國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

新議題在微網誌上的傳播預測

Novel Topic Information Diffusion Prediction on Microblog

洪三權

San-Chuan Hung

指導教授：林守德 博士

Advisor: Shou-De Lin, Ph.D.

中華民國 102 年 1 月

January, 2013

致謝

感謝家人的在精神與物質上支持，人過了 20 歲還能免於經濟壓力煩惱而可以專心念書，是件幸福且奢侈的事。

感謝林守德老師在研究上的教導，增進我對數據的敏感神經，使我學會從數據當中觀察現象，說出一個有道理的故事的方法；也感謝資訊網路與多媒體研究所教過我的老師們，使我在資訊科學有所精進。

感謝社會系的劉華真老師於兩年前願意雪中送炭幫我寫推薦信，使我有進入網媒所的條件。

感謝實驗室的夥伴， 與你們一起討論課業、做研究、玩耍，讓碩士漫長兩年生涯增色不少。特別感謝 Tim 學長和 Weidong，謝謝你們願意和我討論，從你們身上學到很多東西。

感謝 Python 的作者 Guido van Rossum 開發優雅且簡潔的語言，以及 Social Network 社群開發 Python Lib 的努力，特別是 Networkx 。缺乏 Python 和 Networkx 強大的火力支援，研究的路途想必將會深陷於 Segmentation Fault 和 Memory 不足的幽暗深谷。

中文摘要

新議題的傳播預測是社會網路分析領域新興且重要的問題。新議題由於缺乏過去的傳遞記錄，而其傳遞模式變得更加難以預測。相較於過去的研究，我們運用議題語義相似性，並結合異質網路點類型資訊，在更現實的情境條件上解決本問題。根據實驗結果，結合新模型與過去舊有的模型，在評鑑指標：「接收者操作特徵曲線下面積」有 3.4%的增進。

關鍵字：訊息傳遞預測、新議題、社會網路

Abstract

This work brings a marriage of two seemly unrelated topics, natural language processing (NLP) and social network analysis (SNA). Information diffusion prediction on novel topic is a challenging and important task in SNA which is to predict the diffusion of a new topic, and we design a learning-based framework to solve this problem. We open the scenario into a more realistic setting, and exploit the latent semantic information among users, topics, social connections, and heterogeneous node types as features for prediction. Our framework is evaluated on real data collected from public domain. The experiments show 3.4% AUC enhancement from baseline methods.

Keyword: Information diffusion prediction, novel topic, social network

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In recent year, microblog, such as Plurk and Twitter, has become more and more important as online service in daily life. According to Twitter official statistic report in 2006 [1], there were 140 million active users and 340 million Tweets (messages) a day, and the number has been still increasing. Similarly, Plurk is a popular microblog service in Asian with more than 5 million users [2]. Nowadays, microblog has become popular Internet service.

Microblog is a novel Internet communication service. Users can post a message to describe their present status with limited words (only 140 words per message allowed in Twitter). Besides, users can connect with each other in a follower-followee relationship, where follower can subscribe interesting people as their followee such that they can read the message posted by followee instantly and response it.

Today, microblog becomes a common platform for spreading information. People can get latest pop music news of Lady Gaga, Justin Bieber, or Korean famous singer Psy just by following their Twitter accounts. Not only soft information like music would be shared in microblog, but even hard information: politics. American president candidates in 2012, Obama and Romney, both launched Twitter accounts to send instant

messages to their supporters; meanwhile, Taiwan politicians, no matter KMT or DPP, use Plurk accounts to communicate to their supporters.

Thanks to the great popularity of microblog, large amount of data generated by microblog users can help researchers to study phenomena on social network. One of important problems is “diffusion prediction”: how is meme, information, or disease propagated on social network? Who will infect whom?

The diffusion prediction on social network has been studied for decades. Generally, it can be divided into two categories: model-driven approaches and data-driven approaches. For model-driven approaches [3], such as Independent Cascade Model and Linear Threshold Model, these work design models to simulate cascading behavior , usually relying on intuition, without using historical diffusion records.

Researchers has exploited diffusion prediction problem by utilizing historical data; however, most of work [4]-[6] assumes that some of diffusion records of the predicted topics are known. In fact, such assumption is not always held in real world. For example, there was little information propagation about “iPhone” before it was first released in 2007. Besides, it will be more practical and valuable if the unseen topic diffusion can be predicted. For instance, if a company could estimate the propagation result of their new product on social media, they can plan market strategy more precisely like sending coupons to expected high-influence users; if a political party can predict the diffusion result for their half-baked policy, they can send activity invitations to bloggers who can help them to influence more audiences.

