

MODELING CORRELATION BETWEEN SOCIAL CONNECTIONS, TOPICS AND INFORMATION DIFFUSION ON SOCIAL MEDIA

KUNTAL DEY



AMARNATH & SHASHI KHOSLA SCHOOL OF
INFORMATION TECHNOLOGY

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2019

©Indian Institute of Technology Delhi (IITD), New Delhi, 2019

MODELING CORRELATION BETWEEN SOCIAL CONNECTIONS, TOPICS AND INFORMATION DIFFUSION ON SOCIAL MEDIA

by

KUNTAL DEY

Amarnath and Shashi Khosla School of Information Technology

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2019

Dedication

This work is dedicated, with a lot of love and respect to my mother Bibhasri and father Tarak Nath, with a deep gratitude to my wife Sudeshna, and with plenty of affection to my son Monikuntal.

Certificate

This is to certify that the thesis titled “**Modeling Correlation between Social Connections, Topics and Information Diffusion on Social Media**” being submitted by **Kuntal Dey** to the Indian Institute of Technology Delhi for the award of the degree of **Doctor of Philosophy** is a record of original bona fide research work carried out by him. He has worked under my guidance and supervision, and has fulfilled the requirements for the submission of this thesis, which has attained the standard required for a Ph.D. degree of this institute.

The work presented in this thesis has not been submitted to any other university or institute for the award of any other degree or diploma.

Prof. Saroj Kaushik

Department of Computer Science and Engineering

Indian Institute of Technology Delhi

New Delhi 110 016

New Delhi

Acknowledgments

Inception and maturing this doctoral dissertation would have not been possible without the support of several people. I would hereby take the opportunity to express deepest gratitude to them.

First and foremost, I am extremely grateful to my PhD supervisor, Prof. Dr. Saroj Kaushik, Professor and Head of the Department (Computer Science and Engineering), IIT Delhi, for her priceless guidance, consistent encouragement, the deep trust she invested in me, her generous nature that earns an extra-special note of respect from me, and the scholarly inputs that she has provided me throughout the process, including writing scientific papers and this dissertation. Madam was there for me whenever I needed any form of scientific help, technical guidance, logistic support that included all the way from collaborating with other students to ensuring that I get sufficient resources to carry out my work, and on top of it all the positive energy that she passed to me all through the process. She was available for guiding me in spite of her hectic schedule. She is simply incredible, beyond what words can say! My best thanks, respected Madam, for all your blessings - though I know there is no way that I can really thank you enough!

Dr. L V Subramaniam from IBM Research India was instrumental in helping and encouraging me in a wide range of ways. He was also there for me to provide support and inputs whenever I needed, in spite of his busy schedule in IBM Research. Thank you, LVS!

Prof. Dr. Sanjiva Prasad, who was the Head of the Department for Khosla School of Information Technology when I had commenced by PhD, was instrumental in ensuring that my work could start off well on track. Thank you, Prof. Prasad.

I owe a lot to the IIT Delhi Post Graduate Section office, the department staff and a lot of other students in the process. Mr. Rajesh Kumar from IIT Delhi office provided invaluable help in the process of submitting all the semester reports, and also in coordinating so that my administrative process requirements were always taken care of as per the institute schedule.

I also owe to other doctoral students of Prof. Kaushik, namely Shivendra Tiwari, Sunita Tiwari and Priti Jagwani. Sunita and Priti provided me with infinite logistic help and inputs in the process, and along with Shivendra, it was a great team of students working in their respective projects under Prof. Kaushik. I have gained great friends in them, in the process.

I am much indebted to my family. My mother, who used to strongly urge me to do doctoral

studies, and my father, who had also supported the thought till his last day, were instrumental in motivating me. My wife helped me remain motivated, and has taken my habit of working in a completely indisciplined manner with a smiling face on her stride. My son was born at a stage when my work was maturing towards completion, and he brought so much happiness to my life that words cannot express. Overall, had my family not been so supportive, I would not have succeeded in conducting my research towards my doctoral study.

– **Kuntal Dey**

Abstract

The emergence of social discussion platforms has introduced a disruptive change in how information and discussion topics reach people. Interactive online networks such as Facebook, Youtube, Pinterest *etc.*, and microblog platforms, such as Twitter, have been widely adopted by end-users of social media. It is immensely valuable to analyze the user-generated content, the underlying user network, and user attributes, to form a better understanding of the social information movement dynamics, which is the overall aim of this work.

