BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 13 2007, pages 1658–1665
doi:10.1093/bioinformatics/btm161

Data and text mining

A quantitative model for linking two disparate sets of articles in MEDLINE

Vetle I. Torvik and Neil R. Smalheiser*

Department of Psychiatry and Psychiatric Institute (MC912), University of Illinois-Chicago, Chicago, IL 60612, USA

Received on March 1, 2007; revised on April 20, 2007; accepted on April 21, 2007

Advance Access publication April 26, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Background: Identifying information that implicitly links two disparate sets of articles is a fundamental and intuitive data mining strategy that can help investigators address real scientific questions. The Arrowsmith two-node search finds title words and phrases (so-called B-terms) that are shared across two sets of articles within MEDLINE and displays them in a manner that facilitates human assessment. A serious stumbling-block has been the lack of a quantitative model for predicting which of the hundreds if not thousands of B-terms computed for a given search are most likely to be relevant to the investigator.

Methodology/Principal Findings: Using a public two-node search interface, field testers devised a set of two-node searches under real life conditions and a certain number of B-terms were marked relevant. These were employed as 'gold standards;' each B-term was characterized according to eight complementary features that were strongly correlated with relevance. A logistic regression model was developed that permits one to estimate the probability of relevance for each B-term, to rank B-terms according to their likely relevance, and to estimate the overall number of relevant B-terms inherent in a given two-node search.

Conclusions/Significance: The model greatly simplifies and streamlines the process of carrying out a two-node search, and may be applicable to a number of other literature-based discovery applications, including the so-called one-node search and related gene-centric strategies that incorporate implicit links to predict how genes may be related to each other and to human diseases. This should encourage much wider exploration of text mining for implicit information among the general scientific community.

Availability: Two-node searches can be carried out freely at http://arrowsmith.psych.uic.edu

Contact: neils@uic.edu, vtorvik@uic.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The field of text mining has arisen to assist investigators in identifying, summarizing and integrating the key findings relevant to a given scientific question (Cohen and Hersh, 2005; Hunter and Cohen, 2006; Jensen *et al.*, 2006; Krallinger and Valencia, 2005). Whereas information retrieval and information extraction techniques deal with information that is stated explicitly in a scientific article, literature-based discovery techniques are designed to identify assertions that are implicit within or across two or more articles (Hristovski *et al.*, 2005; Smalheiser, 2005; Smalheiser *et al.*, 2006; Srinivasan, 2004; Weeber *et al.*, 2005; Wren *et al.*, 2004).

A common scenario involving implicit information arises, when an investigator finds experimentally that two phenomena, previously thought to be unrelated, are unexpectedly related in some meaningful way, and would like to find existing knowledge that might shed light on potential mechanistic links between them. Alternatively, an investigator may hypothesize that a link exists between two disparate phenomena, and wish to assess whether the existing literature provides any implicit support for the hypothesis that would encourage experimental testing (Smalheiser and Swanson, 1998; Swanson and Smalheiser, 1997). A variety of other daily information-seeking activities also involve looking for items or concepts that are shared by two different sets of articles: for example, a physician may want to compile a list of symptoms that are shared in two different diseases, or a student may wish to browse the literature of an unfamiliar discipline for information that is likely to be relevant to his or her home discipline (Smalheiser *et al.*, 2006). Thus, text mining for implicit information includes situations in which one is searching for known findings, as well as identifying novel hypotheses or previously unreported links between two different sets of documents.

The Arrowsmith two-node search strategy has been proposed to assist in finding and assessing implicit knowledge (Smalheiser and Swanson, 1998; Swanson and Smalheiser, 1997). As illustrated in Figure 1, the key idea is to identify one or more B-terms that link two sets of articles (or 'literatures') A and C. In the biomedical domain, the searcher carries out two different PubMed searches that define two sets of articles A and C—often these may represent disparate disciplines and have few or no articles in common. Then, the system collects all words and 2- or 3-word phrases ('B-terms') that are shared in the titles of A and C, and displays these to the searcher so that the titles in A containing a given B-term can be juxtaposed directly to those containing that B-term in C, thereby allowing one to assess quickly whether they are related meaningfully.

