A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation

Vetle I. Torvik and Marc Weeber
Department of Psychiatry (MC912), University of Illinois at Chicago, 1601 W. Taylor Street, Chicago, IL 60612. E-mail: vtorvik@uic.edu, marc@weeber.net

Don R. Swanson
Division of the Humanities, University of Chicago, Chicago, IL 60637. E-mail: dswanson@uchicago.edu

Neil R. Smalheiser
Department of Psychiatry (MC912), University of Illinois at Chicago, 1601 W. Taylor Street, Chicago, IL 60612. E-mail: smalheiser@psych.uic.edu

We present a model for estimating the probability that a pair of author names (sharing last name and first initial), appearing on two different Medline articles, refer to the same individual. The model uses a simple yet powerful similarity profile between a pair of articles, based on title, journal name, coauthor names, medical subject headings (MeSH), language, affiliation, and name attributes (prevalence in the literature, middle initial, and suffix). The similarity profile distribution is computed from reference sets consisting of pairs of articles containing almost exclusively author matches versus nonmatches, generated in an unbiased manner. Although the match set is generated automatically and might contain a small proportion of nonmatches, the model is quite robust against contamination with nonmatches. We have created a free, public service ("Author-ity": http://arrowsmith.psych.uic.edu) that takes as input an author's name given on a specific article, and gives as output a list of all articles with that (last name, first initial) ranked by decreasing similarity, with match probability indicated.

Introduction

Bio-informatics research databases have dramatically accelerated the pace of discovery in the biomedical sciences. Among these, Medline is the oldest and the best curated, and arguably it contains the most scientific information insofar as it summarizes knowledge that has been published across all biomedical fields. Medline and the most popular search interface, PubMed, have devoted a great deal of attention to the comprehensive retrieval of papers according to their *subject content*. Thus, each paper in Medline is indexed by hierarchical controlled-vocabulary medical subject headings (http://www.nlm.nih.gov/mesh/), and this information is used automatically in query processing by PubMed.

In contrast, relatively little attention has been given to discerning who's who in Medline—that is, which individuals have appeared as author or coauthor on specific papers. Until 2002, Medline fields did not record the full first name of an author, even if that information was given in the paper (NLM Technical Bulletin, 2001), and even now PubMed does not permit searching on author full names. Even knowing first name and middle initial does not suffice to solve the problem of assigning authors, however: There are many different individuals having the name Robert W. Williams. Furthermore, Medline fields generally record only the affiliation of the first-listed author and ignore all coauthors. (And many earlier papers lack affiliations entirely; for example, only 1.6% of papers published in 1986 encoded affiliation data, with the percentage rising to 64% by 1989). For this reason, one cannot selectively retrieve papers written by a given individual, but must request all papers having a given (last name, first initial) and sort through these manually. This can be a significant inconvenience for end users—for example, PubMed finds 208 papers with the name "RW Williams," and if one includes papers in which no middle initial was given, or which used a different middle initial (e.g., a maiden name), one needs to examine 7,500 papers. One cannot rely on affiliation, journal, or even MeSH to restrict the search accurately to identify specific individuals, particularly in the current age of multidisciplinary collaborative research: Robert W. Williams of the University of Tennessee at Memphis has appeared as coauthor on papers listing a variety of institutional affiliations in the same year,

Received April 24, 2003; revised October 16, 2003; accepted February 11, 2004

 • Published online 5 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20105

on topics ranging from visual anatomy to behavioral genetics to activation of T lymphocytes during virus infections.

The current inability to identify which papers bearing the same name (last name, first initial) are written by different individuals is also a critical impediment to research devoted to understanding the publication behavior of biomedical scientists. Citation analyses (Garfield, 1979; Noyons, Moed, & Van Raan, 1999) and collaboration graphs (connecting scientists who have published together) (Grossman, 2002; Newman, 2001) are used heavily in scientometrics and policy studies, yet these currently rely on names (Institute for Scientific Information [ISI] uses only first author names), and this severely distorts the true picture, particularly for common names. Going further, one would like to be able to describe collaboration graphs of scientists that are anchored in specific literatures (e.g., Jones has published in literature A [but not C], and has coauthored one or more papers with Smith, who has published in literature C [but not A]; Torvik, Weeber, Smalheiser, & Swanson, 2002). To develop such tools, and to analyze the structure of the resulting collaborative chains, it is important to disambiguate which individuals have written which papers in Medline. Disambiguation could also assist with everyday scientific tasks of numerous kinds: For example, study sections seeking to choose referees would benefit from identifying all of their coauthors (i.e., identifying conflicts of interest early); journal editors could assign papers for review more readily by knowing the characteristic publication profile of its prospective reviewers; and conference organizers would similarly benefit from knowing the publication profile of prospective invitees.

Here, we present a model for estimating the probability that a pair of author names (sharing a last name and first initial) appearing on two different articles refer to the same individual. We hypothesize that different papers written by the same individual will tend to share certain characteristic features, not only dealing with the author's personal information (name and affiliation attributes) but other attributes of the articles as well, much more so than pairs of papers authored by different individuals.

First, we present a comparison vector, which we call the "similarity profile," that describes, for any two papers bearing the same author (last name, first initial), how similar the two papers are across eight different dimensions: middle initial match, suffix match (e.g., Jr. or III), journal name, language of article match, number of coauthor names in common, number of title words in common after preprocessing and removing *title-stopwords*, number of affiliation words in common after preprocessing and removing *affiliation-stopwords*, and number of MeSH words in common after preprocessing and removing *mesh-stopwords*. These are calculated solely from comparing corresponding Medline fields.

Second, the similarity profile is computed for the members of two large reference sets: a match set, consisting of many (millions) pairs of papers almost exclusively coauthored by the same individual, and a nonmatch set, consisting of many pairs of papers known to be authored by different individuals. These reference sets are the heart of the model and represent Medline as a whole in an unbiased fashion. Given any pair of papers bearing the same author (last name, first initial), the similarity profile is computed, and its relative frequency is observed in the match set versus the nonmatch set. The observed relative frequencies are then smoothed, interpolated, and extrapolated for profiles that were infrequently (or never) observed in the reference sets based on the flexible monotonicity criterion, which takes into account possible nonlinear and interactive effects across dimensions.

If the observed profile is much more frequent in the match set than in the nonmatch set, it is likely that the two papers were written by the same individual.

Third, when a population of articles corresponding to a particular name is considered, the relative frequency is further weighted by an estimate for the a priori probability of match for the given name. This allows for incorporating the variability between individual names because of name frequency and the number of articles per individual into the estimates of the pairwise match probabilities. For example, if the name is very unusual (e.g., D. C. Gajdusek), the chances are better that any two randomly chosen papers with that name are written by the same individual than if the name is very common (e.g., J. Smith).

Fourth, the estimated pairwise probabilities are further improved by using information from three-way comparisons to take into account the "geometric" constraints of authorship. By the laws of probability, if paper A and paper B have a pairwise probability of 0.9 that they were written by the same author, and if paper B and paper C have a pairwise probability of 0.9, then paper A and C must logically have a pairwise probability of at least 0.8 ($=0.9 + 0.9 - 1$). Yet empirically we sometimes observe two papers that both have a high probability of match (0.9) to a third paper nevertheless have a low estimated probability of match (0.2) to each other based solely on pairwise comparisons. For example, suppose a paper by Cohen and O'Reilley and one by Cohen and Nakamura both share coauthors with a paper by Cohen, O'Reilley, and Nakamura. It is likely that it is the same Cohen in all three papers, yet if the (Cohen and O'Reilley) and (Cohen and Nakamura) papers do not share many attributes, the model might give a low estimated pairwise probability value.

Background Information

Medline

Medline is the U.S. National Library of Medicine's (NLM) premier bibliographic database containing 11,299,108 records of biomedical articles as of January 2002, and approximately 2,000 newly completed records are added daily. The National Center for Biotechnology Information (NCBI) maintains the Entrez information retrieval system, also called PubMed (http://www.ncbi.nih.gov/ entrez), which provides free, public access to Medline. When a new article is added to Medline, it is manually indexed by medical subject headings (MeSH). Each Medline record contains the title, author name(s), affiliation (if available), abstract (if available),

journal, date of publication, language, and MeSH corresponding to a particular paper. Starting in 1988, the NLM started consistently recording affiliations corresponding to the first author. The number of authors included in Medline records have changed over time. During 1966–1984, and 2000–present, all authors were included; during 1984–1995 and 1995–2000, the first 10 and 25 authors are included, respectively. Full names are included starting with articles published in 2002 and only when they are listed on the original article.

Authority Control in Bibliographic Databases

The process of maintaining cross-references and consistent forms of fields in bibliographic databases is referred to as *authority control*. Traditionally, authority files have been created and updated manually by librarians. For example, the American Mathematical Society maintains MathSciNet, which has over 380,000 authors of mathematical research articles encoded, a result of manually disambiguating author names (http://www.ams.org/mathscinet/searchauthors). There is no similar database of authors in the biomedical literature.

There are efforts being made to automatically create authority files for fields that are hard to control, such as author names and affiliations. French, Powell, and Schulman (2000) used edit distances to authority control affiliations in an astrophysics bibliographic database with approximately 450,000 records. Warner and Brown (2001) used "commonness" of name, publication date versus author's date of birth or death, and author's affiliation to authority control a collection on approximately 29,000 musical compositions. The U.S. Census Bureau authority controls (referred to as record linkage) names and addresses of administrative records to identify individuals within and across large databases, such as the 1040 and social security files (Judson, 2002; Winkler, 1995).

Word Sense Disambiguation in Free Text

There is a large body of literature on word sense disambiguation in free text. For example, Yu, Hripcsak, and Friedman (2002) identified a list of abbreviations and their full forms from Medline abstracts. There is also a large body of literature on free text authorship attribution. For example, Holmes, Gordon, and Wilson (2001) used stylometry to address the authorship of love letters supposedly written by a confederate general during the Civil War. When handling larger bodies of free text, application specific preprocessing steps (e.g., information extraction techniques) are most often needed to narrow in on words or phrases that are to be disambiguated. Authority control alleviates some of this burden because the fields are more uniform and, as such, help increase the accuracy of disambiguation.

Pairwise Document Similarity Measures

The present model is an example of probabilistic information retrieval, which is guided by the Probability Ranking Principle (Robertson, 1977; Sparck Jones, Walker, & Robertson, 2000):

> If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best to be gotten for the data.

The principle suggests utilizing the available matching information to its fullest and optimally weighting individual matching pieces of this information, while leaving the relative importance of precision and recall as a user adjustable parameter. In the present situation, it is necessary for the underlying model to perform a functional mapping from a multidimensional space (name and article attributes) onto a single dimension that gives an appropriate ranking.

The Data, Model, and Methods

Constructing the Author-Article Database Table from Medline

The 2002 baseline release of Medline comes as XML files in utf-8 format, which includes most international characters. From these files a relational author database using the ASCII character set (English alphabet) was created. This database was actually split into several tables to optimize the various types of SQL queries that are necessary. For illustrative purposes, this section describes one table called AUTHOR_ARTICLES, which contains all 34,128,384 unique (author name, article) pairs defined on the 11 fields listed in the following paragraphs.

Preprocessing the title, affiliation, and MeSH fields include making all characters lowercase, removing nonalphabetic characters (except numbers), and removing single-character words. The sets of stopwords are used to reduce the number of arbitrary similarities between differing individuals. The stopwords for the affiliation and MeSH fields were selected from the topmost frequently used words. For example, "human" appears as a MeSH in approximately 65% of all Medline records, and the word "university" appears in 50% of the affiliations, and, as such, will probably not be of much value in discriminating authors. The three different sets of title stopwords will be used to analyze the effect of varying the extent of stoplisting.

