A Model-based K-means Algorithm for Name Disambiguation

Hui Han[1] **Hongyuan Zha**[1] **C. Lee Giles**[1, 2]

[1]Computer Science & Engineering, The Pennsylvania State University, University Park, PA 16803, USA

[2]School of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16803, USA

{hhan,zha}@cse.psu.edu giles@ist.psu.edu

Abstract

Unambiguous identities of resources are important aspect for semantic web. This paper addresses the personal identity issue in the context of bibliographies. Because of abbreviations or misspelling of names in publications or bibliographies, an author may have multiple names and multiple authors may share the same name. Such name ambiguity affects the performance of identity matching, document retrieval and database federation, and causes improper attribution of research credit. This paper describes a new K-means clustering algorithm based on an extensible Naïve Bayes probability model to disambiguate authors with the same first name initial and last name in the bibliographies and proposes a canonical name. The model captures three types of bibliographic information: coauthor names, the title of the paper and the title of the journal or proceeding. The algorithm achieves best accuracies of 70.1% and 73.6% on disambiguating 6 different "J Anderson" s and 9 different "J Smith" s based on the citations collected from researchers' publication web pages.

1. Introduction

Unambiguous identities of resources and URI (Uniform Resource Identifier) references are important to the construction of semantic web, global knowledge federation and scalable web services (Berners-Lee et al. 2001; Clark 2002; Pepper et al. 2003). Personal identity ambiguity is one of the semantic challenges. In research papers or bibliographies, we observe two types of name ambiguities due to name variation or name misspelling. 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 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).

Such name ambiguity problem can affect the performance of document retrieval and cause incorrect identification of and credit attribution for researchers. For example, the most cited document in CiteSeer (Giles, Bollacker and Lawrence 1998) database as of May 2003 (http://citeseer.nj.nec.com/access.html) has the authors "Pierluigi Crescenzi, Leandro Dardini, Roberto Grossi", but the CiteSeer document web page directs an author homepage to "Mark Jerrum". Another example: "D. Johnson" is the most cited author in computer science according to CiteSeer's statistics in May 2003 (http://citeseer.nj.nec.com/mostcited.html). 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 "Joel T. Johnson". Moreover, the Digital Bibliography & Library Project (DBLP), a large collection of computer science bibliographical records, is also found to list bibliographies from different name entities under the same name. For example, the "Yu Chen" from DBLP refers to at least three different people: Yu Chen from UCLA; Yu Chen from Microsoft at Beijing branch and Yu Chen as the senior professor from Renmin University of China. Therefore, successful name entity disambiguation may greatly help in locating the right researcher and obtaining his/her academic information from the correct homepage, and indexing bibliographic database more accurately and efficiently.

Name ambiguity is a special case of the general problem of identity uncertainty, where objects are not labeled with unique identifiers (Pasula et al. 2002). Such identity uncertainty problem is persuasive in the real-world data, especially in the scenario of heterogeneous data sources such as Internet, and has been approached independently in different research areas. For example, the data cleaning work (Bitton and DeWitt 1983; Hernandez and Stolfo 1998; Lee, Ling and Low 2000; Monge and Elkan 1997; Bilenko and Mooney 2003; Tejada, Knoblock and Minton 2002) has focused on the detection of duplicate records and the elimination and merge/purge problem in databases. In addition, record linkage matches records from two files or

two databases (Fellegi and Sunter 1969; Cohen, Kautz and McAllester 2000). Furthermore, data association assigns new observations to existing trajectories when tracking multiple objects (Bar-Shalom and Fortmann 1988). Lastly, the citation matching or co-referencing problem detects multiple citations that refer to the same publication (Giles, Bollacker and Lawrence 1998; Pasula et al. 2002; Marthi, Milch and Russell 2003; McCallum, Nigam and Ungar 2000).

Our work focuses on addressing the author identity uncertainty problem in the context of citations and proposes the idea of a *canonical* name, i.e. a name that is the minimal invariant and complete name entity for disambiguation. Such a name may have more than just the name of the individual as constituents.

