Chapter 6

Author Name Disambiguation

Neil R. Smalheiser

Vetle I. Torvik

University of Illinois at Chicago

Introduction

For any work of literature, a fundamental issue is to identify the individual(s) who wrote it and, conversely, to identify all of the works that belong to a given individual. Attribution would seem to be a simple process (putting aside those works that are published anonymously) and yet it represents a major, unsolved problem for information science. Consequently, it is necessary to analyze the metadata, and sometimes the text, of a work of literature in order to make an educated guess as to the identity of its authors.

Author name disambiguation comprises four distinct challenges. First, a single individual may publish under multiple names—this includes (a) orthographic and spelling variants, (b) spelling errors, (c) name changes over time as may occur with marriage, religious conversion, or gender reassignment, and (d) the use of pen names. Second, many different individuals have the same name—in fact, common names may comprise several thousand individuals. Third, the necessary metadata are often incomplete or lacking entirely—for example, some publishers and bibliographic databases did not record authors' first names, their geographical locations, or identifying information such as their degrees or their positions. Fourth, an increasing percentage of scholarly articles is not only multi-authored, but also represents multi-disciplinary and multi-

institutional efforts. In such cases, disambiguating some of the authors does not necessarily help assign the remaining authors.

If the importance of assigning authorship was under-appreciated in the past, the situation is much different today. Numerous editorials and workshops have called for more effective disambiguation methods (e.g., CrossRef Author ID meeting [www.crossref.org/CrossTech/2007/02/crossref_author_id_meeting.html]; Workshop on Scholarly Databases & Data Integration [www.scimaps.org/meeting_060830.php]; Andrade, 2006; Bohne-Lang & Lang, 2005). Individual publishers and aggregators have set up their own internal disambiguation efforts on a massive scale (e.g., Thomson Scientific [with Web of Science] and Elsevier [with Scopus]). Information science researchers have proposed numerous models for author name disambiguation within bibliographic databases and on the Web, as will be discussed in this chapter.

This flurry of activity indicates the importance of author name disambiguation as a central issue in mining the literature. Whereas libraries once might have identified authors manually, this strategy fails with the emergence of massive (and hyperlinked) digital libraries. The rise of the Internet, with its use of increasingly sophisticated search engines, calls for conducting searches not simply on *keywords* but centered on *individuals*. National intelligence and law enforcement efforts have emphasized the importance of identifying and tracking individuals. Author disambiguation is needed to assist with everyday scientific tasks of numerous kinds. For example, investigators who are searching for potential collaborators in different disciplines seek authors (and not just their papers) because individuals are a great source of unpublished information—ideas, raw data, failed experiments, personal research notes, and hypotheses that were never followed up. As well, funding agencies seeking to choose

referees would benefit from identifying all of their co-authors (in order to spot potential conflicts of interest in advance). Journal editors could assign papers for review more readily by knowing the characteristic publication profile of its reviewers and conference organizers would benefit from knowing the publication profile of prospective invitees.

Similarly, policy makers would like to follow trainees and grantees over their subsequent careers. Currently, for example, the National Research Council (NRC) assesses doctoral program in the U.S. every ten years using a variety of indicators including the number of publications and citations per faculty member (www7.nationalacademies.org/resdoc/index.html). The NRC uses a faculty questionnaire (www7.nationalacademies.org/resdoc/Faculty__questionnaire_pdf.pdf) that asks for name variants and ZIP codes to help identify publications. Citation information is retrieved from Thomson Scientific.

Disambiguation is needed in order to create links from databases or digital libraries to online resources, such as full-text papers and the authors' home pages (if present). This entails knowing the individuals, not merely the names, on a paper. Thus, the rise of large bibliographic databases has invited data-mining analyses to understand large-scale features of the data as a whole and to extract, re-assemble, and synthesize the raw information to create entirely new knowledge. Author name disambiguation is a fundamental step in mapping knowledge domains (Shiffrin & Börner, 2004) and in other bibliometric and scientometric analyses. It will even be useful to marketers who wish to direct their advertisements to specific individuals. Finally, as we will discuss in some detail, knowing individuals (not merely author names) is crucial for establishing new resources such as citation networks, collaboration networks, and author profiles.

Creating Disambiguation Datasets

Why Not Establish a Registry of Unique Author Identifiers?

Before surveying current research approaches, we might ask: Why not simply set up a registry of author names with unique IDs? A registry would, in principle, solve the problem of author name disambiguation prospectively and, if each author submitted a list of his or her pre-existing publications when joining the system, it would allow one to assign many articles and books retrospectively as well. Technically, it is no more difficult to implement such a registry than to maintain any other Web-based service that relies upon author registration; Dervos, Samaras, Evangelidis, Hyvärinen, and Asmanidis (2006) give a good overview of the security, authentication, and programming issues that are involved in this endeavor. In their scheme, which so far has been implemented as a pilot project, UAI_Sys authors would enter their own metadata (which can be public or not), would choose their own passwords and safety questions, and would be assigned a permanent 16-digit ID number. Authors would then be responsible for using this number in all of their publications (for the rest of their lives); it is assumed that authors will agree to remember their passwords and will update their metadata periodically (this is not mandatory but without it the reliability and value of the metadata will degrade rapidly).

What are the drawbacks? Although the scheme has conceptual simplicity and is technically feasible, it fails to take into account the realities of human behavior. Authors are expected not only to cooperate voluntarily and actively, but also to enter their own data accurately and periodically over a time span of approximately 50 years. For this to work, the vast majority of authors need to participate—even one who was, for example, the seventh-listed author on a single article written while a student technician on a project. Almost half (46 percent) of the individuals in Medline have authored only a single article (Torvik & Smalheiser, 2008);

these are both the most difficult to disambiguate and the least likely to participate. Is it likely that the registry will garner universal support from authors who do not receive any short-term or tangible rewards for participating? In the biomedical arena, relatively few investigators post reprints of their articles, either on their own Web sites (Harnad, 2001), in institutional repositories, or in PubMed Central (Roberts, 2001)—even though these actions would definitely enhance their readership, visibility, and impact (Eysenbach, 2006). We have not even been able to convince our own colleagues to add their middle initials or suffixes to their names when publishing papers, even though this would take almost no time or effort and would assist in disambiguation (Garfield, 1969). Finally, unique person IDs are politically unpopular and strongly resisted in the United States, in spite of their advantages for census, health care, and national security. Authors may have concerns about a global database that keeps track of their affiliations, addresses, and publications throughout their lives.

The registry scheme also fails to take into account the tenuous nature of Web-based resources and their funding (Databases in peril, 2005; Merali & Giles, 2005). Which organization would ensure permanent funding for an author database, along with the staff needed to undertake the assigning of author IDs? In addition to the overt cost of maintaining the database, there is the extra cost of handling cases where a single individual requests more than one ID (e.g., lost or forgot the first ID number) and detecting errors (e.g., the author or publisher records the wrong ID number). Furthermore, the database should ideally cover all countries, all languages, all disciplines, and all forms of publication (journals, magazines, books, and grey literature). It is possible, even probable, that smaller disciplinary communities and publishing groups will establish and maintain their own author IDs so that a single individual may be connected to not one, but several, different ID numbers. However, the more IDs, the more work for authors; and

the less value they possess for disambiguation.

Why Not Carry Out Manual Disambiguation?