Fields in database table AUTHOR_ARTICLES:

(primary key = *pmid*, *order*)
pmid = unique (PubMed) article identification number
order = position of author name on article
last = last name of author
init1 = first initial of author name
init2 = middle initial of author name
suff = suffix of author name
title = set of title words after preprocessing and removing title-stopwords
affl = set of affiliation words after preprocessing and removing *affiliation-stopwords*
jrnl = journal name
lang = language of article
mesh = set of MeSH words after preprocessing and removing *mesh-stopwords*

where

title-stopwords =

Small: PubMed's set of stopwords as of January 2002, which consists 365 commonly used English words, like "the" and "and."

Medium: The small stoplist together with the 1,029 words that appear in over 0.1% of the titles. About 400 of these frequent words were not included in this list because we judged that they may be important for establishing connections between two disparate disciplines.

Large: The small stoplist together with a list of the 8,207 words that are thought not to be important in establishing connections between two disparate disciplines. These words have been accumulated over the years as a part of the Arrowsmith Project (Swanson & Smalheiser, 1997). All words on the medium stoplist were also on the large stoplist.

affiliation-stopwords = small *title-stopwords* ∪ {university, medicine, medical, usa, hospital, school, institute, center, research, science, college, health, new, laboratory, division, national},

mesh-stopwords = {human, male, female, animal, adult, support non-u.s. gov't, middle age, aged, english abstract, support u.s. gov't p.h.s., case report, rats, comparative study, adolescence, child, mice, time factors, child preschool, pregnancy, united states, infant, molecular sequence data, kinetics, support u.s. gov't non-p.h.s., infant newborn}.

Of the records in the AUTHOR_ARTICLES table, 55% of the names have no middle initial, and 0.085% have no first initial. Many of the names without a first initial either contain errors or consist of people or entities that go by a single name, such as Sister Mary and the Duchess of York. There are 2,374,994 distinct names based on last name and first initial, of which 39% appear in only one article, and 20 names (including J Smith, J Lee, J Miller among others) appear in over 6,000 articles.

Defining the Similarity Profile

The objective is to define a similarity profile that captures multiple aspects of similarity between a pair of papers by the same author, to discriminate them from pairs of papers by different authors. Each element of the similarity profile is obtained from a field in Medline (such as title), where the shared terms (ignoring multiple occurrences) are enumerated. The simplicity of the profile makes it easy to calculate and interpret and allows for incorporating interactive (e.g., between Journal and MeSH) and nonlinear effects into the probabilistic model.

Suppose two distinct records obtained from the AUTHOR_ARTICLES table are given by:

$R_A = (pmid_A, order_A, last_A, init1_A, init2_A, suff_A, coauth_A, title_A, affl_A, jrnl_A, lang_A, mesh_A)$,

$R_B = (pmid_B, order_B, last_B, init1_B, init2_B, suff_B, coauth_B, title_B, affl_B, jrnl_B, lang_B, mesh_B)$,

where

$pmid_A \neq pmid_B$,

$coauth_A$ = set of distinct author names (defined by *last* and *init1*) on the article corresponding to $pmid_A$, less ($last_A$, $init1_A$), and

$coauth_B$ = set of distinct author names (defined by *last* and *init1*) on the article corresponding to $pmid_B$, less ($last_B$, $init1_B$).

Comparisons will only be performed for pairs of records that share an author last name and first initial. The similarity profile $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)$ is created by comparing the two records element wise as follows:

x_1 = 3 if $init2_A = init2_B$ and both are given (e.g., (A, A)), 2 if $init2_A = init2_B$ and both are not given (i.e., (∅, ∅)), 1 if $init2_A \neq init2_B$ and one is not given (e.g., (A, ∅)), and 0 if $init2_A \neq init2_B$ and both are given (e.g., (A, B)).

x_2 = 1 if $suff_A = suff_B$ and both are given (e.g., (Jr, Jr)), and 0 otherwise,

$x_3 = |title_A \cap title_B|$,

x_4 = 1 if $jrnl_A = jrnl_B$, and 0 otherwise,

$x_5 = |coauth_A \cap coauth_B|$,

$x_6 = |mesh_A \cap mesh_B|$,

x_7 = 3 if $lang_A = lang_B$ and non-English (e.g., (jpn, jpn)), 2 if $lang_A = lang_B$ and English (i.e., (eng, eng)), 1 if $lang_A \neq lang_B$ and one is English (e.g., (eng, jpn)), and 0 if $lang_A \neq lang_B$ and both are non-English (e.g., (jpn, fre)).

$x_8 = |affl_A \cap affl_B|$,

x_9 = 1 if $affl_A = \emptyset$ or $affl_B = \emptyset$, and 0 otherwise.

Here, the scores x_1 and x_2 describe the similarity of the two names, and will be referred to as *name similarity scores*, while the scores $x_3, x_4, \ldots, x_9$ describe the similarity of the two articles they appear on, and will be referred to as *article similarity scores*. Other attributes like author name position, and the presence of non-ASCII characters in the last name, were excluded because they were found not to have a significant discriminatory power in our preliminary studies.

Outline of the Probabilistic Matching Model

It is our hypothesis that different articles authored by the same individual will share similarities in one or more aspects of the Medline records, more so than articles authored by different individuals.

To formalize this idea mathematically let $\Pr\{\boldsymbol{x}|M\}$ and $\Pr\{\boldsymbol{x}|N\}$ denote the probability of observing similarity profile $\boldsymbol{x}$ given that one knows the two author names being compared refer to the same individual, and different individuals, respectively. Here, M denotes an "author match," and N denotes an "author nonmatch." It is then intuitive that a high value of the ratio $r(\boldsymbol{x}) = \Pr\{\boldsymbol{x}|M\}/\Pr\{\boldsymbol{x}|N\}$ will strengthen the case for saying that the two names refer to the same person. For example, suppose that a particular similarity profile $\boldsymbol{x}$ occurs frequently when comparing same individuals (e.g., $\Pr\{\boldsymbol{x}|M\} = 10\%$), and rarely occurs when comparing different people,

(e.g., $\Pr\{\boldsymbol{x}|N\} = 0.01\%$). In this case $r(\boldsymbol{x})$ is large ($10/0.01 = 1{,}000$), which suggests that when $\boldsymbol{x}$ is observed the two names are most likely referring to the same individual.

The actual probability of the two names referring to the same individual depends on the population of articles that are being compared. For example, when comparing a population of two individuals who have each published five papers, the ratio $r(\boldsymbol{x})$ should be weighted much higher than when comparing a population of 10 individuals with five papers each. To formalize this weighting idea let $\Pr\{M|\boldsymbol{x}\}$ denote the probability that two author names being compared refer to the same individual given that the similarity profile $\boldsymbol{x}$ is observed. Bayes' theorem is then manipulated as follows:

$$\begin{aligned}\Pr\{M|\boldsymbol{x}\} &= \Pr\{\boldsymbol{x}|M\}\Pr\{M\}/\Pr\{\boldsymbol{x}\} \\ &= \frac{\Pr\{\boldsymbol{x}|M\}\Pr\{M\}}{\Pr\{\boldsymbol{x}|M\}\Pr\{M\} + \Pr\{\boldsymbol{x}|N\}\Pr\{N\}} \\ &= \frac{1}{1 + \dfrac{\Pr\{\boldsymbol{x}|N\}\Pr\{N\}}{\Pr\{\boldsymbol{x}|M\}\Pr\{M\}}} \\ &= \frac{1}{1 + \dfrac{1 - \Pr\{M\}}{\Pr\{M\}r(\boldsymbol{x})}}\end{aligned}$$

The resulting expression shows how $\Pr\{M|\boldsymbol{x}\}$ only depends on $r(\boldsymbol{x})$ and $\Pr\{M\}$. The greater the values for $\Pr\{M\}$ and $r(\boldsymbol{x})$ become, the greater $\Pr\{M|\boldsymbol{x}\}$ becomes. Here, the parameter $\Pr\{M\}$ denotes the overall probability of a match for a given name within a population of articles. In other words, it measures the probability that the two names refer to the same individual not knowing anything about the similarity profile $\boldsymbol{x}$. For example, when comparing a population of two individuals each with five papers, there are 45 distinct pairs of papers, of which 20 are matches, leading to $\Pr\{M\} = 4/9$. Similarly, when comparing a population of 10 individuals each with five papers, there are 1,225 distinct pairs of papers, of which 100 are matches, leading to $\Pr\{M\} = 4/49$. One will, of course, not know how many individuals there are with a particular name or how many papers each has written. $\Pr\{M\}$ can be estimated in a number of different ways described in Step 3 of the section called A Stepwise Procedure for Fitting the Probability Model. The reason for this indirect model is because of the fact that $\Pr\{M|\boldsymbol{x}\}$ cannot be accurately and unbiasedly estimated in an easy and direct fashion.

A Stepwise Procedure for Fitting the Probability Model

We have developed a method for generating training data in an unbiased manner, and estimating $r(\boldsymbol{x})$ for all possible profiles $\boldsymbol{x}$ across all author names in Medline, and $\Pr\{M\}$ for each name, based on the following criterion:

The monotonicity criterion: $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$ implies that $\Pr\{M|\boldsymbol{x}\} \leq \Pr\{M|\boldsymbol{y}\}\ \forall\ \boldsymbol{x}, \boldsymbol{y}$ where $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$ holds if and only if $x_i \leq y_i \forall\ i = 1, 2, \ldots, 9$.

The monotonicity criterion can be interpreted as each similarity score x_i having a nonnegative effect on the probability of a match. As an example, consider a pair of papers, both with C Friedman as an author. With all other scores fixed, the more coauthor names the two articles have in common, the more likely it is that the two C Friedman's refer to the same C Friedman. Because $\Pr\{M\}$ is constant for a given name, the function $r(\boldsymbol{x})$ will also satisfy the monotonicity criterion. A nice property of this criterion is that it is not overly restrictive, and provides a manner in which to smooth, interpolate and extrapolate our estimates for $r(\boldsymbol{x})$ to profiles that are rarely (or never) observed in our training sets. The interested reader is referred to Torvik and Triantaphyllou (2002, 2003), and Robertson, Wright, and Dykstra (1988) for further details on the monotonicity property.

The goal is to estimate $\Pr\{M|\boldsymbol{x}\}$, the probability that a pair of names on two different articles refer to the same individual for any possible $\boldsymbol{x}$. To that end, estimates of $r(\boldsymbol{x}) = \Pr\{\boldsymbol{x}|M\}/\Pr\{\boldsymbol{x}|N\}$ are computed and stored for all possible $\boldsymbol{x}$ averaged across all names in Medline. When a pair of names are to be compared, the distribution $\Pr\{\boldsymbol{x}\}$ is computed for that name, and compared to $\Pr\{\boldsymbol{x}|M\}$ and $\Pr\{\boldsymbol{x}|N\}$ for a few representative profiles $\boldsymbol{x}$ to initially estimate $\Pr\{M\}$. Next, the similarity profile $\boldsymbol{x}$ is computed for the pair to be compared. Then, $r(\boldsymbol{x})$ is looked up and weighted by the estimate of $\Pr\{M\}$ to compute initial estimates of $\Pr\{M|\boldsymbol{x}\}$. These estimates are then improved by updating the a priori match probability and imposing "geometric" three-way constraints on the pairwise probabilities. The details of this procedure are described next.

Step 1: Generating reference sets. To estimate $r(\boldsymbol{x})$ one needs to generate training data measured on all nine similarity scores, where one can label, with high confidence, the individual pairs as matches or nonmatches.

