

For conference – subject relationship:

1. Laurel Orr and Jennifer Ortiz: Clustering with the DBLP Bibliography to Measure External Impact of a Computer Science Research Area.

(https://homes.cs.washington.edu/~jortiz16/images/MLProjectPaper.pdf)

1. <https://webdocs.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html>
2. <https://en.wikipedia.org/wiki/List_of_computer_science_conferences>
3. <http://www.cs.jhu.edu/~taochen/SoC_Conference_Ranking.html>
4. <https://www.scimagojr.com/journalrank.php?area=1700&category=1702>
5. Earlier we had planned to use keywords, but not all the research papers have them.
6. Another approach that was decided upon was distant reading.
7. More than 5561 conferences
8. 11940 conferences on dblp
9. 4612 journals on dblp

Subjects:

* AL: Algorithms and Theory.
* LS: Languages and Software.
* CA: Computer Architecture.
* DB: Databases
* DM: Data Mining and Information Retrieval
* AI: Artificial Intelligence.
* NL: Natural Language Processing
* ED: Computer Education
* W3: Web and Information Systems.
* DP: Distributed and Parallel Computing
* GV: Graphics, Vision and HCI..
* MM: Multimedia
* NC: Networks, Communications & Performance.
* SE: Security and Privacy
* OS: Operating Systems / Simulations

For downloading PDFs:

1. Earlier semanticscholar, but sometimes it links to sciencedirect, for which our institute does not have a subscription.
2. Thus, we are now using <http://sci-hub.tw/>. Just DOI is required to be appended to this link. Not sure if this website is legal.

For author affiliations:

1. Earlier, we decided to first search the occurrences of ‘@’ in the document. The logic being, this will give us the email ids of the authors. Now from the domain name of these email ids, we can find out the name of the institute, using this json file:

<https://raw.githubusercontent.com/Hipo/university-domains-list/master/world_universities_and_domains.json>

1. But not all research papers had domain names (actually a very large number). So, we decided to search for keywords like ‘University’ or ‘Institute’ to find the name of author’s affiliation. This also failed as many institutes don’t have such keywords, and even if they have one it is in many cases in their regional languages (like in German).
2. Also, there was a problem with linking, which author should be linked with which email id. Many a times there are more authors than email ids or names of universities. For example, one format of writing email ids that was observed was: [Person\_A, Person\_B]@uni.edu.
3. After referencing few reference papers:
   1. Ozair Saleem, Seemab Latif: Information Extraction from Research Papers by Data Integration and Data Validation from Multiple Header Extraction Sources.
   2. Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, Bela Gipp: Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents.
   3. Dominica Tkaczyk, Pawel Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, Lukasz Bolikowski: CERMINE: automatic extraction of structured metadata from scientific literature.
   4. Zhixin Guo, Hai Jin: Reference metadata extraction from Scientific Papers
4. We found that each of these techniques uses Machine Learning approaches. So, we decided to use any of the mentioned tools in these papers.

For author disambiguation or homonym detection:

1. Marcel R. Ackermann, Florian Reitz: Homonym Detection in Curated Bibliographies: Learning from dblp’s Experience.
2. Fakhri Momeni, Philipp Mayr: Using co-authorship networks for author name disambiguation.
3. E. Amigó, J. Gonzalo, J. Artiles, F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints".
4. A. A. Ferreira, M. A. Gonçalves, A. H. F. Laender, "A brief survey of automatic methods for author name disambiguation".
5. T. Gurney, E. Horlings, P. V. den Besselaar, "Author disambiguation using multi-aspect similarity indicators".
6. P. Mayr, F. Momeni, "An open testbed for author name disambiguation evaluation".
7. H. T. Nguyen, T. H. Cao, "Named entity disambiguation: A hybrid statistical and rule-based incremental approach".
8. All of these papers implemented ML, graphical networks or approaches not known to us. Most of these papers were above our level of understanding.
9. Initially, we had decided to employ unsupervised learning with a suitable clustering algorithm, but after contacting few issues with this approach, we gave upon this idea.
10. Now, we plan to implement a rule-based scoring approach.
11. The rules/algorithms were decided by us solely, without any external help.
12. But when we stumbled upon this research paper:

E. Caron, N. J. van Eck, "Large scale author name disambiguation using rule-based scoring and clustering".

<https://s3.amazonaws.com/academia.edu.documents/42806714/Research_quality_characteristics_of_publ20160218-10100-1wsawux.pdf?response-content-disposition=inline%3B%20filename%3DResearch_quality_characteristics_of_publ.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20191116%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20191116T133929Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=ae8b778c46bb25705a6c717277297b2ebd2fa6260528fd7d42374efabe509112#page=91>

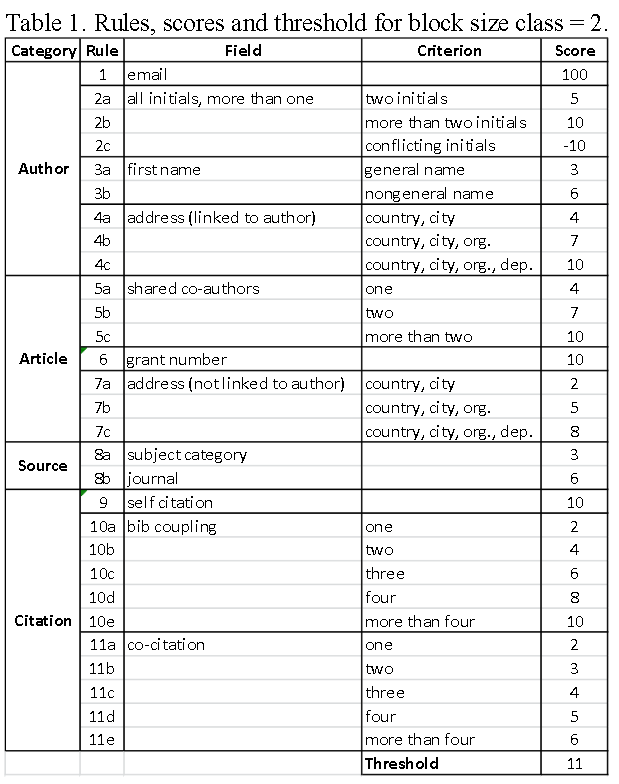
1. We only changed our scoring scheme, according to this paper, everything else remains the same.

For DBLP search:

1. <https://dblp.uni-trier.de/xml/docu/dblpxmlreq.pdf>
2. <https://github.com/scholrly/dblp-python>
3. Issues: [Now resolved by us]
   1. Syntax error
   2. Not running for homonyms

Scoring Schema:

|  |  |  |
| --- | --- | --- |
| Category | Criterion | Score |
| Common Co-Authors  (other than the author in question) | 1 (both papers have less than 4 authors each) | 8 |
| 1 (any of the paper has more than 4 authors) | 4 |
| 2 or more | 4 per common co-author |
| Affiliation  (of the author in question) | matched | 10 |
| Journal/Conference | Exact match | 6 |
| 1 or more corresponding subjects match | 3 |
| Subject | 1 | 2 |
| 2 | 4 |
| 3 or more | 5 |
| Self-Citation |  | 10 |
| Bibliographic coupling | 1 | 2 |
| 2 | 4 |
| 3 or more | 5 |
| Threshold | | 11 |



Ai, computational theo and math, comp graphics and cad, comp net and comm, cs appl, cs misc, comp vision and pattern recog, hardware and arch, info sys, signal processing, software