Although the two-node search has promise for becoming a major means of mining the scientific literature, and has been implemented in free, public web interfaces, a typical search may

*To whom correspondence should be addressed.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1658/223810 by guest on 26 January 2019

produce a list of hundreds to a few thousand B-terms. During field-tester evaluations carried out during the past few years (Smalheiser, 2005; Smalheiser *et al.*, 2006), we progressively introduced a total of eight filtering options that appeared useful for restricting and ranking B-terms; for example, users could specify the minimum number of occurrences of a B-term within the A or C literature, or its desired semantic categor(ies); as well, we defined a literature cohesion score (Swanson *et al.*, 2006), which was employed for ranking the list of B-terms. Some of the features were frequency-based, some not; some were intrinsic to the B-term itself and some varied according to the specific pair of literatures employed in the search (Section 2.2).

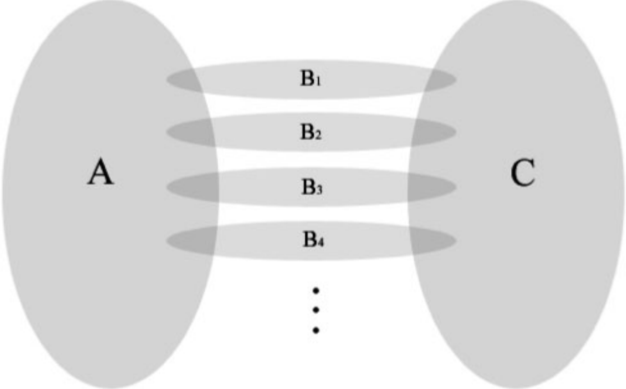

Fig. 1. Venn diagram illustrating the Arrowsmith data mining model, in which two disparate sets of articles (A and C) are implicitly related via title terms (B_i's) that they share.

However, a serious stumbling-block has been the lack of a quantitative model for automatically predicting which B-terms are most likely to be relevant to a given search. Mutual information measures have been proposed to score B-terms (Wren, 2004), but these methods lack a significant corpus of gold-standard searches for evaluation. As shown in the present report, we have accumulated a set of six diverse two-node searches, which permitted us not only to formulate a quantitative model for predicting which B-terms are likely to be relevant to a given search, but to measure the overall amount of implicit information shared by two different literatures.

2 RESULTS

2.1 Two-node searches used as gold standards

Field testers used the Arrowsmith tool in the context of their daily scientific work and their feedback contributed to the development of the software and web interface (Smalheiser *et al.*, 2006). Six real world searches formulated by six different people were selected to serve as 'gold standards' for quantitative modeling (Table 1). These consisted of pairs of topically coherent sets of articles with few or no articles in common, ranging from several hundred to several thousand articles in each set. The scientific contexts of some of these searches were discussed in Smalheiser *et al.* (2006). Each list of B-terms was manually checked by N.S. to identify all terms that were deemed relevant to a specific question, and to ensure consistency in scoring criteria. Note that some of the searches were analyzed twice to answer two different questions (Table 1). For example, the search on lit A = retinal detachment

Table 1. Six real-world two-node searches used as 'gold standards'

A-literature query	C-literature query	B-terms	Relevant B-terms sought
Retinal detachment[ti] $n=5122$	Aortic aneurysm[ti] $n=5687$	$n=2294$	(a) Diseases or syndromes in which both features have been described. $n=30$ (b) Surgical procedures used for diagnosis or treatment of both. $n=26$
mglur5[ti] OR metabotropic glutamate receptor[ti] OR metabotropic glutamate receptors[ti] $n=2032$	Lewy body[ti] OR Lewy bodies[ti] $n=1141$	$n=820$	(a) Signaling molecules that directly or indirectly modulate or are modulated by mGluR5 and that either modulate Lewy bodies or are altered in diseases that have Lewy bodies. $n=19$ b) Specific brain regions studied in both. $n=42$
'Magnesium'[mh] AND magnesium[ti] AND ('1900'[pdat]:' 1987/12/31'[pdat]) $n=6238$	('Migraine disorders' [mh] AND migraine[ti]) AND ('1900'[pdat]: '1987/12/31'[pdat]) $n=3205$	$n=1879$	Terms described as relevant in Swanson *et al.* (2006) excluding two judged too general to be useful (reactivity and spreading). $n=41$
Beta-amyloid precursor protein[ti] OR amyloid precursor protein[ti] OR APP[ti] AND ('amyloid'[mh] OR amyloid [tw]) $n=2118$	Reelin[All Fields] $n=493$	$n=1003$	Genes or proteins shared in Reelin and APP (amyloid precursor protein) signal transduction pathways. $n=54$
('Nitric oxide'[mh] OR nitric oxide[ti]) AND ('mitochondria'[mh] OR mitochondria[ti] OR mitochondrial[ti]) $n=786$	psd[ti] OR psd93[ti] OR psd95[ti] OR psds[ti] OR 'postsynaptic density' [ti] OR 'postsynaptic densities'[ti] $n=545$	$n=584$	Physiological or pathological responses that link the action of nitric oxide on mitochondria and the normal function of post-synaptic densities. $n=51$
Calpain[ti] OR 'calpain'[mh] $n=3352$	Postsynaptic[All Fields] AND density[All Fields] $n=2562$	$n=3131$	Protein substrates that are cleaved by calpain, and that are either localized at or act upon the postsynaptic density. $n=63$

