Research Statement

Peng Cui
cuip@tsinghua.edu.cn

Department of Computer Science and Technology
Tsinghua University
Beijing, China

Research Background

Web social networks provide users with a platform for information communication and sharing, personal opinion delivery, emotion exchange and social activities. The socialization of web information endows the processes of information generation, communication and influence with more social attributes. Only using information science techniques about information processing and communication theory cannot reveal the characteristics of web social network structure and evolution mechanism, and social individuals' behavior patterns. Meanwhile, the Internetization of social network makes the social relations more fuzzy and dynamic, which results in the infectivity of small scale static social network analysis based sociology theory.

In my research, I will combine the sociology theory for social network analysis and the web information mining techniques together to address the following issues: (1) **Social Influence Analysis** for influencer mining and behavioral targeting; (2) **Social Contextual Recommendation** for personal and accurate social items recommendations; and (3) **Behavioral Information Flow Mining** to reveal the diffusion mechanism of social information.

Current Research 1 – Social Influence Prediction

People and information are two core dimensions in a social network. People sharing information (such as blogs, news, albums, etc.) is the basic behavior. In our research, we focus on predicting item-level social influence to answer the question Who should share What, which can be extended into two information retrieval scenarios: (1) Users ranking: given an item, who should share it so that its diffusion range can be maximized in a social network; (2) Web posts ranking: given a user, what should she share to maximize her influence among her friends.

We formulate the social influence prediction problem as the estimation of a

user-post matrix, in which each entry represents the strength of influence of a user given a web post. We propose a Hybrid Factor Non-Negative Matrix Factorization (HF-NMF) approach for item-level social influence modeling, and devise an efficient projected gradient method to solve the HF-NMF problem. Intensive experiments are conducted and demonstrate the advantages and characteristics of the proposed method.

Current Research 2 –Social Contextual Recommendation

Recommender systems eventually play the role of actively feeding the information items (e.g. web posts, products, etc.) to users according to their implicit intentions. Thus the users and items are two core dimensions that we should focus on to improve the recommending performance. Although the collaborative filtering based recommendation methods have been widely adopted for commercial applications, we know little about why users adopt or reject items, which has been the bottleneck for further improving the recommending performances. With the emergence of social network platforms, it is possible that we can understand more on both the user and item sides. Items there are not only the labels or descriptive contents any more, but contain more social attributes like freshness, hotness, the authors or senders, etc. Meanwhile, the users are more transparent, where their preferences, relationships, and influences are exposed by their implicit interactions with items and other users. How to fully exploit these social contextual information to reveal the information adoption mechanism, and accordingly propose a method for social contextual recommendations, is rather challenging and demanded.

To achieve this goal, we propose a probabilistic matrix factorization method to fuse the social contexts and the user-item interaction matrix. In the method, there are three low-rank latent spaces respectively corresponding to the user space, item space and inter-personal influence space. Meanwhile, we have 4 observed matrices directly derived from data: the user-item interaction matrix, the item content similarity matrix, the user preference similarity matrix, and the user-user interaction matrix. The three latent spaces are regularized by the observed matrices in that: (1) the users that are similar in hidden user space have similar preferences (derived from preference similarity matrix); (2) the items that are similar in hidden item space have similar descriptive contents (derived from content similarity matrix); (3) high interpersonal influences in the hidden influence space generates frequent interpersonal interactions; (4) the product of user hidden space and item hidden space corresponds to the users' personal

preferences on the items; (5) the product of interpersonal influence and personal preference is proportional to the probability of item adoption. Empirical results from two real social datasets prove that the implementation of our model can increase the accuracy of recommendation prediction 21% and 17%.

Current Research 3 –Behavioral Information Flow Mining

When a piece of information is generated on social platforms, it will spread in a complex cascaded way, which forms an information flow tree (as in Figure 1). Different from other information flow like in computer network or disease network, the information flows in social network carry rich signals about user behaviours, for example their attitudes, preferences, influences, etc. Thus we started to investigate social network from Behavioural Information Flow (BIF) perspective, and try to make some progress on the intrinsic problem of social network: why the information spreads in that way?

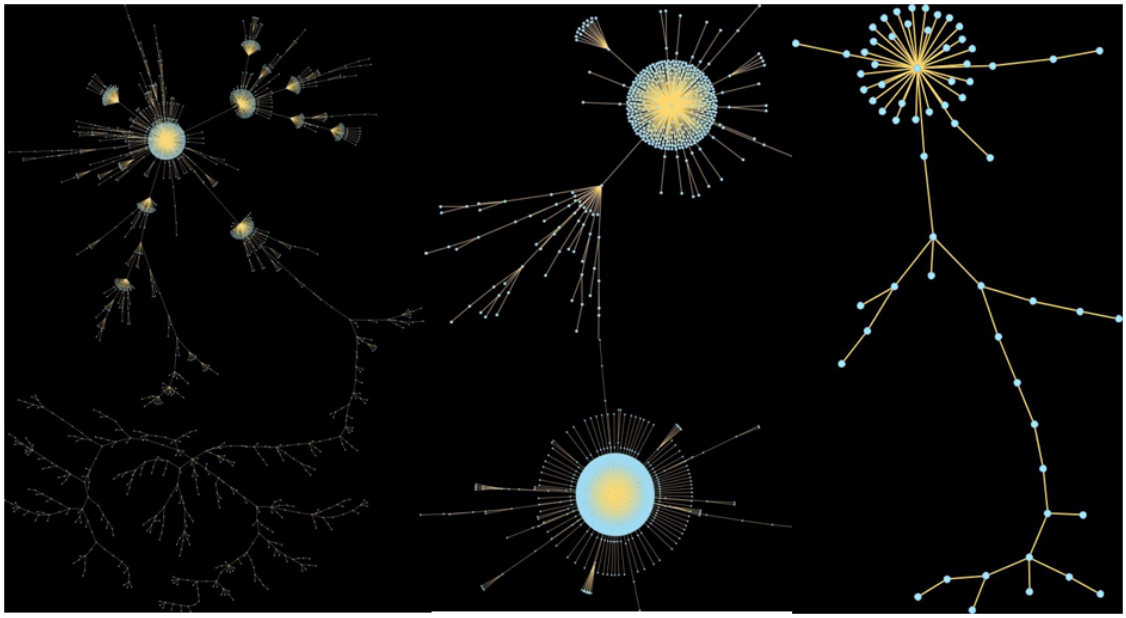

Figure 1. The BIF trees for 4 pieces of information in a microblog platforms.

Now we have accumulated flow trees for almost 10 million microblogs, and start to investigate the following detail problem: (1) The comparison of social network diffusion patterns or statistics with email and disease diffusions, and why; (2) The descriptive feature extraction for the BIFs, and propose a model for BIF reconstruction; (3) Information flow prediction in early stage.

Datasets

We now have rich dataset resources on social network analysis.

(1) Almost complete data of Tencent Weibo (a Twitter style microblog website in China), including user profiles, tweets, comments, user behaviours, complete diffusion parth, etc. The total user volume of Tencent Weibo is more than 100M.

(2) The complete access log of more than 100K users in Renren.com, a Facebook style social network web site in China. In the access log, every click action of the users are recorded. The total user volume of Renren.com is more than 80M.

(3) Other public datasets.

Team

We now have 3 phd students, 3 master students and several undergraduate students working on these research topics. Also, we've organized an interdisciplinary interest group for collaborative research, including three students from Computer Science, 2 students from Journalism and Communication, 2 students from Sociology, and 1 student from Psychology.