First, to generate matches, all author names that have first and middle initials, *and* suffixes were selected. Then, the article similarity scores were computed for all pairs that matched on last name, first and middle initials, *and* suffixes. In most cases, these pairs represent the same person (average r-value $>3{,}500$). To reduce the number of nonmatches within this set, names that have more than one middle initial recorded with suffixes (e.g., JA Smith Jr and JB Smith Jr) were excluded. This resulted in about 27,000 distinct names, and 4.3 million distinct pairs of "matches" defined on the article similarity scores, which is referred to as the *article attribute match set*. Second, to generate nonmatches, 30,000 records from the AUTHOR_ARTICLES table were randomly selected without replacement. Then, the article similarity scores were computed for all pairs that differed on last names and *pmid*. This resulted in a set of 450 million pairs, which probably excludes matches altogether and is referred as the *article attribute nonmatch set*.

To generate an analogous pair of sets of matches and nonmatches for the name similarity scores, 1,000 names were

TABLE 1. Definitions of the match sets and the nonmatch sets.

	Match sets	Nonmatch sets
Yields name similarity profiles (and language similarity score in the match set)	54,000 pairs of articles in which the first author names match on last name and first initial, and share one or more coauthor names, two or more affiliation words, and two or more MeSH.	9.2 million pairs of articles in which the first author names match on last name, first initial, language, and have nothing else in common.
Yields article similarity profiles	4.3 million pairs of articles in which the two author names match on last name, initials and suffix.	450 million pairs of articles in which the two author names do not match on last name.

randomly selected without replacement, and within each name, all pairwise comparisons were computed. This set is referred to as the *mixed set* as it contains a significant proportion of both matches and nonmatches. From the mixed set we extracted the 54,000 pairs that were both first authors and had one or more coauthor names in common, two or more affiliation words in common, and two or more MeSH terms in common. In most cases, these pairs represent the same person (average r-value $>$ 5,000). This set is referred to as the *name attribute match set*. Similarly, we extracted the 9.2 million pairs of articles that shared first author name but did not have anything else in common other than language, nor were they missing the affiliations. These pairs have very high probability of being authored by different individuals (average r-value $<$ 0.01). This set is referred to as the *name attribute nonmatch set*.

Because suffixes are much more common in English names than non-English, the article attribute match set contained an unusually high number of English articles, and it is therefore biased for the language similarity score. To overcome this bias, a separate set of matches were generated in a similar fashion as the name attribute match set, yielding a *language attribute match set*. Table 1 summarizes the manner in which all the training sets were generated.

Step 2: Estimating r(x) for all possible similarity profiles x based on the monotonicity criterion. The goal is to create a lookup table for the estimated ratio $r(\boldsymbol{x}) = \Pr\{\boldsymbol{x}|M\}/\Pr\{\boldsymbol{x}|N\}$ for all possible $\boldsymbol{x}$. Because the individual reference sets for the article similarity scores, language similarity scores, and name similarity scores were generated independently, their individual ratio functions $r_1(x_1)$, $r_2(x_2)$, $r_7(x_7)$, and $r_a(\boldsymbol{x_a})$, respectively, are first estimated. Together these functions can be used to estimate the overall ratio by

$$\hat{r}(x_1, x_2, x_3, \ldots, x_9) = \bar{r}_1(x_1)\bar{r}_2(x_2)\bar{r}_7(x_7)\hat{r}_a(x_3, x_4, x_5, x_6, x_8, x_9)$$

This is based on the criterion that x_1, x_2, x_7, and $\boldsymbol{x_a}$ are all mutually independent in their effect on the probability of match. The validity of this criterion is assessed in the section called Evaluation of the Fitted Model.

Each x_i value for the name and language scores (i.e., $x_1 = 0, 1, 2, 3$; $x_2 = 0, 1$; $x_7 = 0, 1, 2, 3$) was observed a sufficient number of times in both the match and nonmatch sets to provide accurate estimates. The mean estimate satisfies the monotonicity criterion and is given by

$$\bar{r}_i(x_i) = \frac{m(x_i)/\sum_{x_i} m(x_i)}{n(x_i)/\sum_{x_i} n(x_i)}, \quad \text{for} \quad i = 1, 2, 7,$$

where

$m(x_i)$ = number of times x_i was observed in the match set, and
$n(x_i)$ = number of times x_i was observed in the nonmatch set.

Estimating $r_a(\boldsymbol{x_a})$ for the article similarity profiles $\boldsymbol{x_a} = (x_3, x_4, x_5, x_6, x_8, x_9)$ requires more consideration and will be described next. The article attribute reference sets allows us to get a preliminary estimate for $r_a(\boldsymbol{x_a})$ by

$$\bar{r}_a(\boldsymbol{x_a}) = \frac{m(\boldsymbol{x_a})/\sum_{\boldsymbol{x_a}} m(\boldsymbol{x_a})}{n(\boldsymbol{x_a})/\sum_{\boldsymbol{x_a}} n(\boldsymbol{x_a})}, \quad \forall\, \boldsymbol{x_a} \in \boldsymbol{X},$$

where

$\boldsymbol{X} = \{\boldsymbol{x_a}: m(\boldsymbol{x_a}) \text{ and } n(\boldsymbol{x_a}) > 0\}$, the 1,081 profiles observed at least once in both the match and nonmatch sets,
$m(\boldsymbol{x_a})$ = number of times profile $\boldsymbol{x_a}$ was observed in the match set, and
$n(\boldsymbol{x_a})$ = number of times profile $\boldsymbol{x_a}$ was observed in the nonmatch set.

However, the number of observations $m(\boldsymbol{x_a})$ and $n(\boldsymbol{x_a})$ tend to decrease rapidly as $\boldsymbol{x_a}$ increases and the numbers vary because of sampling. Therefore, the estimates $\bar{r}_a(\boldsymbol{x_a})$ become inaccurate as $\boldsymbol{x_a}$ becomes greater. The next three steps, 2a, 2b, and 2c, alleviate this problem via smoothing, interpolation, and extrapolation, respectively.

Step 2a: Smoothing $\hat{r}_a(x_a)$ for each observed x_a via quadratic programming. To improve the accuracy our estimates for $r_a(\boldsymbol{x_a})$, the monotonicity property is enforced by solving the following linearly constrained least squares problem:

$$\text{minimize} \ \sum_{\boldsymbol{x_a} \in \boldsymbol{X}} w(\boldsymbol{x_a})(\bar{r}_a(\boldsymbol{x_a}) - \hat{r}_a(\boldsymbol{x_a}))^2$$
$$\text{subject to } \hat{r}_a(\boldsymbol{x_a}) \leq \hat{r}_a(\boldsymbol{y_a}) \,\forall\, (\boldsymbol{x_a}, \boldsymbol{y_a}): \boldsymbol{x_a} \preccurlyeq \boldsymbol{y_a}$$

where,

$$w(\boldsymbol{x_a}) = m(\boldsymbol{x_a}) + n(\boldsymbol{x_a}),$$
$$\boldsymbol{X} = \{\boldsymbol{x_a}: m(\boldsymbol{x_a}) > 0 \text{ and } n(\boldsymbol{x_a}) > 0\}, \quad \text{and}$$
$$\bar{r}_a(\boldsymbol{x_a}) = \frac{m(\boldsymbol{x_a})/\sum_{x_a} m(\boldsymbol{x_a})}{n(\boldsymbol{x_a})/\sum_{x_a} n(\boldsymbol{x_a})}.$$

The objective of this optimization problem is to find an estimate $\hat{r}_a(\boldsymbol{x_a})$ whose squared distance to $\bar{r}_a(\boldsymbol{x_a})$ is as small as possible, weighted by the confidence in the estimate, while satisfying the monotonicity constraints. The confidence is here measured by $m(\boldsymbol{x_a}) + n(\boldsymbol{x_a})$, the number of times the profile $\boldsymbol{x_a}$ was observed in both the match set and nonmatch set.

This optimization problem belongs to the class of problems known as quadratic programs, in which the objective function is quadratic and the constraints are linear in the unknowns. Quadratic programs can in general be solved by several different optimization algorithms. The Active Set algorithm implemented as a part of the Matlab Optimization Toolbox 2.1 (The Mathworks, Inc., Natick, MA) was used here.

Step 2b: Interpolating $\hat{r}_a(x_a)$ to unobserved x_a that precede the upper profiles. All the profiles in $\boldsymbol{X}$ precede one or both of the profiles (9, 1, 7, 9, 12, 0) and (9, 1, 7, 9, 0, 1), which are referred to as the *upper profiles*. The estimate $\hat{r}_a(\boldsymbol{x_a})$ for each observed profile $\boldsymbol{x_a}$ that precedes an upper profile is interpolated from a preceding profile $\boldsymbol{p_a}$ and a succeeding profile $\boldsymbol{s_a}$:

$$\hat{r}_a(\boldsymbol{x_a}) = \frac{\hat{r}_a(\boldsymbol{p_a}) + \hat{r}_a(\boldsymbol{s_a})}{2} \quad \forall\, \boldsymbol{x_a} \notin \boldsymbol{X}, \quad \text{and}$$
$$\boldsymbol{x_a} \preccurlyeq (9, 1, 7, 9, 12, 0) \quad \text{or} \quad \boldsymbol{x_a} \preccurlyeq (9, 1, 7, 9, 0, 1),$$

where

$\boldsymbol{p_a}$ = the preceding profile (i.e., $\boldsymbol{p_a} \prec \boldsymbol{x_a}$) in $\boldsymbol{X}$ with the maximum ratio $\hat{r}_a(\boldsymbol{p_a})$, and
$\boldsymbol{s_a}$ = the succeeding profile (i.e., $\boldsymbol{x_a} \prec \boldsymbol{s_a}$) in $\boldsymbol{X}$ with the minimum ratio $\hat{r}_a(\boldsymbol{s_a})$.

Note that this will result in interpolated estimates that satisfy the monotonicity property $\hat{r}_a(\boldsymbol{x_a}) \leq \hat{r}_a(\boldsymbol{y_a})$ for any pair $(\boldsymbol{x_a}, \boldsymbol{y_a})$: $\boldsymbol{x_a} \preccurlyeq \boldsymbol{y_a}$. This will also hold true even when both $\boldsymbol{x_a}$ and $\boldsymbol{y_a}$ are unobserved. The $\hat{r}_a(\boldsymbol{x_a})$ values were interpolated for total of 21,319 unobserved profiles, leading to a total of 22,400 stored profiles and their estimated $\hat{r}_a(\boldsymbol{x_a})$ values. This procedure provides the $\hat{r}_a(\boldsymbol{x_a})$ values for all profiles $\boldsymbol{x_a}$ preceding the two upper profiles.

Step 2c: Extrapolating $\hat{r}_a(x_a)$ to unobserved x_a that succeed the upper profiles. To get estimates for $\hat{r}_a(\boldsymbol{x_a})$ for all other possible $\boldsymbol{x_a}$, the estimate for $\hat{r}_a(\boldsymbol{x_a})$ is extrapolated from a preceding profile $\boldsymbol{p_a}$ as follows:

$$\hat{r}_a(\boldsymbol{x_a}) = max_{\boldsymbol{p_a} \in X}\{\hat{r}_a(\boldsymbol{p_a}): \boldsymbol{p_a} \prec \boldsymbol{x_a}\}.$$

Here, $\boldsymbol{p_a}$ denotes the preceding profile (i.e., $\boldsymbol{p_a} \prec \boldsymbol{x_a}$) with the maximum ratio $\hat{r}_a(\boldsymbol{p_a})$. In practice, this step is quite simple, because a profile $\boldsymbol{x_a}$ simply needs to be converted to:

$$\boldsymbol{x_{new}} = (min\{9, x_3\}, min\{1, x_4\}, min\{7, x_5\}, min\{9, x_6\}, min\{12, x_8\}, min\{1, x_9\}),$$

and then $\hat{r}_a(\boldsymbol{x_{new}})$ is simply looked up from one of the 22,400 stored profiles:

$$\hat{r}_a(\boldsymbol{x_a}) = \hat{r}_a(\boldsymbol{x_{new}}).$$

Note that $\hat{r}_a(\boldsymbol{x_{new}})$ is probably an underestimate for $r_a(\boldsymbol{x_a})$. This is of little practical significance because extrapolation is rarely needed. Furthermore, the extrapolated values tend to be very large (order of 10^4) and assigning larger values will not be of practical use.

