R Notebook

Project Description

Term

- Spring 2017

Team

- Group 12

Team Members

- Kai Chen (Presenter)
- Senyao Han
- Kexin Nie
- Yini Zhang
- Chenyun Zhu

Project Summary

In our project, we studied entity resolution through two academic papers which introduce two methods.

Paper 2:

- Using linear svm, with its own unique way of evaluation
- Comparing the performance of different variables through its average accuracy.

Paper 5:

- Error-driven machine learning with a ranking loss function.
- Features selected (6 features): cosine similarity of coauthors,papers and journals; mean Euclidean distance (word2vec) of coauthors, papers and journals.

The detailed methods we applied are as follow: Clusterwise Scoring Function as the partition criterion, Error-driven Online Training to generate training examples and Ranking Perceptron as the loss function.

Paper 2 Evaluation

Feature:

Paper, Journal, Coauthor, Paper and Journal

Average Accuracy:

Paper	Journal	Coauthor	Paper and Journal
0.7275	0.6458	0.8093	0.8980

SD Accuracy:

Paper	Journal	Coauthor	Paper and Journal
0.0901	0.1159	0.0791	0.0444

Average Running time (s):

Paper	Journal	Coauthor	Paper and Journal
5.010	4.293	1.251	6.897

Paper 5 Evaluation

Feature

- Baseline (3 features): Cosine Similarity (coauthor, paper, journal)
- Improved model (6 features): Cosine Similarity, Mean Euclidean Distance

Algorithm

- Score function: weighted sum of cosine similarity of coauthor, paper and journal.
- Clusering predictive algorithm: bottom-up agglomerative clustering, linkage is the average distance or averaged cosine similarity, merge decision is according to maximum score.
- Parameter Update: comparison between the score of a better merge, and the score of the wrongly predicted merge.

Baseline

Model Evaluation (average)

Precision	Recall	F1 score	Accuracy
0.2321	0.7956	0.3478	0.5688

Parameters of features (average)

Cos_Coauthor	Cos_Paper	Cos_Journal
0.3176	0.4308	0.2516

Improved Model

Model Evaluation (average)

Precision	Recall	F1 score	Accuracy
0.4433	0.4930	0.4109	0.7674

Parameters of features (average)

Cos_Coauthor	Cos_Paper	Cos_Journal	dist_coauthor	dist_paper	dist_journal
0.6423	0.5835	-0.1141	-0.2480	0.1969	-0.0605

Model Comparison

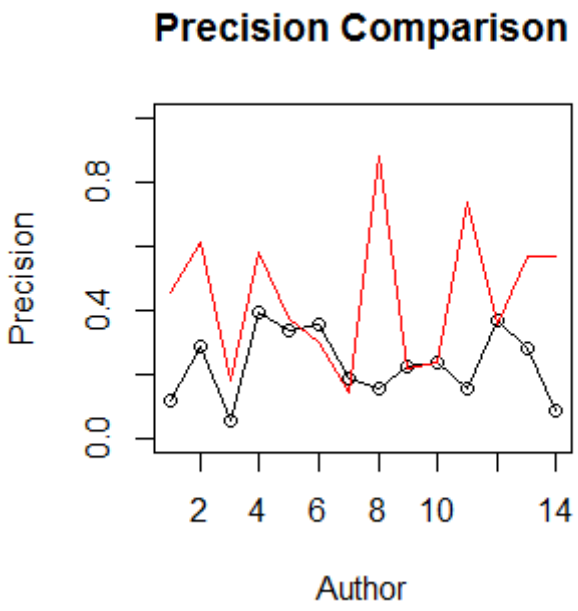

Figure 1:

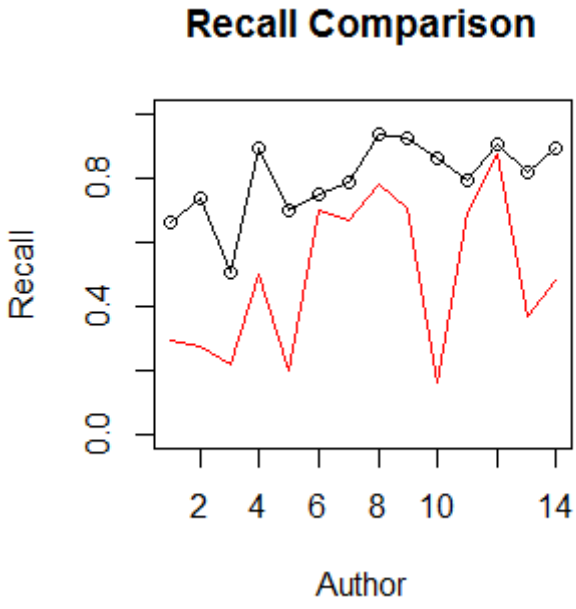

Figure 2:

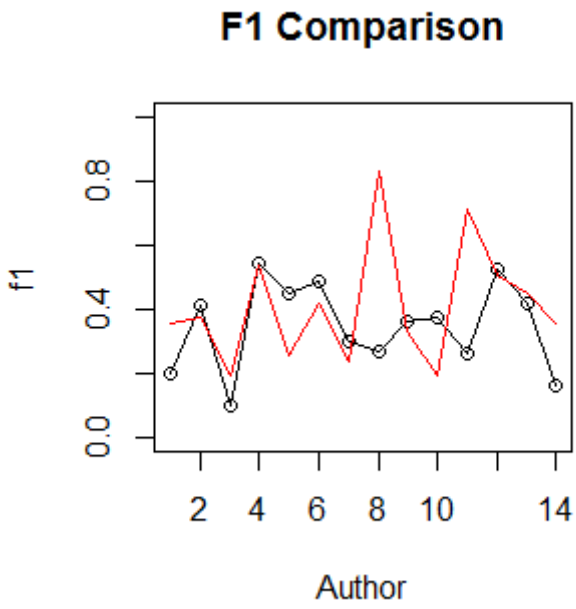

Figure 3:

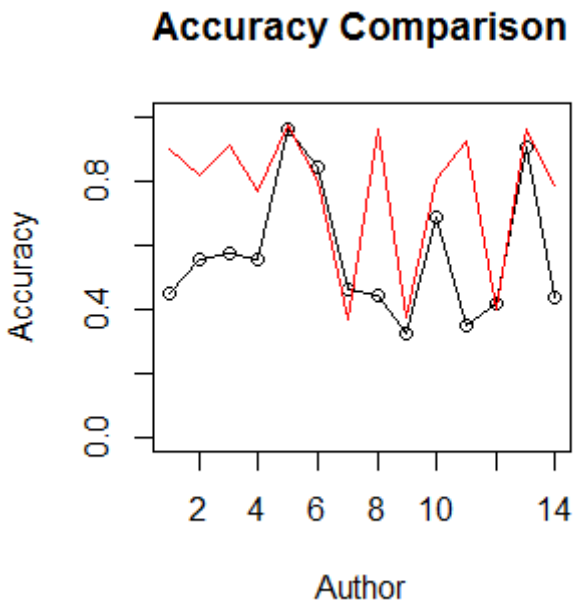

Figure 4: