Peer2ref: a peer-reviewer finding web tool that uses author disambiguation

Miguel A. Andrade-Navarro\textsuperscript{a}, Gareth A. Palidwor\textsuperscript{b}, Carol Perez-Iratxeta\textsuperscript{b,*}

\textsuperscript{a} Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany.

\textsuperscript{b} Ottawa Hospital Research Institute, 501 Smyth Road, Ottawa, Ontario K1H 8L6, Canada.

*Corresponding author:
Carol Perez-Iratxeta
501 Smyth Road,
Ottawa. ON K1H8L6.
Canada
Tel: 1-613-737-8899 (x 79525)
email: cperez-iratxeta@ohri.ca
Abstract

Background:

Reviewer and editor selection for peer review is getting harder for authors and publishers due to the progressive growth of the body of knowledge that carries specialization onto narrower areas of research. Examination of the literature facilitates finding appropriate reviewers but is time consuming and complicated by author name ambiguities.

Results:

We have developed a method called peer2ref to support authors and editors in selecting suitable reviewers for scientific manuscripts. Peer2ref works from a text input, usually the abstract of the manuscript, from which important concepts are extracted as keywords using a fuzzy binary relations approach. The keywords are searched on indexed profiles of words constructed from the bibliography attributed to authors in MEDLINE. The names of these scientists have been previously disambiguated by coauthors identified across the whole MEDLINE. The methods have been implemented in a web server that automatically suggests experts for peer-review among scientists that have authored manuscripts published during the last decade in more than 3,800 journals indexed in MEDLINE.

Conclusion:

peer2ref web server is publicly available at http://www.ogic.ca/projects/peer2ref/hurrengoa.html.

Keywords

- Publishing
- Information Storage and Retrieval
- MEDLINE
- Peer Review, Research
- Natural language processing
**Background**

Effective expert selection for anonymous peer review is a critical step in the process of publishing research in referred scientific journals, which remains the most important platform for the dissemination of scientific knowledge. Demand for peer review is increasing as the number of researchers, journals and publications increases. In the field of biomedicine alone, the estimated number of titles currently published stands between 13,000 and 14,000, of which 5,300 are indexed in the MEDLINE database (Journals database at the NCBI, [1]).

Upon successful submission of a manuscript for publication in a journal, editors attempt to quickly identify suitable reviewers, sometimes with the assistance of authors who may have been prompted to provide several suggestions. This procedure demands a fair knowledge of the experts in the manuscript’s area of knowledge from both authors and editors. Inevitably, authors and editors will concentrate their requests on a small set of referees, usually senior authors that they know and trust. As a result, senior authors are overloaded with demands.

The bibliography offers a resource to find experts. However, the increase in the rate of production of new research makes it increasingly difficult to track all the publications coming out from even narrow fields of research and many authors that could potentially be good reviewers may not be requested. An approach to ease this problem is the development of computational methods to assist authors and editors in reviewer selection based on the literature. Such methods have the
potential to facilitate the task and should produce less biased and more systematic expert selections than manual protocols.

Ideally, we would like the computer to point to potential reviewers for a given manuscript using just the manuscript content as input. One straightforward strategy in this direction is to search the database of (peer-reviewed) scientific literature for the most similar documents to the manuscript we want to review, and then suggest the authors of these documents as experts. A widely used measure of document similarity is the cosine between the abstracts of the documents encoded as vectors of the frequency of words they contain [2]. For example, BioMed Central editorial uses this approach by proposing to the associate editor that would handle the submission the cosine values between the abstract of the submitted manuscript and abstracts referenced in MEDLINE.

There are two web server applications to find similar abstracts in MEDLINE. An early one is eTBLast [3], which uses the same principle with a more elaborated measure of text similarity that takes into account word frequencies and word order in the text. A more recent one is Jane [4], a straight implementation of Lucene’s [5] MoreLikeThis algorithm, which does not take into account the words’ order but their relevance according to their frequencies in a whole corpus.

Here we propose a more comprehensive approach to computational selection of peer reviewers, which relies on comparison of a word profile of the manuscript not to that of other single manuscripts but to the collection of manuscripts authored by each potential peer reviewer. This approach necessitates building word profiles for authors. However, the problem that considerably complicates
the matter of identifying the manuscripts authored by one individual, especially in an automatic way, is that many authors share last names and initials with other authors. Particularities of names across countries further complicate the issue. For example, most Chinese last names are extremely common, with the eleven most common being shared by about 40% of the Chinese population, yet their wide variety of given names is lost by Western abbreviation practices [6]. Ambiguity due to given name abbreviation is a problem that affects other Asian scientists as well [7-8] in a manner well beyond the matter discussed here [9].

