Model Selection Strategies for Author Disambiguation

Roman Kern
Institute of Knowledge Management
Graz University of Technology
Graz, Austria
rkern@tugraz.at

Mario Zechner
Know-Center GmbH
Graz, Austria
mzechner@know-center.at

Michael Granitzer
Know-Center GmbH &
Graz University of Technology
Graz, Austria
mgranitzer@tugraz.at

***Abstract*—Author disambiguation is a prerequisite for utilizing bibliographic metadata in citation analysis. Automatic disambiguation algorithms mostly rely on cluster-based disambiguation strategies for identifying unique authors given their names and publications. However, most approaches rely on knowing the correct number of unique authors a-priori, which is rarely the case in real world settings. In this publication we analyse cluster-based disambiguation strategies and develop a model selection method to estimate the number of distinct authors based on co-authorship networks. We show that, given clean textual features, the developed model selection method provides accurate guesses of the number of unique authors.**

***Keywords*-author disambiguation, model selection**

I. Introduction

Author disambiguation, i.e. identifying unique authors given their names and publications, remains a core challenge for utilizing bibliographic metadata and citation management [1]. Most disambiguation techniques tackling this problem rely on clustering methods [2], [3]. However, an often neglected problem in clustering, which is especially crucial for author disambiguation, is model selection, i.e. detecting the number of unique authors/cluster.

In this publication we address the topic of clustering-based author identification and disambiguation and present the following contributions:

- We developed a new model selection strategy based on authorship co-occurrence
- We compare different cluster-based disambiguation strategies and feature combinations on real world data sets
- We utilize web search engines to extend the features available for disambiguation

As a result, our model selection strategy yields reasonable results in case clean textual features are given. Further, average linking based Hierarchical Agglomerative Clustering (HAC) seem to be the best clustering approach for disambiguation, which raises the need for efficient blocking strategies to decompose the problem into tractable units.

In the following we give an overview over the process of author identification and disambiguation for research papers and present details on features, clustering and model selection afterwards. We end with detailed experiments and a conclusion.

II. Author Identification & Disambiguation

Assuming a given set of documents $\mathcal{D}$, we target the task of extracting all unique authors $\mathcal{A}$ and assigning the correct subset of unique authors $\mathcal{A}_k$ to every publication $d_k \in \mathcal{D}$. Authors are represented in publications or associated publication metadata syntactically via strings; We denote $s_{k,j}$ as author representation of author a_j in document d_k.

Ambiguity arises from the following causes:

1) Two distinct authors $a_i, a_j \in \mathcal{A}$ share the same author representations $s_{*,i} = s_{*,j}$ where the subscript $*$ refers to all possible publications.
2) A single author $a_i = a_j$ is referred to by different orthographic variations $s_{k,i} \neq s_{l,i}$
3) An author has changed her name over time

In the following, we focus on resolving the first two causes, leaving the third cause open for future research. Disambiguating cause 1 and 2 can be broken down into three steps: identification, blocking and disambiguation.

A. Identification

Identification refers to extracting the author representations $\{s_{k,i} \dots s_{k,j}\}$ and corresponding authorship features $\vec{f_k}$ from a publication d_k. The representation of author names may be provided through metadata or may be, for example, extracted through information extraction techniques. Authorship features $\vec{f_k}$ usually contain extractable information such as co-authors, title of the publication, topic etc. needed for later stage of the disambiguation process.

In our implementation we assumed that bibliographic data is given by publishers or aggregators like Mendeley[1]. Bibliographic data includes authors, title, publisher, year etc. In order to obtain full text features for the publication we utilized existing web search engines, namely Google Scholar[2] and Bing[3], and searched for the publication title via

[1]http://mendeley.com
[2]http://scholar.google.com
[3]http://bing.com

phrase searches. The content of the 5 best matching search results was then used as the textual representation of the publication.

B. Blocking

Blocking [4] refers to the process of grouping similar author representations $s_{a,i} \approx s_{b,j}$ and their corresponding authorship features over all publications $\mathcal{D}$ into distinct groups or blocks $\mathcal{B}_l = \{< s_{a,j}, \vec{f_a} > \ldots < s_{b,i}, \vec{f_b} >\}$. The purpose of blocking is to speed up the subsequent disambiguation step by partitioning the potential set of author representations into smaller chunks. Two properties are important for the blocking function: (i) scalability to large sets of strings and (ii) creating blocks containing all author representation, i.e. $s_{*,j}$ of all authors with similar author representations, i.e. $s_{*,j} \approx s_{*,i}, \forall a_j \neq a_i$.

