Mining Author Identifiers for PubMed by Linking to Open Bibliographic Databases

Li Zhang[*], Yong Huang[†], Qikai Cheng[‡] and Wei Lu[§]
School of Information Management
Wuhan University
Wuhan, China
Email: [*]zhangli@whu.edu.cn, [†]yonghuang1991@whu.edu.cn, [‡]chengqikai0806@163.com, [§]weilu@whu.edu.cn

***Abstract*—Author identifier (ID) is essential for many downstream tasks, such as co-author network and scientist mobility analysis. As a widely used database, author ID of PubMed is not officially provided by National Institutes of Health (NIH), that restrict some identifier-based researches or systems. This study exploited three open bibliographic databases Aminer, Microsoft Academic Graph (MAG) and Semantic Scholar (S2) to associate author ID for PubMed. For this purpose, paper linking and author linking was performed in order to mine paper and author links between PubMed and these databases. Performance of author name disambiguation (AND) was evaluated on two datasets. Our findings suggested that, S2 contains full volume of PubMed regarding link completeness. With respect to correctness of author ID, S2 and MAG achieved better performance than Aminer. The best F1 score of there available identifiers is below 90%, indicate AND for large scale database remain as a difficult task and efforts are being need for further improvement. We made the final dataset publicly available for facilitating future research.**

***Index Terms*—PubMed, Author ID, Author Name Disambiguation, Evaluation**

I. Introduction

Author ambiguity is widely presented in many bibliographic databases. For some popular names (e.g., ”John smith”), we are not sure all the search results that share similar name are the same person, and this uncertainty is even higher for abbreviated names, e.g., ”J. smith”. This problem can be eliminated if author can be correctly identified. Identifying unique author is important for many problems, such as author-related queries, co-author network analysis and scientometric measures. As a widely used bibliographic database, PubMed does not officially provide author identifier (ID) for its over 30 million papers. Recent studies [1], [2] reported that, although PubMed has integrated author ID created by Liu et al [3], it is still not publicly available - The ID is not visible on any of the pages that a researcher could access. To identify author uniqueness for PubMed, a number of author name disambiguation (AND) methods have been proposed in recent years [1], [4], [5], [6]. However, high computational complexity makes these methods difficult to perform on PubMed-scale databases. Other measures were taken in a different way. Many unique author systems were developed and became increasingly popular for researches: ORCID, Google Scholar, Mendeley, Scopus, ResearcherID, ResearchGate [1], [7], etc. However, not all of these author identifiers are interconnected, and also not widely used in today’s bibliographic databases, with a large proportion of authors without creating identifier in these systems. To address this problem, three well-known databases, Aminer [8], Microsoft Academic Graph (MAG) [9] and Semantic Scholar (S2) [10] were used, all of them have their own author ID system and include over 170 million articles. We first mined the paper links to PubMed from these databases. Then, author link was matched based on the same author order of the same paper in different databases. Further more, two AND datasets were used for evaluating performance of author IDs. Note that we only included author IDs from the mentioned databases, author ID from other sources (e.g., Web of Science) is not considered due to availability. Recently proposed AND methods [5], [11], [12] were also eliminated for evaluation, since the high complexity of model and computation makes it difficult to disambiguate all authors on PubMed-scale database.

Our contributions are twofold: First, we mined the links to PubMed from Aminer, MAG and S2, paper and author links for PubMed were derived along with three type of author IDs. Second, we found S2 has highest coverage of PubMed, evaluation on two PubMed-related AND dataset also indicated that MAG and S2 author ID performed better than Aminer. We made the author IDs dataset publicly available[1]. As a set of easily accessible and reproducible baselines for AND research community, we believe it could facilitate identifier-based researches.

II. Linking Approach

In this section, we dive into Aminer, MAG and S2 to explore their outlinks to PubMed. Since the papers are collected from the Internet, the source link is also stored. Fig. 1 shows the process to mine the paper links and author links. We first mined the links of the same paper among these databases, then author links are mined using an author matching strategy.

