Author profiling

zhiming Zhou

OUTLINE

- Introduction
- Previous work
- System overview
- Algorithm
- Conclusion

OUTLINE

- Introduction
- Previous work
- System overview
- Algorithm
- Conclusion

OUTLINE

- Introduction
- Previous work
- System overview
- Algorithm
- Conclusion

PREVIOUS WORK

```
quest_papers_by_authorname = 'select P.PaperID, P.NormalizedPaperTitle, P.C
```

```
for node in root.xpath('//id'):
    Paperid = node.text.split(',')
    rs_papers[i][0] = Paperid[0]
    i = i + 1
```

- disambiguate_by_coauthors.py
- Ajay Gupta.xml
- Alok Gupta.xml
- Barry Wilkinson.xml
- Bin Li.xml
- Bin Yu.xml
- Bin Zhu.xml
- Bing Liu.xml
- Bo Liu.xml
- Bob Johnson.xml
- Charles Smith.xml
- Cheng Chang.xml
- Daniel Massey.xml
- David Brown.xml
- David C. Wilson.xml
- David Cooper.xml
- David E. Goldberg.xml
- David Jensen.xml
- David Levine.xml
- David Nelson.xml
- Eric Martin.xml
- F. Wang.xml
- Fan Wang.xml
- Fei Su.xml
- Feng Liu.xml
- Feng Pan.xml
- Frank Mueller.xml
- Gang Chen.xml
- Gang Luo.xml
- Hao Wang.xml
- Hiroshi Tanaka.xml
- Hong Xie.xml
- Hui Fang.xml
- Hui Yu.xml
- J. Guo.xml
- J. Yin.xml
- Jeffrey Parsons.xml
- Ji Zhang.xml
- Jianping Wang.xml
- Jie Tang.xml
- Jie Yu.xml
- Jim Gray.xml
- Jing Zhang.xml
- John Collins.xml
- John F. McDonald.xml
- John Hale.xml
- Jose M. Garcia.xml
- Juan Carlos Lopez.xml
- Kai Tang.xml
- Kai Zhang.xml
- Ke Chen.xml
- Keith Edwards.xml
- Koichi Furukawa.xml
- Kuo Zhang.xml
- Lei Chen.xml
- Lei Fang.xml
- Lei Jin.xml
- Lei Wang.xml
- Li Shen.xml
- Lu Liu.xml
- M. Rahman.xml
- Manuel Silva.xml
- Mark Davis.xml
- Michael Lang.xml
- Michael Siegel.xml
- Michael Smith.xml
- Michael Wagner.xml
- Ning Zhang.xml
- Paul Brown.xml
- Paul Wang.xml
- Peter Phillips.xml
- Philip J. Smith.xml
- Ping Zhou.xml
- Qiang shen.xml
- R. Balasubramanian.xml
- R. Cole.xml
- R. Ramesh.xml
- Rafael Alonso.xml
- Rakesh Kumar.xml
- Richard Taylor.xml
- Robert Allen.xml
- Robert Schreiber.xml
- S. Huang.xml
- Sanjay Jain.xml
- Satoshi Kobayashi.xml
- Shu lin.xml
- Steve King.
- Thomas D.
- Thomas He
- Thomas Me
- Thomas Tra
- Thomas Wo
- Thomas Zim
- Wei Wang.
- Wei Xu.xml
- Wen Gao.xm
- William H. H
- X. Zhang.xm
- Xiaoming W
- Xiaoyan Li.x
- Yan Tang.xm
- Yang Wang.

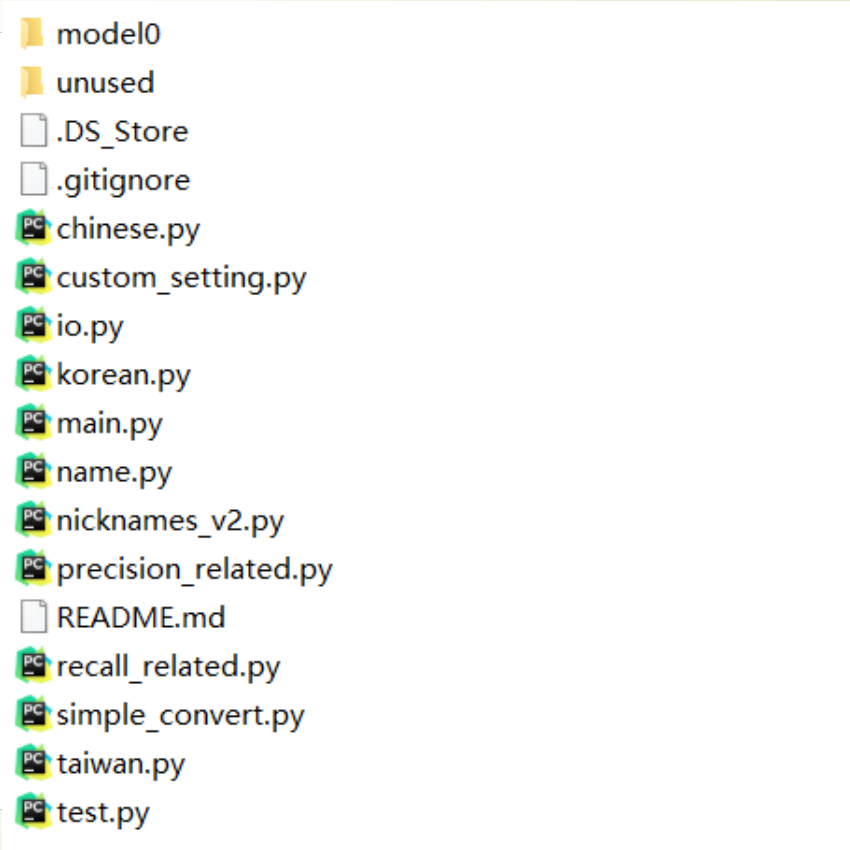
```
    rs_papers[i][4] = organization[0]
    i = i + 1
```