Recently, some work [2] starts to deal with the novel topic diffusion prediction problem; however, past work [2], [6] usually contain a strict constraint that there should

exist diffusion records in training data for nodes pairs to be predicted. The experiment scenario can be described as a diffusion revival problem: predicting whether past diffusion edges will reoccur in the future or not. But, the setting is unrealistic because of following reasons: To begin with, the data may be insufficient and incomplete such that some diffusion paths are missing in training data but they may emerge in testing data. Besides, each topic has its novelty, and may bring up different type of curious users to discuss, so the novel topic diffusion edges may not overlap old topics propagation records.

This work not only focus on novel topic diffusion prediction, but also, most different from past work [2], [6], introduce a new task called "diffusion prediction on obscure networks": predicting diffusion on not only existing edges but also edges without records in past.

The problem is challenging. First, for novel topics, there are not diffusion records for inference. Second, it [2] has shown that the useful features for novel topic diffusion prediction usually based on past diffusion records; however, the predicted node pairs may have no data about that. Moreover, some users may be clam in training data and their data is missing. Missing diffusion and user behavior data weakens the proposed features before.

In this paper, we propose a novel feature called Heterogeneous-Topic Estimating User Behavior (HTB) to address above challenges. To estimating the propagation tendency from one user to another, it combines topic model for novel topic lacking records problem, bases on user-level rather than pair-level for unseen edges, and utilizes heterogeneous node types data to estimate the tendency of calm users in novel topic.

The experiment results show that HTB can help proposed models in past improve in a more realistic scenario.

1.2 Contribution

First, we introduce a new task of information diffusion prediction on novel topic, where we release a constraint that there should be past diffusion records for predicted edges. We argue that the assumption is not realistic; instead, in real scenario, diffusion edges in novel topic may be not existed in past diffusion records.

Furthermore, to tackle the problem, new features are introduced. We proposed features User Behaviors and Hetero-Topic Estimating User Behavior based on topic similarity and heterogeneous node type information.

Besides, we design experiments to evaluate proposed models. The experiment shows that combining new and past features can improve performance in predicting novel topic information diffusion.

1.3 Thesis Organization

The thesis is organized as follows: In Chapter 2, relative work of information diffusion prediction is introduced and reviewed. To tackle the problem, new designed features are introduced in Chapter 3, and the experiment setting and result on Plurk data are discussed in Chapter 4. Finally, this work is concluded in Chapter 5, and we propose future work in Chapter 6.

Chapter 2

Related Work

In this chapter, relative work of information diffusion prediction is introduced. The problem has been studied for decades. Generally, it can be divided into two categories: Model-Driven approaches and Data-Driven approaches.

2.1 Model-Driven Approaches

For model-driven approaches, such models usually simulate the diffusion process based on intuition. Independent Cascade Model (IC) and Linear Threshold Model (LT) [3] are well known diffusion simulating models that have been studied for years. The core idea of IC model is that active nodes propagate information to inactive neighbor nodes with probability. If an inactive neighbor v is activated by an active node u, then v will try to active inactive neighbors like u. Similarly, LT model also assumes that the active nodes will influence their neighbors, but not in a probability way. In LT model, each node has it's own threshold to become active. If one node is activated, it will propagate a real-value score to neighbors. When an inactive node's received score exceeds the threshold, it will become active.

Besides, Heat Diffusion Model (HD) [7] is also a diffusion model, inspired by the physic phenomenon, assuming that information is propagated from the "hot people" to the "cold people", just like heat flowing from high temperature points to cold temperature points. Analog to physic formula, HD model describes the heat flow as follow:

$$\frac{f_i(t+\Delta t)-f_i(t)}{\Delta t}=\alpha \sum_{j:(v_j,v_i)\in E}(f_j(t)-f_i(t)),$$

$f_i(t)$ is the heat of node i at time t, and α is the thermal conductivity-the heat diffusion coefficient. Just like LT model, in HD model each node also has its own threshold. If one node receives heat exceeding its threshold, it will become active.

In sum, model-driven models focus mainly on simulating diffusion process rather than utilizing past record data. Model-driven models usually perform worse than data-driven models due to without learning process.

