Journal of Biomedical Informatics xxx (2014) xxx–xxx

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 xxxx

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

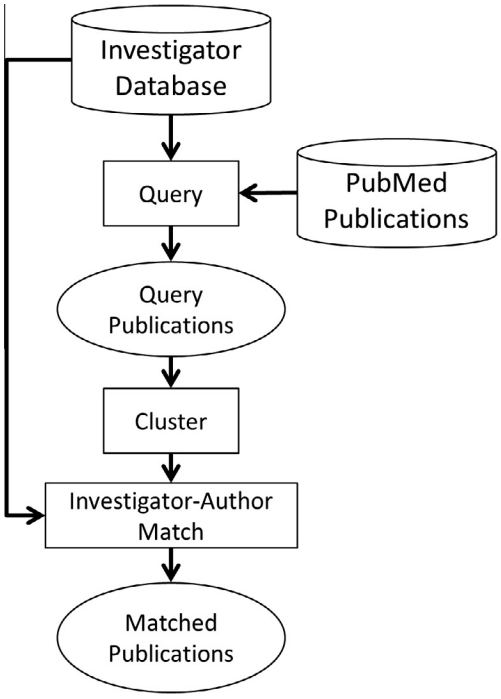

Fig. 1. Data flow of the ReCiter algorithm. Investigator names and departmental affiliations are selected from an institutional database; citations are extracted from PubMed using name-based queries; citations are clustered into separate identifies; the identities most closely matching the investigators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