
The Arrowsmith Project Homepage - Netscape 6
File Edit View Search Go Bookmarks Tasks Help
http://arrowsmith.psych.uic.edu
Search
ARROWSMITH
LINKING
DOCUMENTS, DISCIPLINES, INVESTIGATORS AND DATABASES
Ready to use Arrowsmith?
Start
Introduction and tutorial
Start
NLM
UIC Department of Psychiatry
Document: Done (0.11 secs)

# Identifying Authors that Link Disparate Literatures

Vetle I. Torvik[1], Marc Weeber[2], Neil R. Smalheiser[1], Don R. Swanson[3]

The goal of an Arrowsmith "two node" search is to identify meaningful linkages that may exist between two apparently disparate sets of articles A and C, which may represent e.g. two distinct disciplines. To date, we have focused on identifying title words and phrases (B-terms) that are common to both sets of articles. However, identifying authors that link the two literatures (B-authors) and tracing their research efforts may also point to a significant body of information. We are now developing a new search mode, in which we will identify authors who link the two sets of articles. Such a search tool should identify individuals familiar with the science and scientists across two different disciplines, who may be particularly adept in appreciating new research directions, and in helping establish collaborations among research groups within each discipline. In addition to developing the author search mode, future research will address whether it can provide insights into the structure of the scientific enterprise, and whether it can provide information useful to laboratory scientists.

[1] University of Illinois at Chicago, Department of Psychiatry, Chicago, IL;
[2] Erasmus University of Rotterdam, Institute of Medical Informatics, Rotterdam, The Netherlands;
[3] University of Chicago, Division of the Humanities, Chicago, IL

## We propose to study four different degrees B-authors:

The "zero$^{th}$ degree" identifies individuals that have published papers in the direct A ∩ C literature.

The "first degree" identifies individuals that have published papers in both literatures A and C, but not in the direct A ∩ C literature.

The "second degree" identifies individuals that have published papers in one of the two literatures (either A or C, but not both) and have co-authored papers with people who have published in the other literature.

The "third degree" identifies individuals that have <u>not</u> published in either the A or the C literature, but have co-authored papers with an A-author and a C-author.

B-authors are likely to reveal a mixture of different kinds of relationships between the two literatures. B-authors of the zero$^{th}$ degree may be conducting a research project that crosses disciplines. B-authors of the 1$^{st}$ degree may be conducting somewhat disparate research in each of the two disciplines. B-authors of the 2$^{nd}$ degree may be conducting research in one of the two disciplines, and collaborating with individuals working in the other discipline. Some B-authors may simply be conducting research in a collaborative discipline (e.g., bioethics, statistics, or bioinformatics) which is equally related to many disciplines. Authors of the higher degrees are likely to reveal relationships that the lower degrees may not be able capture, for example, when the two literatures are disparate yet related. We need to perform real author mode searches to determine what kind of relationships are revealed and the ability of the different degrees to capture these relationships.

## Some examples of individuals who would want to identify B-authors and why

Individuals who are looking for a body of information that is related to both the A and the C literature but is not in the public domain (e.g., raw data, failed experiments, personal research notes).

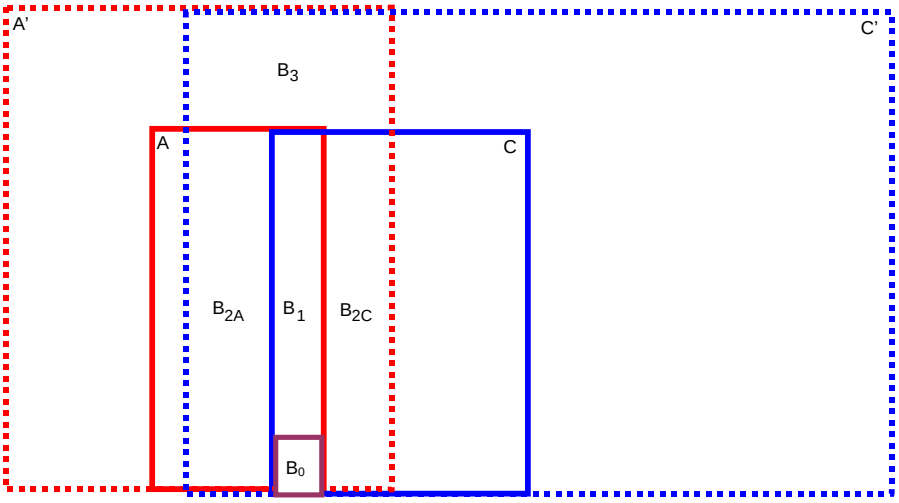
Individuals who are looking for a body of information that is related to both the A and the C literature but is not in Medline articles (e.g., non-English or non-biomedical literature) or hard to identify in Medline. In this case a B-author is a good starting point for further exploration. Tracing the research efforts of a B-author on a global scale or direct contact with a B-author may uncover a whole new body of information that is related to both literatures.

