Retweeting Prediction Using Relationship Committed Adjacency Matrix

Li Weigang Zheng Jianya Liu Yang

TransLab, Computer Science Department of University of Brasilia, Brasilia – DF, Brazil

weigang@unb.br darcy_zheng@hotmail.com gilbertoliu514@msn.com

Abstract. *Considering the interaction and evolution of human activity associated relationship in online social networks (OSNs), we describe the Follow Model to present the relationships between users and further extend this logical formulation to Adjacency Matrix (AM) as Relationship Committed Adjacency Matrix (RCAM). With this relationship presentation by matrix, K-step multiplications of the RCAM can be used to query k-th sequence of the followers of followers etc. The paper establishes several mixed models with target and similarity functions to query who may probably retweet. A framework for retweeting prediction by Conditional Random Fields (CRF) method is developed and implemented together with the data from a Chinese famous micro-blogging, Sina Weibo. The simulation obtained the results of retweeting prediction with the indexes of precision larger than 61% and recall larger than 58% in some case studies.*

1. Introduction

Online social networks (OSNs), such as Twitter and Weibo, have become important human social activities today. Four basic characteristics are considered representative of these kinds of networks: 1) Complex topology composed networks; 2) Big data generation and aggregation; 3) Online instantly dynamic evolution; 4) Internal and external objects random interaction [Cha et al 2010, Hopcroft et al 2011, Yang and Leskovec 2011, Sun et al 2012]. To study any topic of these properties has been recognized as challenge with the necessary to overcome several obstacles. Most recent researches show the state of the art in this field involving physics, mathematics, computer and related engineering etc., we have studied some notable references concerning new concepts and methodologies.

Beside these four aspects and related researches, we still observed fifth property especially from OSNs, i.e., interaction and evolution of human activity associated relationship. In the existing literature, there is a lack of effective description of human behavior and relationships for OSNs. The researchers used to adapt existing mathematical models from graph theory, instead of logical models from artificial intelligence. Most of these formulations commit only to the existence of facts, i.e. the objects (nodes) or relations (edges) separately.

[Sandes et al 2012] proposed a *Follow model* to describe the relationships of follower, followee and r-friends of the users in OSNs. In this paper, initially, we describe the *Follow model* to present the relations of follower, followee and r-friends in OSNs as a new logical formulation. When the *Adjacency Matrix (AM)* is aggregated

with followship relations, we define *Relationship Committed Adjacency Matrix (RCAM)* $\boldsymbol{A_{in}}/\boldsymbol{A_{out}}$. In further multiplications of these matrixes, we have more ontological commitments to the relations of OSNs. The two-step operation of $\boldsymbol{A_{in}^2}$ is with complexity of $O(n^3)$. Even for a big n, once getting $\boldsymbol{A_{in}^2}$, it is useful for many times in querying. In sequence operations of $\boldsymbol{A_{in}^k}$, $n(k)$ will reduce significantly.

In OSNs, the multi-constraint queries such as "who, when and what" are important for mining analysis and retweeting prediction. From the huge relation and message data, it is a common difficulty of $O(n^3)$ computation complexity problem by using adjacency matrix directly. In some especial cases, *Follower AM* $\boldsymbol{A_{in}}$ and its transpose *Followee AM* $\boldsymbol{A_{out}}$ can be used to get much mining information. We develop various queries: *a)* the information about the followers, followees and r-friends from $\boldsymbol{A_{in}}$. *b)* the followers of the followers of all users in the network from $\boldsymbol{A_{in}^2}$; the followees of the followees from $\boldsymbol{A_{out}^2}$; *c)* the followees of the followers from $\boldsymbol{A_{out}A_{in}}$; *d)* the followers of the followees from $\boldsymbol{A_{in}A_{out}}$. We still can combine k-step operations of these matrixes to present more and more relationships of OSNs. This information is also important to estimate the number of times of possible-view of the original tweets and retweets.

In order to predict retweeting, we studied the following human social activities: *a)* One does the thing for his relatives; *b)* One does the thing as his relatives do; *c)* One does something for another as doing for another´s friends. These personal habits influence the retweeting decision. In our mixed models with target ($\boldsymbol{R_{rt}}$) and similarity ($\boldsymbol{R_{sim}}$) functions [Sun et al 2012], case *b)* can be represented as: $v\ \boldsymbol{R_{rt}}\ f(w)\ \boldsymbol{R_{sim}}\ w$, where $f(.)$ is a relation function to describe a subset of the users belong to a same category as defined in *Follow model*. In this situation, user v is likely to retweet w´s tweet as he already retweeted the w´s relatives in $f(w)$.