This paper proposes a model-based K-means clustering algorithm (Hartigan and Wong 1979) for name disambiguation, with a probability model used to compute distance. That is, the probability a cluster of citations of an author produces a citation is the distance between the cluster and the citation. We choose naïve Baye to model the probabilities of citation attributes: coauthors, paper titles and journal/proceeding ("journal") titles for name disambiguation. The motivation is that a researcher usually has research areas that are stable over a period and tends to co-author papers with a particular group of researchers during this period. Such citation attributes contain rich information about the real identity of every author in a citation. The choice of Naïve Bayes model is motivated by its simplicity and extensibility for more identity attributes such as the researcher's affiliation. Such model also avoids the weights tuning for different attributes usually used in the similarity-based methods (Bilenko and Mooney 2003). We test the algorithm on disambiguating two sets of data -- 6 "J Anderson"s and 9 "J Smith"s and achieve highest accuracies of 70.1% and 73.6%.

The rest of the paper is organized as follows: Section 2 introduces the model-based K-means algorithm for name disambiguation; Section 3 reports experiment design and results; Section 4 concludes and discusses the future work.

2. Name disambiguation algorithm

Given a set of ambiguous author names such as "J Smith" and associated citations as shown in Table 1, the goal of our name disambiguation system is to cluster the citations of different name entities, and output identity information such as the research interests and his collaborators.

The methods for extracting citation attributes include regular expression matching, rule-based system (Califf et al. 1999), hidden Markov models (Seymore, McCallum, and Rosenfeld 1999; Skounakis, Craven, and Ray 2003; Takasu 2003) and Support Vector Machines (Han et al. 2003). To minimize the effect of inaccurate citation parsing on the study of the algorithm we propose, our experiments use regular expression matching and manual correction for citation parsing.

J Smith	Citations
1	*Rapid Profiling via Stratified Sampling, S. Sastry, R. Bodik, **J. E. Smith**, 28th Int. Symposium on Computer Architecture, pp. 278-289, June 2001.*
2	*Relationships in Influence Diagrams," Operations Research 41 (1993), 280-297. **Smith, James E.,** "Moment Methods for Decision Analysis", Management Science 39 (1993).*
3	*Henry E.J. and **Smith J.E.** 2002. The Effect of Surface-Active Solutes on Water Flow and Contaminant transport in Variably Saturated Porous Media with Capillary Fringe Effects. Journal of Contaminant Hydrology. Vol. 56 (3-4) p.247-270.*

Table 1. Three entities under the same ambiguous name "J Smith" and associated citations

2.1. The K-means algorithm

Step1. Initialization. Randomize and equally assign N citations (N is the total number of citations in the dataset) into K clusters. As the choice of K can be an independent research issue and to focus our study on the performance of the model-based clustering algorithm, we set K as the number of real name identities in the training dataset the ambiguous name corresponds to.

Step2. Consider each cluster as a "virtual entity" and estimate the prior probability of each cluster and the probability that a certain type of information is produced by the cluster, such as the probabilities that the virtual entity coauthors with a researcher, uses a certain keyword for the paper title, etc.

Step3. For each citation $C \in [1..N]$, estimate the probability that each cluster X_i ($i \in [1, K]$) would have generated the citation C. Then assign C to the clusters with the highest posterior probabilities of producing the citation C. There are two types of citation assignment. "Hard clustering" assigns C to the cluster with the highest posterior probability; "soft clustering" assigns C to multiple clusters. We use "hard clustering" for the final citation cluster assignment, and "soft clustering" during the algorithm iteration, which assigns C to two clusters with the highest posterior probabilities of producing C. And the second cluster has greater than 75% of the highest probability of producing C.

Step4. If the algorithm converges (when fewer than 1% of the citations change the cluster assignment), output each citation cluster and the associated top paper title keywords, journal title keywords and coauthors ranked by the probabilities that they are generated by the cluster; otherwise, go to Step 2 and continue the iteration.

Such unsupervised clustering algorithm expects the initial virtual entity with noisy information to become pure and refers to the real identity when the clustering converges. The algorithm bases on the naïve Bayes modeling, with the probability as the distance for the K-means algorithm.

2.2 The naive Bayes model

We assume that each author's citation data is generated by the naïve Bayes model, and use his/her past citations to estimate the model parameters. Specifically, we estimate the model parameters of each cluster (virtual entity) in Step 2 of the above K-means algorithm. Based on these parameter estimates, we use Bayes rule to calculate the probability that each cluster X_i ($i \in [1, K]$) would have generated each citation and reassign each citation into the clusters in Step 3. Name disambiguation is therefore achieved by clustering the citations of the same identity. For space limitation, we only give an overview of the model here and omit the model parameters and estimation.