2.2 Data-Driven Approaches

On the other hand, some work has started to utilize data such as social network topology, text content, and user behaviors with machine learning model for information diffusion prediction. For instance, [4] and [5] combine social features and text content features to predict whether a message will be retweeted. Similarly, [6] proposed model to predict an URL whether will be diffused by a user; however, most data-driven work needs part of past records for topics and makes unseen topic diffusion cannot be predicted because the past records are unknown.

[8] proposed model can handle the novel topic evolution and discover following diffusion paths. The main difference between their work and this work is that their model mainly handles implicit diffusion paths whose data are usually unavailable. In contrast, our work focuses on explicit diffusion paths prediction.

[2] dealt with novel topic diffusion prediction by utilizing content and social features; but, their work contains a constraint that the proposed model only predicts the

existing edges that edges with diffusion records in past. In other words, their model cannot predict whether unseen edges will appear or not. This work release the constraint and find that proposed useful features in [2] can not cope with the unseen edges prediction problem totally, so we introduce new design features for enhancement in following chapter.

Chapter 3

Methodology

In this chapter, we formally define a few key concepts and a novel task of information diffusion, and then we introduce a machine learning approach to tackle this problem. Some features are proposed by [2]: Topic Information, User Information, User Topic Interaction, and Global Features, and we proposed two further features: User Behaviors and Heterogeneous-Topic Estimating User Behaviors.

3.1 Problem Formulation

3.1.1 Definition

3.1.1.1 Heterogeneous Social Network

A **heterogeneous social network** is a graph $G = (V, E, NT, nt)$ where V is a set of vertices, E is a set of edges among V, NT is a set of node types, and nt is a mapping from vertex to it's node type. In a **social network**, a vertex corresponds to a user and an edge relates to a connection between user i and j. The edges direction can be undirected or directed, and in our setting edges are directed. $NTy = \{Ty_1, Ty_2, Ty_3 ... Ty_N\}$ is the set of node types, $Ty_i = \{ty_{i,1}, ty_{i,2} ... ty_{i,k_i}\}$ is a set of type value for a specific node type i. nt maps each user u to a tuple of node type values. For instance, in Plurk data NT is {Gender, Location, Relationship, Default_Lang }, and a user node type tuple may be {"female", "Taiwan", "single", "zh_TW" }.

3.1.1.2 Topic Words & User Words

A **topic** means a keyword or a set of keywords discussed by users highly, like "iPhone5" or "earthquake", with related word distribution $P(w \mid t_i)_{w \in W}$, and **topic words** means a set of topics word distribution representing in a matrix form $TW = (P(w_j \mid t_i))_{i,j}$.

Similarly, **user words** means a set of users' word distributions represented in a matrix form $UW = (P(w_j \mid u_i))_{i,j}$.

3.1.1.3 User Behavior and Diffusion Records

User behavior means users perform specific actions with ***c*** times in a specific topic ***t*** on microblog. Especially we focus on two actions: the number of posting a message ***mc*** and the number of responding messages ***rc***, and User Behavior are described as $M = \{(v,t,mc) \mid v \in V, t \in T\}$ and $R = \{(v,t,rc) \mid v \in V, t \in T\}$.

Diffusion records are the set of diffusion recording how many times **c** one user ***u*** diffusing a specific topic **t** to another user, which is described as $D = \{(v,u,t,c) \mid v,u \in V; t \in T\}$.

3.1.2 Problem Definition

Give a heterogeneous social network G, user words UW, and a topic set T. T_N is the set of novel topics whose diffusion records and user behavior are unknown, and T_R is the set of the rest of topics, i.e. M_{T_R}, R_{T_R} and D_{T_R} are given. It is also assumed that the topic words TW is given no matter T_N or T_R. The goal is to predict whether a user u will diffuse a novel topic $t \in T_N$ information to a user v. Note that it is not assumed that

D_{T_R} should have the diffusion records of u and v; besides, M_{T_R} and R_{T_R} may not have data of u or v.

3.2 Intuition

The proposed models are based on two intuition: topic similarity and heterogeneous node type coherence; besides, we also study the dataset to support our intuition.

3.2.1 Topic Similarity and Diffusion

Intuitively, similar topics may attract similar group of people to discuss, and diffusion edges in past may revive in similar novel topics. For instance, people loving Apple's products, like "iphone4S" and "iphone5", may have interests in "iphone6."

