Research Policy 43 (2014) 941–955

Contents lists available at ScienceDirect

Research Policy

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

Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010)

Guan-Cheng Li[a], Ronald Lai[b], Alexander D'Amour[c], David M. Doolin[d], Ye Sun[e], Vetle I. Torvik[f], Amy Z. Yu[g], Lee Fleming[h,*]

[a] Fung Institute for Engineering Leadership, UC Berkeley College of Engineering, Berkeley, CA 94550, United States
[b] Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, United States
[c] Department of Statistics, Harvard University, Cambridge, MA 02138, United States
[d] CloudPassage, Inc., 153 Townsend Street, San Francisco, CA 94026, United States
[e] Grantham, Mayo, Van Otterloo & Co. LLC, 40 Rowes Wharf, Boston, MA 02110, United States
[f] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E Daniel Street, Champaign, IL 6182, United States
[g] MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[h] Fung Institute for Engineering Leadership, UC Berkeley College of Engineering, Berkeley, CA 94550, United States

ARTICLE INFO

Article history:
Received 24 November 2011
Received in revised form 18 January 2014
Accepted 19 January 2014
Available online 28 February 2014

Keywords:
Disambiguation
Patents
Networks
Inventors
Careers

ABSTRACT

Research into invention, innovation policy, and technology strategy can greatly benefit from an accurate understanding of inventor careers. The United States Patent and Trademark Office does not provide unique inventor identifiers, however, making large-scale studies challenging. Many scholars of innovation have implemented ad-hoc disambiguation methods based on string similarity thresholds and string comparison matching; such methods have been shown to be vulnerable to a number of problems that can adversely affect research results. The authors address this issue contributing (1) an application of the Author-ity disambiguation approach (Torvik et al., 2005; Torvik and Smalheiser, 2009) to the US utility patent database, (2) a new iterative blocking scheme that expands the match space of this algorithm while maintaining scalability, (3) a public posting of the algorithm and code, and (4) a public posting of the results of the algorithm in the form of a database of inventors and their associated patents. The paper provides an overview of the disambiguation method, assesses its accuracy, and calculates network measures based on co-authorship and collaboration variables. It illustrates the potential for large-scale innovation studies across time and space with visualizations of inventor mobility across the United States. The complete input and results data from the original disambiguation are available at (http://dvn.iq.harvard.edu/dvn/dv/patent); revised data described here are at (http://funglab.berkeley.edu/pub/disamb_no_postpolishing.csv); original and revised code is available at (https://github.com/funginstitute/disambiguator); visualizations of inventor mobility are at (http://funglab.berkeley.edu/mobility/).

1. Introduction

Reasonably complete though raw United States patent data first became available in the 1990s for research in the fields of technology and innovation. Publication of a curated dataset by the National Bureau of Economic Research (NBER) enabled access by a much broader set of researchers (Hall et al., 2001) especially those that lacked the resources and hardware or programming skills to manipulate the raw data. The original NBER database included inventor names, firm name and state level data but did not identify unique inventors over time.

Uniquely identifying inventors presents at least two challenges. First, the United States Patent and Trademark Office (USPTO) does not require consistent and unique identifiers for inventors. For example, the last author of this paper is listed as Lee O. Fleming on patent 5,136,185 (Fleming, 1992) but as Lee Fleming on patent 5,029,133 (Fleming, 1991). Both inventors work for Hewlett Packard, both invent digital hardware, and both live in Fremont, California – without personal knowledge, with what confidence could we infer that this is the same inventor? Moving directly into the second challenge, could we repeat this process for millions of inventors? Accurate and automatic disambiguation of the entire

* Corresponding author. Tel.: +1 510 664 4586.
E-mail addresses: guanchengli@eecs.berkeley.edu (G.-C. Li), laironald@gmail.com (R. Lai), adamour@iq.harvard.edu (A. D'Amour), ddoolin@cloudpassage.com (D.M. Doolin), Edward.Sun@GMO.com (Y. Sun), vtorvik@illinois.edu (V.I. Torvik), ayu@media.mit.edu (A.Z. Yu), lfleming@ieor.berkeley.edu (L. Fleming).

0048-7333/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.respol.2014.01.012

patent record requires careful algorithm design to ensure scalability and, even then, significant computational resources to ensure feasibility. For example, the brute force approach to compare all pairwise inventor-patent records is not feasible at full scale for any but the most powerful computers in existence.

In recent years there has been a flurry of activity surrounding the problem of name ambiguity in bibliographic records such as journal and conference paper collections (reviewed by Smalheiser and Torvik, 2009). Of particular note, and strong motivation for this paper, recent work has highlighted the pitfalls of poor or simplistic author disambiguation; for example: Raffo and Lhuillery (2009) demonstrate differences in econometric inferences, Diesner and Carley (2009) show differences in entity resolution and relationships in newspaper corpora, and Fegley and Torvik (2013) illustrate dramatic distortions in social networks due to non-existent or poor disambiguation. Due to space constraints, we will not make similar comparisons here, but recommend the reader to this literature, and encourage the community to heed this literature's concerns in future analyses.

