

# **Mapping Information Diffusion through Social Computing for Knowledge Discovery**

*Thesis submitted to*  
Cochin University of Science and Technology  
*in partial fulfilment of the requirements for the award of the*  
*degree of*

**DOCTOR OF PHILOSOPHY**

*by*  
**MINI U**

*Under the guidance of*  
**Dr. K. POULOSE JACOB**



**FACULTY OF TECHNOLOGY**  
Cochin University of Science and Technology  
Kochi - 682 022, Kerala, India  
January 2018

# **Mapping Information Diffusion through Social Computing for Knowledge Discovery**

**Ph.D. Thesis in the field of Social Media Mining**

## **Author**

**MINI U**

Department of Computer Science  
Cochin University of Science and Technology  
Kochi– 682 022, India  
E-mail: mini\_u@cusat.ac.in

## **Research Advisor**

**Dr. K. POULOSE JACOB** (Supervising Guide)

Professor, Department of Computer Science  
Cochin University of Science and Technology  
Kochi– 682 022, India  
E-mail: kpj@cusat.ac.in

**JANUARY 2018**

*Dedicated to ...*

*My Family, Friends & Well wishers*

## DECLARATION

I hereby declare that the thesis entitled “ *Mapping Information Diffusion through Social Computing for Knowledge Discovery*” is the authentic record of research work carried out by me, for my Doctoral Degree under the supervision and guidance of Dr. Poulouse Jacob K Professor, Department of Computer Science, Cochin University of Science & Technology and that no part thereof has previously formed the basis for the award of any degree or diploma or any other similar titles or recognition.

KOCHI- 22  
30-01-2018

Mini U



FACULTY OF TECHNOLOGY  
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY  
KOCHI- 682022, INDIA

### *Certificate*

*Certified that the work presented in this thesis entitled “Mapping Information Diffusion through Social Computing for Knowledge Discovery” is based on the bona fide research work done by Ms. Mini U under my guidance in the Department of Computer Science, Cochin University of Science and Technology, Kochi -22 and has not been included in any other thesis submitted previously for the award of any degree.*

*Kochi-22  
30/01/2018*

*Dr. K. Poullose Jacob*

## *Certificate*

*This is to certify that all the relevant corrections and modifications suggested by the audience during the Pre-synopsis seminar and recommended by the Doctoral Committee of the candidate have been incorporated in the thesis.*

*Kochi-22  
30/01/2018*

*Dr. K. Poullose Jacob*

## *Acknowledgements*

The research work leading to PhD has been a long and challenging journey. This work would not have been completed, but for the unstinted support and guidance of many. I thank God Almighty for the blessings which have led to the completion of this research work.

I express my sincere gratitude to my guide Dr. K.Poulose Jacob, Professor, Department of Computer Science, Cochin University of Science and Technology, for his encouragement, guidance and the immense support he extended to me through out. I am privileged to have him as my guide.

I thank Dr. Sumam Mary Idicula, Professor, Department of Computer Science, Cochin University of Science and Technology, for motivating me in pursuing the PhD programme in the department. I am grateful for all her suggestions in my learning process. I am extremely thankful to Mr.K.B.Muraleedharan and Dr.Jacob Philip for the support extended for the completion of my course work.

I had the support and guidance of Dr. Kannan and Dr. Judy, Department of Computer Applications from whom I got constant inspiration to work hard

and to be persistent. Their insightful comments, support and advice has been of immense benefit to me during the most difficult times. They always kept me motivated and inspired during this entire journey.

My sincere thanks is due to Dr. G Santhosh Kumar, Professor, Department of Computer Science, Cochin University of Science and Technology, for all the help he has extended. I express my sincere gratitude to him. I am equally thankful to Dr. Baby.M.D, Dr.M.Bhasi, Dr. K.V. Pramod, Dr. Santosh Kumar M.B, research scholars of Department of Computer Applications and Department of Computer Science for their constant support and motivation. I would like to thank Ms. Liny Varghese who helped me with ideas and final conceptualisation of the thesis.

