Discovery of Information Disseminators and Receptors on Online Social Media

Munmun De Choudhury
School of Computing, Informatics & Decision Systems Engineering
Arizona State University, Tempe
munmun@asu.edu

ABSTRACT

Today, there is significant sharing of information artifacts among users on various social media sites, including Digg, Twitter and Flickr. An interesting consequence of such rich and extensive social interaction is the evolving nature of "roles" that are acquired by users over time, in the context of variegated communication activities, such as commenting, replying, uploading a media artifact and so on. In this paper, we investigate the discovery of two roles that define information dissipation: disseminators and receptors. We propose a computational framework based on factorization of stacked representation of activities and test the outcomes on a dataset from Digg. Experiments show that our approach can, interestingly, reveal correlations with user activities occurring at a future point in time.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Algorithms, Experimentation

Keywords

Information diffusion, information disseminators, social network analysis, social media, Digg.

1. INTRODUCTION

Classically, human communication activity involves mutual exchange of information, and the pretext of any social interaction among a set of individuals is a reflection of how our behavior, actions and knowledge can be modified, refined, shared or amplified based on the information that flows from one individual to another. Thus, over several decades, the structure of social groups, society in general and the relationships among individuals in these societies have been shaped to a great extent by the flow of information in them [1]. A key aspect of understanding such flow of information in a network is therefore to understand the roles acquired by individuals with respect to dissipation and consumption of information over time.

Today, the pervasive use of online social media has made the cost involved in propagating a piece of information to a large audience extremely negligible, providing extensive evidences of large-scale information dispersion [3, 2]. There are multifaceted personal publishing modalities available to users today, where such large scale information contagion is prevalent: such as weblogs, social networking sites like MySpace and Facebook as well as microblogging tools such as Twitter. In this paper, we develop computational methods to understand user activities via roles such as information disseminators and receptors in the context of the social media Digg (http://www.digg.com/).

2. COMPUTATIONAL APPROACH

In this section we present the mathematical model of identifying information disseminators and receptors in a multidimensional activity based social network. Suppose for every user in a set $\{u_i\}$ a set of stories $S_{1,i}, S_{2,i}, \cdots, S_{k,i}$ uploaded by her from time slices t_1 through t_k on a certain topic θ. We need to determine a subset of users in $\{u_i\}$ who are information disseminators and a subset of users who are information receptors.

Our main idea is that information disseminators and receptors can be quantified by two measures respectively: distribution and consumption of information on an emergent theme in the social network, over the course of a long period of time. Note, distribution and consumption properties are dual of each other. We define distribution of information by an information disseminator to be a measure of her degree of triggering another user to engage in social actions, communication on the stories and / or user-user replies in the social network, or generate similar information content in the future. Consumption of information by an information receptor is the degree to which a user accepts the novel information distributed by the disseminator in some form—social actions, communication on the stories and / or user-user replies, or generation of similar content. Distribution and consumption of information due to information disseminators and receptors should have two important characteristics: *reachability* to a wide audience and *persistence* i.e. temporal permanence over time.

Stacked Representation. We construct a stacked graph representation of reachability and persistence properties over the stories introducing information content in the social network: the stories $S_{1,i}, S_{2,i}, \cdots, S_{k,i}$. A stacked graph is a set of graphs capturing how much reachable and temporally persistent the novel stories are to different users in the network. We develop four stacked graphs at time slice t_{k+1} where the graphs correspond to several user activities over time. A graph $G_X(V, E_X; t_{k+1})$ in the stack contains edges over the V users based on their activity X. Note, the user activity X could be: actions (A), comments (C), replies (R) or uploading similar information U. Hence we formally define the graph $G_X(V, E_X; t_{k+1})$ as follows:

HT'10, June 13–16, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0041-4/10/06.

1. A set of nodes V, corresponding to the users $\{u_i\}$.
2. A set of edges E_X, where a directed edge $e_X(u_m, u_p; t_{k+1})$ exists if u_p followed u_m in performing the activity X or if u_p followed u_m in uploading a similar story in the past over t_1 to t_k. Each edge $e_X(u_m, u_p; t_{k+1})$ is also associated with a weight $\omega_X(u_m, u_p; t_{k+1})$ which is given by the ratio of the total number of instances when u_p followed u_m in performing the activity X / uploading similar stories, to their individual number of instances of performing X / uploading similar stories. Let us represent the weighted adjacency matrix corresponding to each of the four graphs as $\mathbf{W}_A(t_{k+1})$, $\mathbf{W}_C(t_{k+1})$, $\mathbf{W}_R(t_{k+1})$ and $\mathbf{W}_U(t_{k+1}) \in \mathbb{R}_+^{|V|\times|V|}$.

Once we construct all such graphs over activities A, C, R and U, that is, $G_A(V, E_A; t_{k+1})$, $G_C(V, E_C; t_{k+1})$, $G_R(V, E_R; t_{k+1})$ and $G_U(V, E_U; t_{k+1})$, we are now interested in determining the information disseminators and receptors from these four stacked graphs at time t_{k+1}.