Step 3: Estimating Pr{M} for a specific author name: based on name frequency, zero profile frequency, or directly from predicted proportion of matches. The a priori probability of match $\Pr\{M\}$ averaged across all the names in Medline is about $1/11$.

However, this estimate is not accurate for any given name, because the number of individuals and the number of papers per individual vary dramatically across names. It is therefore important that the prior is estimated for each name individually. There are several possible ways to accomplish this:

The **first approach** is to compute the proportion of "zero" profiles observed by performing all of the pairwise comparisons for a given name, and then see how close this proportion is to the match and nonmatch sets. Here, we refer to the vectors $\boldsymbol{x_a} = (0, 0, 0, 0, 0, 0)$ or $(0, 0, 0, 0, 0, 1)$ as the zero profile, which accounts for 40.4% of match set (i.e., $\Pr\{\boldsymbol{x_a}|M\} = 0.404$) and 86.6% of nonmatch set (i.e., $\Pr\{\boldsymbol{x_a}|N\} = 0.866$). That is, when the population of articles compared are all authored by different individuals, $\Pr\{\boldsymbol{x_a}\}$ will tend to lie around 0.866, and if they are all authored by the same person, $\Pr\{\boldsymbol{x_a}\}$ will tend to lie around 0.404. Therefore, given an estimate for $\Pr\{\boldsymbol{x_a}\}$, then $\Pr\{M\}$ can be estimated as follows:

$$\Pr\{M\} = \frac{\Pr\{\boldsymbol{x_a}\} - \Pr\{\boldsymbol{x_a}|N\}}{\Pr\{\boldsymbol{x_a}|M\} - \Pr\{\boldsymbol{x_a}|N\}} = \frac{\Pr\{\boldsymbol{x_a}\} - 0.866}{0.404 - 0.866},$$

where

$$\boldsymbol{x_a} = (0, 0, 0, 0, 0, 0) \quad \text{or} \quad (0, 0, 0, 0, 0, 1).$$

This estimate measures the degree of similarity of the articles themselves and, as such, takes into account the variability resulting from the number of individuals as well as the number of articles per individual. However, this estimate will not be accurate when $\Pr\{\boldsymbol{x_a}\}$ lies within the interval (0.404, 0.866) and very close to one of the endpoints. In this case, one can use an estimate based on the average prior (1) for a given name frequency (see later) or (2) across all names

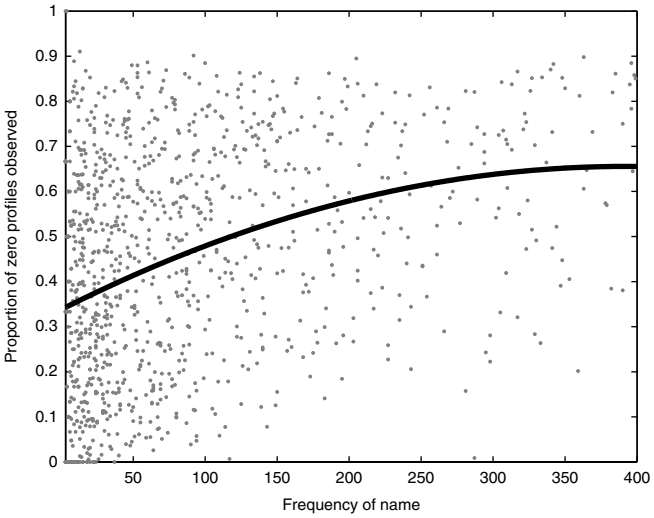

FIG. 1. The proportion zero profiles observed for each name frequency varies greatly across names.

in Medline 0.11. For example, if the observed $\Pr\{\boldsymbol{x_a}\}$ is close to or higher than 0.866, it is likely that all the papers are authored by the same person, and a value for $\Pr\{M\}$ much higher than $1/11$, such as 0.9, should be used. Similarly, when $\Pr\{\boldsymbol{x_a}\}$ close to or lower than 0.404, a value much lower than $1/11$, such as 0.01, should be used.

The **second approach** uses only the frequency of the name itself to predict the prior. To do this one first needs to define the prior probability of match (averaged across all names in Medline) as a function of the frequency of the name. Figure 1 shows the proportion of zero profiles in set of names randomly samples across set of name frequencies ranging from 3 to 400. The best fitting regression curve (shown as a solid line in the figure), given by

$$\Pr\{\boldsymbol{x_a}\}(frq) = 0.338 + 0.00162\,frq - 0.00000206\,frq^2,$$

can then be used to estimate the prior by

$$\Pr\{M\} = \frac{\Pr\{\boldsymbol{x_a}\} - 0.866}{0.404 - 0.866}.$$

From the figure it is clear that for any given name frequency the points vary dramatically, implying that this estimate comes with a high variance. This is expected because some people have published many articles (over 1,000), whereas others have published very few.

The **third approach** is based on the predicted proportion of matches. Once the set of pairwise probabilities have been computed, one can update the estimate simply by the number of pairs labeled matches (with a match probability > 0.5) divided by the total number of pairs. This estimate can be used at any stage when a new set of match probabilites are generated.

Step 4: Updating pairwise probabilities based on three-way "geometric" probability constraints. Suppose we are given three papers A, B, and C where the pairwise model estimates the match probabilities to be $\Pr\{A \text{ and } B \text{ match}\} = p_1$, $\Pr\{A \text{ and } C \text{ match}\} = p_2$, and $\Pr\{B \text{ and } C \text{ match}\} = p_3$. Let $p_{(i)}$ denote the ith largest value (for $i = 1, 2, 3$) of three probabilities $\{p_1, p_2, p_3\}$. Because two out of three pairwise matches is not possible and because of the laws of probability the following constraint has to be satisfied: $p_{(3)} \geq p_{(2)} + p_{(1)} - 1$. Yet empirically we sometimes observe cases where this constraint is violated: If two papers that both have a high probability of match (e.g., 0.9 and 0.8) to a third paper, that nevertheless have a low estimated probability of match (e.g., 0.2) to each other based solely on pairwise comparisons. In the case when the constraint is violated, the two larger probabilities should be made smaller, while the smaller probability should be made larger. This can be accomplished by solving the following least-squares problem:

$$\text{minimize } w(\hat{p}_{(1)} - p_{(1)})^2 + w(\hat{p}_{(2)} - p_{(2)})^2 + (\hat{p}_{(3)} - p_{(3)})^2 \quad \text{subject to } \hat{p}_{(3)} \geq \hat{p}_{(2)} + \hat{p}_{(1)} - 1$$

where $\hat{p}_{(i)}$ denotes our adjusted estimate for $p_{(i)}$. When the constraint is not violated, $\hat{p}_{(i)}$ stays the same at $p_{(i)}$, otherwise they are updated by substituting $\hat{p}_{(3)} = \hat{p}_{(2)} + \hat{p}_{(1)} - 1$ into the objective function and setting its gradient to 0, resulting in the following updates:

$$\hat{p}_{(1)} = [(1 + w)p_{(1)} - p_{(2)} + p_{(3)} + 1]/(2 + w)$$
$$\hat{p}_{(2)} = [(1 + w)p_{(2)} - p_{(1)} + p_{(3)} + 1]/(2 + w)$$
$$\hat{p}_{(3)} = [wp_{(1)} + wp_{(2)} + 2p_{(3)} - w]/(2 + w)$$

Here, the weighting factor w, allows for adjusting the magnitude of change in the lowest probability relative to the two larger probabilities. The correction is weighted more heavily on the low value, reflecting the fact that comparisons with high probability values also have higher confidence than those with low values. The greater value assigned to w, the greater the lower probability has to be increased relative to the two larger probabilities. For example, $\{0.8, 0.2, 0.9\}$, then $p_{(1)} = 0.9, p_{(2)} = 0.8, p_{(3)} = 0.2$, and the constraint is violated because $0.9 + 0.8 - 1 = 0.7 > 0.2$. Therefore, the probabilities are updated to

0.733, 0.633, 0.367, when $w = 1$
0.775, 0.675, 0.450, when $w = 2$,
0.829, 0.729, 0.557, when $w = 5$, and
0.858, 0.758, 0.617, when $w = 10$.

In general, the high probability estimates are more accurate than the lower ones, because we know that many pairs of papers by the same individuals have very little in common, while very few papers by different people have very much in common. Therefore, the weight factor w should be higher than 1. The effect of the weight factor is analyzed in the section titled "What is the optimal geometric constraint weight factor?".

Summary: The Complete Matching Model

The complete model fitted to the training data generated from Medline can now be stated mathematically as follows:

$$\Pr\{M|\boldsymbol{x}\} = \frac{1}{1 + \dfrac{1 - \Pr\{M\}}{\Pr\{M\} r(\boldsymbol{x})}},$$

where

$$\hat{r}(x_1, x_2, \ldots, x_9) = \bar{r}_1(x_1)\, \bar{r}_2(x_2)\, \bar{r}_7(x_7)\, \hat{r}_a(x_3, x_4, x_5, x_6, x_8, x_9), \quad \text{and}$$

$\Pr\{M\}$ is estimated for a given name.

The values for $\bar{r}_1(x_1)$, $\bar{r}_2(x_2)$, and $\bar{r}_7(x_7)$, and a sample of $\hat{r}_a(x_3, x_4, x_5, x_6, x_8, x_9)$ as generated from the match and nonmatch are presented in the next two sections. $\Pr\{M\}$ is initially estimated either based on the zero profile frequency by

$$\Pr\{M\} = \frac{\Pr\{\boldsymbol{x_a}\} - 0.866}{0.404 - 0.866}, \quad \text{for} \quad \boldsymbol{x_a} = (0, 0, 0, 0, 0, 0) \text{ or } (0, 0, 0, 0, 0, 1).$$

After all the pairwise match probabilities are computed, the prior is updated by the predicted proportion of matches: number of pairs with $\Pr\{M|\boldsymbol{x}\} > 0.5$ divided by the total number of pairs. Then a final adjustment of the pairwise match probabilities are imposed by the geometric three-way constraints.

The estimated values for $\Pr\{M|\boldsymbol{x}\}$ can be used to rank all the articles relative to one of the articles by decreasing probability of match. One can also label a pair a match if $\Pr\{M|\boldsymbol{x}\}$ is greater than some threshold p. If the penalty for labeling a match a nonmatch is the same as the penalty for labeling a nonmatch a match, then using a cut-off of $p = 0.5$ is natural. This cutoff would also maximize the expected accuracy across Medline as a whole. Alternatively, one can use an intermediate category of so-called unassignables that have a match probability close to 0.5 (e.g., $0.5 - \delta < \Pr\{M|\boldsymbol{x}\} < \delta + 0.5$).

Results

Distributions of the Individual Similarity Scores Within the Reference Sets

The section on Defining the Similarity Profile defined nine similarity scores that were thought to be potentially useful in discriminating between matching and nonmatching author names. From the generated reference sets, their individual distributions were computed. Table 2 shows how the individual similarity scores distribute within the match and nonmatch sets. Each row gives the distribution of a similarity score x_i within the match set and the nonmatch set, and gives the associated r-values estimated by $\bar{r}_i(x_i)$ for $x_i = 0, 1, \ldots, 5$. For example, the row labeled $\Pr\{x_3|M\}$ *title w/ small stoplist*, gives the proportions 0.6485, 0.2114, 0.08181, and so forth, of the pairs in the match set that have 0, 1, 2, and so forth, title words in common after using PubMed's stoplist of 365 words. It is interesting to notice that nonmatches rarely have something in common other than language.

