Retweet Behavior Prediction Using Hierarchical Dirichlet Process

Qi Zhang, Yeyun Gong, Ya Guo, Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, P.R.China
{qz, 12110240006, 13210240002, xjhuang}@fudan.edu.cn

Abstract

The task of predicting retweet behavior is an important and essential step for various social network applications, such as business intelligence, popular event prediction, and so on. Due to the increasing requirements, in recent years, the task has attracted extensive attentions. In this work, we propose a novel method using non-parametric statistical models to combine structural, textual, and temporal information together to predict retweet behavior. To evaluate the proposed method, we collect a large number of microblogs and their corresponding social networks from a real microblog service. Experimental results on the constructed dataset demonstrate that the proposed method can achieve better performance than state-of-the-art methods. The relative improvement of the the proposed over the method using only textual information is more than 38.5% in terms of F1-Score.

INTRODUCTION

Social media are rapidly growing, with thousands of millions of users participating in them every day. According to the social media report done by Nielsen[1], U.S. audiences spent more than 121.1B minutes in social media in a month. Microblog services usually provide a function, retweet, for users to re-post someone else's microblogs. Through this feature, users can quickly share the microblogs that they view as valuable and important with all of their followers. This causes content to spread from one community to another. Moreover, because microblog services also provide applications for users to share and consume content on mobile devices, information can rapidly spread much faster than any other infectious method. Retweet is usually considered to be the main essence of the viral aspect of content spreading in social media (Rodrigues et al. 2011). Hence, the task of predicting retweet behaviors has been used in various social network applications such as business intelligence (Castellanos et al. 2011; Hoffman and Fodor 2010), microblog retrieval (Chang and Kim 2012), popular event prediction (Hong, Dan, and Davison 2011; Gupta et al. 2012), and so on.

Due to the increasing requirements of predicting retweet behaviors in recent years, the task of how to model retweet behaviors has received considerable attention. Various methods have been proposed from different perspectives, including social influence (Liu et al. 2010; Zhang et al. 2013), textual features (Naveed et al. 2011), social features (Zaman et al. 2010; Petrovic, Osborne, and Lavrenko 2011; Luo et al. 2013), history information (Feng and Wang 2013), visual features (Can, Oktay, and Manmatha 2013), and combinations of different features (Suh et al. 2010). There are also a variety of works studying different models to achieve the task, including Matchbox (Stern, Herbrich, and Graepel 2009; Zaman et al. 2010), multiple additive regression-trees (MART) (Wu et al. 2008), maximum entropy classifier (Artzi, Pantel, and Gamon 2012), autoregressive-moving-average (ARMA) (Luo, Wang, and Wu 2012), factor graph model (Yang et al. 2010), conditional random fields (Peng et al. 2011) et al.

Through analyzing retweet microblogs, we find that even though several microblogs have the same content posted or retweeted by several followers of a user, the user may only retweet one of them. Who post the microblog and other factors may also impact a user's decision. In this paper, we argue that whether a user will retweet a microblog or not is based on the following three main factors: 1) social information (e.g. who post or re-post the microblog); 2) content information (e.g. the topics of the microblog); 3) user properties (e.g. the interests of the user). However, because most existing methods only consider either structural or textual information, it is difficult to effectively process this type of issue.

In this paper, we propose a novel model based on non-parametric statistical methods to incorporate textual, structural, and temporal information to model retweet behavior. For analyzing the retweet phenomenon and evaluating the proposed method, we collected a large dataset from a real online microblog service. We crawled not only the microblog posts itself but also the social networks of the corresponding users. From the constructed dataset, we observe that the retweet behavior is influenced by not only the content itself, but also structural and temporal information. Based on these observations, we adapted the hierarchical Dirichlet process (HDP) (Teh et al. 2006) to define a novel nonparametric method to model retweet be-

[1] http://www.nielsen.com/us/en/reports/2012/state-of-the-media-the-social-media-report-2012.html

haviors. The retweet prediction algorithm uses the estimated model and several simple steps to achieve the task. The experimental results on the constructed dataset demonstrated that the proposed method could achieve significantly better performance than state-of-the-art methods. Moreover, in a comparison of the performances of the methods without incorporating all these factors, we also observed that combining all these factors could significantly improve the effectiveness of prediction. The main contributions of this work can be summarized as follows:

- We propose a novel nonparametric Bayesian model adapted from the hierarchical Dirichlet process to combine textual, structural, and temporal information for predicting retweet behaviors.
- We construct a large collection of microblogs from a real microblogging service. It contains both microblog content and the social network information of related users.
- Experimental results demonstrate that the proposed method can achieve better performance than state-of-the-art methods.

