Journal of Biomedical Informatics 51 (2014) 8–14

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automatic generation of investigator bibliographies for institutional research networking systems

Stephen B. Johnson [a,*], Michael E. Bales [b], Daniel Dine [b,c], Suzanne Bakken [b,c], Paul J. Albert [d], Chunhua Weng [b,c]

[a] Department of Public Health, Weill Cornell Medical College, New York, United States
[b] Department of Biomedical Informatics, Columbia University, New York, United States
[c] The Irving Institute for Clinical and Translational Research, Columbia University, New York, United States
[d] Samuel J. Wood Library, Weill Cornell Medical College, New York, United States

ARTICLE INFO

Article history:
Received 13 December 2013
Accepted 20 March 2014
Available online 30 March 2014

Keywords:
Authorship
Bibliography as topic
MEDLINE
Natural language processing
Pattern recognition
Automated

ABSTRACT

Objective: Publications are a key data source for investigator profiles and research networking systems. We developed ReCiter, an algorithm that automatically extracts bibliographies from PubMed using institutional information about the target investigators.
Methods: ReCiter executes a broad query against PubMed, groups the results into clusters that appear to constitute distinct author identities and selects the cluster that best matches the target investigator. Using information about investigators from one of our institutions, we compared ReCiter results to queries based on author name and institution and to citations extracted manually from the Scopus database. Five judges created a gold standard using citations of a random sample of 200 investigators.
Results: About half of the 10,471 potential investigators had no matching citations in PubMed, and about 45% had fewer than 70 citations. Interrater agreement (Fleiss' kappa) for the gold standard was 0.81. Scopus achieved the best recall (sensitivity) of 0.81, while name-based queries had 0.78 and ReCiter had 0.69. ReCiter attained the best precision (positive predictive value) of 0.93 while Scopus had 0.85 and name-based queries had 0.31.
Discussion: ReCiter accesses the most current citation data, uses limited computational resources and minimizes manual entry by investigators. Generation of bibliographies using named-based queries will not yield high accuracy. Proprietary databases can perform well but requite manual effort. Automated generation with higher recall is possible but requires additional knowledge about investigators.

1. Introduction

One of the goals of the Clinical and Translational Science Award (CTSA) program is to create a virtual community of investigators across institutions and research domains [1]. Toward this end, a number of institutions are developing systems to characterize expertise, and to search for and match potential collaborators. VIVO is a network of profiles of researchers that includes publications, teaching, service, and professional affiliations [2]. *Digital Vita* is a social network that enables users to manage online profiles, curriculum vitae and biosketches [3]. Harvard Catalyst Profiles provides directory information and also illustrates how investigators are connected in the community [4]. Other systems include BiomedExperts and ResearchGate [5].

These systems integrate data from national databases, local databases and user input. Integration of databases is often challenging because no authoritative identifier for researchers exists connecting their publications, grants, patents, mentoring, service and teaching [6]. Publications are a key source of information about investigator expertise. A major obstacle to leveraging publication data is that authors do not have unique identifiers [7,8]. Such identifiers have important implications for determining the different roles of authors and how contributions to science are measured [9,10].

In response, a number of organizations are developing name disambiguation solutions. The International Organization for Standardization (ISO) is developing the International Standard Name Identifier (ISNI). Thomson Reuters Web of Knowledge currently offers ResearcherID, which enables an author to build an online

* Corresponding author. Address: Center for Healthcare Informatics and Policy Weill Cornell Medical Center 425 E 61st St, #317 New York, NY 10065, United States. Fax: +1 646 962 0105.
E-mail address: johnsos@med.cornell.edu (S.B. Johnson).

http://dx.doi.org/10.1016/j.jbi.2014.03.013
1532-0464/

publication list using search services [11]. Thomson Reuters and Nature Publishing Group initiated Open Researcher and Contributor ID (ORCID), a non-profit, central registry of unique identifiers with links to other current identity schemes [12]. Community of Science (COS) Pivot contains a database of profiles submitted by researchers and reviewed by a team of editors [13]. The National Institutes of Health help investigators make their publications available through My NCBI, and link investigators to their eRA Commons accounts [14].