We implemented *RCAM* associated querying and retweeting prediction framework to a real online social network, Sina Weibo in China. The data involves 58.66 million users with 265.11 million followship relations and 369.80 million messages, which 51.62% of them are retweets. The query and prediction results demonstrated the effective and efficiency of the proposal in this research for OSNs.

2. Follow Model

In this section, we brief describe the *Follow model* which was developed by Sandes et al. [Sandes et al 2012] to formulate the followship relations in OSNs. With this meta-model, the different relationships associating with some activities in OSNs can also be extended in this manner, such as tweeting, mentioning and retweeting.

Definition 1. An online social network can be described as a directed graph $G = (V, E)$, the vertex set V contains the users u in node A and v in node B where $A, B \in V$; the directed edge set $E : V \times V$ represents a relation R between the user u in node A and v in node B, where the relation $(u, v) \in R$ means that user u follows user v and the edge $(A, B) \in E$.

Definition 2. If user u follows v, u is named as a follower, or a fan of v; and v is called a followee of u. If u follows v and v follows u, u and v are both defined as r-friends.

Definition 3. The terms of followee, follower and r-friend in *Definitions 1 and 2* are formally presented by the functions: $f_{in}(.)$, $f_{out}(.)$ and $f_r(.)$, here we defined these functions as *Follow model*. There are:

$f_{out}(u) = \{v|(u, v) \in E\}$, is followee function to present the subset, V^* , of all followees of user u, $V \rightarrow V^*$, $V^* \in V$;

$f_{in}(u) = \{v|(v, u) \in E\}$, is follower function to present the subset , V^* , of all followers of user u, $V \rightarrow V^*$, $V^* \in V$;

$f_r(u) = f_{out}(u) \cap f_{in}(u)$, is r-friend function to present the subset, V^* , of all r-friends of user u, $V \rightarrow V^*$, $V^* \in V$.

The functions in definition 2 are formally called as *Follow model* which has following three properties: reverse relationship, compositionality and extensibility.

Definition 4. For users x and y with relationship functions f , $f' \in \{f_{in}, f_{out}, f_r\}$, $V->V^*$, $V^* \in V$, f' is the reverse function of f, if:

$$f'(x) = \{y|x \in f(y)\}$$

The definition of $f_{in}(.)$, $f_{out}(.)$ and $f_r(.)$ functions for online social networks was initially introduced in the research of Sandes and others [Sandes et al 2012].

Reverse relationship. For the followee, follower and r-friend functions, their reverse functions are defined as equation (1).

Definition 5. For the relationship functions $f, f' \in \{f_{in}, f_{out}, f_r\}$, according to the *Definitions* 1-4, there is reverse function f' for each of them:

$$f' = \begin{cases} f_{in} & \text{if } f = f_{out} \\ f_{out} & \text{if } f = f_{in} \\ f_r & \text{if } f = f_r \end{cases} \quad (1)$$

With this definition, the *Follow model* can be more easily used to query the information from OSNs and to develop the optimization algorithms.

3. RELATIONSHIP COMMITTED ADJACENCY MATRIX

In this section, we present the definition and property of *Relationship Committed Adjacency Matrix (RCAM)* for online social networks.

3.1 Adjacency Matrix of Graph

For a directed and unweighted graph described in *Definition 1*, this graph can be presented by an adjacency matrix $\boldsymbol{A}$, where $A(u, v) = 1$, if $(u, v) \in E$; otherwise, $A(u, v) = 0$. If there are n vertices in this graph, the matrix $\boldsymbol{A}$ is with $n \times n$ elements [Foley 1996].

The theory of graph tells us that the number of k-step sequences between vertex u and v in the graph with adjacency matrix $\boldsymbol{A}$ is the (u, v) entry in $\boldsymbol{A}^k$, where $\boldsymbol{A}^k = \boldsymbol{A}\,\boldsymbol{A}\ldots\boldsymbol{A}$ is the multiplication of k times of the matrix $\boldsymbol{A}$.