The ratio $\bar{r}_i(x_i)$ quantifies how each similarity score x_i affects the match probability when one knows nothing about the other scores. Figure 2 shows a visual of these ratios, excluding the title score for the medium and large stoplists, and the *affiliation(s) not given* score (x_9). Notice that the y-axis is given on an exponential scale. It is therefore clear that as the value of each x_i increases, the $r_i(x_i)$, and consequently $\Pr\{M|x_i\}$ increases exponentially.

Figure 2 also shows how the individual similarity scores fare against each other. The number of common coauthor names seems to be the most important, followed by journal name match, and then middle name initial match. It is interesting that coauthor is the most important single parameter in the model, despite the fact that we are performing a match on their *names* without disambiguating whether they correspond to the same individual or not; this potential ambiguity does not seem to be limiting in practice. Although suffix matches are important they are rare and, as such, less useful. The number of common affiliation words, title words and MeSH are tied in fourth place. The flipside is that MeSH, title, and affiliation words are more often in common than coauthor names or journal, and therefore all the variables together will provide a more powerful similarity measure.

Interactions Among Article Similarity Profiles

Figure 3 shows the estimated values for the ratio $r_a(\boldsymbol{x_a})$ on a sample of the most frequently observed article similarity profiles on the four lower-most levels of the partially ordered set (or poset for short). The profiles not shown in the figure succeed these profiles, and, as such, will have greater $\hat{r}_a(\boldsymbol{x_a})$ values. Each vertex in the poset gives the profile in the form $\boldsymbol{x_a} = (x_3, x_4, x_5, x_6, x_8, x_9)$ and the associated ratio $\hat{r}_a(\boldsymbol{x_a})$ is shown below the vector. The lines connecting the vertices represent the precedence relation $\preccurlyeq$. For example, the profile labeled (010000) corresponds to a match on journal only and was estimated to have an r-value of 11.66.

The Case of C Friedman

Let us illustrate how the model works for a typical example of an author disambiguation task in Medline. Although the 2002 baseline release of Medline lists a total of 401 papers with the name C Friedman (with middle initial given or lacking), only the 248 articles that give affiliations are considered here, to allow for accurate manual disambiguation. The C Friedman papers were manually disambiguated and found to comprise 21 distinct individuals (14 with two or more papers: Charles P (59), Craig D (47), Chad I (32),

TABLE 2. Distributions of the match, and nonmatch sets over the individual similarity scores.

Number of attributes in common		0	1	2	3	4	5
$\Pr\{x_1\|M\}$	*init2*	0.00556	0.03004	0.5102	0.4542		
$\Pr\{x_1\|N\}$		0.4143	0.3231	0.2313	0.03130		
$\bar{r}_1(x_1)$		0.01343	0.09295	2.2058	14.5140		
$\Pr\{x_2\|M\}$	*suffix*	0.9978	0.00204				
$\Pr\{x_2\|N\}$		0.9999	$8.4*10^{-6}$				
$\bar{r}_2(x_2)$		0.9978	242.16				
$\Pr\{x_3\|M\}$	*title w/*	0.6485	0.2114	0.08181	0.03290	0.014185	0.005847
$\Pr\{x_3\|N\}$	*small*	0.9019	0.08986	0.007290	0.0008554	0.0001062	0.000017223
$\bar{r}_3(x_3)$	*stoplist*	0.7191	2.353	11.22	38.46	133.5	339.5
$\Pr\{x_3\|M\}$	*title w/*	0.8006	0.1438	0.04243	0.009326	0.002702	0.0007640
$\Pr\{x_3\|N\}$	*medium*	0.9930	0.006691	0.0003200	0.00001665	0.000001197	0.0000001023
$\bar{r}_3(x_3)$	*stoplist*	0.8063	21.49	132.6	560.0	2,258	7,466
$\Pr\{x_3\|M\}$	*title w/*	0.8147	0.1391	0.03709	0.007060	0.001516	0.0003961
$\Pr\{x_3\|N\}$	*large*	0.9935	0.006178	0.0002927	0.00001360	0.0000007408	0.00000004672
$\bar{r}_3(x_3)$	*stoplist*	0.8200	22.51	126.7	519.2	2,046	8,478
$\Pr\{x_4\|M\}$	*jrnl*	0.8875	0.1125				
$\Pr\{x_4\|N\}$		0.9989	0.001087				
$\bar{r}_4(x_4)$		0.8885	103.5				
$\Pr\{x_5\|M\}$	*coauth*	0.8421	0.1137	0.02859	0.009615	0.003585	0.001425
$\Pr\{x_5\|N\}$		0.9998	0.0002313	0.000005862	0.000001126	0.0000003671	0.0000001313
$\bar{r}_5(x_5)$		0.8423	491.7	4,877	8,542	9,766	10,855
$\Pr\{x_6\|M\}$	*mesh*	0.8077	0.1191	0.04334	0.01729	0.007159	0.003035
$\Pr\{x_6\|N\}$		0.9685	0.02830	0.002690	0.0004202	0.00008356	0.00001796
$\bar{r}_6(x_6)$		0.8340	4.207	16.11	41.13	85.68	169.0
$\Pr\{x_7\|M\}$	*lang*	0.000075	0.02872	0.9527	0.01849		
$\Pr\{x_7\|N\}$		0.03797	0.3302	0.6263	0.005515		
$\bar{r}_7(x_7)$		0.001974	0.08700	1.5211	3.3532		
$\Pr\{x_8\|M\}$	*affl*	0.8200	0.02888	0.03490	0.03725	0.03456	0.01968
$\Pr\{x_8\|N\}$		0.9853	0.01271	0.001565	0.0003449	0.00009476	0.00002390
$\bar{r}_8(x_8)$		0.8323	2.272	22.30	108.0	364.7	823.3
$\Pr\{x_9\|M\}$	*affl(s)*	0.2561	0.7439				
$\Pr\{x_9\|N\}$	*not*	0.2815	0.7185				
$\bar{r}_9(x_9)$	*given*	0.9095	1.0354				

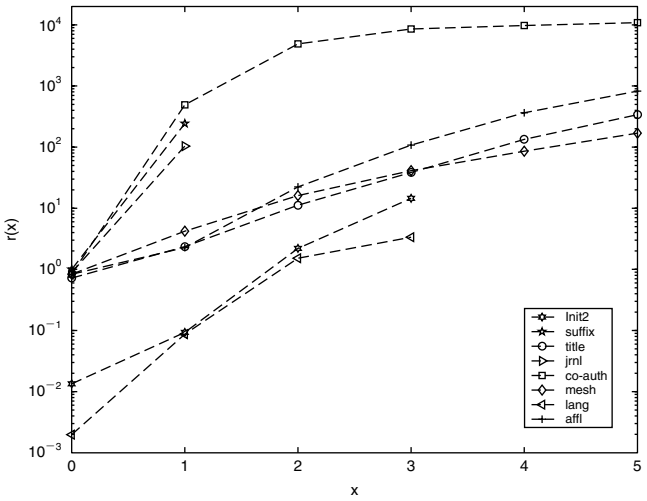

FIG. 2. Distributions of the *r*-values for the individual similarity scores.

Carol (30), Cynthia L (20), Carl J (13), Candace (12), Cindy R (7), Carolyn S (5), Charles A (6), Carola P (4), Clive S (2), CH (2), and C_1 (2), and 7 people with only one paper each: Catherine R, Constance, Colleen B, C_2, CA_1, CA_2, and CT). Some of the first names were found by searching other bibliographic databases (EBSCO, Ovid, and ScienceDirect), and searching for lists of publications on personal or institutional home pages using Google. As a result, partial lists of papers for Charles P (personal CV online), Craig D (coauthor's list of publications), Chad I (personal home page), Carol (personal home page), Cynthia L (laboratory home page), Candace (Community of Science® profile), and Carolyn S (personal home page) were identified. If nothing was found on the individual they were judged to be matches or nonmatches based on all the information available in the Medline record, and considering the groups of the other papers known to be by the same person.

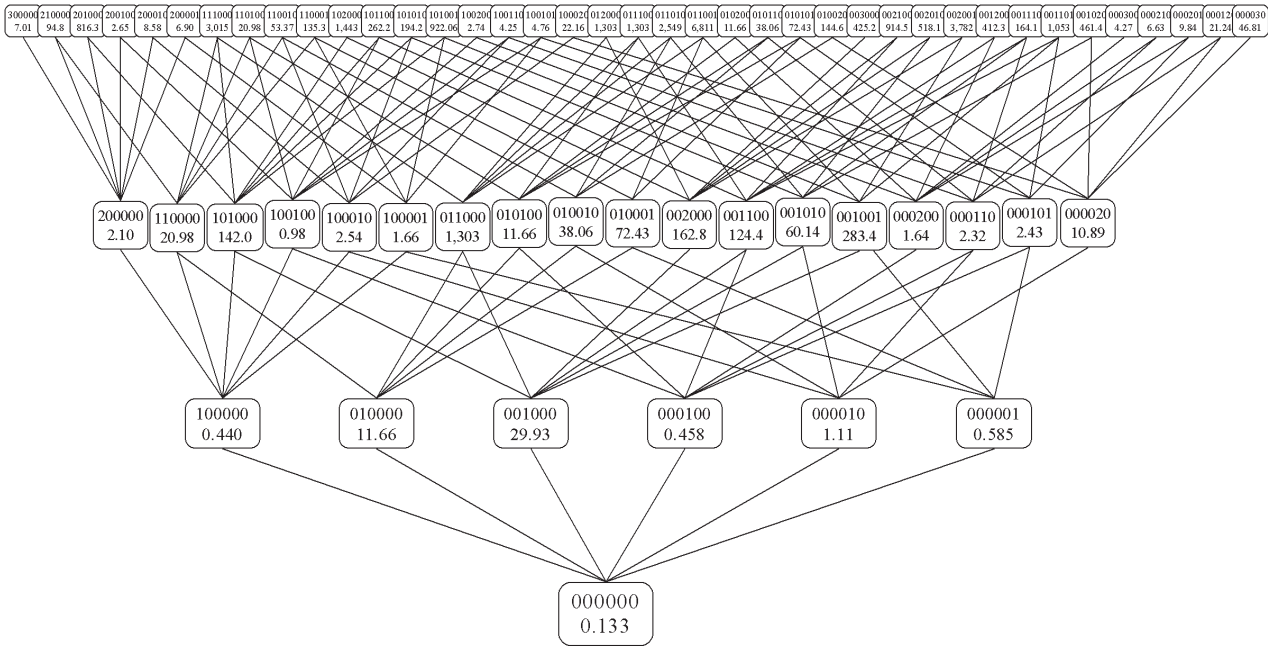

FIG. 3. The estimated *r*-values over a subset of the most frequently observed article similarity profiles in the reference sets.

Carol Friedman, who is currently affiliated with both the Medical Informatics Department at Columbia University and the Computer Science Department at Queens College CUNY, has published numerous articles in the area of medical informatics that are indexed in Medline. On the other hand, Charles P. Friedman, currently of the Center for Biomedical Informatics at the University of Pittsburgh, has also published extensively on similar topics and in overlapping journals, and his middle initial is sometimes omitted. Furthermore, there are at least three additional papers in Medline with the name C Friedman who also have New York affiliations.

