE	X	
MEV	X	X
CCC	X	X
TCFCC	X	X
PAC		X
CPCP	X	X
CPCPR		
CCCP		
CCCPR		
DMVH		
DMVHR		
DLH		
SRLH	X	X

Table 5.17: BF - Prediction Performances Comparison with Other Research

Features	**NBC**	**BN**
OTHER RESEARCH	0.862	0.879
OUR OPTIMAL SUBSET	0.899	0.912

Table 5.18: WG - Compared Feature Subset Performances (ROC-AUC)

Features	**NBC**	**BN**
All Features	0.925	0.955
Subset Selected by Our Method	0.956	0.965
Subset Selected by CMIM	0.954	0.961
Subset Selected by MIFS	0.950	0.961
Subset Selected by MRMR	0.943	0.962

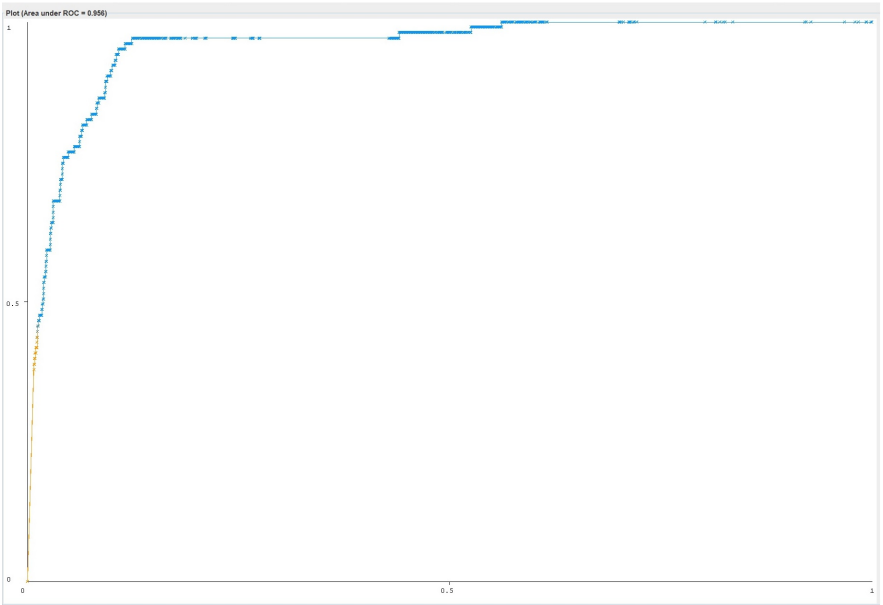

Figure 5.9: WG - ROC Curve of OUR OPTIMAL SUBSET with NBC Algorithm

Table 5.19: WG - Picked Feature Subsets

Feature	**NBC**	**BN**
CFC		X
CFR		
PAF	X	
AACF	X	X
TCFC	X	
SPD		
TSP		X
CPC		
CPR		
PAP		
AACP		
MCC		X
AACPE	X	X
MEV		X
CCC	X	X
TCFCC	X	X
PAC		X
CPCP		
CPCPR		X
CCCP	X	
CCCPR		X
DMVH		
DMVHR		
DLH		
SRLH		X