Given an input test citation C with the implicit omission of the query author, the target function is to find a cluster X_i with maximal posterior probability to author the citation C, i.e.,

$$\max_i P(X_i|C) \quad (1)$$

Using Bayes rule, the problem becomes finding

$$\max_i P(C|X_i)P(X_i)/P(C) \quad (2)$$

where $P(X_i)$ denotes the prior probability of X_i authoring papers, and is estimated as the proportion of the papers of X_i among all the citations. $P(C)$ denotes the probability of the citation C and is omitted since it does not depend on X_i. Then function (2) becomes

$$\max_i P(C|X_i)P(X_i) \quad (3)$$

We assume coauthors, paper titles, and journal titles are independent citation attributes, and different elements in an attribute type (such as different coauthors, keywords) are also independent from each other. Therefore, we decompose $P(C|X_i)$ in function (3) as

$$P(C|X_i) = \Pi_j P(A_j|X_i) = \Pi_j \Pi_k P(A_{jk}|X_i) \quad (4)$$

where A_j denotes different type of attribute; that is, A_1 - coauthor names; A_2 – paper title; A_3 - journal title. Each attribute is decomposed into independent elements represented by $A_{jk(j)}$ ($k \in [0 .. K(j)]$). $K(j)$ is the total number of elements in attribute A_j. For example, $A_1 = (A_{11}, A_{12}, \ldots, A_{1k}, \ldots, A_{1K(1)})$, where A_{1k} indicates the k_{th} coauthor in C.

To avoid underflow, we store log probabilities in implementation, and the target function becomes:

$$\max_i P(X_i|C) = \max_i [\Sigma_j \Sigma_k \log(P(A_{jk}) + \log(P(X_i))] \quad (5)$$

where $j \in [1, 3]$ and $k \in [0, K(j))$. The above attribute independence assumption may not hold for real-world data, since there exists cases such as multiple coauthors always appear together. However, empirical evidence shows that naïve Bayes often performs well in spite of such violation. Friedman, Domingos and Pazzani show that the violation of the word independence assumption sometimes may affect slightly the classification accuracy (Friedman 1997; Domingos and Pazzani 1996).

2.3 Semantic clustering on keywords

The above model takes each word of the paper title or the journal title as an independent element, and estimates its corresponding author-specific probability. The model captures the information such as the research field, keywords in the research direction, and the preference of title word usage of an author X_i.

However, the paper and journal title words are sparse, and an author may not reuse a certain group of words with high probabilities. Therefore, it is reasonable to cluster the semantically similar words and model the probability that an author uses the similar words for his/her paper title. Once the "similar" words are clustered, the cluster label will represent the words in that cluster. The probability estimation on the old and new words will become that on the similar and dissimilar words in such keyword clustering case. Similar word clustering can also be applied to journal titles. Clustering similar words can use the existing word clustering methods, such as the methods based on WordNet (Banerjee and Pedersen 2002), distributional word clustering (Baker and McCallum 1998; Pereira, Tishby and Lee 1993; Dhillon, Manella and Kumar 2002], bipartite word clustering (Zha et al. 2001), etc. Our experiments use the CBC (Clustering By Committee) clustering algorithm by Pantel and Lin (2002) motivated by its good performance on sparse feature space.

3. Experiments

3.1 Experiments design

We conduct two experiments to disambiguate "J Anderson"s and "J Smith"s. Both names correspond to multiple name entities in the databases of our EbizSearch system (Petinot et al. 2003). We query "google" using name information such as "J Anderson", or "James Anderson", and the keyword "pubications". We manually check the returned links from google, and collect their publication web pages to construct our datasets. Table 2 shows the dataset for 9 "J Smith"s. For space limitation, we put the dataset for 6 "J Anderson"s at http://www.personal.psu.edu/users/h/x/hxh190/projects/name_project.htm.