3.2 RCAM for OSNs

Based on the *Follow model* in section 2, we define *Relationship Committed Adjacency Matrix* with the ontological commitments to the users and their relations in OSNs.

3.2.1 Definition of RCAM

For a directed and unweighted graph described in *Definition 1*, this graph can be presented by a *Follower Adjacency Matrix*, where $A_{in}(u, v) = 1$, if u follows v and $(u, v) \in E$; otherwise, $A_{in}(u, v) = 0$; where suffix *in* is according to $f_{in}(.)$ function of *Follow model* to describe the following relation. If there are n vertices in this graph, the matrix $\boldsymbol{A_{in}}$ is with $n \times n$ dimensions.

We can get the transpose matrix, $\boldsymbol{A_{in}}^{\boldsymbol{T}}$, of *Follower Adjacency Matrix* $\boldsymbol{A_{in}}.\boldsymbol{A_{in}}$ reflects a *(Follower, Followee)* matrix, i.e., we read the line of FAM to get the follower relation, and read the column to get the followee relation. $\boldsymbol{A_{in}}^{\boldsymbol{T}}$ reflects a *(Followee, Follower)* matrix, i.e., we read the line of matrix to get the followee relation, and read the column to get the follower relation. According to $f_{out}(.)$ function of *Follow model* to describe the followee relation, we use *Followee Adjacency Matrix* $\boldsymbol{A_{out}}$ to present the *(Followee, Follower)* matrix, where $\boldsymbol{A_{out}} = \boldsymbol{A_{in}}^{\boldsymbol{T}}$.

More generally, we call *Follower/Followee Adjacency Matrix* as *Relationship Committed Adjacency Matrix*, with the abbreviation: RCAM.

3.2.2 Combination of the Operators with A_{in} and A_{out}

With *Follower/Followee Adjacency Matrix,* we can combine many operators to get the information from $\boldsymbol{A_{in}}$ and/or $\boldsymbol{A_{out}}$.

Taking online social networks as examples, we observed that the number of k-step following relations between vertex u and v in the graph with following adjacency matrix $\boldsymbol{A_{in}}$ is the (u, v) entry in $\boldsymbol{A_{in}}^{\boldsymbol{k}}$, where $\boldsymbol{A_{in}}^{\boldsymbol{k}} = \boldsymbol{A_{in}}\,\boldsymbol{A_{in}} \ldots \boldsymbol{A_{in}}$ is the multiplication of k times of the matrix $\boldsymbol{A_{in}}$. This means that from $\boldsymbol{A_{in}}^{\boldsymbol{k}}$ we can get the information in k-steps of the followers of the followers of … of any user within $\boldsymbol{A_{in}}$, if there are. This is the case of $f_{in}^{k}(.)$ in *Follow model*.

We also observed that the number of k-step followee relations between vertex u and v in the graph with followee adjacency matrix $\boldsymbol{A_{out}}$ is the (u, v) entry in $\boldsymbol{A_{out}}^{\boldsymbol{k}}$, where $\boldsymbol{A_{out}}^{\boldsymbol{k}} = \boldsymbol{A_{out}}\,\boldsymbol{A_{out}} \ldots \boldsymbol{A_{out}}$ is the multiplication of k times of the matrix $\boldsymbol{A_{out}}$. This also means that from $\boldsymbol{A_{out}}^{\boldsymbol{k}}$ we can get the information in k-step of the followees of the followees of … of any user within $\boldsymbol{A_{out}}$, if there are. This is the case of $f_{out}^{k}(.)$ in *Follow model*.

In order to know the followees of the followers of any user in $\boldsymbol{A_{in}}$, we have $\boldsymbol{A_{out}}\,\boldsymbol{A_{in}}$. This is the case of $f_{out}f_{in}(.)$ in *Follow model*. For getting the followers of the followees of any user in $\boldsymbol{A_{in}}$, we have $\boldsymbol{A_{in}}\,\boldsymbol{A_{out}}$. This is the case of $f_{in}f_{out}(.)$ in *Follow model*. In these two cases, we can also get more combinations of any steps if there are relations in $\boldsymbol{A_{in}}$.