Table 5.20: WG - Prediction Performances Comparison with Other Research

Features	**NBC**	**BN**
OTHER RESEARCH	0.927	0.941
OUR OPTIMAL SUBSET	0.956	0.965

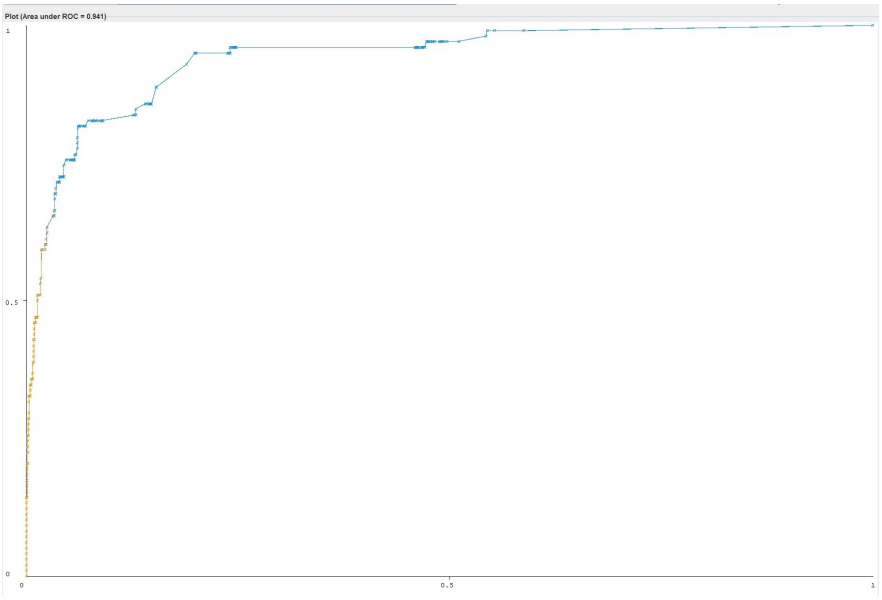

Figure 5.10: WG - ROC Curve of Other Research with BN Algorithm

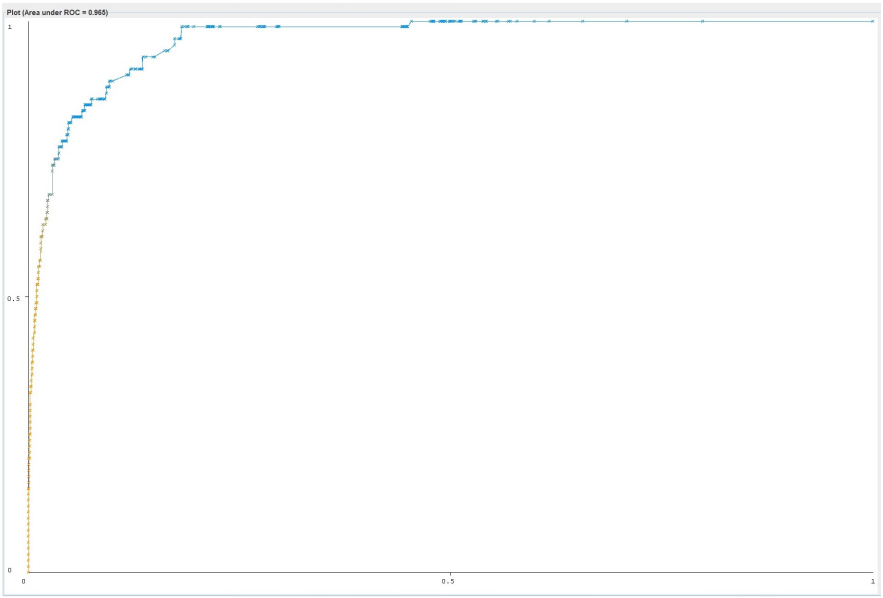

Figure 5.11: WG - ROC Curve of Our Optimal Subset with BN Algorithm

Table 5.21: WG - True Positive Rates for Two Proximity Groups

Features	**NBC**	**BN**
Proximity Group 1	0.679	0.714
Proximity Group 2	0.673	0.764
All Predictions	0.676	0.739

5.2.5.6 Temporal Proximity Analysis for Positive Link Predictions

Our motivation is to improve the LP performance from positive links' prediction accuracy perspective as positively predicted outputs would be used by a recommendation system. Each predictor's performance may also depend on temporal proximity of prediction time to the actual link creation time.

In our research, predictions are done for links that may be created after PPBD with the LBSN data collected before PPBD which can be rephrased as prediction time. Therefore, we performed temporal proximity analysis for positive link predictions by including the link creation times and PPBD into the evaluation. Using BN and NBC algorithms with WG dataset, true positive link prediction rates are compared for varying temporal proximity groups. In such experiment, we couldn't keep calculating ROC-AUC as false positive rates are not independent for the proximity.

Proximity group with lower number covers links which are created earlier than the links in the other proximity group with higher number. **Table 5.21** compares **true positive rates** when links after PPBD are divided into 2 groups based on proximity and **Table 5.22** compares **true positive rates** when links after PPBD are divided into 3 groups based on proximity.

Results from temporal proximity analysis depict that:

- There is not much performance difference for algorithms whether proximity based groups are used or not.
- Our total prediction time is less than 3 weeks, so most of the LBSN data is still fresh at the prediction time and most of the link creation time is not far away.

Table 5.22: WG - True Positive Rates for Three Proximity Groups

Features	**NBC**	**BN**
Proximity Group 1	0.703	0.730
Proximity Group 2	0.649	0.730
Proximity Group 3	0.676	0.757
All Predictions	0.676	0.739

- BN predicts slightly better for links created later.