I am thankful to Mr. Vijay Paul in sharing the data sets during my initial period of research which was extremely helpful in pursuing with the research problem. I am indebted to Mr. Vijay Nair for his data sharing and insightful interventions which helped me to complete the research. I am also thankful to Ms. Delna Gomez for the layout of my thesis.

I have special word of appreciation for Mr. Joe Joseph in providing me the best library support. I also place on record my gratitude to Mr. Lal Paul, Mr. Renjith and Mrs. Manju for providing me all the technical support required for carrying out my research work. I am extremely grateful to Ms.



Girija and all the staff of the department for their encouragement and support. A special gratitude and appreciation is extended to my colleagues and professional friends. My friends are my life. I thank all my friends for the support they had given me in difficult moments. I am deeply indebted to them for this.

My husband Harikrishnan, daughter Keerthana and my father Rajagopal showered unconditional support and encouraged me to pursue my dreams. They have been my source of strength. I could never have reached this point without their encouragement and support. I express my profound gratitude to all who helped me in this journey.

MINI U

## *Abstract*

Computational Social Science is an emerging discipline that uses digital tools to analyze the rich and interactive life we lead. The exponential growth of Internet technologies, World Wide Web and social media in the last few years are generating large amount of social and behavioural data through the social interactions. The data about human behaviour is available for further analysis on a scale never thought about and with tremendous granularity as well as precision. This is helping social scientists to study human behaviour and social interaction in unprecedented detail. These techno-sociological studies need collaborative, interdisciplinary contribution of computer experts, information scientists, physicists, as well as mathematicians and sociologists. The ability to collect and process social data helps researchers to address core questions in social sciences in new ways which opened up nascent areas to explore. Talking to friends in our social network, navigating the Web and forming opinions by listening to others and to the media have become part of daily life. Challenges have arisen recently with the advent of online social media, which produces large amounts of both network and natural language data. Thus understanding, predicting, and enhancing human behaviour in networks pose important research problems for computer and data scientist with practical applications of high impact.

Systems as diverse as genetic networks or the World Wide Web are best described as networks with complex topology.

Computational Social Science uses powerful computer simulations of networks, data collected from online social networks and experiments involving hundreds of thousands of individual conversations to answer questions that were previously impossible to investigate. Networks have been studied as graphs in mathematics, physics, sociology, engineering and computer science, biology and economics. Grounded in graph and system theories, this approach has proven to be a powerful tool for studying networks in physical and social world, including the web. Social media has become an integral part of any business promotions today, unlike the earlier times where it was only considered only as a social networking tool. This growing pace of social media and its impact has evolved in such a way that it is continuously shifting, leaving marketers constantly challenged, and most businesses overwhelmed with the never-ending changes. Social media analysis has become a necessity for business. The challenge for marketers is to find new ways to capture the attention of consumers who are bombarded with too much digital noise or information every day.

## **Objectives of the Research Work**

The objective of the research is to understand the information diffusion process in an online social network, develop a procedure using text mining techniques to do opinion mining, listen to the social media conversations and extract business intelligence. The research is undertaken in three parts:

- Study the application of social network analysis and understand various parameters
- Gain knowledge on how the information is diffused in a social network.
- Investigate the social customer relationship management with a view to evolving a knowledge discovery process leading to business intelligence.

***Information and Activity Diffusion and Propagation:*** The propagation of information and activities through a social network is an area less investigated or studied from a research perspective specifically to India. India being the second largest in the number of Internet users, the business organisations have great interest in such studies to understand the spread of information which may be applied in promoting their products with a view to achieving their goals.

The user interactions are tracked using Facebook Insights. It helps in tracking the number of active users which leads to understanding page performance. Various metrics are analysed to measure the effectiveness of the social media campaign. The social media mining in viral marketing is studied on current and potential adopters. This is compared against the Bass Model, which is the standard diffusion model in order to describe the process of how new products get adopted in the market.