3.2.2 Heterogeneous Node Type Coherence

Users with same type may have interest in similar topics. For example, male users may have more interest to engage in sport topics than in celebrity of celebrity gossip, but female users' preference may be contrary.

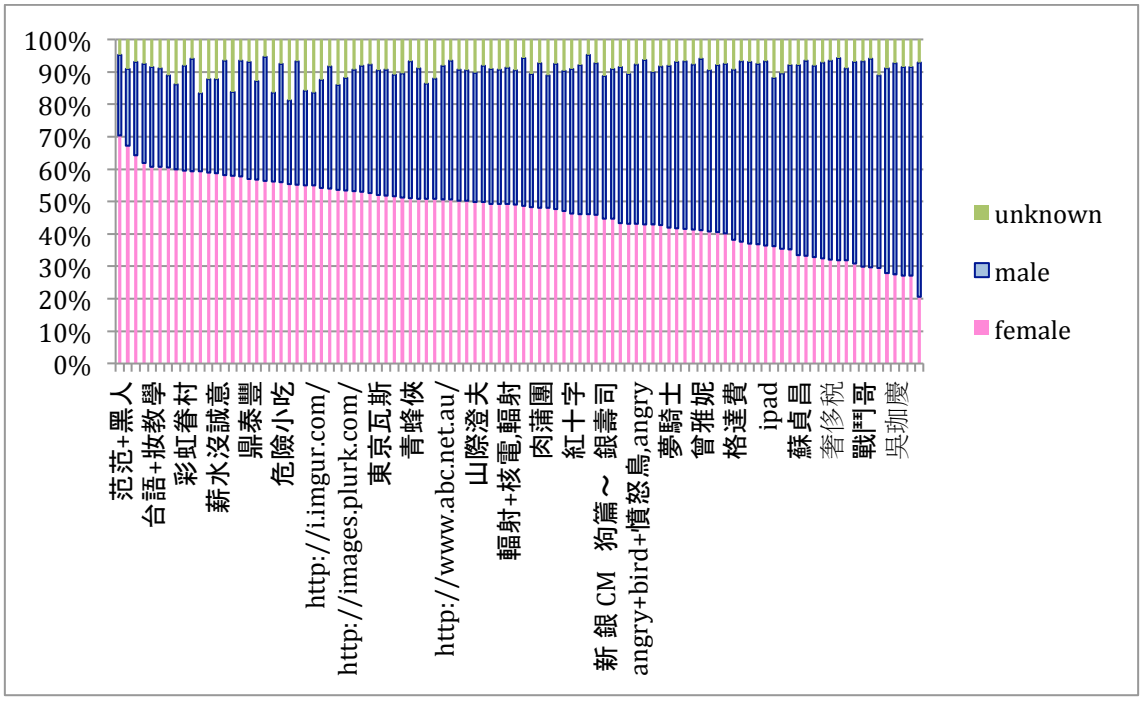

Figure 1 Gender composition of each topic

We calculate the statistic of gender composition in each topic shown above. It shows that gender composition is diverse from topic to topic. Some topics are attractive to female, like celebrity marriage “范范+黑人” or class for make up “台語+妝教學”, but other topics, such as politician “蘇貞昌” or athlete “吳珈慶”, are more charming to male.

The observation shows that some topics are more attractive to one node type than another one, and the result shows that the relationship between node types and topics may be insightful and useful for designing features.

3.3 Feature Generation

We first introduce features proposed by [2] for comparison: Topic Information, User Information, User-Topic Interaction and Global Information; and further, to deal

with diffusion records lacking problem, we propose User Behavior Information features for a more realistic scenario.

3.3.1 Topic Information

We extract hidden topic category information to model *topic signature*. In particular, we exploit the Latent Dirichlet Allocation (LDA) method [9],which is a widely used topic modeling technique, to decompose the topic-word matrix *TW* into hidden topic categories:

$$TW = TH * HW ,$$

where *TH* is a topic-hidden matrix, *HW* is hidden-word matrix, and *h* is the manually-chosen parameter to determine the size of hidden topic categories. *TH* indicates the distribution of each topic to hidden topic categories, and *HW* indicates the distribution of each lexical term to hidden topic categories. Note that *TW* and *TH* include both existing and novel topics. We utilize $TH_{t,*}$, the row vector of the topic-hidden matrix *TH* for a topic *t*, as a feature set. In brief, we apply LDA to extract the topic-hidden vector $TH_{t,*}$ to model *topic signature (TG*) for both existing and novel topics.