5.2.6 Evaluation

Based on the experiment results, we can easily see that an effective LP solution in LBSN requires an additional feature reduction to optimize predictor performance with an optimal feature subset. All 4 subsets had better results for 4 different data groups on both classification algorithms. Improvements are not huge numerically however at the given rates of prediction performance, those improvements are valuable for OSNs targeting millions of people. Also, most of the selected subsets are having less than 10 features which would also contribute from efficiency perspective.

From the perspective of comparing overall LP performance with best performing research from literature, our optimal feature subset which leverages both proposed features and feature reduction is overperforming in all data groups in all algorithms.

If we analyze our proposed feature reduction by comparing with other similar methods, performance results are mostly greater than or equal to the best of them. There is not an obvious differentiation between those 3 methods' results but results depict a slighter improvement from our method for all data groups.

Novelty of our method was formulating an algorithm that combines two critical requirements from our feature set:

- Removal of logical redundant features among others
  - Wanted to minimize the negative performance impact of feature interac-

tions

  - Defined a custom similarity metric and clustered features according to their interaction with others

- Picking the optimal subset

  - Non-monotonic picking from clusters to avoid possible local maxima risk for greedy selection approaches.
  - Enhanced search space by the help of fully customized genetic algorithm applied for feature clusters

CHAPTER 6

FEATURE MINING

Our focus on feature mining was to leverage more from the check-in data to improve link prediction, because that semantically rich information is exclusive for LBSNs. There are some studies that used the common places of two users and basic details (total check-in count) of those places. We wanted to deep dive on the semantics of these places where we can enrich the information gain from check-in data. Category of a place is one of main property of a place and that information is provided by nearly all LBSNs. This chapter covers the feature mining efforts to propose category based features [12] and evaluate their performances.

6.1 Proposed Place Category Based Features

We proposed two new groups of features to understand impact of the place category on link prediction for LBSN users. We calculated **Common Place Check-in Count Product Sum** and **Common Category Check-in Count Sum Product** for each category, while using two users' historical data to see the prediction performance.

Previous features like common place or common category count were combining all places (ignoring category) to extract one feature from all. However, this unification phase may cause an informative data loss as we converted some vectored data to scalar data. For example, there are 283 different categories for the places in our dataset. All of them are not as effective as others in new friendship. Some places/categories are more encouraging for making new friends than others. Therefore, we decided to propose some new features which are calculated for each category while evaluating any link candidate user pair.

Common Place Check-in Count Product Sum (CPCPS) is a feature which will be calculated in each category for a given two candidate users. It is a sum of products where each product is a multiplication of check-in counts for two candidates at one shared place of related category. Each feature will be a good resemblance of how often they went to shared places for related category.

In order to emphasize on the commonness of the place, we multiply check-in count of each candidate at shared place. Check-in count of a candidate (X) at a place (Y) is formulated as CC(X,Y) function in related feature's Equation 61 for category α.

$$CPCPS(A, B, ct_\alpha) = \sum_{z \in (ct_\alpha)} (CC(A, z) * CC(B, z)) \tag{61}$$

For example;

- A and B are link candidates
- A and B has common places X, Y, Z
- X and Y are category1 places and Z is a category2 place
  - *CPCPS(A, B, category1) = [CC(A, X) * CC(B, X)] + [CC(A, Y) * CC(B, Y)]* and
  - *CPCPS(A, B, category2) = [CC(A, Z) * CC(B, Z)]*

Common Category Check-in Count Sum Product (CCCSP) is a feature which will be calculated in each category for a given two candidate users. It is a product of sums where each sum is the total check-in count for a candidate at given category. Each feature will be a good resemblance of how often they went to some place for related category.

In order to emphasize on the commonness of the category, we multiply total check-in count of each candidate. Related feature formula is given at Equation 62.

$$CCCSP(A, B, ct_\alpha) = \sum_{\forall z \in (ct_\alpha)} CC(A, z) * \sum_{\forall z \in (ct_\alpha)} CC(B, z) \tag{62}$$

For example;

- A and B are link candidates
- A visited places K, L and M; B visited X, Y and Z
- K, X and Y are category1 places and L, M and Z are category2 places
  - *CCCSP(A, B, category1) = [CC(A, K)] * [CC(B, X) + CC(B, Y)]* and
  - *CCCSP(A, B, category2) = [CC(A, L) + CC(A, M)] * [CC(B, Z)]*

By the definitions above; we have 566 new feature values from f_0 to f_{565} (2 feature for each 283 categories) that we calculate for each link candidate pairs.

We will use reference keys to identify each feature; for any category with id **i**:

- **CCCSP** feature will be referenced by $F_{(2*i)-2}$ value.
- **CPCPS** feature will be referenced by $F_{(2*i)-1}$ value.