Many of the above approaches rely heavily on the manual labor of individual researchers to perform searches, upload information or edit publication lists. To help reduce this effort, some databases employ automated disambiguation to separate author identities. For example, Elsevier's Scopus assigns a unique number to authors and groups all their documents using an algorithm that analyzes affiliation, publication history, subject area and coauthors [15]. Thomson Reuters' Web of Science performs a similar service. The limitation is that this process only includes authors whose documents are contained in their databases, which (with the exception of some documents such as those published in open access journals) can only be accessed by subscription. CiteSeer automatically acquires, parses and indexes publicly available articles, focusing primarily on computer and information science [16].

To tackle the ambiguity problem in PubMed, a group at the University of Illinois at Chicago developed Authority, which groups papers written by the same author into clusters [17–20]. While an interface is freely available online, the database is static, and is not updated as PubMed changes (the database may be requested for research purposes). Advanced methods such as random forests can achieve good results experimentally, but are not yet available for practical applications and may be computationally intensive [21].

Standards organizations, government agencies and publishers may eventually provide a solution to the author identification problem, but a solution is needed in the interim. This article offers an approach called ReCiter, a method that focuses on the biomedical domain, is freely available, works with changing PubMed content, and does not require extensive manual labor from investigators.

2. Material and methods

ReCiter generates custom bibliographies for a given set of investigators using a bibliographic database. This experiment reports a test of the ability of ReCiter to generate accurate and complete bibliographies for all investigators at Columbia University Medical Center. Below we describe each step in the algorithm, followed by evaluation on a random sample.

The input to the ReCiter algorithm (Fig. 1) is a database of investigators for whom we wish to collect citation data (e.g., faculty, students and research scientists at a given institution), which contains descriptive information (e.g., name and departmental affiliations). The algorithm identifies appropriate articles for each individual by matching information from the local database to a cluster of citations retrieved from a publication database.

2.1. Representation of target investigators

To generate bibliographies, ReCiter requires a list of target investigators, consisting at minimum of the full name of each individual. Ideally, the investigator database is an authoritative source (e.g. curated by a given institution), which ensures formatted data (e.g., components of names properly identified) and correct spelling. The ReCiter algorithm performs better when provided with additional information about each investigator, such as

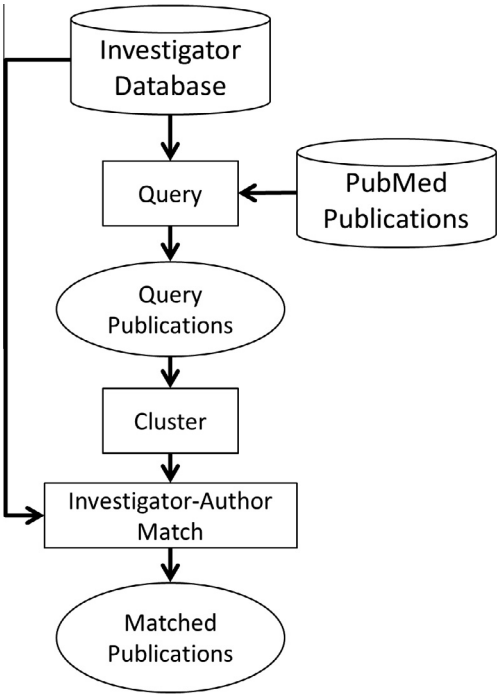

Fig. 1. Data flow of the ReCiter algorithm. Investigator names and departmental affiliations are selected from an institutional database; citations are extracted from PubMed using name-based queries; citations are clustered into separate identifies; the identities most closely matching the investigators are chosen.

departmental affiliations. ReCiter represents each target investigator using the same fields as a citation: authors, institution, journal, keywords, etc. This format makes it possible to supply detailed information about individuals when available, such as prior institutions and departments, alternate names (e.g., short variants of first name, or maiden name), frequent coauthors, and research keywords. In this study, we used the Columbia University human resource database as the source of potential investigators. Names were separated into first, middle and last; prefixes and suffixes were discarded (Dr., Jr., the Third) as were additional middle names. Only current employees were selected, and these were further restricted to faculty, research scientists, postdoctoral fellows and closely related titles. Graduate students were not included in this source. One or more current department affiliations were extracted for each investigator, but prior affiliations at other universities were not available from this source. Note that some individuals with certain job titles may not have any publications.

