Authormagic in INSPIRE

Author Disambiguation in Scholarly Communication

Brooks, Travis C.
SLAC
travis@slac.stanford.edu

Carli, Samuele
Univ. degli Studi di Firenze and CERN
samuele.carli@cern.ch

Dallmeier-Tiessen, Sünje
HU Berlin and CERN
sunje.dallmeier-tiessen@cern.ch

Mele, Salvatore
CERN
salvatore.mele@cern.ch

Weiler, Henning
Univ. Erlangen-Nürnberg and CERN
henning.weiler@cern.ch

ABSTRACT

"Authormagic" is a system designed to solve the systemic challenge of the attribution of scholarly artifacts to unique authors in scientific digital libraries. It relies on the unique combination of machine-based knowledge retrieval and distributed knowledge of the users of the system themselves. Algorithmically computed lists of the authors' publications, disambiguated through a (meta-) data mining approach, allow users to follow an intuitive procedure to validate and improve content to an author's scholarly profile. This approach constitutes the core of a new paradigm for extended author-centric and user-centric services in large-scale scientific digital libraries.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: *Clustering, information filtering* and *relevance feedback*.

General Terms

Algorithms and Human Factors.

Keywords

Digital libraries, author identification and disambiguation, crowdsourcing, subject repositories

1. INTRODUCTION

For decades, knowledge about the attribution of scholarly artifacts to individuals, in the frequent cases of name ambiguity, has been with the authors' peers or knowledgeable experts in libraries and has since been a known and systematic problem for repositories of all kinds and sizes. The digitization of scholarly communication, public interest in science, the international expansion of the scientific community and other factors constitute to a substantial growth of the information environment in both, quantity and complexity. In turn, these advances make unambiguous attribution of scholarly artifacts an even harder task, while making it all more valuable for the discovery and retrieval of information in the environment of scholarly communication and the increased market for evaluation of scientific productivity. There exist several initiatives aiming to find a solution to author ambiguity, the most promising being ORCID, "a community effort to establish an open, independent registry that is adopted and embraced as the industry's de facto standard", which "brings together the most influential universities, funding organizations, societies, publishers and corporations." [1].

In collaboration with DESY [2], Fermilab [3] and SLAC [4], CERN [5] is developing INSPIRE [6]: a complete digital library of all High-Energy Physics (HEP) publications [7]. INSPIRE follows in the steps of the HEP tradition of community-based scientific information infrastructures [8, 9]. INSPIRE offers a unique opportunity to explore possibilities in author disambiguation. The data set comprises the entire corpus of documents produced within the discipline. A hybrid approach of algorithmic author disambiguation and user engagement (i.e. "crowdsourcing") is beginning to enrich the author information in the HEP database and could possibly be extended to other disciplines.

A fully disambiguated set of authors builds a foundation for the development of novel services in information discovery and knowledge extraction. There are opportunities in exploring and understanding precise authorship data and its correlation in various information graphs (co-authorship, citation, affiliation etc.). It also enables meaningful author-centric scientometrics and removes ambiguities from attempts to use bibliographic and citation data to evaluate scientific quality. In INSPIRE, the latter is of real concern within the community, since INSPIRE data is pivotal in community-based hiring and funding decisions.

Crowdsourcing author disambiguation and the attribution of scholarly artifacts stands as a prototype to measure the willingness of users to participate in the curation of their community's scientific information and metadata. An earlier survey [10] highlighted a general interest by community members to spend 30 minutes or more per week in similar activities. If this trend is confirmed, it could point to a paradigm shift in the organization of scientific information.

2. INSPIRE

INSPIRE combines the successful SPIRES database [11], operated by DESY, Fermilab and SLAC, with the Invenio [12] digital library technology developed at CERN. INSPIRE is run by a collaboration of these four labs, and interacts closely with HEP publishers (e.g. APS [13], Elsevier [14], IOPp [15], SISSA [16] and Springer [17]), and sister information services such as arXiv.org [18], NASA-ADS [19] and PDG [20].

Today INSPIRE counts one million records with half a million Open Access full-text documents. This system is in transition to become the main working tool of a community of 50'000 scientists worldwide with a clear roadmap to provide new services in information discovery and retrieval tailored to the community needs.

3. AUTHORMAGIC

3.1 The Algorithm

An author disambiguation algorithm has been designed and implemented as a tool to identify which individual researcher wrote which particular documents about what, when, where and with whom. The algorithm determines metadata similarities (e.g. name, affiliation, keywords, coauthors and others) between all 6'500'000 author signatures (author names along with describing metadata on documents) from INSPIRE holdings. These similarity measures are used to ultimately compute clusters of information, where each cluster represents a person. The algorithm first collects and unifies potentially ambiguous text components to ensure a high level of comparability. Pre-selecting groups of signatures based on their last names allows for the minimization of comparisons needed for the computation. Based on all information found for these last name groups, individual clusters are found by cross-comparing metadata entries for each member of the last name group. The algorithm's performance highly depends on metadata quality, wherein older documents or material harvested from cognate disciplines pose difficulties for accurate assignment decisions.