The Proposed Method

The Problem Statement

Given a microblog m and a user u, the retweet prediction task is to decide whether the user u will retweet the microblog or not. In Twitter-like services, a user u_A can follow other users. If u_A follows u_B, the activities (e.g. tweet, retweet) of u_B are visible to u_A. User u_A is called the *follower* of u_B and u_B is called the *followee* of u_A. Hence, to model the retweet prediction behavior of a user u, we can restore the browsing history of user u through collecting microblogs posted by the followees of u. Let D_u to represent all the microblogs which are posted by the followees of the user u. The dth microblog in D_u consists of a sequence of words $w_d = \{w_{dn}\}_{n=1}^{N_d}$, where N_d is the number of words in the dth microblog, and w_{dn} is one of the word belonging the vocabulary W. a_d denotes the author of the dth microblog. In this work, we only consider the microblogs posted or retweeted by the followees of a user.

Data Analyses

To analyze the retweet behavior and construct evaluation data set, we crawled a large number of microblogs based on the properties of microblog service. According to the design of Twitter-like microblog services (e.g. Twitter, Sina Weibo, et al.), only the microblogs posted by the followees of the user will be shown up for him to read. Hence, for evaluation and analysis, we collected the data set from Sina Weibo[2] with the following ways. Firstly, we randomly select 200 users as the *central users* we studied in this work. To restore what they have seen from the microblog services, we firstly collected the 2-ego network for all the central users based on their follower-followee relationships. Through this step, we constructed a social network which contains 2.07 million users and 299.6 million following relationships among them. Since existing methods have demonstrated that users' profiles are useful for this task and some of works are based on the features extracted from it, we also crawled the following profiles of users: name, gender, residence, birthday, verification status, #followers, #followees, #microblogs, et al. Finally, we crawled the latest 2,000 microblogs (including tweets and retweets) posted by the followees of all central users. Through this step, about 84.8 million microblogs were collected.

Table 1: Statistics of the data set

# Users	2,073,121
# Following-Relationships	299,602,693
# Microblogs	84,768,859
# Retweets w/o reply	34,024,561
# Retweets with reply	16,974,326

Table 1 shows the statistics of the collected data set. We split the retweets into two categories based on whether they contain reply or not. From the table, we can observe that about 60.2% microblogs are retweets. The percentages of retweets in the data set are much higher comparing to the statistics reported in (Boyd, Golder, and Lotan 2010). Among all the retweets, about 33.3% of them contain not only the original microblog but also replies. In (Yu, Asur, and Huberman 2012), Yu et al. reported the similar statistical result as us. We think that these statistics reflect the practices of social medias in different cultures.

For analyzing the retweet behaviors, we firstly restored the *browsing history* for all 200 central users based on the microblogs posted by their followees. The 1-ego networks of these users can be directly extracted from the crawled social networks. A user's browsing history contains the microblogs which were posted by the followees of the user and were sorted by their posting times from near to far. Through the recovered browsing histories of users and the microblogs posted by themselves we can get the following observations:

- The 1-ego network of all 200 central users contain 82,311 nodes in total. It means that the total number of followees of them are 82,311 users. From analyzing the microblogs reposted by the central users, we can only find 52,177 users. It only covers 63.3% of all the followees. Further more, among these users, less than 17.8% users have more than one microblogs reposted by the central users. The frequencies of them almost follow the power law. Moreover, for different central users, the percentages of users involved in the retweet actions vary greatly from less than 5% to more than 90%. From these statistics, we can observe that the authors of microblogs would impact the retweet behavior.
- Among all the microblogs reposted by the central users, 42.1% of them occur more than one times in their browsing history. It means that more than one followees posted or reposted the same microblog. However, users may not retweet the first one he saw from their browsing

[2] It is one of the most popular website providing microblogging service in China. http://www.weibo.com