2.2. Querying citations

Reciter requires access to a bibliographic database that covers the broad research areas of the target investigators and provides information about each citation: authors, article title, institution, journal name, key words, etc. We chose to use PubMed for this study because it is freely available and has broad coverage of biomedical fields.

A custom, name-based query was created for each investigator. The most basic search strategy is to query by the investigator's last name and first initial. However, in some databases, this can return tens of thousands citations for common names. To improve efficiency, ReCiter can be provided with a cut-off number to limit search retrieval results. In this case, ReCiter uses a more restrictive search using the last name, first initial and middle initial. If this strategy still returns too many citations (or the investigator has

no middle name), the program restricts the query further using words extracted from the institutional affiliation.

The resulting set of citations typically contains most (or even all) of the target investigator's articles. However, the set may contain many "false positive" citations not written by the target investigator. For example, if the target investigator is Jane Smith and the search is performed with last name Smith and first initial J, citations for Jack Smith will also be retrieved and must be separated. The goal is to cluster the citations so that all of, and only, Jane's citations are in one cluster, and all of, and only, Jack's are in another.

2.3. Ordering citations by relevance to target investigator

ReCiter's key task is to determine which citations from the query are written by which individuals. Jane Smith's citations are clearly distinct from Jack Smith's, but a citation by J. Smith is ambiguous. There may be a Jane Smith at Harvard and another at Stanford, or a Jane A. Smith and a Jane B. Smith at the same institution.

To accomplish this task, ReCiter processes the citations one at a time, making a decision about the identity of each citation and then moving onto the next. The program sorts the citations from those having the most complete (e.g., having full names of authors) to least complete information. It first considers the most informative citations, postponing clustering the more ambiguous ones as long as possible.

Author's names in citation databases can differ in their completeness, making it more challenging to separate identity. For example, PubMed did not include the fully spelled out names of authors until 2002, so there is no direct way to distinguish Jack Smith from Jane Smith in an older citation with author J. Smith. Records may also lack information in other fields, such as institution, journal and keywords.

To address this, ReCiter assigns to each citation a completeness score. This score's most important factor is how well the name in the author list of the citation matches the target investigator. For example, if the target investigator is Jane Mary Smith, the most points would be assigned to a match on last name, full first name and full middle name; fewer points if an initial is used to represent either first name or middle name; and fewer still if there is neither a full first nor middle name. Additional points are added if the institution field is associated with the target investigator (in PubMed, this is the case if the investigator is the first author). For example, a citation by Jane M Smith that is missing institution information scores 80, while one by Jane Smith as first author that has institution information scores 65.

2.4. Clustering citations into distinct identities

ReCiter considers each citation in the order determined by the previous step, grouping citations into clusters, each of which is intended to represent a separate investigator identity. For each citation, the algorithm decides whether the citation belongs to an established identity (e.g., whether it is by a Jane Smith at Harvard or another at Stanford) or constitutes a new identity (e.g., the first citation found for Jack Smith).

For the citation to match an existing cluster, all parts of the name must match (e.g., Jane A. Smith is a separate identity from Jane B. Smith). If more than one existing cluster matches, ReCiter selects the one with the greatest number of matching co-authors. For example, a citation by J. Smith might list one or more co-authors who published with Jack Smith.

When co-author information is not sufficient to determine the cluster to choose, the algorithm uses text in the institution, journal, title and keyword fields. The frequencies of words in these fields are used to form one vector for the citation and one for the cluster, which are compared using a similarity score (cosine measure). When a citation shares no co-authors with any cluster, and the similarity score falls below a threshold of 0.2, the citation is considered to constitute a new identity, and a new cluster is created for it.

Note that as the clusters increase in size, they represent the combined information of all the citations they contain (name variants, co-authors and keywords). Each new citation under consideration is effectively compared to all the citations in a cluster, which maximizes the amount of information without having to compare every pair of citations.

2.5. Matching target investigators to citation clusters

The final step is to determine which of the different identities best matches the target investigator (if any). As an implementation point, this matching process is identical to the one described above to assign a citation to one of the established identities. Each target investigator is represented using fields such as full name (ideally with middle name), current institutional affiliation (departments, divisions, etc.) and any available additional information, such as prior institutions, the journals in which the investigator typically publishes, research keywords and the names of frequent co-authors.