To mine paper link, we first located the outlinks in each database. For Aminer, The *paper* table in Aminer open data[2] contains the outlinks, which is a string array type that points to alternative links. In mining the target link, keywords

[1] https://zenodo.org/record/3748896#.XwIdK2gzaUk
[2] https://www.aminer.cn/oag2019

978-1-7281-8915-4/20/$31.00 ©2020 IEEE
DOI 10.1109/QRS-C51114.2020.00043

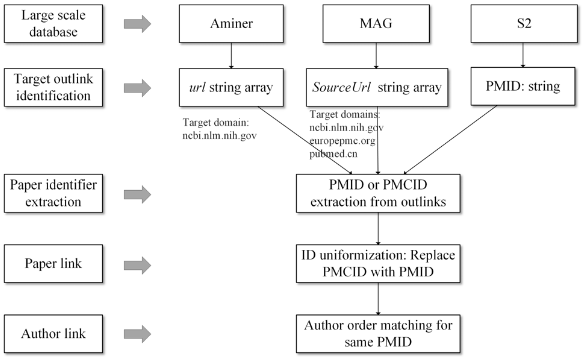

Fig. 1. An overview of Linking Approach, linking Papers and Authors of Aminer, MAG and S2 to PubMed.

TABLE I
NUMBER OF PAPER LINKS OF AMINER, MAG AND S2 TO PUBMED

	# Records link to PubMed	# One to one link[a]	# One to many link
Aminer	27,827,519	27,628,215	199,304
MAG	24,651,530	24,567,423	41,765
S2	30,453,745	30,453,745	0

[a] Note that “One to one” indicates one external paper ID matches a single PMID, and “One to many” indicates one external paper ID matches multiple PMID. These “One to many” paper links were excluded from the following analysis due to ambiguity.

TABLE II
NUMBER OF AUTHOR LINKS OF AMINER, MAG AND S2 TO PUBMED

	# Authors of linked papers	# Author links
PubMed	122,107,254	122,107,254
Aminer	109,371,465	89,466,556
MAG	98,146,543	94,368,705
S2	121,838,090	118,025,121

like “pubmed”, “europe” and “pmc”[3] were used to search in potential outlinks, we found only ”pubmed”, contained in full outlink point to the same domain ”ncbi.nlm.nih.gov”, could locate the target link for PubMed. Similarly, for MAG[4], we observed that their outlinks are mainly from “ncbi.nlm.nih.gov”, “europepmc.org” and “pubmed.cn”. For S2, we did not mine the target outlink, as PMID was carried with in S2 open corpus[5], which is an official identifier of paper in PubMed. Next, we extracted PMID from target outlinks, examples of frequent pattern of outlinks are demonstrated as follows: http://www.ncbi.nlm.nih.gov/pubmed/20313615, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762271/, http://europepmc.org/abstract/PMC/PMC1293358, http://europepmc.org/abstract/MED/13437259 and http://pubmed.cn/25673763. We used a regular expression to extract numbers from these URLs. It should be noted that extracted numbers are not limited to PMID, another alternative identifier PMCID – an official paper identifier for PubMed Central (PMC)[6], was also extracted. As declared by National Library of Medicine (NLM) in its official site[7], majority paper of PMC is included by PubMed, thus, most of PMCID are able to find a corresponding PMID. To match more papers for PubMed, we extracted 2,287,563 PMCID - PMID mappings by parsing PMC archive database[8]. We replaced extracted PMCID with PMID using these mappings. Finally, the paper links to PubMed can be mined from Aminer, MAG and S2. Table I describes the number of mined paper links, MAG and Aminer show a lower coverage of papers. Note that the versions of Aminer, MAG and S2 we used were all released baseline of 2019 year.