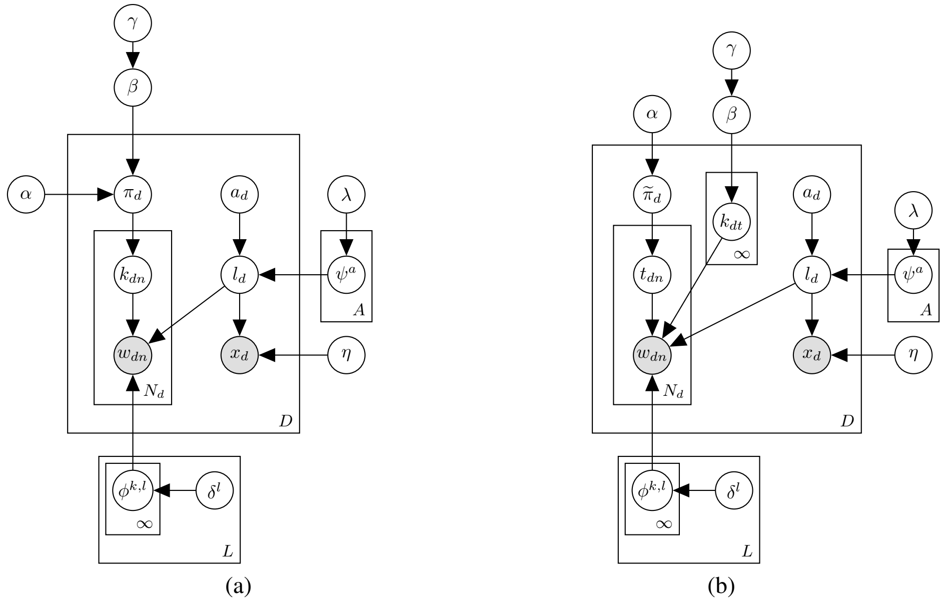

Figure 1: The graphical representation of the proposed models. (a) The original representation. (b) Chinese restaurant franchise construction.

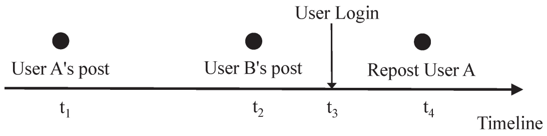

Figure 2: An example of retweet behavior.

history. Figure 2 illustrates this kind of issue. User A and B posted/reposted the same microblog at time t_1 and t_2 respectively. Since the user login at the time t_3, the microblog posted by user B should be shown to the user before user A's. However, the user reposted the microblog posted by user A. According to the statistics, about 37.4% retweet behaviors belong to this type. It can also demonstrate in a manner that retweet behaviors are impacted by not only the content but also who posted it.

- To understand the topics user concentrated on vary over time, we aggregate the microblog posts published in a month together as a document. Then we use HDP to estimate topics of them. From the results, we find that topics user concentrated on are time dependent. For different users, the number of estimated topics varies greatly from less than ten to more than fifty.

The Generation Process

Through descriptions given by the previous sections, we can observe that the retweet behaviors may be impacted by textual, structural and temporal information. Inspired by these observations, in this work, we assume that whether a user retweets a microblog is influenced by:

1. Who posted the microblog.
2. How many followees of the user have posted or retweeted the microblog.
3. The content of the microblog.
4. The time-dependent interests of the user.

Based on these observations, we propose to extend HDP to model the author, structure, and content information. The temporal information is incorporated into the predication stage.

Author Influence. According to the statistics described in the previous section, we can see that users may only retweet microblogs posted by several specific followees of them. Therefore, for user u, we assume that the probability of retweeting microblogs posted by p_{fe_i} for each followee fe_i subject to binomial distribution with Beta prior.

Structure Influence. From data analyses, we also know that the retweet behavior is influenced by their neighbours. Some users prefer to retweet the microblogs which have been retweeted by many their followees, while someone won't like to retweet this kind microblogs. Therefore, we assume that each user corresponds to a distribution. We firstly normalize the retweet times of the microblogs to range from 0 to 1, we use x_d to represent the normalized result. Then, each retweet label is associated with a continuous distribution over x_d. All the results, in this paper, we employ the Beta distribution which can behave versatile shapes.

