Mining Directed Social Network from Message Board

Naohiro Matsumura
Osaka University
Osaka, 560-0043 Japan
+81-6-6850-5231
matumura@econ.osaka-u.ac.jp

David E. Goldberg
UIUC
Urbana, IL 68801
+1-217-333-2346
deg@uiuc.edu

Xavier Llorà
UIUC
Urbana, IL 68801
+1-217-333-2346
xllora@uiuc.edu

ABSTRACT

In the paper, we present an approach to mining a directed social network from a message board on the Internet where vertices denote individuals and directed links denote the flow of influence. The influence is measured based on propagating terms among individuals via messages. The distance with respect to contextual similarity between individuals is acquired since the influence indicates the degree of their shared interest represented as terms.

Categories and Subject Descriptors

C.2 [**Computer Systems Organization**]: Computer - communication networks

General Terms

Algorithms

Keywords

Directed social network, Internet message board

1. INTRODUCTION

With the advent of popular social networking services on the Internet such as Friendster [1] and Orkut [2], social networks are again coming into the limelight with respect to this new communication platform. A social network shows the relationships between individuals in a group or organization where we can observe their social activities. In this paper, we propose a method of mining a directed social network from a message board on the Internet.

2. MINING DIRECTED SOCIAL NETWORKS

In a social network based upon online communication, the distance between individuals does not mean 'geographical distance' because each person lives in a virtual world. Instead, distance can be considered 'psychological distance' and this can be measured by the "influence" wielded among the members of the network.

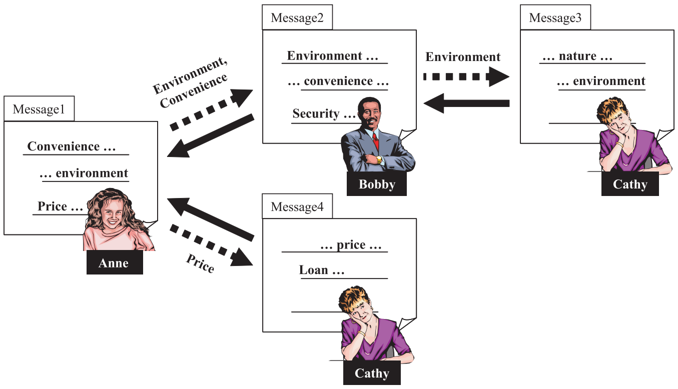

Figure 1: A message chain of four messages sent by three individuals.

The basic idea of measuring influence comes from the IDM (Influence Diffusion Model) [1] in which the influence between a pair of individuals is defined as the sum of propagating terms among them via messages. Our approach simplifies the algorithms of the IDM to make it more intuitively reasonable. Here, let a message chain be a series of messages connected by post-reply relationships, and the influence of a message x on a message y (x precedes y) in the same message chain be $i_{x \to y}$. Then, $i_{x \to y}$ is defined as

$$i_{x \to y} = |w_x \cap \cdots \cap w_y|, \qquad (1)$$

where w_x and w_y are the set of terms in x and y, respectively, and $|w_x \cap \cdots \cap w_y|$ is the number of terms propagating from x to y via other messages. If x and y are not in the same message chain, we define $i_{x \to y}$ as 0 because the terms in x and y are used in a different context and there is no influence between them.

Based on the influence between messages, we next measure the influence of an individual p on an individual q as the total influence of p's messages on other's messages through q's messages replying to p's messages. Let the set of p's messages be α, the set of q's messages replying to any of α be β, and the message chains starting from a message z be ξ_z. The influence from p onto q, $j_{p \to q}$, is then defined as

$$j_{p \to q} = \sum_{x \in \alpha} \sum_{z \in \beta} \sum_{y \in \xi_z} i_{x \to y}. \qquad (2)$$

Here we see the influence of p on q as q's contribution toward the spread of p's messages. The influence of each individual is also measurable using $j_{p \to q}$. Let the influence

[1] http://www.friendster.com/
[2] http://www.orkut.com/

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

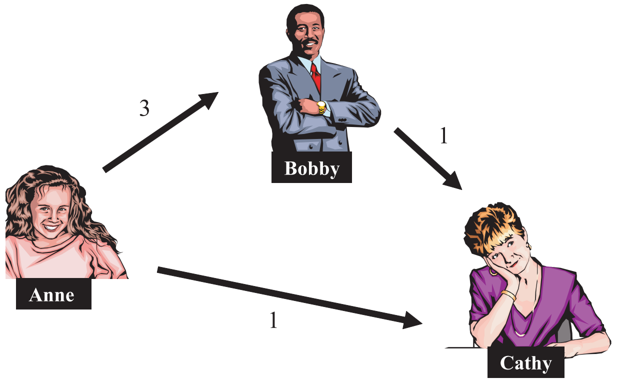

Figure 2: A directed social network showing the influence from Figure 1.

