SyGAR – A Synthetic Data Generator for Evaluating Name Disambiguation Methods

Anderson A. Ferreira, Marcos André Gonçalves, Jussara M. Almeida, Alberto H.F. Laender, and Adriano Veloso

Department of Computer Science
Federal University of Minas Gerais
31270-901 Belo Horizonte-MG Brazil
{ferreira,mgoncalv,jussara,laender,adrianov}@dcc.ufmg.br

Abstract. Name ambiguity in the context of bibliographic citations is one of the hardest problems currently faced by the digital library community. Several methods have been proposed in the literature, but none of them provides the perfect solution for the problem. More importantly, basically all of these methods were tested in limited and restricted scenarios, which raises concerns about their practical applicability. In this work, we deal with these limitations by proposing a synthetic generator of ambiguous authorship records called SyGAR. The generator was validated against a gold standard collection of disambiguated records, and applied to evaluate three disambiguation methods in a relevant scenario.

1 Introduction

It is practically a consensus that *author name disambiguation* in the context of bibliographic citations is one of the hardest problems currently faced by the digital library community. To solve this problem, a disambiguator is applied to correctly and unambiguously assign a citation record to one or more authors, already or not present in the digital library, despite the existence of multiple authors with the same name (or very similar names – polysems), or different name variations (synonyms) for the same author in the data repository.

The complexity of dealing with ambiguities in digital libraries has led to a myriad of methods for name disambiguation [1,2,3,4,5,6]. Most of these methods demonstrated to be effective in specific scenarios with limited, restricted, and static snapshot collections. This leads to the question: *Would any of these methods effectively work on a dynamic and evolving scenario of a living digital library?*

In this paper, we propose a Synthetic generator of ambiguous Groups of Authorship Records (SyGAR) that is capable of generating synthetic authorship records of ambiguous groups, and thus can be used to simulate the evolution of a digital library over time. The use of a synthetic generator to evaluate disambiguation methods makes it possible to generate and simulate several controlled, yet realistic, long term scenarios to assess how distinct methods would behave under a number of different conditions.

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 437–441, 2009.

2 Generating Synthetic Ambiguous Authorship Records

2.1 SyGAR Design

SyGAR takes as input a real collection of ambiguous groups previously disambiguated. Each such authorship record is composed of the author name (*author*), a list of her coauthors' names (*coauthors*), a list of terms present in the work title, and the publication venue title. For each ambiguous group in the input collection (*input group*), the number of unique authors N_A and the total number of authorship records N_R to be generated are also inputs to SyGAR.

As output, SyGAR produces a representative list of synthetically generated authorship records (*output group*) using a set of attribute distributions that characterize the publication profiles of each group and of its individual authors.

Building Author and Group Publication Profiles from Input Groups. Each publication profile of an author a is extracted from the corresponding input group by summarizing her record list into: (1) the distribution of the number of coauthors per a's record - $P^a_{nCoauthors}$; (2) a's coauthor popularity distribution - $P^a_{Coauthor}$; (3) the distribution of the number of terms in a work title by a - P^a_{nTerms}; (4) a's term popularity distribution - P^a_{Term}; and (5) a's venue popularity distribution - P^a_{Venue} (i.e., the distribution of the number of a's records with the same venue title). We assume that these attribute distributions are statistically independent, the terms appearing in the work title are independent from each other, and so are the work coauthors. Finally, we build a group profile with the distribution of the number of records per author - $P^g_{nRecordsPerAuthor}$.

Generating Records for Existing Authors. Each synthetic authorship record is created by following the steps: (1) select one of the authors a of the group according to $P^g_{nRecordsPerAuthor}$; (2) select the number a_c of coauthors according to $P^a_{nCoauthors}$; (3) repeat a_c times: select one coauthor according to $P^a_{Coauthor}$; (4) select the number a_t of terms in the title according to P^a_{nTerms}; (5) repeat a_t times: select one term for the work title according to P^a_{Term}; and (6) select the publication venue according to P^a_{Venue}.

Adding New Authors. SyGAR may be used to create records for new authors. Currently, SyGAR uses a knowledge base with the distribution of the number of records with the same coauthor - $P_{Coauthor}$, and the attribute distributions of the publication profile of each coauthor in the input collection. A new author is created by selecting one of its coauthors a, using $P_{Coauthor}$. The new author inherits a's profile. All generated records will have a as one of its coauthors. This strategy mimics the case of an author who follows the areas of one that will be a frequent coauthor.

2.2 Validation

We validate SyGAR by comparing real ambiguous groups against corresponding synthetically generated groups, assessing whether the synthetic groups capture

Table 1. SyGAR Validation across State-of-the-Art Name Disambiguation Methods

Method	Ambiguous Group	Real		Synthetic	
		MicroF$_1$	MacroF$_1$	MicroF$_1$	MacroF$_1$
SVM	A. Gupta	0.879±0.009	0.650±0.027	0.894±0.009	0.651±0.027
	C. Chen	0.761±0.015	0.611±0.025	0.779±0.012	0.580±0.018
	D. Johnson	0.809±0.027	0.623±0.026	0.817±0.018	0.615±0.029
SLAND	A. Gupta	0.916±0.008	0.809±0.025	0.947±0.006	0.807±0.028
	C. Chen	0.866±0.007	0.781±0.013	0.903±0.007	0.795±0.016
	D. Johnson	0.896±0.028	0.731±0.041	0.905±0.013	0.747±0.023

the aspects that are relevant to disambiguation methods. The real groups used, "C. Chen", "D. Johnson" and "A. Gupta", are selected from the collection of groups extracted from DBLP by Han *et al* [2].