Topic information can be further exploited. To predict whether a novel topic will be propagated through a link, we can first enumerate the existing topics that have been propagated through this link. For each such topic, we can calculate its similarity with the new topic based on the hidden vectors generated above (e.g., using cosine similarity between feature vectors). Then, we sum up the similarity values as a new feature: *topic similarity (TS)*. For example, a link has previously propagated two topics for a total of three times {ACL, KDD, ACL}, and we would like to know whether a new topic, EMNLP, will propagate through this link. We can use the topic-hidden vector to

generate the similarity values between EMNLP and the other topics (e.g., {0.6, 0.4, 0.6}), and then sum them up (1.6) as the value of *TS*.

3.3.2 User Information

Similar to topic information, we extract latent personal information to model *user signature* (the users are anonymized already). We apply LDA on the user-word matrix *UW*:

$$UW = UM * MW ,$$

where *UM* is the user-hidden matrix, *MW* is the hidden-word matrix, and *m* is the manually-chosen size of hidden user categories. *UM* indicates the distribution of each user to the hidden user categories (e.g., age). We then use $UM_{u,*}$, the row vector of *UM* for the user *u,* as a feature set. In brief, we apply LDA to extract the user-hidden vector $UM_{u,*}$ for both source and destination nodes of a link to model *user signature (UG)*.

3.3.3 User-Topic Interaction

Modeling user-topic interaction turns out to be non-trivial. It is not useful to exploit latent semantic analysis directly on the user-topic matrix $UR = UQ * QR$, where *UR* represents *how many times each user is diffused for existing topic R* ($R \in T$), because *UR* does not contain information of novel topics, and neither do *UQ* and *QR*. Given no propagation record about novel topics, we propose a method that allows us to still extract implicit user-topic information. First, we extract from the matrix *TH* (described in Section 3.3.1) a subset *RH* that contains only information about existing topics. Next we apply left division to derive another user-hidden matrix *UH*:

$$UH = (RH \setminus UR^{T})^{T} = ((RH^{T} RH)^{-1} RH^{T} UR^{T})^{T}$$

Using left division, we generate the *UH* matrix using existing topic information. Finally, we exploit $UH_{u,*}$, the row vector of the user-hidden matrix *UH* for the user *u*, as a feature set.

Note that novel topics were included in the process of learning the hidden topic categories on *RH*; therefore the features learned here do implicitly utilize some latent information of novel topics, which is not the case for *UM*. Experiments confirm the superiority of our approach. Furthermore, our approach ensures that the hidden categories in topic-hidden and user-hidden matrices are identical. Intuitively, our method directly models the user's preference to topics' signature (e.g., how capable is this user to propagate topics in politics category?). In contrast, the *UM* mentioned in Section 3.3.2 represents the users' signature (e.g., aggressiveness) and has nothing to do with their opinions on a topic. In short, we obtain the user-hidden probability vector $UH_{u,*}$ as a feature set, which models *user preferences to latent categories (UPLC)*.

3.3.4 Global Information

Given a candidate link, we can extract global social features such as *in-degree (ID)* and *out- degree (OD)*. We tried other features such as PageRank values but found them not useful. Moreover, we extract the *number of distinct topics (NDT)* for a link as a feature. The intuition behind this is that the more distinct topics a user has diffused to another, the more likely the diffusion will happen for novel topics

3.3.5 User Behavior Information

To deal with unseen edges prediction, we introduce two new user-behavior features: user message behavior (UM) and user response behavior (UR), and then incorporate topic information and heterogeneous network information to refine User Behavior features called as Heterogeneous-Topic Estimating User Behavior (HTB).

User Behavior features describe one user's tendency to post a message or to response messages. For a user, user message behavior (UM) and user response behavior (UR) are calculated by aggregating the number of posting message and responding messages in old topics; however, User Behavior features have no variance in each novel topic, so we proposed Heterogeneous-Topic Estimating User Behavior (HTB) model to refine the feature.

Heterogeneous-Topic Estimating User Behavior (HTB) comprises two parts: using heterogeneous network information to enhance user behavior for miss data problem, and estimating user behavior in novel topic with content information.

