To Follow or Not to Follow: A Feature Evaluation

Yanan Zhu
Information Retrieval Lab
Department of Computer Science
Georgetown University
yanan@ir.cs.georgetown.edu

Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
nazli@ir.cs.georgetown.edu

ABSTRACT

The features available in Twitter provide meaningful information that can be harvested to provide a ranked list of followees to each user. We hypothesize that retweet and mention features can be further enriched by incorporating both temporal and additional/indirect links from within user's community. Our empirical results provide insights into the effectiveness of each feature, and evaluate our proposed similarity measures in ranking the followees. Utilizing temporal information and indirect links improves the effectiveness of retweet and mention features in terms of nDCG.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval– *Selection process*

General Terms

Algorithms, Experimentation

Keywords

Twitter; User Recommendation; Retweet; Mention; Temporal Ranking; Personalization; Social Media

1. INTRODUCTION

One of the challenges for the users of social media, such as in Twitter, is the fast growing number of people each user is following. We evaluate features that refine this list of followees by ranking them based on the similarity of the followees to the user. Existing research that ranks information in social media, in particular in Twitter, has focused on ranking the tweets, i.e., the content [1][3], and ranking the users [2][4] globally. We are interested to personalize the ranking of users as it pertains to the community of a given user. We evaluate Twitter features of *retweet* and *mention* as ranking functions; we also introduce two new features, namely *indirect retweet*, and *indirect mention (section 2)*. We further utilize the *temporal* aspect to enhance the performance of features. We provide a comparison in respect to their quality in ranking a user's list of followees.

2. FEATURES & SIMILARITY SCORES

We now define each feature and its corresponding similarity score function.

Temporal Score: We assume that more recently a tweet is published, more important is the tweet. With this premise, we hypothesize that the time associated to a tweet, i.e., published time, affects the similarity score, in measuring the similarity of a user to his/her followees. The interest of a person may change during the time; thus, the importance of the tweet is associated with time. We assign a *temporal score* to each tweet, which is defined as,

$$Temporal(tw_k) = a(t_{cur} - t_{min}) + \frac{1-at_I^2-b}{t_I}(t_{cur} - t_{min}) + \mathrm{b} \qquad (eq.\ 1)$$

where time interval is defined as $t_I = t_{max} - t_{min}$; t_{max} is the time of the most recent tweet and t_{min} is the time of earliest tweet of the user who publishes the tweet tw_k; t_{cur} is the current time of tweet tw_k; a and b are experimental parameters, where $a<0$, $0 \leq b < 1$.

[Direct] Retweet Score (DR): We define a *retweet score* to capture similarity between a target user and his/her followees. This score considers the retweets stemmed from the followees of a user. That is, the ratio of the number of tweets retweeted from a followee u_j by the target user u_i to the number of all the tweets retweeted by the target user u_i. It is defined as,

$$ret(u_i, u_j) = \frac{|retweet_{u_i}(u_j)|}{|retweet_{u_i}(u_{all})|} \qquad (eq.\ 2)$$

where $\mathrm{retweet}_{u_i}(u_j)$ is a function that returns the set of tweets that retweeted by user u_i from user u_j, and $\mathrm{retweet}_{u_i}(u_{all})$ returns the entire set of tweets that retweeted by user u_i from all the users in the community of user u_i.

Indirect Retweet Score (IR): To capture the hidden potential connection/similarity between a target user u_i and his/her followee u_j, we further consider an *indirect retweet* score. That is, we consider the retweets of the other followees in the community of a target user $u_i$that have retweeted the same tweet as the target user has retweeted. We define the *indirect retweet* score as a function of both *[direct] retweet* score (first term in equation 3), and summation of scores based on *indirect retweet* (second term in *eq. 3*).

$$IR(u_i, u_j) = t * ret(u_i, u_j) + (1-t)\sum_{tw_k \in ret(u_j)} \frac{ret(u_i, user(tw_k))}{|tw_k|} \qquad (eq.\ 3)$$

In (*eq. 3*), $t<1$, $\mathrm{ret}(\mathrm{u_i}, \mathrm{u_j})$ is the *[direct] retweet* score between user u_i and u_j defined in (*eq. 2*); $\mathrm{tw_k}$ is all tweets that are retweeted from user u_j and $\mathrm{ret}(\mathrm{u_i}, \mathrm{user}(\mathrm{tw_k}))$ returns the *[direct] retweet* score between user u_i and publisher of $\mathrm{tw_k}$.

Indirect Retweet with Temporal Score (IRT): To evaluate the effect of temporal feature, we incorporate the *temporal* score into *indirect Retweet (IR)* score. We call this IRT score and define as,

$$IRT(u_i, u_j) = t * ret(u_i, u_j) + (1-t) *$$

$$\sum_{tw_k \in ret(u_j)} Temporal(tw_k) \frac{ret(u_i, user(tw_k))}{|tw_k|} \qquad (eq.\ 4)$$

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

where the first term is the *[direct] retweet* score and the second term is the *indirect retweet* score which is associating each tweet tw_k with the *temporal score* $Temporal(tw_k)$.

[Direct] Mention Score (DM): Similar to the retweet feature, we define *mention* score. We assume that a target user shares similar interest with a followee if s[he] mentions that followee by using mention symbol (@). For user u_i, *mention* score for user u_j is:

$$men(u_i, u_j) = \frac{|mention_{u_i}(u_j)|}{|mention_{u_i}(u_{all})|} \qquad (eq.5)$$

The $mention_{u_i}(u_j)$ returns the set of tweets that mention user u_j by user u_i, and $mention_{u_i}(u_{all})$ returns the entire set of tweets that mention any users by user u_i.

