

Name Disambiguation in Author Citations using a K-way Spectral Clustering Method

Hui Han^{1,2}
¹Yahoo! Inc.
701 First Avenue
Sunnyvale, CA, 95129
huihan@yahoo-inc.com

Hongyuan Zha²
²Department of Computer
Science and Engineering
The Pennsylvania State
University
University Park, PA, 16802
zha@cse.psu.edu

C. Lee Giles^{2,3}
³School of Information
Sciences and Technology
The Pennsylvania State
University
University Park, PA, 16802
giles@ist.psu.edu

ABSTRACT

An author may have multiple names and multiple authors may share the same name simply due to name abbreviations, identical names, or name misspellings in publications or bibliographies (citations)¹. This can produce name ambiguity which can affect the performance of document retrieval, web search, and database integration, and may cause improper attribution of credit. Proposed here is an unsupervised learning approach using K -way spectral clustering that disambiguates authors in citations. The approach utilizes three types of citation attributes: co-author names, paper titles, and publication venue titles². The approach is illustrated with 16 name datasets with citations collected from the DBLP database bibliography and author home pages and shows that name disambiguation can be achieved using these citation attributes.

Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Name Disambiguation, Feature Selection, Unsupervised Learning, Spectral Clustering

1. INTRODUCTION

Name disambiguation can have several causes. Because of name variations, identical names, name misspellings or pseudonyms, two

¹<http://www.library.umass.edu/reference/glossary.html#cite>

²By “publication venue titles”, we mean the titles of known publication sources, such as proceedings and journals.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL'05, June 7–11, 2005, Denver, Colorado, USA.

Copyright 2005 ACM 1-58113-876-8/05/0006 ...\$5.00.

types of name ambiguities in research papers and bibliographies (citations) can be observed. The first type is that an author has multiple name labels. For example, the author “David S. Johnson” may appear in multiple publications under different name abbreviations such as “David Johnson”, “D. Johnson”, or “D. S. Johnson”, or a misspelled name such as “Davad Johnson”. The second type is that multiple authors may share the same name label. For example, “D. Johnson” may refer to “David B. Johnson” from Rice University, “David S. Johnson” from AT&T research lab, or “David E. Johnson” from Utah University (assuming the authors still have these affiliations).

Name ambiguity may affect the quality of scientific data gathering, can decrease the performance of information retrieval and web search, and even may cause incorrect identification of and credit attribution to authors. In the web DBLP (Digital Bibliography & Library Project) [1], for example, we found that the author page of “Yu Chen” in the DBLP contains citations from three different people with the same name: Yu Chen from University of California, Los Angeles, Yu Chen from Microsoft at Beijing branch, and Yu Chen as the senior professor from Renmin University of China. Another example of the errors in DBLP is in the author page that refers to “Jia Li” from the Department of Statistics at the Pennsylvania State University. However, the “Home Page” link in her author page erroneously points to a faculty member from Department of Mathematical Sciences at the University of Alabama in Huntsville with the identical name. In another well-known web site, CiteSeer [23], we also observe several errors. For example, “D. Johnson” is ranked as the most cited author in Computer Science according to CiteSeer’s statistics in June 2003³. However, the citation number that “D. Johnson” obtained in CiteSeer’s statistics is actually the sum of several different authors such as “David B. Johnson”, “David S. Johnson”, and even “Joel T. Johnson”.

Given a set of citations that have the same name label, how do we disambiguate authors if the name label refers to a single author, or different authors with identical names? We consider two approaches: supervised and unsupervised machine learning. In supervised learning, each canonical author name⁴ can be considered as a class and name disambiguation then classifies citations into their author classes [26]. However, supervised learning methods need labeled data and do not always have the authors’ previous citations or identification information to train the classifiers. With

³<http://citeseer.ist.psu.edu/mostcited.html>

⁴A name that is the minimal invariant and complete name entity for disambiguation. [26]

unsupervised learning methods, we would not need labeled data for training. The name disambiguation problem can be formulated as partitioning collections of citations into clusters, with each cluster containing only citations authored by the same author, thus disambiguating authorship in citations to induce author name identities.

We propose using K -way spectral clustering [52], a graph model that has been successfully applied to data mining and cluster analysis, for name disambiguation in citations as described in detail in section 3. Table 1 shows an example of partial citation clusters with disambiguated authorships resulting from our algorithm. Complete citation clusters are not shown due to space limitations.

The rest of the paper is organized as follows: section 2 discusses prior work; section 3 introduces the K -way spectral clustering method; section 4 reports experiments and results; and section 5 concludes and discusses future work.

Cluster	Author citations
1	Rapid Profiling via Stratified Sampling, S. Sastry, R. Bodik, J. E. Smith , 28th Int. Symposium on Computer Architecture, 2001.
	Relational Profiling: Enabling Thread-Level Parallelism in Virtual Machines, Timothy Heil and J. E. Smith , 33rd Int. Symp. on Microarchitecture, 2000.
	Concurrent Garbage Collection using Hardware Assisted Profiling, Timothy Heil and J. E. Smith , International Symposium on Memory Management, 2000.
2	Smith, James E. , "Moment Methods for Decision Analysis", Management Science 39 (1993).
	Smith, James E. , "Generalized Chebychev Inequalities: Theory and Applications in Decision Analysis", Operations Research 43 (1995).
	Smith, James E. , Samuel Holtzman and James E. Matheson, "Structuring Conditional Relationships in Influence Diagrams", Operations Research 41 (1993).
3	Henry E.J. and Smith J.E. 2002. The Effect of SurfaceActive Solutes on Water Flow and Contaminant transport in Variably Saturated Porous Media with Capillary Fringe Effects. Journal of Contaminant Hydrology.
	Henry E.J., Smith J.E. , and Warrick A.W. 2002. Two-Dimensional Modeling of Flow and Transport in the Vadose Zone with Surfactant-Induced Flow. WATER RESOURCES RESEARCH.
	Smith, J.E. and Zhang F.Z. 2001. Determining Effective Interfacial Tension and Predicting Finger Spacing for DNAPL Penetration into Water-Saturated Porous Media. Journal of Contaminant Hydrology.

Table 1: Partial citation clusters of three disambiguated authors of the same name label "J. E. Smith".

2. PRIOR WORK

Name ambiguity is a special case of the general problem of *identity uncertainty*, where objects are not labeled with unique identifiers [37]. Much research has been done to address the identity uncertainty problem using different methods, such as record linkage [21], duplicate record detection and elimination [10, 31, 35], merge/purge [27], data association [8], database hardening [13], citation matching [34, 34], name matching [9, 45, 11], name equivalence identification [20], address matching [15], and name authority control in library cataloging practice [48, 17, 24]. At the concept level these methods include word sense disambiguation [47, 30].