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1658/223810 by guest on 26 January 2019

versus lit C = aortic aneurysm was analyzed to identify diseases or syndromes in which both conditions had been described (not necessarily in the same paper or in the same patient); and, separately, analyzed to create a list of surgical procedures used for diagnosis or treatment in both conditions. In most cases, the relevance of a B-term was readily apparent from observing the corresponding titles of the AB and BC papers. However, in some cases, scanning the abstracts of these papers or utilizing other sources of knowledge was necessary to verify the relevance of a B-term (e.g. some B-terms were not familiar to users and had to be looked up). If a B-term was marked relevant, its close variants were automatically considered to be relevant as well (e.g. microtubule-associated protein-2, MAP-2, MAP2 and MAP 2). Note that some meaningful terms do not appear explicitly on the B-list because they are longer than three words. For example, in the case of '*N*-methyl *D*-aspartate receptor subunit,' we marked the shorter fragments on the B-list that most clearly pointed to its meaning; '*N*-methyl *D*-aspartate receptor' and 'NMDA receptor' would be marked positive, but '*N*-methyl *D*-aspartate' and 'receptor subunit' would not. Each search was also manually inspected to mark a similar number of B-terms as non-relevant that were not expected to be relevant in any search scenario that could be envisioned. B-terms marked non-relevant tended to be common, general words, e.g. 'necessary,' 'newly' or 'underlying'.

All B-terms were divided into two sets: (1) terms marked relevant, and (2) all other terms (the 'mixed' set). Note that the 'mixed' set potentially includes at least 3 different types of B-terms: (a) B-terms that were not marked, but might have been scored as relevant by a different person or in response to a different stated query need; (b) B-terms that might be relevant in reality, but were not marked due to the incomplete state of our current scientific knowledge; and (c) B-terms that are truly non-relevant to the task of linking A and C in meaningful manner—some of which were marked as non-relevant and some which were not marked. All in all, the pooled data set consisted of 9711 B-terms across all gold-standard searches, including 326 in the marked relevant set, and 9385 in the 'mixed' set, of which 212 were marked non-relevant.

2.2 Selecting features

As discussed earlier, eight complementary features of B-terms had previously been developed for manually filtering the list of B-terms on the Arrowsmith website (Smalheiser, 2005), including features that capture various aspects of absolute and relative frequencies of the terms within each literature, recency of the term when first appearing in MEDLINE, cohesion (roughly, whether the term refers to a highly specific or a general concept) (Swanson *et al*., 2006), concepthood (i.e. whether the term maps to any UMLS semantic categories or not), shared MeSH headings between the AB and BC subliteratures (Swanson *et al*., 2006), and presence on a 1400 word stoplist of common words [see Smalheiser (2005) and Supplementary Table S1 for further discussion]. Note that five of the features are global, i.e. the feature value for a given B-term is the same across all searches, whereas three features vary according to the specific pair of literatures employed in the search. The list of features is somewhat arbitrary, insofar as each could have been scored using somewhat different formulas, and the list is certainly incomplete, insofar as one can imagine additional features that could have been scored (see Discussion section). However, the 8 features gave excellent performance (see subsequently) suggesting that they are sufficient to serve as a basis for initial modeling. Each feature was scored as described in Supplementary Table S1, such that each B-term was represented by an 8-dimensional vector.

