Identifying Influential Users in Social Network with Review Data

Yilin He

Machine Learning Department
School of Computer Science
Carnegie Mellon University, Pittsburgh, PA, 15213

Tóm tắt nội dung

Background: Social networks have been widely utilized by a variety of popular review websites, and different users have a different impact on how information propagates through the network. Identifying influential users in such system could be beneficial for advertising purposes.

Aim: To estimate the influence propagation probability for social ties in the network using review data which contains rating and temporal information. Then to identify influential users in the social network. The new method should be able to tolerate some basic attacks.

Data: The Yelp dataset was used in this analysis, and we focused on a subset of 5691 users who have at least one of 40077 reviews about 2724 businesses in Pittsburgh. Synthetic dataset for attacks is generated on top of the Yelp dataset where up to 10000 fake social ties are added in the network.

Method: The convex network inference model was adopted and applied the algorithm with prior knowledge of the social network. We then used a new method combining the two modifications on top of this method: the self-exploring model which supports user's information discovery without knowledge propagation; the rating model which incorporated rating data. The credit distribution model for influence maximization was used for comparison.

Result: Our new model achieved the best outcome in finding a set of influential users that have a high influence on the network. We also simulated attacks on the social network which changed 0.55% to 55% of the social network. Our results show that the total weight of the entire network can be changed up to 80%, but the influence per review for a set of the 100 most influential users only changes by 18% with this attack. If attackers focus on using users with only high number of reviews, this change could go up to 26.6%.

Conclusion: The network inference model is adopted and modified to estimate the influence propagation probability with review data, which is shown to perform well in identifying influential users in the network while being able to tolerate attacks on the social network without a severe impact on the result.

1 Introduction

The social network has been largely utilized by a variety of popular review websites. A lot of such websites supports login via social network sites such as Facebook or Google plus, and by doing so, it can quickly create a friend list for users to share the content with friends without suffering from a cold start problem. For example, popular review websites such as Yelp and Goodreads both support login through other social media account and will automatically link users' account with their friends' using outside social network information. Once the network is created, users can see their Facebook friends' reviews on Yelp or Goodreads without going through the process of manually adding friends after account creation [1].

The core features of these sites and traditional social network sites is the content creation and sharing among users. Based on positive or negative feedback given from friends, users might take different actions toward an individual product. Since users have different behavior pattern where some create content and some solely consume content, different users have a different impact on how information is propagated through the network.

2 Related Work

Richardson et al. [2] first started the study of influence maximization problem using proba- bilistic approaches and later Kempe et al. formulated the problem as finding a small subset of nodes k that maximizes the expected number of influenced nodes under a stochastic cas- cade model. This problem is proved to be NP-hard and a greedy algorithm is provided to solve the problem. However, this algorithm has a huge drawback of efficiency, so a lot of more recent work have been focusing on improving the scalability of the algorithm. In [3], 2Leskovec et al. proposed "lazy forward" algorithm which largely improved the efficiency of the algorithm by exploiting the submodularity property of the objective matrix. Chen et al. [4] later proposed yet another greedy algorithm for with new degree discount heuristics that further improves the efficiency even further while achieving a matching performance to the original greedy algorithm [5].

3 Problem Definition

To formally state the problem, we were given an undirected social network graph where represents the set of users in the social network and edge represents a social connection between user u and user . We also have a set of subject S where s S is the subject for user to review, such as a restaurant or a book. A review tuple is.

Some mathematics formula

$$F(x) = \arg \max_{y \in GEN(x)} w \cdot f(x, y)$$

4 Data

In this paper, we used a data set from Yelp Dataset Challenge. The entire data set contains a network of 552 thousand users with 3.5 million edges and their 2.2 million reviews for 77 thousand businesses from ten cities. A subset of the dataset is used in our analysis which contains all 2724 businesses which are located in Pittsburgh. Among the 17124 users have reviewed at least one of the selected businesses, we then selected the largest connected components in this subset of the social network containing 5691 users. Rest of the connected component only contains eight users or less and is thus ignored for rest of the study. The final data contains 5691 users with 18115 social ties and their 40077 reviews for 2724 Pittsburgh businesses. For our analysis, we used the following information from the dataset:

User file:

uid An unique identifier for each user

friends A list of uid for current user's friends in the social network. This friendship

Review file:

uid The unique identifier of reviewer.

sid The unique identifier of the business being reviewed.

review timestamp The date of which the review was made, counting from 1/1/2000 in the unit of days.

rating User's rating for this business, ranging from 1 star to 5 stars in integers.

Tài liệu

[1] Mario Baroni, Simon Maguire, and William Drabkin. The concept of musical grammar. *Music Analysis*, 1983.

[2] Mario Baroni, Simon Maguire, and William Drabkin. The concept of musical grammar. *Music Analysis*, 1983.

[3] Mario Baroni, Simon Maguire, and William Drabkin. The concept of musical grammar. *Music Analysis*, 1983.

[4] Mario Baroni, Simon Maguire, and William Drabkin. The concept of musical grammar. *Music Analysis*, 1983.

[5] Mario Baroni, Simon Maguire, and William Drabkin. The concept of musical grammar. *Music Analysis*, 1983.