Exploiting Latent Information to Predict Diffusions of Novel Topics on Social Networks

Tsung-Ting Kuo[1]*, San-Chuan Hung[1], Wei-Shih Lin[1], Nanyun Peng[1], Shou-De Lin[1], Wei-Fen Lin[2]

[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taiwan

[2]MobiApps Corporation, Taiwan

*d97944007@csie.ntu.edu.tw

Abstract

This paper brings a marriage of two seemly unrelated topics, natural language processing (NLP) and social network analysis (SNA). We propose a new task in SNA which is to predict the diffusion of a new topic, and design a learning-based framework to solve this problem. We exploit the latent semantic information among users, topics, and social connections as features for prediction. Our framework is evaluated on real data collected from public domain. The experiments show 16% AUC improvement over baseline methods. The source code and dataset are available at http://www.csie.ntu.edu.tw/~d97944007/diffusion/

1 Background

The diffusion of information on social networks has been studied for decades. Generally, the proposed strategies can be categorized into two categories, model-driven and data-driven. The model-driven strategies, such as independent cascade model (Kempe et al., 2003), rely on certain manually crafted, usually intuitive, models to fit the diffusion data without using diffusion history. The data-driven strategies usually utilize learning-based approaches to predict the future propagation given historical records of prediction (Fei et al., 2011; Galuba et al., 2010; Petrovic et al., 2011). Data-driven strategies usually perform better than model-driven approaches because the past diffusion behavior is used during learning (Galuba et al., 2010).

Recently, researchers started to exploit content information in data-driven diffusion models (Fei et al., 2011; Petrovic et al., 2011; Zhu et al., 2011). However, most of the data-driven approaches assume that in order to train a model and predict the future diffusion of a topic, it is required to obtain historical records about how this topic has propagated in a social network (Petrovic et al., 2011; Zhu et al., 2011). We argue that such assumption does not always hold in the real-world scenario, and being able to forecast the propagation of novel or unseen topics is more valuable in practice. For example, a company would like to know which users are more likely to be the source of 'viva voce' of a *newly* released product for advertising purpose. A political party might want to estimate the potential degree of responses of a half-baked policy before deciding to bring it up to public. To achieve such goal, it is required to predict the future propagation behavior of a topic even *before* any actual diffusion happens on this topic (i.e., no historical propagation data of this topic are available). Lin et al. also propose an idea aiming at predicting the inference of implicit diffusions for novel topics (Lin et al., 2011). The main difference between their work and ours is that they focus on implicit diffusions, whose data are usually not available. Consequently, they need to rely on a model-driven approach instead of a data-driven approach. On the other hand, our work focuses on the prediction of explicit diffusion behaviors. Despite the fact that no diffusion data of novel topics is available, we can still design a data-driven approach taking advantage of some explicit diffusion data of known topics. Our experiments show that being able to utilize such information is critical for diffusion prediction.

2 The Novel-Topic Diffusion Model

We start by assuming an existing social network $G = (V, E)$, where V is the set of nodes (or user) v, and E is the set of link e. The set of topics is

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 344–348,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

denoted as T. Among them, some are considered as novel topics (denoted as N), while the rest (R) are used as the training records. We are also given a set of diffusion records $D = \{d \mid d = (src, dest, t)\}$, where *src* is the source node (or diffusion source), *dest* is the destination node, and t is the topic of the diffusion that belongs to R but not N. We assume that diffusions cannot occur between nodes without direct social connection; any diffusion pair implies the existence of a link $e = (src, dest) \in E$. Finally, we assume there are sets of keywords or tags that relevant to *each* topic (including existing and novel topics). Note that the set of keywords for novel topics should be seen in that of existing topics. From these sets of keywords, we construct a topic-word matrix $TW = (P(word_j \mid topic_i))_{i,j}$ of which the elements stand for the conditional probabilities that a word appears in the text of a certain topic. Similarly, we also construct a user-word matrix $UW = (P(word_j \mid user_i))_{i,j}$ from these sets of keywords. Given the above information, the goal is to predict whether a given link is active (i.e., belongs to a diffusion link) for topics in N.

