An Influence Field Perspective on Predicting User's Retweeting Behavior

Yi Shen[1,2], Jianjun Yu[2], Kejun Dong[2], Juan Zhao[2], and Kai Nan[2](✉)

[1] University of Chinese Academy of Sciences, Beijing, China
shenyi@cnic.ac.cn
[2] Computer Network Information Center,
Chinese Academy of Sciences, Beijing, China
{yujj,kevin,zhaojuan,nankai}@cnic.ac.cn

Abstract. User's retweeting behavior, which is the key mechanism for information diffusion in the micro-blogging systems, has been widely employed as an important profile for personalized recommendation and many other tasks. Retweeting prediction is of great significance. In this paper, we believe that user's retweeting behavior is synthetically caused by the influence from other users and the post. By analogy with the concept of electric field in physics, we propose a new conception named "influence field" which is able to incorporate different types of potential influence. Based on this conception, we provide a novel approach to predict user's retweeting behavior. The experimental results demonstrate the effectiveness of our approach.

Keywords: Retweet · Field theory · Influence · Electric field strength

1 Introduction

Due to its dual role of social network and news media [1], the micro-blogging systems have become an important platform for people to acquire information. In a typical micro-blogging system (e.g., Twitter), users are allowed to publish short messages with a limitation of 140 characters (i.e., posts or tweets). Unlike other social-networking services, micro-blogging introduces a new relationship named "follow", which enables users to access a filtered timeline of posts from anyone else they care about without any permission. The retweet mechanism, which provides a convenient function for users to share information with their followers, has become a hot spot in the field of social network analysis.

There are many previous studies focus on how to employ diverse features to address the retweeting prediction problem. Intuitively, people tend to retweet the posts with the content they are interested in. Therefore, the content-based

Y. Shen—This research was supported by Special Items of Information, Chinese Academy of Sciences under Grant XXH12503; and by Around Five Top Priorities of "One-Three-Five" Strategic Planning, CNIC under Grant CNIC_PY-14XX; and by NSFC Grant No.61202408.

J. Li and Y. Sun (Eds.): WAIM 2015, LNCS 9098, pp. 3–16, 2015.
DOI: 10.1007/978-3-319-21042-1_1

features are widely used in the existing works [2–4]. Actually, the reality is much more complicated because besides users' intrinsic interests, users' behaviors on social networks may be also caused by the influence of other users [5,6]. For example, a user retweets a post about "World Cup 2014", the reason behind this retweeting might be: (1) he is a soccer fan and he is interested in "World Cup 2014" (2) he is attracted by a post about "World Cup 2014" because it has been retweeted by his intimate friends. As a result, social-based features are also important for addressing the retweeting prediction task [7,8]. However, the social-based features applied in the previous studies are mainly author-oriented, in other words, they mainly focus on the relationship between the current user and the author of the post, while the influence from other influential users (especially the close friends of the current user) is neglected.

In this paper, we interpret retweeting behavior as an outcome of the influence. In fact, everyone has a certain amount of potential influence [9], and the influence of user A to user B is negatively correlated with the "distance" between them. For instance, the influence of A to one of his closest friends is likely to be greater than the influence to a long-forgotten acquaintance. Inspired by this, we make an analogy between the users in the micro-blogging environment and the electric charges in an electric field and assume that every user has an "influence field" around himself, then we employ the field theory in physics to interpret users' retweeting behavior.

The major contributions of our paper are as follows:

(1) We propose the conception of "influence field" and borrow the field theory in traditional physics to model the influence between the users in the micro-blogging.

(2) By defining different types of influence, we apply the "influence field" to address the retweeting prediction task. Our "influence field" model not only considers the influence of all the potential influential-users, but also takes account of the inherent relationship among different features. We evaluate our approach on a large dataset from Sina Weibo, which is the most popular micro-blogging in China. The experimental results demonstrate that our "influence field" model is indeed effective.

(3) Although our work has been done in the context of Sina Weibo, we expect that the "influence field" conception would hold for many other social-network platforms(e.g., Twitter and Facebook).