We conduct linguistic preprocessing on the citation dataset. All the author names in citations are simplified as "First name initial + Last name", e.g. "Robert L. Winkler" is simplified as "R Winkler". We stem the words of paper

J Smith	Identity information		Publication website	Size
1	James E. Smith	U. of Wisconsin (Computer Sciences)	http://www.engr.wisc.edu/ece/faculty/smith_james.html	30
2	James E. Smith	Stanford University (Business)	http://www.fuqua.duke.edu/faculty/alpha/jes9.htm	14
3	James Smith	Unilever Cambridge Centre (Chemistry)	http://www.cus.cam.ac.uk/~js252/publications.html	14
4	James E. Smith	McMaster University (Geology)	http://www.science.mcmaster.ca/geo/faculty/smith/publications.html	29
5	James W. Smith	University of Washington, (Business)	http://faculty.washington.edu/jws4/publications.htm	22
6	John R. Smith	Columbia University (Electrical Engineering)	http://www.ctr.columbia.edu/~jrsmith/html/publications.htm	31
7	John Lindsay Smith	University of York (Chemistry)	http://www.york.ac.uk/depts/chem/staff/jrlspub.html	27
8	Jonathan A Smith	University of London (Psychology)	http://www.psyc.bbk.ac.uk/people/academic/smith_j/	35
9	Judith E. Smith	University of Leeds (Biology)	http://www.fbs.leeds.ac.uk/?publications=JES	25

Table 2. Citation dataset of 9 "J Smith"s. "Identity information" contains the full name of each "J Smith", his affiliation or research area; "Publication website" is the website of his publication list; "Size" is the number of citations.

title and journal names using Krovetz stemmer (Krovetz 1993), and remove the stop words such as "a", "the", etc. We also replace the conference or journal name abbreviations by their full names for more information. The full names of the conference or journal names are obtained from DBLP websites (http://www.informatik.uni-trier.de/~ley/db/conf/indexa.html and http://www.informatik.uni-trier.de/~ley/db/journals/index. html).

x	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	26		2			2			
2		2		1	3		2	6	
3	1		1	1	7	1	1	2	
4			1	28					
5		10			11	1			
6						31			
7			1			2	24		
8		1		5	7			22	
9	1	2	3				9	1	9

Table 3. Confusion matrix M used in an experiment of disambiguating 9 "J Smith" s. The number x in the row or column headers means "J Smith x". The empty cells represent 0, and the non-empty cell M[i, j] (i Є [1, 9] and j Є [1, 9]) is the number of "J Smith i" predicted as "J Smith j" .

We evaluate the name disambiguation performance based on confusion matrix as shown by Table 3, and define the clustering accuracy as the sum of the diagonal elements divided by the sum of all the elements in the matrix.

3.2 Experiments on two datasets

We apply the model-based K-means algorithm on each dataset before and after using CBC word clustering algorithm. For each case, we run 10 times experiments. Table 4 &5 show the results on two datasets.

Table 4 shows that the K-means algorithm achieves the average accuracy of 52.5% on disambiguting "J Anderson"s starting with the initial cluster score of 26.5%. Clustering semantically similar words using CBC algorithm further improves the name disambiguation accuracy to 60.0%. Table 4&5 also show the best performance on disambiguting 6 "J Anderson"s (70.6% accuracy) and 9 "J Smith"s (73.6% accuracy).

Table 6 shows an example of 3 clusters corresponding to the confusion matrix in Table 3, which gives an overview of the identity of a potentially real author from the cluster-associated high probability keywords and coauthors.

	Best results		Average results of 10 experiments	
	Original citations	CBC word clustering	Original citations	CBC word clustering
Initial score	28.4	25.5	26.5	25.7
At convergence	64.7	**70.6**	52.5	60.0

Table 4. The performance(%) on disambiguting 6 "J Anderson"s .

	Best results		Average results of 10 experiments	
	Original citations	**CBC word clustering**	**Original citations**	**CBC word clustering**
Initial score	19.4	19.4	18.9	19.1
At convergence	57.7	**73.6**	48.1	64.2

Table 5. The performance(%) on disambiguting 9 "J Smith"s.