Most cases to date in which author names have been disambiguated have tended to involve manual curation. Librarians have traditionally carried out authority control on book collections (Maxwell, 2002), for example through: Library of Congress Authorities (authorities.loc.gov), Virtual International Authority File (viaf.org) (Tillett, 2002), and OCLC Research FictionFinder (fictionfinder.oclc.org). *Mathematical Reviews* has disambiguated over 2 million publications manually since 1940 (www.ams.org/mr-database/mr-authors.html) and faculty publication lists have been created in this manner (Scoville, Johnson, & McConnell, 2003). The National Library of Medicine maintains a Web site (www.LocatorPlus.gov) with almost a half million authority records for personal names (mostly book authors). The Union List of Artist Names (ULAN; www.getty.edu/research/conducting_research/vocabularies/ulan/about.html), developed by the Getty Research Institute, is an artist-centered database containing about 120,000 artists uniquely identified by assigned ID numbers; it is regularly updated by contributions from various institutions and projects. The database encodes name variants (including pseudonyms, names in different languages, variant spellings, and changes over time) and other information about artists.

Several initiatives use a combination of automatic and author-supplied or community-supplied input. For example, DBLife (DeRose, Shen, Chen, Lee, Burdick, Doan, et al., 2007) extracts author information from within a defined database research community and displays it in a standardized format that is subject to manual correction. The FOAF (Friend of a Friend) initiative is a community driven effort to define a Resource Discovery Format (RDF) vocabulary

for expressing metadata about people, their interests, relationships, and activities (www.foaf-project.org). Several Web-based services allow users to register and create profiles that are linked to their papers. For example, Community of Science (www.cos.com) has almost 480,000 profiles and Research Papers in Economics (RePEc; repec.org) has about 15,000 registered authors.

Nevertheless, manual disambiguation is a surprisingly hard and uncertain process, even on a small scale, and is entirely infeasible for common names. For example, in a recent study we chose 100 names of Medline authors at random and randomly selected a pair of articles for each name; these pairs were disambiguated manually, using additional information as necessary and available (e.g., author or institutional home pages, the full text of the articles, Community of Science profiles, Google searches). Two different raters did the task separately. In over a third of the cases it was not possible to be sure whether the two papers were written by the same individual. In a few cases, one rater said that the two papers were "definitely by different people" and the other said they were "definitely by the same person."

One might think that at the very least authors themselves would be able to point out their own publications from among a list of papers bearing their names but we have found that they, too, can be grossly unreliable. For example, the Community of Science Profile asks each person setting up a profile to approve publications from a list (of all papers bearing the same name) with a series of manual clicks. When there are more than 300 papers on the list, people tend to accept all of them rather than to examine them carefully. Conversely, based on our personal experiences working with scientists, we have found that they sometimes forget about articles that they have authored and even vigorously deny authorship until shown otherwise, particularly when they were not the senior or corresponding author or when the article was not peer-reviewed.

Research Approaches to Author Name Disambiguation

Author name disambiguation involves some of the same issues as other kinds of entity recognition and resolution. For example, many active research efforts are devoted to recognizing named entities within texts and on Web pages (Cohen & Hersh, 2005; Mann & Yarowsky, 2003; Vu, Masada, Takasu, & Adachi, 2007), disambiguating word sense (Schuemie, Kors, & Mons, 2005), and identifying co-reference mentions. Record linkage is a related problem insofar as it involves deciding whether two different entries (in the same or different databases) refer to the same person. Nevertheless, author name disambiguation is potentially a much richer enterprise than these other tasks because it goes beyond particular mentions or particular articles to provide an overall characterization of an individual. Whereas named entity recognition may attempt to identify which George W. Bush is being mentioned in a particular article, author disambiguation incorporates information across all of an individual's works and also includes features that involve extensive computation and outside knowledge from external sources. Thus, we envisage the process of author name disambiguation going considerably beyond component tasks such as classification or clustering, to provide an in-depth analysis or profile of the person.

At the abstract level, most research approaches to author name disambiguation share the broad outlines of predictive machine learning (Mitchell, 1997), which is designed to cluster or classify a body of works of literature corresponding to the individuals who wrote them. Machine learning generally requires: (1) acquiring training sets that provide positive and negative examples, (2) extracting one or more features from the works or their metadata, (3) employing a decision procedure of optimization or learning that acts upon the features, and (4) evaluating system performance. However, existing systems vary greatly in how these steps are formulated

and carried out. At least ten different approaches have been described in the past few years; these will be reviewed, compared, and contrasted in this and the following section. At the outset, it is important to keep in mind that different systems should not be compared on the basis of performance parameters (e.g., recall and precision) alone because each system was developed for a different type of disambiguation task and dataset. However, each of the methods could potentially be applied to any of the major bibliographic databases such as DBLP, CiteSeer, arXiv, Medline, Google Scholar, Web of Science, Scopus, Astrophysics Data System (ADS), Libra (Academic Search), or RePEc.

Most disambiguation approaches fall into one of the two machine learning paradigms: supervised or unsupervised. Supervised approaches take as input a set of training examples consisting of pairs of articles that are labeled as either positive (author match) or negative (not author match); unsupervised approaches do not use labeled training examples. In general, supervised approaches perform better because they are tuned specifically to the data (e.g., to determine the relative importance and interactive effects of different features such as a co-author vs. journal name vs. title word vs. affiliation).

Having sufficient training data is critical to the performance of any predictive model that will be extrapolated to new heretofore-unseen examples. The amount of data sufficient for training depends on the complexity of the model. Generating training sets does not have to be a manual, tedious, and error-prone process; in fact, it can be done in an automatic fashion (e.g., Torvik, Weeber, Swanson, & Smalheiser, 2005; Yin, Han, & Yu, 2007). For instance, one can generate positive examples by grouping papers that share personal e-mail addresses. Also, a name suffix (e.g., Jr. or 3rd) is a highly distinguishing feature of an individual, particularly when the first and last names are also unusual. Importantly, training sets should represent the entire

database and not exhibit bias toward certain values of the predictive features (e.g., using personal e-mail addresses will bias the dataset toward newer papers and using suffixes will give a bias toward English names). Thus, the bias needs to be measured, and accounted for, if significant correlations with predictive features are detected. Training sets can also be generated using a hybrid of manual and automatic methods as in the paradigm of active learning (e.g., Kanani, McCallum, & Pal, 2007; Torvik & Triantaphyllou, 2002), a strategy in which the learning algorithm iteratively detects the most informative examples for manual curation and the disambiguation model is updated after each iteration.

Most groups have tackled disambiguation by some sort of blocking mechanism (e.g., Bilenko, Kamath, & Mooney, 2006; On, Lee, Kang, & Mitra, 2005), such as considering only pairs of names that match on last name and first name initial. Blocking dramatically reduces the computational cost (for Medline this reduces the number of pairwise comparisons by a factor of about 100,000).

The features that are available for prediction vary across datasets (e.g., Medline does not have citation information, except for a small percentage of papers that are cross-listed in PubMed Central). Feature selection is perhaps the most important aspect of designing a model for disambiguation because it determines the upper limit of accuracy. A good principle is to use as many of the available and useful features as possible because one or just a few features are likely to limit the recall of the approach. Each feature can be encoded several different ways: For example, the similarity of two affiliation strings can be based on the number of shared affiliation words (Torvik et al., 2005), exact match on the institution after pre-processing the affiliation into canonical form (French, Powell, & Schulman, 2000), and geodesic distance using a geocoder such as Google Maps (maps.google.com). Each approach is associated with a different spectrum

of errors (e.g., Google Maps returns University of Chicago when given the input University of Illinois, Chicago). There are also a variety of ways to weight the features (e.g., Jaccard, TF*IDF, Jaro-Winkler, Levenstein) or to assess partial matches between the features (e.g., based on edit-distance vs. BLAST similarity [Krauthammer, Rzhetsky, Morozov, & Friedman, 2000]).