After mining the paper links, we mined author link for PubMed authors. Since the authors share the same order in the same paper indexed in different bibliographic databases was considered to be the same person, author ID allocated by other databases can therefore be assigned to the PubMed authors. The three databases also record author order along with the author ID. MAG has *AuthorSequenceNumber*, Aminer and S2 combine author order and author name into Json array format. It has been verified by us that the stored order in the Json array is in line with the nature order of the published paper. We mined author links for PubMed by examining whether two conditions are fulfilled simultaneously: the same PMID and the same author order, then author IDs were matched based on the author link. A statistical results of author link are shown in Table II. Compared to Aminer and MAG, S2 contains the maximum number of authors that link to PubMed.

III. EVALUATION

In this section, we evaluated the performance of matched author IDs. At the time of writing this paper, we did not find a study that mined author IDs from three large databases (Aminer, MAG and S2) for entire PubMed. It was unclear the extent to author ID correctness. Due to this, we measured the performance of these author IDs. A few evaluation datasets for AND are publicly available. In recent years, manually labeled gold standard datasets are available: SONG [4] and GS [5], both of them were carefully curated. To determine whether two similar names point to the same individual in SONG, multiple iterations were used in the annotation process to ensure quality. GS is the first unbiased gold standard database based on 1900 publication pairs from PubMed, the gold standard shows close similarity to MEDLINE in terms of chronological distribution and information completeness (e.g., coverage of East Asian last names). It is worth mentioning that the disambiguated dataset Author-ity [13] is available for academic use, our evaluation did not include it, since the latest version of Author-ity does not contains any paper published after 2009 year and

[3] “europe” was used to identify whether the source links are from Europe PMC, which is an open science platform. “pmc” is abbreviation of PubMed Central.

[4] https://www.openacademic.ai/oag/

[5] http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/download/

[6] https://www.ncbi.nlm.nih.gov/pmc/

[7] https://www.nlm.nih.gov/bsd/difference.html

[8] ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/

TABLE III
PERFORMANCE OF AUTHOR NAME DISAMBIGUATION OF AMINER, MAG AND S2 ON SONG DATASET

Database	ACC (%)	P (%)	R (%)	F1 (%)
Aminer	87.38	94.25	20.44	33.59
MAG	96.13	94.42	78.35	85.64
S2	93.73	93.07	65.57	76.94

TABLE IV
PERFORMANCE OF AUTHOR NAME DISAMBIGUATION OF AMINER, MAG AND S2 ON GS DATASET

Database	ACC (%)	P (%)	R (%)	F1 (%)
Aminer	41.62	97.86	17.15	29.18
MAG	71.76	99.07	56.32	71.81
S2	85.16	94.01	81.77	87.46

up-to-date disambiguated dataset is not available. A survey [14] suggested the Author-ity has a very restrictive licence and is not comparable in terms of its availability. For SONG, authors are organized as groups according to last name, then multiple iterations of disambiguation are performed to further divide a author group into different author sets. the authors within the same set are the same individuals and the authors across different sets are different individuals. This dataset, has 385 author sets among 36 author groups from 2,875 publications. For evaluation, we transformed the form of group organized into the form of author pair organized like Vishnyakova [5] did. In doing so, any two authors belonging to the same author set are enumerated as the same authors (positive samples), and any two authors belonging to different author sets but within the same author group are enumerated as different authors (negative samples). After transformation, 28,925 positive samples and 154,765 negative samples were generated. For GS, which contains 1,900 pair wise samples, we found 10 samples were not correctly labeled, we removed them from the dataset, leaving 1,202 positive samples and a 688 negative samples. Note that all instances in SONG dataset are the first author of a specific paper. It is not clear the author order of GS dataset. We used the following strategies to match author order: The two instances that come from GS and PubMed share the same PMID and same last name and same name initials are considered to be the same individual, identical PMID and last name and name initials is used as a condition for author order matching for GS. Note that not every instances in the GS was matched with a author order, 161 mismatched instances were removed.

