Predicting Jobs of Users' Interest

Vincent Yun Lou
Computer Science Department
Stanford University
yunlou@stanford.edu

Jinchao Ye
Computer Science Department
Stanford University
jcye@stanford.edu

Abstract

Nowadays, more and more companies are dedicated to help people to find ideal jobs. Such websites includes LinkedIn, GlassDoor, CareerBuilder and etc. Good job recommendation will benefit not only applicants, but also employers. Therefore, good job recommendation will make such companies more competitive.

Current algorithms usually categorize an applicant and then recommend popular jobs in that category to that applicant. However, many factors will affect which jobs an applicant will likely to apply, including location, salary, employer's reputation, the applicant's working history, the applicant's recent applications and so on.

We plan to use several algorithms to approach this problem. One algorithm is that we can extract a feature vector from the job description of each job and use logistic regression / support vector machine to classify whether a specific user will likely to apply. Or we can also form this problem as a clustering problem. We can cluster either similar users or similar jobs a user has applied. Then we can use such clusters to help us predict which jobs a user is likely to apply according to the distance between a job and the center of the cluster.

Future Distribution Permission

Data

The project idea is from a competition on Kaggle. We also got the data there. There are 6 files:

window_dates.tsv contains information about the timing of each window. Each row corresponds to a window, and has the date and time that the training period begins, that the training period ends, and that the test period ends.

users.tsv contains information about the users. Each row of this file describes a user. The UserID column contains a user's unique id number, the WindowID column contains which of the 7 windows the user is assigned to, and the Split column tells whether the user is in the Train group or the Test group. The remaining columns contain demographic and professional information about the users.

test_users.tsv contains a list of the Test UserIDs and windows, for your convenience. All of the information in this file can be found in users.tsv.

user_history.tsv contains information about a user's work history. Each row of this file describes a job that a user held. The UserID, WindowID, and Split columns have the same meaning as before. The JobTitle column represents the title of the job, and the Sequence column represents the order in which the user held that job, with smaller numbers indicating more recent jobs.

jobs.tsv contains information about job postings. Each row of this file describes a job post. The JobID column contains the job posting's unique id number, and the WindowID column contains which of the 7 windows the job was assigned to. The other columns contain information about the job posting. Two of these columns deserve special attention, the StartDate and EndDate columns. These columns indicate the period in which this job posting was visible on careerbuilder.com. Each job was visible for part of its 13-day window, but not necessarily for the entire 13 days. Users can only apply to a job between its StartDate and EndDate, so don't predict that a user applied for a job if the job was not visible for at least part of the 4-day Test period.

splitjobs.zip is a directory containing jobs1.tsv, jobs2.tsv, ... , jobs7.tsv, each of which contain all jobs in a given window. Thus, for example, jobs3.tsv contains all jobs in Window 3. This directory contains the exact same information as jobs.tsv, in the same format, and is provided merely for your convenience.

apps.tsv contains information about applications made by users to jobs. Each row describes an application. The UserID, WindowID, Split, and JobID columns have the same meanings as above, and the ApplicationDate column indicates the date and time at which UserID applied to JobId.