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

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 stuff 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 it 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।

<u>सामयिक होमोफिली और इसकी विशेषता:</u>
वर्तमान कार्य द्वारा किया गया एक अन्य मुख्य योगदान सामयिक होमोफिली को परिभाषित करने और पहचानने की ओर है, और इस तरह के होमोफिल्ली 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।

<u>सामाजिक सामुदायिक गतिशीलता के परिणाम के रूप में विषय</u>
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।

<u>सामयिक सूचना प्रसार:</u>
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

List of Figures

List of Tables

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
VBN	Verb, Past Participle
VBP	Verb, Non-3rd Person Singular Present
VBZ	Verb, 3rd Person Singular Present
WCC	Weakly Connected Component