In this sense, we may understand that the *Follower/Followee Adjacency Matrix* $\boldsymbol{A_{in}}$ or $\boldsymbol{A_{out}}$ can be considered as an operator to reflect the relationship between the users in $\boldsymbol{A_{in}}$ or $\boldsymbol{A_{out}}$.

3.3 Example of Using RCAM

In this subsection, we provide *step-by-step* illustration on how to use *RCAM* for information querying from a simple social networks in figure 1.

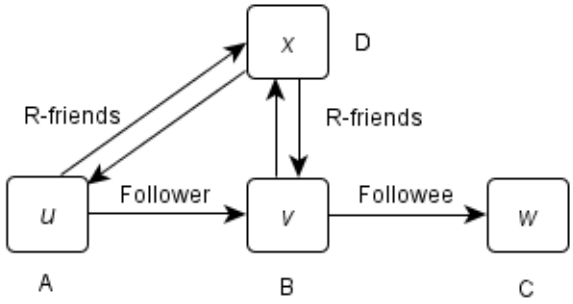

Figure 1. Follower, Followee and R-friends relationships.

3.3.1 Presentation of A_{in} and A_{out}

First, we present the follower/followee adjacency matrix $\boldsymbol{A_{in}}$/$\boldsymbol{A_{out}}$ of the network in figure 1. What information can be obtained from $\boldsymbol{A_{in}}$ or $\boldsymbol{A_{out}}$? As the detail presentation of $\boldsymbol{A_{in}}$ and $\boldsymbol{A_{out}}$ in figure 2(a) and 2(b) separately, remember $\boldsymbol{A_{in}}^{T} = \boldsymbol{A_{out}}$. We can obtain the information showing in table 1. In figure 2, *u/A* means that the user *u* is in the node *A*, etc. The value of each element of $\boldsymbol{A_{in}}$ (or $\boldsymbol{A_{out}}$) is based on the definition in 3.2.1 and with the relationship of figure 2.

$$\begin{array}{c} \\ u/A \\ v/B \\ w/C \\ x/D \end{array}\!\!\begin{array}{c} \begin{array}{cccc} u/A & v/B & w/C & x/D \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ u/A \\ v/B \\ w/C \\ x/D \end{array}\!\!\begin{array}{c} \begin{array}{cccc} u/A & v/B & w/C & x/D \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

Figure 2. (a) Follower AM $\boldsymbol{A_{in}}$; (b) Followee AM $\boldsymbol{A_{out}}$.

In table 1, from the line of $\boldsymbol{A_{in}}$ and from the column of $\boldsymbol{A_{out}}$, we can obtain the follower information, such as *u* follows *v*, etc. From the column of $\boldsymbol{A_{in}}$ and from the line of $\boldsymbol{A_{out}}$, we can obtain the followee information, such as *u* is a followee of *x*, etc. With the symmetry property of the follower/followee relationship, we can obtain the re-friends relations, such as *u* and *x* are r-friends.

Table 1. Information represented by $\boldsymbol{A_{in}}$ or $\boldsymbol{A_{out}}$

A_{in} / A_{out}	Column (follower)	Line (followee)	R-friends
Line (follower)	*u* follows *v* *u* follows *x* *v* follows *w* *v* follows *x* *x* follows *u* *x* follows *v*	-	-
Column (followee)	-	*u* is a followee of *x* *v* is a followee of *u* *v* is a followee of *x* *w* is a followee of *v* *x* is a followee of *u* *x* is a followee of *v*	-
R-friends	-	-	*u* and *x* *v* and *x*

3.3.2 Two –step Operations: $A_{in}\,A_{in}$ and $A_{out}\,A_{out}$,

To get the subset of the followers of followers of any user, if there are, we operate the multiplication of following adjacency matrix, such as $A_{in}^2 = A_{in}\,A_{in}$. Equation (2) shows the multiplication process.

$$A_{in}^2 = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

Equation (2). Two-step operation of follower AM: A_{in}^2.

To get the subset of the followers of followers of any user, if there are, we operate the multiplication of following adjacency matrix, such as $A_{out}^2 = A_{out}A_{out}$. Equation (3) shows the multiplication process.

$$A_{out}^2 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 \end{bmatrix}$$

Equation (3). Two-step operation of followee Am: A_{out}^2.