Most, but not all, approaches to disambiguation involve collapsing all of the feature scores into a single numeric value that indicates the degree of similarity between a pair of papers. However, combining features by a simple weighted sum of the individual features will introduce errors if the features are not independent of each other—if they show partial redundancy or other interactive effects (Torvik et al., 2005). Some models deal with this issue by transforming features into sets of latent (or hidden) variables that are not correlated. For example, in probabilistic latent semantic analysis (PLSA) (e.g., Hofmann, 1999; Song, Huang, Councill, & Giles, 2007), the relationships among documents, names, and text words are connected by a set of independent variables, each of which represents a latent topic. Another similar approach is latent Dirichlet allocation (LDA) (Bhattacharya & Getoor, 2006, 2007; Blei, Ng, & Jordan, 2003; Song et al., 2007), a hierarchical Bayesian model that captures entities at multiple levels, using a hidden (or latent) label for a group of people that tend to publish papers together and a hidden label for individuals, where the hidden labels are to be inferred simultaneously. In contrast, some models employ sets of features that are independent of each other, a situation which is simpler to analyze but requires vigilance to detect and deal with cases of non-independence (e.g., Torvik & Smalheiser, 2008; Torvik et al., 2005).

When a collection of articles is examined in a pairwise fashion, it is not uncommon to find that different pairwise judgments give results that are inconsistent with each other, especially if the assignments are performed independently. For example, a paper written by J.

Thompson and N. Willow does not share any co-authors with one written by J. Thompson and W. Fried, so they might be rated as having low similarity. Yet, if there exists a third paper by J. Thompson, N. Willow, and W. Fried, it is quite likely that the same J. Thompson wrote all three papers. One way to detect and correct these so-called transitivity violations is to look at sets of three papers at once and assess the transitivity (Huang, Ertekin, & Giles, 2006; Soler, 2007; Torvik et al., 2005). Higher-order constraints can also be examined (Culotta, Kanani, Hall, Wick, & McCallum, 2007). It is important to resolve transitivity violations because they may represent the most difficult cases and often the most important ones. For example, transitivity violations may arise when one person has written many articles with two different, non-overlapping sets of co-authors—creating a challenge to determine whether all of these papers were written by the same person. In addition, the process of forming clusters of articles assigned to each individual (agglomerative clustering) breaks down even for a small rate of transitivity violations.

Authorship Attribution and Stylometry

Authorship attribution is the problem of deciding who wrote certain documents. Stylometry (e.g., Holmes, 1998; Holmes, Robertson, & Paez, 2001) or computational stylistics refers to the process by which one creates literary "signatures" (Binongo, 2003, p. 16) of authors by summarizing the writing style in their documents, often relative to other, comparable authors. These signatures most often include the frequencies of common, topically non-specific words but may also include the sentence structure and grammar, the context (a blog vs. scholarly article), statements of fact vs. opinion, and positive vs. negative flavor of the comments. In general, stylometry operates on a substantial body of author-generated text, whereas author name disambiguation of bibliographic records has generally considered relatively short, formatted

metadata. Stylometry methods are often used by literary scholars to assign authorship to documents that are anonymous or disputed (e.g., Shakespeare vs. Marlowe, The Federalist Papers). The problem is most often formulated as assigning authorship of a given document to a small number of possible authors but is also used to characterize individual authors, for example, a novelist's change in literary style over time (Can & Patton, 2004). Authorship attribution has applications in criminal cases (e.g., to detect plagiarism or to identify authors of ransom notes or malicious computer code) and statistical tests have been proposed for authorship attribution (akin to DNA testing) (Madigan, Genkin, Lewis, Argamon, Fradkin, & Ye, 2005). Stylometry has also been used to uncover an author's gender (Koppel, Argamon, & Shimoni, 2002), age, native language, the nature of their opinions, and so forth. However, stylometry is not well suited to the analysis of multi-authored texts and has not (to our knowledge) been applied to author name disambiguation of the scholarly literature.

Entity Resolution and Word-Sense Disambiguation

Entity resolution refers to the process of identifying multiple references to the same object and distinguishing them from mentions of different objects. For example, a single name may appear on different Web pages, representing many different individuals (Mann & Yarowsky, 2003). Zoominfo (www.zoominfo.com) has created a database and search engine designed to identify and disambiguate mentions of companies and related individuals on the Web. A closely related research problem is that of word sense disambiguation (e.g., Schuemie et al., 2005) where the goal is to determine the meaning of a word instance based on the context in which it occurs. Gene names (Chen, Liu, & Friedman, 2005) and abbreviations (Zhou, Torvik, & Smalheiser, 2006) tend to be particularly ambiguous across scientific papers (e.g., the word

"cold" could refer to the common cold, chronic obstructive lung disease, freezing temperatures, etc.). Entity resolution and word sense disambiguation operate on natural language text; author name disambiguation operates primarily on metadata about authors and articles.

Record Linkage in Administrative Databases

Record linkage refers to the process by which one identifies multiple records in a database, or across two different databases, as referring to the same individual. Record linkage is an important issue in public health records (Jaro, 1995) and census records (Winkler, 1995, 1999); it has a long and rich history based on the work by Fellegi and Sunter (1969) in statistical modeling. This work has been transferred into the field of author name disambiguation, including similarity measures (e.g., the Jaro-Winkler similarity measure) and the conditional independence assumption (contribution of features are independent) resulting in a model that is characterized by a linear combination of features. Record linkage research focuses on name variants and address normalization (Churches, Christen, Lim, & Zhu, 2002) and is often based on information such as mailing address, telephone number, birth date, and gender: items generally unavailable for author name disambiguation. The scope of record linkage is also different from that of author name disambiguation, in which it is common for an author to have written hundreds of papers, on a variety of topics, with many different co-authors, spanning many years, with several different affiliations.

Giles and Colleagues

At Pennsylvania State University, C. Lee Giles runs the CiteSeer project (citeseer.ist.psu.edu), which has a query interface to a collection of full-text manuscripts (pdf and

postscript documents) that was generated by crawling the Web (Giles, Bollacker, & Lawrence, 1998; Li, Councill, Lee, & Giles, 2006). Giles and colleagues have published an impressive body of research articles outlining various methods for author name disambiguation. Their earlier approaches emphasized scalability and computational efficiency and suffered from poor prediction accuracy, probably because they used a limited set of similarity features (only co-author names, title of paper, and title of publication venue). They published two different unsupervised approaches: the spectral clustering method (Han, Zha, & Giles, 2005), which requires pre-specifying the estimated number of author clusters, and the DBSCAN approach (Huang et al., 2006), which is highly efficient and resolves transitivity violations. They also described an active learning approach to generating training data to reduce error rates and in an earlier study (Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004) they described two supervised methods using a hybrid of naïve Bayes and support vector machines (Burges, 1998). Their most recent approach (Song et al., 2007) utilizes more extensive metadata of citations and authors, as well as the full first page of the paper to associate authors with topics (a collection of words) using two different unsupervised latent models: probabilistic latent semantic analysis and latent Dirichlet allocation. This approach shows much improvement from their previous unsupervised methods when applied to a sample of names from the CiteSeer database.

Getoor and Colleagues

Bhattacharya and Getoor (2006) adapted LDA to author name disambiguation; in this scheme, authors are thought of as belonging to one or more groups of individuals who tend to co-author papers. Their approach simultaneously discovers clusters of author-individuals and clusters of papers by these individuals. They used an unsupervised training method and an

Expectation Maximization (EM) algorithm coupled with Gibbs sampling for parameter estimation. Introducing groups into the model comes at a significant computational cost, however, because the method was about 100 times slower than an alternative approach (Bhattacharya & Getoor, 2007).