Dealing with author name ambiguity remains a hard problem. For the biomedical community, an obvious and ideal solution would be to have each author assigned a unique identifier in MEDLINE upon their first publication. However, this solution has no trivial implementation, as it would require the combined effort from a coordinator organization, such as the NCBI, and the whole body of the scientific community. Even if implemented today, this would not resolve the name ambiguity of the large body of prior literature.

Meanwhile, the problem is only worsening due to the ever-increasing number of scientists. Computational efforts, mostly industry-led, are being made by implementing algorithms that, by parsing MEDLINE, would partially address this matter. Some initiatives combine registration of users with profile generation and their degree of integration with companies and accessibility to their methods and data are very heterogeneous. ResearcherID from Thomson Reuters (http://www.researcherid.com/) is a company resource linked to other Thomson databases such as ISI Web of Knowledge. BioMedExperts from Collexis
(http://www.biomedexperts.com/) contains a collection of automatically generated profiles for authors in MEDLINE based, according to their web site, on the concepts associated to the identifiers and on co-authors. Users can register and modify the profiles but the system is not fool-proof (personal observation). Author-ity is an academic effort to generate author profiles [10] and offers a database of disambiguated author names in MEDLINE for download and a web interface to query it (http://arrowsmith.psych.uic.edu/arrowsmith_uic/author2.html).

Among the strategies to disambiguate authors that share the same name are the use of keywords that identify a particular subject of research, collaborators co-signing publications with the authors (networks of collaborators), physical location extracted from the affiliation data usually complemented with the years of publication, journal subject class (i.e. journals in the area of cardiology), and even co-citations in web pages [11].

Here, we have chosen to implement an approach based on co-authorship because it is straightforward, and can be easily applied in an unbiased manner to every single name. Accordingly, we disambiguated every author name in MEDLINE by co-authors and assigned different identifiers to each disambiguated instance. Next, for each identifier we derived profiles of keywords extracted from the abstracts of the references in MEDLINE associated to it. Given a manuscript, our method uses these profiles to suggest peer-reviewers based on the similarity to the keyword profile deduced from the manuscript.
Methods

Author name disambiguation

A major complication to identifying referees for peer review in an automatic manner is that many MEDLINE authors share both last name and initials. To address this hurdle in our expert finding tool, we have implemented a straightforward algorithm to disambiguate MEDLINE author names using coauthor names. The procedure was applied to every name in MEDLINE, defined as “last name_initials” in MEDLINE’s field [Author] and hereafter called simply “name”, and produced disambiguated profiles that consist of the set of references attributed to an author identifier. The algorithm is as follows. Given a name, we gather all MEDLINE records with that name in the author list and look for the most frequent coauthor name in the set. The subset of references with both the name of the author and most frequent coauthor is assigned to an identifier and constitutes a disambiguated author’s profile. Next, we do the same for the next most frequent coauthor name in the remaining subset of references. This procedure continues until we do not find more name repetitions among coauthors. The remaining references are singletons with different identifiers assigned to them. The application of this algorithm to MEDLINE’s 2010 release of references authored by more than one scientist resulted in 11,394,787 disambiguated authors. For a number of disambiguated authors we created a profile that describes their expertise by extracting all nouns from the abstracts of their associated references. We did this for those with at least an associated manuscript as first or last author published from 2000 to 2010 (MEDLINE release
2010) in one of more than 3,800 journals indexed in MEDLINE. The complete list of journals is available in the tool's website. As a result we obtained a set of 2,660,235 profiles, expectedly experts on a broad collection of subjects that covers most biological and biomedical research.

Keyword computation from abstracts

Given an abstract we assign to each of its words a keyword score that reflects its importance in the text. To do so we consider the relationships between all words in the text as deduced from a simple model of fuzzy binary relations [12], which is based on co-occurrence within sentences. Keywords tend to be frequent words that have more and stronger relations with other words. Details of the computation can be found in [13]. The more and stronger relations a word has, the higher its keyword score will be. Computations are done just over the nouns, after filtering out stop words since this works well for the classification of scientific text [14]. Nouns and sentences are detected using a standard grammatical tagger (TreeTagger; Helmut Schmid, IMS, Stuttgart University http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/).