First, we assume nodes in same heterogeneous node type group share similar user-behavior information, and the group's aggregated information can enhance individual feature if the group has consistent property. For a user u, the heterogeneous-enhanced user-behavior (HB) of u is described as follows:

$$HB(u,k,Ty) = (1-\alpha_{Ty(u),k}) * B_{u,k} + \alpha_{Ty(u),k} * Bmean_{Ty(u),k}$$

where Ty(u) is user's value of node type Ty, k is the topic that records in training data, $B_{u,k}$ is the user-behavior of u in topic k, $Bmean_{Ty(u),k}$ is average user-behavior in topic k of Ty(u) type nodes, and $\alpha_{Ty(u),k}$ is the confidence coefficient of node type value Ty(u)

information in topic k. The $\alpha_{Ty(u),k}$ is defined as follows: $\alpha_{Ty(u),k} = 1 - \frac{Bstd_{Ty(u),k}}{\max_{t \in Ty}(Bstd_{t,k})}$ where $Bstd_{Ty(u),k}$ is the user-behavior standard deviation of Ty(u) type nodes in topic k. The intuition is that if the distribution of user-behavior in node type value Ty(u) is more concentrated, the average user-behavior of Ty(u) is more believable.

Next, to estimate the user-behavior in novel topic k_{new}, it needs to connect the relationship between known topics and the novel topic by topic similarity information. The main idea is to trust more on similar known topics' records but less on irrelevant topics' records. For instance, to estimate one user's user-behavior in novel topic "iPhone 6", user records in "iPhone 5" and "iPhone 4" are more believable than in records in "Tsunami". To do so, first, obtain the topic-hidden TH distribution by Latent Dirchlet Allocation (LDA) [9], the state of the art method. Second, calculate similarity on TH between old topics and the novel topic, and we choose cosine distance as similarity score. Finally, to estimate the novel topic user-behavior, aggregate heterogeneous-enhanced user-behavior in old topic and weight by the topic similarity. The whole process can be summarized as following formula:

$$HTB(u, k_{new}, Ty) = \sum_{k} tsim(k_{new}, k) * HB(u, k, Ty)$$

where $tsim(k_{new},k)$ is the topic similarity between k_{new} and k in the topic-hidden distribution.

Chapter 4

Experiment

This chapter describes experiment details to evaluate the proposed model effectiveness, and compare it with baseline models.

4.1 Plurk Data

4.1.1 Data Preparation

First, 100 most popular topics (e.g., tsunami) are identified from Plurk microblog site between 01/2011 and 05/2011 and related users posts and response are crawled. Then 100 topics are manually separated into 7 groups based on semantic meaning: disaster, URL sharing, entertainment, domestic politics, daily life, global politics, and sports.

4.1.2 Diffusion Records & Underlying Social Network

The positive diffusion records are generated based on the post-response behavior. That is, if a person *x* posts a message containing one of the selected topic *t*, and later there is a person *y* responding to this message, we consider a diffusion of *t* has occurred from *x* to *y* (i.e., *(x, y, t)* is a positive instance).

The dataset contains a total of 699,985 positive instances out of 100 distinct topics; the largest and smallest topic contains 117,201 and 1,102 diffusions, respectively.

The underlying social network is created using the post-response behavior as well. We assume there is an acquaintance link between *x* and *y* if and only if *x* has responded to *y* (or vice versa) on at least one topic.

4.1.3 Heterogeneous Node Type

We choose Plurk public available user information as node type for each user: gender, location, relationship, and default language. Note that status are self-reported by users. Therefore, users can choose not to provide status, and a few users' information is missing because their accounts are expired. For the missing data, we add an "unknown" type for each node type. The following table summarizes four kinds of node types.

Node Type	Description	Size (Include Unknown)
Gender (G)	User's gender	3
Location (L)	User's current location	210
Relationship (R)	User's current relation with others status	11
Default Language (DL)	Default language the user using	39

Table 4-1 Heterogeneous Node Type

4.1.4 Corpus Processing

Furthermore, the sets of keywords for each topic are required to create the *TW* and *UW* matrices for latent topic analysis; we simply extract the content of posts and

responses for each topic to create both matrices. We set the hidden category number h = m = 7, which is equal to the number of topic groups.

We remove stop-words, use SCWS [10] for tokenization, and MALLET [11] and GibbsLDA++ [12] for LDA.