2 Related Work

A better understanding of the user retweeting behavior is of great significance. Since retweeting plays a crucial role in the information propagation in the micro-blogging systems, researchers have completed a lot of interesting work through analyzing users' retweeting behavior, such as rumor identification [10] and break news detection [11]. Furthermore, retweet is also employed as an important profile to build the user interest model in many personalized applications such as recommender system [12].

A bulk of studies try to understand why people retweet. For example, Boyd et al. [13] pointed out that there are diverse motivations for users to retweet, such as for spreading information to other users, saving the valuable posts for future personal access, commenting on someone's post in the form of incidental retweeting, etc. Macskassy and Michelson [14] studied a set of retweets from Twitter, they claimed that the content based propagation models could explain the majority of retweeting behaviors they saw in their data.

Plenty of previous works focus on extracting various kinds of features to predict the users' retweeting behavior as well as analyze the effects caused by different features. Some of these works [15–17] predicted retweeting from a global perspective which aimed to forecast whether a post will be popular in the future. There are also some studies which conducted the prediction to the individual level and manage to answer the question whether a post will be retweeted by a specific user. For example, Luo et al. [3] proposed a learning to rank based framework to discover the users who are most likely to retweet a specific post. Peng et al [2] applied conditional random fields(CRF) to model and predict the users' retweet patterns. The authors of [18] address this problem by means of executing constrained optimization on a factor graph model. There are also many works [4,7,8] in which the authors considered the individual level prediction as a classification task and then built effective classifiers to address this problem.

The difference between our work and most of the previous studies is mainly reflected in two aspects. One is that we emphasize the impact of all the influential users rather than only the author of the post, the other is that we integrate diverse features as well as their correlation to model user's retweeting behavior.

3 Methodology

In this section, we first describe the conception of "influence field" in detail, and then present the calculation of the elements in the "influence field" model. Finally, we introduce the classifier based on "influence field" to address the retweet prediction task.

3.1 The Conception of Influence Field

The intuitive idea behind our approach is that a user's behavior is usually influenced by others. Take retweeting as an example, we suppose that user A has a certain probability $P_{u_A}^{po}$ to retweet a post po about "data mining". If A notices that another user B has retweeted po, then $P_{u_A}^{po}$ may increase due to the influence of B. If B is a famous expert on "data mining", $P_{u_A}^{po}$ will be even larger. Moreover, as people are easy to be affected by their close friends in many cases [19], for this reason, if B happens to be a good friend of A, $P_{u_A}^{po}$ may be significantly increased. To sum up, user's (e.g., B) impact on another user (e.g., A) is positively correlated to B's influence and negatively correlated to the "distance" between them. Thus, this effect can be written as follows:

$$E = K\frac{I_u}{R^{\rho}} \tag{1}$$

Where E is used to measure the impact caused by u. K is a constant. I_u is the global influence of the influential user. R represents the distance between the two users. ρ is a coefficient to tune the effect of the distance. It is worth noting that this formula is very similar to the definition of "electric field strength" in physics. As depicted in Figure 1(a), an electric charge with power Q will generate an electric field around itself, and the field strength at a point r far away is calculated as: $E' = k\frac{Q}{r^2}$. Inspired by the electric field theory, we assume that everyone on micro-blogging has his/her own "influence field", which makes himself/herself as the center.

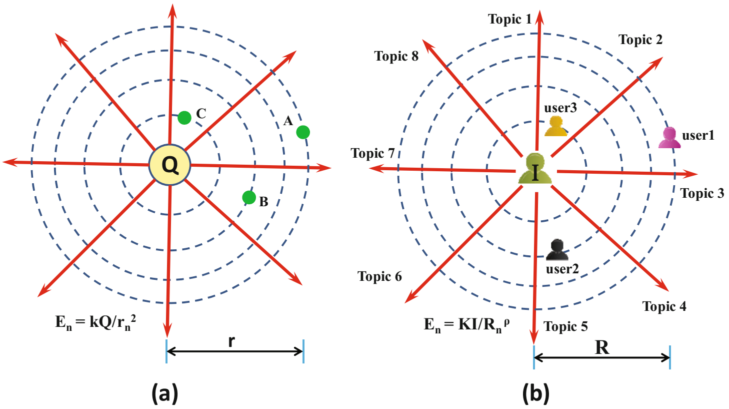


Fig. 1. Examples of the electric field (a) and the influence field (b)

As portrayed in Figure 1(b), the user at the center will affect the behaviors of user1-user3 through the influence field of himself and the strength will attenuate as the distance increases. The field strength of the "influence field" can be defined according to Equation 1.