Shown in Figure 4 is one of several pairs of articles (where one has Carol as a coauthor, while the second has Charles P as a coauthor) that shared journal name and 2 MeSH terms, where the pairwise model assigned a low match probability (0.2). This demonstrates the benefit of bringing interactive effects into the model. According to the model, the two MeSH terms do not increase the *r*-value when the journal already matches (see profiles 01000 and 010200 in Figure 2). If MeSH was assumed to be independent of journal, then 2 MeSH terms would yield an increase by a factor of 16.11 (see MeSH row in Table 2), and as a result, the assigned match probability would be close to 0.9,

PMID: 7895136
Title: A continuous-speech interface to a decision support system: II. An evaluation using a Wizard-of-Oz experimental paradigm.
Authors: Detmer WM, Shiffman S, Wyatt JC, **Friedman CP**, Lane CD, Fagan LM
Affiliation: Section on Medical Informatics, Stanford University School of Medicine, CA 94305-5479.
Journal: **J Am Med Inform Assoc**. 1995 Jan-Feb;2(1):46-57.
MeSH: - Adolescent; Algorithms; Animal; *Decision Making, Computer-Assisted; Dogs; Human; ***Natural Language Processing;** Prospective Studies; Reference Values; **Semantics**; Speech; Support, Non-U.S. Gov't; Support, U.S. Gov't, P.H.S.; Terminology; *User-Computer Interface

PMID: 7719797
Title: A general natural-language text processor for clinical radiology.
Authors: **Friedman C**, Alderson PO, Austin JH, Cimino JJ, Johnson SB.
Affiliation: Columbia University, New York, NY, USA.
Journal: **J Am Med Inform Assoc**. 1994 Mar-Apr;1(2):161-74.
MeSH: Diagnosis, Computer-Assisted; Human; Medical Records; ***Natural Language Processing**; *Radiology Information System; **Semantics**; Support, Non-U.S. Gov't; Support, U.S. Gov't, P.H.S.

FIG. 4. A pair of nonmatching papers (one by Carol and one by Charles P) that share several similarities.

PMID: 7882309
Title: Assignment of the human p27Kip1 gene to 12p13 and its analysis in leukemias.
Authors: Pietenpol JA, Bohlander SK, Sato Y, Papadopoulos N, Liu B, **Friedman C**, Trask BJ, Roberts JM, Kinzler KW, Rowley JD, et al.
Affiliation: **Johns Hopkins Oncology Center, Baltimore, Maryland** 21231.
Journal: Cancer Res. 1995 Mar 15;55(6):1206-10.

PMID: 8646710
Title: A double-blind comparison of the efficacy of two dose regimens of oral granisetron in preventing acute emesis in patients receiving moderately emetogenic chemotherapy.
Authors: Ettinger DS, Eisenberg PD, Fitts D, **Friedman C**, Wilson-Lynch K, Yocom K
Affiliation: The **Johns Hopkins Oncology Center, Baltimore, Maryland** 21287, USA.
Journal: Cancer. 1996 Jul 1;78(1):144-51.

FIG. 5. A pair of nonmatching papers (one by Cynthia L and one by Carl J) that share several similarities.

and labeled match (unless the triplet adjustment would have fixed it).

Shown in Figure 5 is a pair of articles (one has Cynthia L Friedman as a coauthor, while the second has Carl J Friedman as a coauthor) where the affiliation of the first author on both papers is Johns Hopkins Oncology Center, Baltimore, Maryland. Both articles are missing middle initials, and both are on the subject of cancer research published in similar journals (although Cynthia's is on genetics and Carl's is a clinical drug trial). As expected, before the geometric constraints are enforced, the model estimates the probability of match to be very high (0.99). However, after the geometric constraints are enforced, the model estimates the match probability to be 0.3. This shows that, in addition to bringing many of the true matches together, the triplet adjustment also helps separate the nonmatches.

The manual disambiguation can be used as a gold standard to evaluate the performance of the model based on the following measures:

Precision = the total number of distinct pairs correctly labeled as matches divided by the total number of distinct pairs labeled as matches

Recall = the total number of distinct pairs correctly labeled as matches divided by the total number of distinct pairs of true matches

Accuracy = number of distinct pairs correctly labeled (as matches or nonmatches) divided by the total number of distinct pairs

These measures can be thought of as the retrievals averaged over all distinct pairs of papers, where each paper represents a query, and the other papers are retrieved if their match probabilities are above 0.5. For example, in the case of C Friedman, the pairwise model (using an estimated prior of 0.122 and geometric constraint weight factor of 4—the following sections show how these parameters were obtained) labels a total of 3,836 pairs labeled matches, out of which 3,778 are true matches and 58 pairs are true nonmatches, for which precision = 98.5%, recall = 91.9%, and accuracy = 98.7%.

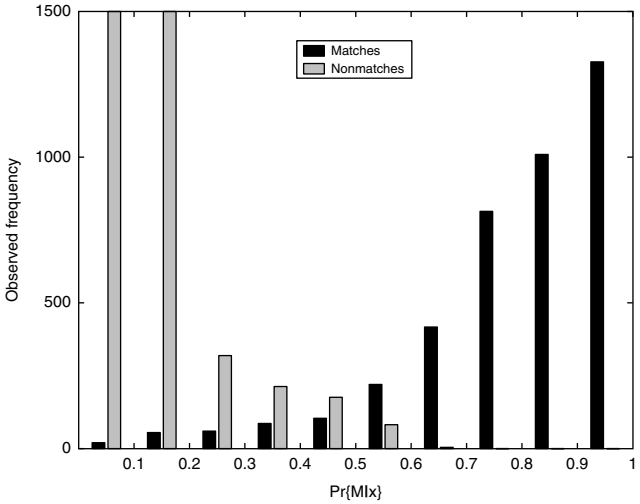

FIG. 6. The distribution of match probabilities in the C Friedman articles after using the estimated Pr{M} of 0.122 and a geometric constraint weight factor of 4.

Figure 6 shows the distribution of match probabilities in this case. Note that the y-axis was cut off at 1,500, to be able to see the low frequencies, albeit some of the frequencies do go above. It is clear that the model correctly assigns the great majority of the matches with high match probabilities and the great majority of the nonmatches with low match probabilities.

What is the best way to estimate Pr{M}? Out of the 30,628 distinct pairs of papers in the C Friedman case, there are 4,112 that we know are matches and 26,516 that we know are nonmatches. That is, the true a priori probability of match is Pr{M} = 0.134. This value can be used as a gold standard to assess the quality of the different estimates given in Table 3. The estimate based on name frequency alone is about 4 times greater than the true value, showing, as expected, that it comes with a high variance. The estimate based on the zero profile frequency 0.205 is more accurate

TABLE 3. Different estimates for the a priori probability of match Pr{M} in the case of C Friedman.

Method	Estimated Pr{M}
True value	0.134
Based on name frequency	0.550
Based on zero profile frequency	0.205
Predicted proportion of matches —using 0.205 prior, without enforcing three-way geometric constraints	0.104
Predicted proportion of matches —using 0.104 prior, and 4 for the three-way geometric constraint weight	0.122

(53% higher than the true value) because it takes into account the similarity of the C Friedman articles. Using this 0.205 in computing the pairwise probabilities yields 0.104 predicted proportion of matches, which is 22% below the true value. Using 0.104 as the prior and enforcing the geometric constraints yields a final estimate of 0.122, which is only off by 9%. In summary, the model is robust to inaccurate initial estimates because the majority of articles by the same people have very high match probabilities, and the majority of articles by different individuals have very low match probabilities, especially after imposing the three-way geometric constraints.

What is the optimal geometric constraint weight factor? Figure 7 shows the precision, recall, and accuracy of the model as a function of the "geometric" probability constraint weight factor w, using the true a priori match probability 0.134. In comparing all triplets of the 248 C Friedman papers, only about 1% violate this constraint, and most of these violations occur when comparing three papers authored by the same individual. For example, more than 60% of the triplets coming from Carol's 30 papers violate this constraint. Therefore, w should be set higher than 1. Based on Figure 7, the precision and recall break point occurs when w is 8, although this varies from name to name. One may also argue that the cost of incorrect labels are higher for true nonmatches than for true matches, implying that w should be set lower than the break point.

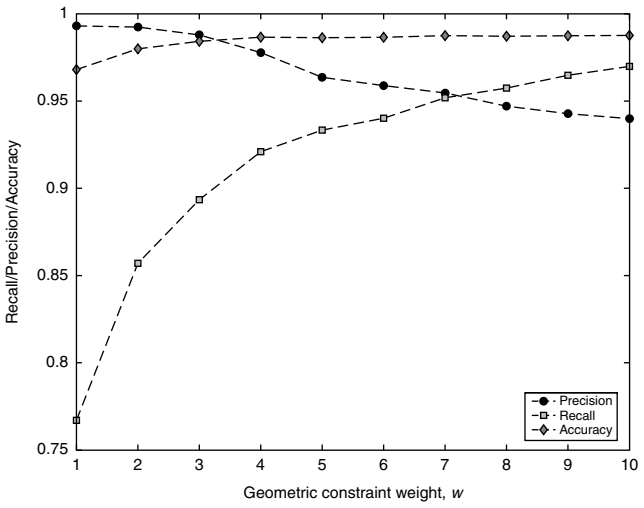

FIG. 7. The performance of the model in the case of C Friedman across different values for the geometric constraint weight.

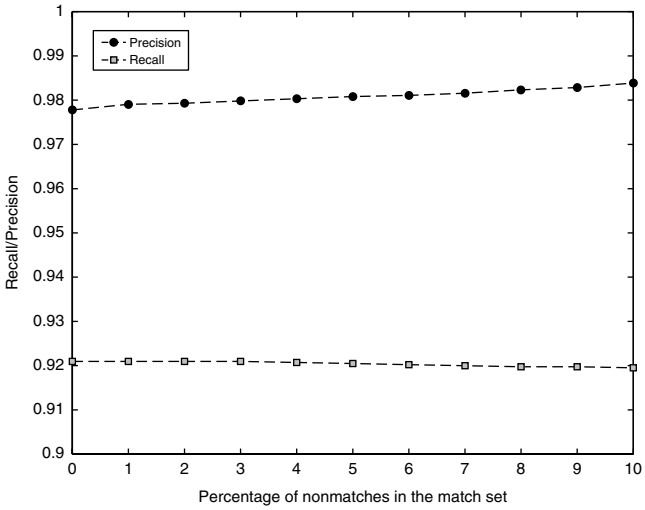

FIG. 8. The robustness of the model across different levels of match set contamination.

What happens when the match set is contaminated with nonmatches? Figure 8 shows how the precision and recall are affected when the r-values are adjusted to reflect cases when the match set is contaminated with a fixed proportion p of nonmatches as follows:

$$r_{\text{new}}(x) = (\Pr\{x|M\}\,(1-p) + \Pr\{x|N\}p)/\Pr\{x|N\}$$
$$= (1-p)r(x) + p.$$

We believe that the true level of contamination of the match set with nonmatches is probably very small (0.1% or less). However, even high levels of contamination (up to 10%) do not have a significant effect on recall and precision (Figure 8). This is a consequence of the fact that the majority of matches are assigned a high match probability and the majority of nonmatches are assigned a low high match probability (as shown in Figure 6).

Evaluation of the Fitted Model

There are three major criteria that need to be satisfied for the probability model to work well. In this section, these criteria are outlined and evaluated, and their resulting benefits and possible pitfalls are discussed.

Criterion 1: Is the similarity profile definition "the best possible"? The goal is to define a similarity profile that is as simple as possible while capturing as much information

necessary to create authorship fingerprints. This section addresses the issue of whether there is additional information available in the Medline records or other information processing methods that may be useful for matching authors.

Criterion 1a: Are there other attributes present in the Medline records that may be useful? We included as much information available in Medline as was thought to be useful. In the process of defining the best possible set of similarity scores, two attributes ended up being excluded, namely, the position on the article (e.g., first or last author) and presence of non-ASCII characters (e.g., Æ) in their last name. The two authors' orders may yield a slight decrease in probability of match when they are both first and there is nothing else in common between the two. However, this decrease was not significant enough to include in the final model. Also, a mismatch between two non-ASCII last names was so rarely observed in the reference sets that it was not included in the final model.