2.1 The Framework

The main challenge of this problem lays in that the past diffusion behaviors of new topics are missing. To address this challenge, we propose a supervised diffusion discovery framework that exploits the latent semantic information among users, topics, and their explicit / implicit interactions. Intuitively, four kinds of information are useful for prediction:

- *Topic information*: Intuitively, knowing the signatures of a topic (e.g., is it about politics?) is critical to the success of the prediction.
- *User information*: The information of a user such as the personality (e.g., whether this user is aggressive or passive) is generally useful.
- *User-topic interaction*: Understanding the users' preference on certain topics can improve the quality of prediction.
- *Global information*: We include some global features (e.g., topology info) of social network.

Below we will describe how these four kinds of information can be modeled in our framework.

2.2 Topic Information

We extract hidden topic category information to model *topic signature*. In particular, we exploit the Latent Dirichlet Allocation (LDA) method (Blei et al., 2003), 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.

2.3 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)*.

2.4 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 2.2) 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 2.3 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).*

2.5 Global Features

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.

2.6 Complexity Analysis

The complexity to produce each feature is as below:

(1) *Topic information*: $O(I * |T| * h * B_t)$ for LDA using Gibbs sampling, where I is # of the iterations in sampling, $|T|$ is # of topics, and B_t is the average # of tokens in a topic.

(2) *User information*: $O(I * |V| * m * B_u)$, where $|V|$ is # of users, and B_u is the average # of tokens for a user.

(3) *User-topic interaction*: the time complexity is $O(h^3 + h^2 * |T| + h * |T| * |V|)$.

(4) *Global features*: $O(|D|)$, where $|D|$ is # of diffusions.

3 Experiments

For evaluation, we try to use the diffusion records of old topics to predict whether a diffusion link exists between two nodes given a new topic.

3.1 Dataset and Evaluation Metric

We first identify 100 most popular topic (e.g., earthquake) from the Plurk micro-blog site between 01/2011 and 05/2011. Plurk is a popular micro-blog service in Asia with more than 5 million users (Kuo et al., 2011). We manually separate the 100 topics into 7 groups. 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 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). Our dataset contains a total of 1,642,894 positive instances out of 100 distinct topics; the largest and smallest topic contains 303,424 and 2,166 diffusions, respectively. Also, the same amount of negative instances for each topic (totally 1,642,894) is sampled for binary classification (similar to the setup in KDD Cup 2011 Track 2). The negative links of a topic t are sampled randomly based on the absence of responses for that given topic.

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. Eventually we generated a social network of 163,034 nodes and 382,878 links. 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 use area under ROC curve (AUC) to evaluate our proposed framework (Davis and Goadrich, 2006); we rank the testing instances based on their likelihood of being positive, and compare it with the ground truth to compute AUC.

3.2 Implementation and Baseline

After trying many classifiers and obtaining similar results for all of them, we report only results from LIBLINEAR with c=0.0001 (Fan et al., 2008) due to space limitation. We remove stop-words, use SCWS (Hightman, 2012) for tokenization, and MALLET (McCallum, 2002) and GibbsLDA++ (Phan and Nguyen, 2007) for LDA.

There are three baseline models we compare the result with. First, we simply use the total number of existing diffusions among all topics between two nodes as the single feature for prediction. Second, we exploit the independent cascading model (Kempe et al., 2003), and utilize the normalized total number of diffusions as the propagation probability of each link. Third, we try the heat diffusion model (Ma et al., 2008), set initial heat proportional to out-degree, and tune the diffusion time parameter until the best results are obtained. Note that we did not compare with any data-driven approaches, as we have not identified one that can predict diffusion of novel topics.

3.3 Results

The result of each model is shown in Table 1. All except two features outperform the baseline. The best single feature is *TS*. Note that *UPLC* performs better than *UG*, which verifies our hypothesis that maintaining the same hidden features across different LDA models is better. We further conduct experiments to evaluate different combinations of features (Table 2), and found that the best one (*TS + ID + NDT*) results in about 16% improvement over the baseline, and outperforms the combination of all features. As stated in (Witten et al., 2011), adding useless features may cause the performance of classifiers to deteriorate. Intuitively, *TS* captures both latent topic and historical diffusion information, while *ID* and *NDT* provide complementary social characteristics of users.