If co-authors are provided, the cluster with the greatest number of matching co-authors will be selected. Otherwise, the cluster with highest similarity score is selected, based on word frequencies in the investigator's institution, journal, title and keyword fields. If no cluster's score exceeds the 0.2 threshold, the algorithm determines that the given target investigator does not have any publications in the source database.

2.6. Evaluation of the ReCiter algorithm

A random sample of 200 investigators was selected from the investigator database (Fig. 2). The ReCiter algorithm queried PubMed, clustered the resulting citations and matched the clusters to investigators as described above. For comparison, we created a set of citations using name-based queries. These queries used the same search strategies as ReCiter (adding middle initial or institutional keywords), and tuned them so that they would generate similar numbers of citations per investigator (under 200). The concept was to provide a reasonable alternative method for automatically generating an investigator bibliography. The more restrictive strategies increase the precision of the queries while reducing recall.

We also used the names from the random sample to query the Scopus database. We chose Scopus because it also attempts to partition citations into separate identities [15], and because one of our institutions had access to this proprietary source. An administrator entered the first and last name for each investigator and selected an identity based on institutional affiliation, field, city and country. When Scopus presented multiple identities matching the institutional affiliation, all the identities were selected.

Results from name-based queries, ReCiter and Scopus were combined (removing duplicates) and presented to human reviewers without indicating which citations came from which method. Five members of the research team were chosen as reviewers and each citation was reviewed by three reviewers, so each reviewer assessed the retrieved citations for 120 investigators. A reviewer saw the combined list of citations for each investigator, starting with the most recent, along with the investigator's full name and departmental affiliations. Reviewers were free to use external sources, such as PubMed, to assist their decisions, but did so rarely. For each citation, the reviewer had to make a binary decision: yes, the publication was written by the target

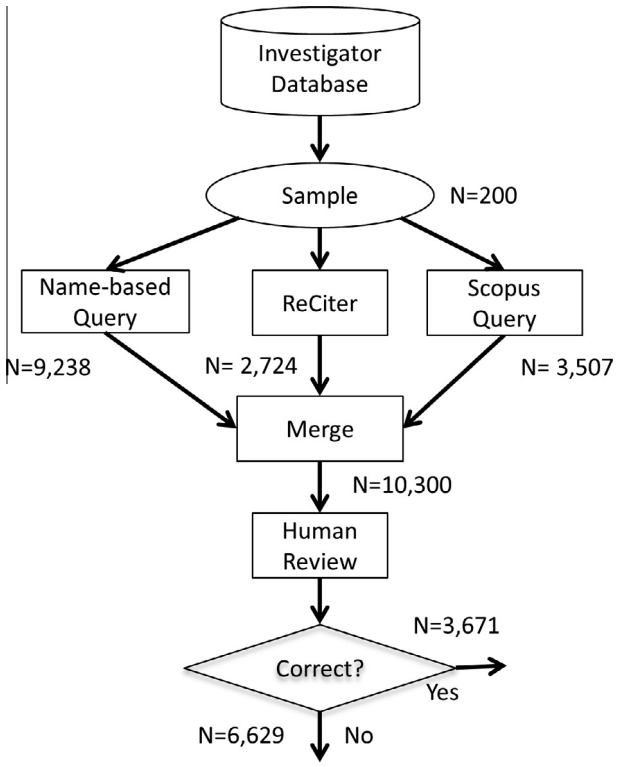

Fig. 2. Data flow of the evaluation of the ReCiter algorithm: select a random sample of investigator names and departmental affiliations from an institutional database; extract citations from PubMed using name-based queries; cluster citations and match to investigators; restrict name-based queries to ensure less than 200 citations per investigator; merge restricted and matched citations; assess correctness of merged citations; extract citations from Scopus using name-based queries; extract overlap between Scopus and reviewed citations; determine which non-overlapping publications are in PubMed; merge non-overlapping PubMed citations with reviewed citations.

investigator; or no, the publication was not written by the target investigator. For investigators with no publications, the reviewer had to make another binary decision: yes, the investigator does not actually have any publications; or no, the investigator does have publications. We created a gold standard using majority vote (e.g., a total of two or three "yes" votes was interpreted to mean that a citation was written by the investigator) and used this to compare the performance of the three methods.