The electric field strength in physics is a vector with its direction. In order to be consistent with this concept, we define "topic" as the direction of "influence field". This design is to match the fact that each user may has uneven influence on different topics. For example, David Beckham is a famous soccer player, and his influence on the topic "sports" is far greater than on "cloud computing".

We try to employ "influence field" discussed above to interpret and predict users' retweeting behavior in the micro-blogging systems. We assume that user u will be affected by a force (f_u) when he/she is reading post po, and u has a threshold θ_u^{po} for retweeting it. Besides the influential users, the influence from po should also be considered. Finally we utilize an inequality to predict whether u will retweet po: $f_u \geq \theta_u^{po}$

Based on Equation 1, we define f_u as follows:

$$f_u = \sum_{u' \in U} \frac{K I_{u'} I_{po}}{R(u', u)^\rho} \tag{2}$$

Where U is the collection of the users who are able to influence u, I_{po} is the influence of po ,which is used to measure the importance of po. Generally speaking, the posts with high influence are more likely to be retweeted. We can see that Equation 2 shares the similar structure with the famous Coulomb's law($F = kQq/r^2$) in physics.

3.2 The Calculation of the Elements

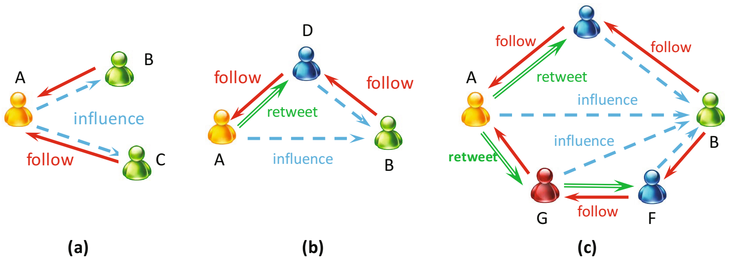
In this subsection, we mainly introduce how to calculate the elements of the "influence field". First, we will talk about how to identify those users (U) who will influence the current user u when he/she is reading a post. Next, we will elaborate the calculation of I_u, $R(u', u)$, I_{po} and θ_u^{po} respectively.

Identify the Influential Users (U). As depicted in Figure 2, three types of users as following will influence the current user u when po is exposed to him/her.

The author of po. Consider the simplest case shown in Figure 2(a), both user B and user C follow user A. Once A has published a post po, B and C will be influenced by user A to some extent since po will appear in their separate timelines.

The followees of u . There are also many other cases that u does not follow the author of po directly. As portrayed in Figure 2(b), user A is the author of po, B only follows user D and receives po through the "retweeting-path" from A to D. In this case, we consider that both A and D will influence B because B will perceive their retweeting behaviors. In Figure 2(c), two followees of B have retweeted po and we believe that both of them(i.e., E and F) will influence B.

The mentioned users in po . Besides the author and the followees, there is another type of users on the "retweeting-path". Take user G in Figure 2(c) as an instance, although he is neither the author nor the followee of B, he is also able to influence B because his nickname will be mentioned in po with a prefix of "@" symbol.



(a)
(b)
(c)

Fig. 2. Different types of influential users

To sum up, the influential users U includes all those users appear on the "retweeting-paths" from the author of po to u. The duplicate users are removed.

Calculate the Influence of User(I_u). Since we have defined the directions of "influence field" as the topics, it is necessary for us to find some way to measure user's influence on each topic. TwitterRank [20] provided an approach based on the PageRank algorithm and the topic model which is able to calculate the influence on a single topic for each user on Twitter. It measures the influence taking both the link structure and the topical similarity between users into account. Since TwitterRank has proved its effectiveness on a large dataset of a typical micro-blogging platform (i.e., Twitter), we decide to adopt it to calculate the users' topic-level influence in this paper.