OUTLINE

- Introduction
- Previous work
- System overview
- Algorithm
- Conclusion

SYSTEM OVERVIEW

- Maximize the recall
- Maximize the precision



OUTLINE

- Introduction
- Previous work
- System overview
- Algorithm
- Conclusion

ALGORITHM

- **Pre-processing**
  - Clean the data:
    1. Noisy First or Last Names
    2. Mistakenly Separated or Merged Name Units

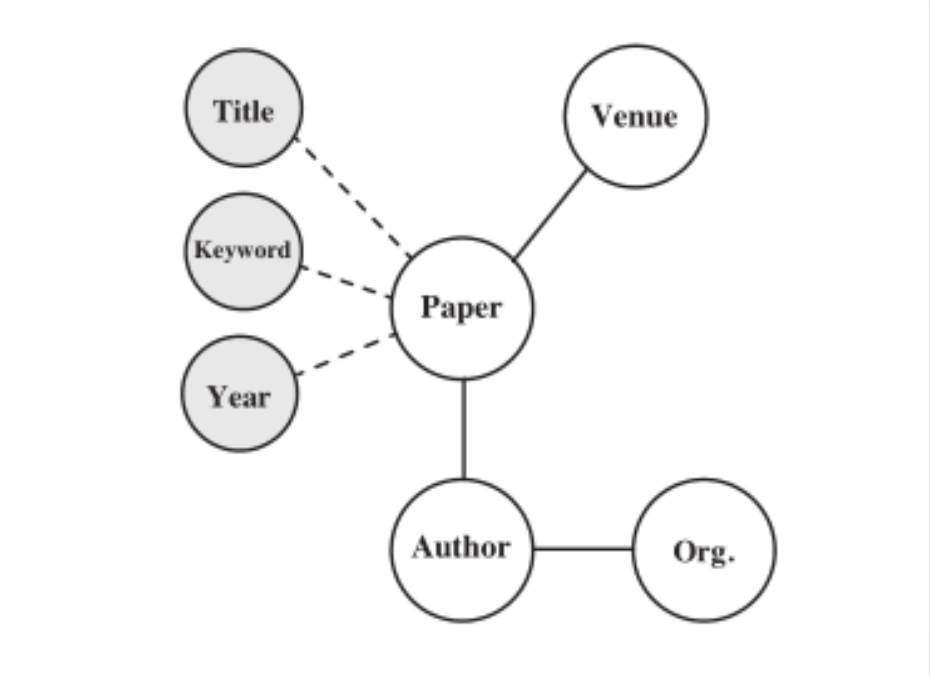
ALGORITHM

- **Improving the Recall**
  - String-based Consideration:
    1. Levenshtein Edit Distance
    2. Soundex Distance
    3. Overlapping Name Units
  - Name-Specific Consideration:
    1. Name Suffixes and Prefixes
    2. Nicknames
    3. Name Initials
    4. Asian Names and Western Names

ALGORITHM

- **Improving the Precision**
- Meta-Path-based Similarity:

