De-lurking in Online Communities Using Repost Behavior Prediction Method

Omid Reza Bolouki Speily*
Department of Information Technology & Computer Engineering, Urmia University of Technology, Urmia, Iran
speily@uut.ac.ir

Received: 10/Sep/2016 Revised: 01/Sep/2017 Accepted: 05/Sep/2017

Abstract

Nowadays, with the advent of social networks, a big change has occurred in the structure of web-based services. Online community (OC) enable their users to access different type of Information, through the internet based structure anywhere any time. OC services are among the strategies used for production and repost of information by users interested in a specific area. In this respect, users become members in a particular domain at will and begin posting. Considering the networking structure, one of the major challenges these groups face is the lack of reposting behavior. Most users of these systems take up a lurking position toward the posts in the forum. De-lurking is a type of social media behavior where a user breaks an "online silence" or habit of passive thread viewing to engage in a virtual conversation. One of the proposed ways to improve De-Lurking is the selection and display of influential posts for each individual. Influential posts are so selected as to be more likely reposted by users based on each user's interests, knowledge and characteristics. The present article intends to introduce a new method for selecting k influential posts to ensure increased repost of information. In terms of participation in OCs, users are divided into two groups of posters and lurkers. Some solutions are proposed to encourage lurking users to participate in reposting the contents. Based on actual data from Twitter and actual blogs with respect to reposts, the assessments indicate the effectiveness of the proposed method.

Keywords: De-Lurking; Post Similarity; Lurker; Online Community.

1. Introduction

Online community (OC)[1] refers to a group of users who have a common interest in a particular subject area, produce, and share online knowledge and information [1] - [3]. Most of these communities transfer their knowledge in the form of texts, hypertext and multimedia on Web Platforms or forums [1]-[3]. With the development of such forums as yahoo answers, Twitter or stack overflow, etc., the fundamental challenge is increased information sharing in these groups [4]. OC's are classified as direct and indirect communications [5]. Indirect communication is usually done through the establishment of two-way friendship and direct communication between two users is made by following other users for the latest posts created and not necessarily is this type of communication two-way. Despite the diversity in implementation, all these groups follow the same process: a user posts a message and the other user reposts it if he/she likes it. These posts are virally shared on the net. Accordingly, a certain post is made available to a large number of users [6]. In the present study, considering the nature of OCs, the communication among users is considered to be direct. Experts from different subfields have gathered together in these groups and the users follow the posts made by other users in shared areas according to their interests [3]. Nevertheless, there are many users who do not participate actively in these online communities. According to previous studies, users are divided into two groups: lurkers (nonparticipants) and posters (participants) [4]. In various papers, different definitions of lurkers have been provided [5]-[7]. Emphasizing a characteristic, each of these definitions has described such users. Generally, these users join an online community or group consciously; however, they do not post anything. Creating a post in online communities indicates users' participations. Regarding the level of this participation, no clear definitions have been provided in different references. In many references, the users who do not post anything in online communities are known as lurkers [5]; however, a time limit has been made for this nonparticipation in some other references [8]. For instance, if a user did not (re) post a post in the past year, s/he is considered a lurker. In addition, there are many disagreements over the effects of such users [4], [8], [9]. Some references have interpreted them as free riders who use sources free without providing communities with any benefits [10].

In the present study, we are looking for a method to select k influential posts for each user to be displayed when signing in. The idea has been addressed in several papers under titles such as post ranking, post refining, etc.[7] - [9]. The issue of selecting k posts for display does not mean to simply maximize the number of posts reposted but, in fact, its concern is to De-lurking in OCs.

1. The online community considered in this article has a similar structure to Twitter. In such online communities it is possible to track other users as well as view, comment, and re-post their posts.

* Corresponding Author

Delurking is a type of social media behavior in which a user breaks an "online silence" or habit of viewing an inactive thread to interact with a virtual conversation. The term means that the user typically does not participate in social media or online social activities [11].

Those posts are selected to be displayed that are more likely to be reposted by sub-users and cause de-lurking. To put it more formally, the purpose is to display for the user those posts that maximize a tree of reposts of which this user is the root.