3.3.3 Operation of Followers of Followees: $A_{in}A_{out}$

To get the subset of the followers of followees of any user, if there are, we operate the multiplication of following adjacency matrix: $A_{in}A_{out}$.

Equation (4) shows the multiplication process, where, we can obtain the information of followers of followees of any user, if there are. From the line or column of $A_{in}A_{out}$, the followers of followees of u are v and x; the follower of followee of v is u; the follower of followee of x is u.

$$A_{in}A_{out} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 \end{bmatrix}$$

Equation (4). Operation of followers of followees: $A_{in}A_{out}$.

3.3.4 Operation of Followees of Followers: $A_{out}\,A_{in}$

To get the subset of the followees of followers of any user, if there are, we operate the multiplication of following adjacency matrix: $A_{out}A_{in}$.

Equation (5) shows the multiplication process, where, we can obtain the information of followees of followers of any user, if there are. From the line or column of $A_{out}A_{in}$, the followee of follower of u is v; the followees of followers of v are u and x; the followee of follower of w is x; the followees of followers of x are v and w.

$$A_{out}A_{in} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

Equation (5). Operation of followees of followers: $A_{out}A_{in}$.

4. Retweeting Prediction

In this section, we first establish mixed models with target and similarity functions to represent who may probably retweet, and then present a framework for retweeting prediction using the method of Conditional Random Fields [Lafferty 2001, Peng 2011].

4.1 Who May Retweet?

In OSNs, a user retweets the tweets of his friends and he may do the same thing as his friends do. In some situations, if he retweeted the messages from some users, he may also do that for a relative of these users. We are now formulating these processes.

4.1.1 Target and Similarity Relations

When studying Heterogeneous Information Network, [Sun et al 2012] proposed a method to define the meta path-based topology, such as the target and similarity relations. We will use their idea considering the retweeting activities in OSNs.
1) When considering retweeting activities, the target relation is defined as $R_T = (v, w)$, which means that user v retweets the tweet or another retweet from w.
2) Another relation R_{sim} is defined as the similarity between the users or objects, or even two actions in different situations as below study.

4.1.2 One Does the Thing for His Relatives

In the situation of one does the thing for his relatives, if two users belong to a same category, one may retweet another with a certain probability. The formulation of this relation can be represented as following:

$$v \in f(w)\ R_{rt}\ w \tag{6}$$

where $f(.)$ is a relation function to describe a subset of the users belong to a same category as defined in *Follow model*. In this situation, user v is likely to retweet w´s tweet. We extend the $f(.)$ in more detail cases:

1) $v \in f_r(w)\ R_{rt}\ w$. In w´s r-friends subset $f_r(w)$, v is more likely to retweet w´s message.

2) $v \in f_{in}(w)\ R_{rt}\ w$. In w´s follower subset $f_{in}(w)$, v is likely to retweet w ´s message.

3) $v \in m_{out}(w)\ R_{rt}\ w$. User v is in w´s mentionee subset $m_{out}(w)$, v is probable to retweet w ´s message after mentioning.

4.1.3 One Does the Thing as His Relatives Do

The situation of one does the thing as his relatives do can also be found in OSNs. If the relatives of a user retweet a message, this user may also retweet this message with a certain probability. The formulation of this situation can be represented as following:

$$v\ R_{sim}\ f(v)\ R_{rt}\ w \tag{7}$$

where $f(.)$ is a relation function to describe a subset of the users as defined in the *Follow model*. In this situation, describing as R_{sim}, v is like to retweet w´s tweet if some of v´s relatives in the subset $f(v)$ do that. We extend the $f(.)$ in more detail cases:

1) $v\ R_{sim}\ f_r(v)\ R_{rt}\ w$. The intuition is that if some of v´s r-friends retweet w´s tweets, then v is likely to do the same thing with w too.

2) $v\ R_{sim}\ f_{out}(v)\ R_{rt}\ w$. In case of many of *v´*s followee retweet *w´*s tweet, *v* is likely to retweet *w´*s tweet too.

3) $v\ R_{sim}\ f_r^2(v)\ R_T\ w$. Considering of two-step followship, in this case, the intuition is that *v* is likely to retweet *w´*s tweets if many of *v´*s r-friends of r-friends do that.

4) $v\ R_{sim}\ f_{out}^2(v)\ R_T\ w$, Also considering of two-step followship, user *v* is likely to retweet *w´*s tweets if many of *v´*s followees of followees retweet the *w´*s messages.