In the latter paper, Bhattacharya and Getoor (2007) used the assignment of papers to one person as information to help assignment of papers to other authors. They refer to this as *collective entity resolution*: Given two papers both with J. Smith and C. Blatt as authors, if one can determine that the two instances of C. Blatt do *not* refer to the same person, then it is much less likely that the instances of J. Smith refer to the same person as well. Taking this effect into account will help most in the contexts where there is a high level of ambiguity (e.g., where C. Blatt may refer to many different people) and this principle can be used to help resolve ambiguity that is present in other types of features such as affiliations. To fit the model they carried out bootstrapping (i.e., forming initial estimates of the model's parameters by sampling from the dataset) and considered high vs. low ambiguity domains separately. They started with the most confident assignments before addressing the less confident ones. The overall similarity measure between a pair of records is computed as a weighted combination of the feature similarity (based on pairwise comparisons) and the relational similarity (based on previously disambiguated people). This requires a manually adjusted weighting parameter that does not seem to have a single optimal value across different contexts; clustering requires specifying a similarity threshold. They also compared several different relational similarity measures drawn from Liben-Nowell and Kleinberg (2003, 2007), including simply counting the number of common co-authors. This method is an improvement on their earlier model (Bhattacharya & Getoor, 2006) but scalability is still a problem: Computation time is more than ten times longer

than needed for baseline computations (straightforward pairwise comparisons) on relatively small datasets and this difference should increase with the number of relations and size of the dataset.

Bilgic, Licamele, Getoor, and Schneiderman (2005) described an interactive disambiguation system called D-Dupe (www.cs.umd.edu/projects/linqs/ddupe), which takes as input a set of papers (with metadata) and presents the authors in a co-authorship network, highlighting potential ambiguity in order to facilitate manual disambiguation. The underlying system also allows for manually adjusting the weights given to each feature. The D-Dupe software is free to use for non-commercial purposes, with licensing options for commercial uses.

McCallum and Colleagues

McCallum and colleagues have published a series of influential papers on author name disambiguation and related problems and methods (Culotta & McCallum, 2006; Culotta et al., 2007; Kanani et al., 2007) and have created a digital library called Rexa (rexa.info) that is focused on computer science literature (including National Science Foundation grants) currently containing 7 million records. The site supports searching and browsing results of their information extraction and disambiguation efforts.

The group has described methods for representations that go beyond pairwise comparisons to include 3-way and higher-order simultaneous comparisons among documents (Culotta & McCallum, 2006). Kanani and colleagues (2007) described a method for supplementing author and paper metadata with information drawn from the Web by using active learning to reduce the manual effort involved in assessing Web queries. In a complementary strategy, Culotta and colleagues (2007) presented a model that takes advantage of aggregate

constraints associated with a cluster of papers associated with an author. For example, they noted that in any given year, an author is unlikely to publish more than 30 papers and is likely to have only one or two affiliations or e-mail addresses. Primary features of the model include first and middle names, rarity of last name (based on census data), several measures of title similarity, e-mail, affiliation, and venue of publication when available, as well as several higher-order features (defined as a function encompassing several primary features). They showed improved performance over baseline on the Rexa and DBLP datasets when including aggregate constraints in the model, but only when an appropriate training paradigm was used (error-driven training examples and the MIRA [margin infused relaxed algorithm] ranking loss function).

Other Approaches

Malin, Airoldi, and Carley (2005) and Malin (2005) used a network model for the Internet Movie DataBase (IMDB) to define features that are based upon the social network of individuals who have worked together. Like Bhattacharya and Getoor (2007), they found that features based on social network data do contribute substantially to disambiguation.

Hill and Provost (2003) addressed the problem of identifying the author of a paper given just the citations listed on the paper; they showed that the author can be identified in approximately 25 to 45 percent of the cases based solely on this information, as assessed using a sample of approximately 30,000 papers from the High-Energy Physics Literature (SPIRES-HEP) database. Giles and Councill (2004) have also discussed the use of acknowledgments extracted from articles for improving disambiguation.

Tan, Kan, and Lee (2006) used ambiguous author names within DBLP as the starting point for Google searches (on title and author name) and examined the ten top URLs that

resulted from each search. This information alone sufficed to disambiguate names with an accuracy of about 80 percent.

Yin and colleagues (2007) have focused on the problem of object distinction, that is, the case where two distinct individuals share the same name (in contrast to the problem of merging different name instances that correspond to the same individual). It may not be a coincidence that one of the co-authors of this paper (Philip S. Yu) has a name that is associated with the greatest number of publications in DBLP. They utilized two different similarity measures. One measure is based on matches on shared features such as co-author names and venue (e.g., name of conference proceedings) using the Jaccard similarity coefficient. The other measure is based on a "collective random walk probability" (Yin et al., 2007, p. 1245). The two measures were then combined using a geometric average.

DiLauro, Choudhury, Patton, Warner, and Brown (2001) and Warner and Brown (2001) describe a project to automate the authority control process involving a collection of approximately 29,000 pieces of sheet music, using "commonness" of name, publication date versus author's date of birth or death, and author's affiliation.

Thomson Scientific and Elsevier, two commercial competitors, each maintain a subscription-based bibliographic database (Web of Science and Scopus, respectively) containing more than 30 million records that undergo predictive author name disambiguation. The clusters of papers by predicted author-individuals have been incorporated into their query interfaces but their predictive methods have not been described in enough detail to discuss in this review.

Summary

In summary, author name disambiguation within bibliographic databases is a very active

area of research in the computer science community. Many different features have been employed for modeling and several quite imaginative and powerful approaches have been proposed that include higher-order comparisons among documents, groups of co-authors and other social network data, and external information obtained from Web pages. The limiting factor in performance is not access to sufficient information, but rather the computational load involved in taking all of the available information into account, which currently limits their extension to very large databases or digital libraries.

In the next section, we discuss our own approach to author name disambiguation in Medline. Like several other groups, the Author-ity model employs features taken from the metadata (the Medline record fields). However, instead of collapsing the feature scores to a single number, the features are encoded as a multi-dimensional vector. This allows for very detailed modeling of each feature as well as investigating how each feature interacts with the others. Because the results of all possible vector scores are computed in advance, a pair of articles can be disambiguated by fast look-up. Thus, as we shall see, the Author-ity model combines high performance (in terms of both recall and precision) with high scalability and efficiency, provided that the bibliographic metadata are highly accurate and complete, and provided that the features of documents added to the database over time are relatively stable.

The Author-ity Project

The Author-ity project arose as an off-shoot of the Arrowsmith two-node search tool (Smalheiser & Swanson, 1998; Smalheiser, Torvik, Bischoff-Grethe, Burhans, Gabriel, Homayouni, et al., 2006; Swanson & Smalheiser, 1997; Torvik & Smalheiser, 2007). In the two node search, a user looks for items or concepts that link two disparate sets of Medline articles

implicitly, yet in a meaningful way. A common scenario involving implicit information arises when an investigator finds experimentally that two phenomena, previously thought to be unrelated, are unexpectedly related in some way, and would like to find existing knowledge that might shed light on potential mechanisms that may link them. Alternatively, an investigator may hypothesize that a link exists between two disparate phenomena and wish to assess whether the existing literature provides any implicit support for the hypothesis that would encourage experimental testing. A variety of daily information-seeking activities also involves looking for items or concepts that are shared by two different sets of articles: For example, a physician may want to compile a list of symptoms that are shared in two different diseases or a student may wish to browse the literature of an unfamiliar discipline for information that is likely to be relevant to his or her home discipline (Smalheiser et al., 2006). Thus, text mining for implicit information includes situations in which one is searching for known findings, as well as identifying novel hypotheses or previously unreported links between two different sets of documents (Smalheiser, 2005; Torvik & Smalheiser, 2007).