We combined new features with best performing topological and interaction features from literature that we collected and developed in our previous study [9]. Related reference keys are given in **Table (6.1)** where numbers after proposed features are selected intentionally.

6.1.1 Performance Evaluation

Here are the steps applied to analyze the proposed feature performances for every data group:

1. Use **Best Performing Feature Set (BPFS)** specific for related dataset from previous studies.
2. Use each new feature together with BPFS to train a Bayesian Network to observe link prediction performance change compared to the case just using the BPFS.

Table 6.1: Existing Features

Reference Key	**Feature**
F_{566}	Common Friends Count
F_{567}	Common Friends Ratio
F_{569}	Adamic Adar of Common Friends
F_{570}	Total Common Friend Closeness
F_{572}	Total Shortest Paths
F_{573}	Common Place Count
F_{574}	Common Place Ratio
F_{577}	Min Check-in Count of Common Places
F_{579}	Min Entropy of Common Places
F_{580}	Common Check-in Count
F_{581}	Adamic Adar of Common Friend Common Check-in
F_{582}	Preferential Attachment Check-in Counts
F_{583}	Common Place Check-in Counts Product
F_{584}	Common Place Check-in Counts Product Ratio
F_{585}	Common Category Check in Counts Product
F_{586}	Common Category Check in Counts Product Ratio
F_{589}	Distance Of LatLon Homes
F_{590}	Sum of Radius Length from Lat-Lon Homes

3. By selecting features which improved the performance; we define a **Filtered New Feature Set (FNFS)** and calculated the performance by using all **FNFS and BPFS**.

4. Apply feature selection to the newly proposed features; obtain best performing features: BPFS + subset(FNFS).

After applying the steps given above, we could get best performance improver category feature subset for every data group.

6.1.1.1 Friend-of-Friends Group (FOF) Performances

Table 6.2 shows the FOF group performances (ROC-AUC value) of predictors with mined category features.

6.1.1.2 Place-Friends Group (PF) Performances

Table 6.3 shows the PF group performances (ROC-AUC value) of predictors with mined category features.

6.1.1.3 Both-Friends Group (BF) Performances

Table 6.4 shows the BF group performances (ROC-AUC value) of predictors with mined category features.

6.1.1.4 Whole Group (WG) Performances

Table 6.5 shows the WG performances (ROC-AUC value) of predictors with mined category features.

Table 6.2: FOF Group Results with Mined Category Features

NO	**DETAILS**	**AUC of ROC**
1	From our previous studies we have BPFS (reference keys): F_{567}, F_{569}, F_{572}, F_{577}, F_{580}, F_{581}. Prediction performance for BFS is given.	0.960
2	Calculated each feature's performance together with BPFS. Find best performing feature is F_{192}. Prediction perf. for BFS + best new feature is given.	0.962
3	Following 14 new category features (FNFS) had increased the prediction performance when used each one together with BPFS: F_{32}, F_{54}, F_{94}, F_{108}, F_{114}, F_{120}, F_{192}, F_{196}, F_{222}, F_{258}, F_{304}, F_{308}, F_{364}, F_{412}. Prediction performance for BFS + FNFS is given.	0.957
4	Following FNFS subset is selected after performing a feature selection: F_{114}, F_{192}, F_{308}. Prediction performance for BFS + subset(FNFS) is given.	0.963

Table 6.3: PF Group Results with Mined Category Features

NO	**DETAILS**	**AUC of ROC**
1	From our previous studies we have BPFS (reference keys): F_{566}, F_{569}, F_{572}, F_{574}, F_{577}, F_{579}, F_{580}, F_{581}, F_{585}. Prediction performance for BFS is given.	0.976
2	Calculated each feature's performance together with BPFS. Find best performing feature is F_{45}. Prediction perf. for BFS + best new feature is given.	0.976
3	Following 27 new category features (FNFS) had increased the prediction performance when used each one together with BPFS: F_1, F_3, F_9, F_{16}, F_{34}, F_{39}, F_{45}, F_{49}, F_{53}, F_{61}, F_{97}, F_{103}, F_{107}, F_{109}, F_{130}, F_{153}, F_{168}, F_{180}, F_{189}, F_{200}, F_{214}, F_{256}, F_{257}, F_{362}, F_{374}, F_{415}, F_{454}. Prediction performance for BFS + FNFS is given.	0.977
4	Following FNFS subset is selected after performing a feature selection: F_3, F_9, F_{16}, F_{39}, F_{45}, F_{49}, F_{97}, F_{153}, F_{180}, F_{200}, F_{362}, F_{415}, F_{454}. Prediction performance for BFS + subset(FNFS) is given.	0.978