In the proposed method, in addition to the structural characteristics of users on the network, the attention is paid to the attitude of reposting among users. Users are divided into two categories of posters and lurkers based on their reposting attitude. Lurkers are those who become a member in an online community but do not post (or simply post very little), are only readers, and are not active. Considering this type of users, the present study intends to reduce pure lurking given the time the important posts are available online, that is, the posts, which are more likely to be reposted by users, are displayed longer for lurkers than for posters.

In the second section of this article, the literature on the subject is reviewed. In the third section, the proposed method is presented according to the nature of lurkers. The fourth section contains a detailed evaluation of the proposed method and the fifth section is devoted to summing up the study done.

2. Literature Review

On the whole, numerous articles have been presented to solve the problem of reposting based on the identification of influential individuals in the network. In this section, we try to review the literature on the subject, especially the recent one. The article [12] evaluates the transfer of posts among bloggers. By analyzing timestamps on each post and transferring them on the network of Weblog users, it suggests that this sharing follows a specific model referred to in this article as waterfall model. The identification of these models is of great importance in the study on the sharing or transferal of information in (online and offline) communities. It is important to identify these models. The study [13] on blogsphere, addresses “epidemic” interests among different blogs with regard to the content cited or copied from another blog. By studying the cases, it estimates the relationship between two similar blogs. By relationship, it means the use of another post in the form of citation or copy. Another important point addressed in the study is the evaluation of the influence of a blog on another blog via a post. [14] is a study on sharing small pieces of text (for example, in connection with the news) used in other articles and other texts. For this purpose, a method has been implemented by which the source of each piece can be specified in the network. This makes it possible to study the structure of sharing in the network. The study aims to find the sources by which a post or posts are influenced.

Major constraints faced in the literature are, however, the absence of detailed knowledge on the network structure in which sharing takes place. In more recent researches, attempt has been made to obtain certain information about the network in which sharing takes place in addition to the collection of the data on sharing. For example, in [15], certain studies have been done on sharing tree in Facebook fan pages. In [16], Bakhshi conducted a study on sharing information in the online social game “second life”. In this article, by tracking the sharing of "gesture" an information unit in the game (which can be copied by other users), he obtained interesting results. Using simulation, Bakhshi showed that the possibility of transferring this information between two friends is more than that between two users who do not know each other. It has also been shown that certain users play a more important role in sharing the game information, called "adaptive users" in this study. Most of the methods presented on this subject are based upon the identification of individuals in the group influentially involved in sharing information. In this respect, the posts are mostly displayed to influential users so that, on reposting, more users may view these posts. These methods are mainly based on individuals’ statistical characteristics such as the number of follower users and number of reposts [10]. It is proven in [11] that there is an insignificant relationship between popularity (in terms of number of followers) and influence. Studies such as [20] and [21] have proposed more effective methods in terms of scalability and run-time compared to the ambitious method of Camp. In [22], a study has been conducted on a subset of Twitter data and a method has been presented for evaluating the influence of individuals regarding the characteristics in each issue. In [23], various criteria used to determine the influence of users have been investigated. Accordingly, this article has carried out a comprehensive study to explain the accuracy of the criteria such as indegree, repost and mentioning. The first criterion is the number of users following the subjects of the intended user. Repost refers to the number of times users repost the issues raised by the relevant users. Mentioning refers to the number of times that users mention the relevant user’s name. According to the experiments conducted in this study, having even more than a million followers does not guarantee a user’s influence. Given the existing studies and the nature of OCs, the present article tries to present a method for displaying the best information posts online needed by users leading to increased participation of the users. In this method, the importance of the posts displayed is taken into account in terms of post subject, user interest and information level in addition to the type of user (active or lurking).

3. *k* Influential Posts Selection

This section explains selection of k influential posts in OCs and discusses its characteristics.

3.1 Statement of Problem

Suppose that the OC is implemented under a social network. These groups are typically displayed as graphs of followers-followees. Users follow other users considering their interests and expertise needed. This directed graph is defined as G = (V, E) where E represents the relationship between users and V represents users. $(u,v) \in E$ shows that the user u follows the user v. If P represents the total of posts created in the whole online community, then an online social event (post) occurs when user u, creates post p at $t \in T$ time, represented as $post(u,p,t)$. In the same manner, when the user v shares the post p through the user u at time t, "reposting" occurs, represented as $repost(v,u,p,t)$. According to the definitions provided, the probability of repost can be defined as a function of the probability of repost $p : P \times V \times V \times T \rightarrow [1,0]$. In this function, T is the temporal domain. The probable repost of P posts by any user from the V users group, within the temporal domain T, includes the values between zero and one.