Table III and Table IV demonstrated the evaluation results of AND on two datasets. S2 and MAG outperformed Aminer by a large margin. A lower recall of positive samples on two datasets both suggested that, two papers written by a same author were more frequently determined as being from different authors.

IV. CONCLUSION

This paper identified the problem of lack of publicly available author ID of PubMed, we addressed this by associating three well-known bibliographic databases, outlinks of them that point to PubMed were extracted for mining paper and author links. Author disambiguation performance of these author IDs was evaluated on two PubMed-related datasets. Our finding is that, Semantic Scholar covers more portions of PubMed than Aminer and MAG. Evaluation result of author disambiguation suggested that, although S2 and MAG performed better than Aminer on two datasets. The best F1 score is under 90%, indicate that more efforts are being need for author disambiguation research. Although AND has been studied for many years, little work has been done to create author IDs for entire PubMed. We have made this dataset publicly available for academic and industrial use. We believe that our dataset could facilitate author identities-based researches, especially for those tasks that have a strong need for author ID but not requiring very high precision. In the future work, we intend to investigate some methods to improve performance of incremental author name disambiguation, which may be more challenging as a large number of literature continuously pouring into bibliographic databases.

REFERENCES

[1] M. J. Lerchenmueller and O. Sorenson, “Author disambiguation in pubmed: Evidence on the precision and recall of author-ity among nih-funded scientists,” *PLoS One*, vol. 11, no. 7, p. e0158731, 2016.

[2] D. K. Sanyal, P. K. Bhowmick, and P. P. Das, “A review of author name disambiguation techniques for the pubmed bibliographic database,” *Journal of Information Science*, p. 0165551519888605, 2019.

[3] W. Liu, R. Islamaj Doğan, S. Kim, D. C. Comeau, W. Kim, L. Yeganova, Z. Lu, and W. J. Wilbur, “Author name disambiguation for p ub m ed,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 765–781, 2014.

[4] M. Song, E. H.-J. Kim, and H. J. Kim, “Exploring author name disambiguation on pubmed-scale,” *Journal of informetrics*, vol. 9, no. 4, pp. 924–941, 2015.

[5] D. Vishnyakova, R. Rodriguez-Esteban, and F. Rinaldi, “A new approach and gold standard toward author disambiguation in medline,” *Journal of the American Medical Informatics Association*, vol. 26, no. 10, pp. 1037–1045, 2019.

[6] K. Kim, A. Sefid, B. A. Weinberg, and C. L. Giles, “A web service for author name disambiguation in scholarly databases,” in *2018 IEEE International Conference on Web Services (ICWS)*. IEEE, 2018, pp. 265–273.

[7] A. M. Harrison and A. M. Harrison, “Necessary but not sufficient: unique author identifiers,” *BMJ innovations*, vol. 2, no. 4, pp. 141–143, 2016.

[8] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, “Arnetminer: extraction and mining of academic social networks,” in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 990–998.

[9] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, “An overview of microsoft academic service (mas) and applications,” in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 243–246.

[10] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha *et al.*, “Construction of the literature graph in semantic scholar,” *arXiv preprint arXiv:1805.02262*, 2018.

[11] L. Peng, S. Shen, D. Li, J. Xu, Y. Fu, and H. Su, “Author disambiguation through adversarial network representation learning,” in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[12] K. Kim, S. Rohatgi, and C. L. Giles, “Hybrid deep pairwise classification for author name disambiguation,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2369–2372.

[13] V. I. Torvik and N. R. Smalheiser, “Author name disambiguation in medline,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 3, pp. 1–29, 2009.

[14] M.-C. Müller, F. Reitz, and N. Roy, “Data sets for author name disambiguation: an empirical analysis and a new resource,” *Scientometrics*, vol. 111, no. 3, pp. 1467–1500, 2017.