The core idea of TwitterRank is to make an analogy between the influence of a user and the authority of a web page, then extend the PageRank algorithm with topical information to calculate the user's topic-specific influence.In detail, the TwitterRank algorithm consists of the following steps:

1. First of all, The Latent Dirichlet Allocation (LDA) [21] model is applied to distill the topics that users are interested in based on the posts they have published.
2. Secondly, a directed graph of users is formed according to the following relationships among them.
3. Each topic corresponds to a probability transition matrix P_t, and the specific transition probability of the random walk from u_i to his followee u_j is defined as:

$$P_t(i,j) = \frac{|\tau_j|}{\sum_{a:a\in i's followee} |\tau_a|} * sim_t(i,j) \tag{3}$$

Where $|\tau_j|$ denotes the number of posts published by u_j and the denominator part represents the total number of posts published by all u_i's followees. $sim_t(i,j)$ is the similarity of u_i and u_j in topic t, which can be calculated as:

$$sim_t(i,j) = 1 - |P^t_{u_i} - P^t_{u_j}| \tag{4}$$

Here, $P^t_{u_i}$ and $P^t_{u_j}$ are both topic distributions generated by LDA.

4. Finally, based on the transition probability matrix, the topic-specific influence of the users, which is denoted as $\overrightarrow{I_t}$, can be iteratively calculated by:

$$\overrightarrow{I_t} = \gamma P_t \times \overrightarrow{I_t} + (1-\gamma)E_t \tag{5}$$

E_t is the teleportation vector which is introduced to tackle the case in which the users may follow one another in a loop. The authors normalized the t-th column of the user-topic matrix obtained by LDA so as to represent E_t. γ is a parameter between 0 and 1 to tune the probability of teleportation.

As presented in [20], the user's global influence can be measured as an aggregation of his/her influence in different topics, which can be calculated as:

$$I_u = \sum_t r_t \cdot I_u^t \tag{6}$$

I_u^t is the corresponding element of the user u in vector $\overrightarrow{I_t}$. r_t is the weight assigned to topic t, which is the probability of user u is interested in t, i.e., P_u^t generated by LDA.

Measure the Distance between Users($\boldsymbol{R(u_1, u_2)}$). The distance between two users is largely determined by their relationship. In the previous research, the relationship is measured by "tie strength" [22]. The relationships between users can be divided into two categories: the weak ties and the strong ties [23–25]. The weak ties include our loose acquaintance, new friends and our 2-degree friends (i.e., friends of our friends). Strong ties refer those people we are most concerned, such as our family and our trusted friends. In fact, most of our communication on social networks is with our strong ties [26]. Not surprisingly, people are disproportionately influenced by the strong ties, and the strongest influence is between mutual best friends [27]. As a result, we believe that the users' retweeting behavior is also influenced by the strong ties. Intuitively, the more frequent the interaction between two users, the stronger their tie strength. In addition, a large number of common friends may also mean strong ties [28]. In micro-blogging, we define friends as the users who have followed with each other.

$$R(u_1, u_2) = \frac{\lambda r(u_1, u_2)}{lg[(N_0 + 1)\sqrt{(N_1 + 1)(N_2 + 1)}]} \tag{7}$$

Finally, we model the distance between two users according to Equation 7, where λ is a constant coefficient, $r(u_1, u_2)$ is the "router distance" between u_1 and u_2. For example, if A follows B, then $r(A, B) = 1$, and more, if B follows C and A doesn't follow C, then $r(A, C) = 2$, etc. N_0 is the number of the common friends of u_1 and u_2. N_1 is the frequency of interaction from u_1 to u_2 and N_2 denotes the frequency of interaction from u_2 to u_1 . The interaction here refers the behaviors include retweeting, mention and comment on the micro-blogging system. The closer N_1 and N_2, the stronger the tie strength, then the nearer the distance. The square root and the logarithmic function are used to smooth the final result.