Name authority control, name matching, and name equivalence identification are the work most similar to ours. Name authority control aims to find the authoritative form of names, i.e., the unambiguous reference to an individual [17]. Getty's ULAN (Union

List of Artist's Names) [2] and the Library of Congress name authority file [3] are good examples of such authorized names. Name authority control usually provides a set of rules and standardized terms for consistent name representation, e.g. the form of the name to be used. A "canonical name" [26] includes an authorized name as a special case. Though much work in name authority control has used manual analysis, automated systems are being considered [17, 26]. Such automated systems use supervised learning methods, relying much on a priori knowledge of ambiguous name entities or name word lists.

Name matching [9, 11, 13, 45] usually identifies a name entity with different name labels from duplicate records of different syntactic formats. For example, "Bart Selman" and "B. Selman" are ambiguous name labels of the same person who authored the work cited as "Critical behavior in satisfiability" [13]. Name matching does not focus on the case of different name entities which have identical name labels. Our method disambiguates names from different records (citations) authored by the same name entity, and addresses both types of name ambiguities previously mentioned.

Name equivalence identification [20] addresses both types of name ambiguities previously mentioned. The work derives heuristic rules from purely name strings to identify equivalent names, i.e., names that refer to the same person. Our method exploits person identity information from sources that is not limited to person names, such as coauthor names, paper titles, and publication venue titles. In digital libraries, publications usually reflect research fields of the authors. Authors are often seen to coauthor with a certain group of other authors. Our method works in conjunction with previous work on name matching and name equivalence identification, which usually use string-based comparisons to induce author identity and addresses name misspellings and abbreviations.

Clustering methods appear to be a natural solution for disambiguation problems. Feitelson [20] uses cliques to represent a group of names that refer to the same person in his name equivalence identification work. In the task of word sense disambiguation, a sense is often seen to correspond to a cluster, and instances of words with the same sense are expected to be part of the same cluster [38, 14, 32, 18, 33, 49]. K-means, naive Bayes and Gaussian mixture model are widely used clustering methods. However, these methods are prone to local minima, and initial data partitions can seriously impact the clustering results [50]. Spectral clustering methods use eigen-decomposition techniques and find an approximation of the global optimal solution in terms of defined criteria function [50, 52]. Spectral clustering is often found to give better results than traditional clustering methods, e.g. k-means [50, 52].

Our contribution is the selection of features for name disambiguation and a novel application of a K -way spectral clustering method to name disambiguation in author citations. Through extensive experimentation, we gain insights in the factors that affect the name disambiguation performance, and propose possible solutions for disambiguation performance improvement.

3. K-WAY SPECTRAL CLUSTERING WITH QR DECOMPOSITION

Spectral clustering methods compute eigenvalues and eigenvectors of a Laplacian matrix (or singular values and singular vectors of certain matrix) related to the given graph, and construct data clusters based on such spectral information [19, 28, 36, 40, 42]. Recent research on theoretical understanding of spectral methods found that important algebraic structures in general exist in the eigenvectors and in the singular vector matrices for data clusters [4, 52]. In particular, Zha et. al [52] found that minimizing a sum-of-

square cost function can be reformulated as a trace maximization problem associated with the Gram matrix of the data vectors. They show that a partial eigen decomposition of the Gram matrix obtains the *global* optimal solutions for a relaxed version of the trace maximization problem. Accordingly, the cluster assignment for each data vector can be found by computing a pivoted QR decomposition of the eigenvector matrix. The K -way spectral clustering with QR decomposition is shown in their experiments to outperform the K -means algorithm [52].

As motivation for our work, we tried the K -means algorithm on our web collected publication list dataset (described in detail in section 4.1), together with LSI dimension reduction and variations of feature weight assignments. The worse performance achieved on these datasets by using k -means algorithm compared to the spectral clustering conforms with previous practice [50, 52].

Next we describe the spectral clustering method for experiments to cluster citations with the same name label but different authors. We model each citation as a node in an undirected graph. Each edge (i, j) in the graph is assigned a weight that reflects the similarity between two citations i and j . The name disambiguation problem for author citations is defined as a partition of the graph so that citations that are more similar to each other, e.g. authored by the same author, belong to the same cluster.

3.1 Citation Matrix and Feature Design

We observe that an author’s citations usually reveal his or her identification information, such as the author’s research area, and his or her individual patterns of co-authoring. We use three types of citation attributes to design features for name disambiguation: co-author names, paper titles, and publication venue titles. A feature is a component of a citation attribute, e.g., one co-author name or one pre-processed word in the title of a paper or publication venue. It should be noted that our technique can also be extended to use other information, e.g., the affiliations and addresses of authors.

We construct citation vectors for each name dataset. With m features in the name dataset, each citation can be represented as a m -dimensional vector, i.e., $M = (\alpha_1, \dots, \alpha_m)$. If the i th feature in the dataset appears in citation M , α_i is the feature i ’s weight. Otherwise, $\alpha_i = 0$. We study two types of feature weight assignment, the usual “TFIDF”; and the normalized “TF” (“NTF”), where $ntf(i, d) = freq(i, d) / \max(freq(i, d))$ refers to the term frequency of feature i in a citation d . $\max(freq(i, d))$ refers to the maximal term frequency of feature i in any citation d . With the “NTF” scheme, the weight of features with different ranges of values is normalized. The normalized “TF” scheme has been shown to improve the classification performance [25]. Using completely unsupervised learning methods, we do not have training data to learn the weights for different type of features. However, we propose combining supervised learning methods in our future work for automatic feature weight assignment. The Gram matrix of the citation vectors represents the pairwise cosine similarities between citations. We apply the K way spectral clustering algorithm to the Gram matrix as described in the following two subsections.

3.2 Spectral Relaxation

Given a set of m -dimensional citation vectors $\alpha_i, i = 1, \dots, n$, we form the m -by- n citation matrix $A = [\alpha_1, \dots, \alpha_n]$. A partition Π of the citation vectors can be written in the following form

$$AE = [A_1, \dots, A_k], A_i = [\alpha_1^{(i)}, \dots, \alpha_{s_i}^{(i)}], \quad (1)$$

where E is a permutation matrix, and A_i is m -by- s_i , i.e., the i th cluster contains the citation vectors in A_i . For a given partition Π in Equation 1, the associated sum-of-squares cost function is defined

as

$$ss(\Pi) = \sum_{i=1}^k \sum_{s=1}^{s_i} \|\alpha_s^{(i)} - m_i\|^2, m_i = \sum_{s=1}^{s_i} \alpha_s^{(i)} / s_i, \quad (2)$$

i.e., m_i is the mean vector of the citation vectors in cluster i . It was shown in [52] that the minimization of the above sum-of-square cost function can be formulated as a relaxed maximization problem

$$\max [trace(X^T A^T A X)], \quad (3)$$

where $X^T X = I_k$ and X can be an arbitrary orthonormal matrix. It turns out that the above trace maximization problem has a closed-form solution.