There is also a range of situations in which one would like to identify not concepts, but investigators who link two disparate literatures in a meaningful way. This includes identifying investigators who have published in both literatures, as well as those who have a less direct relationship, for example, have published in one literature and have collaborated with people who have published in the other literature. In order to be able to examine such connections, it is first necessary to identify the individuals involved, not merely their names. Thus, the first step in this project was to model and disambiguate all author names in Medline.

It should be emphasized that, because of size and scope, the difficulty of disambiguating authors in Medline is much greater than for the other examples discussed in the section on

research approaches to author name disambiguation. For example, name ambiguity is much more pervasive in Medline than in CiteSeer (Bhattacharya & Getoor, 2007). For common names, such as W. Wu, scores of different individuals share the same name and publish on the same topics; and several different individuals have similar or even identical affiliations.

Early Work

In phase I of this project, we created a statistical model that predicts, for any two Medline articles sharing the same author (last name, first initial), the probability that they are written by the same individual (Torvik et al., 2005). The model is based on a comparison vector incorporating six features of the Medline record—shared title words, journal name, co-author names, medical subject headings, language, and affiliation—as well as two distinctive features of the author name itself (presence of middle initial and suffix). This vector method takes into account nonlinear and interactive effects across features; moreover, positive and negative training sets are very large and constructed automatically, which allows for very robust results. Name and article attributes are assumed to be independent of each other (some exceptions occurred and were dealt with), which allows us to use training sets based on name to characterize article attributes, and vice versa.

Thus, given any pair of papers bearing the same author (last name, first initial), we compute the comparison vector and observe its relative frequency in the positive vs. negative training sets (the r-value). If the observed profile is much more frequent in the positive set than in the negative set, it is likely that the two papers were written by the same individual. However, the r-value is insufficient for estimating the probability that a pair of articles is written by the same individual: One also needs an estimate of the a priori probability of match for the given

name (Torvik et al., 2005). For example, if the name is very unusual (e.g., N. Smalheiser), the chances are better that any two randomly chosen papers with that name are written by the same individual than if the name is common (e.g., J. Larson). The pairwise model has been implemented as a public Web-based query interface called the "Author-ity" tool (arrowsmith.psych.uic.edu). The user enters a specific (last name, first initial) and is shown a list of articles bearing that name; when the user chooses a specific paper from the list, the output displays the articles ranked in descending order of probability that they were authored by the same individual.

More Recent Advances

Phase 2 of the Author-ity project proceeded with support from a two-year grant from the National Library of Medicine. The original pairwise model was enhanced in a variety of ways; for example, first names were added as a new dimension of the comparison vector. First names were extracted from Medline records when present (these began to be recorded in 2002) and supplemented with information extracted from publishers' pages on the Internet (only the public, unrestricted pages of online journals containing tables of contents were used for extraction). In scoring a match between the first names present on two different articles, the match score was weighted according to the frequency of that name in Medline as whole and partial matches; name variants and nicknames were scored appropriately as well (Torvik & Smalheiser, 2008). E-mail addresses and their variants were also incorporated into the model (e-mail addresses were also employed to create an alternative set of positive and negative training sets). Some violations of the independence assumption were found and corrected; for example, people named Kim tend to have an affiliation in Korea more than expected by chance. This would bias the probability

estimates unfairly and was corrected by looking for cases in which there were significant cross-correlations between last name and affiliation country, then removing the country name from consideration when computing the comparison vector for that name. Finally, we improved the methods of estimating the prior probability for a given name and optimized weighting parameters for handling transitivity violations (Torvik et al., 2005, see also Huang et al., 2006; Reuther, Weber, Walter, Ley, & Klink, 2006).

Using the enhanced and corrected pairwise model (Torvik & Smalheiser, 2008), pairwise comparisons were made for all papers in Medline that shared a last name and first initial within the author name field. Agglomerative clustering was carried out in two different ways: stopped at a point of high precision or at the maximum likelihood point (Torvik & Smalheiser, 2008). Whereas some other disambiguation efforts have chosen a high-precision endpoint, we found that high-precision clusters very often split the papers written by one individual into multiple clusters, corresponding to different groups of co-authors, different topics, different affiliations, and so on. This problem was minimized with the maximum-likelihood clustering strategy and, surprisingly, we found that maximum-likelihood clusters contained relatively little lumping of distinct individuals into the same cluster. Thus, the maximum-likelihood strategy was adopted generally in our project.

The resulting disambiguation dataset was evaluated extensively from several different perspectives (Torvik & Smalheiser, 2008). We estimate that this clustering strategy has a recall of 98.8 percent of all papers written by the same individual (1.2 percent of papers can have a different last name, either due to variant spelling, mis-spelling, or different last names given). The precision and the accuracy of assigning a given paper to a given author-individual cluster are almost 98 percent across the dataset as a whole. Lumping (putting two different individuals into

the same cluster) affects less than 0.5 percent of all clusters, whereas splitting (assigning papers written by the same individual to two or more clusters) appears to affect less than 2 percent of all papers. The clustering solution was highly robust against small changes in pairwise probability estimates. The major predicted author-individual cluster corresponding to a given person fails to capture about 1.6 percent of that person's output, namely, those articles that are highly divergent from the others. Nevertheless, the current dataset shows excellent performance and appears to be suitable for initial use by the scientific community. The dataset will be available upon request to any nonprofit academic research group; we are in the process of creating a public Web interface to the dataset as part of the Arrowsmith suite of informatics tools (hosted at arrowsmith.psych.uic.edu).

Incorporating relatively simple extensions to the model in the future should result in even better recall and precision. For example, recall should improve by taking into account idiosyncrasies in the way that author names are encoded in Medline, such as spelling errors and alternative spellings in last names, compound last names, or nicknames and alternative first names that do not share the first initial (e.g., Jerry vs. Gerald). Precision should improve by employing additional types of information from the Medline record; for example, we have not yet utilized similarity in terms used in the abstract field as a feature in the statistical model (cf. Wilbur & Yang, 1996). Furthermore, instead of simply counting similarity in words used in the affiliation fields, words should ideally be mapped to canonical forms of affiliations (French et al., 2000) or mapped onto a geographical system to distinguish between words that describe countries, cities, institutions, departments, streets, an so on.

We have not yet utilized some of the kinds of information that have been employed by other groups and reviewed in the section on research approaches to name disambiguation (e.g.,

co-authorship groups or aggregate constraints), nor have we employed information that is external to the bibliographic record (e.g., using information from the Web). This is probably one of the reasons why the Author-ity model is very efficient and scalable to very large, highly ambiguous datasets. However, adding some of these other types of information should improve performance further and we hope to explore how to do so without increasing computation time excessively. Also, we still need to explore how best to update the dataset as new publications are added to Medline each week.

Strengths and Weaknesses of Existing Approaches: Challenges for Future Research

There is no single paradigmatic author name disambiguation task—each bibliographic database, each digital library, and each collection of publications has its own unique set of problems and issues. Collections differ in size, author diversity, and curation reliability, as well as in the types of metadata that are assigned to each publication, the cultural context in which the data are used, and the rate of growth of new items. For certain purposes (e.g., awarding the Nobel Prize to the author of a breakthrough), it may be very important to achieve a high level of accuracy in disambiguation. For other purposes (e.g., as an aid to routine information retrieval), it may suffice to assign a high proportion of a person's articles correctly, with little penalty occurring if some articles are missed or mis-assigned.