If σ is taken as the selection procedure of k influential post from among the candid posts (v, t) for the user v at any t time, the output $\sigma(v,t)$ is the k influential post to the user v at the time t. the total candidate posts for the user v are those already created or reposted the users of group V followed by the user v. Equation (1) shows the initial set of candidate posts for $t' < t$.

$$init(v,t) = \{p \in P \mid post(u,p,t') \cap (u,v) \in V\} \quad (1)$$

From this series of posts, the duplicate posts already displayed for the user in the previous time $t' < t$ should be removed. The candidate posts, as shown in equation (2), are as follows:

$$candid(v,t) = init(v,t) - \{p \in P \mid post(v,p,t') \cup \sigma(v,t')\} \quad (2)$$

Here, a formal definition is given for selection of the k post.

Definition (1): (selection of k influential posts) consider the graph online community as a graph G = (V, E) where V is the total of users and E is the total of follow-up communications among users. $k \in \mathrm{N}$ is an input issue and represents the number of influential posts selected from among candidate posts for display. This requires the definition of the procedure $\sigma : V \times T \rightarrow 2^P$ selecting for each user $v \in V$, and each timestamp t ($t \in T$), a k series of influential posts as $\sigma(v,t) \subseteq cand(v,t)$ where $|\sigma(v,t)| = k$ for display to the user v. If the number of posts is considered as |P|, there is $2^{|P|}$ different modes (sharing or not sharing any posts) for selection. The selection should be in a way that the number of network reposts depending on the model number of posts is maximized regarding the model of post sharing (section 2.3.3). Eq. (3) gives the mathematical expression for selection of influential posts.

$$Max \sum_{p \in P} |\{w \in V \mid \exists t \in T : post(w,p,t)\}| \quad (3)$$

3.2 Heuristic Method For Selection Of K Post

We provide an heuristic method for selection of k influential posts considering the computational capabilities and simplicity. This procedure is designed so as to be operational in online environments. In this section, the proposed method is introduced.

A- Similarity Between Posts (Post-Post)

For this purpose, two methods of topic similarity and geographical similarity are applied. For topic similarity, the posts of a user are considered as a set of words. According to the definition, this candidate post t_{wi} is topically similar to the posts of user v if it is related to his interests. $P_v = \{tw_1, tw_2, ..., tw_n\}$ is the total posts created by the user ***v***. To determine the relationship between a post and the topic of interest to the user, TF/IDF method and cosine of the angle between vectors of the words t_{wi} and P_v are used. Owing to the diversity of words employed, this method has low accuracy. For this purpose, different themes can be categorized in the texts using Latent Dirichlet Allocation. Each theme includes a set of words $M_{topic\#} = \{w_1, w_2, ..., w_L\}$, of which the probability of occurrence is specified in the relevant theme. To increase the accuracy of thematic similarity of candidates to previous posts of the user ***v*** ($P_v = \{tw_1, tw_2, ..., tw_n\}$), the angle between these two vectors is measured through the cosine, using Equation 4.

$$topsim(tw_i, P_v) = \frac{M_{topic_{tw_i}} . M_{topic_{P_v}}}{\| M_{topic_{tw_i}} \| . \| M_{topic_{P_v}} \|} \quad (4)$$

In this article, in addition to calculating the lexical similarity between the post tw_i and the posts used by the user, geographical similarity is taken into account. Normally, the influence of the posts addressing regional issues is higher than the Tweets of other regions. Therefore, in this article, maxmind data set was used including 4 million names of cities and regions along with additional information of the country, location, etc. to find the words related to cities and geographic regions. To find words related to cities from the post tw_i, all the words in the post but additional words are used (even the hashtags, the symbol # excluded). If Loc_{tw_i} be the set of cities plus country and the region used in the post and Loc_{u_i} be the set of cities, countries and regions used by the user u_i in all the posts created, Equation 5 shows the geographical relationship between the post tw_i and the user u_i.

$$Locsim(tw_i, u_i) = \frac{|(Loc_{tw_i} \bigcap_{u_i \in U \,\&\, tw_i \in P} Loc_{u_i})|}{|(Loc_{tw_i} \bigcup_{u_i \in U \,\&\, tw_i \in P} Loc_{u_i})|} \quad (5)$$