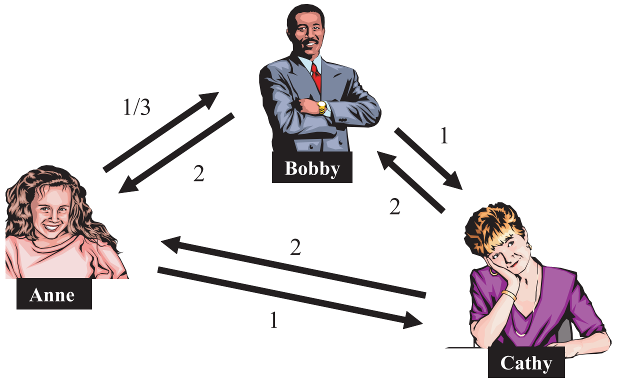

Figure 3: A directed social network showing the distance from Figure 1.

of p be k_p, and all other individuals be γ. Then, k_p is defined as

$$k_p = \sum_{q \in \gamma} j_{p \to q}. \tag{3}$$

As an example of measuring the influence, let us use the simple message chain shown in Figure 1 where Anne posted Message 1, Bobby posted Message 2 as a reply to Message 1, and Cathy posted Message 3 and Message 4 as replies to Message 2 and Message 1, respectively. In the figure, solid arrows show the replies to previous messages, and dotted arrows show the flows of influence. Here, the influence between a pair of individuals is as follows.

- The influence of Anne on Bobby is 3 (i.e., $j_{Anne \to Bobby} = 3$), because two terms ("Environment" and "Convenience") were propagated from Anne to Bobby, and one term ("Environment") was propagated from Anne to Cathy via Bobby.
- The influence of Anne on Cathy is 1 (i.e., $i_{Anne \to Cathy} = 1$), because one term ("Price") was propagated from Anne to Cathy.
- The influence of Bobby on Cathy is 1 (i.e., $i_{Bobby \to Cathy} = 1$), because one term ("Environment") was propagated from Bobby to Cathy.
- The influence of Bobby on Anne and of Cathy on Anne is 0 (i.e., $i_{Bobby \to Anne} = 0$ and $i_{Cathy \to Anne} = 0$), because no term was propagated to Anne from either Bobby or Cathy.

Note that we ignore the influence of Anne on Cathy, even though a term "Environment" was propagated from Anne to Cathy via Bobby, because we want to measure direct influence between individuals. Instead, we consider the indirect influence of Anne on Cathy via Bobby as the contribution of Bobby, and add it to the influence of Anne on Bobby.

By mapping the influence between individuals, we can obtain a social network showing influence as in Figure 2 where their relationships are shown as directional links and the influence between them.

The influence between individuals also shows the distance between them with respect to contextual similarity since the influence indicates the degree of their shared interest represented as terms. The influence and contextual distance between individuals are inversely related; i.e., the greater the influence, the shorter the distance. Here, let us define the length of a link (i.e., distance) as follows.

Definition 1. The distance from an individual p to an individual q, $d_{p \to q}$, is defined as the value inversely proportionate to the influence from p to q; i.e., $d_{p \to q} = 1/j_{p \to q}$.

The distance is between 0 and 1 when the influence is more than 0. However, the distance cannot be measured by the above definition if the influence is 0. In that case, we define the distance $n - 1$ (n is the number of individuals participating in communication) as the case of the weakest relationships; i.e., the diameter of a social network where all individuals are connected linearly with maximum distance. In this way, a social network with distance is extracted from message chains as shown in Figure 3.

Based on the forward and backward shortest distances between all the pair of individuals in a social network, we can define "communication gaps" as an indicator to understand the state of communication. I revealed the existence of three types of communication, i.e., interactive communication, distributed communication, and soapbox communication, by examing 3,000 social networks. Unfortunately, we skip the details of communication gaps in this paper because of the space limitation. Please see [2] for more details.

3. CONCLUSION

Human beings are social creatures and we could not survive without cooperating with others. Therefore, understanding how relationships are created and function is essential to make our lives happier and richer. we hope this study will contribute to the realization of a better way of life through social network research.

4. ACKNOWLEDGEMENTS

We thank the Chance Discovery Consortium for providing financial support for the stay in IlliGAL at UIUC.

5. REFERENCES

[1] N. Matsumura. Topic Diffusion in a Community. *Chance Discovery*, pages 84–97. Springer Verlag, 2003.

[2] N. Matsumura, D. E. Goldberg, and X. Xlorà. Mining social networks from message boards. IlliGAL Report No. 2005001, 2005.