The thesis performs analysis of Twitter text (tweets) and social connections (followerships), from the perspective of modeling correlations between (a) social connections and structures such as communities, (b) the lifecycle of topics of discussion, and (c) the movement of topical information on the social networks. In this work, we provide a novel definition of topics using user-given explicit hashtags, latent semantic content and temporal continuity of the content, and empirically discover and model several interesting and correlated attributes of such topics. This includes analyzing the lifecycle of these topics, assessing user roles in the lifecycle, finding the impact of social communities in such lifecycles, finding topical homophily and discovering its characteristics, modeling topical information flow on the social network and assigning hashtags to tweets for better topic discovery. Prior to that, we establish in the background that a correlation exists between the language usage of individual users of social networks, and the social communities they implicitly belong to. In one direction, we observe that, language used by individuals is influenced by the language used by the social communities they belong to. In the other direction, we empirically show that, language usage similarity is a reasonable approximator for social communities. Subsequently, the core of the work is laid out over the following research modules.

Topic Finding and Topic Lifecycle Analysis: Since topics are not well-defined in Twitter and there is no universally accepted external definition, we propose a novel definition of topics, using a combination of user-given explicit hashtags, implicit (latent) content semantics and temporal continuity of content semantics. Further, we model the lifecycle of topics thus defined, and perform an exploration towards user participation in the topics. The lifecycle analysis addresses how these topics get born, become prominent, and evolve in terms of its constituents (thereby, morph), rather than simply getting born and dying too soon. The user participation angle of our analysis shows

how the behavior characteristics of the participants play roles in shaping up the lifecycle of topics thus derived.

Topical Homophily and Its Characterization: Another core contribution made by the present work is towards defining and identifying topical homophily, and characterizing such homophily. As per the classical definition of homophily, “similarity breeds familiarity”. Our work defines topical homophily as familiarity being bred by user pair similarity computed over their participation in similar topics. We empirically demonstrate the existence of topical homophily, and further show that such homophily is characterized by its non-monotonic nature, as well as, it peaks faster for tighter definitions of topical similarity of semantic concepts.

Topic Lifecycle as an Outcome of Social Community Dynamics: The novel definition of topics, and the significant presence of topical homophily at different levels of familiarity, motivate us to investigate whether social communities, rather than only individual users in isolation (discounting social connections), have a role to play in lifecycle of topics. We show that, topics associate with communities at any given time, by modeling the correlation between social communities and the constituent hashtags of the topics. We show that the constituent hashtags are independently used across communities, and the topics evolve atomically (independently) within communities. Empirical demonstration of the lifecycle of topics as a property of social community dynamics is novel in the literature.

Topical Information Diffusion: Information movement is one of the core properties of social networks. The prior works mainly account for two factors: (a) the movement of information along the social connections and (b) individual users getting sufficiently excited (“activated”) to move the received information further forward. On the other hand, we believe that whether or not a user chooses to forward the information they receive will also depend upon the user’s affinity towards the tweet information topic content. We thereby propose *TopSPA*, a novel information diffusion model that underneath uses the spreading activation (SPA) technique, but blends it with user affinity towards its information content.

Topic Augmentation: While the use of hashtags is integral to our definition of topics, a large fraction of the tweets does not contain any hashtag. This requires assigning hashtags to such tweets, which is equivalent to recommending hashtags for tweets. We propose a model using a combination of three components: NLP, a temporal factor and a socio-temporal factor. The NLP factor analyzes the semantic content of the test tweet with respect to the concept space covered known training tweets. The socio-temporal factor assumes that, individuals are more likely to use hashtags that have been used more recently, by their stronger social connections. The temporal factor assumes that, new topics that break out as bursts in real life are likely to reach individual users from other sources, such as the outside world, instead of via social friends.

Our work gives its end-user to an end-to-end execution pipeline to correlate social connections,

topics and information diffusion on Twitter. The key contribution of this thesis is an empirical establishment of two fundamental hypotheses: (a) topics and their lifecycles are community-level properties, and (b) information diffusion is governed by the topic of the content.

अध्याय सार

सामाजिक चर्चा प्लेटफार्मों के उद्भव ने लोगों में जानकारी और चर्चा के विषय कैसे पहुंचते हैं, में एक विघटनकारी परिवर्तन पेश किया है।

फ़ेसबुक, यूट्यूब, इत्यादि जैसे इंटरएक्टिव ऑनलाइन नेटवर्क और ट्विटर जैसे माइक्रोब्लॉग प्लेटफ़ॉर्म को व्यापक रूप से सोशल मीडिया के उपयोगकर्ताओं द्वारा अपनाया गया है।

सामाजिक जानकारी आंदोलन की गतिशीलता की बेहतर समझ बनाने के लिए, उपयोगकर्ता द्वारा उत्पन्न सामग्री, अंतर्निहित उपयोगकर्ता नेटवर्क और उपयोगकर्ता विशेषताओं का विश्लेषण करना बेहद मूल्यवान है, जो इस काम का समग्र उद्देश्य है।