***Crowd sourcing and Opinion mining in Social Media:*** - Customer sentiment analysis is a method of processing information, generally in text format, often from social media sources, to determine customer opinion and responses. Analysis of the data allows organizations to assess whether customer reaction to a new product was positive or negative, or whether owners of a product are experiencing major technical difficulties. Social Customer Relationship Management (CRM) can be integrated with the website to use social networking portals for furthering the business using viral marketing. This is a two sided weapon which can promote your product via a viral positive feedback, or do the opposite by a similar negative one. Hence promoting bi-directional communication between companies and customers using social media and analysing the customer feedback to use crowd sourcing for business benefit is the need of the hour. Analysis of aggregated data over time provides insights into trends, while analysis of individual cases in near real time lets companies address and

resolve customer issues quickly. At the heart of customer sentiment is text analysis, a complex process based on statistical and linguistic analyses.

A tool **SENTIMATCH** is developed to understand opinion, emotion and sentiment in contextual level beyond keyword searching. The domain selected is eco-tourism. This tool is integrated into a portal which maps all the conversation and this promote and monitor public participation. Sentiment analysis is done on these conversations which gives a good snapshot of what the relationships within and between groups in a country are, at a specific point in time. The methods used here can be extended for any products with public opinion.

***Applicability to Social Intelligence:*** Organisations have begun to use social media, to enable participation and knowledge sharing a relatively recent phenomenon, with the aim of improving business operations. Knowledge Management is a process of blending the internal and external information of an organisation to acquire an actionable knowledge using the different forms of technological platform. Social media can support a range of knowledge management (KM) practices. Despite a number of researchers recognizing the importance of social media to KM, there are currently few studies reported. New ways of social interaction through computer systems are transferable from the general context of the Internet to corporate intranets, where they provide support to Knowledge Management.

The framework realises the theory presented in the work. The focus is on possible applications of the automated sentiment analysis and how the framework can be helpful in providing insightful views.

## **CONTENTS**

<b>No.</b>	<b>Chapter</b>	<b>PageNo</b>
<b>CHAPTER – I</b>		
1.	INTRODUCTION	
1.1	Overview	1
1.2	History, Growth and Application of Social Networks	2
1.3	Online Social Networking Sites	4
1.4	Information Mining from Social Media	7
1.5	Business Intelligence from Social Media	10
1.6	Challenges in Processing Social Media	12
1.7	Internet Usage Pattern in India	15
1.8	Addressing the Research Gap	17
1.9	Research objective	20
1.10	Research Questions	21
1.11	Outline of the Thesis	22



## **CHAPTER -II**

2	RELATED WORK AND LITERATURE REVIEW	25
2.1	Introduction	25
2.2	Introduction to Graph Theory and Complex Networks	25
2.3	Properties of Complex Networks	26
2.4	Social Network Sites and Social Media	29
2.5	Diffusion of Information	31
2.6	Metrics	34
2.7	Growth of Online Social Networks	38
2.8	Information Diffusion Models	43
2.9	Visualisation Tools	48
2.10	Sentiment Analysis	49

## **CHAPTER -III**

3	OPINION MINING AND PREDICTIVE ANALYTICS	53
3.1	Social Media Landscape	53
3.2	Scenarios	54
3.3	Sentiment analysis – Work Flow	55

3.3.1	Data acquisition and Pre- Processing	58
3.3.2	Procedure for Analysing Sentiments with semantic search	59
3.4	Ontology Modelling	64
3.5	Computing Similarity	65
3.6	Providing Search Capabilities with Graph	66
3.6.1	Algorithms behind the tagging service	67
3.7	Predictive Analytics and Recommender System	70
<b>CHAPTER -IV</b>		
4	<b>INFORMATION DIFFUSION AND PROPOGATION</b>	72
4.1	Introduction	72
4.2	Social media and brand promotion	73
4.2.1	Facebook as Social Media Platform	74
4.3	Brand Awareness	75
4.4	Modelling Diffusion of Information	77
4.5	Research Methodology	78
4.6	Measuring the impact in Facebook promotion	80
4.7	Diffusion Model	81