B- Similarity Between Users (User - User)

This criterion is very important for OCs. People with different specializations share their posts in the network. It is very beneficial to find users with common fields. For both users $u, v \in V$, the degree of similarity is equal to the degree of similarity between the posts already created. The essential thing about sharing information in OCs is to find people with the same level of information in addition to similar posts. For example, a user who has created more than 100 posts about smart phones applications is different from someone who has just had a few posts or reposts in the same field. For either user, the action vector can be defined (Equation (6)). This vector contains n keywords created or reposted by the user $v \in V$. Weighted cosine is used to determine the similarity between the vectors of users. In this respect, the coefficients i (number of keyword repeated by the user $v \in V$) is determined for n keyword till time t. Considering the coefficients i, the level of users' knowledge on a specific area is determined according to the number of posts made by them. The two users are examined and taken into account for determining the similarity given the repetition of the keywords in the posts.

$$vector_u^t = i_1^t keyword_1 + ... + i_n^t keyword_n \quad (6)$$

The degree of similarity between the users u and v is equal to the value of cos for vectors of these two users.

$$sim(u,v) = \cos(vector_u, vector_v, t) = \frac{vector_u^t \, vector_v^t}{||vector_u^t|| . ||vector_v^t||} \quad (7)$$

In equation (6), $||vector_v^t||$ or/and $||vector_u^t||$ respectively represent the value of action vector for the users v and u. If any of these values is zero, it means that the relevant user has had no action (not created nor reposted). In that case, the similarity between two users is not defined. Accordingly, in collecting data, only those users are taken into account who have at least created one post or reposted. Based on this assumption, there is no problem in calculating this similarity. It should be noted that the action vector of a user changes as the time changes. In this respect, the action vector of users at the time t is used in each determination of the similarity between two users.

C- Using Logistic Regression to Predict the Probability of Reposting

Logistic regression is used to estimate the probability of reposting. Typically, based on input features in logistic regression, a linear function is defined which, based on these features (similarity of candidate post, geographical similarity, and similarity between user and user), predicts the influence (being reposted) of a post using the sigmoid (Logistics) function (Equation 8).

$$y_i = h(w^T x_i) = \frac{1}{1 + \exp(-w^T x_i)} \quad (8)$$

In this equation, y_i is the prediction based on the x_i inputs. w^T is the vector of coefficients of each feature obtained from training data. Equation 8 shows the error function of logistic regression algorithm. In this equation, N is the number of reposts used in the training data set.

$$J(w^T) = \frac{1}{N} \sum_{i=1}^{N} \text{Cost}(h_w(x_i), y_i) \quad (9)$$

In Equation 9, $\text{Cost}(h_w(x_i), y_i)$ is the algorithmic cost function. Since binary classification algorithm is used, $J(w^T)$ is written as Equation 10. To obtain the weight of each feature, (w^T) should be determined based on the input data x_i and real data related to the users' repost in the data set y_i in such a way that $J(w^T)$ is minimized ($\min_{w^T} J(w^T)$).

$$J(w^T) = \frac{1}{N}[\sum_{i=1}^{N} y_i \log h_w(x_i) + (1-y_i)\log(1-h_w(x_i))] \quad (10)$$

To cope with over fitting in the equation above, regularization term is taken into account. The value of error function regarding this regularization term is given in the Equation (11). *m* is the number of features used for classification.

$$J(w^T) = \frac{1}{N} \times \sum_{i=1}^{N} y_i \log h_w(x_i) + (1-y_i)\log(1-h_w(x_i)) + \sum_{j=1}^{m} (w_j)^2 \quad (11)$$

The reason behind using logistic regression is its capability with respect to the issue mentioned above. In addition to classification, this method has also probable output. For example, $h_w(x_i) = 0.8$, that is, the probability of retweet is at 80% for the sample ($h_w(x_i) = p(y_i = 1 | x_i; w^T)$, which is very useful for the question in this article considering the need to estimate the probability of reposting.

