Author Name Disambiguation Algorithm

ZhiCheng Ji
Tsinghua University
jizc19@mails.tsinghua.edu.cn

Zaiyuan Lu
Tsinghua University
lzc19@mails.tsinghua.edu.cn

1 Introduction

The growth of scientific literature makes the problem of Name Disambiguation more difficult and urgent. Although this subject has been extensively studied in academia and industry, the problem has not been solved well due to the clutter of data and the complexity of the same name scenario. AMiner together with Biendata held a competition during the last months of 2019 aiming for new algorithms to tackle this problem. The competition has finished and the winning team(Ziyue Quao and Hanxue Wang) achieved a final F1 score of 0,90327. Our new research tries to improve the existing method and find more useful informations to distinguish different authors

2 Data acquiring

The data will only be acquired from the Aminers Who is Who competition data. This set of information contains a trainset and a validation set with a substantial size.

3 Work planning

- 6-8 weeks
  learn the algorithm used in the competition
- 9-10 weeks
  try the model in the competition
- 11-15 weeks
  improve the model

4 Algorithms used

There are several indications to distinguish whether a paper belongs to a certain author. One of which is reviewing coauthors. If a paper is written by author A and B, and another paper is also written by the authors the same name, then it is very likely that the authors A and B has written these papers. Besides coauthoring, another feature would be the specific research organization the author is in. These are rules based matching techniques. Then, a clustering analysis will be conducted using DBSCAN

5 Testing

We plan to use the evaluation method of the competition based on the Macro Pairwise-F1.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

$$\text{PairwisePrecision} = \frac{\#PairsCorrectlyPredictedToSameAuthor}{\#TotalPairsPredictedToSameAuthor}$$

$$\text{PairwiseRecall} = \frac{\#PairsCorrectlyPredictedToSameAuthor}{\#TotalPairsToSameAuthor}$$

$$\text{Pairwise}F_1 = \frac{2 \times \text{PairwisePrecision} \times \text{PairwiseRecall}}{\text{PairwisePrecision} + \text{PairwiseRecall}}$$

Figure 1: Evaluation Method