Certainly, the existing machine-learning models discussed in the sections on research approaches to author name disambiguation and the Author-ity Project have room for further improvement in precision and recall, either by encompassing additional features or by combining aspects of different models into one. However, optimizing performance is only one of the goals of future research. A quick-and-dirty algorithm may still be preferred over a high performing one

if it is scalable, efficient, rapid, and easy to pre-compute (so that disambiguation does not need to be computed in real time). Each of these represents a major computing challenge. Moreover, in cases where new publications or Web sites arrive in an ongoing stream, they ideally should be disambiguated and clustered in an online fashion (e.g., Bhattacharya, Geetor, & Licamele, 2006).

When assessing machine learning models that employ unsupervised methods, it should be noted that high accuracy can often be achieved simply by applying the methods to datasets that have relatively low levels of ambiguity or by reporting the results on test examples that correspond to the optimal choice of algorithm parameters (such as the pre-specified number of clusters desired). Thus, performance parameters need to be taken with a grain of salt until they are demonstrated to be robust across different collections and show other desirable features (e.g., scalability, efficiency).

Furthermore, different disambiguation datasets need to be linked together; for example, it is desirable to identify the same individuals and their publications when they appear across a variety of (overlapping) public and proprietary biomedical databases and when they appear in other disciplines (engineering/computer science databases) or in other types of publications such as books, magazines, and patents. It is desirable to link the disambiguation datasets to other types of information related to the individuals—their home pages on the Web, news articles that mention them, and so forth (DeRose et al., 2007). However, some record linkage projects (e.g., involving sensitive data on health, finance, and crime) have met with public outcry because they were perceived as being secret, revealing, imposed, inaccurate, or precursors to administrative action (U.S. General Accounting Office, 2001). Thus, it is important to keep the public in mind when embarking on an author disambiguation project. At the very least, author disambiguation research should be transparent.

Using Author Name Disambiguation Datasets to Create and Analyze Networks that Relate Authors and Literatures

The methods used to create disambiguation datasets are similar, in principle, to methods that have been applied to resolution and disambiguation of words and other kinds of entities. But authors are *people*—intelligent, emotional, beings who think, learn, and change over time; collaborate with others; choose scientific topics; change jobs occasionally; make discoveries (as often as possible); and make mistakes (all too often). An author potentially can be attached to many different types of complementary data. Official data include the list of publications assigned to them in the disambiguation dataset (and their accompanying features), as well as job titles, affiliations, memberships in societies, awards, co-authors, and so on. Personal data include birthplace, family situation, educational level and institution, mentors, gender, disabilities, ethnicity, marriage status, occupation of spouse, and so on. Thus, the factors that influence scientific and publication behavior are far more complex, rich, and interactive than those needed to disambiguate the sense of a word in a specific passage of text. Author name disambiguation is the first and pivotal step in opening up a major new field of analysis that is person-centered, not just document-centered.

The information in disambiguation datasets is valuable for several different types of analyses; for example, in bibliometrics and scientometrics, one would like to understand which factors determine productivity and to analyze publication patterns and trends of individuals and larger groups (e.g., centers, institutes, geographical regions, or disciplines). The Author-ity model does not simply carry out disambiguation as a yes/no assignment of papers to persons, but employs a metric of document similarity that can help to characterize the nature of a person's

research output; this helps characterize the papers that are known to belong to a given individual (e.g., to identify and characterize those papers that represent outliers). In policy studies, one would like to trace the impact of specific interventions (e.g., funding initiatives, training grants) on subsequent productivity or patent activity. Social scientists would like to model the factors that influence scientific discovery, collaborations, and formation of informal or formal social networks (e.g., conferences, societies, journals). These factors may vary by discipline (e.g., biology vs. mathematics), culture (e.g., Western Europe vs. East Asia), laboratory structures (e.g., individual principal investigators vs. large centers), all of which require author disambiguation data for accurate analyses. Last, but not least, individual scientists would like to have access to disambiguation data (and attached links, e.g., to personal home pages on the Web, author profiles, and CVs) in order to retrieve articles by a particular person or to find potential collaborators.

We expect that author name disambiguation projects will begin to inform how names and other metadata are encoded in publications. At present, each publisher may encode names in a different manner; there is no standard way to represent affiliations and there is no general capturing of certain types of information (e.g., job titles, degrees held) that would be very helpful for disambiguation. Moreover, at present the metadata are not generally made public for others to read and extract, which would assist in coordinating disambiguation efforts across publishers and disciplines. Author name disambiguation should also help global efforts that, for example, seek to identify a set of highly distinguishing attributes to encode authors in metadata fields (Hickey, O'Neill, & Toves, 2002; IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998).

Investigator-Literature and Literature-Literature Networks

In line with the original motivation to develop Author-ity, once a disambiguation dataset is available (in this case, for Medline authors), it becomes feasible to create and analyze networks that relate each individual *I* to any given discipline or topic *T*. (This idea can be extended to include any coherent collection of articles or any other individual [who can be represented by the set of articles *T* that he or she has authored].) This is a mixed network, in that both individuals and sets of articles are nodes, which are connected by directed links from an individual I_i to a set of articles T_i. Each link is associated with an integer, depending on the nature of the association between the two nodes. For example, if an individual I_i has authored one or more papers within a set of articles T_i, then the link is assigned a value of 0. If an individual I_i has not authored any papers within a set of articles T_i, but has co-authored one or more papers with someone who has published within *Ti*, then the link is given a value of 1. If an individual I_i has not co-authored one or more papers with someone who has published within *Ti*, but has co-authored papers with someone that has link 1 status, then the link is assigned a value of 2, and so on. Thus, each individual I_i is said to have *n*th degree status with relation to a given set of articles T_i.

Networks that relate authors to literatures can be constructed in a number of ways. One can assign link values to examine implicit topics, items, or concepts rather than co-authors: For example, if an individual I_i has not authored any papers in T_i but has written papers on some of the same topics, the link may be assigned a value of 1. Yet another approach is to consider networks in which sets of articles *T* (rather than individuals) are nodes and examine how any pair of literatures T_i and T_j are linked via the investigators who have either written articles in T_i and T_j or have co-authored with these individuals.

An important aspect of the network approach is that at least three types of information can be used to encode features that are attached to a node in the network. First, one can use information that is internal to the datasets (e.g., topics taken from the articles assigned to each individual). Second, one can compute a variety of network properties (Liben-Nowell & Kleinberg, 2007) (e.g., whether the node is a hub). Third, information can be extracted from external sources (e.g., the age or address of the investigator at the time a given article was written). External information also includes the individual's status at the time (as a student, postdoc, independent investigator, or center director) as well as the nature of the work—certain types of investigators such as statisticians, bioethicists, or microarray facility supervisors straddle a variety of disciplines. Citation information can be employed as well. Potentially one could utilize external information arising from a person's blog or their purchasing behavior on Amazon.com. Such a highly multi-dimensional set of features provides a very rich vein for data mining but requires the use of sophisticated multi-level, multi-dimensional network analysis tools (Contractor, Wasserman, & Faust, 2006; Monge & Contractor, 2003) and data-mining approaches (Liu, Han, Xin, & Shao, 2006).

Regardless of the specific network, the main idea is that any investigator I_i can be placed, directly or indirectly, in relation to any set of articles T_i, and that pairs of literatures and pairs of investigators can be similarly related to each other within the same framework. Then, one can proceed to model and analyze the factors that determine these relationships. Because the approach is so general, a wide variety of questions needs to be addressed:

- For example, certain investigators may act as a bridge between two disparate disciplines, either directly (they have published in both subjects) or indirectly (they have published on

related topics or collaborated with investigators researching these subjects). Can we identify factors that distinguish these bridge investigators from others?

- Conversely, certain topics may be regarded as interdisciplinary, or frontier areas, or as emerging burst areas of high current interest. Can we identify factors that would detect these automatically?
- Certain aggregate groups of investigators (centers, institutes, schools) are much more productive than others. Can we identify factors that would allow one to predict, in advance, which groups are likely to be more productive, or at least identify factors that correlate with high productivity?