Table 6.4: BF Group Results with Mined Category Features

NO	**DETAILS**	**AUC of ROC**
1	From our previous studies we have BPFS (keys): F_{566}, F_{567}, F_{569}, F_{570}, F_{577}, F_{579}, F_{580}, F_{581}, F_{582}, F_{583}, F_{590}. Prediction performance for BFS is given.	0.913
2	Calculated each feature's performance together with BPFS. Find best performing feature is F_{26} Prediction perf. for BFS + best new feature is given.	0.915
3	Following 14 new category features (FNFS) had increased the prediction performance when used each one together with BPFS: F_{26}, F_{36}, F_{38}, F_{45}, F_{92}, F_{95}, F_{100}, F_{107}, F_{138}, F_{139}, F_{150}, F_{175}, F_{252}, F_{259}. Prediction performance for BFS + FNFS is given.	0.917
4	Following FNFS subset is selected after performing a feature selection: F_{26}, F_{45}, F_{95}, F_{100}, F_{138}, F_{139}, F_{150}. Prediction performance for BFS + subset(FNFS) is given.	0.919

Table 6.5: WG Results with Mined Category Features

NO	**DETAILS**	**AUC of ROC**
1	From our previous studies we have BPFS (reference keys): F_{569}, F_{572}, F_{573}, F_{577}, F_{579}, F_{580}, F_{581}, F_{582}, F_{584}, F_{586}, F_{589}. Prediction performance for BFS is given.	0.965
2	Calculated each feature's performance together with BPFS. Find best performing feature is F_{200} Prediction perf. for BFS + best new feature is given.	0.965
3	Following 24 new category features (FNFS) had increased the prediction performance when used each one together with BPFS: F_3, F_{16}, F_{36}, F_{49}, F_{52}, F_{109}, F_{130}, F_{150}, F_{153}, F_{158}, F_{159}, F_{170}, F_{180}, F_{192}, F_{198}, F_{200}, F_{202}, F_{230}, F_{242}, F_{257}, F_{362}, F_{374}, F_{415}, F_{504}. Prediction performance for BFS + FNFS is given.	0.966
4	Following FNFS subset is selected after performing a feature selection: F_3, F_{16}, F_{36}, F_{49}, F_{109}, F_{130}, F_{153}, F_{170}, F_{198}, F_{200}, F_{242}, F_{257}, F_{362}, F_{415}. Prediction performance for BFS + subset(FNFS) is given.	0.967

6.1.2 Evaluation

Proposed features using categories improved the link prediction performance. Interesting observation from the results is the difference of improvement maker features and selected best feature subsets between datasets. Some categories seem to be helping for the link candidates when they have shared friend. However, some other categories seem to be helping more for the link candidates when they have shared place. That behavior is expected as both category information and data subsets have their semantic data organically. Here are some findings:

- CPCPS features not improved the performance for FOF dataset as expected (no shared place). In general, more CCCSP features have improvements than CPCPS ones.
- FOF and WG datasets share improver F_{192}, CCSP of "Gas & Automotive" category
- FOF and BF datasets share "Church" category as improver with different feature group
- For BF dataset both CCSP and CPCPS features for "Dive Bar" category has improvements on prediction.
- BF and WG datasets share improvers F_{36} and F_{150}, CCSP of "Cineplex" and "Convention Center" categories.
- BF and PF datasets share improvers F_{45} and F_{107}, CPCPS of "Italian" and "Dessert" categories. In addition, they share "Other - Shopping" category as improver with different feature group
- PF and WG datasets share improvers on "Craftsman", "Chipotle", "Bookstore", "Apple Store", "Other - Entertainment", "Terminal", "Ultra-Lounge", "Antique Hotel", "Plaza / Square", "Arcade", "Other-Airport", "Accessories" and "Other-Services" categories.

Motivation was to evaluate impact of check-in place categories based on the performance of related features. Such an evaluation enabled us to create link predictor

models with higher performance by enriched information gain from new category based features. Our results depict that some place categories are more correlated with new friendship and some are not. By this way, we could enhance our prediction performance by usage of category semantic while calculating check-in related features.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusions

In this thesis, we researched the link prediction problem for Location Based Social Networks. First, we formulated problem as a binary classification problem to decide existence of future friendship between two users looking at their topological and interaction features extracted from the social network.