4.8	Case Studies	82
4.8.1	Case: 1 Brand Promotion of XYZ Hypermarket	83
4.8.2	Case 2: Brand promotion of a new Cinema Multiplex	86
4.8.3	Case study 3 Brand promotion of a Movie	87
4.9	Discussion	88
4.10	Findings	90
4.10	Conclusion and future directions	92
<b>CHAPTER V</b>		
5.	OPINION MINING FROM SOCIAL CONVERSATIONS	98
5.1	Overview	98
5.2	Challenges	99
5.3	Scenario	99
5.3.1	Developing a Response Review Strategy	101
5.4	SENTIMATCH -The Tool	104
5.5	Experiment set up- Algorithm	108
5.5.1	Software Set up	108

5.5.2	Model Framework for feature extraction	112
5.5.3	Data acquisition and Pre- Processing	112
5.5.4	Data Curation	113
5.5.5	Subjectivity Classification.	115
5.5.6	Tokenising	117
5.5.7	POS Tagger	118
5.5.8	Parsing	120
5.6	Sentimental Analysis and Opinion Mining	122
5.6.1	Named Entity Recogniser	123
5.6.2	Identifying Opinions with feature selection.	124
5.6.3	Rule Engine	124
5.6.4	Feedback	126
5.6.5	Impact of Social Media findings	127
5.7	Data Analysis	128
5.8	Result Analysis	129
5.9	Converting Visitors to Customers	140
5.10	Extracting information from Clickstream data	141
5.10.1	Fitting a Markov Chain	145

5.10.2	Click Prediction	146
5.10.3	Clustering Clickstream Data	146
5.11	Findings	148
5.12	Further Enhancement	149
<b>CHAPTER VI</b>		
6	MONITORING PUBLIC PARTICIPATION IN MULTI-LATERAL INITIATIVES USING SOCIAL MEDIA INTELLIGENCE	150
6.1	Introduction	151
6.2	Collaboration and Consultation Portal	152
6.3	Social Media and Public Participation	154
6.4	Challenges of the Consultation Hub	156
6.5	Related Technologies	157
6.5.1	Social Media Mining	157
6.5.2	Lexical and Quantitative Analysis	158
6.5.3	Text Analysis , Semantic tagging and Analysis	159
6.4	Other Analytic Techniques	160
6.5	Open Source and Commercial Tools Used	160

6.6	Research Methodology	161
6.7	Case study	162
6.8	Architecture of the Automated Monitoring and Evaluation Tool for Multilateral Development Agency	163
6.9	Solution Architecture	164
6.10	Technical Architecture	166
6.11	Working of Consultative Hub Portal	166
6.12	Text Analysis	167
6.13	Deployment of the Collaboration Portal	170
6.14	Findings	172
6.14	Conclusion	175

## **CHAPTER – VII**

7	KNOWLEDGE MANAGEMENT FOR COLLABORATIVE SOCIAL MEDIA INTELLIGENCE	176
7.1	Introduction	176
7.2	Challenges of Knowledge Management	180
7.3	Knowledge Management Framework	182

7.4	Data Collection	185
7.5	Semantic Search platform for the Knowledge Management Solution	186
7.6	Developing and implementing Knowledge Management Frameworks	190
7.6.1	Knowledge Management assessment and benchmarking	190
7.6.1.1	Knowledge Audit	190
7.6.1.2	Needs Assessment	190
7.6.1.3	Readiness Assessment	191
7.6.1.4	Benchmarking of the Current-state of KM	191
7.6.2.	Knowledge Management Strategy development	192
7.6.2.1	Create Knowledge Management Framework	192
7.6.3	Gap Analysis and Change Management Strategy	193
7.7	The impact of social media on knowledge management	193
7.8	Social Media Monitoring	194
7.9	Social Media Intelligence	195

7.10	Findings	196
7.11	Future Thoughts	197
<b>CHAPTER- VIII</b>		
8	CONCLUSION AND FUTURE SCOPE	199
8.1	Main Contributions	200
8.2	Challenges faced	202
8.3	Conclusion	203
8.4	Future thoughts	205
	References	206
	Publications	224