We can also have more combination situations such as r-friends of followees ($f_r f_{out}(.)$) and followers of r-friends ($f_{out} f_r(.)$) in the cases 3) and 4) etc.

4.1.4 One Does Something for Another as Going for Another´s Friends

One does something for another as doing for another´s friends. This scenario can be found in OSNs and described as equation 8,

$$v\ R_{rt}\ f(w)\ R_{sim}\ w \tag{8}$$

where, user *v* may retweet *w´*s tweets, as *v* already retweeted the tweets of *v´*s relatives defined by $f(w)$. We extend $f(w)$ in detail by the following cases:

1) $v\ R_{rt}\ f_r(w)\ R_{sim}\ w$, in this case, if *v* always retweets many of the r-friends of *w*, then *v* is more likely to retweet *w´*s messages in the future.

2) $v\ R_T\ f_{in}(w)\ R_{sim}\ w$, in this case, the intuition is that if *v* always retweet many of the followers of *w*, then *v* is likely to retweet w´s messages in the future.

3) $v\ R_{rt}\ f_r^2(w)\ R_{sim}\ w$, Considering two-step followships, if *v* always retweeted messages of the r-friends of r-friends of user *w*, then it is reasonable for *v* to retweet *w´*s tweets.

4) $v\ R_{rt}\ f_{in}^2(w)\ R_{sim}\ w$, Also considering two-step followships, if *v* retweets the messages of the followers of the followers of user *w*, then it is reasonable for *v* to retweet *w´* messages.

We can also have more combination cases such as r-friends of followers ($f_r f_{in}(.)$) and followers of r-friends ($f_{in} f_r(.)$) in the cases 3) and 4) etc.

4.2 Retweeting Probability

We introduce Conditional Random Fields [Lafferty et al 2001, Peng et al 2011, Junior et al 2012] as a basic method in the retweeting prediction based on the mixed models of the target and similarity functions in 4.1. Among several open source software packages that implement CRF algorithm, we chose the CRF++ in our research due to its use simplicity and efficiency. There are also some related works on the retweeting studies [Yang et al 2012, Petrovic et al 2011, Comarela et al 2012, Xu et al 2012].

4.3 Retweeting Prediction Framework

Based on RCAM and above mixed models with target and similarity functions, we construct a retweeting prediction framework with five procedures:

1) ***RCAM construction***. From the data set of relationship network, we construct *Relationship Committed Adjacency Matrixes*: A_{in}, A_f and A_m (*Mention Adjacency Matrix*). This step provides basic relationship matrixes for further computation and can be used to obtain statistical and mining information from OSNs.

2) ***RCAM computation***. With the matrixes from previous step, we can calculate two-step multiplication of RCAM $\boldsymbol{A_{in}^2}$, $\boldsymbol{A_{in}A_{out}}$ and $\boldsymbol{A_{out}A_{in}}$ etc.

3) ***Information querying***. Using RCAMs, we can execute most of the queries about the information of the users and tweets, such as the number of followers, the common followers, the friends of friends, and the estimated number of viewing of the tweets etc. These querying results are fundamental information for the prediction of retweeting action.

4) ***Feature abstraction***. With the information of the followship network and the message dataset, we abstract six features which are the basic elements for building the dataset for the experiments of retweeting prediction.

5) ***Retweeting prediction***. As the last step, we choose the suitable toolkit, for example CRF++ in this paper, to perform CRF algorithm for retweeting prediction.

5. CASE STUDY OF SINA WEIBO

In this section, we apply the proposed prediction framework to Sina Weibo, one of the largest online social networks in China. We first describe the data and then show the prediction results from three projected cases.

5.1 Data Description

We use the data collected from Sina Weibo which was published in the WISE2012 Challenge [WISE 2012]. The data consists of two separate sets: 1) the Relationship Network and, 2) the Message List. In order to maintain user privacy, all the data has been anonymized and all identifiers are presented by the respected numerical indexes.

There are 58.66 million users with 265.11 million follow relationships and 369.80 million messages. Adopting these data, we abstract a dataset that consists of 19,142 users and 17,474 retweets from 1,472 unique original tweets by the content about Steve Jobs, especially related to his death.