थीसिस ट्विटर टेक्स्ट (ट्वीट) और सोशल कनेक्शन (फेलोशिप) का विश्लेषण करती है, (ए) सामाजिक कनेक्शन और संरचनाओं जैसे समुदायों के बीच संबंध के परिप्रेक्ष्य से (बी) चर्चा के विषयों का जीवनचक्र, और (सी) सामाजिक नेटवर्क पर सामयिक जानकारी की आवाजाही।

इस काम में, हम उपयोगकर्ता द्वारा दिए गए स्पष्ट हैशटैग, अव्यक्त शब्दार्थ सामग्री और सामग्री की लौकिक निरंतरता का उपयोग करते हुए विषयों की एक उपन्यास परिभाषा प्रदान करते हैं, और ऐसे विषयों के कई रोचक और सहसंबद्ध विशेषताओं को अनुभव और खोज करते हैं।

इसमें इन विषयों के जीवन चक्र का विश्लेषण करना, जीवनचक्र में उपयोगकर्ता की भूमिका का आकलन करना, ऐसे जीवनचक्र में सामाजिक समुदायों के प्रभाव का पता लगाना, सामयिक समरूपता की खोज करना और इसकी विशेषताओं की खोज करना, सामाजिक नेटवर्क पर सामयिक जानकारी का प्रवाह और बेहतर विषय के लिए ट्वीट के लिए हैशटैग प्रदान करना शामिल है। खोज।

इससे पहले, हम पृष्ठभूमि में स्थापित करते हैं कि सामाजिक नेटवर्क के व्यक्तिगत उपयोगकर्ताओं की भाषा के उपयोग के बीच एक सहसंबंध मौजूद है, और वे जो सामाजिक समुदाय हैं, उनका संबंध है।

एक दिशा में, हम मानते हैं कि, व्यक्तियों द्वारा उपयोग की जाने वाली भाषा उन सामाजिक समुदायों द्वारा उपयोग की जाने वाली भाषा से प्रभावित होती है, जिनसे वे संबंधित हैं।

दूसरी दिशा में, हम अनुभवजन्य रूप से बताते हैं कि, भाषा का उपयोग समानता सामाजिक समुदायों के लिए एक उचित सन्निकटन है।

इसके बाद, निम्नलिखित अनुसंधान मॉड्यूल पर काम का मूल निर्धारित किया गया है।

विषयगत खोज और विषय जीवनचक्र विश्लेषण:

चूंकि विषय ट्विटर में अच्छी तरह से परिभाषित नहीं हैं और कोई सार्वभौमिक रूप से स्वीकृत बाहरी परिभाषा नहीं है, इसलिए हम उपयोगकर्ता द्वारा दिए गए स्पष्ट हैशटैग, निहित (अव्यक्त) सामग्री शब्दार्थ और सामग्री शब्दार्थ की अस्थायी निरंतरता के संयोजन का उपयोग करते हुए विषयों की एक अच्छी तरह से गोल परिभाषा का प्रस्ताव करते हैं। ।

इसके अलावा, हम इस प्रकार परिभाषित विषयों के जीवनचक्र को मॉडल करते हैं, और विषयों में उपयोगकर्ता की भागीदारी के लिए एक अन्वेषण करते हैं।

जीवनचक्र विश्लेषण से पता चलता है कि ये विषय कैसे पैदा होते हैं, प्रमुख बनते हैं, और अपने घटकों (इसलिए, रूप) के संदर्भ में विकसित होते हैं, बजाय केवल जन्म लेने और मरने के जल्द ही।

हमारे विश्लेषण के उपयोगकर्ता सहभागिता कोण से पता चलता है कि प्रतिभागियों के व्यवहार की विशेषताएं इस प्रकार व्युत्पन्न विषयों के जीवनचक्र को आकार देने में भूमिका निभाती हैं।

सामयिक होमोफिली और इसकी विशेषता:

वर्तमान कार्य द्वारा किया गया एक अन्य मुख्य योगदान सामयिक होमोफिली को परिभाषित करने और पहचानने की ओर है, और इस तरह के होमोफिली की विशेषता है।

समरूपता की शास्त्रीय परिभाषा के अनुसार, "समानता नस्ल परिचित"।

हमारे काम में सामयिक समरूपता को परिभाषित किया जाता है क्योंकि उपयोगकर्ता जोड़ी द्वारा समान विषयों में उनकी भागीदारी पर गणना की गई समानता से परिचित किया जाता है।