3.3 Probability of Reposting Based on User Type

The calculated probability of repost of post p of the user u by the user v at timestamp t (put formally) $p(repost(v,u,p,t))$ is shown in the algorithm (1). In algorithm (1), t_u is the time a post is reposted by the user u. t is the run-time of the algorithm and γ_v is the average time interval between the posts of user v. α is the adjusted coefficient of γ_v. At the zero line of this algorithm by calculating the elapsed time since the repost of the posts by the user u, if the time is less than the

average time interval between the posts of the user v ($\alpha\gamma_v$), there is still the probability that the user v has not seen the post p. In this respect, the probability of reposting by the user v is equal to max($p^p_{u,v}, \varepsilon$). The line 2 evaluates a condition in which the time elapsed since the repost of the posts by user u is longer than the average time interval between the posts of the user v ($\alpha\gamma_v$). In this case, the user had most likely seen the post but was unwilling to repost it. Considering this fact, it is least probable that the user v repost the post p of user v in the timestamp t ($p(repost(v,u,p,t)).\varepsilon$ is equal to the very minimal value at 10^{-3}.

Algorithm (1) calculating the probable repost of post p of the user u by the user v at timestamp t	
Input: t_u is the time the post is shared by user *u*, the present time *t*, γ_v, $p^p_{u,v}$	
Output: *p*(*repost* (*v, u, p, t*)) probable repost of post p of the user *u* by the user *v* at timestamp *t*	
Definitions: ε is the minimal value, γ_v is the average time interval between the posts of the user *v*, $p^p_{u,v}$ is the probable repost of post *p* (from posts of u) by the user v, α is the coefficient	
00	If $t\text{-}t_u <= \alpha\gamma_v$
01	*p*(*repost* (*v, u, p, t*))=max($p^p_{u,v}, \varepsilon$)
02	if otherwise
03	*p*(*repost* (*v, u, p, t*))=ε
04	End

Accordingly, if the time elapsed since the creation of the post is longer than the time the user is inactive in the online community, it is least likely to be displayed and its value is equal to ε . The coefficient of α allows more opportunity for lurking users by adjusting the impact of γ_v. Normally, γ_v of lurking users is far longer than γ_v of other users. For ease of calculation, the coefficient α of γ_v is considered to be 1 and 2 for ordinary users and lurking users, respectively. In this respect, when a user is lurking, he/she has more time to read and repost the posts with higher probability.

4. Results

In this section, we will discuss in detail the experimental data and simulation framework as well as the results of this research.

4.1 Experimental Data & Simulation Framework

Twitter data were used to test the proposed method. With more than 200 million users, Twitter witnesses a million posts per day. This social networking site is a directed graph including users who freely follow other users. Each user is able to create a post called “Tweet”. These tweets contain photos, URL links, texts and so on. Each tweet contains a maximum of 140 words. Tweets are registered in the user page after they are created and can be viewed by the users who follow the former ones. Re-Tweets refers to the repost of a user’s Tweet by other users. Each user Retweets a Tweet according to his interest in it. Twitter as a huge data base is in the focus of many researchers in this field [7], [30], [31]. For collecting the data of Twitter Search API, all public tweets related to the field of IT technology with different keywords (8 keywords) were extracted from September 20 to December 24 of 2011. The data contains 94 thousand tweets. By making a Twitter API query for each user ID, its metadata including the followers and the followees were extracted. To map the social graph of users, only those users were taken into account who had at least Tweeted or Retweeted a post.

For testing, the simulation method similar to [7], [32], [33] was used. All the parameters and variables are based on actual data. On obtaining the social network graph and the set of tweets as input, the probability of reposting $P(repost(v,u,p,t))$ was calculated. Then, reposting is begun randomly and based on the model of post sharing. In order to carry out a simulation based on real data, an hour scale is considered for the time in the proposed method. This makes it possible to select k influential posts for each user per hour. First, the simulation is started from an augment node for λ followed by all nodes. This makes it possible to uniformly enter the posts of the data set. The rate at which posts are entered into the simulation environment acts as an input parameter. The other assessment input parameter is the number of iterations occurred for the simulation. In general, simulation is started by displaying the posts created by λ for each user at any timestamp t. Then, by calculation of $\sigma(v,t)$ using the proposed method, the k influential posts are determined for the user v. At each attempt, λ begins to uniformly create posts for users. Depending on the model of reposting, the users begin to randomly repost the posts (taking into account the probability of biased reposting), as was mentioned earlier.