Criterion 1b: Why not weight the words by specificity? Some may argue that words should be weighted by their specificity (Salton, Wong, & Yang, 1975; Wilbur & Yang, 1996). For example, the MeSH term "child" should carry less weight because it less specific compared to other terms such as "schizophrenia." However, allowing for weights leads to a continuous domain for the similarity scores and makes multivariate analysis difficult without imposing further assumptions on the model such as linearity or independence. These assumptions are too restrictive and will lead to inappropriate estimates (see Criterion # 3). Even if it is possible to weight the terms, it will make the matching rules more complex and hard to evaluate manually. It is much easier just to count the number of terms in common without having to weight each term.

To address the issue of term weighting, the results of three different stoplists (small, medium and large) on the title words were compared. The words added to the larger stoplists, would be weighted lower than the ones that are on the smaller stoplist. In this sense, varying the extent of stoplisting captures the idea of a simplified weighting scheme. When the medium stoplist was applied to the match set, the frequency of the case having no title words in common increased from 64.9% to 80.1%, and as a result the *r*-value for 1 or more common title words increased by about 10 times. This creates a dilemma because a higher *r*-value is more indicative of a match, but at the cost of about 15% pairs of matches that lose all the title words they had in common. Even though these words are frequent in Medline, they may be important for disambiguation. In a weighting scheme the word "cell" will receive much smaller weight than the word "5-HT," for example, but within a population of authors that share a name, both words may be just as important. For example, if the population consists of a single person with lots of papers that use the word "cell," then that word will be important and should carry a high weight. In general, the problem is to find items in common because even articles written by the same individual only have something in common on either title, journal, coauthors, MeSH, *or* affiliation only about 60% of the time, on average. A weighting scheme could potentially remove important connections.

Criterion 1c: Potential ways to improve the similarity profile? Our model could definitely benefit from supplementary information not present in Medline but available from other sources such as the Web and journal publishers. For example, it would be very useful to have the first names spelled out, as well as the list of affiliations and which author name each corresponds to (instead of just the first author's affiliation as given in Medline). As well, reference citations are not included in Medline, although they may play an important role in matching authors because people often cite themselves.

It remains for future investigations to assess whether the model could be significantly improved by phrase processing or employing natural language processing techniques to define a disciplinary metric of similarity for MeSH or titles, or a geographic distance metric for affiliations (Churches, Christen, Lim, & Zhu, 2002). The title, MeSH, and affiliation fields may potentially benefit from a metric that is able to capture the similarity of words that are written differently.

Criterion 2: Do the reference sets accurately and unbiasedly represent Medline? The reference sets build the foundation for the probability model, and it is therefore essential that they unbiasedly and accurately represent Medline as a whole. Ideally, one would want to compute the exact probability distributions $\Pr\{\boldsymbol{x}|M\}$ and $\Pr\{\boldsymbol{x}|N\}$. Exact computations are practically impossible as they would require individuals encoded in the database, which is the goal in the first place. Traditionally, the matching process would be done manually or by selecting small, possibly biased subsets. In contrast, we chose to generate large reference sets in an automatic manner.

The key to reducing bias is to make sure that (1) the proportion of matches and nonmatches is very high in match sets and nonmatch sets, respectively, and (2) the reference sets do not contain a population of names whose similarity profile distributions vary significantly from the overall distributions. The article attribute match set was generated from author names that contained suffixes with an average *r*-value greater than 3,500. As we show next, the presence of suffixes is more common in English-speaking population groups, but otherwise provides a broad slice of Medline and is not biased on other attributes. The article attribute nonmatch set was generated by comparing author names with different last names and, as such, excludes matches altogether. The name attribute reference sets were also randomly selected from Medline as a whole and the targeted matches and nonmatches were defined by article similarity scores that yielded high *r*-values ($>$5,000) and low *r*-values ($<$0.01), respectively.

As Table 4 shows, the MeSH, title word, and affiliation distributions in the suffix set are quite similar to that of

TABLE 4. The % occurrence of the 20 topmost frequent words in the affiliation, MeSH, title and language fields in the suffix set (i.e., article attribute matchset) versus Medline overall.

Overall	Suffix	Affiliation	Overall	Suffix	MeSH	Overall	Suffix	Title	Overall	Suffix	Lang
49.96	58.29	university	2.25	2.32	In Vitro	6.64	6.75	effect	76.69	99.63	eng
22.02	32.89	medicine	2.20	2.25	Liver	6.41	5.69	cell	4.84	0.082	ger
18.51	42.35	usa	2.17	1.89	Base Sequence	5.75	4.89	study	0.31	0.064	por
16.40	25.5	medical	2.14	2.18	Rabbits	5.07	5.27	patient	1.37	0.062	spa
16.29	11.67	hospital	2.06	2.04	Amino Acid Sequence	4.33	4.57	human	4.40	0.026	rus
15.31	21.58	school	2.03	3.42	Follow-Up Studies	3.68	3.59	treatment	0.32	0.024	dut
12.85	11.05	institute	1.97	1.88	Cells, Cultured	3.50	3.58	disease	2.66	0.017	jpn
11.61	25.73	center	1.91	2.90	Methods	3.11	2.68	case	1.54	0.016	ita
8.97	8.95	research	1.90	2.17	Age Factors	2.93	2.56	protein	0.23	0.015	nor
8.67	9.23	science	1.86	2.30	Diagnosis, Differential	2.79	2.25	rat	0.34	0.0063	dan
7.48	0.40	japan*	1.84	3.84	Dogs	2.50	2.02	activity	0.01	0.0034	mul
6.51	9.94	college	1.79	1.86	Brain	2.50	2.47	clinical	0.51	0.0024	chi
6.41	9.99	health	1.78	1.59	Rats, Inbred Strains	2.22	2.26	factor	0.31	0.0019	swe
6.08	9.74	new	1.60	2.20	Aged, 80 and over	2.21	2.13	acid	0.49	0.0014	cze
5.39	12.64	surgery*	1.58	1.86	Cattle	2.06	2.45	induced	0.13	0.00097	slo
4.96	5.34	laboratory	1.55	1.88	Risk Factors	2.05	2.58	cancer	0.19	0.00097	hun
4.79	9.30	division	1.55	1.63	Cell Line	1.97	2.01	analysis	0.13	0.00048	bul
4.58	<0.1	uk*	1.54	1.44	Microscopy, Electron	1.97	1.67	new	0.16	0.00048	scr
4.42	5.74	national	1.48	3.04	Postoperative Complications	1.97	1.81	blood			
4.29	4.12	biology	1.47	1.86	Prognosis	1.84	1.90	receptor			

Medline as a whole. However, 99.6% of the articles in the suffix set were originally written in English versus 76.7% in Medline overall. Clearly, the suffix set is biased towards English articles. We overcame this by generating a separate language match set and incorporating the resulting language similarity score into the model based on the criterion that it is independent of the other article similarity scores.

Criterion 3: Could a linear or an independent model perform as well? In general, the multivariate distribution tends to enhance the effect of comparing matches to nonmatches. When matches have something in common, they tend to have several items in common, and this phenomenon does not occur for nonmatches. In other words, the positive correlation between variables tend to be higher for matches, than for nonmatches. For example, most journals are only published in one language, suggesting that most often when a match has the journal name in common, it also has the language in common more often than nonmatches. Figure 9 shows the pairwise interactive effects of the article similarity scores. Although this figure is complex, the observations made from the figure can be summarized as follows:

1. *Affiliation(s) not given* is the only similarity score that has a positive interactive effect with each of the other scores. That is, it tends to enhance the effect of the other variables, even though it is not very useful by itself. For example, if there are no affiliation words in common, then a match is less likely if both affiliations are given, than if one or both are missing.
2. The journal, MeSH, and title similarity scores have negative interactive effects on matching. It is intuitive that articles within a particular journal have similar title words and MeSH terms more so than different journals. It also makes sense that title words and MeSH have a redundant effect because both describe the topic of the article. Therefore, having MeSH terms in common does not add as much when the articles already have title words in common, and vice versa.
3. The effect of the affiliation similarity score is independent of the title, MeSH, and journal scores. This can be observed by the parallel curves in the column corresponding to affiliation. As a result, affiliation is more powerful for disambiguation, even when articles have title, MeSH or even journal in common.

These observations suggest that imposing linearity or independence constraints on the model may be inappropriate. To fully appreciate this fact, two alternate models were fitted to the training data, namely, a log-linear model and a product model, which assumes independence among all the similarity scores. The log-linear model was fitted using forward variable selection of all main effects and all possible interactive effects using S-Plus 6.1 (Insightful Corp, Seattle, WA). Each profile $\boldsymbol{x_a}$ was weighted by the number of times it was

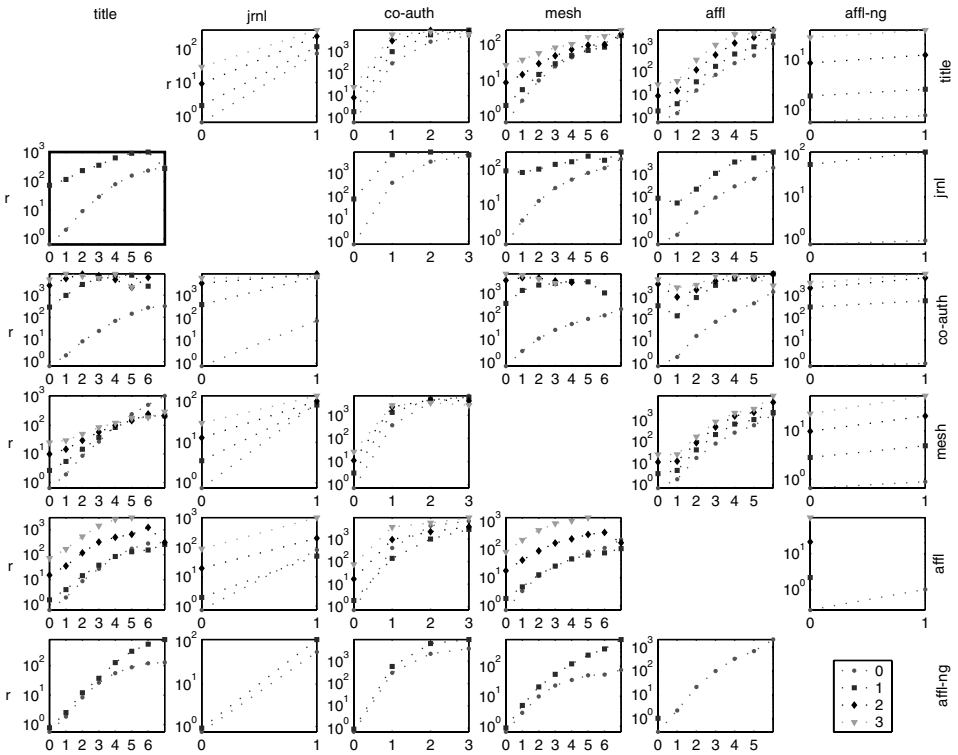

FIG. 9. Pairwise interactive effects of the article similarity scores. In each plot, the y-axis corresponds to the ratio $r(x_i, x_j)$, where x_i is the value of the row attribute and x_j is the value of the column attribute. The x-axis corresponds to the value of the column attribute, and the legends (shown on the bottom right) indicate the value of the row attribute. Within each panel, pairwise interactions are independent when the set of curves run parallel to each other, are positive when the set of curves tend to diverge (going from left to right), and are negative when they tend to converge.

observed in the match set and nonmatch set (i.e., $m(\boldsymbol{x_a}) + n(\boldsymbol{x_a})$. This resulted in the following model:

$$\log_{10}(r(x_3, x_4, x_5, x_6, x_8, x_9)) = 0.87 + 0.49x_3 + 1.93x_4 + 1.85x_5 + 0.57x_6 + 0.82x_8 + 0.64x_9 - 0.14x_3x_5 - 0.47x_4x_5 - 0.30x_4x_8 - 0.33x_5x_8 - 0.11x_6x_8,$$

yielding a multiple R^2 of 0.9854. However, this model does not satisfy the monotonicity criterion because of the negative interactive effects. For example, the title score (x_3) will have a negative effect when there are four or more MeSH terms (i.e., $x_6 > 4$) in common.