Factorization of Reachability and Persistence. We conjecture that the emergent roles are an artifact of their measures of information distribution and consumption respectively. Hence each of the observed stacked graphs corresponding to reachability and persistence, that is, the various user activities like social actions, communication on the stories and / or user-user replies, or generation of similar content $G_A(V, E_A; t_{k+1})$, $G_C(V, E_C; t_{k+1})$, $G_R(V, E_R; t_{k+1})$ and $G_U(V, E_U; t_{k+1})$, would be a representation of the measures of distribution and consumption of the information disseminators and receptors respectively.

We therefore define a non-negative matrix factorization based framework to determine the measures from the stacked graphs of user activity that incorporate reachability and persistence. Given the weighted adjacency matrices of the stacked graphs as $\mathbf{W}_A(t_{k+1})$, $\mathbf{W}_C(t_{k+1})$, $\mathbf{W}_R(t_{k+1})$ and $\mathbf{W}_U(t_{k+1}) \in \mathbb{R}_+^{|V|\times|V|}$, we factorize each of them into a column vector of measures of information distribution by the disseminators $\mathbf{D}_A(t_{k+1})$, $\mathbf{D}_C(t_{k+1})$, $\mathbf{D}_R(t_{k+1})$ and $\mathbf{D}_U(t_{k+1}) \in \mathbb{R}_+^{|V|\times 1}$, and a row vector of measures of information consumption by the receptors $\mathbf{H}_A(t_{k+1})$, $\mathbf{H}_C(t_{k+1})$, $\mathbf{H}_R(t_{k+1})$ and $\mathbf{H}_U(t_{k+1}) \in \mathbb{R}_+^{1\times|V|}$.

For the purpose, we develop an optimization scheme based on an additive objective function over the distribution and consumption measures of all users for each type of activity, $\mathbf{D}_A(t_{k+1})$, $\mathbf{D}_C(t_{k+1})$, $\mathbf{D}_R(t_{k+1})$ and $\mathbf{D}_U(t_{k+1})$ for the disseminators, and $\mathbf{H}_A(t_{k+1})$, $\mathbf{H}_C(t_{k+1})$, $\mathbf{D}_H(t_{k+1})$ and $\mathbf{D}_H(t_{k+1})$ for the receptors. The objective is to maximize the measures of both the measures over each type of user activity, subject to appropriate weights for each type. Finally, the users in $\{u_i\}$ with high measures corresponding to the optimal weights would be the information disseminators and receptors at that time slice, t_{k+1}.

3. EXPERIMENTAL STUDIES

We present some quantitative experimental studies in this section on a dataset crawled from the popular news-sharing social media Digg. We seeded our crawling from the stories in the featured category 'Popular' on the Digg website. We crawled all stories in this list and which submitted over August and September 2008 and used a snowballing technique to expand the social graph. In total, this dataset comprises 187,277 stories, 7,622,678 diggs, 687,616 comments and 477,320 replies over a set of 51 topics in this time range.

The observed user activities in a social network, like, social actions, comments, replies and uploading news stories are likely to be the effect of their mutual interactions. Such interactions we conjecture are likely to be an effect of the phenomenon of information distribution and consumption due to a handful of disseminators and receptors. Hence to validate the two kinds of roles identified in this paper, we use the measures of distribution and reception of information by each of them at a time slice t_{k+1}, to find correlations with the corresponding user activities at that time slice.

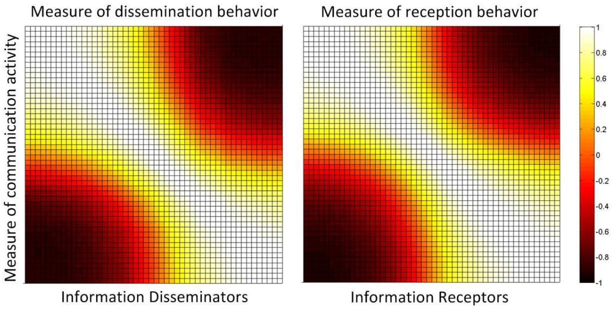

Figure 1: Correlation coefficient between measure of information dissemination and reception, and measure of communication activity of users in Digg dataset. Results are averaged over the four types of user activities discussed here: actions (A), comments (C), replies (R) or uploading similar information U.

The experimental results conducted on the dissemination / reception and communication activity correlations are shown in Figure 1 over a period of two months (August and September 2008). We observe that there is significant correlation between the two measures; implying that our method of discovery of information disseminators and receptors, quantitatively, is meaningful in capturing the temporal dynamics of social actions.

4. CONCLUSIONS

In this paper, we have proposed a computational framework to discover information disseminators and receptors in multi-dimensional activity based social media. Our method formulated information dissipation and consumption as dual of each other and used a non-negative matrix factorization based framework to determine scalar measures of dissipation and consumption for each user in the network, over actions, comments, replies or uploading similar information. Experiments on a large Digg.com dataset show that our discovered information roles are able to yield high correlation with communication activities of the users over time.

5. REFERENCES

[1] Scott L. Feld. The focused organization of social ties. *American Journal of Sociology*, 86(5):1015–1035, 1981.

[2] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501, New York, NY, USA, 2004. ACM.

[3] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03*, pages 137–146, 2003.