Calculate the Influence of Post ($\boldsymbol{I_{po}}$). Intuitively, a post with rich information usually has a strong influence. We model the influence of a post also at the topic granularity. We employ "topic entropy" to measure the amount of information contained in *po* on each topic and the higher the value of "topic entropy", the stronger the influence. We consider each post as a document, based on the "bag of words" assumption, LDA is applied to represent each post as a probability distribution over a certain number of topics, while each topic is expressed as a probability distribution over the words. Finally, the topic-level influence of post *po* on topic *t* is calculated as the following equation:

$$H_{po}^{t} = -\sum_{i=1}^{K} P(w_i|t)\log_2 P(w_i|t) \tag{8}$$

Where w_i denotes the words in *po*. K is the total number of words in *po*.

For each retweet, we calculate the intervals between the retweeting timestamp and the generation timestamp of corresponding *po* in our dataset. According to our statistics, most retweeting behaviors happened during a short period after the original posts have been published. About 50% of the intervals are less than 1 hour and over 90% of them are less than 1 day(1440 minutes), which means the probability of user to retweet *po* will gradually diminish over time. For this reason, we apply an exponential time factor to discount the influence of those old posts as Equation 9, where Δt_0 is the interval between the published timestamp of *po* and the current timestamp, μ is a decay coefficient.

$$I_{po} = \sum_{t\in T} I_{po}^{t} = \sum_{t\in T} H_{po}^{t} \times e^{-\mu\Delta t_0} \tag{9}$$

The Threshold for User to Retweet a Post ($\boldsymbol{\theta_u^{po}}$). Different users have their respective "accepted thresholds" even for the same post, additionally, single user has different "accepted thresholds" for different posts [19]. Therefore, the threshold θ_u^{po} should be determined by both the current user and the post.

In general, for a specific user u, the threshold of the possibility to retweet post *po* is positively correlated to their divergence. We still employ LDA to generate latent topic distributions of u and *po* respectively, then adopt Kullback Leibler divergence as Equation 10 to measure the distance between them.

Finally, the threshold θ_u^{po} is defined as Equation 11, where σ is a constant, $freq(u)$ stands for the percentage of retweeted posts in the latest 100 posts of u, which is used to describe how much a user prefer to retweet posts recently.

$$D_{KL}(u||po) = \sum_{t\in T} P(t|u)\cdot log\frac{P(t|u)}{P(t|po)} \tag{10}$$

$$\theta_u^{po} = \frac{\sigma D_{KL}(u||po)}{freq(u)} \tag{11}$$

In summary, there are four elements in the "influence field" model. I_u and I_{po} represent the global influence of the influential users and the post respectively. The distance R is applied to model the relationship between users, which projects I_u to an individual level. Analogously, the θ_u^{po} stands for the correlation between the post and the current user, which captures the effect of I_{po} at an individual level.

3.3 Predict the Retweeting Behavior

The retweeting behavior prediction can be considered as a classification task. For a given triplet (u,po,t_0), our goal is to correctly categorize it. The outcome is denoted as Y_{u,po,t_0}. $Y_{u,po,t_0} = 1$ means that user u will retweet the post po before timestamp t_0, and $Y_{u,po,t_0} = 0$ otherwise.

Since we have defined the calculation of all the elements necessary in the model, the decision inequality $f_u \geq \theta_u^{po}$ could be written as:

$$Q(u,po,t_0) = \frac{f_u}{\theta_u^{po}} - 1 = \sum_{u' \in U} \sum_{t \in T} \frac{Kr_t \cdot I_u^t I_{po}^t}{\theta_u^{po} R(u',u)^{\rho}} - 1 \geqslant 0 \tag{12}$$

We merge the constant coefficients, and $Q(u,po,t_0)$ could be written as:

$$Q(u,po,t_0) = \phi Q^{'}(u,po,t_0) - 1 \tag{13}$$

Where $\phi = \frac{K\lambda}{\sigma}$.