हम अनुभवजन्य रूप से सामयिक होमोफिली के अस्तित्व को प्रदर्शित करते हैं, और आगे बताते हैं कि इस तरह के होमोफिली को गैर-मोनोटोनिक प्रकृति की विशेषता है, साथ ही, यह अर्थ संबंधी अवधारणाओं की सामयिक समानता की तंग परिभाषाओं के लिए तेजी से चोट करता है।

सामाजिक सामुदायिक गतिशीलता के परिणाम के रूप में विषय

विषयों की उपन्यास परिभाषा, और परिचित के विभिन्न स्तरों पर सामयिक समरूपता की महत्वपूर्ण उपस्थिति, हमें यह जांचने के लिए प्रेरित करती है कि क्या सामाजिक समुदाय, अलगाव में केवल व्यक्तिगत उपयोगकर्ताओं (सामाजिक कनेक्शनों को छूट देने) के बजाय, विषयों के जीवनचक्र में भूमिका निभाते हैं।

हम बताते हैं कि, विषय किसी भी समय समुदायों के साथ जुड़ते हैं, सामाजिक समुदायों और विषयों के घटक हैशटैग के बीच संबंध को मॉडलिंग करके।

हम बताते हैं कि घटक हैशटैग स्वतंत्र रूप से समुदायों में उपयोग किए जाते हैं, और विषय समुदायों के भीतर परमाणु (स्वतंत्र रूप से) विकसित होते हैं।

सामाजिक सामुदायिक गतिकी की संपत्ति के रूप में विषयों के जीवनचक्र का अनुभवजन्य प्रदर्शन साहित्य में उपन्यास है।

सामयिक सूचना प्रसार:

सूचना आंदोलन सामाजिक नेटवर्क के मुख्य गुणों में से एक है।

पूर्व काम मुख्य रूप से दो कारकों के लिए जिम्मेदार हैं: (ए) सामाजिक कनेक्शन और (बी) के साथ जानकारी के आंदोलन व्यक्तिगत रूप से प्राप्त जानकारी को आगे बढ़ाने के लिए पर्याप्त रूप से

उत्साहित (" सक्रिय ") हो रहे हैं।

दूसरी ओर, हम मानते हैं कि उपयोगकर्ता प्राप्त जानकारी को अग्रेषित करना चाहते हैं या नहीं, यह भी ट्वीट सूचना विषय सामग्री के प्रति उपयोगकर्ता की आत्मीयता पर निर्भर करेगा।

इसके बाद, हम एक उपन्यास सूचना प्रसार मॉडल का प्रस्ताव करते हैं, जो कि प्रसार सक्रियण (एसपीए) तकनीक का उपयोग करता है, लेकिन इसे अपनी सूचना सामग्री के प्रति उपयोगकर्ता के साथ मिश्रण करता है।

विषयगत विचलन:

जबकि हैशटैग हमारे विषयों की परिभाषा का अभिन्न अंग है, लेकिन ट्वीट्स के एक बड़े हिस्से में कोई हैशटैग नहीं है।

इसके लिए ऐसे ट्वीट्स को हैशटैग असाइन करना पड़ता है, जो ट्वीट्स के लिए हैशटैग की सिफारिश करने के बराबर है।

हम तीन घटकों के संयोजन का उपयोग करके एक मॉडल का प्रस्ताव करते हैं: एनएलपी, एक अस्थायी कारक और एक सामाजिक-अस्थायी कारक।

एनएलपी कारक परीक्षण स्थान के सिमेंटिक कंटेंट का विश्लेषण कॉन्सेप्ट स्पेस कवरेड ट्रेनिंग ट्वीट के संबंध में करता है।

सामाजिक-अस्थायी कारक मानता है कि, व्यक्ति अपने मजबूत सामाजिक कनेक्शन द्वारा हाल ही में उपयोग किए गए हैशटैग का उपयोग करने की अधिक संभावना रखते हैं।

लौकिक कारक यह मानता है कि, नए विषय जो वास्तविक जीवन में फटने के रूप में सामने आते हैं, सामाजिक स्रोतों के माध्यम से अन्य उपयोगकर्ताओं, जैसे कि बाहरी दुनिया, से अलग-अलग उपयोगकर्ताओं तक पहुंचने की संभावना है।

हमारा काम ट्विटर पर सामाजिक कनेक्शन, विषयों और सूचना प्रसार को सहसंबंधित करने के लिए अपने अंतिम-उपयोगकर्ता को एंड-टू-एंड निष्पादन पाइपलाइन देता है।

इस थीसिस का मुख्य योगदान दो मौलिक परिकल्पनाओं का एक अनुभवजन्य स्थापना है: (ए) विषय और उनके जीवनचक्र समुदाय स्तर के गुण हैं, और (बी) सूचना प्रसार सामग्री के विषय द्वारा शासित है।