Contextual feature analysis, feature reduction and feature mining approaches are applied for the problem to improve effectivity and/or efficiency of the formulated classifiers. Naive Bayes Classifier, Bayesian Networks and Random Forest algorithms are used as supervised learners to train mentioned classifiers.

Used a public dataset from a location based social network [31]. To enrich the measurements and evaluations data is grouped into 4 based on the link candidates state at prediction time: having common friend, having common place, etc.

A end-to-end framework is developed for the whole research which takes the dataset as input, calculates extracted features, applies various algorithms to determine feature subset, trains link prediction learners, collects performances and evaluate outputs.

After a comprehensive research, 20 features from literature are collected and implemented in the framework. Observed that LBSN specific contextual information (check-in) was not leveraged enough at those features. We proposed novel features which are calculated by using check-in time, check-in category and common friend details of candidates. We can conclude that; proposed features improve the link prediction performance at all data groups as they make use of available information in

LBSN data that cannot be utilized by the existing features.

Adding features are not always the best approach for a classification problem from both effectivity and efficiency perspectives. Picking optimal feature subsets based on the optimization goal was another research are we studied:

- For efficiency goal, we worked on a greedy method that focuses on individual feature's extraction time cost and considers its prediction accuracy. From the results, we can summarize that great time cost reduction could be achieved without losing much from prediction with that proposed method.

- For effectively goal, we observed accuracy improvements for all data groups and all algorithms after applying proposed two step feature reduction:

  - Removal of logical redundant features among others by clustering features with custom similarity metrics and

  - Formulating optimal feature subset by picking a feature from each cluster by leveraging non-monotonic genetic algorithm with a high coverage of search space.

With the motivation of finding features with exclusive information gain, we also performed feature mining focused on place category information. Two proposed groups of features were calculated separately for each category in dataset for a link candidate pair. Our results depict that some place categories are more correlated with new friendship and we could enhance prediction performance by usage of category semantic through features from those categories.

To summarize, we could help LP problem literature for LBSNs with 3 feature based approaches. Some of the proposed features and proposed feature reduction approaches can be also leveraged in other research areas with similar problem constraints. Even working on a problem with very high performing existence studies, we could move the literature one step forward with propositions and their outputs.

7.2 Future Work

Current research is based on using single location social network data for feature extraction to predict friendship links between users. As mentioned users are real people, multiple external data sources should also be available online; websites, archives and other social networks. Towards the link prediction goal, leveraging other social networks to make use of the information gain seems highly promising especially from social networks with different type of social interaction. One of the future work area would be this hybrid data consumption.

Throughout this thesis, features are calculated from LBSN interactions which are assumed to be isolated and identical. However, check-ins can be dependent to each other (as a sequence) and identical check-in of a user in a place impact the new links differently as a result of human behavior. Statistical analysis of such dependency between check-ins may help on finding finer grained features which is another possible future area of research.

REFERENCES

[1] C. C. Aggarwal, ed., *Social Network Data Analytics*. Springer, 2011.

[2] J. Heidemann, M. Klier, and F. Probst, "Online social networks: A survey of a global phenomenon," *Computer Networks*, vol. 56, no. 18, pp. 3866 – 3878, 2012. The WEB we live in.

[3] Y. C. Xu, Y. Yang, Z. Cheng, and J. Lim, "Retaining and attracting users in social networking services: An empirical investigation of cyber migration," *The Journal of Strategic Information Systems*, vol. 23, no. 3, pp. 239 – 253, 2014.

[4] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, (New York, NY, USA), pp. 556–559, ACM, 2003.

[5] E. Bütün, M. Kaya, and R. Alhajj, "A supervised learning method for prediction citation count of scientists in citation networks," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, (New York, NY, USA), pp. 952–958, ACM, 2017.

[6] H. Yin, Z. Hu, X. Zhou, H. Wang, K. Zheng, Q. V. H. Nguyen, and S. Sadiq, "Discovering interpretable geo-social communities for user behavior prediction," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 942–953, May 2016.

[7] O. Roick and S. Heuser, "Location based social networks – definition, current state of the art and research agenda," *Transactions in GIS*, vol. 17, no. 5, pp. 763–784, 2013.

[8] N. Li and G. Chen, "Analysis of a location-based social network," in *2009*

International Conference on Computational Science and Engineering, vol. 4, pp. 263–270, Aug 2009.

[9] A. E. Bayrak and F. Polat, "Contextual feature analysis to improve link prediction for location based social networks," in *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, SNAKDD'14, (New York, NY, USA), pp. 7:1–7:5, ACM, 2014.