In the subset of these 17,474 messages, we take 2/3 of them as the training set, and the others are the testing set. All experiments are conducted on a Linux-based machine with 4 Intel Core 2.80GHz cores and 4G Memory.

5.2 Three Experiment Cases

In this study, we design three experiment cases to evaluate the performance of our proposed method.

1) In the scenarios of *One does the thing for his relatives*, *One does the thing as his relatives do*, and *One does something for another as doing for another´s friends*, where involving *Follower Adjacency Matrix* $\boldsymbol{A_{in}}$, we predict the retweeting actions based on the basic relationship and other essential features. We call this case as Base RCAM procedure.

2) In the same scenarios mentioned above, the second experiment involves more combinations of RCAMs and two-step multiplication, such as $\boldsymbol{A_{in}^2}$. We call this case as Two-step RCAM procedure. After obtain the results of retweeting prediction, we compare to the results from Base RCAM procedure.

3) In the last case, we use *Mention Adjacency Matrix* $\boldsymbol{A_m}$ as a basic source to abstract the features in retweeting prediction. We call this case as Mention RCAM procedure, which is used as a performance testing of our proposed method.

5.3 Retweeting Prediction Results

Generally, to evaluate the result of prediction result, the *precision*, *recall* and *F1 score* are adopted as the indexes to measure the prediction performance. In our analysis, the precision is defined as the ratio of the true positive value in our prediction result (the number of correct prediction, of which the users do the retweeting action) to the all positive value in the prediction results. The recall is defined as the ratio of the true positive value in prediction result to the number of all retweet actions in the sample dataset. The performance indexes of the retweeting prediction are represented in table 2.

In the case of Base RCAM, with the information of section 4, we take the first 4 features as the basic metrics; with these features we obtained the results of retweeting prediction with the precision 65.5%, recall 70.3% and F1 score 67.8%.

Table 2: The prediction result using CRF++

	Precision	Recall	F1 score
Base RCAM	65.5%	70.3%	67.8%
Two-step RCAM	62.1%	67.2%	64.5%
Mention RCAM	66.2%	69.7%	67.9%

In the case of Two-step RCAM, we query the information of two-step relationships from two-step multiplication of *Follower Adjacency Matrix*, $\boldsymbol{A_{in}^2}$, then further to see the retweeting activity from the followers of followers.

In this case, see Equation (4), $\boldsymbol{A_{in}^2 = A_{in} \times A_{in}}$, we obtained the information of two-step relationship with comparison between $\boldsymbol{A_{in}}$ and $\boldsymbol{A_{in}^2}$. The position of matrix which its value is 0 in $\boldsymbol{A_{in}}$ while 1 in $\boldsymbol{A_{in}^2}$, this means that the number of line and the column have the two-step relationships (excepting the diagonal, because it is meaningless in our experiment).

With $\boldsymbol{A_{in}^2}$, we add two more features in our dataset which are the friends of the friends and the followers of the followers. With these two more features, we got the results with the precision 62.1%, recall 67.2% and F1 score 64.5%. This performance is a little worse than the Base RCAM case, because the relationship network in our dataset is incomplete. The advance of the two-step relations did not demonstrate the advantage as we expected in retweeting prediction.

In the case of Mention RCAM, we appended the mention relationship between users into our dataset. With this feature, we obtained the prediction results with better performance. The difference between this result and Base RCAM is still not distinct, because the mention relationships in the dataset are very small comparing to the following relationship. Although the improvement is not significantly, it really shows that the tendency of mention relationship can improve the prediction performance. In this experiment, the indexes of precision, recall and F1 score are 66.2%, 69.7% and 67.9% respectively, see table 2.

In this experiment, we found that the result is better than the case of Base RCAM with the introducing the mention relationship. This result approves our proposal in section 3.4, i.e. the mentioner is more likely to retweet the original tweet which mentioned him.

Due to the social relationship network is not complete in this dataset and the results of two-step RCAM procedure didn't come to our expectation. We added 15% relationships in the data set, which were generated randomly as the missing links to supplement the network. With this modification, performance results of retweeting prediction are listed in the table 3.