Results

Using disambiguated author word profiles for referee selection

We propose that the adequacy of a disambiguated author as referee of a given manuscript can be measured in terms of the similarity between their word profile and that of the abstract of the query manuscript. To test this we designed the following benchmark. Given a manuscript, we reasoned that, in general, its last (corresponding usually to senior) author should be deemed an appropriate expert
to review it, and hence should have a profile more similar to the one of the manuscript than authors of other manuscripts. In view of that, we constructed a benchmark dataset by randomly selecting 200 manuscripts from *BMC Bioinformatics* from the period 2001-2005. Last authors were identified, and for each of these authors a word profile was produced using the set of manuscripts assigned to their disambiguated profiles but excluding the manuscript included in the benchmark dataset. For each of the 200 abstracts in the benchmark, we ranked all 200 last authors according to the cosine distance between the word profile vectors of author and abstract and recorded the position of the abstract’s last author. This gave us an average ranking of 29.5 (1 as the best possible ranking), with 63 correct identifications. We repeated the analysis using profiles of non-disambiguated authors, which resulted in an average ranking of 48 (Figure 1, left panel), and only 22 correct identifications. In general, the disambiguated profiles resulted in greater cosine similarities between manuscript and author profiles than the non-disambiguated ones (Figure 1, right panel), which partially explains the larger discriminative power of the disambiguated profiles over the non-disambiguated profiles. We note that this benchmark was relatively hard because as all manuscripts were selected from the same journal and period, there was some overlap between the topics dealt with in those manuscripts and some of these authors could well be appropriate referees of each other’s publications.
Web tool implementation

Our algorithm has been implemented as the web tool peer2ref. To run peer2ref users have to paste some text from their manuscript in the input page from which at most 50 keywords with scores higher than 0.05 are used to build the profile of the manuscript. Usually a well written abstract is enough to properly reflect the content of a manuscript. However, an unstructured or too short abstract will not provide the necessary keywords for a successful search. In these cases, as well as when the research subject is highly original, additional input from the main text may be helpful (see additional information in the web tool's Supplement section).

Next, users can optionally select the broad subject(s) of the manuscript. Appropriate reviewers will be selected among authors that have published in the journals associated with these subjects. This allows to target referees in the particular context of a narrower subject of interest, which may not be selected in a global context due to slightly lower scores. For example, a word such as “heart” would become less discriminating in the context of the subject “Cardiology” than in a global context. Subject selection may also be handy when looking for referees for multidisciplinary research, for example, when it is necessary to summon the expertise of both geneticists and computational biologists.

Classifying journals by subject is not a trivial problem. Here we are using a ready-made classification from the US National Library of Medicine which consists of 123 broad subjects.

Apache Lucene version 2.9.4, a Java based indexing and search library, was used to index the disambiguated author profiles for rapid searching. The resultant
index is large, approximately half the size of the source data, however searches using the index take only seconds. The 350 authors with the closest profiles to the query manuscript profile, according to the Lucene's *Similarity* method [5], are selected as potentially suitable referees, ranked by Lucene's scores and displayed in the results page, together with the most relevant keywords from their profiles (Figure 2).

PubMed links to the references associated with each profile allow a quick assessment of the suitability of the suggested candidates. The built-in search function of web browsers can be used to quickly find the highest ranked reviewers associated with a particular keyword. A table of journals where the resulting reviewers have published is linked from the results page. These may be considered suggestions of appropriate journals for submission of the query manuscript. In the Supplement section of the peer2ref website we have included a step by step example as well as further details and usage tips.

**Discussion**

Finding experts for anonymous peer review is becoming harder due to progressive specialization of scientists on narrow research fields and other issues, such as author name ambiguity. Computational methods can facilitate this task and additionally, introduce an element of fairness into the process of referee selection as the same objective test is applied to every author to determine their adequacy as reviewer.

We have developed a fully automatic method to provide reviewer suggestions for manuscripts using a small amount of representative text, expectedly their
abstracts. To address the problem of author name ambiguity, our method uses a pre-computed dataset of disambiguated authors deduced automatically from all references in MEDLINE. According to our benchmark, the effectiveness of author profiles derived from their associated manuscripts in identifying them as experts is far superior when considering their disambiguated profiles, indicating that disambiguated name identifiers point to more accurate information. Obviously, a more rigorous evaluation of the disambiguated profiles would consist of comparing a large enough set of automatically derived profiles to their corresponding manually curated versions. This would be extremely time consuming and out of the scope of this work. We are aware that, because the way it was produced, i.e. using solely coauthors, our database of disambiguated authors is highly atomized, meaning that, in general, several author profiles will correspond to the same scientist, giving a high precision-low recall disambiguation. However, given its intended purpose of finding experts in narrow areas of research, this feature of our algorithm is advantageous. Often, scientists have different sets of collaborators with whom they work on somewhat different matters. This results in authors publishing manuscripts in the same small area of research with the same group of coauthors. Our disambiguated profiles partially capture this effect.

To summarize, we have developed a fully automatic method to provide suggestions of reviewers for scientific manuscripts, using text from the abstract of the manuscript and disambiguated authors profiles deduced automatically from the whole MEDLINE. We have implemented it in a public web server to make it
accessible to authors and editors. In the future, we plan to further develop the
computation of similarity scores and to add further functionality to the web tool.