Implementing the blocking process naively by computing all pair-wise comparisons yields a time complexity of $O(n^2)$. Hashing strategies of author names however allow to reduce the time complexity to $O(n)$. The choice of the hash function remains crucial. For example, a phonetic hash function like SOUNDEX may be used to merge author names that sound alike but have different orthographic structures. More complex, similarity preserving hash functions like for example fuzzy fingerprints or Semantic Hashing [5], [6] may utilize an even more tuneable similarity measure.

Note that ideally the similarity measure should consider ambiguities from single authors with different orthographic variations by grouping them into the same block. In general, decreasing the required similarity of author representations within a block increases computational complexity for subsequent disambiguation, but also increases the potentially achievable recall.

In our experiments we used surname information as the blocking criterion. All authors with the same last name are grouped together in one block. This has been possible since the bibliographic data we used distinguished between forename and surname.

C. Disambiguation

Disambiguation refers to the process of identifying the group of author representations $s_{*,k}$ contained in a block $\mathcal{B}_l$ referring to one distinct author $a_k \in \mathcal{A}$.

Our work focuses on solving the disambiguation stage via clustering and on utilizing special model selection strategies for guessing the correct number of authors. The details are presented in the next section.

III. Clustering based Disambiguation and Model Selection

A block consists of a set of publications for similar or identical author representations, e.g. "M. Granitzer" and "Michael Granitzer". The intuition is that each author is publishing within a certain scientific field. Her publications will therefore contain a set of words, keywords or other natural language features which span a specific topic within the field. Clustering the publications within a block should therefore give us sets of publications which cover the same topic.

Therefore, we extract natural language features from the textual representation of the publication, title and author names. By applying OpenNLP [4] we extract tokens, sentences and part of speech tags (see [7]) and stem and normalize the token via the Snowball Stemmer [5].

In order to reduce noise in the textual representation, we applied a keyword extraction algorithm based on calculating page rank related measures of words (see [8]).

The following features have been used to span a vector space with TFIDF weights

- Noun, adjective and adverbs of a publication's plain text(s)
- Keywords extracted from a publication's plain text(s)
- The tokenized title text without any part of speech filtering
- The tokenized author names normalized to lower-case.

Further, for tokens, keywords and title text we applied all different combinations of stemming and normalization.

Clustering based on textual features alone is not sufficient in the case that two distinct authors share a similar name and publish in the same field. We therefore also exploit co-authorship information and other metadata (e.g. publication year) to be able to distinguish between distinct authors in such a case.

Clustering: In order to compare different clustering methods for disambiguation, we applied the Batch K-Means, Growing K-Means [9] - utilizing the K-Means++ seeding strategy [10] in both cases - and the Hierarchical Agglomerative Clustering (HAC) using single, complete and average linking. All clustering algorithms utilized the simple cosine similarity measure which has been proven valuable in text clustering.

Model Selection: Model selection, i.e. guessing the correct number of clusters, is a challenging, and often neglected problem. Results on author disambiguation reported in the academic literature mostly assume the correct number of clusters (i.e. unique authors) to be given, which is not the case in reality. Hence, we have developed a co-author based model selection strategy along with additional implementation details.

The initial model selection strategy we used was a stability-based criterion as proposed in [11]. However, this criterion showed very low accuracy of guessing the correct k, which led us to the development of two new model selection strategies specific to disambiguating authors of scientific publications.

[4] http://opennlp.sourceforge.net/

[5] http://snowball.tartarus.org/

Both approaches are motivated as follows: the clustering of publications is performed for a large range of cluster numbers, usually from a single cluster to one cluster per publication. Each clustering will group textually similar publications of a block. To decide which clustering result is the best fit we want to use external information, in this case co-authorship information. The more co-authors within a cluster overlap and the less they are spread over other clusters the more probable it is that this cluster represents a single author. We perform this for each cluster in a clustering and average the result. We then can compare these measures for each clustering and select the one with the best co-author overlap.

We implemented this strategy with two statistical methods: conditional probabilities and point-wise conditional entropy.