Theorem. (Ky Fan) *Let H be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, and the corresponding eigenvectors $U = [u_1, \dots, u_n]$. Then*

$$\lambda_1 + \dots + \lambda_k = \max_{X^T X = I_k} trace(X^T H X). \quad (4)$$

Moreover, the optimal X^* is given by $X^* = [u_1, \dots, u_k]Q$ with Q an arbitrary orthogonal matrix.

It follows from the above theorem that we need to compute the largest k eigenvectors of the Gram matrix $A^T A$. Let X_k be the n -by- k matrix consisting of the largest eigenvectors of $A^T A$. Each row of X_k corresponds to a citation vector, and the above process can be considered as transforming the original citation vectors in a m -dimensional space to new citation vectors in the k -dimensional space. However, the goal here is not to reconstruct the citation matrix using a low-rank approximation but rather to capture its cluster structure, as shown in the next subsection.

3.3 Cluster Assignment Using Pivoted QR Decomposition

Assume that the best partition of the citation vectors in A that minimizes $ss(\Pi)$ is given by $A = [A_1, \dots, A_k]$, where each sub matrix A_i corresponds to a cluster. The Gram matrix of A can be written as

$$A^T A = \begin{pmatrix} A_1^T A_1 & 0 & \cdot & 0 \\ 0 & A_2^T A_2 & \cdot & 0 \\ 0 & 0 & \cdot & A_k^T A_k \end{pmatrix} + E \equiv B + E. \quad (5)$$

When the overlap among clusters represented by the sub matrices A_i is small, the norm of E will be small compared with the block diagonal matrix B in the above equation. Let the largest eigenvector of $A_i^T A_i$ be y_i , and

$$A_i^T A_i y_i = u_i y_i, \|y_i\| = 1, i = 1, \dots, k, \quad (6)$$

then the columns of the matrix

$$Y_k = \begin{pmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_k \end{pmatrix} \begin{pmatrix} y_1 & & & \\ & y_2 & & \\ & & \ddots & \\ & & & y_k \end{pmatrix} \quad (7)$$

span an invariant subspace of B . Let the eigenvalues and eigenvectors of $A^T A$ be $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n, A^T A x_i = \lambda_i x_i, i = 1, \dots, n$. After some manipulation, it can be shown that

$$X_k^T \equiv [x_1, \dots, x_k] = Y_k V + O(\|E\|), \quad (8)$$

where V is an k -by- k orthogonal matrix. Ignoring the $O(\|E\|)$ term, we see that

$$X_k^T = \underbrace{[y_{11}v_1, \dots, y_{1s_1}v_1, \dots]}_{\text{cluster 1}}, \underbrace{[y_{k1}v_k, \dots, y_{ks_k}v_k]}_{\text{cluster k}}, \quad (9)$$

where $y_i^T = [y_{i1}, \dots, y_{is_i}]$, and $V^T = [v_1, \dots, v_k]$. A key observation is that all v_i are orthogonal to each other. Once we have selected a v_i , we can jump to other clusters by looking at the orthogonal complement of v_i . Also notice that $\|y_i\| = 1$, so the elements of y_i can not be all small. A robust implementation of the above idea can be obtained as follows: we pick a column of X_k^T which has the largest norm, say, it belongs to cluster i ; we then orthogonalize the rest of the columns of X_k^T against this column. For the columns belonging to cluster i the residual vector will have small norm, and for the other columns the residual vectors will tend to be not small. We then pick another vector with the largest residual norm, and orthogonalize the other residual vectors against this residual vector. The process can be carried out k steps, and it turns out to be exactly QR decomposition with column pivoting applied to X_k^T , i.e., we find a permutation matrix P such that

$$X_k^T P = QR = Q[R_{11}, R_{12}], \quad (10)$$

where Q is a k -by- k orthogonal matrix, and R_{11} is a k -by- k upper triangular matrix. We then compute the matrix

$$\hat{R} = R_{11}^{-1}[R_{11}, R_{12}]P^T = [I_k, R_{11}^{-1}R_{12}]P^T. \quad (11)$$

The cluster membership of each citation vector is determined by the row index of the largest element in absolute value of the corresponding column of \hat{R} .

4. EXPERIMENTS ON CITATION DATASETS

4.1 Datasets Used

We collected two types of citations in different ways for experiments. The first type of citations are downloaded from the DBLP Computer Science bibliography which contains more than 400,000 citation records with parsed citation attributes in the XML format. We formed the three attributes in each citation as a string, and then clustered author names with the same first name initial and the same last name. Each name is associated with the citations where the name appears. We sorted the formed name clusters by the number of name variations contained. Top ranked ambiguous names are popular names from Asia, such as “J. Lee”, “S. Lee”, “Y. Chen” and “C. Chen”. Besides these four name datasets, we also used other 10 sets of ambiguous names from the DBLP bibliography as shown in Table 2. The other type of citation database has been manually extracted from publication lists from researchers homepages resulting from “J Anderson” and “J Smith” queries into a Search Engine. This type of citation contains two name sets: 15 “J Anderson” of 229 citations and 11 “J Smith” of 338 citations. These authors have diverse research areas and probably more than the typical authors in the DBLP bibliography. The complete datasets of 16 names is available upon request.

For evaluation, we carefully manually labeled the canonical name entities and associated citations. Citations listed in an author’s publication home page are considered as being written by the same author. Authors with the same name and same affiliation, or same email address are considered to be the same. Authors of the same name that also have the same co-author names (in a complete name format) are very likely the same author. Citations that have the same name label, and are about the same topic are likely to be written by the same author. We also sent emails to some authors to confirm their authorship of citations. The citations for which we had insufficient information to be judged were eliminated. Moreover, we populated the datasets with publication lists downloaded from

the available home page URLs of authors in the datasets. Duplicate citations were detected and removed using CiteSeer’s citation matching algorithm [23].