Content Influence. The content influence is modeled by hidden topics. Since the actual number of topics is an unknown priori, we adopt HDP-based topic model to achieve the task. We use t to represent the cluster contained in the microblog, and k represents the topic for each cluster. We can use a Gibbs sampler to get latent variables, and then the generate probabilities of the microblog d can be calculated for each label respectively as follows:

$$p_c(w_d|\mathbf{t},\mathbf{k},\mathbf{l}) = \prod_{n=1}^{N_d} f^{k_{dt_{dn}},l_d}(w_{dn}), \quad (1)$$

where w_d are the words in microblog d, N_d is the number of words in microblog d, w_{dn} is the nth word in microblog d, $k_{dt_{dn}}$ is the topic for cluster t_{dn} in microblog d, l_d is the retweeted label for microblog d, $f^{k_{dt_{dn}},l_d}(w_{dn})$ is the likelihood of generating w_{dn} for topic $k_{dt_{dn}}$ and retweet label l_d.

Incorporate author, structure, and content information into hierarchical Dirichlet process, we propose the model ASC-HDP. A represents author, S represents structure, and C represents content. Let A_u to represent the number of followees of user u, and D_a to represent the number of microblogs of user a. $m_{.k}$ represents the number of clusters previously assigned with topic k. n_{dt} is the number of words of cluster t in microblog d. We use $p(w_{dn} = w|t_{dn}, l_d, k_{dt_{dn}})$ to represent the generative probability of word w_{dn} given cluster t_{dn}, retweeted label l_d and the topic $k_{dt_{dn}}$ for the cluster t_{dn}. We can get the generation process of our model as follows:

For each followee of the user u. $a = 1, 2, ..., A_u$
  1. Draw $\psi_a \sim Beta(\lambda)$
  2. For each microblog of followee a. $d = 1, 2, ..., D_a$.
    a. Draw a retweet label $l_d \sim Binomial(\psi^a)$
    b. Draw a normalized retweet times $x_d \sim Beta(\eta_{l_d})$
    c. For each word $n = 1, ..., N_d$
      i. Choose an exist cluster $t_{dn} = t \propto n_{dt}$
      ii. Choose a new cluster $t_{dn} = t^{new} \propto \alpha$
      iii. If choose an existing cluster then
        Draw a word $w_{dn} = w$ according to probability $p(w_{dn} = w|t_{dn}, l_d, k_{dt_{dn}})$
      iv. If choose a new cluster then
        i) Choose an existing topic $k_{dt_{dn}^{new}} = k \propto m_{.k}$
        ii)Choose a new topic $k_{dt_{dn}^{new}} = k^{new} \propto \gamma$
        iii) Draw a word $w_{dn} = w$ according to probability $p(w_{dn} = w|t_{dn}, l_d, k_{dt_{dn}^{new}})$

Fig. 1 shows the graphical representation of the generation process.

Learning and inference

To learn the parameters of ASC-HDP, we use Gibbs sampler to sample cluster assignments $\mathbf{t}$ and topic assignments $\mathbf{k}$ for the training data. For the test data, we also need to sample the retweet label $\mathbf{l}$. Adding a superscript $\neg i$ to a variable, indicate the same quantity it is added to without the contribution of object i. For example, $n_{dt}^{\neg dn}$ represents the number of words of cluster t in microblog d without the contribution of word w_{dn}.

Sampling t. To resample cluster t_{dn} for the nth word in microblog d, we make use of exchangeability and imagine word n being the last word in microblog d. We obtain the conditional posterior for t_{dn} by combining the conditional prior distribution for t_{dn} with the likelihood of generating w_{dn}. For the previously used cluster t, the conditional probability of $t_{dn} = t$ is as follows:

$$p(t_{dn} = t|\mathbf{t}_{\neg dn}, \mathbf{k}, \mathbf{l}) \propto \frac{n_{dt}^{\neg dn}}{n_{..}^{\neg dn} + \alpha} f_{\neg w_{dn}}^{k_{dt},l}(w_{dn}), \quad (2)$$