4.2 Evaluation of Post Selection Method Using Simulation

To test the proposed method, 72 hours simulation was carried out. In each hour simulation for the first 10 hours, λ created 10 posts on average. In the first 10 hours, a total of 100 Twitter posts were examined. The social graph of the data collected contains 4.9 thousand users. A variety of implementations was done for various modes, to be evaluated as follows.

, to evaluate the increasing of information propagation in OC, 4 more methods, other than the proposed one, are also employed. 1) Latest post method: based on the time a post has created, the learners visit the latest k post. 2) Random k post selection method for a learner. 3) Post-post similarity method: The k posts with the highest similarity to the learner’s posts. 4) learner-Learner similarity method: k post selection method with the highest similarity of the posts authors with the learner.

As shown in Table (1), the results of reposting activities done according to different methods is given.

The Random refers to a method that randomly selects k posts from among candidate posts for display of the node v. Similarly, Recency (Rec) selects the recent k posts from candidate posts for display. These two basic methods in [8], [19] are used for comparison. As is clear from the graph, the social similarity method in the data set under study had the best performance compared to the similarity of users. It should be mentioned that the method of post-to-post similarity performed better than the method of individual's influence. As predicted, the two Random and Recency methods had the worst performance for k=5 (selection of five influential posts for each user).

Table 1 shows the results for the two modes k = 3 and k = 8. As specified in the table, Random method had no acceptable performance for k = 8. This may be due to low rate of post entry. This method cannot be efficient for high rates of posts. The results in k = 5 are roughly the same for k = 3.

Table 1. Repost action per thousand for two selection modes of 3 and 8 influential posts.

k=8	***k=3***	***Methods***
activity (1000=K)	***Activity (1000=K)***	
1.5	3	***Random***
3.6	2.3	***Recent***
6.8	2.9	***Post-Post Similarity***
6.9	2.7	***User-User Similarity***
8.2	4.1	***Proposed Method***

4.3 Repost Prediction Method Evaluation

The correlation of these features with repost behavior is studied in this paper. If the values of these features are significantly related to repost behavior, the relationship can be measured based on conventional learning. To this end, a method similar to Pearson's correlation method is employed. Since each feature has continuous values and the repost behavior is a binary variable (0 or 1), Pearson's method cannot be used. The point-biserial method is utilized for such problems. If the values of each feature of each post are a continuous variable (x), and repost behavior for the post is a binary variable (y), then the point-biserial correlation coefficient is based on formula (11), where M1 shows the mean feature value for posts, leading to repost behavior y=1. Similarly, M0 is the mean of features of posts, not leading to repost behavior y=0. Moreover, n1 shows the number of posts in the samples except for posts leading to repost behavior (y=1), and n0 denotes the number of posts in samples which do not result in repost behavior (y=0). In addition, n is the total number of samples examined, sn is the standard deviation for values of features of all of the studied post samples. This correlation coefficient varies between -1 and 1, where 1 shows maximum positive correlation between measured values and user behavior and -1 shows maximum negative correlation between measures values and user behavior. Zero (0) also shows independence of the features from user repost behavior.

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \tag{11}$$

Table (2) presents results of the coefficient of correlation between features and reposts behavior. These results are reflective of a positive correlation between these features and repost behavior. The maximum correlation belongs to post-post similarity.

Table 2. Coefficient of correlation between proposed features and repost behavior

	Post-Post Similarity	User-User Similarity
r_{pb}	0.63	0.54

Two data sets from the Twitter social media were used to assess the proposed method. The data was selected based on the following selection criteria: 1) All users should have more than 30 items in their list of followers and followee; 2) Users creating or retweeting at least 20 tweets a week. Each of the selected data sets has a one-month history of general tweets and meets the mentioned criteria. These two data sets were selected twice in two weeks in 2010 and 2011. These two data sets were selected because due to the content-based nature of the proposed method, it was tried to conduct assessments when the topics were not the same. Finally, after preprocessing procedures, a total of 23038 active users were identified in the two data sets, which contained 1211347 tweets and 92307 retweets. Table (3) presents overall specifications of the data sets collected for assessment.

