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

**Vetle I. Torvik, PhD[1], Marc Weeber, PhD[1], Don R. Swanson, PhD[2], Neil R. Smalheiser, MD, PhD[1]**
**[1]Department of Psychiatry, University of Illinois at Chicago, Chicago, IL**
**[2]Division of the Humanities, University of Chicago, Chicago, IL**

**Abstract**

We present a model for automatically generating training sets and estimating the probability that a pair of Medline records sharing a last and first name initial are authored by the same individual, based on shared title words, journal name, co-authors, medical subject headings, language, and affiliation, as well as distinctive features of the name itself (i.e., presence of middle initial, suffix, and prevalence in Medline).

**Background**

Each author name recorded in Medline before 2002 is given by last name, first initial (if available), middle initial (if available) and suffix (such as Jr. or 3rd, if available). Although one can readily retrieve all articles containing a given name, many different people may share the same last name, first and even middle initial. Disambiguating author names in Medline is necessary to improve the efficiency of searches on the author field, as well as other bibliometric analyses (e.g. citation rankings and collaboration graphs).

**Methods**

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 invididuals and enough to create distinctive authorship fingerprints in most cases. To assess this hypothesis we generated four reference sets of pairs of articles containing almost exclusively author matches versus almost exclusively non-matches, in a manner that does not require manual assessment of articles, according to the following table:

| | Matchsets | Non-Matchsets |
|---|---|---|
| Yields name similarity vectors $\mathbf{x}_n$ | 54K pairs of 1st authors that match on last name, 1st initial, 1 or more co-authors, 2 or more affiliation words, and 2 or more MeSH | 9.2M pairs of 1st authors that match on last name, 1st initial, and have nothing else in common |
| Yields article similarity vectors $\mathbf{x}_a$ | 4.3 M pairs that match on last name, initials and suffix | 450M pairs that do not match on last name |

The vector **x** defines the similarity between a pair of names (matching on last name and first name initial) and the two articles they appear on, by the number of items they have in common. The title, MeSH and affiliation fields were conservatively stoplisted. Middle initial matches are further conditioned on whether they are missing (e.g., 0:[A,B], 1:[A,-] , 2:[-,-], 3:[A,A]). Similarly, language matches are conditioned on whether they are non-English. The reference sets were used to estimate the ratio $r(\mathbf{x}) = \Pr\{\mathbf{x}|M\}/\Pr\{\mathbf{x}|N\}$ by $r_n(\mathbf{x}_n)r_a(\mathbf{x}_a)$ for each observed vector $\mathbf{x} = (\mathbf{x}_n, \mathbf{x}_a)$ having verified that the name similarities ($\mathbf{x}_n$) and article similarities ($\mathbf{x}_a$) are independent. The constraint $r_a(\mathbf{x}_a) \geq r_a(\mathbf{y}_a)$ was validated and imposed on all pairs $(\mathbf{x}_a, \mathbf{y}_a)$: $\mathbf{x}_a \geq \mathbf{y}_a$ for smoothing, interpolation and extrapolation purposes. The estimated $r(\mathbf{x})$ can then be used to estimate the probability of match given any **x**:

$$\Pr\{M|x\} = \frac{1}{1 + \left(1 - \Pr\{M\}\right) / \left(\Pr\{M\} r(x)\right)}$$

Here, Pr{M}is estimated for the population of articles to be disambiguated, and as such takes into account the variability due to the frequency of the last name and the number of articles per individual.

**Results**

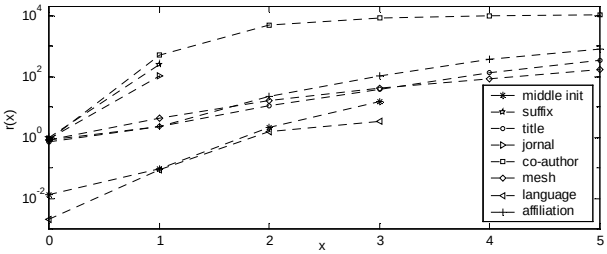


**Figure 1** Individual attributes' similarity distributions estimated by comparing the matchsets versus the non-matchsets.

**Conclusions**

We found that pairs of different articles authored by the same individuals (i.e., matches) do share similarities in one or more aspects of the Medline records, more so than non-matches. The most powerful measure for distinguishing matches from non-matches is the number of common co-authors, followed by journal match, and then middle initial match. Although suffix matches are important they are rare and, as such, less useful. The number of common affiliation words, title words, MeSH are tied for fourth place. However, about 40% of the pairs in the matchset have nothing in common other than last name, initials, and language. This can be partly attributed to missing data (e.g., affiliations belong only to the 1st authors and only about 40% of the records have them, and middle initials are often missing). One can expect better results when a clustering strategy is utilized in addition to the pairwise comparisons. However, it is likely that supplementary information (e.g., from personal webpages and publishers) will be necessary to fully disambiguate author names in Medline.