For a new cluster t^{new}, the conditional probability of $t_{dn} = t^{new}$ is as follows:

$$\begin{aligned} &p(t_{dn} = t^{new}|\mathbf{t}_{\neg dn}, \mathbf{k}, \mathbf{l}) \\ &\propto \frac{\alpha}{n_{..}^{\neg dn} + \alpha} p(w_{dn}|\mathbf{t}_{\neg dn}, t_{dn} = t^{new}, \mathbf{k}, \mathbf{l}), \end{aligned} \quad (3)$$

where $\frac{\alpha}{n_{..}^{\neg dn}+\alpha}$ is the prior probability that t_{dn} takes at a new cluster t^{new}. $p(w_{dn}|\mathbf{t}_{\neg dn}, t_{dn} = t^{new}, \mathbf{k}, \mathbf{l})$ is the likelihood due to w_{dn} for $t_{dn} = t^{new}$, which can be calculated by integrating out the possible values of $k_{dt^{new}}$ as follows:

$$\begin{aligned} &p(w_{dn}|\mathbf{t}_{\neg dn}, t_{dn} = t^{new}, \mathbf{k}, \mathbf{l}) \\ &= \sum_{k=1}^{K} \frac{m_{.k}}{m_{..} + \gamma} f_{\neg w_{dn}}^{k,l}(w_{dn}) + \frac{\gamma}{m_{..} + \gamma} f_{\neg w_{dn}}^{k^{new},l}(w_{dn}), \end{aligned} \quad (4)$$

where $m_{..}$ is the total number of clusters assigned to any topic. $\frac{\gamma}{m_{..}+\gamma}$ is the prior probability that assigned to a new topic k^{new}. $f_{\neg w_{dn}}^{k^{new},l}(w_{dn})$ is the prior density of w_{dn}.

If $t_{dn} = t^{new}$, the probability of exist topic k for this table is calculated by the following equation:

$$p(k_{dt^{new}} = k|\mathbf{t}, \mathbf{k}^{\neg dt}, l_d) \propto \frac{m_{.k}}{m_{..} + \gamma} f_{\neg w_{dn}}^{k,l_d}(w_{dn}), \quad (5)$$

and the probability of new topic k^{new} can be calculated by the following equation:

$$p(k_{dt^{new}} = k^{new}|\mathbf{t}, \mathbf{k}^{\neg dt}, l_d) \propto \frac{\gamma}{m_{..} + \gamma} f_{\neg w_{dn}}^{k^{new},l_d}(w_{dn}), \quad (6)$$

Sampling k. Since changing k_{dt} actually changes the component membership of all data items in the cluster t, so that the conditional probability for exist topic k is

$$p(k_{dt} = k|\mathbf{t}, \mathbf{k}^{\neg dt}, l) \propto \frac{m_{.k}^{\neg dt}}{m_{..} + \gamma} f_{\neg \mathbf{w}_{dt}}^{k,l_d}(\mathbf{w}_{dt}) \quad (7)$$

and for a new topic k^{new} is

$$p(k_{dt} = k^{new}|\mathbf{t}, \mathbf{k}^{\neg dt}, l) \propto \frac{\gamma}{m_{..} + \gamma} f_{\neg \mathbf{w}_{dt}}^{k^{new},l_d}(\mathbf{w}_{dt}), \quad (8)$$

where $\mathbf{w}_{dt}$ is the words of cluster t in microblog d.

Sampling l. The component membership of all the data items in microblog d will change when l_d changes, so

$$\begin{aligned} &p(l_d = l|\mathbf{t}, \mathbf{k}, w_d, \mathbf{l}_{\neg d}, x_d) \propto \frac{N_a^l + \lambda_1}{N_a + \lambda_1 + \lambda_2} \\ &\cdot \frac{(1 - x_d)^{\eta_{l1}-1} {x_d}^{\eta_{l2}-1}}{B(\eta_{l1}, \eta_{l2})} \cdot \prod_{n=1}^{N_d} f(w_{dn}|\phi^{k_{dt_{dn}},l}), \end{aligned} \quad (9)$$

where N_a is the number of microblogs of user a saw by user u. N_a^l is the times of user u retweet from user a.

For simplicity, we update η after each sample by the method of moments, detailed as follows:

$$
\begin{aligned}
\eta_{l1} &= \overline{x}_d^l\left(\frac{\overline{x}_d^l(1-\overline{x}_d^l)}{\delta_l^2}-1\right)\\
\eta_{l2} &= (1-\overline{x}_d^l)\left(\frac{\overline{x}_d^l(1-\overline{x}_d^l)}{\delta_l^2}-1\right),
\end{aligned}
\tag{10}
$$

where $\overline{x}_d^l$ is the sample mean and δ_l^2 is the biased sampled variance of the retweet times belonging to retweet label l. Sparsity is a serious problem for parameter estimation of η, so when the user have retweeted less than twenty times we set $\eta_{l1} = \eta_{l2} = 0.5$ to avoid the error caused by estimation error.