Table of Contents

Certificate	i
Acknowledgments	iii
Abstract	v
Table of Contents	ix
List of Figures	xvii
List of Tables	xix
Acronyms	xxi
Chapter 1 Introduction	1
1.1 The Primitive Constituents of Twitter	2
1.2 Deriving Non-Primitive Constituents of Twitter	3
1.3 Computational Disciplines of Research	7
1.4 Twitter Socio-Topical Information Research Pipeline	8
1.5 Research Literature Addressing Twitter	10
1.6 Open Challenges in Twitter Research	13
1.6.1 Topic Detection and Spreading	13
1.6.2 Evolution of Topics - Lifecycle of Topical Discussions	16
1.6.3 Predicting Social Structures	16
1.6.4 Impact of Social Structures on Twitter Dynamics	17
1.6.5 The Role Played by Social Influence	18
1.6.6 Homophily	19
1.6.7 Data Augmentation with Hashtag Assignment	19
1.7 Motivation	20
1.8 Background Work: Interplay of Language and Communities on Twitter	21
1.9 Overview of Our Work	26

1.10 Conclusion	30
Chapter 2 Topic Finding and Topic Lifecycle Analysis	33
2.1 Introduction	33
2.2 Related Work	36
2.3 Overview of Solution Approach	37
2.4 Defining Topics	38
2.4.1 Semantically Related Hashtag Identification	39
2.4.2 Topic Cluster Creation	40
2.5 Topic Lifecycle Analysis	42
2.5.1 Detecting Morphing of Topics over Dominant Hashtags	44
2.6 Assessing User Roles in Topic Lifecycles	45
2.6.1 User Timeline Creation for Their Participation in Each Topic	45
2.6.2 User Role Analysis	46
2.7 Graph Characterization	47
2.8 Experiments	47
2.8.1 Data Description	48
2.8.2 Experimental Setup	48
2.8.3 Characterizing Topic Lifecycle: Dominance, Evolution and Intensity	50
2.8.4 Performing Qualitative Examination on an Example Topic	50
2.8.5 Characterizing Participation of Users	52
2.8.6 Characterization at the Graph Level	54
2.8.7 User Characterization using Social Graphs	56
2.9 Conclusion	58
Chapter 3 Correlating Topical Similarity with Familiarity	61
3.1 Introduction	61
3.2 Related Work	63
3.3 Solution Approach	64
3.3.1 Finding User Participation in Topics	65
3.3.2 User Pair Similarity Finding	66
3.3.3 Correlating Similarity with Familiarity	66
3.4 Experiments	67
3.4.1 Dataset Description	68
3.4.2 Experimental Setup	68
3.4.3 Validating the First Hypothesis	70
3.4.4 Validating the Second Hypothesis	70

3.4.5	Validating the Third Hypothesis	72
3.4.6	Community-Level Familiarity Based Topical Homophily	72
3.5	Outcome Summary of Investigation of Topical Homophily	72
3.6	Conversation Politeness of Topical Opinion Homophily Groups	73
3.7	Problem Overview: Politeness in Topical Opinion Homophily Groups	74
3.8	Related Work: Politeness in Topical Opinion Homophily Groups	75
3.9	Solution Approach: Politeness in Topical Opinion Homophily Groups	76
3.9.1	Computing Semantic Similarity using Content Embedding as Input	76
3.9.2	Pairwise Politeness versus Stance Similarity	79
3.9.3	Politeness versus Stance Similarity Given Social Familiarity	79
3.10	Experiments: Politeness in Topical Opinion Homophily Groups	80
3.10.1	Tools and Resources: Politeness in Topical Opinion Homophily Groups	80
3.10.2	Correlating Stance Similarity and Social Familiarity given Pairwise Politeness	81
3.10.3	Pairwise Politeness vs. Stance Similarity Given Communities	81
3.11	Conclusion	83
Chapter 4	Topic Lifecycle: Function of Social Communities	85
4.1	Introduction	85
4.2	Related Work	88
4.3	Solution Approach	89
4.3.1	Creating Topic and Hashtag Timelines for Communities	91
4.3.2	Community-Centric Analysis of Topic Lifecycles	93
4.3.3	Methodology to Validate Our Hypotheses	94
4.4	Experiments	97
4.4.1	Dataset	97
4.4.2	Experimental Setup	97
4.4.3	Experimental Validation of Our Research Hypotheses	98
4.5	Conclusion	104
Chapter 5	Topical Information Diffusion	107
5.1	Introduction	107
5.2	Related Work	109
5.3	Solution Approach	112
5.3.1	Overview of Our Approach	112
5.3.2	The Traditional SPA Model	112
5.3.3	Proposing the <i>TopSPA</i> Model	116
5.4	Experiments	122