In the conditional probabilities case we calculate the (inverse) conditional probability of one co-author belonging to a given cluster (1 - $P(author|cluster)$) as well as the conditional probability of a cluster containing a given co-author (1 - $P(cluster|author)$). We average both inverse conditional probabilities for all author/cluster combinations ($P_{avg}(author|cluster), P_{avg}(cluster|author)$) and merge these two probabilities via a simple arithmetic mean ($(P_{avg}(author|cluster) + P_{avg}(cluster|author))/2$) to arrive at a final model selection value for the clustering. The clustering with the minimum model selection value is selected as the final model of the block disambiguation.

The point-wise conditional entropy approach works exactly the same, we just replace the probabilities with entropies. We define point-wise conditional entropy as follows (it shares same intuition behind point-wise mutual information, to incorporate only the positive samples, in this case to use only clusters a co-author is found at least a single time, denoted as $Clusters_{Co-Author}$):

$$H_{pointwise}(Y|X) \stackrel{\text{def}}{=} -\sum_{C} p(x,y) \log\, p(y|x) \quad (1)$$

$$C = \{c|c \in Clusters_{Co-Author}\} \quad (2)$$

Figure 1 shows the 3 curves generated via the point-wise conditional entropy on a test dataset with 16 distinct authors. The y-axis depicts the entropy, the x-axis the number of clusters.

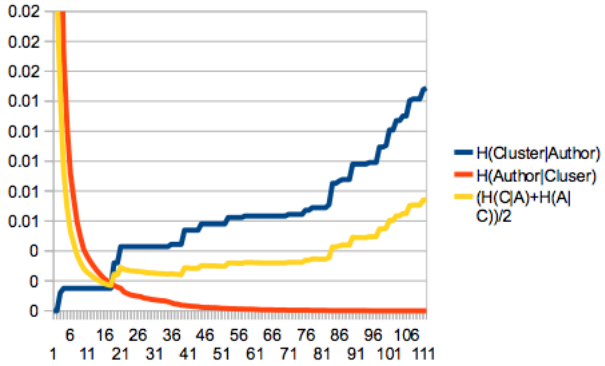

Figure 1. Model selection output for the Martin subset of the Giles set (16 distinct authors). The x-axis represents the number of clusters, the y-axis the point-wise conditional entropy.)

IV. Experiments

A. Data Set

For evaluating the disambiguation quality of the research prototype we use two gold standards: the Giles dataset and a dataset provided by Mendeley Ltd.

1) Giles Dataset: The Giles dataset (c.f. [12]) is a publicly available gold standard for evaluating author disambiguation in the context of scientific citations. Ten highly ambiguous author names were selected from the Citeseer database. For each of the ten author names a set of publications was created and manually disambiguated. For each author a text file encodes a number of publications in form of citations. The citations give information on the authors, the title and venue of the cited publication. Each citation has a label identifying the distinct author. There is a total of 8453 citations in the dataset, split among the ten author names.

2) Mendeley Dataset: The Mendeley dataset was provided by Mendeley and was created from author profiles found in the Mendeley system curated by the respective author herself. The raw dataset consists of author profile information of distinct authors and a list of ids of the publications that belong to a specific author. While the whole data set contains over 4000 authors and 43984 publications, our experiments have been conducted only on a subset (i.e. block) of all authors with highly ambiguous names like "Lee".

B. Results

We performed two types of evaluation on the datasets. In the first evaluation we test the performance of the clustering itself, fixing the number of clusters at the real number of authors of a block. We vary the used clustering algorithm as well as the features used to create the vector spaces. This evaluation step should tell us which algorithm along with which feature combination is best suited. Since the model selection only selects one of the clusterings produced the output of this evaluation stage is also the best result achievable after model selection.

1) Clustering Results: We present the clustering evaluation results on selected subsets of the Giles and Mendeley datasets, with a focus on identifying suitable feature combinations. Each subset relates to one block $\mathcal{B}_l$ of authors which share the same surname. We present the results on the Martin, Kumar and Gupta subsets of the Giles dataset along with the results on the Lee subset of the Mendeley

Author	Title	Keyword	Stem	Normalize	Purity	F1
✓	✓	✓		✓	0.92	0.9
✓	✓	✓	✓	✓	0.91	0.89
		✓		✓	0.87	0.84
		✓			0.87	0.84
✓		✓	✓	✓	0.88	0.83

Table I
BEST 5 RESULTS USING HAC CLUSTERING ON THE GILES-MARTIN SUBSET. SORTED BY F1. 16 DISTINCT AUTHORS, 112 PUBLICATIONS.