4.1.5 Cross Validation & Negative Sampling

We use topic-wise 4-fold cross validation to evaluate our method, because there are only 100 available topics. For each group, we select 3/4 of the topics as training and 1/4 as validation.

The positive instances (u,v) can be categorized into three types: "existing edges", "unseen edges", and "edges with calm users". Existing edges mean the edges with past diffusion records. If one user u and another user v did not diffuse information, but they both posted or responded others before, and then we call (u,v) as an unseen edge. If one of users or both users are calm in past, not posting or responding ever, we call (u,v) as an edge with calm users. The following figure shows three types of edges.

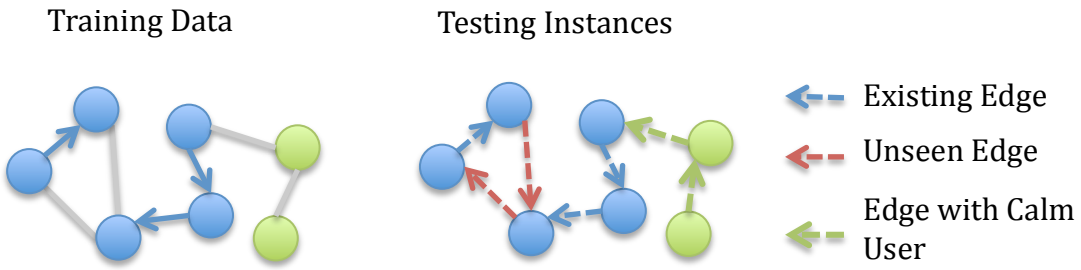

Figure 2 Edge Type

The same amount of negative instances for each topic (totally 699,985) is sampled for binary classification. For each type of edges, to prepare the similar status to

corresponding positive instance, we introduce three different negative sampling methods.

For existing edges, negative instances of a topic *t* are sampled randomly based from known topics' diffusion records. For unseen edges, we sample two not calm users who are not connected in network. And for edges with calm users, if both users are calm users, we sample two other calm users from network randomly as a negative instance; if one user is calm user but the other is not, we sample a calm user and a user with diffusion records as a negative instance.

4.2 Baseline

In the following experiments, we compare User Behavior and Heterogeneous-Topic Estimating User Behavior with the proposed models in [2]: Topic Information, User Information, Topic-User Interaction, Global Features, because those are data-driven learning methods in predicting explicit diffusion edges for novel topic.

4.3 Evaluation Metric

We choose under ROC curve (AUC) [13] as evaluation metric; Testing instances are ranked based on their likelihood of being positive, and compare it with the ground truth to compute AUC.

4.4 Experiment Result

4.4.1 Single Feature Comparison

Type	Features	AUC
Proposed Features	TG (Topic Signature)	50.00%
	DF (Existing Diffusion)	55.32%
	NDT (Number of Distinct Topics)	55.97%
	TS (Topic Similarity)	57.54%
	UG (User Signature)	57.94%
	UPLC (User Preference to Latent Categories)	59.13%
	OD (Out-Degree)	64.44%
	ID (In-Degree)	**67.00%**
User Behavior	UM (User Message Behavior)	67.17%
	UR (User Response Behavior)	68.31%
Heterogeneous-Topic Estimating User Behavior (HTB)	HTB_R_M	69.17%
	HTB_G_M	69.53%
	HTB_L_M	69.54%
	HTB_DL_M	69.54%
	HTB_R_R	69.74%
	HTB_L_R	70.36%
	HTB_DL_R	70.40%
	HTB_G_R	**70.63%**

Table 4-2 Single Feature Comparison

HTB is described as follows: for HTB_X_Y, X is the heterogeneous node type, and Y is the user behavior type. For example, HTB_G_R uses gender as node type and responding action as user behavior type.

To begin with, we compare different single features performance shown above. Both user behavior and heterogeneous-topic estimating user behavior features outperform than proposed features from [2]. The best single feature is HTB_G_R, heterogeneous-topic estimating user behavior with gender as node type and responding

action as user behavior. Comparing with best single feature (ID) in baseline, HTB_G_R improve 3.63%, from 67.00% to 70.63%.

It [2] was showed that features exploiting users pair diffusion records, such as Existing Diffusion (DF), Topic Similarity (TS), and Number of Distinct Topics (NDT), have better performance than the others; however, those features do not performs well in our experiment scenario. The reason is that users pair diffusion records may be missing in real scenario, which causes those useful features based on diffusion records performance decreasing.