We employ the logistic regression model as our classifier for learning the value of ϕ. We make use of $Q(u,po,t_0)$ as the decision boundary, then the logistic function can be written as:

$$P(Y_{u,po,t_0} = 1 | Q(u,po,t_0)) = \frac{1}{1 + e^{-(\omega_1 Q(u,po,t_0) + \omega_0)}} \tag{14}$$

Where ω_1 is the weight coefficient and ω_0 is the bias, both of which can be learned by minimizing the cost function of the logistic regression model. For convenience, we modify Equation 13 as:

$$P(Y_{u,po,t_0} = 1 | Q^{'}(u,po,t_0)) = \frac{1}{1 + e^{-(\omega_1^{'} Q^{'}(u,po,t_0) + \omega_0^{'})}} \tag{15}$$

Where $\omega_1^{'} = \phi\omega_1$ and $\omega_0^{'} = \omega_0 - 1$.

4 Experimental Evaluation

4.1 Data Preparation and Experimental Setting

We crawled the latest 200 posts of 61,736 users as well as their follow-relationship from Sina Weibo. The dates of these posts range from 2012.3.2 to 2013.8.28. After removing the inactive users and the fake accounts, 37,931 users are left. Here, the inactive users refer to those users whose total posts number is less than 20 or the followee number is less than 10. We employ the approach introduced in [29] to detect those fake accounts.

We select 300 popular posts which are published on 2013.05.01, 2013.05.11 and 2013.05.21 (100 posts each day). By analyzing these posts, we obtain 15,831 retweets and 650,212 non-retweets and merge them as the dataset for the experiment. For a triple (u,po,t_0), we consider it as a non-retweet if it meets four conditions:

(1) po is exposed to u.
(2) u publishes or retweets another post at t_0.
(3) po is published with 5 hours before t_0.
(4) po is not retweeted by u before t_0.

We first treat every post as a document and apply LDA to calculate the probability of generating a post from each topic. Users' topical distributions are also estimated by the same LDA model, then the TwitterRank algorithm is executed on the global dataset to generate users' topical influence. During these processes, the teleportation parameter(γ) in TwitterRank is set to 0.85 as suggested in [20]. There are three parameters in LDA, i.e., the topic number T and two Dirichlet hyper-parameters α, β, they are set as $T = 50$, $\alpha = 50/T$, $\beta = 0.1$ respectively. When calculating the field strength of influence field, ρ is initially set to 2 via making an analogy with the electric field and the decay coefficient μ is heuristically set to 0.6.

For each triple (u, po, t_0), we first collect corresponding influential user set which is denoted as U for each u. Then we calculate $Q^{'}(u, po, t_0)$ as elaborated in Section 3. Finally, 10-fold validation for the logistic regression classifier is executed on the dataset. u will be predicted to retweet po before t_0 if the probability returned by logistic regression is larger than 0.5.

4.2 Performance of the Proposed Approach

In order to verify the effectiveness of our model based on "influence field", we utilize the logistic function as the classifier, employ the force function $Q^{'}(u, po, t_0)$ as the feature, then compare with several baseline approaches as follows:

(1)Only consider the influence of the author in the force function $(Q^{'}_{A}(u, po, t_0))$.

(2)Consider the influence of the author and the mentioned users in the force function $(Q^{'}_{A+M}(u, po, t_0))$.

(3)Consider the influence of the author and u's followees as the influential users in the force function $(Q^{'}_{A+F}(u, po, t_0))$.

(4)Use the features listed in [7] and [2] as the basic features.

(5)Combine $Q^{'}(u, po, t_0)$ and the basic features as the advanced features.

The evaluation results are listed in Table 1. We can see that $Q^{'}_{A+F}(u, po, t_0)$ obviously outperforms the result of the case in which only consider the influence of the author, which means beside the author, the followees who have retweet po indeed play an important role to influence u. This may be caused by two reasons: for one reason, users tend to trust their close friends as well as the posts retweeted by these friends, and most of these intimate friends are usually in their respective followee-lists; for the other reason, we tend to subconsciously imitate other people's behavior, especially those people we perceive to be like us [19]. Since people usually follow many users who share similar preferences with themselves due to the homophily [30], as a result, these followees may influence the current user to retweet a post which matches all their tastes. For example, a

PhD student on data mining may retweet a post about a new algorithm because he noticed that many famous data scientists he has followed have retweeted this post. However, those "mentioned users" can hardly improve the performance. This is probably because u is not familiar with these users. Moreover, only their "nickname" appended before the original post is visible to u, therefore, it may be difficult for them to catch u's attention.