We used regular expression matching and manual correction to parse the citations collected from web pages. Citation attributes can also be extracted by other methods such as rule-based system [12], hidden Markov models [41, 43, 44], or Support Vector Machines [25]. We pre-processed all datasets as follows. All the author names in the citations were simplified to first name initial and last name. For example, “Yong-Jik Kim” was simplified to “Y. Kim”. A reason for such simplification is that the first name initial and last name format is popular in citation records. Since more name information usually helps name entity disambiguation, we think that insufficient name information from a simplified name format would be good test for evaluating our algorithm. Besides, such simplified name representation helps constructing ambiguous name datasets, and may avoid in some cases of name misspellings. We stemmed the title words of publication venues using the Porter’s stemmer [22], and removed the stop words such as “a”, “the”, etc. We also added from the DBLP bibliography⁵ full names of the abbreviated publication venue titles.

4.2 Experiment Design

For each name dataset, we vary the size of the datasets in two different ways. The first selects the authors associated with at least a minimal number of citations (as shown by the columns of Table 2). The second randomly selects a percentage (from 10% through 100%, with step size of 10%) of the citations of each author from the dataset containing authors that have at least 10 citations. We compare the disambiguation accuracy achieved in each size variation of the datasets to study the effect of dataset size on name disambiguation. In each size variation of the dataset, we applied the K -way spectral clustering algorithm and we compare two schemes of feature weighting: the “TFIDF” and “NTF” schemes. We also study the contribution of each citation attribute on name disambiguation, by using co-author names alone, paper title words alone, and publication venue title words alone, respectively. Then we investigate the effect of the amount of name information on disambiguation, by representing the first name with first name initial and first three characters of the first name, respectively. As the choice of number of clusters could be an important yet separate research issue, and is not the focus of our current work, we pre-defined the number of clusters as labeled. That is, if there are N correct clusters, the dataset is clustered into N clusters.

4.3 Evaluation Method

We evaluate experimental results based on the confusion matrix, where $A[i, j]$ represents the number of “Author i ” predicted as “Author j ” in matrix A . $A[i, i]$ represents the number of correctly predicted names for “Author j ”. We define the disambiguation accuracy as the sum of diagonal elements divided by the total number of elements in the matrix.

4.4 Name disambiguation on DBLP citations

4.4.1 Effect of dataset size on name disambiguation

Disambiguation accuracy, as shown in Figure 1, changes with the two types of dataset size variations, as described in section 4.2. For each dataset from the second type of size variation, we report the average accuracy of 10 times experiments, where in each experiment we randomly select a certain percentage of the citations of

⁵<http://www.informatik.uni-trier.de/~ley/db/conf/indexa.html> and <http://www.informatik.uni-trier.de/~ley/db/journals/index.html>

Name	≥ 2		≥ 3		≥ 4		≥ 5		≥ 6		≥ 7		≥ 8		≥ 9		≥ 10	
	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N	C
A. Gupta	26	577	22	569	18	557	17	553	17	553	17	553	14	532	12	516	11	507
A. Kumar	14	244	11	238	9	232	7	224	7	224	6	218	6	218	5	210	5	210
C. Chen	61	800	50	778	40	748	35	728	29	698	27	686	25	672	22	648	20	630
D. Johnson	15	368	11	360	10	357	9	353	8	348	7	342	6	335	6	335	6	335
J. Lee	100	1417	91	1399	58	1300	55	1288	46	1243	44	1231	40	1203	38	1187	38	1187
J. Martin	16	112	13	106	11	100	6	80	6	80	5	74	5	74	4	66	4	66
J. Robinson	12	171	10	167	8	161	8	161	7	156	7	156	7	156	6	148	6	148
J. Smith	31	927	25	917	21	905	19	897	17	887	16	881	15	874	14	866	12	848
K. Tanaka	10	280	9	278	8	275	8	275	6	275	6	265	5	265	5	258	5	258
M. Brown	13	153	13	153	10	144	8	136	7	131	7	131	7	131	5	115	5	115
M. Jones	13	259	12	257	11	254	10	259	9	245	9	245	9	245	6	221	6	221
M. Miller	12	412	10	408	7	399	7	399	5	389	5	389	5	389	5	389	5	389
S. Lee	86	1458	74	1439	56	1385	45	1341	40	1316	38	1304	36	1290	36	1290	36	1290
Y. Chen	71	1264	61	1244	48	1205	42	1181	36	1151	30	1115	27	1094	25	1078	22	1051

Table 2: The 14 DBLP name datasets varied by size. “ $\geq i$ ” means that the dataset contains authors who have at least i citations. In each size variation of the dataset, the column “N” lists the number of authors each name label corresponds to. For example, the dataset that contains “J. Lee” of at least 2 citations has 100 different “J. Lee”, such as “Jaejin Lee”, “Jon Lee”, etc. The column “C” lists the total number of citations in the corresponding dataset.

each author. The results show that the increase of author citations generally improves the disambiguation performance. For example, the accuracy of disambiguating “J. Martin” increases from 82% to 96.8% when we increase the percentage of citations of each “J. Martin” from 10% to 100%. However, results on the “J. Robinson” dataset show the opposite trend. We observe that two “J. Robinson”’s with the largest number of citations both publish papers on the topic of “databases”. These two “J. Robinson”’s are always clustered together. It appears that the increase in citations in this case introduces errors and decreases the disambiguation accuracy. To resolve this, we probably need more features, such as author’s affiliations for successful disambiguation. Overall, the experiments on disambiguating “M. Jones”, “D. Johnson”, “M. Brown” and “M. Miller” achieved higher accuracies than on other names, such as the four popular Asian names, “J. Lee”, “S. Lee”, “C. Chen”, and “Y. Chen”. Table 3 shows the detailed results on each name dataset with the second type of dataset size variation.

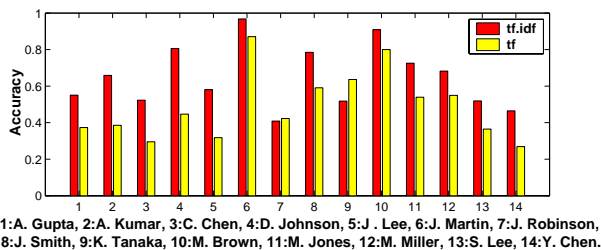


Figure 2: Name disambiguation accuracies (Y axis) of using two feature weighting schemes. X axis represents 14 names. For each name, the left bar represents the usual “TFIDF” and the right bar represents the “NTF”.