3. Results

Names, titles and departmental affiliations were extracted for 10,471 investigators from the Columbia University human resource database as of November 2010. Table 1 shows each title ordered by the number of individuals with that title. In addition, the table shows the total number of citations matched to each title using the ReCiter algorithm. As expected, full professors had the highest number of matching citations (59,905) and of citations per investigator (64). Assistant professors had more matching citations (19,872) than associate professors (17,691), but many fewer citations per person (7 vs. 22). Research scientists (including senior) had many fewer matched citations (2159) than either assistant (2901) or associate (812) professors, but many more citations per person (26–30 vs. 7–22). The overall average number of citations per investigator was 12 (SD 35).

The ReCiter program extracted citations from PubMed using three different search strategies (Table 2). For the majority of investigators (93%), the queries used last name and first initial. If a query returned more than 4000 citations, more restrictive search strategies were used: middle initial (3%) and words extracted from institutional affiliation text (4%).

The named-based queries used by ReCiter produced from 0 to 4000 citations. The ReCiter program grouped the citations into clusters, where each cluster represents a distinct investigator identity.

For each investigator, the ReCiter algorithm attempted to select one of the clusters as the matching identity. Table 3 shows that for slightly more than half of the investigators, no cluster was selected; that is, the algorithm determined that the investigator did not have any citations in PubMed. The majority of the remaining investigators (45%) had 1–69 citations. The last 5% had 70–700 citations.

For the evaluation, a random sample of 200 investigators was extracted from the investigator database (Fig. 2). The name-based queries reviewed by human judges were more restrictive to generate a comparable number of citations (Table 2). Roughly half the investigators required a more restrictive search query (strategy 2 or 3), resulting in a total of 9238 citations. The ReCiter algorithm produced 2724 citations using clustering and matching. The Scopus database identified 3507 citations. The citations were merged using the Pubmed identifier, resulting in 10,300 unique citations.

Each reviewer was responsible for evaluating the merged publications for 120 target investigators and had to make about 6000 decisions about whether a retrieved publication was authored by the target investigator. Interrater agreement (Fleiss' kappa) among the five judges was 0.81. Table 4 compares the three methods in terms of how well retrieved citations match the target investigator. Scopus queries achieved the best recall (sensitivity) of 0.81 (0.79–0.82), name-based queries had 0.78 (0.76–0.79), while ReCiter had 0.69 (0.67–0.70). ReCiter attained the best precision (positive predictive value) of 0.93 (0.92–0.94) while Scopus had 0.85 (0.83–0.86) and name-based queries had 0.31 (0.30–0.32).

4. Discussion

The purpose of ReCiter is to automatically generate bibliographies for a given set of investigators. One of our goals in this study was to show that it is feasible to cluster citations into separate identities using common co-authors and text similarity scores, as has been done in other tools described in the introduction. However, the real challenge is to determine which of these identities (if any) is the one affiliated with the institution. This is the crucial step if we seek to automate the process of populating research networking systems, with a minimum of input from users.

4.1. Handling change

The algorithm incorporates a number of heuristics that might be employed by humans to partition citations into separate identities: start with the most informative citations (e.g., having full names); exploit common co-authors; and look for similar words in text fields (title, journal, institution or keywords). The algorithm must then determine whether any of these identities match the target investigator, by comparing name variants, keywords for the investigator's field, and institutional affiliation. These heuristics must contend with the fact that investigators are dynamic, changing names, research areas and institutions:

- An investigator's name can vary across citations. The algorithm can handle variants due to name completeness (as described in Section 2.3), but also different names (e.g., before and after marriage). However, the latter is only possible if there is a source of information about alternative forms. The human resource

Table 1
Number of investigators in each title category (titles are specific to the human resource database for the institution); number of citations selected as matching identity for each title; average number of citations per investigator; and number of investigators who have at least citation.