Peer2ref is available for public use at

http://www.ogic.ca/projects/peer2ref/hurrengoa.html

Competing interests

The authors declare that they have no competing interests.

Author's contributions

M. A. A.-N. and C. P.-I. conceived the method and developed it. C. P.-I. produced the
disambiguation and programmed the web server. G. A. P. implemented the Lucene
indexing and searching. All authors contributed to the manuscript.

Acknowledgments

We are grateful to Helmut Schmid for developing TreeTagger and making it
publicly available. We are grateful to the NCBI for their MEDLINE distribution and
in particular to Jane Rosov for her suggestion of using the Journals database. M.
A. A.-N. acknowledges funding from the Helmholtz Alliance on Systems Biology
(Germany).
References

1. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S et al: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2011, 39:D38-51.

2. Manning C, Schütze H: *Foundations of Statistical Natural Language Processing*. MIT Press Cambridge, MA; 1999.

3. Errami M, Wren JD, Hicks JM, Garner HR: e**TBLAST**: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Res*. 2007, 35:W12-15.

4. Schuemie MJ, Kors JA: Jane: suggesting journals, finding experts. *Bioinformatics* 2008, 24:727-728.

5. Hatcher E, Gospodnetić O: *Lucene in Action*. Manning Publications, Greenwich; 2005.

6. Xu Z, Nicolson DH: Don't Abbreviate Chinese Names. *Taxon* 1992, 41:499-504.

7. Puniamoorthy N, Jeevananthinee J, Kutty SN: Give south Indian authors their true names. *Nature* 2008, 452:7187.

8. Qiu J: Scientific publishing: identity crisis. *Nature* 2008, 451:7194.

9. Kurien BT: Name variations can hit citation rankings. *Nature* 2008, 453:7194.

10. Torvik VI, Smalheiser NR: Author Name Disambiguation in MEDLINE. *ACM Trans Knowl Discov Data*. 2009, 3:11.
11. Yang KH, Peng HT, Jiang JY, Lee HM, Ho JM: *Research and Advanced Technology for Digital Libraries* 2008, *5173*:185-196.

12. Miyamoto S: *Fuzzy sets in information retrieval and cluster analysis.* Kluwer Academics Publishers, Dordrecht; 1990.

13. Perez-Iratxeta C, Bork P, Andrade MA: *Computing fuzzy associations for the analysis of biological literature.* *Biotechniques* 2002, *32*:1380-1385.

14. Shah PK, Perez-Iratxeta C, Bork P, Andrade MA: *Information extraction from full text scientific articles: where are the keywords?* *BMC Bioinformatics* 2003, *4*.
**Figure legends**

**Figure 1:** (Left panel) Distribution of the rank of last authors in the benchmark. (Right panel) Distribution of cosine similarity between word profiles of last authors and their manuscripts. Red: disambiguated profiles. Blue: not-disambiguated profiles. For full data see Supplement Data table *benchmark_results.xls.*

**Figure 2:** Snapshot of the results obtained by peer2ref. A step by step example is available at the web tool site.

**Supplementary Data**

File name (Excel table .xls): *benchmark_results.xls:*

Title of Data: Table containing benchmark results.

Description: **PMID**: PMID identifier; **Disambiguated**: cosine value between the word profile vectors of last disambiguated author and benchmark abstract. **Non-disambiguated**: cosine value between the word profile vectors of last (non-disambiguated) author and benchmark abstract. **Rank disambiguated**: order in which the corresponding author ranked when using disambiguated profiles. **Rank non-disambiguated**: order in which the corresponding author ranked when using non-disambiguated profiles.
Figure 1
| score | author ID | author   | coauthor   | keywords                                                                 | abstracts |
|-------|-----------|----------|------------|--------------------------------------------------------------------------|-----------|
| 0.34  | J704529   | Jaenisch R | Meissner A | cell dna gene methylation stem histone chromatin genome pattern protein mouse promoter factor modification differentiation oct4 transcription state change cpg derivation expression chip nanog es | 25        |
| 0.326 | S428194   | Santos F  | Dean W     | methylation dna development cell modification genome chromatin embryo gene totipotency region role novo demethylation expression blastocyst stage protein origin mark pronucleus histone h3 domain dnmt3b | 12        |
| 0.325 | H158919   | Hanna J   | Jaenisch R | cell methylation dna gene genome factor stem histone mouse cpg expression pattern state modification differentiation representation promoter sequencing change island insight chromatin sequence pluripotency molecule | 11        |
Additional files provided with this submission:

Additional file 1: benchmark_results.xls, 23K
http://www.biodatamining.org/imedia/9489515935706228/supp1.xls