Researchers working in a discipline A who wish to identify possible collaborators that are specialists in a different discipline C. 2$^{nd}$ degree B-authors may be particularly adept to appreciate such a collaboration.

Administrators (e.g., program directors for funding agencies) looking for individuals that may facilitate research collaborations across two disciplines. B-authors are likely to be familiar with the science and scientists across disciplines.

## How to identify the B-authors

To carry out these searches efficiently, we have installed a local database of Medline. This allows us to quickly retrieve the lists of author names in the A, the C, and the A ∩ C literatures, as well as their co-author names. From the list of distinct A author names, a list of distinct co-author names is generated and labeled A'. Similarly a C' list is generated from the C author names. Then, the B-authors (of degree 0, 1, 2 or 3) are those names in common to both A' and C'.




$B_0$ = set of author names in the A ∩ C literature
A (or C) = set of author names in the A (or C) literature
A' (or C') = A (or C) author names and their co-authors
$B_1$ = (A ∩ C) / $B_0$
$B_{2A}$ = A ∩ ~C ∩ C'
$B_{2C}$ = C ∩ ~A ∩ A'
$B_3$ = A' ∩ C' ∩ ~A ∩ ~C

## Notes on identifying B-authors

If an author a $\in$ A , then a $\in$ A', and if an author c $\in$ C , then c $\in$ C'. That is, an author is considered a co-author of himself/herself.

An author qualifies for $B_3$ when he/she has co-authored with a $B_1$-author or a $B_0$-author. In a sense, the $B_1$-author or the $B_0$ author already establishes the link between A and C. Instead of removing such $B_3$-authors, it may be beneficial to subdivide the $B_3$-authors into additional categories until we can identify what type relationships they reveal.

Single-paper authors (about 43% of the authors in Medline) and can be removed from consideration for 1$^{st}$ and 2$^{nd}$ degree searches. In addition, single-author papers (about 26% of the articles in Medline) can be removed for the 3$^{rd}$ degree searches.

In reality an author may belong to several B categories while our definitions characterize an author by his/her smallest degree. As we begin to examine the B-authors empirically we can also assess whether it will be beneficial to rank them according to their "linkage strength". There are a number of parameters that may affect the linkage strength, such as the numbers of AC, A, and C articles, the numbers of A and C co-authors.

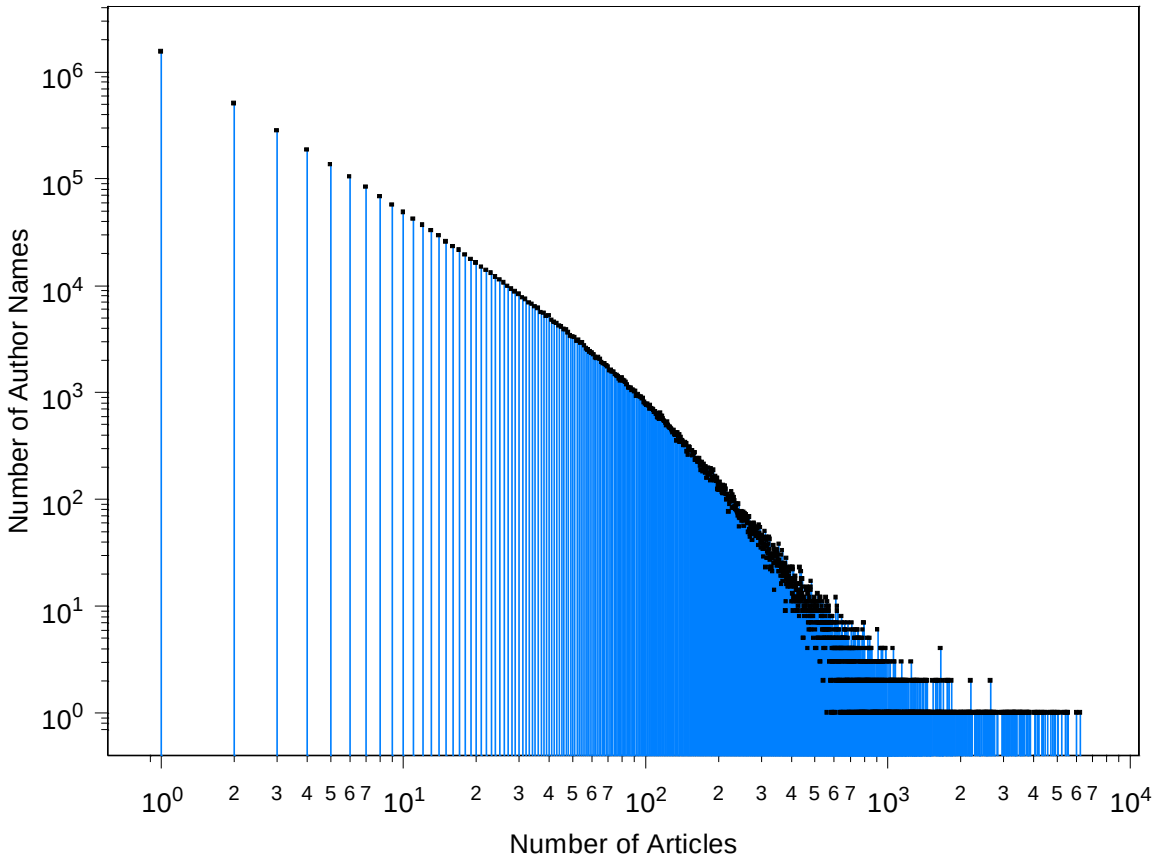
## Automatic identification of distinct authors in Medline