Investigator title	Individuals		Matched citations		Cit./Indiv.	Individuals with citations	
Assistant professor	2901	28%	19,872	16%	7	1460	50%
Instructor	1107	11%	1753	1%	2	269	24%
Postdoctoral residency fellow	946	9%	2042	2%	2	349	37%
Professor	929	9%	59,905	48%	64	717	77%
Associate professor	812	8%	17,691	14%	22	579	71%
Postdoctoral research scientist	652	6%	1943	2%	3	317	49%
Associate research scientist	648	6%	5423	4%	8	417	64%
Postdoctoral clinical fellow	562	5%	660	1%	1	235	42%
Postdoctoral research fellow	497	5%	1690	1%	3	245	49%
Lecturer	417	4%	7225	6%	17	205	49%
Assistant	293	3%	847	1%	3	57	19%
Staff associate	285	3%	1304	1%	5	129	45%
Senior staff associate	117	1%	1225	1%	10	86	74%
Associate	101	1%	435	0%	4	38	38%
Affiliate physician	96	1%	69	0%	1	10	10%
Research scientist	53	1%	1582	1%	30	38	72%
Senior lecturer	30	0%	529	0%	18	5	17%
Senior research scientist	22	0%	577	0%	26	13	60%
Associate research scholar	2	0%	0	0%	0	0	0%
Research scholar	1	0%	13	0%	13	1	100%
Total	10,471	100%	124,785	100%	12		50%

Table 2
Number of investigators requiring each type of search strategy for the database as a whole and for the random sample. Strategy 1 uses last name and first initial; strategy 2 adds middle initial; strategy 3 adds key words from institutional affiliation (names of school, department, division, etc.).

Search strategy	Last name	First initial	Middle initial	Institution keywords	Total investigators		Sample investigators	
1	x	x			9763	93%	99	50%
2	x	x	x		301	3%	49	24%
3	x	x		x	407	4%	52	26%
Total					10,471	100%	200	100%

Table 3
Number of investigators for whom the specified number of citations were selected as matching their identity. More than half did not match a cluster. Most of the remaining investigators matched a cluster containing 1–69 citations.

Pubs	Investigators	Percent
0	5301	50.6
1–69	4690	44.8
70–139	318	3.0
140–209	96	0.9
210–279	33	0.3
280–349	19	0.2
350–419	6	0.1
420–489	3	0.0
490–559	3	0.0
560–629	1	0.0
630–669	1	0.0
Total	10,471	100.0

Table 4
Comparison of citation matches between human reviewers and name-based queries, ReCiter and Scopus. Values marked as recall in the first column give the sensitivity, while values marked precision in the last column give the positive predictive value.

		Human			
		Yes	No	Total	
Name-based queries	Yes	2863	6375	9238	0.31 (precision)
	No	808	254	1062	0.24
	Total	3671	6629	10,300	
		0.78 (recall)	0.04		
ReCiter	Yes	2523	201	2724	0.93 (precision)
	No	1148	6428	7576	0.85
	Total	3671	6629	10,300	
		0.69 (recall)	0.97		
Scopus	Yes	2965	542	3481	0.85 (precision)
	No	706	6087	6819	0.90
	Total	3671	6629	10,300	
		0.81 (recall)	0.92		

database used in this study did not contain information about alternate names, so they would be treated as different identifies. A human reviewer would draw the same conclusion if unaware of the link between the names.

- An investigator can conduct research in two or more distinct fields, and might publish in different journals with different co-authors in each field. If these lines of research were pursued at the same institution, the algorithm would combine the citations into a single cluster. If performed at two distinct institutions with no overlap in content, it would be very difficult even for a human reviewer to know that this was the same individual.
- An investigator can move around between different institutions. For example, an investigator might have joined the current institution recently, and have no citations there, but have many citations from previous places of employment. ReCiter would be able to identify this individual if the previous citations contain similar affiliation words (e.g., extracted from the department or division name). Likewise, if an investigator changed institutions a lot, but maintained similar co-authors, the algorithm would be able to cluster the citations together. Without

evidence of common discipline or co-authors, it would be difficult for a human reviewer to rule out the possibility that the sets of citations are produced by distinct individuals.

ReCiter was not provided with information about prior affiliations, and therefore must infer this from citation data, making the connection by means of similar words in titles, journals or keywords. A trusted source of information about all institutional affiliations for an investigator would improve clustering of citations into identities and also enable matching one of those identities to the investigator. For example, free-text extracted from resumes could be used [22]. ReCiter can automatically convert this text to word frequency data for use in clustering citations and matching identities.