4.4.2 “TFIDF” v.s. “NTF”

In each size variation of the 14 DBLP name datasets, we compare two feature weighting schemes: “TFIDF” and the “NTF”. Experimental results show that “TFIDF” performs better than “NTF” in general. Figure 2 shows an example on the 14 DBLP name

datasets containing authors that have at least 10 citations. Experiments on other size variations of the datasets show similar results. “TFIDF” outperforms “NTF” because of the nature of the weighting schemes. “TFIDF” considers not only the frequency of a feature in a citation but also the distribution of a feature in all the citations of a name dataset. “NTF” considers only the feature frequency in one citation, which is limited by the fact that there seems to be very few words that are repeated in a single citation. As such, the “TFIDF” scheme better captures features specific to an author than “NTF” does. This indicates that a good feature weighting is important to the performance of name disambiguation. Improvements may be achieved using better feature weighting techniques such as Log Entropy [5].

4.4.3 Effect of amount of name information on disambiguation

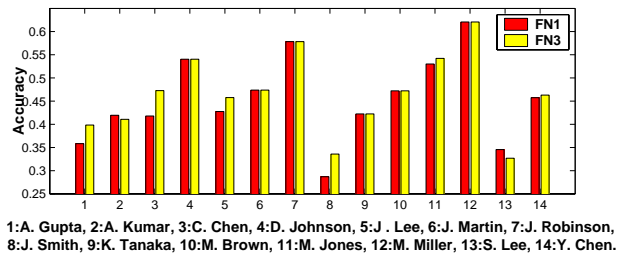


Figure 3: Name disambiguation accuracies (Y axis) using different amount of first name information. The X axis represents 14 names. For each name, the left bar (FN1) represents the result of using the first name initial; the right bar (FN3) represents the result of using first three characters of the first name.

Simplifying each name with the first name initial and last name introduces name ambiguity. For example, the names “Sung Jin Kim” and “Seon-Kyu Kim” are simplified to the same name label “S. Kim”. To investigate this effect, we did another set of experiments, representing the first name by its first three characters. We observe that most names from the DBLP database have complete

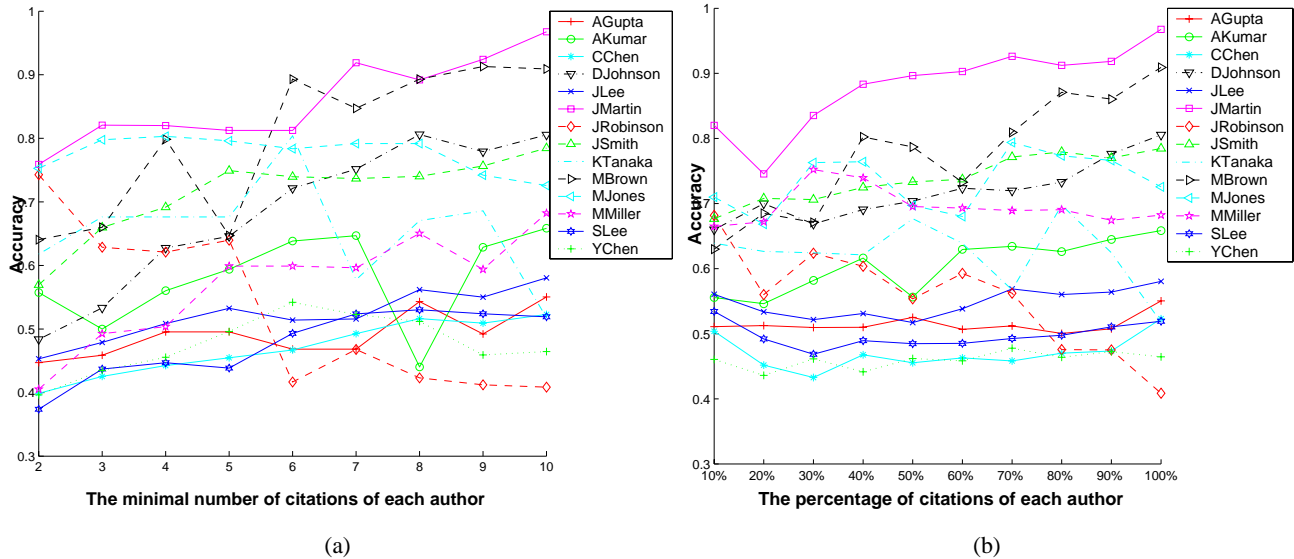


Figure 1: Name disambiguation accuracy change with the variation of the dataset size. X axis in Figure (a) and (b) shows the two different types of dataset size variations. Y axis represents name disambiguation accuracy with the “TFIDF” feature weighting. Lines of different colors and shapes represent different ambiguous names.

first name information, while web collected publication lists contain many names that are in the format of first name initial and last name. Such inconsistent name formats cause one author to be represented by two different features and introduces name ambiguity. Therefore, we only report experimental results on all citations collected from the DBLP database Bibliography in which we vary the representation of first names. The citation vectors are constructed with only co-author names, and we do not consider the cases when an author has no co-authors. Figure 3 shows the results on the datasets that contain all author citations. Representing the first name by its first three characters improves disambiguation accuracy for most names, e.g. “A. Gupta”, “C. Chen”, “J. Lee”, “J. Smith”, “M. Jones” and “Y. Chen”. We observe that many different co-authors in these datasets have the same name label in the simplified format of first name initial and last name, e.g. 18.7% different co-authors in the “C. Chen” dataset, 29.5% co-authors in the “J. Lee” dataset, and 12% co-authors in the “Y. Chen” dataset. Therefore, adding additional name information may decrease name ambiguity, and improve the disambiguation accuracy. However, we notice the classification accuracy drops on the specific name datasets “A. Kumar” and “S. Lee” when representing the first name by its first three characters. Two reasons may explain why. The first is that DBLP citations still have inconsistent name representations for the same author, for example, the two formats “W. Tsai” and “Wen Tsai” for the same author. Simplifying names in the first name initial and last name format, however, represents the above two names as the same. The second reason is name misspellings. For example, “Kohji Zettsu” is misspelled in some citations as “Koji Zettsu”. Representing the first name by its first three or more characters keeps such name misspelling, and incorrectly recognizes the above two name expressions as different. This indicates that combining techniques on duplicate string detection [9, 45, 11, 39] may improve the name disambiguation performance.