We also notice that the approach using $Q^{'}(u, po, t_0)$ with all the influential users does not significantly outperforms the approach with basic features. The reason may be that the basic features contain some other factors associated with users' retweeting behavior. For example, the features "how many times po has been retweeted" and "whether po contains a hashtag" may reflect the impact of some breaking news. However, we have not considered them yet. After incorporating $Q^{'}(u, po, t_0)$ and the basic features, we are able to obtain an improvement of performance with about 3.7% in terms of F1 value. It means that our force function $Q^{'}(u, po, t_0)$ is significant for retweeting prediction.

Table 1. The results of approaches with different features

Features	**Precision(%)**	**Recall(%)**	**F1(%)**
$Q^{'}_{A}(u, po, t_0)$	59.74	62.32	61.01
$Q^{'}_{A+M}(u, po, t_0)$	60.03	62.19	61.09
$Q^{'}_{A+F}(u, po, t_0)$	62.56	65.67	64.08
$Q^{'}(u, po, t_0)$	62.78	65.63	64.17
basic features (BF)	61.41	67.65	64.38
$Q^{'}(u, po, t_0)$ + BF	**65.15**	**71.27**	**68.07**

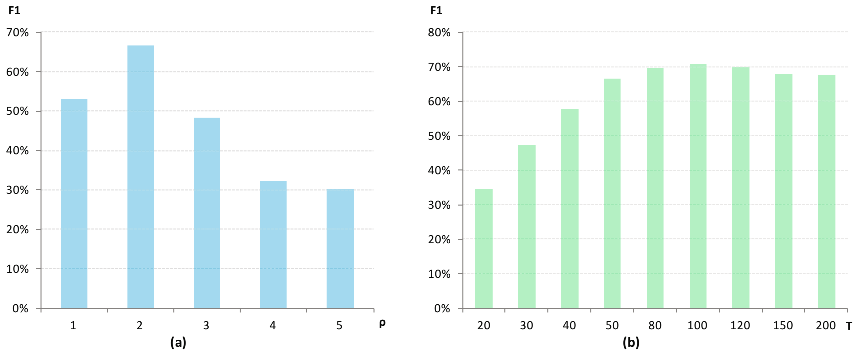


Fig. 3. The impact of distance coefficient (a) and topic number (b)

4.3 Impact of Model Parameters

We investigate the impact of distance tuning coefficient ρ and the topic number T on the model performance in this subsection. The results of different ρ are portrayed in Figure 3(a). The vertical axis is the F1 value. It is obvious that result outperforms others when ρ is set to 2. We note that it is exactly the same coefficient in Coulomb's law.

The number of topics would also influence the result. As depicted in Figure 3(b), the performance of our model tends to be stable when $T \geqslant 50$. However, we should note that the complexity of the model will increase as we add more topics, as a result, we decide to set $T = 50$.

5 Conclusion

In summary, inspired by the field theory in physics, we present a novel approach based on "influence field" with various features to predict the retweeting behavior in the micro-blogging system. We interpret retweeting behavior as an outcome of the influence from the post and the influential users. During our approach, many features are integrated together to generate the force which effects on the user and the inherent correlations among the features are also considered. The experimental results show that our strategy is indeed effective for the retweet prediction task.

There are also some limitations of our method. Firstly, because we are not able to get the timestamp of user's following behavior, as a result, the global topology structure of users' following relationship we used in the TwitterRank algorithm may deviate from the exact state at the timestamp of users' retweeting behavior in our experiment. Secondly, during our experiments, we simply assume that all the posts exposed to the current user have been read. In fact, we do not know whether the user did not want to retweet a post or he did not even see this post. These may hurt the overall performance to some extent. We will attempt to address these problems in the future.