Indirect Mention Score (IM): Similarly, we take *indirect mention* into consideration by the same intuition that users who mention a same person may indeed have similarity.

$$IM(u_i, u_j) = p * men(u_i, u_j) + (1-p) \sum_{tw_k \in men(u_j)} \frac{men(u_i, user(tw_k))}{|tw_k|} \qquad (eq.6)$$

The first term is the *[direct] mention* score and the second one is the summation of scores based on *indirect mention*, defined analogous to *indirect retweet*, where $men(u_i, user(tw_k))$ is the *[direct] mention score* between u_i and publisher of tw_k.

Indirect Mention with Temporal Score (IMT): By incorporating temporal feature to IM score, we define IMT as:

$$IMT(u_i, u_j) = p * men(u_i, u_j) + (1-p) * \sum_{tw_k \in men(u_j)} Temporal(tw_k) \frac{men(u_i, user(tw_k))}{|tw_k|} \qquad (eq.7)$$

Once again, and analogous to IRT, first term is *[direct] mention* score, and the second term is the *indirect mention* with *temporal* score.

3. EVALUATION

Dataset: There is no benchmark dataset with user-specific relevant judgment for evaluating personalized recommendation research in micro-blogging platform. Not only knowing the ground truth of user's interest is a difficult task due to the privacy concerns but also a crawled data set may not provide a good coverage of all the potential scenarios. For these reasons, we created our data set by simulating communities of 50-100 followees for 10 target users. The user interests are known apriori (10-30 topics of interest), based on which we create communities of followees. The probability of two users sharing the same interest is set from 0.2 to 0.8. To gain a better coverage, we control the probability of following back (0.2-1) to produce different type of social structures (Data available upon request).

Ground Truth: We take advantage of users' *interests* to define the $realSimilarity(u_i, u_j)$ between users as,

$$realSimilarity(u_i, u_j) = \frac{|commonInterest(u_i, u_j)|^2}{|interest(u_i)| * |interest(u_j)|} \qquad (eq.\ 8)$$

where the $commonInterest(u_i, u_j)$ returns the set of interests user u_i and u_j share; $interest(u_i)$ returns the set of interests of user u_i.

According to the $realSimilarity(u_i, u_j)$, we then produce a ranked list, served as a ground truth. We assign the relevant score to each followee of the target user in community according to the followee's ranked position in this ranked list as:

$$rel(u_i) = n + 1 - position(u_i) \qquad (eq.\ 9)$$

We use commonly used nDCG metrics where IDCG is the DCG score of ideal situation when ranked list is sorted by relevance.

Results: We evaluate the effectiveness of *retweet* and *mention* features with and without the indirect link and temporal aspect on ranking the followees, in terms of nDCG score. Figure 1 illustrates nDCG of methods based on *[direct] retweet* (DR), *indirect retweet* (IR) *and indirect retweet with temporal* score (IRT) across 10 communities of followees. *DR* method is improved by 3.5% using IR method (0.88 vs. 0.85), and is further improved by 7.1% using IRT method (0.91 vs. 0.85). Similarly, figure 2 shows the comparison among the *[direct] mention* (DM), *indirect mention* (IM) and *indirect mention with temporal* aspect (IMT). IM method improves DM by 3.6% (0.84 vs. 0.81). Temporal aspect further improves DM by 7.4% using IMT (0.87 vs.0.81). Figure 3 illustrates the average nDCG scores and standard deviations over all 10 tested communities. The results indicate the following ordering among methods: IRT > IR > IMT > IM > DR > DM. Considering standard deviation, IRT (std: 0.02) is shown to be the most stable method, indicating its excellent performance is relatively stable across all the communities.

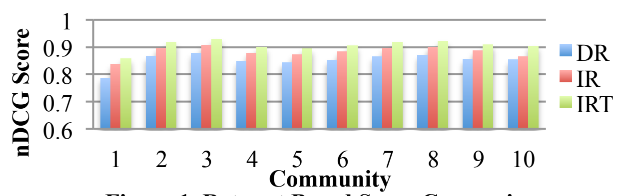

Figure 1. Retweet Based Score Comparison

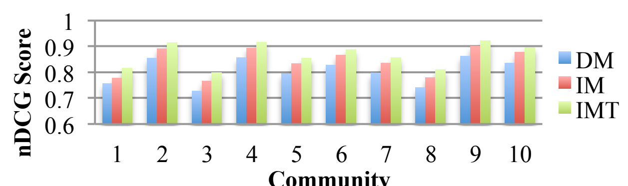

Figure 2. Mention Based Score Comparison

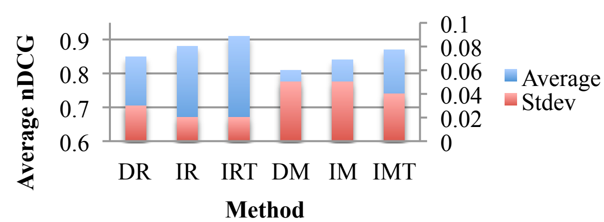

Figure 3. Average nDCG and Standard Deviation

4. REFERENCES

[1] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. Measuring User Influence in Twitter: The Million Follower Fallacy. *4th AAAI Conf. on Weblogs and Social Media*, 2011.

[2] Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H. An empirical study on learning to rank of tweets. *23rd ACL Conf. on Computational Linguistics (COLING)*, 2010.

[3] Kwak,H.,Lee, C., Park, H., Moon,S. What is Twitter, a Social Network or a News Media? *19th World Wide Web Conf. (WWW),* 2010.

[4] Welch,M., He, D., Schonfeld ,U.,and Cho, Junghoo.Topical Semantics of Twitter Links. *4th ACM Conf. on Web search and Data Mining (WSDM),* 2011.