Name Disambiguation via Multiple Rounds of Reduction

Stephen Carley, Alan Porter and Jan Youtie

Name 1

Name 2

Author A

Name 1

Author B

Name 1

Name 2

Name 3

Author C

Name 1

Name 2

Name 3

Author D

Name 5

Name 4

Name 6

Why does it matter?

- To measure the impact of a given author we need a way to build a dataset of scholarship unique to this individual
- For metrics to have meaning we need to accurately consolidate scholarship of individual authors (or institutions/disciplines/countries/etc)
- A Web of Science (WOS) search for:
  - Alan Porter based on his ORCID ID results in too few records
  - AU="Porter, Alan" drops true positives and introduces noise
  - AU="Porter, A" retains all true positives but introduces **significant** noise

Common Surnames

- According to the U.S. Census Bureau:
  - Carley is the 6,887th most common U.S. surname
  - Wang is the 282nd most common U.S. surname
  - Porter is the 159th most common U.S. surname

- According to the Chinese Ministry of Public Security:
  - Wang is the #1 most common surname in mainland China

Method

- Cast a broad initial net – i.e. WOS search for last name comma first initial
  - This almost always collects 100% of the true positives along with LOTS of false positives
  - We proceed deductively with our mission being to drop as many false positives and as few true positives as possible
- This procedure looks at authors with last name and first initial and then makes matches based on commonalities in fielded data

Method (cont.)

- A match is made if two author names (with the same last name and same 1st initial) share fielded data (from the match field – e.g. the same affiliation) with an author name the user knows to be (and selects as) a true positive.
- The initial dataset (based on a last name, first initial search) is iteratively reduced by applying this procedure to Match Field 1, Match Field 2, Match Field 3, etc.

Example 1: ZL Wang (Georgia Tech)

- An April 2017 search for him across publication years 2009 to 2017 yields:
  - 4,810 records
  - 428 authors with last name Wang and first initial Z
- How to identify the needle in the haystack – i.e. Georgia Tech's ZL Wang?

True and False Positive Coverage

- Of the 4,810 records there are 700 true positives (I cheated)
  - 15% true positives
  - 85% false positives
- A WOS search for ZL Wang based on his ORCID iD yields 626 records (89%)
- A WOS search for ZL Wang based on his ResearcherID yields 618 records (88%)

There are 25 match fields which drop 0 true positives for ZL Wang:

- Journal
- Source
- ISSN
- Title
- Author Affiliations (1st)
- Authors 1st
- Coauthors
- Author Affiliation (City, Country)
- Author Affiliation (Organization and City, Country)
- Organization Names Reprint
- Author Affiliations (Organization Only)
- Web of Science Category
- Keywords Plus
- Research Area
- Cited Journal
- Number of Authors
- Number of Author Affiliations
- Author City
- Cited References
- Combined Keywords + Phrases
- Document Type
- Countries
- Publication Type
- Publication Year
- Times Cited Matches

1st Round of Reduction: Journal

- Of the all the fields that drop 0 true positives, the field that drops the most false positives is Journal (216 false positives dropped), followed by ISSN/Title/Author Affiliations (1st)/1st Author/Coauthors (in that order).
- Matching based on Journal reduces our Authors list (of Wang, Zs) from 307 to 91.

2st Round of Reduction: 1st Author Affiliation

- Of these 91, the field that drops the most false positives (while retaining all true positives) is 1ST Author Affiliation (25 false positives dropped), followed by Title/1st Author/Author Affiliation Organization and City, Country (in that order)
- Matching based on 1st Author Affiliation further reduces our Authors list (of Wang, Zs) from 91 to 66.

3rd Round of Reduction: Title

- Of these 66, the field that drops the most false positives (while retaining all true positives) is Title (9 false positives dropped), followed by Source/1st Author/Coauthors (in that order)
- Matching based on 1st Author Affiliation further reduces our Authors list (of Wang, Zs) from 66 to 57.

4th Round of Reduction: Source

- Of these 57, the field that drops the most false positives (while retaining all true positives) is Source (2 false positives dropped), followed by Coauthors and ISSN (and those are the only two remaining fields of the ones that drop 0 true positives but eliminate false positives)
- Matching based on 1st Author Affiliation further reduces our Authors list (of Wang, Zs) from 57 to 55.

5th and 6th Rounds of Reduction: Coauthors & ISSN

- Of these 55, Coauthors drops 1 false positive (while retaining all true positives) and ISSN does the same, resulting in a final list of 53 authors names (6 of which are true positives). The initial list consisted of 307 Wang, Zs (254 of which were dropped)
- So we went through a total of 6 rounds of reduction, shrinking our original list by 83% before we hit the point where we couldn't drop anymore false positives (without also dropping true positives)

Example 2: Alan Porter (Georgia Tech)

- An July 2017 search for AU=Porter, A yields:
  - 3,617 records
  - 174 authors with last name Porter and first initial A
- How to identify the needle in the haystack – i.e. Georgia Tech's Alan Porter?

True and False Positive Coverage

- Of the 3,617 records there are 234 true positives (I cheated)
  - 6% true positives
  - 94% false positives
- A WOS search for Alan Porter based on his ORCID iD yields 93 records (40%)
- A WOS search for ZL Wang based on his ResearcherID yields 93 records (40%)

Match Results for Alan Porter

MATCH FIELD	# TRUE POSITIVES RETAINED	DATASET REDUCTION
Journal	234 (100.0%)	1,240 (34.3%)
Coauthors	233 (99.6%)	902 (24.9%)
Title	233 (99.6%)	791 (21.9%)
ORCID ID	234 (100.0%)	1,414 (39.1%)
ResearcherID	234 (100.0%)	1,535 (42.4%)
Email (9% coverage)	116 (49.6%)	1,524 (42.1%)
Publication Year	234 (100.0%)	712 (19.7%)
ISSN	234 (100.0%)	1,131 (31.3%)

True Positives Retained: Porter v Wang

MATCH FIELD	# TRUE POSITIVES RETAINED FOR ALAN PORTER	# TRUE POSITIVES RETAINED FOR ZL WANG
Journal	234 (100.0%)	700 (100.0%)
Coauthors	233 (99.6%)	700 (100.0%)
Title	233 (99.6%)	700 (100.0%)
ORCID ID	234 (100.0%)	695 (99.3%)
ResearcherID	234 (100.0%)	695 (99.3%)
Email	116 (49.6%) (field coverage: 9%)	695 (99.3%)
Publication Year	234 (100.0%)	700 (100.0%)
ISSN	234 (100.0%)	700 (100.0%)

Dataset Reduction: Porter v Wang

FIELD	% DATASET REDUCTION FOR ALAN PORTER	% DATASET REDUCTION FOR ZL WANG	DIFFERENCE
Journal	34%	25%	9%
Title	22%	21%	1%
Coauthors	25%	13%	12%
ISSN	31%	26%	5%
Publication Year	20%	10%	10%
ORCID ID	39%	37%	2%
ResearcherID	42%	33%	10%
Email	42%	30%	12%

Discussion

- Factors which make it difficult to isolate scholarship include:
  - Common names
  - Variation in name spelling (for the same author)
  - Voluminous scholarship
- Pros:
  - Significantly reduces manual effort
- Cons:
  - Unlikely to reduce a very large initial dataset down to a sub-dataset consisting of only true positives
  - When match field coverage isn’t high true positives often get dropped