	Keywords	**Journal words**	**Coauthors**
Cluster 1	Predict surface process effective model trace superscalar thread stratify enabling relational machine rapid garbage hardware assist concurrent parallelism collection profile virtual tn low program design instruction complexity protein bandwidth	survival conference international memory workshop int high fordham microarchitecture language practice compilation compute india annual design performance Barcelona proceedings technique parallel	E. Rotenberg Q. Jacobson T. Heil S. Sastry Y. Sazeides A. Dhodapkar
Cluster 4	Media contaminant spacing hysteresi dnapl mobile interfaci effective predict horizontal process porous saturate effects error surface experiment penetration active fingering fractal velocity induce soluble unsaturated finger model variable fringe capillary immobile visualize concept tension determine	Francis journal, taylor theory event cgu millennium meeting geophysic scientific ugc union quebec wetland Canada eos transactions soil hydrology psychology britain health resources contaminant	E. J. Henry A. W. Warrick P. Flowers P. Sheeran N. Beail A. S. Crowe
Cluster 7	aliphatic influence silica tert solution radical das butoxy tetra eatalyse size iodosylbenzene amino hinder aromatic cumyloxyl covalent phenyl ethylbenzene marked autoxidation dioxygen cycloalkene exchange axi porphyrin steric butyl photocleavage methylpyridyl acid phenolate carboxyl pentafluorophenyl ring iron	chem mol kinet tran soc perkin cat dalton faraday mol internat event analysis lewis dean eds drug diversity operation practical survival	B. Gilbert A. Dunn P. Taylor R. Terry J. Oakes G. Hodges

Table 6. Example of three clusters formed by the K-means algorithm.

4. Conclusions and discussions

This paper describes a K-means algorithm with a simple extensible naïve Bayes model to disambiguate names from citations. The algorithm clusters the citations based on three types of bibliographic information: coauthor names, the paper title words and the journal title words. High probability keywords, journal words and coauthor names give an overview of the potential identity of the cluster. The preliminary experiments achieve 70.1% best accuracy on disambiguating 6 different "J Anderson" s and 73.6% best accuracy on disambiguating 9 different "J Smith" s. Clustering semantically similar keywords using CBC (Clustering by Committee) algorithm shows the promise of improving the name disambiguation performance.

Within the probability framework, we believe further improvements can be obtained, e.g., the model can be extended for more attributes such as researcher's affiliation. It is also worthwhile to compare our model-based K-means algorithm (where probability defines distance) with the similarity-based K-means algorithm (where similarity defines distance) for a more objective evaluation on the name disambiguation algorithm. We would also work further on the choice of K, initial cluster assignment based on name information (e.g., the citations associated with "J. E. Anderson" and "J. H. Anderson" will be in two different clusters), and experiments on larger datasets.

Clustering documents based on domains (or sub domains) may also help eliminate author ambiguity, as authors in different domains could be identified as different. The choice of domains would be important to the clustering.

We also see extensions to many types of name disambiguation in digital documents, i.e. potential applications in homepage disambiguation. To check whether two homepages H_1 and H_2 with ambiguous owner names (and publication lists in citation format) really belong to the same author, we can use the cumulative probability of all citations in the publication list as the probability of the corresponding homepage, or we can regard all citations in a homepage as a meta-citation. Then we use the citations in H_1 to train a model of Author 1, and compute the probabilities of Author 1 authoring the citations of H_2, and vice versa. If both the two probabilities are large, then H_1 and H_2 are for the same author.

5. Acknowledgement

We would like to acknowledge partial support from NSF Grant 0121679 and helpful discussions with Cheng Li.

6. References

Banerjee, S., and Pedersen, T. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet. *Proc. the 3rd Intl. Conf. on Intelligent Text Processing and Computational Linguistic.*

Baker, L. D., and McCallum, A. K. 1998. Distributional Clustering of Words for Text Classification. *Proc. 21st ACM Intl. Conf. on Research and Development in Information Retrieval.*

Berners-Lee, T. Hendler, J. and Lassila, O. 2001. The Semantic Web. *Scientific America.*

Bilenko, M., and Mooney, R. J. 2003. Adaptive Duplicate Detection using Learnable String Similarity Measures. *Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.*

Bitton, D., and DeWitt, D. J. 1983. Duplicate Record Elimination in Large Data Files. *ACM Transactions on Database Systems* 8(2):255—265.

Califf, M. E., and Mooney, R. J. 1999. Relational Learning of Pattern-Match Rules for Information Extraction. *Proc. 16th National Conf. on Artificial Intelligence (AAAI-99).*

Clark, K. G. 2002. *Identity Crisis.* XML.com.

Cohen, W. W., Kautz, H. A., and McAllester, D. A. 2000. Hardening Soft Information Sources. *Proc. 6th Intl. Conf. on Knowledge Discovery and Data Mining* 255-259.

Dhillon, I., Manella, S., and Kumar, R. 2002. A Divisive Information Theoretic Feature Clustering for Text Classification. *J. of Machine Learning Research* 3 1265-1287.