4.4.4 Coauthor name v.s. paper title v.s. publication venue title

Name	Coauthor 1	Coauthor 2	PTitle	Venue title
A. Gupta	37.9%	39.8%	47.7%	24.7%
A. Kumar	25.7%	34.0%	61.0%	45.2%
C. Chen	33.3%	37.3%	43.7%	23.7%
D. Johnson	31.9%	41.2%	53.4%	50.0%
J. Lee	38.8%	45.1%	38.1%	19.6%
J. Martin	37.9%	62.5%	50.0%	65.2%
J. Robinson	41.2%	53.0%	43.2%	37.2%
J. Smith	46.7%	58.4%	44.0%	24.7%
K. Tanaka	49.6%	54.5%	68.6%	46.5%
M. Brown	50.4%	57.4%	61.7%	36.5%
M. Jones	43.9%	61.8%	50.2%	33.5%
M. Miller	52.4%	53.7%	52.4%	53.0%
S. Lee	34.3%	36.1%	37.7%	30.4%
Y. Chen	37.3%	43.1%	31.2%	19.8%
Mean	40.1%	48.4%	48.8%	36.4%
Std	7.7%	10.0%	10.3%	13.9%

Table 4: Name disambiguation accuracies using co-author information alone, paper title words alone (PTitle) and publication venue title words (Venue title) alone. “Std” - Standard Deviation. Column “Coauthor 1” considers the names that do not have co-authors as being incorrectly disambiguated; “Coauthor 2” does not consider the cases where names have no co-authors.

We achieved different disambiguation accuracies using each citation attribute alone. Table 4 shows an example on the 14 DBLP name datasets containing authors who have at least 10 citations. Experiments on other size variations of the datasets show similar results. Because the names without co-authors can not be disambiguated by using co-author names alone, we evaluate the performance using two methods. The first method considers the names that do not have co-authors as being incorrectly disambiguated, as

Name	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
A. Gupta	51.1	51.3	51.0	51.0	52.5	50.7	51.2	50.1	50.8	53.9
A. Kumar	55.6	54.6	58.2	61.6	55.7	63.0	63.4	62.6	64.5	64.3
C. Chen	50.4	45.2	43.3	46.8	45.6	46.3	45.8	47.0	47.3	50.6
D. Johnson	66.0	70.0	66.9	69.1	70.4	72.4	72.0	73.3	77.6	79.1
J. Lee	56.1	53.4	52.2	53.1	51.7	53.8	56.9	56.0	56.4	56.2
J. Martin	82.0	74.5	73.6	88.3	89.7	90.3	92.6	91.1	91.8	96.8
J. Robinson	68.2	56.0	62.4	60.4	55.4	59.3	56.2	47.6	47.5	39.2
J. Smith	67.7	70.8	70.6	72.5	73.3	73.7	77.2	77.9	77.0	77.4
K. Tanaka	63.9	62.7	62.4	62.1	67.7	63.9	56.7	69.8	62.5	50.8
M. Brown	63.0	68.5	67.1	80.2	78.7	73.3	81.0	87.1	86.0	87.0
M. Jones	71.0	66.8	76.3	76.4	70.0	68.0	79.4	77.3	76.6	70.6
M. Miller	66.5	67.2	75.2	74.0	69.5	69.3	68.9	69.1	67.4	67.4
S. Lee	53.4	49.2	46.9	48.9	48.5	48.5	49.3	49.8	51.1	50.4
Y. Chen	46.1	43.6	46.1	44.2	46.2	45.9	47.8	46.4	47.3	45.5
Mean	61.5	59.6	60.9	63.5	62.5	62.7	64.2	64.7	64.6	63.5
Std	9.4	9.9	11.0	13.2	13.0	12.4	14.0	14.9	14.7	16.8

Table 3: The disambiguation accuracy (%) change with the dataset size variation of the 14 DBLP name datasets. $i\%$ means that $i\%$ citations of each author is randomly selected. “Std” - Standard deviation.

shown in column “Coauthor 1”. The second method, shown in column “Coauthor 2”, does not consider the cases when authors have no co-authors. “Coauthor 2” in Table 4 shows that using co-author information alone outperforms using only paper title words or publication venue title words in most name datasets. We hypothesized that the publication venue title information is more stable than paper title information, because an author may not reuse certain keywords for paper titles, and paper titles usually contain sparse information. Some paper titles, for example, “Where am I?”, give little information about the author’s research topic. Surprisingly, Table 4 shows that using publication venue title words alone generally performs worse than using paper title words alone. The possible reasons are the following. First, the publication venue title information is not always available in the datasets, or is parsed wrong. Second, “Ph.D. Dissertation”, as parsed as the publication venue title, does not reveal the author’s research area. Third, different publication venue titles may share the same abbreviation. For example, “IJCS” can refer to “International Journal of Comparative Sociology”, or “International Journal of Communication Systems”. Simply mapping the publication venue title abbreviation to the entry of a publication venue title full name database may introduce misleading information. It would be helpful if we could disambiguate publication venue title abbreviations by the context information such as the topic of the paper. Fourth, the full publication venue title information we obtain does not cover all the publication venue title abbreviations in the datasets. This may under-exploit the publication venue information. Fifth, since most authors from DBLP datasets are from the Computer Science community, different authors are very likely to have the same or similar research area and publish papers in the same place. In this case, the publication venue title information is not discriminative. According to the above different contributions made by different citation attributes, we can automatically tune different weights for different attributes for improvement, as shown in previous work [9, 45, 11].

4.4.5 Effect of author research area diversity on disambiguation

We observe that many authors from the DBLP datasets have close research areas. For example, over 25% of authors in each name dataset of all author citations publish papers about “networks”. For example, 36.1% (31 out of 86) “S. Lee” and 39.4% (28 out

of 71) “Y. Chen” publish papers about “networks”. “Databases” is another popular research topic. 24.0% (24 out of 100) of “J. Lee”, 29.0% (9 out of 31) “J. Smith”, and 33.3% (4 out of 12) of “J. Robinson” publish papers about “databases”. Correspondingly, many authors share words of the same word stem such as “network”, “database”, “comput” and “system”. Different authors also publish papers in the same publication venues. Such common words from publication venue titles can be considered as “ambiguous information”, and make accurate clustering hard. It is even more challenging to distinguish two authors of the same name label who co-author the same paper, as shown by the following example, “Chien-Chang Chen, Chaur-Chin Chen. Filtering methods for texture discrimination. Pattern Recognition Letters. 1999.”

We consider each author as a class, and plot the within-class and the cross-class similarity distributions using the histogram for each name dataset. The ideal case is that the within-class similarity is distributed around “1” and the cross-class similarity is distributed around “0”. Figure 4 shows that the difference between the within-class and cross-class similarity distribution for “C. Chen” is less than that of the “J. Martin” dataset. This explains why the disambiguation accuracy on “C. Chen” is worse than that on “J. Martin”. A possible solution for improvement can be a set of features that enlarge the differences between citations of different authors.