4.2. Predicting participation in research

No limit was placed on the number of citations that the ReCiter algorithm could assign to any given cluster; it was possible in principle for a cluster to have grown into thousands of citations. However, we found instead that the average number of citations in a cluster was around 10, most investigators had less than 70 citations and the maximum was less than 700 (Table 3), suggesting that the approach is finding realistic groupings. The algorithm also was not provided with any data about job titles, such as which investigators were professors. It determined that these individuals were the most prolific, and that full professors generated the most citations in aggregate and individually (Table 1), confirming what is expected.

The most surprising result is that half the investigators in this sample (5301) did not match any citations (Table 3). The citations of 4713 (89%) of these individuals did not contain any words related to the current institution; thus it is difficult to determine whether they published elsewhere or did not publish at all.

This suggests that it is challenging to develop a definitive list of individuals engaged in research for a given institution using a wide variety of job titles. The last column of Table 1 provides a break down by job title for the number of investigators who have at least one citation. Certain titles (staff members, fellows, and lecturers) show a low percentage of having citations, making it hard to predict whether they participate in research or not based on this data alone. Even for full professors the algorithm was able to find citations for only 77% of individuals. This suggests that clusters exist for the remaining professors, but that they did not match due to changes in name, research area or institution.

4.3. Trading effort for recall

In this study, the ReCiter and Scopus tools employed fundamentally different approaches.

Given data about investigators, ReCiter generated bibliographies automatically, while with Scopus, an administrator had to enter names and select identities manually. This manual effort is similar to what a new user of ORCID (Open Researcher and Contributor Identifier) must do when first populating a profile with citation records from Scopus. Because the administrator would often combine multiple identities in Scopus, the measure of recall may appear inflated, especially in cases where surnames in citations have not been represented consistently.

ReCiter was able to achieve relatively high levels of precision and recall without relying on affiliation information beyond the first author of a citation. In contrast, Scopus had access to affiliation data for all co-authors. As the PubMed database expands to include affiliations for all co-authors, ReCiter will be able to significantly improve the accuracy of clustering and identification.

4.4. Related applications

The author identity problem bears some similarity to establishing unique identifiers in patient care [23], health information exchange [24], and biomedical research [25]. These approaches share a common mathematical approach involving the calculation of similarity based on a set of attributes (name, date of birth, gender, etc.). However, they differ enormously from the present problem in the nature of the attributes that are used, their variation, utility and impact on matching. Moreover, the workflow and privacy concerns are completely different, e.g. clinical systems require human decisions. In contrast, the purpose of this paper is to reduce manual labor for investigators as far as possible when building research profiles.

The use case presented here focuses on building an institution-specific system. The ReCiter algorithm has also been used successfully to identify scientific communities that cross institutional boundaries [26]. This is very useful when a researcher wants to study or explore a particular scientific community, and is able to create a list of names and affiliations from other sources, for example, lists of members who attend particular scientific meetings. In this case, the goal is to identity individuals and their citations that share research themes, rather than a single institution. Another application of our tool is to investigate patterns of co-authorship. This work has potential to determine how team characteristics affect the impact of publications [27].

5. Conclusion

The ReCiter algorithm was able to retrieve and cluster the publications authored by each of a large group of investigators from one university health sciences center with relatively high precision, but with lower recall in comparison with author name queries in PubMed and the proprietary Scopus database. ReCiter is able to automatically select a group of citations from Pubmed, while Scopus requires manual effort by an administrator or an investigator. Our approach exploits local institutional knowledge to generate bibliographies. Enriching this knowledge with information about previous institutional affiliations would improve recall by linking work completed at different sites. The method can be adapted to citation databases that support name-based queries. The tool can be used to enrich a research-networking system centered on a single institution and to explore research communities that span multiple institutions.