Author	Title	Keyword	Stem	Normalize	Purity	F1
✓	✓	✓	✓	✓	0.72	0.50
✓	✓	✓		✓	0.70	0.48
✓	✓	✓	✓		0.70	0.46
	✓	✓	✓	✓	0.61	0.42
✓	✓	✓			0.66	0.39

Table II
BEST 5 RESULTS ON HAC CLUSTERING ON THE GILES LEE SUBSET. SORTED BY F1. 100 DISTINCT AUTHORS, 1419 PUBLICATIONS.

dataset. We selected subsets in a way that represent the best and worst cases.

We refrain from reproducing the results for all clustering algorithm and feature combinations and concentrate on the behaviour of the overall best performing clustering algorithm with regards to different feature combinations. In our experiments on the above data sets we found that partitioning clustering algorithms (Batch and Growing K-Means) did exhibit worse performance than the HAC alternatives in all cases and hence we report only results obtained by the HAC Algorithm with average linking strategy here. We use purity and the F1 measures to assess the quality of the clustering.

Tables I-III provide the detailed results for the different features Author, Title, Keywords, and whether Stemming and/or Normalization have been applied.

The best overall feature combinations over all datasets all contain the authors, titles and keywords. Stemming and normalization only have a marginal impact on the clustering performance.

Due to the varying accuracy, we analysed the textual representation of the different subsets. Especially the results on the Giles Martin subset are very encouraging. However, we attribute this to the quality of the plain text we fetched in the web searches for this dataset. The plain texts for each publication are very clean and contain information specific

Aut.	Tit.	Keyw.	Stems	Norm.	Purity	F1
✓	✓	✓	✓	✓	0.76	0.61
✓	✓	✓	✓		0.75	0.56
✓	✓	✓			0.73	0.51
✓	✓	✓		✓	0.73	0.5
	✓	✓	✓	✓	0.67	0.5

Table III
BEST 5 RESULTS ON HAC CLUSTERING ON THE MENDELEY LEE SUBSET. SORTED BY F1. 49 DISTINCT AUTHORS, 217 PUBLICATIONS.

Dataset	K_{real}	$F1_{best}$	K_{guess}	$F1_{guess}$	$F1_{real}$
Mendely-lee	49	61%	44	28%	27%
Giles-martin	16	90%	16	84%	84%
Giles-gupta	26	65%	14	43%	65%
Giles-kumar	14	70%	14	44%	44%
Giles-chen	61	46%	12	10%	37%
Giles-johnson	15	78%	11	60%	75%
Giles-lee	100	50%	21	5%	38%

Table IV
MODEL SELECTION RESULTS USING POINT-WISE CONDITIONAL ENTROPY ON KEYWORDS ONLY

to that publication.

This is not true for the other datasets, for which we could not fetch high quality plain texts due to getting blocked by various high quality sites. This is also supported by the fact that full-text tokens did not provide significant results on any of the different subsets, except for the Giles Martin subset. Hence we omitted them from being presented in the results table.

In the case of the Mendeley dataset the search results were not as distinctive as in the case of the Giles dataset. The reason for this is the quality of the titles in that dataset which we use for the plain text search. These are very noisy and thus not suitable for web searches as well as the clean titles of the Giles dataset.

Based on the above results we recommend using the features "Authors" "Title" and "Keywords" in a weighted manner with stemming and normalization enabled or disabled depending on the corpus. We believe that high quality plain texts for the other datasets will yield similar results as produced for the Giles Martin subset above.

V. MODEL SELECTION RESULTS

We evaluated both our model selection strategies (conditional probabilities and point-wise conditional entropy) on a subset of the ground truth datasets. We used the average link HAC implementation and limited our evaluation to the feature combinations of taking only keywords into account and a combination of all features (normalized). Combining all features have shown best performance previously. Since there is an overlap between the features for model selection and clustering we focused on disjunctive features for the clustering and the model selection step.

Tables IV and V present the results for point-wise conditional entropy; For the conditional probability we present the results in table VI. In the tables K_{real} depicts the real number of authors, $F1_{best}$ the best achieved F1 measure for any feature combination using the real number of authors, while $F1_{real}$ provides the $F1$ measure using either all features or keywords only on the real number of authors. K_{guess} shows the number of authors guessed and $F1_{guess}$ the corresponding $F1$ measure.