Table 3: The prediction result by CRF++ with modified data

	Precision	Recall	F1 score
Base RCAM	61.1%	58.5%	59.8%
Two-step RCAM	63.2%	59.3%	61.2%
Mention RCAM	63.4%	59.2%	61.2%

From table 3, we note that the two-step RCAM relationship influenced the precision of retweeting prediction. At the same time, with the involving the mention relationship, using of the Mention RCAM procedure is also obtaining a better performance than the Two-step RCAM procedure. In this sense, we still need more experiments and theory analysis to approve this proposal.

6. Conclusions

In this research, we developed a new formulation based on the *Follow model* and *Relationship Committed Adjacency Matrix* to study the query and retweeting prediction in Online Social Networks.

Using RCAM, we developed various mix models of target and similarity functions to estimate the possible users who may retweet. By means of the method of Conditional Random Fields, we implemented a framework to predict retweeting actions.

The experiments based on the data from Sina Weibo shown the effective of our proposed methods using RCAM. The measurement indexes were acceptable, such as the precision is larger than 61% and recall is larger than 58% in the most of the results of retweeting prediction.

With complexity of $O(n^3)$ of the operation of $\boldsymbol{A}_{in}^{2}$ and other k-step RCAMs, more efficient formulations and algorithms should be developed for the huge scale networks.

Conditional Random Fields is a useful method for retweeting prediction, but to develop more methods is still necessary. We may also implement an independent and integrated system instead of the application of CRF++ software package.

References

Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K. (2010) "Measuring user influence in twitter: The million follower fallacy". *In Proceedings of 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 10–17.

Hopcroft, J. Lou, T and Tang, J. (2011) “Who will follow you back”?: reciprocal relationship prediction. In Proceedings of ACM Conference on Information and Knowledge Management, 1137-1146.

Yang, J, Leskovec, J. (2011) “Patterns of temporal variation in online media”. Proceedings of the fourth ACM international conference on Web search and data mining. New York, NY, USA: ACM, 2011: 177–186.

Sun, Y., Han, J., Aggarwal, C. C. and Chawla, N.V. (2012) “When Will It Happen?:Relationship Prediction in Heterogeneous Information Networks”. In *Proceedings of Int. Conf. on Web Search and Data Mining*, WSDM'12, 663-672.

Sandes, E., Weigang, L and Melo, A. (2012) “Logical model of relationship for online social networks and performance optimizing of queries”. In Proceedings of Web Information Systems Engineering - WISE 2012, X.S. Wang, I. Cruz, A. Delis, et al. Springer Berlin Heidelberg, LNCS: 726–736.

Foley, J. D., Dam, A., Feiner, S. K. and Hughes, J. F. (1996) “Computer Graphics: Principles and Practice”. Second Edition.

Lafferty, J. McCallum, A and Pereira, F.C.N. (2001) “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”. In *Intl. Conf. on Machine Learning*, 282-289.

Peng, H. K., Zhu, J., Piao, D., Yan, R and Zhang, Y. (2011) “Retweet modeling using conditional random fields”. In: *Data Mining Workshops (ICDMW)*, 336-343.

Junior, J., Almeida, L., Modesto, F., Neves, T. and Weigang, L. (2012) “An Investigation on Repost Activity Prediction for Social Media Events”. In Proceedings of Web Information Systems Engineering - WISE 2012, X.S. Wang, I. Cruz, A. Delis, et al. Springer Berlin Heidelberg, LNCS: 715–725.

WISE 2012 Challenge (2012) http://www.wise2012.cs.ucy.ac.cy/challenge.html.

Z. Yang, Jingyi G., Keke C., Jie T., Juanzi L., Li Z. and Zhong S. (2010) “Understanding Retweeting Behaviors in Social Networks”. In Proceedings of CIKM’10, Toronto, Canada, 1633-1636.

Petrovic, S., Osborne, M., Lavrenko, V. (2011) “RT to Win! Predicting Message Propagation in Twitter”. In Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM, Barcelona, Catalonia, Spain. The AAAI Press.

Comarela, G., Mark C., Almeida, V., Benevenuto, F. (2012), “Understanding factors that affect response rates in Twitter”. In Proceedings of the ACM SIGWEB Conference on Hypertext and Social Media (HT12). Milwaukee, WI, USA, 123-132.

Xu, Z., Zhang, Y, Wu, Y., Yang, Q. (2012) “Modeling user posting behavior on social media sigir”, In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, 545-554.