A Collective Approach to Scholar Name Disambiguation

PennState College of Information Sciences And Technology

北京航空航天大學 BEIHANG UNIVERSITY

Dongsheng Luo, Shuai Ma, Yaowei Yan, Chunming Hu, Xiang Zhang, and Jinpeng Huai

Motivation & Introduction

- Name ambiguity is very common.
  - 2000 papers written by over 200 different “Wei Wang” in DBLP.
  - 300 names are used by more than 114 million people in the United States.
- Name ambiguity causes problems in:
  - publication and author search
  - scholar entity ranking
  - document retrieval

Problem Definition

Given a citation record set, each citation contains author names, title, publication year and venue, the task of name disambiguation is to, for each name, partition its paper set, so that each subset contains all and only all papers written by an author in the real world.

Related Work

Supervised Methods

- human-labeled data
- High accuracy compared to Unsupervised methods.
- time-consuming. Impractical to label data when the dataset is large.

Unsupervised Methods

- Clustering or topic modeling
- Without manually labelled data.
- Relatively low accuracy.

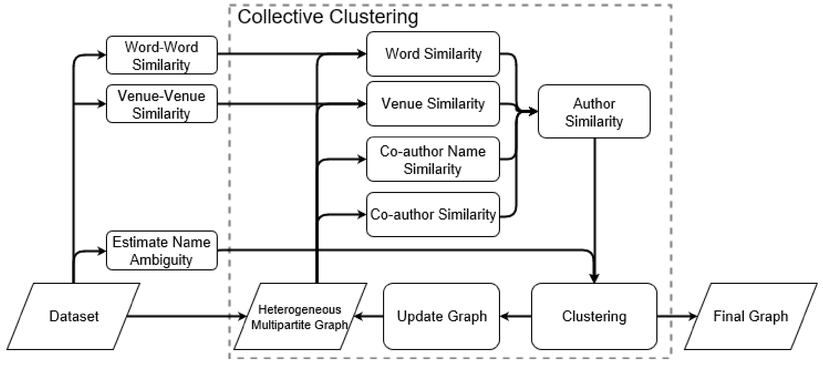
Challenges

- Limited Information is available.
  - Most digital libraries only provide basic citation information.
- Disambiguation results of names affect each other.
  - Coauthor names are also ambiguous.

Method

- Build a heterogeneous graph to represent the dataset.
- Extend author’s attributes by considering venue-venue similarity and word-word-similarity.
- Collective clustering
  - Step 1: Initialize a queue q by entering all names.
  - Step 2: Pop a name m from q and disambiguate m.
  - Step 3: Update the graph according to the disambiguate result of m.
  - Step 4: If the graph is stable, return the final graph. Otherwise, go to step 2.

Framework

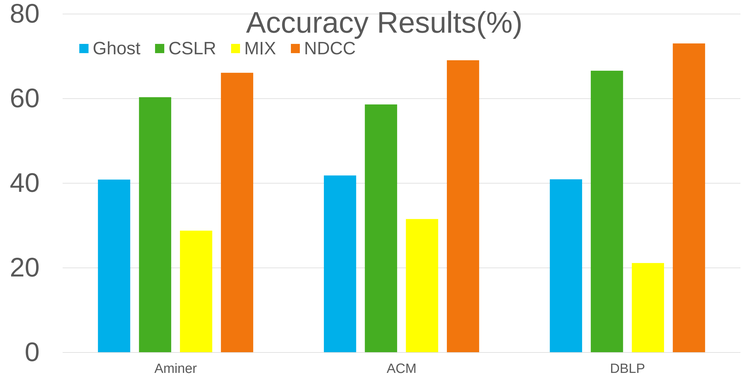


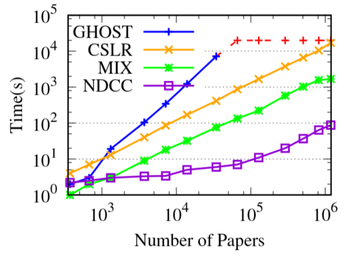
Dataset

Name	# Author Names	# Paper
AMiner	1,062,896	1,397,240
ACM	2,002,754	2,381,719
DBLP	1,871,439	3,566,329

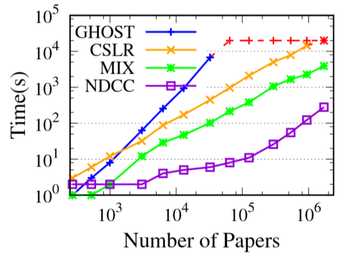
Experimental Results

Comparing our proposed method NDCC with three the state-of-art methods: Ghost, CSLR and MIX.

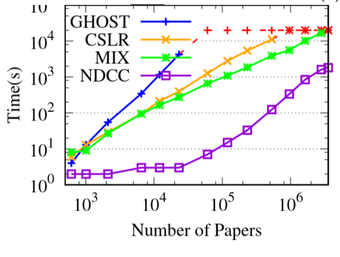




(a) AMiner



(b) ACM



(c) DBLP