**Wiley Team 4: Bibliometrics Analysis of Authors**

Jiahui Bi, Shan Gao, Hongyi Chen, Yuyi Yan

The overall US publishing industry is financially healthy, the average revenue of scientific journals has grown by nearly 60% from 2010 to 2017.We are looking for methods to identify future leading authors in a specific scientific field. The data is downloaded from Web of Science (core collection) under the topic of Biomaterials.

Our project firstly analyzed the attributes that may indicate whether an author would become one of the top authors in the future or not. The aspects we considered are funding, the number of publications and impact factor. We also used social network methodology to build co-author network and add network features to the model. We found that network features like centrality and degrees, having a great importance. Since the core part of the project is to build an effective predictive model that will tell us the top authors in the future. We created a time gap in order to make prediction by using current features. The data from 2010-2012 is used to constructed features and 2013-2015 is used to create target variable. The response variable we used here is the sum of impact factor of certain author from second time period (2013-2015). That is to say we use features from former time period to predict these authors’ future performance measured by the sum of impact factor from next time period.

We used two kinds of regression models to do the prediction, Linear Regression and Random Forest Regression. The model predicted the next period impact factor of each author. Compared with the real number of impact factor, the linear regression with elasticNet has the lowest RMSE score. However, the scatter plots (real value vs predicted value) shows that linear regression model is more conservative in predicting top authors than random forest. To verified, we sorted the predicted values and gave a rank to each author for each model respectively. We compared these rankings with real ranking from both time periods. We measured the model performance in two ways: one is how many top authors we predicted, considering the real top N list, the other is how many potential authors (those did not appear on the top author list in the first time period) we predicted. The result shows that the random forest regression is more likely to pick up those potential authors.

For the future work, since we did not consider topic changing and funding was an important part of our feature space, we analyzed the data after 2010. For future analysis, more aspects can be considered, like citation network analysis, academic year, combined with topic analysis. Thus, we can expand our time span, more authors would be taken under consideration in a long run. In our work, we used impact factor to evaluate each author’ work during a certain time period, the distribution is right skewed and that resulted in a very unstable model. For future work, other variables can be chosen to measure the quality of each author’ work, like H-index.