[10] A. E. Bayrak and F. Polat, "Mining individual features to enhance link prediction efficiency in location based social networks," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 920–925, Aug 2018.

[11] A. E. Bayrak and F. Polat, "Effective feature reduction for link prediction in location-based social networks," *Journal of Information Science*, vol. 45, no. 5, pp. 676–690, 2019.

[12] A. E. Bayrak and F. Polat, "Examining place categories for link prediction in location based social networks," in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '16, (Piscataway, NJ, USA), pp. 976–979, IEEE Press, 2016.

[13] J. Scott, "Social network analysis," *Sociology*, vol. 22, no. 1, pp. 109–127, 1988.

[14] H. Gao and H. Liu, *Data Analysis on Location-Based Social Networks*, pp. 165–194. New York, NY: Springer New York, 2014.

[15] L. Ball, "Automating social network analysis: A power tool for counter-terrorism," *Security Journal*, vol. 29, pp. 147–168, Apr 2016.

[16] D. Hristova, A. Noulas, C. Brown, M. Musolesi, and C. Mascolo, "A multilayer approach to multiplexity and link prediction in online geo-social networks," *CoRR*, vol. abs/1508.07876, 2015.

[17] N. Benchettara, R. Kanawati, and C. Rouveirol, "Supervised machine learning applied to link prediction in bipartite social networks," in *2010 International Conference on Advances in Social Networks Analysis and Mining*, pp. 326–330, Aug 2010.

[18] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[19] D. Davis, R. Lichtenwalter, and N. V. Chawla, "Multi-relational link prediction in heterogeneous information networks," in *2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 281–288, July 2011.

[20] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 73–80, Oct 2011.

[21] C. Lee, B. Nick, U. Brandes, and P. Cunningham, "Link prediction with social vector clocks," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, (New York, NY, USA), pp. 784–792, ACM, 2013.

[22] W. Almansoori, S. Gao, T. N. Jarada, A. M. Elsheikh, A. N. Murshed, J. Jida, R. Alhajj, and J. Rokne, "Link prediction and classification in social networks and its application in healthcare and systems biology," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 1, pp. 27–36, Jun 2012.

[23] A. E. Mislove, *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. PhD thesis, Rice University, 2009.

[24] J. Zhu, Q. Xie, and E. J. Chin, "A hybrid time-series link prediction framework for large social network," in *Database and Expert Systems Applications* (S. W. Liddle, K.-D. Schewe, A. M. Tjoa, and X. Zhou, eds.), (Berlin, Heidelberg), pp. 345–359, Springer Berlin Heidelberg, 2012.

[25] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds, "An evolutionary algorithm approach to link prediction in dynamic social networks," *Journal of Computational Science*, vol. 5, no. 5, pp. 750 – 764, 2014.

[26] D. Shin, S. Si, and I. S. Dhillon, "Multi-scale link prediction," in *Proceedings*

of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, (New York, NY, USA), pp. 215–224, ACM, 2012.

[27] P. R. S. Soares and R. B. C. PrudêNcio, "Proximity measures for link prediction based on temporal events," *Expert Syst. Appl.*, vol. 40, pp. 6652–6660, Nov. 2013.

[28] T. Tylenda, R. Angelova, and S. Bedathur, "Towards time-aware link prediction in evolving social networks," in *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, SNA-KDD '09, (New York, NY, USA), pp. 9:1–9:10, ACM, 2009.

[29] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, (New York, NY, USA), pp. 243–252, ACM, 2010.

[30] Z. Bao, Y. Zeng, and Y. C. Tay, "sonlp: Social network link prediction by principal component regression," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pp. 364–371, Aug 2013.

[31] M. Allamanis, S. Scellato, and C. Mascolo, "Evolution of a location-based online social network: Analysis and models," in *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, (New York, NY, USA), pp. 145–158, ACM, 2012.

[32] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring Millions of Footprints in Location Sharing Services," in *Proceedings of the Fifth International Conference on Weblogs and Social Media*, (Menlo Park, CA, USA), AAAI, July 2011.

[33] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks," in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[34] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, (New York, NY, USA), pp. 1046–1054, ACM, 2011.

[35] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history," *ACM Trans. Web*, vol. 5, pp. 5:1–5:44, Feb. 2011.

[36] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: Relationship prediction in heterogeneous information networks," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, (New York, NY, USA), pp. 663–672, ACM, 2012.

[37] O. J. Mengshoel, R. Desai, A. Chen, and B. Tran, "Will we connect again? machine learning for link prediction in mobile social networks," in *Eleventh Workshop on Mining and Learning with Graphs, MLG 2013*, 2013.