Retweet Predication

Given the unlabeled microblogs a user saw, we firstly do sampling. After the hidden variables of words become stable, we can estimate the topic distribution of the cluster in the microblog d through Eq.(7) and Eq.(8). Besides that, as described in the previous sections, the topics user interested in may vary over time. Hence, we propose to increase the weights for the hot topics in recent month by $p(k|c_d)$, which is the probability of topic k in the current month. If the topic k is not appear in the current month, $p(k|c_d)$ will be 0. We propose to use the following equation:

$$
p^*(k_{dt}|\mathbf{t},\mathbf{k}_{\neg dt},l) = p(k_{dt}|\mathbf{t},\mathbf{k}_{\neg dt},l)(p(k_{dt}|c_d)+1). \tag{11}
$$

Based on these steps, we can calculate the the probability of retweeting as follows:

$$
\begin{aligned}
&p(l_d = l|\mathbf{t},\mathbf{k}_{\neg dt}, w_d, \mathbf{l}_{\neg d}, x_d) \\
&\propto \frac{N_a^l+\lambda_1}{N_a+\lambda_1+\lambda_2}\,\frac{(1-x_d)^{\eta_{l1}-1}{x_d}^{\eta_{l2}-1}}{B(\eta_{l1},\eta_{l2})} \\
&\sum_{k=1}^{K} p^*(k_{dt_{dn}}|\mathbf{t},\mathbf{k}_{\neg dt},l)\prod_{n=1}^{N_d} f(w_{dn}|\phi^{k_{dt_{dn}},l})p(w_{dn}|w_d),
\end{aligned}
\tag{12}
$$

where $p(w_{dn}|w_d)$ is the weight of the word in the microblog d, which can be estimated by the TF·IDF score of word w_{dn}.

Experiments

Experiment Configurations

For each user, we randomly selected about 70% of all microblogs in their browsing history as training data and 10% as development data. The other 20% are used as the test data. For evaluation metrics, we use precision (P), recall (R), and F1-score (F_1) to evaluate the performance. Precision is calculated based on the percentage of "retweets truly identified" among "retweets labeled by system". Recall is calculated based on the "retweets truly identified" among "golden standard retweets". F1-score is the harmonic mean of precision and recall. We ran our model with 500 iterations of Gibbs sampling. In the HDP-based model, we used $\gamma \sim Gamma(1,1)$ and $\alpha \sim Gamma(1,1)$ as prior distributions for the concentration parameters. The base measure H for both retweet labels used is a symmetric Dirichlet distribution with parameters of 0.5. In the LDA-based model, we use $\alpha = 50.0/K$ and $\delta = 0.1$ for both retweet labels, after trying a few different numbers of topics, we empirically set the number of topics to 20. In both of the two model we set parameter $\lambda_1 = \lambda_2 = 0.1$.

Table 2: The performances of different methods in the test dataset.

Method	Precision	Recall	F1-Score
Naive Bayes	0.362	0.617	0.456
SVM^{Rank}	0.485	0.450	0.467
LRC-BQ	0.547	0.638	0.589
C-LDA	0.474	0.507	0.490
AC-LDA	0.752	0.545	0.632
SC-LDA	0.446	0.650	0.529
ASC-LDA w/o t	0.680	0.624	0.650
ASC-LDA	0.694	0.636	0.664
C-HDP	0.514	0.541	0.527
AC-HDP	**0.809**	0.533	0.643
SC-HDP	0.456	**0.727**	0.561
ASC-HDP w/o t	0.686	0.695	0.691
ASC-HDP	0.746	0.715	**0.730**

For comparison with the proposed model, we also evaluated the following methods on the constructed dataset:

- **Naive Bayes**: The retweet predication task can be formalized as a binary classification task, where each microblog is assigned either positive or negative label to represent whether it will be retweeted or not. Hence, we applied Naive Bayes to model the posterior probability of labels given a microblog.
- **SVM^{Rank}**: We implemented the method proposed in (Luo et al. 2013), where microblog content, retweet history, followers status, followers active time and followers interests are incorporated under the learning-to-rank framework, to achieve the task.
- **LRC-BQ**: It combines the influence locality function and additional features (including personal attributes, instantaneity topic propensity, et al.) (Zhang et al. 2013). Based on their description, we also collected the properties of users needed and implemented this method.
- **C-HDP**: It represents the model only taking the content information into consideration.
- **SC-HDP**: It omits the influence of the author of microblog. Given a microblog, we assume that whether a user u retweet a microblog subject to binomial distribution, which is not relate to its author.
- **AC-HDP**: It denotes the model which does not consider the influence of actions of social network of this microblog.
- **ASC-HDP w/o t**: In this model, we don't use Eq.(11) in the predication stage of the model ASC-HDP.
- **ASC-LDA**: The proposed methods can also be adopted based on Latent Dirichlet Allocation (LDA) model, where each document is viewed as a mixture of topics. We regard

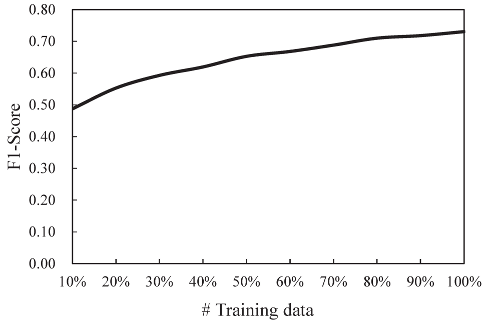

Figure 3: The influence of the number of training data for ASC-HDP.

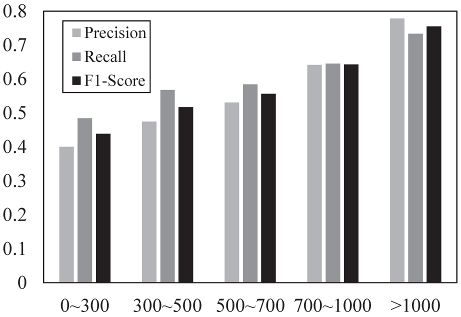

Figure 4: The influence of the number of microblogs posted by users.

microblogs as general documents and learn an LDA-based model from microblogs.

Experimental Results

Table 2 shows the comparisons of the proposed method ASC-HDP with the state-of-the-art discriminative and generative methods on the constructed evaluation dataset. From the results, we can observe that the proposed method is significantly better than the other methods. Through comparing the results between ASC-HDP and C-HDP, AC-HDP, SC-HDP, and ASC-HDP w/o t, we can see the ASC-HDP achieves the significantly better F1-score than other the methods. The results demonstrate that all the factors which we proposed in the previous section can impact the performance of the prediction. From the results of AC-HDP and SC-HDP, we can observe that the author information and neighbourhood information can only benefit the precision and recall aspects respectively. Comparing F1-score of C-HDP with ASC-HDP's, we can see the temporal, structural and author information together bring about 38.5% relative improvement. For LDA based models, we can also observe that the author and social information can also benefit a lot. The performances of LRC-BQ in the constructed dataset is lower the performances reported in (Zhang et al. 2013). This is due to the reason that we do not limit the popularity of the retweets. The topics and content are much more diverse. The results also show that discriminative methods achieve worse results than generative methods. We think that the noisy and varieties content is one of the main reasons of the low performances.

Figure 3 shows the inference of the number of training data. From the results, we can observe that the proposed method ASC-HDP can achieve better performance than other methods with only 30% training data. The results also suggests that the performances of ASC-HDP can further increase with more training data. We have the intuition that if user have posted a large number microblogs, we can accurately estimate their behaviors. To investigate this intuition, we split the users into five groups based on the number of microblogs posted by them. Figure 4 shows the results. We can see that the more microblogs posted by users, the more accuracy we can get.

From the description of the LDA based model, we can know that there are several hyperparameters in ASC-LDA. Among them the number of topics is one the most important factors. Table 3 shows the influence of the number of topics. From the table, we can observe that the proposed model obtains the best performance when the number of topics is set to 20. While, when the number of topics is set to 10 and 30, the performances can compare with it. From analyzing the estimated topics of HDP, we find that the number of topics for most of the users are also in this range. However, the number of topics of different users vary greatly. Hence, HDP-based methods are more suitable for this task. The experimental results also demonstrate it.

Table 3: The inferences of different number of topics for ASC-LDA.