5.4.1	Data Description	122
5.4.2	Identifying Semantic Concepts (LDA-Topics)	122
5.4.3	User Seeding and Assigning Initial Energy	123
5.4.4	Executing and Validating <i>TopSPA</i>	123
5.4.5	Comparing <i>TopSPA</i> vs. Traditional SPA Performances	124
5.5	Conclusion	126
Chapter 6 Data Augmentation by Hashtag Assignment		129
6.1	Introduction	129
6.2	Related Work	131
6.3	Solution Approach: SemTagger	134
6.3.1	Data Cleaning	134
6.3.2	Preprocessing	135
6.3.3	Training	135
6.3.4	Recommending Hashtags	136
6.4	Solution Approach: EmTagger	139
6.4.1	Model 1: Learning Word Embeddings Per Hashtag	139
6.4.2	Testing Model 1	141
6.4.3	Model 2: Learning Word Embeddings Across Hashtags	142
6.4.4	Testing Model 2	145
6.5	Solution Approach: SocioTempR	146
6.5.1	Modeling the Temporal Burst Aspect	148
6.5.2	Modeling the Socio-Temporal Aspect	149
6.5.3	Combining the Scores and Generating Final Ranks	150
6.6	Experiments	150
6.6.1	Dataset Description	151
6.6.2	Tools and Resources	152
6.6.3	Validation Metrics	152
6.6.4	Tuning Hyperparameters	153
6.6.5	Establishing the Baseline and Defining Lift	154
6.6.6	Selecting between SemTagger and EmTagger	155
6.6.7	Executing SocioTempR	155
6.6.8	Comparing with The Literature	158
6.7	Conclusion	159
Chapter 7 Conclusions and Future Directions		161

Bibliography	163
Appendices	183
Appendix A Dataset Used for Experiments	185
A.1 Data Source	185
A.2 Data Preprocessing	186
A.3 Data Overview	186
A.4 Data Characteristics	186
Appendix B Correlating Language Usage and Communities	189
B.1 Introduction	189
B.2 Introduction: Evolution of Language Usage of Users in Community	190
B.3 Related Work: Evolution of Language Usage of Users in Community	191
B.3.1 Models for Language Usage Change of Users	191
B.3.2 Influence Models	192
B.3.3 User Behavior Evolution Models	192
B.4 Approach: Evolution of Language Usage of Users in Community	193
B.4.1 Creating the <i>Man-O-Meter</i> Model Framework	193
B.4.2 Metrics for Characterization	196
B.4.3 Predicting Ranks	197
B.5 Experiments: Evolution of Language Usage of Users in Community	199
B.5.1 Identifying Experts	199
B.5.2 Data Description	200
B.5.3 Experimental Setup	200
B.5.4 LDA-Topic Detection and User-Topic Association Computation	200
B.5.5 Examining User Interest	202
B.5.6 Interest at Category Level	203
B.5.7 Rank Prediction	204
B.6 Summary: Evolution of Language Usage of Users in Community	206
B.7 Introduction: Language Usage Similarity of Users for Approximating Communities	206
B.8 Related Work: Language Usage Similarity of Users for Approximating Communities	208
B.9 Approach: Language Usage Similarity of Users for Approximating Communities	209
B.9.1 Creating Documents and Extracting LDA-Topics	210
B.9.2 Creating Content Graph	211
B.9.3 Social Quality Measurement of Predicted Links	213
B.10 Experiments: Language Usage Similarity of Users for Approximating Communities	215

B.10.1 Data Description	215
B.10.2 Tools Used	215
B.10.3 Language Model Evaluation	216
B.10.4 Evaluation of Granularities	216
B.10.5 Community Structure Inspection	216
B.11 Summary: Language Usage Similarity of Users for Approximating Communities . . .	217
B.12 Conclusion	219
Appendix C Stance Detection	221
C.1 Introduction	221
C.2 Related Work	223
C.3 Central Idea	226
C.3.1 Preprocessing	227
C.3.2 The Deep Neural Network Model	227
C.3.3 The Traditional Machine Learning Model	229
C.4 Experiments	232
C.4.1 Dataset Description	232
C.4.2 Comparing our Best Models with The Literature	233
C.4.3 Fine-Tuning of our DL Model	233
C.4.4 Impact of The Features	234
C.4.5 Generalization of Our ML and DL Models	236
C.5 Discussion	237
C.5.1 Novel, Effective and Ineffective ML Features	238
C.5.2 DL: Approach and Effectiveness	238
C.5.3 Choice of Learning Model: Impact of Data Availability	238
C.5.4 Distribution of Errors	238
C.5.5 Future Work	239
C.6 Conclusion	239
Appendix D Paraphrase and Semantic Similarity Detection . . .	241
D.1 Introduction	241
D.2 Related Work	243
D.2.1 Prior Work on Clean Text Corpus	243
D.2.2 Prior Work on Noisy Short-Text Corpus	244
D.3 Solution Approach	246
D.3.1 Preprocessing	246
D.3.2 Feature Selection	248