The author search mode raises challenges inherent in identifying an author by name only. On one hand, an author's name may have been written differently across the A and C literatures, or they may change their name over time (e.g. through marriage). Furthermore, the author's middle initial is dropped on occasion, and some journals convert international characters to English while others do not. These issues may result in missed author links. On the other hand, many people have the same names. Medline lists author names by last name, first name initial, and middle name initial, when available. For example, Medline currently has over 10,000 articles authored by "Smith_J", which obviously refers to more than one person. If one person has published in the A literature only, and another person with the same name has published in the C literature only, then this will lead to a false author link of the first degree. These facts significantly reduce the discriminating power of the Medline author names.

## Automatic identification … (continued)

Some Medline statistics (Baseline at the end of year 2001)
11,299,108 articles
3,659,842 unique authors names in ISO-latin1 (western European chars)
3,578,472 unique author names in ASCII (English chars)
1,543,803 author names in latin1 with one publication
2,948,380 articles with a single author name
6,230 max # of articles associated with a one name
259,402 articles with no persons listed as authors.
743 authors on a single article with the highest number
72,536 ASCII names with more than one latin1 variant

Figure: Frequency of author names in Medline articles.



## Automatic identification … (continued)

To identify distinct authors in Medline we will cluster author names based on information given in other fields such as co-authors, affiliations, MeSH terms, article type, journals, keywords, use of international symbols and language (if not English). For example, Medline contains over 1,100 articles with Smith_JA as an author, of which 5 contain Chan_CC as a co-author. Buggage_RR is also a co-author on several of these papers. Searching for papers authored by Smith/Buggage identified another article that does not include Chan as co-author. This identifies 6 articles in which Smith_JA is more likely to be the same person than an arbitrary pair of Smith_JA's. The clustering will not cause a loss of information, though possibly a single individual may be assigned to two different clusters (e.g. if they switched fields, affiliations, and co-authors entirely during their career).

## Automatic identification … (continued)

We plan to perform clustering for each distinct author name in Medline, and store this information in a database for efficient retrieval when need for identifying B-authors. An author name within a cluster will be assigned a unique author id. The goal is to determine the clusters so as to minimize the number of false matches and false non-matches. There is no way to achieve 100% accuracy, and the accuracy can at best be estimated from a small sample of author id's that are evaluated by a human.

The clustering will be performed iteratively using intuitive clustering rules based on a set of given variables (co-authors, etc.) until a termination criterion is met. Clustering based on co-authors only will be explored first since it is expected to provide the better results than any other variable alone. The termination criterion will be optimized, after which it will be determined whether additional variables significantly increase the accuracy of the clustering algorithm.

About 2,000 new articles are added to Medline per day. In addition, existing articles are revised and deleted. The new clusters are likely to be similar to the old clusters. Therefore, it is not necessary to recompute all the clusters for all authors in Medline from scratch each time the database is updated. Instead, an incremental version of the clustering algorithm will also be designed.

## Observations that will factor into the design of the clustering algorithm

We would rather catch a link than miss it. Therefore, we would rather have fewer than more clusters.

Individuals with many co-authors (a result of articles with many authors and authors who have published many articles) may have published with different individuals with the same name. Resolution: change rule to match a fraction of the number of co-authors, or simply remove say top 1%.

Affiliations are often not recorded and change over time.

Not matching a pair of author names based on non-matching MESH terms may negatively affect the 1$^{st}$ degree author links when the A and C literature is defined in terms of single MESH categories.

Computational complexity of clustering is likely to be greater than quadratic in the number of (author name, article) pairs.

Matching based on co-author names is powerful. An author name that appears on two different articles are more likely to refer to the same individual if the two articles share the same co-author names, than if not. Also, articles that have author names in Medline have all the author names listed (i.e., no et al.'s.)