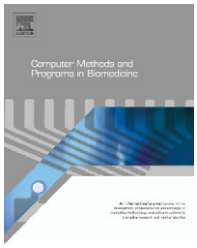
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 94 (2009) 190–197

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE

Neil R. Smalheiser, Vetle I. Torvik, Wei Zhou*

Department of Psychiatry and Psychiatric Institute, MC912, University of Illinois at Chicago, 1601W. Taylor Street, Chicago, IL 60612, USA

ARTICLE INFO

Article history:
Received 7 September 2008
Received in revised form
17 October 2008
Accepted 12 December 2008

Keywords:
Text mining
Web server
Hypothesis
Literature-based discovery

ABSTRACT

The Arrowsmith two-node search is a strategy that is designed to assist biomedical investigators in formulating and assessing scientific hypotheses. More generally, it allows users to identify biologically meaningful links between any two sets of articles A and C in PubMed, even when these share no articles or authors in common and represent disparate topics or disciplines. The key idea is to relate the two sets of articles via title words and phrases (B-terms) that they share. We have created a free, public web-based version of the two-node search tool (http://arrowsmith.psych.uic.edu), have described its development and implementation, and have presented analyses of individual two-node searches. In this paper, we provide an updated tutorial intended for end-users, that covers the use of the tool for a variety of potential scientific use case scenarios. For example, one can assess a recent experimental, clinical or epidemiologic finding that connects two disparate fields of inquiry—identifying likely mechanisms to explain the finding, and choosing promising follow-up lines of investigation. Alternatively, one can assess whether the existing scientific literature lends indirect support to a hypothesis posed by the user that has not yet been investigated. One can also employ two-node searches to search for novel hypotheses. Arrowsmith provides a service that cannot be carried out feasibly via standard PubMed searches or by other available text mining tools.

1. Introduction

The Arrowsmith two-node search tool [1–5] is designed to assist biomedical investigators in formulating and assessing scientific hypotheses. More generally, it allows users to identify biologically meaningful links between any two sets of articles A and C in PubMed, even when A and C share no articles or authors in common and represent disparate topics or disciplines. This fundamental text mining strategy provides a service that cannot be carried out feasibly via standard PubMed searches. Although other ways to link disparate literatures have been studied [e.g., 6,7], to our knowledge, no other two-node search tool has been made freely available to the scientific community.

The key idea, as shown in Fig. 1, is to relate two sets of articles, or literatures (A and C) via title words and phrases (B-terms) that they share. To carry out a two-node search, the user is asked to input two separate PubMed queries that define A and C. Any articles present in both A and C are removed so that the analysis will consider only *indirect* linkages between the

* *Corresponding author*. Tel.: +1 312 413 4581; fax: +1 312 413 4569.
E-mail address: neils@uic.edu (N.R. Smalheiser).
0169-2607/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2008.12.006

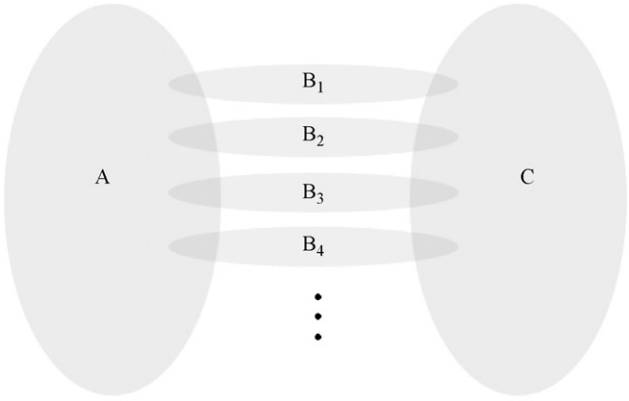


Fig. 1 – Venn diagram illustrating the Arrowsmith data mining model. Two disparate sets of articles (A and C) are implicitly related via title terms (B_i's) that they share. Reprinted from Ref. [5] with permission.

two sets of articles. Then the software identifies all words and 2- and 3-word phrases that are found in the titles of articles in both A and C. These so-called B-terms are processed and ranked according to the predicted probability that they will be relevant for some user in pointing to a meaningful link across the two literatures [5]. The interface then displays the ranked list of B-terms; clicking on a B-term opens a new window that displays the titles that contain A and B juxtaposed to the titles that contain B and C. In this fashion, the user can readily see whether, and how, the two sets of articles are related.