D.3.3	Learning	251
D.4	Experiments	251
D.4.1	Dataset Description	252
D.4.2	Tools and Resources	252
D.4.3	System Performance	253
D.4.4	Contrasting Our Solution with Existing Ones	254
D.4.5	Examining the Generalization of Our Solution on Clean-Text Corpus	255
D.5	Conclusion	257
Publications out of Thesis		259
Brief Résumé		261

List of Figures

Figure 1.1	Research Space Overview	10
Figure 1.2	TIR and TIRO Based User Divergences	23
Figure 1.3	Range of Average Cross Entropies of Users for Different Categories	24
Figure 1.4	Content Models as Approximators of Social Structures	25
Figure 1.5	Effect of Corpus Granularity Selection	26
Figure 1.6	A Visual Overview of Our Contribution	30
Figure 2.1	Overall Workflow of Our Approach	38
Figure 2.2	Workflow: Topic Detection and Topic Lifecycle Analysis	43
Figure 2.3	Creation of Various Timelines	46
Figure 2.4	Topic Intensity Over Time: Two Datasets	51
Figure 2.5	Evolution of Dominant Hashtags: Two Datasets	52
Figure 2.6	Qualitative Analysis of Lifecycle of A Hashtag	53
Figure 2.7	User Distribution Using Dominant vs. Other Hashtags	54
Figure 2.8	User Influence Computed Using Page Rank	55
Figure 2.9	Degree Distribution: Two Datasets	56
Figure 2.10	Network Graph Characteristics: June Data, 400-Topic Granularity	56
Figure 2.11	Network Graph Characteristics: June Data, 500 Topic-Granularity	57
Figure 2.12	Network Graph Characteristics: July Data, 1600-Topic Granularity	57
Figure 2.13	Network Graph Characteristics: July Data, 2000-Topic Granularity	57
Figure 2.14	Average Degree of Users Using Only Dominant vs. Only Other Hashtags	58
Figure 3.1	Solution Architecture for Topical Homophily Finding	65
Figure 3.2	Topical Homophily using Edge Level Familiarity	68
Figure 3.3	Average Topical Similarity vs. Mutual Friends: June	69
Figure 3.4	Average Topical Similarity vs. Mutual Friends: July	70
Figure 3.5	Modularity: Topical Communities vs. Underlying Familiarity Graph	71
Figure 3.6	User Pair Stance Similarity and Interaction Politeness Distribution	82
Figure 4.1	Our Solution Architecture	90

Figure 4.2	Hashtag-Level Community Timeline Creation	92
Figure 4.3	Topic-Level Community Timeline Creation	93
Figure 4.4	Temporal Evolution: Given Pairs of Semantically Related Hashtags	98
Figure 4.5	Time Series of Hashtags of the <i>Iran</i> Topic	100
Figure 4.6	Time Series of the <i>Iran</i> Topic	101
Figure 4.7	Time Series of Hashtags of the <i>Drugs</i> Topic	102
Figure 4.8	Time Series of the <i>Drugs</i> Topic	103
Figure 5.1	Architecture of <i>TopSPA</i>	113
Figure 5.2	Schematic Diagram of SPA Energy Dissipation	114
Figure 5.3	Initial Seed Energy Assignment to End-Users	118
Figure 5.4	Schematic Diagram of <i>TopSPA</i> Energy Dissipation	120
Figure 5.5	<i>Top SPA</i> vs. SPA Using Mean Precision on Libya data	124
Figure 5.6	<i>Top SPA</i> vs. SPA Using Mean Precision on Egypt data	125
Figure 5.7	<i>Top SPA</i> vs. SPA Using Jaccard Coefficient on Libya data	125
Figure 5.8	<i>Top SPA</i> vs. SPA Using Jaccard Coefficient on Egypt data	126
Figure 6.1	Data Cleaning and Preprocessing	134
Figure 6.2	Training and Representative Tweet Set Identification	135
Figure 6.3	Semantic Similarity Based Hashtag Recommendation	138
Figure 6.4	Joint Distribution Maximization Based Hashtag Recommendation	138
Figure 6.5	System Architecture Diagram with EmTaggeR Model 1	140
Figure 6.6	System Architecture Diagram with EmTaggeR Model 2	144
Figure B.1	Histogram: Number of Tweets Across Each Time Segment	201
Figure B.2	Cross Entropy Range for Users Corresponding to Top 5 <i>TIRO</i>	202
Figure B.3	Divergence Range for Users and Topics Corresponding to Top 5 <i>TIR</i>	202
Figure B.4	Range of Average Cross Entropies of Users for Different Categories	203
Figure B.5	Average Loss Function of Model Over Different Categories	205
Figure B.6	Average Zero-One Loss Function of Model Over All Categories	205
Figure B.7	The Architecture of Our Link Prediction Framework	210
Figure B.8	Content Models as Approximators of Social Structures	217
Figure B.9	Effect of Corpus Granularity Selection	218
Figure C.1	Our Deep Neural Network Based Solution Architecture	228
Figure C.2	Our Traditional Machine Learning Based Solution Architecture	229
Figure D.1	Architecture of Our System	246

