Protocol

Authorship Weightage Algorithm for Academic publications: A new calculation and ACES webserver for determining expertise.

Wei-Ling Wu[1] , Owen Tan[1], Kwok-Fong Chan[1], Nicole Bernadette Ong[1], David Gunasegaran[1], Samuel Ken-En Gan[1],*

[1] Antibody & Product Development Lab, EDDC-BII, A*STAR; Wu_Weiling@eddc.a-star.edu.sg; ozruen@gmail.com ; kwokfong29@gmail.com ; 1701734D@student.tp.edu.sg; Davidsegar2@gmail.com (left to right)

* Correspondence: samuel_gan@eddc.a-star.edu.sg; samgan@apdskeg.com

Abstract: Finding experts in any field can be difficult, especially when relying on academic publications given the use of jargons despite publication lists being publicly available. Within the use of publications, discernment of expertise by authorship positions is often absent in the many publication-based expert search platforms available which could function as another discernment filter of expertise. Given that it is common in many academic fields for the research group lead or lab heads to take the position of the last author, the existing authorship scoring systems that assign a decreasing weightage from the first author would not reflect the last author correctly. To address these mentioned problems, we incorporated natural language processing **(**Common Crawl using fastText) to identify related keywords for a search compatible to using jargons as well as an authorship positional scoring with the option to provide greater weightage to the last author. The resulting output is a ranked scoring system of researchers upon every search which we implemented as a webserver for internal agency use called 'APD lab Capability & Expertise Search (ACES)' webserver which can be accessed from webserver.apdskeg.com/aces.

Keywords: Authorship; Natural Language Processing; Score; Weightage

1. Introduction

The surge of research article publications in recent years and the use of jargon often make it challenging for those not within the field to navigate through them to find and discern experts. Yet, finding *bona fide* experts continues to grow in importance in a world of ever-increasing misinformation and fake news that can lead to serious consequences [1] as observed during the COVID19 pandemic [2]. The need to identify the right experts in the many public/private sectors for consultancy and collaborations can make the difference between success and failure. However, the definition of 'expert' is contentious, varying from experience (time in field) to the very nuanced measurement of achievements that can include social impact, journal impact factors, citations, patents, commercialized products among others that have different values to the various needs. For academic uses, the work here will focus on research publications which are publicly available for easier filtering for expertise. Although publicly available, navigating through them is not easy, especially when considering the authorship issues discussed in previous articles [3-6].

Currently, there is a plethora of online platforms available that help identify experts based on research publications such as ResearchGate, Google Scholar, Web of Science, ORCID, Publons, among others. To an extent, this includes LinkedIn, which although not publication based, targets a broad user base beyond academia. While expert identification based on publications is convenient, existing caveats exist in that the contribution of co-authors are typically assumed to be equal in these scoring/ranking platforms [7]. A contentious topic that is still debated in academia [7,8], authorships are originally intended to reflect the level of contribution by the authors to the research [8-12] and would therefore be a good discernment filter in determining expertise. In fact, authorship is one of the key issues discussed in collaborations as well as dedicated resources providing advice to younger scientists on this area [13].

Many scoring and ranking methods as those in the existing platforms lack features to differentiate highly collaborative academics with only middle-authorships from high performing academics with similar number of publications but as mostly first or last-authors. Although discerning eyes could quickly differentiate that the latter is more likely to have higher expertise than the former, this discernment is not reflected in the current metrics nor conveniently accessible in the current platforms.

To include authorship as an added discernment filter, we incorporated a new scoring method that credits the last author as much as the first, and implemented it as a web server for assessment. Incorporating a common Crawl using fastText methods to group search input keywords, the problem of jargons usage is mitigated through relate word search. The resulting suite complements our other digitalisation efforts [14] to build an expert identification system to be used by journals and grant offices to identify suitable peer reviewers. As a demonstration, this authorship and natural language processing keyword search for expertise is currently based on the 2018 Google Scholar data of researchers from the Agency for Science, Technology, and Research, Singapore and incorporated as the APD lab Capability and Expertise Search (ACES) webserver.