The first observation is that model selection works well with datasets that have clean textual features (i.e. the Martin

Dataset	K_{real}	$F1_{best}$	K_{guess}	$F1_{guess}$	$F1_{real}$
Mendely-Lee	49	61%	18	10%	50%
Giles-Martin	16	90%	29	68%	90%
Giles-Gupta	26	65%	80	42%	63%
Giles-Kumar	14	70%	36	53%	70%
Giles-Chen	61	46%	48	38%	42%
Giles-Johnson	15	78%	39	77%	78%
Giles-Lee	100	50%	47	13%	48%

Table V
MODEL SELECTION RESULTS USING POINT-WISE CONDITIONAL ENTROPY AND ALL FEATURES

Dataset	K_{real}	$F1_{best}$	K_{guess}	$F1_{guess}$	$F1_{real}$
Mendely-Lee	49	61%	44	28%	27%
Giles-Martin	16	90%	18	82%	84%
Giles-Gupta	26	65%	26	65%	65%
Giles-Kumar	14	70%	45	71%	44%
Giles-Chen	61	46%	85	37%	37%
Giles-Johnson	15	78%	26	75%	75%
Giles-Lee	100	50%	22	6%	38%

Table VI
MODEL SELECTION RESULTS USING CONDITIONAL PROBABILITY AND FEATURE COMBINATION FOR KEYWORDS ONLY

subset). These features have been obtained by crawling the publisher sites directly and not by using search engines. However, in general such an approach is not possible, since access to the publisher cite is usually restricted.

For keyword features only the point-wise conditional entropy approach underestimates the number of authors while the conditional probability model selection tends to overestimate that number.

In the case of all features the point-wise conditional entropy approach also overestimates the real number of authors. This can be attributed to the fact that both the clustering and the model selection rely on the same feature, namely the authors of a publication.

In summary model selection varies greatly among datasets and feature combination. However, in case of clean textual features both model selection strategies given reasonable good results. Depending on the use case one might resort to using conditional probability model selection if the number of clusters is of less concern than the purity of clusters.

VI. CONCLUSION & FUTURE WORK

Our results show, that given clean textual features, the developed model selection provides appropriate guesses of unique authors within one block. Further, HAC based clustering approaches outperform partition based clustering in disambiguation tasks, which is opposite to results from standard text clustering and raises the question for efficient blocking methods. However, disambiguation remains challenging and future work will focus on entity disambiguation beyond authors of research papers as well as increasing accuracy through improved feature selection and cleaning.[6]

REFERENCES

[1] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, *Two supervised learning approaches for name disambiguation in author citations*. New York, New York, USA: ACM Press, 2004.

[2] H. Zha, X. He, C. Ding, M. Gu, and H. Simon, "Spectral relaxation for k-means clustering," *Advances in Neural Information Processing Systems*, vol. 2, p. 10571064, 2002.

[3] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[4] B.-W. On, D. Lee, J. Kang, and P. Mitra, "Comparative study of name disambiguation problem using a scalable blocking-based framework," in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05*. New York, New York, USA: ACM Press, 2005, p. 344.

[5] B. Stein, "Fuzzy-fingerprints for text-based information retrieval," in *Proceedings of the 5th International Conference on Knowledge Management IKNOW 05 Graz Journal of Universal Computer Science*, 2005, pp. 572–579.

[6] R. Salakhutdinov and G. Hinton, "Semantic hashing," *International Journal of Approximate Reasoning*, vol. 50, no. 7, pp. 969–978, Jul. 2009.

[7] J. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," in *Proceedings of the 5th Conference on Applied Natural Language Processing*, 1997.

[8] R. Mihalcea and P. Tarau, *TextRank: Bringing Order into Texts*, 2004, pp. 404–411.

[9] M. Daszykowski, B. Walczak, and D. L. Massart, "On the optimal partitioning of data with k-means, growing k-means, neural gas, and growing neural gas." *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 6, pp. 1378–1389, 2002.

[10] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[11] T. Lange, M. L. Braun, V. Roth, and J. M. Buhmann, "Stability-based model selection," in *In Advances in Neural Information Processing Systems*. MIT Press, 2002, pp. 617–624.

[12] H. Han, H. Zha, and C. L. Giles, "Name Disambiguation in Author Citations using a K-way Spectral Clustering Method," in *JCDL*. Denver: ACM, 2005, pp. 334–343.

[6]Acknowledgements: This work has been partially funded by the European Commission as part of the TEAM IAPP project (grant no. 251514) within the FP7 People Programme. The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG). We also thank Mendeley Ltd. for preparing the data for the experiments.