Table 3. Data sets specifications

Data Set 1		***Data Set 2***	
Tweet	Retweet	Tweet	Retweet
787376	62191	423971	30116

Three well-known measures, namely "precision", "recall", and "F measure", are used to assess the predictions [12], [13]. These measures are used for prediction problems of the binary class. Formulas 12 to 14 show these measures.

$$Precision = \frac{|\{\text{Predicted RT}\} \cap \{\text{True RT}\}|}{|\{\text{Predicted RT}\}|} \tag{12}$$

$$\text{Recall} = \frac{|\{\text{Predicted RT}\} \cap \{\text{True RT}\}|}{|\{\text{True RT}\}|} \tag{13}$$

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

Using the conventional supervised classification method, the data sets were divided into the training and test groups. Table (4) presents specifications of the training and test groups of the assessment data sets.

Table 4. Specifications of test and training data sets

	Data Set 1		Data Set 2	
	Tweet	Retweet	Tweet	Retweet
Training	623340	49235	335643	23841
Test	164036	12956	88328	6275

For the purpose of binary classification the decision tree method, Naïve Bayes and proposed logistic regression method were used as the bases. Table (5) presents analysis results of the two data sets.

Table 5. Experiment results

	Data set 1			*Data Set 2*		
	Precision	***Recall***	***F measure***	***Precision***	***Recall***	***F measure***
Proposed Method	0.941	0.68	0.789	0.88	0.61	0.720
Naïve Bayes	0.92	0.65	0.763	0.867	0.51	0.702
Decision tree	0.79	0.54	0.746	0.832	0.46	0.593

5. Limitation & Conclusion

The most obvious limitation of this study was the absence of a specific standard in the implementation and design of OCs. Therefore, the suggested method is designed based on the content of posts to make it applicable in various kinds of OCs. Other factors such as external events (such as trends and news), context of the OC (such as health forums or technical forums), security concerns and OC design problems can be effective in users' repost behavior. This article seeks to solve the problem of selecting k influential posts for the user v from among the posts a user like u has created and the user v follows. These posts should be selected in such a way that the entire reposts increase in the network. On reviewing the literature in this field, the complexities of the issue were discussed. The existing methods have failed to be implemented online so far. Since this issue has been defined for knowledge management in online environments, these methods were not usable. In addition to online implementation, the proposed method is suitable for use in OCs. Certain features such as post-post similarity, user-user similarity, and geographic similarity are among major parameters taken into account in the selection of k influential posts. The proposed method focuses on the problem of online communities that is lurking users. An evaluation based on real data and different scenarios makes the proposed method more efficient compared to other methods.

References

[1] W. Guechtouli, J. Rouchier, and M. Orillard, "Structuring knowledge transfer from experts to newcomers," J. Knowl. Manag., vol. 17, no. 1, pp. 47–68, 2013.

[2] M. G. Wilson, J. N. Lavis, R. Travers, and S. B. Rourke, "Community-based knowledge transfer and exchange: Helping community-based organizations link research to action," Implement. Sci., vol. 5, no. 1, p. 33, 2010.

[3] C. K. K. Chan and Y. Y. Chan, "Students' views of collaboration and online participation in Knowledge Forum," Comput. Educ., vol. 57, no. 1, pp. 1445–1457, 2011.

[4] B. Nonnecke and J. Preece, "Lurker demographics: Counting the silent," Proc. SIGCHI Conf. …, vol. 2, no. 1, pp. 1–8, 2000.

[5] B. Nonnecke and J. Preece, "Why lurkers lurk," AMCIS 2001 Proc., pp. 1–10, 2001.

[6] P. G. Kilner and C. M. Hoadley, "Anonymity options and professional participation in an online community of practice," Proc. 2005 Conf. Comput. Support Collab. Learn., pp. 272–280, 2005.

[7] Y. Amichai-Hamburger, T. Gazit, J. Bar-Ilan, O. Perez, N. Aharony, J. Bronstein, and T. Sarah Dyne, "Psychological factors behind the lack of participation in online discussions," Comput. Human Behav., vol. 55, pp. 268–277, 2016.

[8] P. Sloep and L. Kester, "From lurker to active participant," in Learning Network Services For Professional Development, 2009, pp. 17–25.

[9] M. Takahashi, M. Fujimoto, and N. Yamasaki, "The active lurker: influence of an in-house online community on its outside environment," Group, pp. 1–10, 2003.

[10] P. Kollock and M. Smith, "Managing the Virtual Commons: Cooperation and Conflict in Computer Communities," Comput. Commun. Linguist. Soc. Cross-Cultural Perspect., pp. 109–128, 1996.

