Summary 1 of Yelp Challenge 2014

Xiao (Cosmo) Zhang

March 23, 2014

Contents

1 Introduction

All the basic information of the Yelp challenge 2014 can be found on this website: http://www.yelp.com/dataset_challenge

2 Influential Social network

Figure 1 shows a small fraction of a imaginary social network that can be restored from the yelp dataset. All the nodes are representing people in the network, while solid lines between two nodes indicte these two people are friends. We here propose a research topic is related to the prediction of a user's rating of one business, by using the information of this influential social network. A influential social network is defined as the social network where two nodes are influcing each other's attitude and formation of opnion towards certain objects. (ref here)

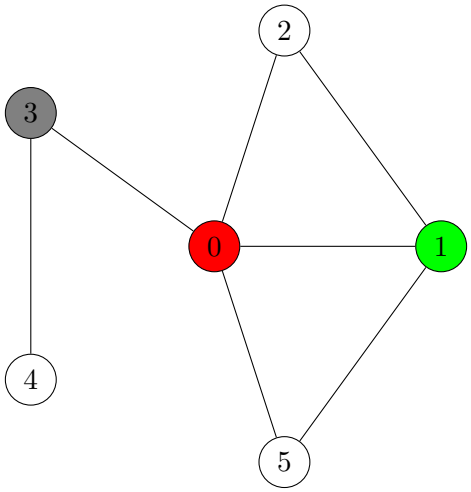

Figure 1: An Illustration Graph

For example, in Figure 1, node 1 and node 0 are inluencing each other on a certain business, and node 0 and node 3 as well. Suppose we are trying to predict the rating of node 0 on a certain business, and we are taking into consideration all the influence from his friends, and then all the four nodes, 1, 2, 3, and 5 will have an effect on him. We believe that friends' rating of a certain business will impact on any individual's making of decision of rating, but with diffent weights of diffeent friends. Hereby we currently propose three different perspectives that may strongly related with the weights of the influence of a friend's rating on this individual: they are the **the number of common friends**, **the similarity of this individual and the friend**, and **the persuasive style of the friend** (ref)

2.1 The number of common friends

We believe the more common friends two people have, the more chance these two people are within a same cluster, which implies these two people are having a deeper relationship. In this case, a friend of this type may have a stronger influence on the individual. For instance. in Figure 1, node 0, 1, 2, and 5 formed a small cluster, and node 0 and node 1 have two common friends, while node 0 and node 3 have no common friends. In this case, we assume node 1 will have a stronger influence on node 0. Here we denote the number of common friends by a.

a can be found by first forming a adjacent linked list. However, since a naive algorithm going through all the two pairs needs $\binom{n}{2} = O(n^2)$ time, and find all the common friends of a pair will also cosume $O(n^2)$ time, a total of $O(n^4)$ time will be need to perform this task. Therefore, transforming it into a adjacent matrix may bring some smart algorithm to reduce the time complexity.

2.2 The persuasive style of the friend

We believe different persuasive styles of an individual's friends will also have different weights on the decision making. For example, in Figure 1, the gray color of node 3 means node 3 is a very persuasive person, and likes give comments to things. In this case, appently from this percpective node 3 will have a stronger influence on node 0. We denote the The persuasive style of the friend by b, whose value is high is the friend is more persuasive.

We try to dig the persuasive style of the friend by using two methods, one is just the number of comments (or words of comments) he gave, which is denoted by w; the other is by mining the review text of a person, we try to identify his persuasive personnality, which is denoted by v. By doing a mining in the text of review, we might need some reliable NPL (Natural Language Processing) tools. Therefore, we can denote b as a function of w and v ($b = g(w, v)$).

2.3 The similarity of this individual and the friend

We use c to denote the similarity between two nodes. We assume, if two people have a high similarity, we will focus on similar topics and elements of the business, and be interested in similar aspects of a business. c can be obtained by using a inner product of two vectors as $c = \zeta_1^{\mathrm{T}}\zeta_2$, if we are using a vector ζ to represent the properties of a node. (ref)

Then our task will be how to dig the properties of a node. We propose the propoties are hidden topics in the review comments. Therefore, a LDA (Latent Dirichlet allocation) technology will be extremely useful to mine the hidden topics fdrom the review text. Or maybe Collaborative Filtering will also be useful.

3 Modeling

We can buid the model in such a way: First we propose $\beta_i = H(a_i, b_i, c_i)$, give a, b, and c, which is the weight function of the individual's ith friend. Also, if we consider the trade-off effect, we can wrtie $\beta_i = H(a_i + (1 - \alpha_1)b_i + (1 - \alpha_2)c_i)$, where α_1 and α_2 are importance coefficients. Then we can construct a regression model:

$$y = J(\sum_i \beta_i x_i) = J(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$$

, where $\mathbf{x}$ is his friend's ratings on a certain business, $\boldsymbol{\beta}$ is the influential weights of his friends as a vector, and y is the rating of prediction. Here we consider a multinomial logit regression, or a multinomial probit regression will be a good choice, because the output y is both discrete and ordinal.

4 Conclusion

We have already constructed the linked list. And next step we are trying to see how many common business people have ratings in this network.