Collaboration Networks

Whereas literature-centered networks are created to ask questions about publication behavior, a different (and simpler) type of network is more suited for asking questions about collaboration behavior: Each investigator *I* is a node; if I_i and I_j have jointly co-authored one article, they are joined by a non-directed link of strength 1. If they have co-authored two articles, the link has strength 2, and so on. Again, a very large number of features can be associated with each node/investigator: internal features, inherent network features, and external information. One can even utilize information that is obtained from the investigator-literature networks (e.g., if investigator I_i stands in *n*th degree relation to another investigator I_j in an investigator-literature network, then this fact can be used as one of the features in the collaboration network).

The study of scientific collaboration is an entire field in itself (Sonnenwald, 2007) and collaboration networks can be analyzed in many different ways. One can try to understand which

factors determine whether two persons will collaborate (resulting in a joint publication). One can also examine networks as they evolve over time. These basic modeling studies set the stage for creating user-friendly tools that will allow a person to find potentially good collaborators for a given problem. Because one person might be an excellent potential collaborator for a large number of people, far too many to work with simultaneously, it is necessary to consider constraints and limiting factors as well.

Conclusions

The term data mining is based on a metaphor in which nuggets of knowledge are sought within a large stack of irrelevant facts—the idea being that data mining identifies and refines something that is already present from the outset. It is true that information cannot be created out of nothing and that once lost, it cannot be recovered. Yet it is also true that systems can evolve radically (e.g., from primordial soup to man) without contradicting the laws of physics. Small re-arrangements of existing elements can create a new entity that, in turn, can have a major innovative impact in some new arena. Introducing movable type created the printing press, which in turn created a literate populace leading to a cultural revolution within Europe. Applying the principles of another modern printing process (lithography) to biological chemicals led to the development of the DNA microarray, which has revolutionized the study of gene expression (Lenoir & Giannella, 2006). Another way to create innovation is to assemble large sets made of noisy, imperfect, unreliable elements, which achieve a certain level of usefulness through redundancy and validation: In the scientific arena, an example of this is the formation of expressed sequence tag (EST) databases; in the textual arena, Wikipedia comes to mind.

In this vein, we believe that attaching a person to a set of documents is a key step toward

a major breakthrough in information science. Linking together a large number of heterogeneous, disparate data elements (descriptive of, or relating to, a particular individual and that reside within Web pages or across many different repositories, databases, or text collections) creates a very rich arena for data mining that would not otherwise exist. Author name disambiguation employs some elements of bricolage insofar as assignments are made using a combination of existing types of features; and it employs redundancy insofar as it makes use of implicit and higher-order interactions (see the sections on research approaches to author name disambiguation and the Author-ity Project).

Author name disambiguation has strategic importance that goes far beyond knowing who wrote what. The case of collaboration networks is merely the simplest example of how disambiguation data can underlie the creation of new resources and tools that open up novel types of investigation. As information science becomes progressively more person centered, not just document centered, we expect to see ripples that will affect the world of publishing, the semantic Web, the design of search engines, and the indexing of data collections.

Acknowledgments

The Human Brain Project/Neuroinformatics research was funded jointly by the National Library of Medicine and the National Institute of Mental Health. We thank our colleagues Clement Yu, Wei Zhou, Don Swanson, and Marc Weeber for stimulating discussions.

References

Andrade, M. (2006). *WikiAuthors*. Retrieved January 20, 2008, from meta.wikimedia.org/wiki/WikiAuthors

Bhattacharya, I., & Getoor, L. (2006). A latent Dirichlet model for unsupervised entity resolution. *Proceedings of the SIAM 6th International Conference on Data Mining*, 47–58.

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, *1*, 1–36.

Bhattacharya, I., Getoor, L., & Licamele, L. (2006). Query-time entity resolution. *Proceedings of the ACM SIGKDD 12th International Conference on Knowledge Discovery and Data Mining*, 529–536.

Bilgic, M., Licamele, L., Getoor, L., & Schneiderman, B. (2005). D-Dupe: An interactive tool for entity resolution in social networks. *Proceedings of the IEEE Symposium on Visual Analytics, Science and Technology,* 43–50.

Bilenko, M., Kamath, B., & Mooney, R. J. (2006). Adaptive blocking: Learning to scale up record linkage. *Proceedings of the IEEE Computer Society 6th International Conference on Data Mining*, 87–96.

Binongo, J. G. N. (2003). Who wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, *16*, 9–17.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 951–991.

Bohne-Lang, A., & Lang, E. (2005). Do we need a Unique Scientist ID for publications in biomedicine? *Biomedical Digital Libraries*, *2*, 1.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167.

Can, F., & Patton, J. M. (2004). Change of writing style with time. *Computers and the*

Humanities, *38*, 61–82.

Chen, L., Liu. H., & Friedman, C. (2005). Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, *21*, 248–256.

Churches T., Christen, P., Lim, K., & Zhu, J. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making*, *2*, 9.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, *6*, 57–71.

Contractor, N., Wasserman, S., & Faust, K. (2006). Testing multi-theoretical multilevel hypotheses about organizational networks: An analytic framework and empirical example. *Academy of Management Review*, *31*, 681–703.

Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. *Proceedings of the AAAI 6th International Workshop on Information Integration on the Web*, 32–37.

Culotta, A., & McCallum, A. (2006). Tractable learning and inference of high-order representations. *Proceedings of the International Conference on Machine Learning Workshop on Open Problems in Statistical Relational Learning*. Retrieved January 11, 2008, from www.cs.umd.edu/projects/srl2006/papers/srl06-culotta.pdf

Databases in peril. (2005). *Nature Cell Biology*, *7*, 639.

DeRose, P., Shen, W., Chen, F., Lee, Y., Burdick, D., Doan, A., et al. (2007). DBLife: A community information management platform for the database research community. *Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research*, 169–172. Retrieved January 12, 2008, from www.cidrdb.org/cidr2007/papers/cidr07p19.pdf

Dervos, D. A., Samaras, N., Evangelidis, G., Hyvärinen, J., & Asmanidis, Y. (2006). The Universal Author Identifier System (UAI_Sys). *Proceedings of the 1st International Scientific Conference, eRA: The Contribution of Information Technology in Science, Economy, Society and Education*. Retrieved January 12, 2008, from dlist.sir.arizona.edu/1716

DiLauro, T., Choudhury, G. S., Patton, M., Warner, J. W., & Brown, E. W. (2001). Automated name authority control and enhanced searching in the Levy collection. *D-Lib Magazine, 7*. Retrieved January 20, 2008, from www.dlib.org/dlib/april01/dilauro/04dilauro.html

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology*, *4*, e157.

Felligi, I., & Sunter A. (1969). A theory for record linkage. *Journal of the American Statistical Society*, *64*, 1183–1210.

French, J. C., Powell, A., & Schulman, E. (2000). Using clustering strategies for creating authority files. *Journal of the American Society for Information Science, 51*, 774–786.

Garfield, E. (1969). British quest for uniqueness versus American egocentrism. *Nature*, *223,* 763.

Giles, C. L., Bollacker, K., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In *Proceedings of the 3rd ACM Conference on Digital Libraries*, 89–98.

Giles, C. L., & Councill, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing. *Proceedings of the National Academy of Sciences of the United State of America*, *101*, 17599–17604.

Han, H., Giles, C. L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 296–305.

Han, H., Zha, H., & Giles, C. L. (2005). Name disambiguation in author citations using a K-way

spectral clustering method. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 334–343.

Harnad, S. (2001). The self-archiving initiative. *Nature*, *410*, 1024–1025.