References

1. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: WWW 2010 Conference Proceedings, pp. 591–600 (2010)
2. Peng, H.K., Zhu, J., Piao, D., Yan, R., Zhang, Y.: Retweet modeling using conditional random fields. In: ICDMW 2011 Conference Proceedings, pp. 336–343 (2011)
3. Luo, Z., Osborne, M., Tang, J., Wang, T.: Who will retweet me?: finding retweeters in twitter. In: SIGIR 2013 Conference Proceedings, pp. 869–872 (2013)
4. Uysal, I., Croft, W.B.: User oriented tweet ranking: a filtering approach to microblogs. In: CIKM 2011 Conference Proceedings, pp. 2261–2264 (2011)
5. Xu, Z., Zhang, Y., Wu, Y., Yang, Q.: Modeling user posting behavior on social media. In: SIGIR 2012 Conference Proceedings, pp. 545–554 (2012)

6. Zhang, J., Liu, B., Tang, J., Chen, T., Li, J.: Social influence locality for modeling retweeting behaviors. In: IJCAI 2013 Conference Proceedings, pp. 2761–2767 (2013)
7. Xu, Z., Yang, Q.: Analyzing user retweet behavior on twitter. In: ASONAM 2012 Conference Proceedings, pp. 46–50 (2012)
8. Petrovic, S., Osborne, M., Lavrenko, V.: Rt to win! predicting message propagation in twitter. In: ICWSM 2011 Conference Proceedings, pp. 586–589 (2011)
9. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on twitter. In: WSDM 2011 Conference Proceedings, pp. 65–74 (2011)
10. Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: Identifying misinformation in microblogs. EMNLP **2011**, 1589–1599 (2011)
11. Phuvipadawat, S., Murata, T.: Breaking news detection and tracking in twitter. In: WI-IAT 2010 Conference Proceedings, pp. 120–123 (2010)
12. Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., Yu, Y.: Collaborative personalized tweet recommendation. In: SIGIR 2012 Conference Proceedings, pp. 661–670 (2012)
13. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: conversational aspects of retweeting on twitter. In: HICSS 2010 Conference Proceedings, pp. 1–10 (2010)
14. Macskassy, S.A., Michelson, M.: Why do people retweet? anti-homophily wins the day! In: ICWSM 2011 Conference Proceedings, pp. 209–216 (2011)
15. Suh, B., Hong, L., Pirolli, P., et al.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: SocialCom 2010 Conference Proceedings, pp. 177–184 (2010)
16. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: WWW 2011, pp. 57–58 (2011)
17. Kupavskii, A., Ostroumova, L., Umnov, A., et al.: Prediction of retweet cascade size over time. In: CIKM 2012 Conference Proceedings, pp. 2335–2338 (2012)
18. Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In: CIKM 2010 Conference Proceedings, pp. 1633–1636 (2010)
19. Adams, P.: Grouped: How small groups of friends are the key to influence on the social web. New Riders (2012)
20. Weng, J., Lim, E.P., Jiang, J., et al.: Twitterrank: finding topic-sensitive influential twitterers. In: WSDM 2010 Conference Proceedings, pp. 261–270 (2010)
21. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of machine Learning research **3**, 993–1022 (2003)
22. Gilbert, E., Karahalios, K.: Predicting tie strength with social media. In: SIGCHI 2009 Conference Proceedings, pp. 211–220 (2009)
23. Granovetter, M.: The strength of weak ties. American Journal of Sociology **78**, 1360–1380 (1973)
24. Krackhardt, D.: The strength of strong ties: The importance of philos in organizations. Networks and Organizations: Structure, Form, and Action **216**, 239 (1992)
25. Haythornthwaite, C.: Strong, weak, and latent ties and the impact of new media. The Information Society **18**, 385–401 (2002)
26. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope (2008). arXiv preprint arXiv:0812.1045
27. Marsden, P.V.: Core discussion networks of americans. American Sociological Review, 122–131 (1987)
28. Xiaolin, S., Lada, A., Martin, S.: Networks of strong ties. Physica A: Statistical Mechanics and its Applications **378**, 33–47 (2007)

29. Shen, Y., Yu, J., Dong, K., Nan, K.: Automatic fake followers detection in chinese micro-blogging system. In: Tseng, V.S., Ho, T.B., Zhou, Z.-H., Chen, A.L.P., Kao, H.-Y. (eds.) PAKDD 2014, Part II. LNCS, vol. 8444, pp. 596–607. Springer, Heidelberg (2014)
30. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology, 415–444 (2001)