For each real group, ten synthetic groups were generated. The number of authors and records per author in the synthetic group are set to be the same as in the input group. Table 1 shows average results of the disambiguation with 95% confidence intervals, with two supervised methods, an SVM-based method [1] and SLAND [6], under micro and macro F_1 measures. For all metrics, methods and groups, the results obtained for the real group are very close to those for the corresponding synthetic group, with a maximum error under 6%. In fact, six out of the twelve pairs of results are statistically indistinguishable with 95% of confidence. Thus, SyGAR is able to accurately capture aspects of real groups that are key to evaluate state-of-the-art name disambiguation methods.

3 Evaluating Disambiguation Methods with SyGAR

To illustrate a use of SyGAR, we evaluate three state-of-the-art disambiguation methods, namely an SVM-based method, a K-way spectral clustering method (KWAY), and SLAND, in realistic scenarios that encompass a live digital library (DL) evolving over a period of ten years. We perform experiments on ambiguous group "A. Gupta". The DL, at its initial state s_0, consists of records from the real group. At the end of each year, a load is performed into the DL with synthetic records generated by SyGAR, parameterized with the real group as source of author profiles. The distribution of the number of records generated to each author in the group is built based on the distribution of the average number of publications per year per (existing and new) author. These distributions were extracted from the DBLP for the analyzed group during the period of 1984-2008.

Starting at state s_i, for each new load, SyGAR generates records to authors already in the DL as well as to new authors (it is specified as a fraction f equals to 0%, 2%, 5%, and 10% of the total number of authors in the DL at the state s_i). If either SVM or SLAND is used, all the records making up state s_i are used as training data, and the data in the new load are used as test data for the disambiguators. If KWAY is used, the generated records are first incorporated into the current state of the DL and the disambiguation is done with all records using the correct number of authors in the DL. The DL evolves then into a new state s_{i+1}, and the micro-F_1 values are calculated for the whole DL in state s_{i+1}.

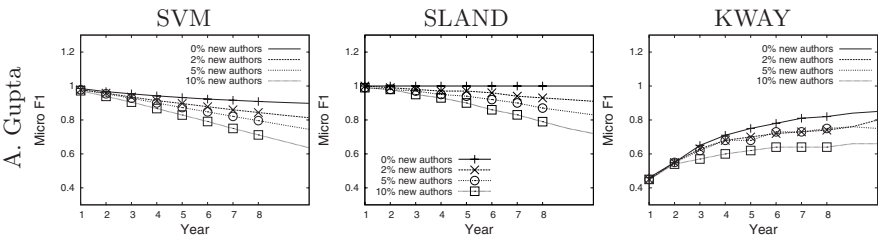

Fig. 1. Evolving DL and Addition of New Authors

The results reported next are averages of five runs, with a standard deviation typically under 5% (and at most 15%) of the mean.

Figure 1 shows the results in each state of the digital library over the ten-year period. There is an increase in the ambiguity for both SVM and SLAND and all values of f (but f=0 for SLAND) with sucessive data loads. Moreover, in any state of the DL, the increase in the ambiguity is higher for larger values of f, as expected. In comparison with SVM, SLAND makes fewer erroneous predictions during its application, dealing better with new authors.

Interestingly, KWAY tends to improve over time, as there is incrementally more information about each author, helping it to better characterize them. However, we also see a trend for performance stabilization typically after the 5^{th} or 6^{th} data load. Nevertheless, KWAY slightly outperforms SVM after ten years for $f > 2\%$, although the improvement does not exceed 10%. In comparison with SLAND, KWAY is inferior in all cases.

4 Conclusions and Future Work

In this paper, we presented SyGAR, a synthetic generator of ambiguous groups of authorship records that is capable of generating synthetic records, and used it to evaluate three state-of-the-art disambiguation methods in scenarios that capture relevant aspects of real-world bibliographic digital libraries.

As future work, we intend to further experiment with other disambiguators and scenarios, enhance SyGAR with more sophisticated mechanisms to add new authors and to dynamically change existing author profiles, and investigate the robustness of several disambiguators to errors in the original input collection.

Acknowledgments. This research is partially funded by projects INCTWeb (grant number 573871/2008-6) and InfoWeb (grant number 55.0874/2007-0), and by the authors' individual grants from CAPES and CNPq.

References

1. Han, H., Giles, C.L., Zha, H., Li, C., Tsioutsiouliklis, K.: Two supervised learning approaches for name disambiguation in author citations. In: JCDL, pp. 296–305 (2004)

2. Han, H., Zha, H., Giles, C.L.: Name disambiguation in author citations using a k-way spectral clustering method. In: JCDL, pp. 334–343 (2005)
3. Huang, J., Ertekin, S., Giles, C.L.: Efficient name disambiguation for large-scale databases. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML/PKDD 2006. LNCS (LNAI), vol. 4213, pp. 536–544. Springer, Heidelberg (2006)
4. On, B.W., Lee, D., Kang, J., Mitra, P.: Comparative study of name disambiguation problem using a scalable blocking-based framework. In: JCDL, pp. 344–353 (2005)
5. Song, Y., Huang, J., Councill, I.G., Li, J., Giles, C.L.: Efficient topic-based unsupervised name disambiguation. In: JCDL, pp. 342–351 (2007)
6. Veloso, A., Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F., Meira Jr., W., Belém, R.: Cost-effective on-demand associative name disambiguation in bibliographic citations. Technical Report RT DCC.001/2009, DCC-UFMG (under review) (2009)