List of Tables

Table 2.1	Data Description	48
Table 2.2	Examples of Topic Clusters	49
Table 3.1	Distribution of Similarity, Familiarity and Politeness	80
Table 3.2	An Intuitive View of The Raw Results	83
Table 3.3	Stance Similarity and Interaction Politeness at a Community Level	83
Table 4.1	Examples of Clusters with Different k-Values	98
Table 5.1	The Egypt and Libya Datasets	122
Table 5.2	<i>TopSPA</i> vs. SPA Using Mean Precision	123
Table 5.3	<i>TopSPA</i> vs. SPA Using Jaccard Coefficient	124
Table 6.1	Main Dataset for Hashtag Assignment Experiments	151
Table 6.2	Auxiliary Dataset for Comparing SemTagger and EmTaggeR	152
Table 6.3	F1 Scores of the Two EmTaggeR Models	154
Table 6.4	Comparing F1 Scores over M Using ALOC	156
Table 6.5	Comparing F1 Scores over M Using MuC	157
Table 6.6	Results for the Top-K Based Method Using ALOC	157
Table 6.7	Results for the Top-K Based Method Using MuC	157
Table 6.8	MuC Lifts Over the LDA Baseline	158
Table A.1	Data Description	186
Table A.2	Examples of Topic Clusters	187
Table B.1	Topic Association Vector of User and Corpus: An Example	196
Table B.2	Data Collection Categories	199
Table B.3	Dataset Statistics	200
Table B.4	Three Randomly Selected Example Topics	201
Table B.5	Our SVM ^{Rank} Execution Setup	204
Table B.6	Basic Statistics of the Event Datasets	215

Table C.1	Example Tweets with Stances	226
Table C.2	Description of the Stance Detection Dataset	232
Table C.3	Performance of Stance Detection Solutions	234
Table C.4	Performance of The Two-Phase DL Solutions	235
Table C.5	F1 Scores for DL Stance Detection Models	235
Table C.6	Ablation Test for ML Approach	235
Table C.7	Comparing with the Literature for Generalization	236
Table C.8	Opinion Towards Target vs. Others	237
Table D.1	The SemEval 2015 Benchmark Dataset	252
Table D.2	Feature Ablation Test	253
Table D.3	Examples of Outputs and Their Correctness	254
Table D.4	Comparing Performances of Our Solution with Others	255

Acronyms

AI	Artificial Intelligence
API	Application Program Interface
ASCII	American Standard Code for Information Interchange
BGLL	Blondel, Guillaume, Lambiotte and Lefebvre
BLEU	Bilingual Evaluation Understudy
BOW	Bag Of Words
CE	Cross Entropy
CNN	Convolutional Neural Network
CoNLL	Computational Natural Language Learning
DL	Deep Learning
IOB	Inside-Outside-Beginning
KSC	K-Spectral Centroid
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
LSTM	Long Short-Term Memory
MALLET	MAchine Learning for Language Toolkit
ML	Machine Learning
MPQA	Multi-Perspective Question Answering
NBOW	Neural Bag Of Words
NE	Named Entity
NER	Named Entity Recognizer
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
NMI	Normalized Mutual Information
OSN	Online Social Network
PoS	Part of Speech
REST	Representational State Transfer

RNN	Recurrent Neural Network
RT	Retweet
SAX	Symbolic Aggregate ApproXimation
SCC	Strongly Connected Component
SGD	Stochastic Gradient Descent
SIR	Susceptible-Infectious-Recovered
SIS	Susceptible-Infectious-Susceptible
SNA	Social Network Analysis
SPA	Spreading Activation
SVM	Support Vector Machine
USA	United States of America
VB	Verb, Base Form
VDN	Verb, Past Participle
VBP	Verb, Non-3rd Person Singular Present
VBZ	Verb, 3rd Person Singular Present
WCC	Weakly Connected Component