4.5 Name disambiguation on web collected publication lists

Given DBLP’s narrowness of topical coverage, name disambiguation on DBLP databases seems to be more challenging than it might be using other citation databases. The task of name disambiguation appears to be even more difficult when the nationalities that occur most frequently have a relatively small set of family names. To see results from other domains where the population of possible names may be larger, we have also conducted another two sets of experiments on the second type of citations, i.e. 11 “J. Smith” of 338 citations and 15 “J. Anderson” of 229 citations. We achieved 84.3% and 71.2% accuracy respectively on these two datasets using spectral clustering, better than the 75.4% and 67.0% accuracy we achieved using K-means algorithm, with “TFIDF” feature weighting schema. This also shows that higher disambiguation accuracy can be achieved using only the three citation attributes when ambiguous authors have more diverse research areas.

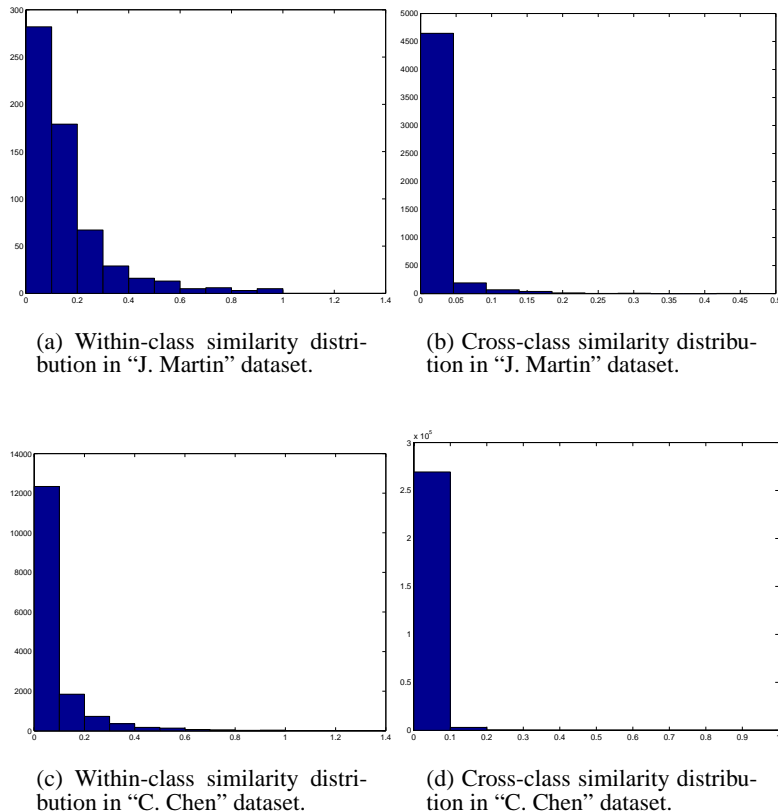


Figure 4: The histogram of within-class and cross-class similarity distribution in "J. Martin" and "C. Chen" datasets. X axis represents the similarity value. Y axis represents the number of citation pairs from the same class (within-class) or from different classes (cross-class) that have the corresponding similarity value.

5. CONCLUSION AND DISCUSSION

We investigate name disambiguation in author citations using a K -way spectral clustering method with QR decomposition for cluster assignment. We also study several factors that may affect the disambiguation performance, such as feature weight assignment, dataset size, the amount of name information, and the author research area diversity. We show that spectral methods outperform k -means for the data sets collected from publication lists.

We also show that as expected, the more features used (coauthor names, paper and publication venue title words) in author classification, the better the classification accuracy. We achieved 61.5% to 64.7% average accuracy on 14 DBLP name datasets with a variety of sizes. The highest accuracy 96.8% is achieved on the "J. Martin" dataset containing "J. Martin"s that have at least 10 citations. The disambiguation accuracies on "S. Lee", "J. Lee", "Y. Chen" and "C. Chen" are lower than on other names. The possible reason is that these datasets contain more ambiguous authors than other datasets, and many authors from these datasets have close research areas. Experiments on disambiguating 11 "J. Smith"s and 15 "J. Anderson"s (citations are publication lists collected mainly from authors' home pages) show 84.3% and 71.2% accuracy respectively.

Further improvements could be obtained through semantic word clustering on paper titles and publication venue titles [51]. We observe that a researcher usually has a research area or areas that do not change over a period of time, and his/her paper or submitted

publication venue titles are closely related to his/her research topic. However, the paper and publication venue title words are sparse, and an author may not reuse a certain group of title words. Moreover, our current work does not recognize the similarity between words such as "Neurocomputing" and "NeuroScience". Therefore, it is reasonable to cluster "similar" title words into research fields, and use a new set of features that summarize similar words. Such a word cluster reduces feature sparseness, and usually has more robust probability estimates by averaging statistics for similar words [6]. Existing word clustering methods we can apply include methods based on the Word Net [7], distributional word clustering [6, 38, 14, 16], bipartite word clustering [53], committee-based word clustering [33], and other word similarity measures [46, 29]. Research keywords classification schemes such as the ACM classification may also help to map different title words into a research category.

In our hand-labeling of the datasets, we used extra information such as affiliations, email addresses, resumes, home pages, and some human judgment. Therefore, in order to improve the name disambiguation performance, we most likely need more features as those that are used in our hand-labeling than the three citation attributes that we currently use. Words and bigrams from the paper abstracts may also provide useful information for disambiguation. We would also like to address the issue of automatically choosing the number of name clusters.

We wish to combine both unsupervised and supervised learning

methods, to build a practical reinforced name disambiguation system in the future. We would like to add string-based components [9, 45] for better representation of author names and for finding the canonical name of an author. We would also like to disambiguate similar corporate names appearing in academic and publishing world, such as “Loyola College.” It may also be useful to extend our name disambiguation systems in digital documents to other applications, especially in the academic, patent, medical records, or genealogy fields.

6. ACKNOWLEDGMENTS

We wish to thank Mark Stefik for his valuable comments, and Pradeep Teregowda for helping in the preparation of the dataset used in this work. We acknowledge partial support from NSF (Grants 0121679 and CCF-0305879) as well as Yahoo! (traveling expenses).

