Also By The Same Author: AKTiveAuthor, a Citation Graph Approach to Name Disambiguation

Duncan M. McRae-Spencer
Electronics and Computer Science
University of Southampton
Highfield, Southampton, UK
(+44) (0) 23 80598347
dmms03r@ecs.soton.ac.uk

Nigel R. Shadbolt
Electronics and Computer Science
University of Southampton
Highfield, Southampton, UK
(+44) (0) 23 8059 3523
nrs@ecs.soton.ac.uk

ABSTRACT
The desire for definitive data and the semantic web drive for inference over heterogeneous data sources requires co-reference resolution to be performed on those data. In particular, name disambiguation is required to allow accurate publication lists, citation counts and impact measures to be determined. This paper describes a graph-based approach to author disambiguation on large-scale citation networks. Using self-citation, co-authorship and document source analyses, AKTiveAuthor clusters papers, achieving precision of 0.997 and recall of 0.818 over a test group of eight surname clusters.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms
Algorithms

Keywords
Name Disambiguation, Self-Citation, Metadata Analysis

1. INTRODUCTION
As automated information extraction systems become increasingly common, there is an increased demand to know whether two similar names refer to the same real-world object or not. This phenomenon is particularly problematic when considering author names of research papers or bibliography citations. Two specific problems exist. Firstly, one author may have multiple aliases, such as 'Nicholas Jennings', 'N. Jennings' and 'Nick R. Jennings'. Secondly, multiple authors may have a similar name, such as David L. Harris (Professor of Engineering at Harvey Mudd College) and David L. Harris (Infrastructure Systems Engineering Department, Sandia Labs, Albuquerque).

JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.

Typical approaches taken to solve this name disambiguation problem are based on text and language processing: for example, Li et al (2005)[3]. However, the approach described in this paper is based on metadata rather than textual context, and more closely matches the approach taken by Han et al (2004, 2005)[1]. This paper details AKTiveAuthor, a novel approach to the problem of automated name disambiguation in the specific context of a large-scale citation network. Our approach is centred around the observation that within these citation graphs, there is a tendency for authors to cite their own previous work. This approach can be used to iteratively tie together papers within a citation graph to eventually yield a collection of papers that should be by the same author. Combining this self-citation approach with already-established procedures for author disambiguation (co-authorship networks and source URL metadata), a graph of an author's work can be produced that is generally very complete.

The rest of this paper deals with the specifics of the methodology used in the experiments (section 2), the results (section 3) and finally the conclusions and future work in this area (section 4).

2. METHODOLOGY AND EXPERIMENTS
To test the effectiveness of our method, it is necessary to check the results against real-world data, which means checking by hand. While it is envisaged that an application based on the algorithm presented here will be run over an entire citation network dataset (such as that of Citeseer[2], our data source for this work), for the purposes of experimentation it is necessary to break down the work into sets of papers small enough that they can be checked by hand. Family-name clusters were chosen for this purpose as the clusters could be produced by simple parsing, and would be small enough to allow for by-hand checking for each cluster.

The first pass at tying the name-cluster papers together is to apply the self-citation graph. Initially, each paper is put in a 'collection' of size one. Each paper in the name-cluster is successively tested against every other paper within that name-cluster to see if the second paper is in the bibliography of the first, or vice versa. If it is found that the two papers are linked by a citation relationship, the second paper is added to the collection associated with the first, a link that is only confirmed when the sanity name-check is performed. The second and third passes are performed in a similar manner, using co-authorship and source URL metadata as a means of testing to see if two authors are the same. Again, these steps require the confirmation of a sanity name-check.

This name-check is based on the observation that while it is not possible to say that two people who share a name are the same person (for instance, manual checking reveals nine distinct David Johnsons within Citeseer's dataset), it is certainly feasible to suggest that two people with different names may be considered different people: Norman L. Johnson is different from David E. Johnson in almost every conceivable case. Therefore before committing to tying together two author 'collections', the full names are checked against each other to see if they are obviously not the same person.

2.1 Metrics

Before considering the results, it is important to explain their format. Unlike Han et al, we are considering the results from the point of view of the real-world authors rather than from the collection of test papers we are looking to classify. As such, a straight accuracy measure of 'how many papers did we match with the correct canonical author' does not work: we are looking to *create* 'canonical authors' as part of the process. Our results therefore more closely reflect information retrieval work and yield three metrics: precision, recall and f-measure[4]. A result is obtained for each metric for each individual paper, and these results are then combined as appropriate.

3. RESULTS

Table 1 shows the overall results for the AKTiveAuthor system against the eight chosen name-clusters. It is important to note that authors who have written exactly one document are not included in these results: they produce an automatic result of 1.000 for both precision and recall, which is not useful.

Name (cluster size)	Precision	Recall	F-measure
Carr (242)	1.000	0.754	0.860
Giles (414)	0.998	0.935	0.965
Glaser (79)	1.000	0.824	0.904
Hall (644)	0.996	0.783	0.877
Harris (477)	0.992	0.705	0.824
Jennings (389)	1.000	0.852	0.920
Johnson (2201)	0.991	0.806	0.889
Lawrence (353)	1.000	0.883	0.938
Arith. Mean	**0.997**	**0.818**	**0.899**

Table 1. Results for the eight name-clusters tested.

3.1 Analysis of results

The first thing to note is that the precision is consistently much higher than the recall. This is in line with expectations: self-citation will lead to very high precision but will not draw in documents that are outside an author's main citation network.

Secondly, results are better where one author dominates a name-cluster. For example, 277 of the 389 Jennings documents (71.2%) are authored by Nick Jennings, while the next highest contributor to the group is Jim Jennings from IBM who has 33 papers in Citeseer. By contrast, the Harris group shows a recall of only 0.705, where the most dominant member of that group (John G Harris of the University of Florida) authors only 34 out of 477 documents (7.1%).

Finally, the results for Johnson (2201 total papers) and Glaser (79 papers) are consistent enough to show that size of name-cluster does not appear to be a factor either way in these analyses. As more data is added to the system, it is expected that more complete datasets (both large and small) will yield better results.

4. CONCLUSIONS AND FUTURE WORK

The purpose behind this work to disambiguate authors is to provide a number of services based on the citation graph and document metadata held in Citeseer. Some of these services would include "view my papers", "count my citations" and perhaps in time "calculate my impact". In terms of usefulness of data for these services, the results have been promising, although it has also been necessary to create a correction facility within the 'view my papers' service allowing manual disambiguation. Overall, however, the results are good enough to prove useful and therefore can be considered a success. Other future work in the area of author disambiguation includes investigating other methods of tying papers together as well as moving from a relational database data source to an RDF-based triplestore and data asserted against a standard ontology, allowing easier integration of future data from heterogeneous sources.

5. ACKNOWLEDGMENTS

This work is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. We would like to thank Professor C. Lee Giles and Isaac Councill of Penn State University for the provision of Citeseer and for ongoing helpful comments. We would also like to thank the reviewers of this paper for their helpful guidance and suggestions.

6. REFERENCES

[1] Han, H., Giles, C. L., Zha, H., Li, C., Tsioutsiouliklis, K. Two supervised learning approaches for name disambiguation in author citations. *In proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, Tuscon, AZ, USA, (2004). 296-305.

[2] Lawrence, S., Bollacker, K., Giles, C.L. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6), (1999). 67-71.

[3] Li, X., Morie, P., Roth, D. Semantic Integration in Text: From Ambiguous Names to Identifiable Entities. *AI Magazine*, 26(1), (2005). 45-58.

[4] van Rijsbergen, C. J. *Information Retrieval*. London: Butterworth, (1979)