The product model is based on the assumption that the similarity scores act independently in their effect on the probability of match, and the *r*-values are estimated by

$$r(x_3, x_4, x_5, x_6, x_8, x_9) = \bar{r}_3(x_3)\bar{r}_4(x_4)\bar{r}_5(x_5)\bar{r}_6(x_6)\bar{r}_8(x_8)\bar{r}_9(x_9),$$

where the individual $\bar{r}_i(x_i)$ functions are taken from Table 2. This model does satisfy the monotonicity criterion; however, it is unable to capture the interactive effects of the article attributes.

Figure 10 shows a plot of the estimated versus the observed *r*-values for the three different models. The size of the points is proportional to the logarithm of the observed frequency in the nonmatch set and the match set. Ideally, the larger points would lie on or near the straight line with slight variations as a result of sampling. The majority of the log-linear based estimates above 10 tend to be significantly lower than the observed values. The product model seems to have the opposite effect, in that most the estimates above 10 tend by significantly greater than the observed values. Note that a similar pattern was observed for the log-linear model when interactions were taken out of the model. In contrast to the product model and the log-linear model, the estimates based on the monotonicity criterion are much closer to the observed values.

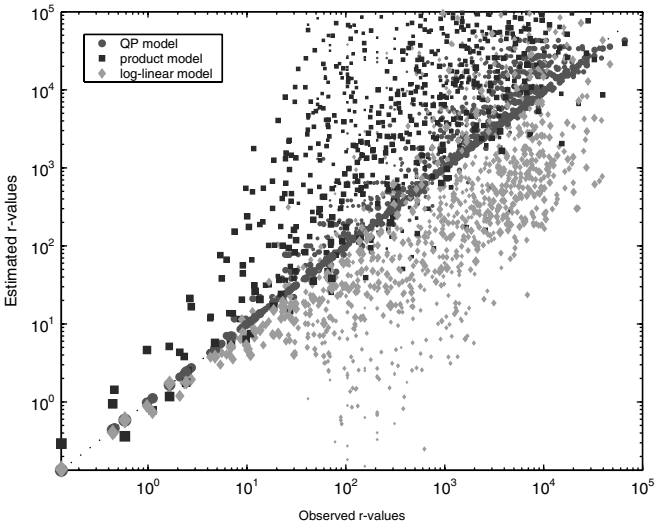

FIG. 10. The distribution of the residuals for the quadratic programming, log-linear, and product models.

Discussion

Summary of Methods and Results

This article introduces and validates a model for author name disambiguation that is based on pairwise comparisons of articles using information that is encoded in Medline fields. Although the present model takes advantage of features that are specific to Medline, the similarity profile approach outlined here can be applied to a wide variety of large-scale data-mining tasks and is not restricted to bibliographic or even textual databases. The model has several noteworthy features: (a) Massive training sets are automatically generated with both positive and negative examples, (b) nonlinear and interactive effects are incorporated across multiple variables, (c) the similarity profile is computed as a probability, whose constraints allow triplet corrections, (d) the method of counting words in common is simpler to compute than the calculation of term statistics, (e) the model is highly intuitive in interpretation—one can see the involvement of each parameter used in the model and its impact on overall probability value, and (f) the model retains the ability to incorporate outside knowledge without changing the basic framework.

An important feature of our approach is that we are not simply aiming for high performance in disambiguation, which might be achieved with sufficient brute-force manual effort or with simple empirical rules of thumb based largely on matching author attributes (e.g., first names and affiliations). Rather, the premise is to disambiguate papers in a manner that is not only automatic, but that permits analyzing fundamental patterns of publishing behavior. Including article attributes such as MeSH headings, title words, journal, language, and coauthors allows one to ask how many papers are published per person per year on average, how strong is the tendency to publish in the same journal over time, and how often scientists collaborate across disciplines or institutions—both across Medline as a whole and in selected subgroups of scientists. Thus, the model should be useful even if one has a list of papers that are known to be written by a single individual. If one takes one paper as the index paper for comparison, all of the other papers can be ranked in order of similarity, in a manner that maps the multidimensional nature of the Medline fields onto a single parameter. This ranking can serve as the basis for clustering an individual's papers by overall similarity, and for identifying "outliers" that differ significantly from the others, e.g., when individuals have moved, changed fields, or collaborated widely with many other groups.

The observed similarity profile distributions were very different in the match vs. nonmatch sets, and the test case of C Friedman showed very high precision (=98.5%), recall (=91.9%), and accuracy (=98.7%), suggesting that the present model is, indeed, adequate for disambiguating the

majority of articles, even though we did not encode first names or affiliations of each author. Although the match set is generated automatically and hence might contain a small proportion of nonmatches (~0.1% or less), we found that the model is quite robust: Even if the match set were to be contaminated with up to 10% nonmatches, this would not affect performance detectably in terms of precision and recall.

The most powerful measure for distinguishing matches from nonmatches is the number of coauthor names in common, followed by match on journal name, and then middle name initial match. Although suffix matches are important they are rare and as such less useful. The number of common affiliation words, title words, and MeSH are tied in fourth place. However, the affiliation similarity score tends to be independent of the journal, title, and MeSH similarity scores, indicating that it is more powerful for disambiguation when articles already have other items in common.

Planned Steps for Creating Author-Individual Clusters: Clustering Algorithms, Supplemental Information, and Assessing Name Variations

The present model for estimating the pairwise match probabilities is not intended to give optimal performance by itself. Rather, it is just the first of several planned steps toward our long-term goal of completely partitioning Medline into unique authors.

Because it is likely that supplementary information will be necessary to fully disambiguate author names, in the **second step**, we will supplement the Medline database with information extracted from personal and publishers' Web pages (Lawrence, Giles, & Bollacker, 1999). Author first names, affiliations and e-mail addresses will be obtained from online providers, when available, for all authors (not just the first author). We also plan to attempt to find online lists of publications by that individual. Such lists cannot be used as a primary means of disambiguation because they are often missing, incomplete, include non-Medline articles, and/or are not up to date. Yet they do provide a "gold standard" for validating that different articles on the list are written by the same person, and for identifying situations where two different author-individual clusters refer to the same individual.

Because one can expect much better results when a clustering strategy is used in addition to pairwise comparisons, in the **third step**, clustering algorithms (e.g., Jain & Dubes, 1988; Karypis, Han, & Kumar, 1999; Taskar, Segal, & Koller, 2001) will be employed on papers bearing the same (last name, first initial) to form clusters of papers that can be assigned to distinct author-individuals. The distribution of pairwise probabilities within a set of papers belonging to a specific name will provide constraints that allow one to adjust the estimated pairwise probabilities more accurately.

The most common reason that a paper may be misassigned is probably caused by missing data, but sometimes the journal prints the wrong name spelling or the wrong middle initial. As well, some names can be written in several different ways, for example, oriental names such as Wei Zhang are spelled Zhang Wei in some journals, and Medline sometimes indexes the first name as the last name. Hispanic and Slavic hyphenated surnames are often written in multiple nonstandard ways (Ruiz-Perez, Delgado Lopez-Cozar, and Jimenez-Contreras, 2002). In the **fourth step**, we plan to see whether changing the first and middle initials, the spelling of the last name (using a short edit distance), or reversing the first and last names would result in a high probability of match with some larger cluster with that name.

Although full disambiguation may never be possible using automatic methods alone, the approaches outlined in the present paper should greatly improve the efficiency of Medline searches on the author field, bibliometric studies (e.g., citation rankings), and characterizing individual scientists' authorship profiles and their collaboration networks over the medical literature.

Concluding Remarks

We have created a free, public service ("Authority": http://arrowsmith.psych.uic.edu) that takes as input an author's last name and first initial given on a specific article in Medline, and gives as output a list of all articles with that name ranked by decreasing similarity, with match probability indicated.

Acknowledgments

This research was funded jointly by the National Library of Medicine and the National Institute of Mental Health (R01LM07292) through the Neuroinformatics/Human Brain Project. We also thank the National Library of Medicine for graciously providing us with the 2002 baseline release of Medline.

References

Churches, T., Christen, P., Lim, K., & Zhu, J.X. (2002). Preparation of name and address data for record linkage using hidden Markov models. BMC Medical Informatics and Decision Making, 2, 9. Retrieved August 28, 2003, from http:// www.biomedcentral.com/1472–6947/2/9

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

Garfield, E. (1979). Citation indexing: Its theory and application in science, technology, and humanities. New York: Wiley.

Grossman, J.W. (2002). The evolution of the mathematical research collaboration graph. Congressus Numerantium, 158, 201–210.

Holmes, D.I., Gordon, L.J., & Wilson, C. (2001). A widow and her soldier: Stylometry and the case of the Pickett letters. Literary and Linguistic Computing, 16, 403–420.

Jain, A.K., & Dubes, R.C. (1988). Algorithms for clustering data. New York: Prentice-Hall.

Judson, D.H. (2002, June 14, 2002). Adventures in Bayesian record linkage. Paper presented at the annual meeting of the Classification Society of North America.

Karypis, G., Han, E.H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. IEEE Computer, 32, 68–75.

Lawrence, S., Giles, C.L., & Bollacker, K. (1999). Autonomous citation matching. Proceedings of the Third International Conference on Autonomous Agents (pp. 392–393). New York: ACM.

Newman, M.E.J. (2001). Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E, 64, 016131.

NLM Technical Bulletin (2001, November-December). Retrieved March 23, 2003, from http://www.nlm.nih.gov/pubs/techbull/

Noyons, E.C.M., Moed, H.F., & Van Raan, A.F.J. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. Journal of the American Society for Information Science, 50, 115–131.

Robertson, S.E. (1977). The probability ranking principle in IR. Journal of Documentation, 33, 294–304.

Robertson, T., Wright, F.T., & Dykstra, R.L. (1988). Order restricted statistical inference. New York: Wiley.

Ruiz-Perez, R., Delgado Lopez-Cozar, E., & Jimenez-Contreras, E. (2002). Spanish personal name variations in national and international biomedical databases: Implications for information retrieval and bibliometric studies. Journal of the Medical Library Association, 90, 411–30.

Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18, 613–620.

Sparck Jones, K., Walker, S., & Robertson, S.E. (2000). A probabilistic model of information retrieval: Development and status. Information Processing and Management, 36, 779–808 (part I), 809–840 (part II).

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

Taskar, B., Segal, E., & Koller, D. (2001). Probabilistic classification and clustering in relational data. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (pp. 870–876). San Francisco: Kaufmann.

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

Torvik, V.I., & Triantaphyllou, E. (2003). Guided inference of nested monotone Boolean functions. Information Sciences, 151, 171–200.

Torvik, V.I., Weeber, M., Smalheiser, N.R., & Swanson, D.R. (2002). Identifying authors that link disparate literatures. Poster presented at the NIH Human Brain Project annual meeting. Retrieved May 8, 2002, from http://arrowsmith. psych.uic.edu

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

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. (1995). Matching and record linkage. In B.G. Cox, D.A. Binder, & B.N. Chinnappa (Eds.), Business survey methods (pp. 355–384). New York: Wiley.

Yu, H., Hripcsak, G., & Friedman, C. (2002). Mapping abbreviations to full forms in biomedical articles. Journal of the American Medical Informatics Association, 9, 262–272.