Ten years ago, we published a tutorial on the practical use of the two-node search tool in this journal [2]. Since then, however, the two-node search tool has undergone extensive development; its underlying algorithms have been greatly improved and its interface is now substantially more advanced [1–5]. We believe the time is right to provide an updated tutorial intended for end-users, that covers a variety of potential scientific use case scenarios.

2. Materials and methods

The UIC team worked with several groups of neuroscience field testers who carried out searches during the course of their daily scientific work [4]. Their feedback led to improvements and permitted us to acquire a set of "gold standard" searches that were employed for quantitative modeling [5]. In a gold standard search, a user marked all B-terms as "relevant" that were useful in answering the question that motivated the search. We then characterized each B-term according to eight different features [3,5], and formulated a quantitative model that optimally separated the set of relevant B-terms from other terms. For each B-term, the model gives the estimated probability that it will be deemed as relevant by some user [5]. The model gave significantly better performance than other proposed methods both on the original set of gold standard searches and on 20 external gold standards derived from TREC Genomics Track 2006 queries [5].

A simplified version of the PubMed query box was imported into the two-node search web interface, so that users input two PubMed queries in order to define the two sets of articles (or literatures) A and C. To retrieve MEDLINE records corresponding to user queries quickly and automatically, a local customized database of MEDLINE was created and updated weekly. When a query is entered, the article ID numbers are downloaded from PubMed and the full MEDLINE records are retrieved from the local database, including a tokenized and stoplisted version of each article title. Articles not found in the local database are downloaded from PubMed as XML files, processed and stored in the local database. Note that the web interface currently processes only the most recent 50,000 articles retrieved for a given PubMed query. B-terms and their feature values are computed in a parallel fashion by processing the sets of tokenized and stoplisted titles in chunks on separate processors, and merging the results when each process is done. The baseline 2005 version of MEDLINE was processed to identify all terms (words and up to 3-word phrases) in titles. Wherever possible, B-term features were pre-computed and stored in the term database for fast look-up [5].

3. Results

The two-node search can contribute to at least five scientific use case scenarios:

1. Identifying concepts or items that have been studied in both A and C (albeit possibly from different points of view).
2. Assessing a recent experimental, clinical or epidemiologic finding that connects two disparate fields of inquiry: (a) identifying likely mechanisms to explain the finding, and (b) choosing promising follow-up lines of investigation.
3. Assessing a novel, but hitherto un-investigated, hypothesis, to learn if the existing scientific literature lends indirect support.
4. Integrating information regarding a single concept or phenomenon that has been studied in two different isolated contexts.
5. Searching for new hypotheses (e.g., by assessing the results of a one-node search).

3.1. *Identifying concepts or items that have been studied in both A and C (albeit possibly from different points of view)*

Perhaps the simplest task for the two-node search tool is to enumerate a list of concepts or items that have been studied in both A and C. For example, studies of microRNAs and of Xist (a noncoding RNA expressed on the X chromosome) comprise two distinct fields of research. Suppose a user wanted to identify a list of entities that have been discussed with regard to both classes of RNAs. A PubMed search on [microRNA AND Xist] finds three articles. However, this will only identify the set of papers that discuss both RNAs, and will miss the many papers that discuss only microRNAs or only Xist. A two-node search gives a much more complete answer: Two PubMed searches are conducted with input query A = microRNA (3582 articles) and C = Xist (576 articles). A large number of B-terms are identified (1127) of which 354 are predicted to be relevant (Fig. 2). These include a heterogeneous mix of entities stud-