4.4.2 Feature Combination Comparison

Type	Features Combination	AUC
Proposed Features Single	ID	67.00%
Proposed Features	TH+TS+ID+OD	69.43%
Proposed Features+ User Behavior+ HTB	TS+HTB_L_M+HTB_G_R+ HTB_L_R+HTB_R_R+HTB_DL_R	72.88%

Table 4-3 Combination Result

Secondly, we compare features combination performance shown above. Containing proposed features, User Behavior, and HTB, the best combination results in 72.88%, which outperforms than proposed single feature with 5.88% and proposed features combination with 3.43%. It shows that HTB and User Behavior features can enhance proposed features by [2] performance.

Chapter 5

Conclusion

Predicting information diffusion in novel topic is challenging: First, there are not diffusion records of novel topics for inference. Second, [2] has shown that the useful features for novel topic diffusion prediction usually based on past diffusion records; but, in real world, predicted node pairs may have no data about that. Furthermore, some users may be calm in observable data, and their data are limited so their behaviors become hard to predict.

We propose User Behavior and Heterogeneous-Topic Estimating User Behavior (HTB) to address above challenges. At begin, we assume that there is a correlation between topic similarity and diffusion repetition, and we observe that different heterogeneous node types have different tendency in each topic. Based on above intuition, HTB combines topic model for novel topic lacking records problem, and utilizes heterogeneous node types information to estimate the tendency of calm users in novel topic. We shift the experiment into a more realistic scenario by releasing constraint, and experiment results show that User Behavior and HTB can help proposed models by [2] to improve 3.4% in the new scenario.

Chapter 6

Future Work

6.1 Time-Sensitive Model

To deal with novel topic lacking diffusion records problem, we focus on utilizing topic information to connect between known topics information and novel topics, and the experiment setting is based on topic-wise; however, it could be more complete if time information is considered. How to utilizing time information to predict novel topic diffusion path can be a next challenging problem.

Chapter 7

Reference

[1] "Twitter Blog: Twitter turns six," *blog.twitter.com*. [Online]. Available: http://blog.twitter.com/2012/03/twitter-turns-six.html. [Accessed: 30-Nov-2012].

[2] T.-T. Kuo, S.-C. Hung, W.-S. Lin, N. Peng, S.-D. Lin, and W.-F. Lin, "Exploiting latent information to predict diffusions of novel topics on social networks," presented at the ACL '12: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, 2012, vol. 2.

[3] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, 2003.

[4] S. Petrovic, M. Osborne, and V. Lavrenko, "RT to Win ! Predicting Message Propagation in Twitter," *ICWSM*, vol. 13, no. 513435, pp. 586–589, 2011.

[5] J. Zhu, F. Xiong, D. Piao, Y. Liu, and Y. Zhang, "Statistically Modeling the Effectiveness of Disaster Information in Social Media," presented at the 2011 IEEE Global Humanitarian Technology Conference (GHTC), 2011, pp. 431–436.

[6] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, "Outtweeting the twitterers-predicting information cascades in microblogs," presented at the Proceedings of the 3rd Workshop on Online Social Networks (WOSN 2010), 2010.

[7] H. Ma, H. Yang, M. R. Lyu, and I. King, "Mining social networks using heat diffusion processes for marketing candidates selection," presented at the CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, 2008.

[8] C. X. Lin, Q. Mei, Y. Jiang, J. Han, and S. Qi, "Inferring the Diffusion and Evolution of Topics in Social Communities," *mind*, 2011.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[10] Hightman, "SCWS - 簡易中文分詞系統 - hightman.cn," *hightman.cn*. [Online]. Available: http://www.hightman.cn/index.php?scws. [Accessed: 30-Nov-2012].

[11] McCallum and A. Kachites, "MALLET: A Machine Learning for Language Toolkit.," *mallet.cs.umass.edu*, 30-Nov-2002. [Online]. Available: http://mallet.cs.umass.edu. [Accessed: 30-Nov-2012].

[12] X.-H. Phan and C.-T. Nguyen, "GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA)," *gibbslda.sourceforge.net*, 30-Nov-2007. [Online]. Available: http://gibbslda.sourceforge.net. [Accessed: 30-Nov-2012].

[13] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," presented at the Proceedings of the 23rd international conference …, 2006.