2. Materials & Method

2.1. Scoring Method

The default scoring method for publication N co-authors uses the harmonic authorship credit model previously described by Hagen [15] that gives the highest weight age to the first author followed by the subsequent authors. The position of the author of interest (p) is determined from the left of the author list and is an element of positive real numbers (1, 2, 3, 4…).

$$\mathrm{score} = \frac{\frac{1}{\mathrm{p}}}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\mathrm{i}}} \tag{1}$$

With selection of the option to credit the last author and second last with equal weightage as that of the first and second respectively, the author score would be calculated using equations (2) to (5) represented by the position of the author of interest (p) in the (i) first or last position:

$$\mathrm{score} = \frac{1+\frac{1}{2}}{2\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\mathrm{i}}\right)} \tag{2}$$

(ii) second or second last position:

$$\mathrm{score} = \frac{1+\frac{1}{2}+\frac{1}{3}}{6\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\mathrm{i}}\right)} \tag{3}$$

(iii) any position between the second and second last position of the author list:

$$\text{score} = \frac{\frac{1}{p+2}}{\sum_{i=1}^{N} \frac{1}{i}} \tag{4}$$

and an exception for (iv) N = 3 where the author of interest is in the middle position.

$$\text{score} = \frac{\frac{1}{3}}{\sum_{i=1}^{N} \frac{1}{i}} \tag{5}$$

2.2. WebServer Design

ACES is a web application developed using Hypertext Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript for the front-end, and Python version 3.7.5 using Flask version 1.1.1 framework for the back-end server.

When the user first keys in a subject keyword on the main page, ACES retrieves relevant publications in the database by relevance and assigns a score to the publication based on distance to the keyword, followed by tabulating the scores of the researcher based on authorship position. The results are then displayed showing the ranked scores of the relevant researchers. The total number of publications, citations and most related publications pertaining to the search query are then displayed under each author.

2.3. Database

The sampling of 700 public Google Scholar profiles of researchers affiliated with the Agency for Science, Technology, and Research (A*STAR), Singapore were collected in 2019 and stored in a NoSQL database. The back-end server receives the data in JavaScript Object Notation (JSON) format containing the researchers' and their publication information of: names, affiliations, total number of citations and publications, research article titles with the authors list, and 30 most important keywords associated with the researcher. Applying the term frequency–inverse document frequency (tf-idf), a numerical statistic to determine the importance of a word in a given document [16] onto the publication titles as a whole, relevance to keywords was computed.

2.4. Search

The search algorithm consists of three main phases: query processing, researcher retrieval and ranking. It utilizes the Natural Language Toolkit (NLTK) version 3.4.4 and Gensim version 3.8.0 libraries.

2.4.1. Query Processing

Figure 1. Main page of the website with the option to check the 'last author corresponding' option.

Upon activating the search (see Figure 1), the search query processing phase starts by filtering out punctuation marks and stop words i.e. insignificant words that appear frequently in English that typically frame sentences [16]. Precompiled stop words were supplied by NLTK and further modified to enable stringent keyword filtering and increase the accuracy of the search query.

2.4.2. Retrieval of Researchers

ACES scores the relevance of the processed search query in keywords of every researcher. Using Gensim NLP library (https://radimrehurek.com/gensim/), two-million-word vectors with 300 dimensions were loaded on the back-end server. These word vectors were trained on Common Crawl (https://commoncrawl.org) using fastText [17], and are used to map the processed query to its corresponding values.

By computing the cosine similarity between word vectors of the search query and the 30 most important keywords of each researcher, a relevancy score is calculated. For faster result output, a batch of the top 20 high scoring researchers based on keywords are first shown.

2.4.3. Ranking

The ranking phase comprises of three steps: word processing, article scoring and the ordering of researchers. The publication titles are processed in the same way as the search query, where the punctuation marks and stop words are filtered out before the cosine similarity between the word vectors are computed. This is followed by the authorship computation which utilizes the harmonic allocation method where the highest score is assigned to the first author and decreases for each subsequent author in the list [15].

Excluding publications with low relevance to the query keyword, overall scores are assigned to the researcher based on the sum of relevance and authorship position scores. The top 20 researchers are ranked based on the combined scores and displayed on the webpage, together with their top five relevant publications and other information available in the database. Publications of low relevance are greyed out, allowing the user to focus only computationally deemed relevant ones (see Figure 2). The researcher names and publications are shown as hyperlinks to a Google search for easily follow-up and twenty researchers are displayed per page.

Figure 2. Example of a researcher in the results page.

3. Results and Discussion

In this work, we built the ACES web server as a resource for discerning and finding academic experts that considers publication authorship positions. While there are existing algorithms such as the harmonic authorship credit model that score researchers based on the authorship positions, the last author is mis-credited. This is especially so where in many academic disciplines, the research group lead and in some cases, the person who conceptualized and directed the research, especially in PhD student publications, is the last author. In cases of collaborations between groups where the other research group senior author occupies the second last position, we assigned weightage similar to that of the second author. Beyond authorship to facilitate the search of jargons in publication titles, a natural language processing toolkit for word matching of the search keyword was added alongside other utility features.

3.1. Exact word matching and Autocorrect

The search algorithm in ACES excludes phrases or words with ambiguous meanings to prevent confounding the scoring. For exact word search, users can utilize double quotation marks to filter out publication titles.