[11] A. Schneider, G. Von Krogh, and P. J??ger, "What's coming next? Epistemic curiosity and lurking behavior in online communities," Comput. Human Behav., vol. 29, no. 1, pp. 293–303, 2013.

[12] H. Zhang, Q. Zhao, H. Liu, K. Xiao, J. He, X. Du, and H. Chen, "Predicting retweet behavior in Weibo social network," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012, vol. 7651 LNCS, pp. 737–743.

[13] X. Tang, Q. Miao, Y. Quan, J. Tang, and K. Deng, "Predicting individual retweet behavior by user similarity: A multi-task learning approach," Knowledge-Based Syst., vol. 89, pp. 681–688, 2015.

[14] J. Leskovec, L. Adamic, and B. Huberman, "The Dynamics of Viral Marketing," ACM Trans. Web, vol. 1, no. 1, pp. 1–39, 2007.

[15] E. Sun, I. Rosenn, C. a Marlow, and T. M. Lento, "Gesundheit! Modeling Contagion through Facebook News Feed Mechanics of Facebook Page Diffusion," Proc. Third Int. ICWSM Conf., no. 2000, pp. 146–153, 2009.

[16] E. Bakshy, B. Karrer, and L. A. Adamic, "Social Influence and the Diffusion of User Created Content," in Electronic Commerce, 2009, pp. 325–334.

[17] P. Domingos and M. Richardson, "Mining the Network Value of Customers," in Proceedings of the Seventh {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining, 2001, pp. 57–66.

[18] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. KDD 02, vol. 02, no. 3, p. 61, 2002.

[19] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03, 2003, p. 137.

[20] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," Proc. VLDB Endow., vol. 5, pp. 73–84, 2011.

[21] U. Feige, V. S. Mirrokni, and J. Vondrák, "Maximizing non-monotone submodular functions," in Proceedings -

Annual IEEE Symposium on Foundations of Computer Science, FOCS, 2007, pp. 461–471.
[22] S. Stieglitz and L. Dang-Xuan, "Emotions and Information Diffusion in Social Media - Sentiment of Microblogs and Sharing Behavior," J. Manag. Inf. Syst., vol. 29, no. 4, p. 217, 2013.
[23] S. Ye and S. F. Wu, "Measuring message propagation and social influence on Twitter.com," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6430 LNCS, pp. 216–231.
[24] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6913 LNAI, pp. 18–33.
[25] M. Magnani, D. Montesi, and L. Rossi, "Information Propagation Analysis in a Social Network Site," 2010 Int. Conf. Adv. Soc. Networks Anal. Min., pp. 296–300, 2010.
[26] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Effects of user similarity in social media," in Proceedings of the fifth ACM international conference on Web search and data mining WSDM 12, 2012, p. 703.
[27] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," Proc. third ACM Int. Conf. Web search data Min. - WSDM '10, p. 241, 2010.
[28] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," The Journal of Mathematical Sociology, vol. 15. pp. 193–206, 1990.
[29] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social Influence Analysis in Large-scale Networks," in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 807–816.
[30] M. Nagarajan, H. Purohit, and A. Sheth, "A Qualitative Examination of Topical Tweet and Retweet Practices," Artif. Intell., pp. 295–298, 2010.
[31] C. G. Knight and L. K. Kaye, "'To tweet or not to tweet?' A comparison of academics' and students' usage of Twitter in academic contexts," Innov. Educ. Teach. Int., no. April 2015, pp. 1–11, 2014.
[32] C. Haeussler, "Information-sharing in academia and the industry: A comparative study," Res. Policy, vol. 40, no. 1, pp. 105–122, 2011.
[33] W. S. Hwang, S. W. Kim, D. H. Bae, and Y. J. Do, "Post ranking algorithms in blog environment," in Proceedings of the 2008 2nd International Conference on Future Generation Communication and Networking, FGCN 2008, 2008, vol. 2, pp. 64–67.

Omid Reza Bolouki Speily received the B.Sc. degree in Computer Engineering from Urmia University, the M.Sc. & Ph.D. degrees in Information Technology from the AmirKabir University of Technology. He worked as a researcher at the Iran Telecommunication Research Center (ITRC). Since 2009 he joined the Urmia University of Technology as a faculty member of Information Technology & Computer Engineering Department. His research interest includes the dynamics complex networks, graph theory, intelligent system, e-Services.