7. REFERENCES

- [1] Digital bibliography & library project.
<http://WWW.Informatik.Uni-Trier.DE/~ley/db/index.html>.
- [2] Getty’s ULAN (Union List of Artist’s Names).
http://www.getty.edu/research/conducting_research/vocabularies/ulan/.
- [3] The library of congress name authority file.
<http://www.loc.gov/marc/authority/index.html>.
- [4] Y. Azar, A. Fiat, A. R. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proceedings of 33rd ACM Symposium on Theory of Computing*, pages 619–626, 2001.
- [5] R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval. 1999.
- [6] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, 1998.
- [7] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, 2002.
- [8] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.
- [9] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, 2003.
- [10] D. Bitton and D. J. DeWitt. Duplicate record elimination in large data files. *ACM Transactions on Database Systems*, 8(2):255–265, 1983.
- [11] L. K. Branting. Name-matching algorithms for legal case-management systems. *Journal of Information, Law and Technology (JILT)*, 1, 2002.
- [12] M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction. In *Proceedings of the 16th National Conference on Artificial Intelligence*, pages 328–334, 1999.
- [13] W. W. Cohen, H. A. Kautz, and D. A. McAllester. Hardening soft information sources. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining*, pages 255–259, 2000.
- [14] I. Dagan, F. C. N. Pereira, and L. Lee. Similarity-based estimation of word cooccurrence probabilities. In *Meeting of the Association for Computational Linguistics*, pages 272–278, 1994.
- [15] L. Daniel and J. Slezak. Street talk: the word on address-matching. *Business Geographics*, pages 26–33, 1995.
- [16] I. Dhillon, S. Manella, and R. Kumar. A divisive information-theoretic feature clustering for text classification. *Journal of Machine Learning Research (JMLR)*, 3:1265–1287, 2003.
- [17] T. DiLauro, G. S. Choudhury, M. Patton, J. W. Warner, and E. W. Brown. Automated name authority control and enhanced searching in the levy collection. *D-Lib Magazine*, 7(4), 2001.
- [18] W. B. Dolan. Word sense ambiguity: Clustering related senses. Technical report, 1994.
- [19] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *SODA: ACM-SIAM Symposium on Discrete Algorithms*, pages 291–299, 1999.
- [20] D. G. Feitelson. On identifying name equivalences in digital libraries. *Information Research*, 9(4):192, 2004.
- [21] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.
- [22] W. B. Frakes and R. Baeza-Yates. *Information Retrieval, Data Structures & Algorithms*. Prentice-Hall International (UK) Limited, London, 1992.
- [23] C. L. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the 3rd ACM Conference on Digital Libraries*, pages 89–98, 1998.
- [24] P. Gillman. National name authority file: Report to the national council on archives. Technical Report British Library Research and Innovation Report 91, The British Library Board, 1998.
- [25] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries*, pages 37–48, 2003.
- [26] H. Han, C. L. Giles, H. Zha, C. Li, and K. Tsioutsouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital libraries*, 2004.
- [27] M. A. Hernandez and S. J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1):9–37, 1998.
- [28] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. In *Proceedings of the 41st Foundations of Computer Science*, pages 367–380, 2000.
- [29] J. Karlgren and M. Sahlgren. From words to understanding. In *Kanerva et al. (eds.) Foundations of Real World Intelligence. CSLI publications*, pages 294–308, 2001.
- [30] R. Krovetz and W. B. Croft. Word sense disambiguation using machine-readable dictionaries. In *Proceedings of the 12th Annual ACM SIGIR Conference*, pages 127–136, 1989.
- [31] M.-L. Lee, T. W. Ling, and W. L. Low. Intelliclean: a knowledge-based intelligent data cleaner. In *In 6th International Conference on Knowledge Discovery and Data Mining*, pages 290–294, 2000.
- [32] H. Li and N. Abe. Word clustering and disambiguation based on co-occurrence data. In *Proceedings of the 17th*

- International Conference on Computational Linguistics*, pages 749–755, 1998.
- [33] D. Lin and P. Pantel. Concept discovery from text. In *Proceedings of Conference on Computational Linguistics*, pages 577–583, 2002.
- [34] A. McCallum, K. Nigam, and L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Knowledge Discovery and Data Mining*, pages 169–178, 2000.
- [35] A. E. Monge and C. Elkan. An efficient domain-independent algorithm for detecting approximately duplicate database records. In *Research Issues on Data Mining and Knowledge Discovery*, pages 23–29, 1997.
- [36] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of Advances in Neural Information Processing Systems*, pages 849–856, 2001.
- [37] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *Proceedings of Neural Information Processing Systems: Natural and Synthetic*, number 15, 2002.
- [38] F. C. N. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.
- [39] A. Pirkola, J. Toivonen, H. Keskustalo, K. Visala, and K. Jarvelin. Fuzzy translation of cross-lingual spelling variants. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 345–352, 2003.
- [40] A. Pothen, H. D. Simon, and K.-P. Liou. Partitioning sparse matrices with eigenvectors of graphs. 11:430–452, 1990.
- [41] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden Markov model structure for information extraction. In *Proceedings of AAAI 99 Workshop on Machine Learning for Information Extraction*, 1999.
- [42] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [43] M. Skounakis, M. Craven, and S. Ray. Hierarchical hidden markov models for information extraction. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2003.
- [44] A. Takasu. Bibliographic attribute extraction from erroneous references based on a statistical model. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries*, pages 49–60, 2003.
- [45] S. Tejada, C. Knoblock, and S. Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 350–359, 2002.
- [46] E. L. Terra and C. L. A. Clarke. Frequency estimates for statistical word similarity measures. In M. Hearst and M. Ostendorf, editors, *HLT-NAACL 2003: Main Proceedings*, pages 244–251, 2003.
- [47] H. R. Turtle and W. B. Croft. Uncertainty in information retrieval systems. *Uncertainty Management in Information Systems*, pages 189–224, 1996.
- [48] J. W. Warner and E. W. Brown. Automated name authority control. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital libraries (JCDL01)*, 2001.
- [49] D. Widdows and B. Dorow. A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics*, pages 1093–1099, Taipei, Taiwan, August 2002.
- [50] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th ACM International Conference on Research and Development in Information Retrieval (SIGIR03)*, pages 267–273, 2003.
- [51] Y. Y. Yao, S. Wong, and L. S. Wang. A non-numeric approach to uncertain reasoning. *International Journal of General Systems*, 23(4):343–359, 1995.
- [52] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering. In *Neural Information Processing Systems (NIPS 2001)*, pages 1057–1064, 2001.
- [53] H. Zha, X. He, C. Ding, M. Gu, and H. Simon. Bipartite graph partitioning and data clustering. In *Proceedings of ACM CIKM 2001, the 10th International Conference on Information and Knowledge Management*, pages 25–32, 2001.