Since typographical/spelling errors are inevitable we incorporated a 'suggested word' feature using the Python package autocorrect version 0.4.4. Although a search with the error would still be performed, an autocorrect suggestion below the search bar will be displayed with the the results (see Figure 3) to which a new search with the suggested correct word can be re-performed by clicking on the hyperlink.

Figure 3. Autocorrect suggestion for the misspelled word of 'cancer'

3.2. Deep Search

ACES was designed to provide nearly instantaneous results with a slight compromise of accuracy. For higher accuracy, a deep search feature was added (see Figure 1 & 3) to skip the phase of fetching just the top 20 researchers based on their keywords, but to perform matching throughout the whole database. While this will take a slightly longer time to compute the results, all the researchers in the database would be computed**.** We have incorporated this default fast search given that the current small database of about 700 researchers is likely to grow by many folds when implemented towards a global use than from our current internal usage.

3.3. Authorship contribution

The reflection of contribution by authorship positions varies in different research fields [11,12], making the harmonic allocation [15] method insufficient for many academic fields where the last author position is the principal investigator or research lead who can play as important if not more than the first author [12]. To allow for better credit to last authors, a checkbox below the search bar labelled 'Last author corresponding' (see Figure 1) was included as an option. This allows the field-specific weightage given to the last author a high score equal to the first author (Table A1) and the second-last as the second. This is made possible via a modified version of the default harmonic allocation. Hence, this option will yield different ranking results for researchers with more last author publications. While it remains contentious whether the last authors should have more or equal credit to the first authors, we have attributed equal corresponding weightage for ease of implementation. Nonetheless, this option should be used only for the relevant fields to prevent confounding the outcomes.

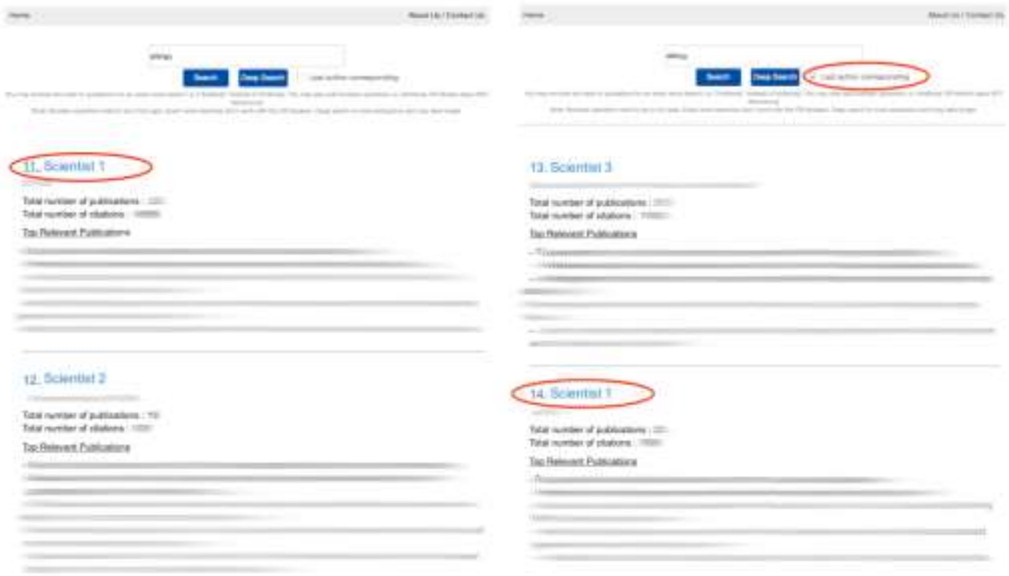

Figure 4. Search result display of "allergy" before (left) and after (right) last author option is selected showing different ranking when considering last authorship.

Since the general user is likely to focus only on the top few ranked individuals for expertise, the order can matter. As a blinded demonstration (Figure 4), when applying the last author weightage, there would be a reordering of 'Scientist 1' which may have more middle-author publications than the other preceding researchers. While the movement here is perhaps small, such reshuffling could be fairly notable in a larger list, demonstrating the relevance of last author weightage scoring as an added discernment filter.

We acknowledge that the scoring method is not comprehensive and that there remains many variations in last authorship and the possibility of slighting the first authors further as well as the even more contentious second last authorship which may not reflect a senior author in many cases. These problems arise from the lack of a universal standard and nuances are likely to persist alongside the definition of 'expertise' that is also dependent on the very different need of different groups. Yet, in the meantime, for the purposes of expert identification within academic settings, the weightage adjustment for the last author may allow new upcoming research leads better visibility.

The current ACES website is available on webserver.apdskeg.com/aces on both desktop and smartphone browsers. A video demonstrating the use of ACES is shown in https://fb.watch/57DYA4KBrj.