Method	Feature	AUC
Baseline	Existing Diffusion	58.25%
	Independent Cascade	51.53%
	Heat Diffusion	56.08%
Learning	**T**opic Sig**n**ature (*TG*)	50.80%
	Topic **S**imilarity (*TS*)	**69.93%**
	User Sig**n**ature (*UG*)	56.59%
	User **P**references to **L**atent **C**ategories (*UPLC*)	61.33%
	In-**d**egree (*ID*)	65.55%
	Out-**d**egree (*OD*)	59.73%
	Number of **D**istinct **T**opics (*NDT*)	55.42%

Table 1: Single-feature results.

Method	Feature	AUC
Baseline	Existing Diffusion	58.25%
Learning	ALL	65.06%
	TS + UPLC + ID + NDT	67.67%
	TS + UPLC + ID	64.80%
	TS + UPLC + NDT	66.01%
	TS + ID + NDT	**73.95%**
	UPLC + ID + NDT	67.24%

Table 2: Feature combination results.

4 Conclusions

The main contributions of this paper are as below:

1. We propose a novel task of predicting the diffusion of unseen topics, which has wide applications in real-world.
2. Compared to the traditional model-driven or content-independent data-driven works on diffusion analysis, our solution demonstrates how one can bring together ideas from two different but promising areas, NLP and SNA, to solve a challenging problem.
3. Promising experiment result (74% in AUC) not only demonstrates the usefulness of the proposed models, but also indicates that predicting diffusion of unseen topics without historical diffusion data is feasible.

Acknowledgments

This work was also supported by National Science Council, National Taiwan University and Intel Corporation under Grants NSC 100-2911-I-002-001, and 101R7501.

References

David M. Blei, Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res., 3.993-1022.

Jesse Davis & Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang & Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. J. Mach. Learn. Res., 9.1871-74.

Hongliang Fei, Ruoyi Jiang, Yuhao Yang, Bo Luo & Jun Huan. 2011. Content based social behavior prediction: a multi-task learning approach. Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, Scotland, UK.

Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic & Wolfgang Kellerer. 2010. Outtweeting the twitterers - predicting information cascades in microblogs. Proceedings of the 3rd conference on Online social networks, Boston, MA.

Hightman. 2012. Simple Chinese Words Segmentation (SCWS).

David Kempe, Jon Kleinberg & Eva Tardos. 2003. Maximizing the spread of influence through a social network. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C.

Tsung-Ting Kuo, San-Chuan Hung, Wei-Shih Lin, Shou-De Lin, Ting-Chun Peng & Chia-Chun Shih. 2011. Assessing the Quality of Diffusion Models Using Real-World Social Network Data. Conference on Technologies and Applications of Artificial Intelligence, 2011.

C.X. Lin, Q.Z. Mei, Y.L. Jiang, J.W. Han & S.X. Qi. 2011. Inferring the Diffusion and Evolution of Topics in Social Communities. Proceedings of the IEEE International Conference on Data Mining, 2011.

Hao Ma, Haixuan Yang, Michael R. Lyu & Irwin King. 2008. Mining social networks using heat diffusion processes for marketing candidates selection. Proceeding of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

Sasa Petrovic, Miles Osborne & Victor Lavrenko. 2011. RT to Win! Predicting Message Propagation in Twitter. International AAAI Conference on Weblogs and Social Media, 2011.

Xuan-Hieu Phan & Cam-Tu Nguyen. 2007. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA).

Ian H. Witten, Eibe Frank & Mark A. Hall. 2011. Data Mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann Publishers Inc.

Jiang Zhu, Fei Xiong, Dongzhen Piao, Yun Liu & Ying Zhang. 2011. Statistically Modeling the Effectiveness of Disaster Information in Social Media. Proceedings of the 2011 IEEE Global Humanitarian Technology Conference.