1.1. *Existing work and contribution*

Our paper contributes (1) an application of the Author-ity disambiguation approach (Torvik et al., 2005; Torvik and Smalheiser, 2009) to the US utility patent database, (2) a new iterative blocking scheme that expands the match space of this algorithm while maintaining scalability, (3) a public posting of the algorithm and code, and (4) a public posting of the results of the algorithm in the form of a database of inventors and their associated patents. The work builds directly on prior efforts by a variety of innovation researchers (Fleming and Juda, 2004; Singh, 2005; Trajtenberg et al., 2006; Raffo and Lhuillery, 2009; Carayol and Cassi, 2009; Lai et al., 2009; Pezzoni et al., 2012). The database provides unique identifiers for each patent's inventors from 1975 through 2010. It also provides social network measures by each inventor, by three-year blocks over the same time period. To illustrate applications of the data, we provide movies of inventor mobility across large U.S. states since 1975. The algorithms and code are made public to encourage further development and improvement by the community of patent and innovation investigators. In addition to improved disambiguation, the Harvard Dataverse Network (DVN) website provides a network interface that enables a researcher to subset the co-authorship networks of inventors.[1] Output formats support both regression analysis and graphical network programs.

1.2. *Precís*

The second section of the paper ("*Overview of dataset preparation*") provides an explanation on how the inventor dataset is created; the third section ("*Disambiguation: overview, theory, and implementation*") provides a non-technical overview and explanation of the disambiguation processes; the fourth section ("*Results and accuracy metrics*") describes how we report results and accuracy; the fifth section ("*Disambiguated data and illustrative applications*") illustrates applications of the data. Appendices include patent data descriptions, listings of data and results distributed through the Harvard Dataverse Network and schemas used in and produced by the disambiguation.

2. Overview of dataset preparation

Fig. 1 illustrates an overview of the patent disambiguation data preparation process. Source data come from the NBER database (Hall et al., 2001), directly from the USPTO weekly publications, and secondary sources. Dataset preparation consists of obtaining, parsing, and cleaning the raw data, creating four preliminary datasets containing inventor, patent, assignee, and classification data, and consolidating all data into a single database with inventor-patent instances.

2.1. *Primary data sources*

The final inventor, assignee, patent, and class datasets were built using primary data sources from the USPTO and the NBER.[2] The USPTO makes up-to-date patent data available on their public web resource[3] through collaborations with the European and Asian patent offices. The weekly data file is a concatenated list of granted patents, where each patent is represented by an XML document (that is, all files are merged chronologically). The NBER patent database contains patents granted from 1975 to 1999 and is publicly available.[4] Since the patent office only began automating data storage in 1975,[5] we are utilizing information from 1975 onwards. To the best of our knowledge, there is no freely available and comprehensive computer database containing U.S. inventor information before 1975, though bulk download of images and OCR text (of variable quality) files are available.[6]

2.2. *Secondary data sources*

In addition to the primary data sources, we used data from secondary public data sources to help identify inventors. These secondary data sources include the USPTO CASSIS dataset,[7] the National Geospatial-Intelligence Agency country files,[8] the US Board on Geographic Names[9] and NBER File of Patent Assignees.[10]

When a patent is granted, the USPTO assigns multiple alphanumeric codes to classify the technology. As technology advances, the USPTO creates new classifications and updates previously coded patents. These classification changes are indicated in CASSIS, a dataset that is updated bimonthly. Classifications reflect the November 2009 concordance. Geographic metrics are sourced from public databases such as the National Geospatial-Intelligence Agency and the US Board on Geographic Names, current through 2009 (recent efforts have improved upon this, see Johnson, 2013).

[1] Original data are stored at http://dvn.iq.harvard.edu/dvn/dv/patent. More recent disambiguation code and updated data are available at Fung Institute and GitHub websites: https://GitHub.com/funginstitute/downloads.

[2] Some of the early NBER data are missing and are supplemented by the 1998 Micropatent CD product (http://www.micropat.com/static/index.htm). We would like to acknowledge the donation of these data from Corey Billington and Ellen King of Hewlett-Packard. This completes approximately 70,000 gaps in data for records from 1975 to 1978.

[3] USPTO provides weekly Bibliographic Information for Patent grants through its Sales Order Management System (SOMS) Catalog. https://EIPweb.uspto.gov/SOMS.

[4] See Hall et al., 2001 at http://www.nber.org/patents/.

[5] NBER provides limited data from 1963 to 1999 but only provides inventor data from 1975 to 1999. Since inventor information is necessary in our disambiguation algorithms, we have only matched inventors to patents granted after 1975. Further information about the inventor dataset can be found at: http://www.nber.org/patents/inventor.txt.

[6] Google Books: http://www.google.com/googlebooks/uspto-patents.html.

[7] Patents CLASS: Current Classifications of US Patent Grant Publications 1790 to Present' (Code: EIP-2050P-DD): http://www.uspto.gov/web/offices/ac/ido/oeip/catalog/products/pp-o2w-3.htm#classP2050dd.

[8] Country Files (GNS) is a public database that contains Longitudinal and Latitude information for cities and locations around the world. http://earth-info.nga.mil/gns/html/namefiles.htm.

[9] States, Territories, Associated Areas of the United States is a National file that contains Longitudinal and Latitude information for cities across the states. http://geonames.usgs.gov/domestic/download_data.htm.

[10] https://sites.google.com/site/patentdataproject/Home/downloads.