5. Conclusion

ACES demonstrates the incorporation of NLP for jargon search and a modified last-author scoring method for identification of academia experts from their publications. With first and last-author scorings, due credit is also given to research leads, and provides an added layer of discernment. Together, ACES allows for a search for experts based on publicly available lists of publication. With added features of deep search, it can be easier for those outside the specific fields to find experts for business consultancy and collaborators in business, public health, and research.

Author Contributions: **Software:** WLW, KFC, OT, NBO and DG. **Supervision:** WLW, KFC and SKEG. **Data curation:** KFC. **Conceptualization:** SKEG

Funding: This research was funded by A*STAR Office of Grants Administration (OGA).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable as the data is from public domain.

Data Availability Statement: ACES belongs to the Agency for Science, Technology, and Research of Singapore, and is for internal use. Interested parties may write to the corresponding author who may make the request to the relevant departments.

Acknowledgments: We acknowledge Ms Ei-leen Tan of the A*Office of Grants Administration, Science and Engineering Research Council (SERC) grant office for the funding of this work. SKEG thanks Ebenezer for support behind the scenes.

Conflicts of Interest: There is no conflict of interest.

References

1. Salvi, C.; Iannello, P.; Cancer, A.; McClay, M.; Rago, S.; Dunsmoor, J.E.; Antonietti, A. Going Viral: How Fear, Socio-Cognitive Polarization and Problem-Solving Influence Fake News Detection and Proliferation During COVID-19 Pandemic. Frontiers in Communication 2021, 5, doi:10.3389/fcomm.2020.562588.
2. Teovanović, P.; Lukić, P.; Zupan, Z.; Lazić, A.; Ninković, M.; Žeželj, I. Irrational beliefs differentially predict adherence to guidelines and pseudoscientific practices during the COVID-19 pandemic. Applied Cognitive Psychology 2021, 35, 486-496, doi:https://doi.org/10.1002/acp.3770.
3. Cho, M.; McKee, M. Authorship in biomedical research—realities and expectations. Science's Next Wave. 2007.
4. Gasparyan, A.Y. Authorship and contributorship in scholarly journals. J Korean Med Sci 2013, 28, 801-802, doi:10.3346/jkms.2013.28.6.801.
5. Sharma, H.; Verma, S. Authorship in biomedical research: A sweet fruit of inspiration or a bitter fruit of trade. Trop Parasitol 2018, 8, 62-69, doi:10.4103/tp.TP_27_18.
6. Harsoor, S.S. Art and science of authorship for biomedical research publication. Indian J Anaesth 2016, 60, 679-683, doi:10.4103/0019-5049.190626.
7. Vavryčuk, V. Fair ranking of researchers and research teams. PloS one 2018, 13, e0195509.
8. Venkatraman, V.; Arzbaecher, R.; Marušić, M.; Marušić, A. Conventions of scientific authorship. Science 2010, 12, doi:10.1126/science.caredit.a1000039.
9. Ballantine, J.; Guo, X.; Larres, P.; Yu, M. Chinese authorial identity: a model for scoring the Student Authorship Questionnaire. Studies in Higher Education 2018, 43, 2385-2397, doi:10.1080/03075079.2017.1333493.
10. Boyer, S.; Ikeda, T.; Lefort, M.-C.; Malumbres-Olarte, J.; Schmidt, J.M. Percentage-based Author Contribution Index: a universal measure of author contribution to scientific articles. Research Integrity and Peer Review 2017, 2, 18, doi:10.1186/s41073-017-0042-y.
11. Patience, G.S.; Galli, F.; Patience, P.A.; Boffito, D.C. Intellectual contributions meriting authorship: Survey results from the top cited authors across all science categories. PLoS One 2019, 14, e0198117, doi:10.1371/journal.pone.0198117.
12. Tscharntke, T.; Hochberg, M.E.; Rand, T.A.; Resh, V.H.; Krauss, J. Author sequence and credit for contributions in multiauthored publications. PLoS Biol 2007, 5, e18, doi:10.1371/journal.pbio.0050018.
13. Gaffey, A. Determining and negotiating authorship. Psychological Science Agenda 2015, 29.
14. Ong, N.B.; Wu, W.-L.; Chan, K.-F.; Gan, S.K.-E. Application Notes: AI-based Research Grant Audits-A* Grant Audit Flagging System (A* GAFS). 2020, doi:10.30943/2020/09012020.
15. Hagen, N.T. Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis. PLoS One 2008, 3, e4021, doi:10.1371/journal.pone.0004021.
16. Leskovec, J.; Rajaraman, A.; Ullman, J.D. Mining of massive data sets; Cambridge university press: 2020.
17. Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; Joulin, A. Advances in Pre-Training Distributed Word Representations. ArXiv 2018, abs/1712.09405.