Social Media Information Diffusion with ProfileRank

Shenghan Yu
Stu. ID. 515030910562

Abstract

With the development of Internet, there have been more and more algorithms on the rank of webpages. Focusing on the reference relationship proves to be an effective way to help people find relative important webpages, like PageRank and HIPS algorithms. In the flied of social media, ProfileRank uses similar methods to deal with information diffusion by identifying influential users and relevant content.

1 Introduction

Nowadays, social media is taking over traditional media. Social media has a far higher speed to create information, and this feature is very similar to Internet. Therefore, how to identify useful information becomes a critical problem. It is an important step towards the design of information diffusion mechanisms, recommendation system, and viral marketing campaigns.

Since the information diffusion in the social media is similar to the one on the internet. We can try to use similar algorithm to deal with the data. On the other way, users have a more important level in the social media. Every user can create and spread content, and for each time, there is always one piece of content for one user creating or spreading. Therefore, the key factors in social media information diffusion are influential users and relevant content.

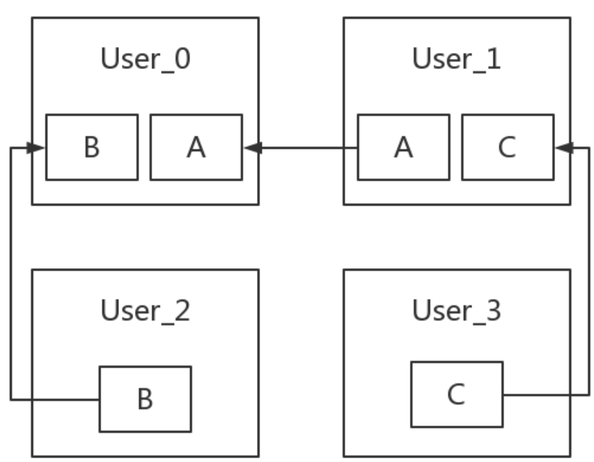

Figure 1: A typical content relationship graph in Twitter.

2 Methodology

ProfileRank considers both influential users and relevant content so that there are two relation matrixes which can be extracted from the original data, a user-content matrix M and a content-user matrix L. There is also a bipartite graph$G(U, C, F, E)$ from the original data, where U is the user set, C is the content set, and E and F are sets of edges that related users to content and the other way around. For Figure 1, the bipartite graph$G$ can be shown as follow:

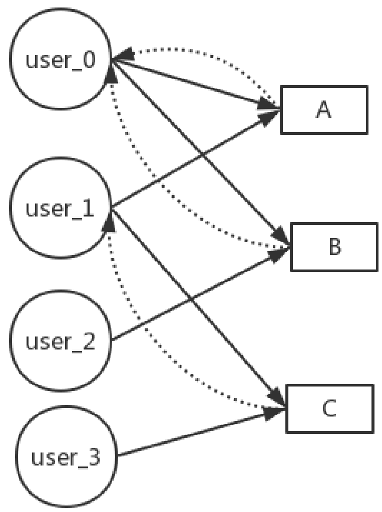

Figure 2: The bipartite graph$G$ for Figure 1.

The user-content matrix $M = (m_{i,j})$ is a $|U| \times |C|$ matrix where $m_{i,j} = \frac{1}{q_i}$ and q_i is the number of pieces of content the users u_i has created or propagated. The content-user matrix $L = (l_{i,j})$ is a $|C| \times |U|$ matrix where $l_{i,j} = 1$ if the user u_j created the piece of content c_i and $l_{i,j} = 0$, otherwise. For the data from Figure 1, the matrix M and L can be written as:

$$M = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

At the same time, to make use of these two matrixes, there is a principle: *A piece of content c is relevant to a user u if it is created and propagated by users that are influential to u and a user v is influential to u if v creates content that is relevant to u.* We can get the relationship between content r and users i with this principle.

$$r = iM$$
$$i = rL$$

Like PageRank, there are two possible issues: dangling users and buckets. A dangling usesr never propagates content from other users. We create an edge (u, c) from every dangling user to a ghost piece of content c and add an edge (c, u) from the ghost content to all the users. In Figure 1, $user_0$ is the dangling user so that the Figure 2 will become the graph shown as follow:

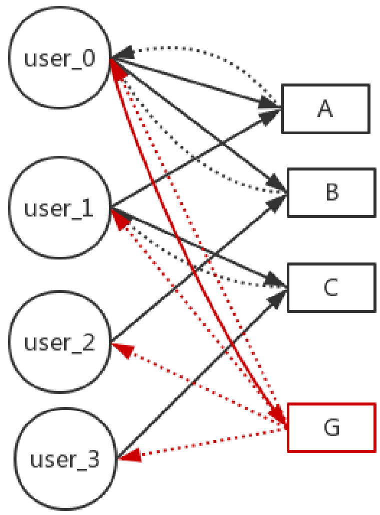

Figure 3: The bipartite graph$G$ handling dangling users.

A bucket is a strongly connected subgraph of the bipartite graph. By defining a damping factor d, the random tweeter will teleport from current to a random profile. Therefore, the definition of r and i can be modified as:

$$r = (1 - d)u(I - dLM)^{-1}$$
$$i = (1 - d)u(I - dML)^{-1}$$

ProfileRank computes user influence and content relevance based on a user-content bipartite directed graph, instead of the directed graph employed by PageRank and HITS. ProfileRank's graph is represented by two matrices, a user-content M and a content-user L matrix, and enables the computation of different score functions, influence and relevance, for different types of nodes, users and content,based on diffusion data.

3 Conclusion and Future Work

ProfileRank is an effective information diffusion model that measures content relevance and user influence in the bipartite graph. This model takes the unique features in the social media into consideration, which is also similar to the algorithms for webpages.

The key factors for ProfileRank are the two relation matrixes. However, the definition of two matrixes, especially for the user-content matrix, is not very clear, so that there might be a few problems in the calculation formula. I think that the user-content matrix may be rewrite as:

$M = (m_{i,j})$ is a $|U| \times |C|$ matrix where $m_{i,j} = \frac{1}{q_i}$ and q_i is the number of pieces of content the users u_i has created or propagated if the user u_i created or propagated the piece of content c_j, and $q_{i,j} = 0$, otherwise. In this way, the relationship between user and content can be defined more clearly, and the following formula can also be rewrote as the future work.

4 Reference

1. L.Adamic and E.Adar. Friends and neighbors on the web. *Social network*:211-230, 2003.
2. J.M.Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA*, 1998.
3. L.Page, S.Brin, R Motwani, and T.Winograd. The pagerank citation ranking: Bringing order to the web. *Technical Report* 1999-66, Stanford InfoLab, 1999.