3.2 Engaging The Users

Social media and the web2.0 paradigm play an increasingly important role in the professional scientific context. This presents an opportunity to adopt a crowdsourcing approach to empower users of INSPIRE to claim documents for themselves or to suggest assignments of documents to other authors the users are related to (by co-authorship, for instance), thus solving faulty algorithm assignments. The main driver for authors to participate in the endeavor is to ensure that their scientific production is accurately represented to the community. Further uses of citation information as well as other services such as an automated generation of lists of publications for CVs are also important to users. This driver has been observed from a previous analysis of the researchers' e-mail feedback to SPIRES. While previous experience of author claiming publications do exist, INSPIRE is in the unique position to leverage its central role as information hub in a discipline, and the recognized role as the first port of call, for instance, for prospective employers to assess the contribution of prospective hires to the discipline.

As part of the active crowdsourcing experiment, first, groups of users will be defined. This will be done analyzing ancillary databases in INSPIRE, which contain user-submitted contact information of most HEP researchers. Then, based on the analysis of demographic information (e.g. discipline, seniority and institutional affiliation) more custom-tailored and targeted mailings will be sent to each group in order to raise awareness and ask the members of these groups for an active participation to correct their publication lists in INSPIRE. Finally, more web marketing techniques will be used to further draw attention to this and potentially other author- and user-centric services.

3.3 A Hybrid Approach

Actions taken by the users will immediately feed back to the algorithm with a two-fold purpose. First, such user-confirmed data will be locked from further modification (i.e. the algorithm may not re-assign the document behind an assignment to another person). Second and most important, user-submitted metadata will be fed back with a high weight into subsequent re-runs of the algorithm, allowing, through co-authorship networks, to further improve other, previously ambiguous, decisions.

WebSci '11, June 14-17, 2011, Koblenz, Germany.

4. FIRST RESULTS

Knowledgeable experts assessed the performance of the algorithm, finding a correct clustering for 96% of 16'500 documents by 127 authors. The algorithm was then applied on the entire INSPIRE holdings, reducing the 6'500'000 unique author signatures, appearing in the INSPIRE metadata for its million records, to 210'000 potential unique authors. This enhanced data is now in production, greatly enhancing the author search and retrieve facility in INSPIRE. At the same time, without explicit publicity, the crowd-sourced claiming functionality has also been made available for serendipitous user-discovery. Within days, 14'557 attributions of documents, involving 351 unique authors have already been recorded. Out of those, the algorithm's suggested attribution was correct over 95% of the times. The conference poster presents an update of those preliminary results.

5. CONCLUSION

Preliminary results validate the Authormagic approach, both in terms of the accuracy of a disambiguation algorithm based on high-quality metadata, and the potential for large-scale crowd-sourced curation of the content of scientific digital libraries.

If the trends observed in the user participation are confirmed, a paradigm shift in the management of curation of repositories could emerge. In turn, a virtuous feedback circle would emerge; allowing other algorithms to leverage the increasing amount of high-quality user contributed metadata enabling novel author-centric and user-centric services, e.g., accurate citation metrics or automated generation of users' bibliographies and scientific biographies. And bring digital libraries in the XXIst century.

6. REFERENCES

[1] Open Researcher and Contributor ID: http://orcid.org
[2] DESY: http://desy.de
[3] Fermilab: http://www.fnal.gov
[4] SLAC: http://slac.stanford.edu
[5] CERN: http://cern.ch
[6] INSPIRE: http://inspirebeta.net/
[7] Available Online: http://www.projecthepinspire.net/
[8] **Gentil-Beccot, A., Mele, S., & Brooks, T. C.** 2009. *Citing and reading behaviours in high-energy physics: how a community stopped worrying about journals and learned to love repositories* http://arxiv.org/abs/0906.5418
[9] **Heuer, R., Holtkamp, A., Mele, S.** 2008. *C Innovation in scholarly communication: Vision and projects from High-Energy Physics* http://arxiv.org/abs/0805.2739
[10] **Gentil-Beccot, A., et al.** 2008 *Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course* http://arxiv.org/abs/0804.2701
[11] SPIRES: http://www.slac.stanford.edu/spires/
[12] Invenio: http://invenio-software.org
[13] American Physical Society: http://www.aps.org/
[14] Elsevier: http://www.elsevier.com/
[15] Institute of Physics publishing: http://publishing.iop.org/
[16] Sissa University: http://www.sissa.it/
[17] Springer: http://springer.com
[18] arXiv.org: http://arxiv.org
[19] NASA-ADS: http://adswww.harvard.edu/
[20] Particle Data Group: http://pdg.lbl.gov/