## List of Tables

<b>No.</b>	<b>Table Id</b>	<b>Description</b>
1.	Table 1.1	Internet Users by Country comparison for three years
2.	Table 5.1	Document Level Classification
3.	Table 5.2	Sentence Level Classification
4.	Table 5.3	Sentiment Scores
5.	Table 5.4	Precision and Recall of Facebook Posts
6.	Table 5.6	Most informative features
7.	Table 5.7	Precision and Recall for Positive and Negative reviews
8.	Table 5.8	Precision and Recall for filtered bag of words
9.	Table 5.9	Bigram Co-location
10	Table 5.10	More Informative Features
11	Table 6.1	Co-Relation between Specific words

## List of Figures

<b>No</b>	<b>Fig Id</b>	<b>Description</b>
1.	Fig 1.1	Social Networks
2.	Fig 1.2	Online Social Network Sites
3.	Fig 1.3	Popular Social Networks ranked by number of active users
4.	Fig 1.4	Face Book Users by Country
5.	Fig 3.1	Sentiment Classification Techniques
6.	Fig 3.2	Sentiment Analysis – Workflow
7.	Fig 4.1	Increase in the number of Adopters over a period of one year
8.	Fig 4.2	Increase in the Number of Adopters (in a short span of one month)
9.	Fig 4.3	Daily viral reach of the Page by story type. (Unique Users)
10	Fig 4.4	Daily viral reach of the Page by story type. (Unique Users) for the first month.

11.	Fig 5.1	Flow diagram of the procedure
12.	Fig 5.2	Sentence Classification
13.	Fig 5.3	Document Level Classification
14.	Fig 5.4	Sentence Level Classification
15.	Fig 5.5	UI for Converting Text into Tokens
16.	Fig 5.6	Tokeniser
17.	Fig 5.7	Tagged Text
18.	Fig 5.8	Phrases and Named Entities
19.	Fig 5.9	Parsed Sentence
20.	Fig 5.10	Parse Tree Structure
21.	Fig 5.11	Text Analysis
22.	Fig.6.1	Consultation Hub
23.	Fig. 6.2	Solution Architecture
24.	Fig. 6.3	UI for channel Monitoring

25.	Fig. 6.4	Screen shot of Crowd Sourcing
26.	Fig. 6.5	Crowd Sourcing in Tree Structure
27.	Fig. 6.6	Graphical View of Dashboard
28.	Fig.6. 7	Comments on a consultation
29.	Fig 6.8.	Analysis of Comments
30.	Fig.6. 9.	Result Analysis
31.	Fig.7.1	SECI-model
32.	Fig.7.2	SECI-model in the context of Big Data
33.	Fig.7.3.	Knowledge Management and Social Software
34.	Fig.7.4	Integrated Knowledge Management Cycle

35.	Fig.7.5	Integrating Informal Networks into Knowledge portal
36.	Fig 7.6	Knowledge Portal
37.	Fig 7.7	Social media competitive analytics framework with sentiment benchmarks for industry-specific marketing intelligence.

### **List of Abbreviations**

SNA	Social Network Analysis
OSN	Online Social Networks
SNM	Social Network Mapping
SCC	Strongly connected component
WCC	Weakly connected component
UGC	User Generated Contents
CRM	Customer Relationship Management
KM	Knowledge Management
WOM	Word of Mouth
NER	Named Entity Recognition
KD	Knowledge Discovery
BI	Business Intelligence
UIMA	Unstructured Information Management Applications

CAQDAS	Computer Assisted Qualitative Data Analysis Tools
JDD	Joint Degree Distribution
RST	Rhetorical Structure Theory
PMI	point wise mutual information
SVM	Support vector machine
NB	Naive Bayes
NLP	Natural language processing
IE	Information Extraction
RDF	Resource Description Framework
OWL	Web Ontology Language
TF- IDF	Term frequency – inverse document frequency
URI	Uniform Resource Identifier

FB	Facebook
BM	Bass Model
ROI	Return on Investment
G/SG	Gamma/shifted Gompertz distribution
UIMA	Unstructured Information Management Architecture
NLTK	Natural Language ToolKit
ETL	Extract, Transform and Load
VADER	Valence Aware Dictionary and sentiment Reasoner
POS	Part Of Speech
NER	Named Entity Recognizer
CRM	Customer relationship management
SKOS	Simple Knowledge Organization System



RSS	Rich Site Summary
CSV	Comma Separated Value