# Topics	Precision	Recall	F1-Score
10	0.681	0.625	0.651
20	**0.694**	0.**636**	**0.664**
30	0.692	0.629	0.659
50	0.659	0.608	0.633
70	0.645	0.593	0.618

Conclusions

In this paper, we proposed a novel method, which is based on hierarchical Dirichlet process, to combine structural, textual, and temporal information to predict retweet behavior. For evaluating the proposed method, we collected a large number of microblogs and their corresponding social networks from a microblog service. Experimental results demonstrate that the proposed method can achieve better performance than state-of-the-art methods. The relative improvement of the the proposed ASC-HDP over C-HDP is more than 38.5% in terms of F1-Score.

Acknowledgement

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 61473092 and 61472088), Shanghai Leading Academic Discipline Project (B114).

References

Artzi, Y.; Pantel, P.; and Gamon, M. 2012. Predicting responses to microblog posts. In *Proceedings of NAACL-HLT '12*.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of HICSS '10*.

Can, E. F.; Oktay, H.; and Manmatha, R. 2013. Predicting retweet count using visual cues. In *Proceedings of CIKM '13*.

Castellanos, M.; Dayal, U.; Hsu, M.; Ghosh, R.; Dekhil, M.; Lu, Y.; Zhang, L.; and Schreiman, M. 2011. Lci: a social channel analysis platform for live customer intelligence. In *Proceedings of SIGMOD '11*.

Chang, J., and Kim, H.-j. 2012. Twitter search methods using retweet information. In *Proceedings of BUSTECH '12*, 67–71.

Feng, W., and Wang, J. 2013. Retweet or not?: personalized tweet re-ranking. In *Proceedings of WSDM*.

Gupta, M.; Gao, J.; Zhai, C.; and Han, J. 2012. Predicting future popularity trend of events in microblogging platforms. *Proceedings of ASIS&T*.

Hoffman, D. L., and Fodor, M. 2010. Can you measure the roi of your social media marketing. *MIT Sloan Management Review* 52(1):41–49.

Hong, L.; Dan, O.; and Davison, B. D. 2011. Predicting popular messages in twitter. In *Proceedings of WWW '11*.

Liu, L.; Tang, J.; Han, J.; Jiang, M.; and Yang, S. 2010. Mining topic-level influence in heterogeneous networks. In *Proceedings of CIKM '10*.

Luo, Z.; Osborne, M.; Tang, J.; and Wang, T. 2013. Who will retweet me?: Finding retweeters in twitter. In *Proceedings of SIGIR '13*.

Luo, Z.; Wang, Y.; and Wu, X. 2012. Predicting retweeting behavior based on autoregressive moving average model. In *Web Information Systems Engineering-WISE 2012*. 777–782.

Naveed, N.; Gottron, T.; Kunegis, J.; and Alhadi, A. C. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*, 8. ACM.

Peng, H.-K.; Zhu, J.; Piao, D.; Yan, R.; and Zhang, Y. 2011. Retweet modeling using conditional random fields. In *Proceedings of ICDMW '11*.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2011. Rt to win! predicting message propagation in twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Rodrigues, T.; Benevenuto, F.; Cha, M.; Gummadi, K.; and Almeida, V. 2011. On word-of-mouth based discovery of the web. In *Proceedings of SIGCOMM '11*.

Stern, D. H.; Herbrich, R.; and Graepel, T. 2009. Matchbox: large scale online bayesian recommendations. In *Proceedings of WWW '09*.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of SocialCom '10*.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association* 101(476).

Wu, Q.; Burges, C. J.; Svore, K. M.; and Gao, J. 2008. Ranking, boosting, and model adaptation. *Tecnical Report, MSR-TR-2008-109*.

Yang, Z.; Guo, J.; Cai, K.; Tang, J.; Li, J.; Zhang, L.; and Su, Z. 2010. Understanding retweeting behaviors in social networks. In *Proceedings of CIKM '10*.

Yu, L. L.; Asur, S.; and Huberman, B. A. 2012. Artificial inflation: The real story of trends and trend-setters in sina weibo. In *Processings of SocialCom-PASSAT '12*.

Zaman, T. R.; Herbrich, R.; Van Gael, J.; and Stern, D. 2010. Predicting information spreading in twitter. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*.

Zhang, J.; Liu, B.; Tang, J.; Chen, T.; and Li, J. 2013. Social influence locality for modeling retweeting behaviors. In *Proceedings of IJCAI'13*.