Link Matching and Prediction in Bipartite Networks– Based on Yelp Reviews

Sidi Chang
schang03@email.wm.edu

March 24, 2016

1 Introduction

For the last five years, social network has become an important part of our daily life. As smart phones become increasingly popularized, Facebook, Google and Uber have changed our behaviors, as well as Yelp. Matching and predicting novel links has many applications in different areas, such as Ads, recommendation system and network analysis [4] [3]. Yelp Business rating prediction challenge [7] in 2015 gave us great datasets to analyze. In this paper, Motivation, prior related work, data collection and aggregation will be introduced. Methodology and algorithms for research plan will be discussed and concluded clearly and thoughtfully.

2 Motivation

My research interest is initially triggered by a debate concerning about whether Uber will make NYC traffic worse [2]. It made me curious about how Uber can affect NYC traffic in different areas and how Uber can recommend drivers to customers in peak hours. After searching and analysis of open source data from Github, the incompleteness of Uber data, lack of open API from Uber made my research hard to continue. Later, I saw Yelp competition in Kaggle which is similar to the task with Uber Data. Yelp Data Challenge [7] gives me an opportunity to conduct research on this field, evaluating on Link Prediction in different areas and recommendation of businesses to each customer.

3 Prior Related Work

The big data and its techniques is discussed in the work of H.V. Jagadish's group [3]. In this Review Article, the authors talked about key insights of big data. The insights states that most researchers only focus on part of creating value from Big Data rather than following all the steps. The methods surveyed show detailed steps on big data life cycle, from data acquisition to interpretation. The conclusion shows that even we have entered an era of Big data, not only key technical challenges but also social and political challenges need to be addressed.

The visual exploration concerning visualization methodology proposed by Nivan Ferreira's group [1]. In the article, the authors not only focus on methodology of methods and challenges to address, but they also propose a new visual query model that supports complex queries over

origin-destination(OD) date. With this implementation, people can access and visualize data much easily. This article also provides us new strategy to generate clutter-free visualization for large results with interactive response times.

The stock aware recommendation algorithm is a link prediction method proposed by Zhong's group [8]. Unlike the relatively simple methods surveyed in previous paper which only cares about the accuracy of matching and prediction, stock aware recommendation algorithm is a hybrid approach that combines bipartite matching and predicting with inventory size. The algorithm can be considered in two phases, in which recommendations for all users based on both user preference and inventory of different items are made as first step, then a dual method that can reduce the number of variables from n^2 to n is applied to improve significantly computing time complexity.

Previous work for a different Yelp Data Challenge Dataset is proposed as a final project for Stanford CS224W class [4]. Various methods such as preference attachment, Jaccard Coefficient are used in this paper. The authors address this problem in detail and tryed different methodology and algorithms to implement. Then the authors also compare experimental results and conclude their decision based on those results. Since this paper addresses several methodology and algorithm which is central to my work, I will describe it detailed in later section.

Lastly, Personalized PageRank is an extension to PageRank proposed by Page, Brin et al. in their original Page Rank algorithm [5]. Whereas the teleport vector is uniform over all nodes in basic PageRank, one can instead choose to specify a different probability distribution over the set of nodes, yielding personalized results for different individuals.

In my project, the difference between my work and other papers has two phases. The first phase is that I would follow some methodology and algorithms of those papers. [3] [1] give me some basic insight of big data procedure. [4] [5] give me methodology and algorithms I can use to make prediction and recommendation. I will combine those ideas and choose the most fit one on my project. The second phase is that I would follow some paper such as [8] and come up with some new idea with recommendation, my recommendation will mainly base on Near-Neighbor Search algorithm.

4 Problem Definition

Part of the problem is a previous challenge called RecSys Challenge 2013 which focuses on Yelp business rating prediction [7]. The source of my dataset is also from there. To transform this into an easily analyzable form, I will first make the following definitions:

- We assume there exists two bipartite sets: set of customers, set of business. Because two sets are bipartite, there does not exist any connection inside each set
- We say that there is an edge between customer i and business j if customer i gave business j a rating before

My problem is specified in two parts: Link prediction on a business rating, and business recommendation for customers.

Link prediction on a business rating. Given a set of businesses B, a set of customers C, I construct the B2C graph $G = (V, E)$ as follow: Let V be the set of $|B| + |C|$. For each customer i and business j, if customer has ever rated business, then edge $value(i, j) = rating$(1 Star to 5 Star). For the graph G, we consider the following features: a) The number of customers who have

rated to several businesses b) the number of customers who have interacted the same business. The link prediction task is then to predict the rating for each business in the future.

Recommendation Based on the previous graph, we need to make recommendations for each customers with their highest potential businesses they may go.

5 Data Collection and Processing

I collected Yelp Dataset Challenge [7] to serve as base data for my project.

5.1 Dataset Information

The Yelp data contains two main datasets: Training set and Test set. In training set, 11,537 businesses, 8,282 checkin sets, 43,873 users, 229,907 reviews; in test set, 1,205 businesses, 734 checkin sets, 5,105 users, 22,956 reviews.

5.2 Preprocessing

Training data and test data have the same format. The first part of my project is to do the prediction which involves the rating info, the second part is about recommendation which mainly focus on stars, reviews and categories. Dataset is stored in JSON object format. I will preprocess dataset into CSV format suitable for network analysis.

6 Methodology and Algorithms for research

As I mentioned before, the paper proposed by Kevin (Junhui) Mao, Xiao Cai and Ya Wang [4] discussed methodology and algorithms clearly and thoughtfully. Insight from their paper, I would use three classes of algorithms in Link prediction and one class of algorithm in Recommendation:

- scores based on proximity metrics using node neighbor features(Both prediction and recommendation)
- scores based on matrix factorization using customer ratings on businesses
- scores based on random network that serves as baseline

The following algorithms would be considered:

- Power Law [6]
- Preference Attachment [6]
- PageRank [5]
- Jaccard Coefficient [6]

Because only five weeks are left for finishing this project. During the first two weeks, I would finish Yelp prediction algorithm design and implementation first as several previous work has been done. Then I would spend another two weeks analyzing recommendation system. For the last week, I would write my Project paper and finish my presentation if there would be one.

References

[1] Nuno Ferreira, Jorge Poco, Huy T Vo, Juliana Freire, and Cláudio T Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2149–2158, 2013.

[2] Business Insider. Uber isn't making nyc traffic worse.

[3] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.

[4] Kevin (Junhui) Mao, Xiao Cai, and Ya Wang. Link prediction in bipartite networks - predicting yelp reviews.

[5] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[6] Anand Rajaraman, Jeffrey D Ullman, Jeffrey David Ullman, and Jeffrey David Ullman. *Mining of massive datasets*, volume 1. Cambridge University Press Cambridge, 2012.

[7] Yelp. Yelp business rating prediction.

[8] Wenliang Zhong, Rong Jin, Cheng Yang, Xiaowei Yan, Qi Zhang, and Qiang Li. Stock constrained recommendation in tmall. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2287–2296. ACM, 2015.