[38] J. Ye, Z. Zhu, and H. Cheng, *What's Your Next Move: User Activity Prediction in Location-based Social Networks*, pp. 171–179. SIAM, 2013.

[39] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, Mar. 2003.

[40] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1, pp. 131 – 156, 1997.

[41] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004.

[42] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Trans. Neur. Netw.*, vol. 5, pp. 537–550, July 1994.

[43] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Dec. 2004.

[44] G. Wang and F. H. Lochovsky, "Feature selection with conditional mutual information maximin in text categorization," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, (New York, NY, USA), pp. 342–349, ACM, 2004.

[45] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, Aug 2005.

[46] M. Prakash and M. Murty, "Feature selection to improve classification accuracy using a genetic algorithm," *Journal of the Indian Institute of Science*, vol. 77, pp. 85–93, 01 1997.

[47] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognition*, vol. 43, no. 1, pp. 5 – 13, 2010.

[48] Y. Xu and D. Rockmore, "Feature selection for link prediction," in *Proceedings of the 5th Ph.D. Workshop on Information and Knowledge*, PIKM '12, (New York, NY, USA), pp. 25–32, ACM, 2012.

[49] Z. Xu, R. Jin, J. Ye, M. R. Lyu, and I. King, "Non-monotonic feature selection," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), pp. 1145–1152, ACM, 2009.

[50] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 301–312, March 2002.

[51] R. Butterworth, G. Piatetsky-Shapiro, and D. A. Simovici, "On feature selection through clustering," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 4 pp.–, Nov 2005.

[52] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 1–14, Jan 2013.

[53] A. Appice, M. Ceci, S. Rawles, and P. Flach, "Redundant feature elimination for multi-class problems," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, (New York, NY, USA), pp. 5–, ACM, 2004.

[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, Nov. 2009.

[55] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, pp. 131–163, Nov. 1997.

[56] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[57] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.

[58] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *ReCALL*, vol. 31, no. HPL-2003-4, pp. 1–38, 2004.

[59] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[60] M. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, p. 025102, 2001.

[61] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211 – 230, 2003.

[62] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[63] S. Tabakhi and P. Moradi, "Relevance-redundancy feature selection based on ant colony optimization," *Pattern Recogn.*, vol. 48, pp. 2798–2811, Sept. 2015.

[64] Z. Zeng, H. Zhang, R. Zhang, and C. Yin, "A novel feature selection method considering feature interaction," *Pattern Recogn.*, vol. 48, pp. 2656–2666, Aug. 2015.

[65] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *In KDD Workshop on Text Mining*, 2000.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Bayrak, Ahmet Engin
Nationality: Turkish (TC)
Date and Place of Birth: 01.08.1984, Ankara
Phone: +90 312 322 96 22
E-mail: engin.bayrak@metu.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
M.S.	METU Computer Engineering	2010
B.S.	METU Computer Engineering	2006
High School	Atatürk High School	2002

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2014 - Present	Microsoft	Senior Software Engineer
2009 - 2014	Minder	Co-Founder / Senior Software Engineer
2006 - 2009	Milsoft	Software Engineer

PUBLICATIONS

International Conference Publications

A. E. Bayrak and F. Polat, Reducing Features to Improve Link Prediction Performance in Location Based Social Networks, Non-Monotonically Selected Subset from Feature Clusters, 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, 2019.

A. E. Bayrak and F. Polat, Mining Individual Features to Enhance Link Prediction Efficiency in Location Based Social Networks, 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, 2018.

A. E. Bayrak and F. Polat, Examining place categories for link prediction in Location Based Social Networks. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining(ASONAM), San Francisco, CA, 2016.

A. E. Bayrak and F. Polat, Contextual Feature Analysis to Improve Link Prediction for Location Based Social Networks. In Proceedings of the 8th Workshop on Social Network Mining and Analysis (SNAKDD'14). ACM, New York, NY, USA, 2014.

M. Azak and A. E. Bayrak, A new approach for Threat Evaluation and Weapon Assignment problem, hybrid learning with multi-agent coordination, 2008 23rd International Symposium on Computer and Information Sciences, Istanbul, 2008.

International Journal Publications

A. E. Bayrak and F. Polat, Effective feature reduction for link prediction in location-based social networks, Journal of Information Science, vol. 45, no. 5, pp. 676–690, 2019.

A. E. Bayrak and F. Polat, Employment of an evolutionary heuristic to solve the target allocation problem efficiently, Information Sciences, Volume 222, 2013.