Domingos, P. and Pazzani, M. 1996. Beyond Independence : Conditions for the Optimality of the Simple Bayesian Classifier. *Proc. 13th Intl. Conf. on Machine Learning.*

Fellegi, P. and Sunter, A. B. 1969. A Theory for Record-linkage. *J. of the American Statistical Association* 64:1183-1210.

Friedman, J. H. 1997. On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality. J. of *Data Mining and Knowledge Discovery* 1(1): 55-77.

Giles, C. L., Bollacker, K., and Lawrence, S. 1998. CiteSeer: An Automatic Citation Indexing System. *Digital Libraries (DL'98).*

Hartigan, J. A. and Wong, M. A. 1979. A K-means Clustering Algorithm. *Applied Statistics.* 28:100-108.

Han, H., Giles, C. L., Manavoglu, E., Zha, H, Zhang, Z., and Fox, E. 2003. Automatic Document Metadata Extraction Using Support Vector Machines. *Proc. ACM/IEEE Joint Conf. on Digital Libraries.*

Hernandez, M. A. and Stolfo, S. J. 1998. Real-world data is dirty: Data Cleansing and the Merge/Purge problem. *J. of Data Mining and Knowledge Discovery* 2(1):9-37.

Krovetz, R. 1993. Viewing Morphology as an Inference Process. *Proc. 16th ACM Intl. Conf. on Research and Development in Information Retrieval (SIGIR).*

Lee, M. L., Ling, W., and Low, W. L. 2000. Intelliclean: A knowledge-based Intelligent Data Cleaner. *Proc. 6th Intl. Conf. on Knowledge Discovery and Data Mining* 290-294.

Marthi, B., Milch, B., and Russell, S. 2003. First-Order Probabilistic Models for Information Extraction. *IJCAI 2003 Workshop on Learning Statistical Models from Relational Data.*

McCallum, A., Nigam, K., and Ungar, L. 2000. Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. *Proc. 6th ACM SIGKDD international conference on Knowledge discovery and data mining* 169-178.

Monge, A. E. and Elkan, C. 1997. An Efficient Domain-independent Algorithm for Detecting Approximately Duplicate Database Records. *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.*

Pasula, H., Marthi, B., Milch, B., Russell, S. and Shpitser, I. 2002. Identity Uncertainty and Citation Matching. *Advances in Neural Information Processing Systems 15.*

Petinot, Y., Teregowda, P. B., Han, H., Giles, C. L., Lawrence, S., Rangaswamy, A., and Pal, N. 2003. eBizSearch: an OAI-Compliant Digital Library for eBusiness. *Proc. of ACM/IEEE Joint Conf. on Digital Libraries.*

Pepper, S. and Schwab, S. 2003. Curing the Web's Identity Crisis: Subject Indicators for RDF. *Ontopia Technical Report.*

Pantel, P. and Lin, D. 2002. Document Clustering with Committees. *Proc. 25th ACM Intl. Conf. on Research and Development in Information Retrieval (SIGIR).*

Pereira, F., Tishby, N., and Lee, L. 1993. Distributional Clustering of English Words. *30th Annual Meeting of the Association for Computational Linguistics.*

Bar-Shalom, Y. and Fortmann, T. E. 1988. Tracking and Data Association. *Academic Press.* 1988.

Seymore, K., McCallum, A., and Rosenfeld, R. 1999. Learning Hidden Markov Model Structure for Information Extraction. *Proc. of AAAI-99 Workshop on Machine Learning for Information Extraction.*

Skounakis, M., Craven, M., and Ray, S. 2003. Hierarchical Hidden Markov Models for Information Extraction. *Proc. 18th Intl Joint Conf on Artificial Intelligence.*

Takasu, A. 2003. Bibliographic Attribute Extraction from Erroneous References Based on a Statistical Model. *Proc. of ACM/IEEE Joint Conf. on Digital Libraries.*

Tejada, S., Knoblock, C. A. and Minton, S. 2002. Learning Domain-independent String Transformation Weights for High Accuracy Object Identification. *Proc. 8th ACM Intl Conf. on Knowledge Discovery and Data Mining.*

Zha, H., He, X., Ding, C., Gu, M., and Simon, H. 2001. Bipartite Graph Partitioning and Data Clustering. *Proc. 10th Intl. Conf. on Information and Knowledge Management.*