References

[1] Clinical and Translational Science Awards. <http://www.ctsaweb.org/> [accessed 20.04.12].
[2] VIVO. An interdisciplinary national network. <http://www.vivoweb.org/> [accessed 20.04.12].
[3] Spallek H, Schleyer T, Butler BS. Good partners are hard to find: the search for and selection of collaborators in the health sciences. Presented at: Project Management and User Engagement. Fourth IEEE International Conference on eScience. Los Alamitos: IEEE Computer Society; 2008. p. 462–7.
[4] Harvard Catalyst Profiles. <https://connects.catalyst.harvard.edu/profiles/about> [accessed 20.04.12].
[5] ResearchGate; 2011. <http://www.researchgate.net/> [accessed 16.04.12].
[6] Rovner SL. A question of identity: multiple initiatives aim to unambiguously identify individual scientists so they're credited for their work. Chem Eng News 2010;88(21):36–7.
[7] Conlon M. Opportunities for federated identity management in scholarly work. National Institutes of Health Workshop on Identifiers & Disambiguation in Scholarly Work. Places and spaces – mapping science; March 2010. <http://scimaps.org/100318> [accessed 20.04.12].
[8] Bennett DB, Publications, identity and disambiguation. National Institutes of Health Workshop on Identifiers & Disambiguation in Scholarly Work. Places and spaces – mapping science; March 2010. <http://scimaps.org/100318> [accessed 20.04.12].
[9] Friedberg EC. Good news on the horizon: the Open Researcher and Contributor ID (ORCID). DNA Repair (Amst). 2010;9(2):102.

[10] Editorial. Credit where credit is due. Nature 2009;462:825.
[11] ResearcherID. Thompson Reuters web of knowledge. <http://www.researcherid.com/> [accessed 20.04.12].
[12] Open Researcher and Contributor ID (ORCID). Thompson Reuters and Nature Publishing Group. <http://www.orcid.org/> [accessed 20.04.12].
[13] Community of Science Pivot. ProQuest, LLC. <http://pivot.cos.com/> [accessed 20.04.12].
[14] MyNCBI. National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov/sites/myncbi/> [accessed 20.04.12].
[15] Scopus. Elsevier Publishing. <http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authoridentifier> [accessed 20.04.12].
[16] CiteSeer. Pennsylvania State University. <http://citeseerx.ist.psu.edu/> [accessed 20.04.12].
[17] Authority. University of Illinois at Chicago. <http://arrowsmith.psych.uic.edu/> [accessed 02.11.11].
[18] Torvik VI, Weeber M, Swanson DR, Smalheiser NR. A probabilistic similarity metric for MEDLINE records: a model for author name disambiguation. AMIA Annu Symp Proc.; 2003:1033.
[19] Torvik VI, Smalheiser NR. Author name disambiguation in MEDLINE. ACM Trans Knowledge Discover Data 2009;3(3):11.
[20] Smalheiser NR, Torvik VI. Author name disambiguation. In: Cronin B, editor. Annual review of information science and technology, no. 43; 2009. p. 287–313.
[21] Treeratpituk P, Giles CL. Disambiguating authors in academic publications using random forests. ACM/IEEE-CS joint international conference on Digital libraries; 2009. p. 39–48.
[22] McKibbon KA, Friedman PW, Friedman CP. Use of a MeSH-based index of faculty research interests to identify faculty publications: an IAIMSian study of precision, recall, and data reusability. Proc AMIA Symp 2002:514–8.
[23] McCoy AB, Wright A, Kahn MG, Shapiro JS, Bernstam EV, Sittig DF. Matching identifiers in electronic health records: implications for duplicate records and patient safety. BMJ Qual Saf; January 29, 2013.
[24] Zhu VJ, Overhage MJ, Egg J, Downs SM, Grannis SJ. An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. J Am Med Inform Assoc 2009;16(5):738–45.
[25] Johnson SB, Whitney G, McAuliffe M, Wang H, McCreedy E, Rozenblit L, et al. Using global unique identifiers to link autism collections. J Am Med Inform Assoc 2010;17(6):689–95.
[26] Bales ME, Johnson SB, Keeling JW, Carley KM, Kunkel F, Merrill JA. Evolution of coauthorship in public health services and systems research. Am J Prev Med 2011;41(1):112–7.
[27] Bales MS, Dine DC, Merrill JA, Johnson SB, Bakken S, Weng C. Associating Co-authorship Patterns with Publication Impact. J Biomed Inform; 2014 [submitted revision].