Handling noise in scholarly search

Challenges & a case study

Anurag Acharya
Overview

Sources of errors

Processes that generate errors

Processes that amplify errors

Approaches for noise-resistance

Author disambiguation: a case study
Sources of errors

Comprehensive index requires blending diverse sources
- Publishers (1000s), libraries (1000s), repositories (1000s), databases (100s)
- Different shortcuts, hacks, agglomeration of errors

Comprehensive index covers diverse fields
- Conflicting standards, conventions - each may be consistent, but...
- Never been designed to be grouped with others
Sources of errors

Comprehensive index covers many languages & cultures
  Each language & culture has its own conventions, some conflict
Print based publication norms favored brevity over clarity
  Packing as much as possible in a few pages -> ambiguous data
Sloppy authors
Just plain wrong data
Processes that generate errors

Overly concise citation styles

E.g., “M Jones, TBH 123, 23, 1923, 1961”

Easy to get things wrong, hard to identify errors, hard to recover

Authors “creative” in citations

Reordered/dropped authors, partial titles, half-remembered names

>900 ways to refer to Proc of National Academy of Sciences
Processes that generate errors

Unfamiliarity with cultural norms

Consider Spanish name: Juan Martinez Gomez

Martinez - father’s last name, Gomez - mother’s last name

Can be abbreviated as J Martinez Gomez or J Martinez

Often mistakenly abbreviated as JP Gomez or J Gomez
Processes that generate errors

Portuguese cultural norms the other way around

José Eduardo Tavares Silva

Silva - father’s last name, Tavares - mother’s last name

Can be José Silva, José Eduardo Silva, JE Silva

Or José Tavares Silva, JE Tavares Silva

But not José Tavares, José Eduardo Tavares or JE Tavares
Processes that generate errors

Transliteration of names combined with abbreviation

E.g., 200 Chinese last names cover 95% of population

Three Korean last names cover over 50% of population

Publication process presents names as FirstInitial LastName

Author names become inherently ambiguous

李开复 -> K Lee
Processes that generate errors

Name presentation variations

Name order & capitalization: JE Smith, WU Huican, P Sundar, Sundar Pichai
Multi-component firstnames: Kai-Fu Lee is likely not the same as Kai Lee
Different name order in one citation: Smith John, Ben Neller
Commingled affiliations: “Depts of Biology & Chemistry and Neurology, USC”
Abbreviated affiliations: “Biology, USC”
Processes that generate errors

Desire to fill missing fields with “something” - authors, libraries, publishers

“Online date” included as publication date

Advisors included as “contributors” on theses

Issue number included for journals that have only volumes
Processes that amplify errors

Citing references directly from bibliography
Citing from shared bibtex/endnote libraries
Citing from databases that have aggregated errors
Viral propagation of errors

Seen cases with > 200 citations, all identical, all wrong
Processes that amplify errors

Aggregation without due caution

- Not normalizing enough -> large number of duplicates
- More popular the item, more the variations
- Popular books have 100s of variations
- Popular conference names can have 1000s
Processes that amplify errors

Aggregation without due caution

   Over normalizing -> conflate distinct entities

   1000s of “B Lee”s, 10s of IJBC

Once the distinction is lost, hard to recover

   All further computation, all derived data tainted

Aggregated data is then released and fed back into publication
Approaches for noise-resistance

Preserve all source data, do entire recomputation periodically

Many systems build & maintain a “canonical” dataset

Errors once made can be hard to recover from

Periodic recomputation allows global improvements

New algorithms, more data, new ways of cross-checking
Approaches for noise-resistance

Compute frequency statistics for all components

Add monitoring for outliers and significant changes

Titles that occur too often are probably wrong

Single date for too many papers is probably wrong

Large change in citation counts over a short period is probably wrong
Approaches for noise-resistance

- Propagate as much source data through processing as possible
  - Avoid committing to “canonical” values early in processing
  - Later stages in processing may have a broader view, could second guess