Hickey, T. B., O'Neill, E. D., & Toves, J. (2002). Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR). *D-Lib Magazine*, *8*. Retrieved January 20, 2008, from www.dlib.org/dlib/september02/hickey/09hickey.html

Hill, S., & Provost, F. (2003). The myth of the double-blind review? Author identification using only citations. *ACM SIGKDD Explorations*, *5*, 179–184.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the ACM SIGIR 22nd Annual International Conference on Research and Development in Information Retrieval*, 50–57.

Holmes, D. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, *13*, 111–117.

Holmes, D. I., Robertson, M., & Paez, R. (2001). Stephen Crane and the New York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, *35*, 315–331.

Huang, J., Ertekin, S., & Giles, C. L. (2006). Efficient name disambiguation for large-scale databases. *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 536–544.

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records: Final report* (UBCIM Publications, New Series, 19). Munich: K.G. Saur. Retrieved January 22, 2008, from www.ifla.org/VII/s13/frbr/frbr.htm

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine, 14*, 491–498.

Kanani, P., McCallum, A., & Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the Web. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 429–434.

Koppel, M., Argamon, S., & Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*, 401–412.

Krauthammer, M., Rzhetsky, A., Morozov, P., & Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene, 259*, 245–252.

Lenoir, T., & Giannella, E. (2006). The emergence and diffusion of DNA microarray technology. *Journal of Biomedical Discovery and Collaboration, 1*, 11.

Li, H., Councill, I., Lee, W., & Giles, C. L. (2006). CiteSeerx: An architecture and Web service design for an academic document search engine. *Proceedings of the ACM 15th International Conference on World Wide Web*, 883–884.

Liben-Nowell, D., & Kleinberg, J. (2003). The link-prediction problem for social networks. *Proceedings of the ACM 12th International Conference on Information and Knowledge Management*, 556–559.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*, 1019–1031.

Liu, H., Han, J., Xin, D., & Shao, Z. (2006). Top-down mining of interesting patterns from very high dimensional data. *Proceedings of the IEEE Computer Society 22nd International Conference on Data Engineering*, 114.

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005). Author identification on the large scale. *Annual Meeting of the Classification Society of North America*. Retrieved January 11, 2008, from stat.rutgers.edu/~madigan/mms/authorid-csna05.pdf

Malin, B. (2005). Unsupervised name disambiguation via social network similarity. *Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security, in conjunction with the SIAM International Conference on Data Mining*, 93–102.

Malin, B., Airoldi, E., & Carley, K. M. (2005). A network analysis model for disambiguation of names in lists. *Computational and Mathematical Organization Theory*, *11*, 119–139.

Mann, G., & Yarowsky, D. (2003). Unsupervised personal name disambiguation. *Proceedings of the ACL SIGNLL 7th Conference on Computational Natural Language Learning,* 33–40.

Maxwell, R. L. (2002). *Maxwell's guide to authority work*. Chicago: American Library Association.

Merali, Z., & Giles, J. (2005). Databases in peril. *Nature, 435*, 1010–1011.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw Hill.

Monge, P., & Contractor, N. (2003). *Theories of communication networks*. New York: Oxford University Press.

On, B.-W., Lee, D., Kang, J., & Mitra, P. (2005). Comparative study of name disambiguation problem using a scalable blocking-based framework. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 344–353.

Reuther, P., Weber, A., Walter, B., Ley, M., & Klink, S. (2006). Managing the quality of person names in DBLP. *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries*, 508–511.

Roberts, R. J. (2001). PubMed Central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences of the United States of America*, *98,* 381–382.

Schuemie, M. J., Kors, J. A., & Mons, B. (2005).Word sense disambiguation in the biomedical domain: An overview. *Journal of Computational Biology*, *12*, 554–565.

Scoville, C. L., Johnson, E. D., & McConnell, A. L. (2003). When A. Rose is not A. Rose: The vagaries of author searching. *Medical Reference Services Quarterly*, *22*(4), 1–11.

Shiffrin, R. M., & Börner, K. (2004). Mapping knowledge domains. *Proceedings of the National Academy of Sciences of the United States of America*, *101,* 5183–5185.

Smalheiser, N. R. (2005). The Arrowsmith project: 2005 status report. In A. Hoffmann, H. Motoda, & T. Scheffer (Eds.), *Discovery science 2005* (Lecture Notes in Artificial Intelligence, 3735) (pp. 26–43). Berlin: Springer-Verlag.

Smalheiser, N. R., & Swanson, D. R. (1998). Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, *57*, 149–153.

Smalheiser, N. R., Torvik, V. I., Bischoff-Grethe, A., Burhans, L. B., Gabriel, M., Homayouni, R., et al. (2006). Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators. *Journal of Biomedical Discovery and Collaboration*, *1*, 8.

Soler, J. M. (2007). Separating the articles of authors with the same name. *Scientometrics*, *72*, 281–290.

Song, Y., Huang, J., Councill, I. G., Li, J., & Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 342–351.

Sonnenwald, D. H. (2007). Scientific collaboration: Challenges and solutions. *Annual Review of Information Science & Technology, 41,* 643–681.

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence, 91,* 183–203.

Tan, Y. F., Kan, M. Y., & Lee, D. (2006). Search engine driven author disambiguation. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries,* 314–315.

Tillett, B. B. (2002). A virtual international authority file. *Workshop on Authority Control among Chinese, Korean and Japanese Languages,* 117–139. Retrieved January 14, 2008, from www.nii.ac.jp/publications/CJK-WS/cjk3-08a.pdf

Torvik, V. I., & Smalheiser, N. R. (2007). A quantitative model for linking two disparate sets of articles in Medline. *Bioinformatics, 23,* 1658–1665.

Torvik, V. I., & Smalheiser, N. R. (2008). Author name disambiguation in Medline. *Journal of Biomedical Discovery and Collaboration*. Manuscript submitted for publication.

Torvik, V. I., & Triantaphyllou, E. (2002). Minimizing the average query complexity of learning monotone Boolean functions. *INFORMS Journal on Computing, 14,* 144–174.

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology, 56,* 140–158.

U.S. General Accounting Office. (2001). *Record linkage and privacy: Issues in creating new federal research and statistical information* (GAO Special Publications: Evaluation Research and Methodology. GAO-01-125SP). Washington, DC: The Office. Retrieved January 11, 2008, from www.gao.gov/new.items/d01126sp.pdf

Vu, Q. M., Masada, T., Takasu, A., & Adachi, J. (2007). Personal name disambiguation in Web

search using knowledge base. *Database Society of Japan Letters, 5*, 53–56.

Warner, J. W., & Brown, E. W. (2001). Automated name authority control. *Proceedings of the ACM/IEEE First Joint Conference on Digital Libraries*, 21–22.

Wilbur, W. J., & Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Computers in Biology and Medicine, 26*, 209–222.

Winkler, W. E. (1995). Matching and record linkage. In B. G. Cox et al. (Eds.), *Business Survey Methods* (pp. 355–384). New York: Wiley.

Winkler, W. E. (1999). *The state of record linkage and current research problems* (U.S. Census Bureau Statistical Research Report Series, R99/04). Washington, DC: The Bureau. Retrieved January 11, 2008, from www.census.gov/srd/papers/pdf/rr99-04.pdf

Yin, X., Han, J., & Yu, P. S. (2007). Object distinction: Distinguishing objects with identical names by link analysis. *Proceedings of the IEEE 23rd International Conference on Data Engineering*, 1242–1246.

Zhou, W., Torvik, V. I., & Smalheiser, N. R. (2006). ADAM: Another database of abbreviations in Medline. *Bioinformatics*, *22*, 2813–2818.