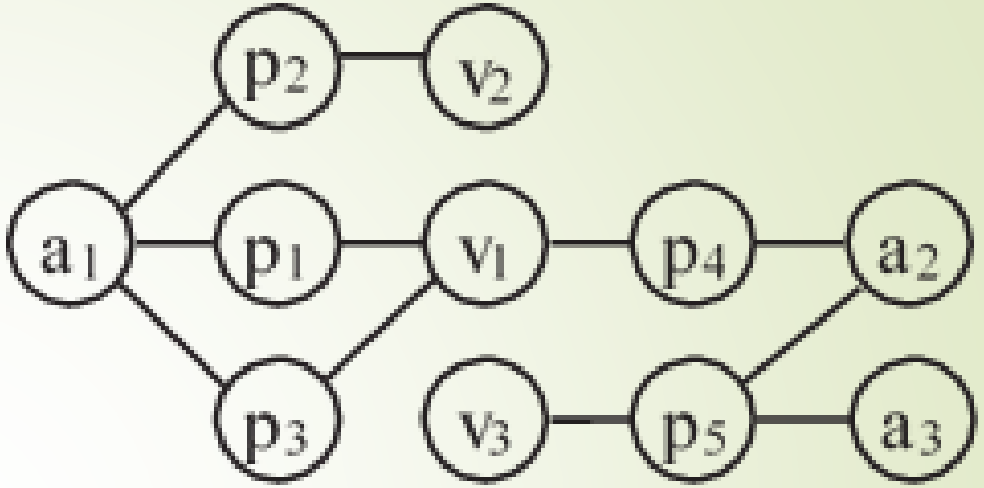
  The selected meta-paths are APA, AOA, APAPA, APV PA, APKPA, APTPA and APY PA. The weights for them are decreasing progressively.



ALGORITHM

- **Improving the Precision**
- Meta-Path-based Similarity:

  The selected meta-paths are APA, AOA, APAPA, APV PA, APKPA, APTPA and APY PA. The weights for them are decreasing progressively.



$$M_{A,P} = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & p_4 & p_5 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

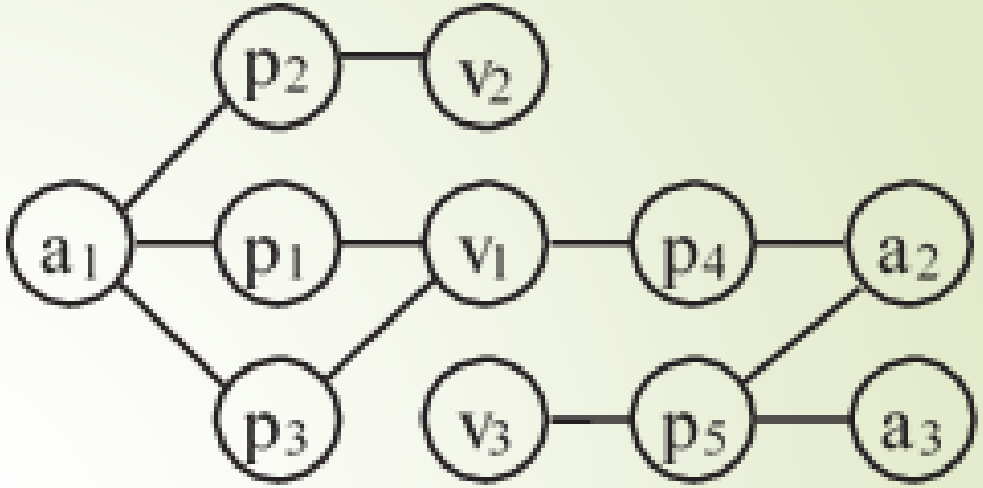
$$M_{P,V} = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{ccc} v_1 & v_2 & v_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$\overline{M_{A,V}} = Normalize(M_{A,P} \times M_{P,V})$$

ALGORITHM

- **Improving the Precision**
- Meta-Path-based Similarity:

  The selected meta-paths are APA, AOA, APAPA, APV PA, APKPA, APTPA and APY PA. The weights for them are decreasing progressively.



Author ID Pair	Similarity	Rank
(1, 2)	0.6325	2
(1, 3)	0	3
(2, 3)	0.7071	1

ALGORITHM

- **Improving the Precision**
- Ranking-based Merging

We do a scan from the top ranked ID pair to the lower ranked ones to help infer the author entity. And we will skip the conflict IDs, find one that has high similarity but also passes the name matching comparison, we believe these two IDs having high probability to be the real duplicate. After that, if A is the duplicate of B and B is the duplicate of C, we will consider that a is the duplicate of C.

Another important strategy is to expand the author names corresponding to the IDs once we are confident about two IDs to be the duplicate. This idea is useful because it can help avoid the mistakenly detected conflicts.

ALGORITHM

- **Post-processing**

  Unconfident duplicate author IDs should be removed even though their names are compatible and their meta-path- based similarity scores are acceptable. This step is crucial in that the later iterative framework requires highly confident output to gradually refine the results.

ALGORITHM

- **Iterative Framework**
  - An iterative framework which takes the detected duplicates of the last iteration as part of the input:
    1. we are able to generate much better meta-path-based similarity scores
    2. recall the name expansion module introduced at the end of the p-step

OUTLINE

- Introduction
- Previous work
- System overview
- Algorithm
- Conclusion

CONCLUSION

We have tried to disambiguation the author name, and we have found a better algorithm which is undoubtedly practical in KDD Cup Data Mining Contest 2013. But there is still lots of work need to be done. In the future, we need to adjust the code to our database, and we need to change some of the parameters to obtain the best result. I am looking forward to the day we complete the work, and I am firmly believed that our work will turn out to be a very important improvement of the Acemap.

Q&A

Thank You!