- Build domain-specific glass-box models
  - General purpose black box models can be hard to diagnose
  - Leverage knowledge of scholarly publishing space
Approaches for noise resistance

Design explicitly for linguistic/cultural variation

E.g., not require initials, custom handling of different kinds of names

Look for multi-component mutual support

E.g., allow multiple author name formats if titles match well
Case study: author disambiguation

Core component of Scholar profiles

Allows profile setup within few minutes

Automated updates reduce author maintenance effort
Richard Thaler
University of Chicago, Booth School of Business
Verified email at chicagobooth.edu
Behavioral Economics

<table>
<thead>
<tr>
<th>TITLE</th>
<th>CITED BY</th>
<th>YEAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nudge: Improving decisions about health, wealth, and happiness</td>
<td>9224</td>
<td>2008</td>
</tr>
<tr>
<td>RH Thaler, CR Sunstein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constitutional Political Economy 19 (4), 356-360</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the stock market overreact?</td>
<td>7883</td>
<td>1985</td>
</tr>
<tr>
<td>WFM Bondt, R Thaler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Journal of finance 40 (3), 793-805</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mental accounting and consumer choice</td>
<td>5811</td>
<td>1985</td>
</tr>
<tr>
<td>R Thaler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marketing science 4 (3), 199-214</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toward a positive theory of consumer choice</td>
<td>5489</td>
<td>1980</td>
</tr>
<tr>
<td>R Thaler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of Economic Behavior &amp; Organization 1 (1), 39-60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anomalies: The endowment effect, loss aversion, and status quo bias</td>
<td>4439</td>
<td>1991</td>
</tr>
<tr>
<td>D Kahneman, JL Knetsch, RH Thaler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Journal of economic perspectives 5 (1), 193-206</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental tests of the endowment effect and the Coase theorem</td>
<td>4116</td>
<td>1990</td>
</tr>
<tr>
<td>D Kahneman, JL Knetsch, RH Thaler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of political Economy 98 (6), 1325-1348</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fairness as a constraint on profit seeking: Entitlements in the market</td>
<td>3617</td>
<td>1986</td>
</tr>
<tr>
<td>D Kahneman, JL Knetsch, R Thaler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American economic review, 728-741</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A survey of behavioral finance</td>
<td>3389</td>
<td>2003</td>
</tr>
<tr>
<td>N Barberis, R Thaler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handbook of the Economics of Finance 1, 1053-1128</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Author disambiguation

Build statistical model grouping articles by author with same name

Author lists, journals, research areas, affiliations, co-authors, article text

Multi-dimensional model, customized for each dimension

Shoot for high precision, very good recall
Author disambiguation

Many authors work in many areas, interests shift over time

  With many co-authors, many journals/conferences

  Trying to group all of these will over-cluster others

Allow different enough groups to remain separate

  Give up some recall for precision
Author disambiguation

Bring human recognition to bear for final disambiguation

Recognizing your own papers is easy & quick

Make it trivially easy to recognize

Allow authors to select & merge multiple groups
Human disambiguation

Two steps of disambiguation under the covers

Model groups papers likely by “same author”

Says “Same John Smith” not which one

Step 1: select “my” paper group

Step 2: (if needed) select multiple paper groups
Human disambiguation

Authors can also remove & add individual articles

Author’s choices fed back into automated update computation

- Improves precision/recall for profile updates
- Automated updates -> need higher precision/recall
How well does it work?

Large number of Scholar profiles created worldwide

- Includes many authors with 100s of papers
- Most profile setups take 5-10 minutes
- Most profiles automatically maintained
Noise-resistance approaches used

Preserve all source data, recompute paper groups frequently

Custom glass-box model, name structure, co-authorship, text, journals

Avoid committing to “canonical” values early - authors do final selection

Design explicitly for linguistic/cultural variation
There is no such thing as “clean” data

There is “noisy” data and “less noisy” data :)

Thank you!