2.3 Combining the features into a single B-term score

Our goal was to derive a mathematical formula for mapping each 8-dimensional vector to a single B-term score, so that this score optimally distinguishes the marked relevant B-terms from all other B-terms. A simple and intuitive formula is the weighted combination of the features:

$$\text{B-term score}: y(\mathbf{x}) = w_1x_1 + w_2x_2 + \cdots + w_8x_8,$$

where $\mathbf{x}$ denotes the B-term feature vector $<x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8>$, x_i denotes the value of the *i*-th feature value of a B-term and w_i denotes the corresponding weight. The weights w_i's capture the effect of each feature regardless of the nature of the B-terms (e.g. whether they refer to genes, diseases, etc.). In order to assess whether this formula could be appropriately applied in this situation, we first checked that the individual features satisfied a linear model and did not show strong interactive effects (see Methods section). Logistic regression was used to estimate the weight parameters because it related the B-term score directly to its probability of being relevant; because it readily allowed for assessing the statistical significance of each parameter; and because it allowed for taking into account the fact that each gold-standard search had a different proportion of marked relevant terms.

Each individual feature showed a statistically significant association with B-term relevance (Supplementary Table S2) and using all features together was much more powerful than any single feature alone. However, the stoplist feature (x_8) tended to be redundant with a combination of other features, and caused some marked relevant terms (e.g. Down syndrome) to be scored poorly. Thus, we decided to exclude the 1400 word stoplist feature from the final multi-dimensional model. (Note that we did employ a short 365 word PubMed stoplist to remove the most common words from titles prior to processing and scoring B-terms.) The weight parameters were then computed using logistic regression, resulting in the following formula for calculating the score of a B-term with 7-dimensional feature vector:

$$\begin{aligned}\text{B-term score}: y = {} & 0.73x_1 + 0.99x_2 + 1.32x_3 + 13.8x_4 \\ & + 0.59x_5 + 0.040x_6 + 0.19x_7.\end{aligned}$$

It is worth discussing why the statistical model was designed to optimally separate the set of marked relevant B-terms from all other B-terms (the 'mixed' set), rather than to separate the marked relevant B-terms from the marked non-relevant B-terms. Whereas the relevant B-terms were marked with high confidence, the non-relevant B-terms were subject to potential bias since marking them requires a person to assess whether a given term should be non-relevant to ALL users under ANY likely scenario. Our approach fits into the machine

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1658/223810 by guest on 26 January 2019

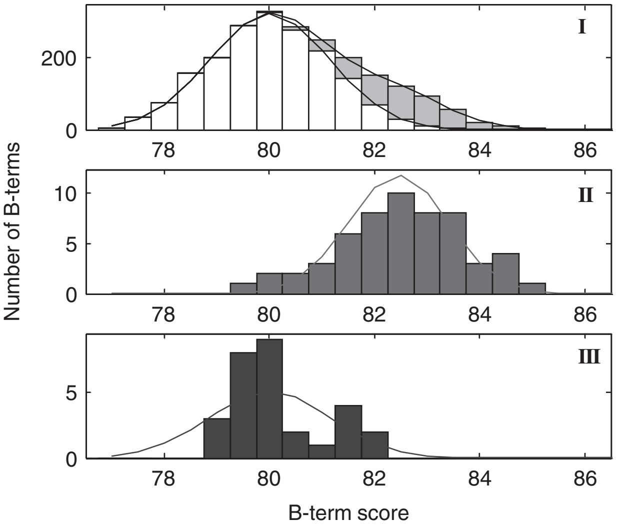

Fig. 2. Distribution of B-term scores for a representative gold-standard two-node search (retinal detachment versus aortic aneurysm). Observed data are represented by bars, and predicted values are represented by curves. (**I**) The overall histogram of B-term scores is fitted by a curve comprising a mixture of two normal distributions. The shaded area represents the proportion of predicted relevant B-terms, and the white area represents the predicted non-relevant B-terms. (**II**) Histogram of marked relevant B-terms (note that it is centered at the mean of the predicted relevant curve). (**III**) Histogram of marked non-relevant B-terms (note that it is centered at the mean of the predicted non-relevant curve).

learning paradigm of classification using a mixture of labeled and unlabeled examples (Nigam *et al.*, 2000). The feasibility of this approach requires that most of the B-terms in the mixed set consist of negative examples (Li and Liu, 2005), and this appears to be satisfied in the present case since, as shown below, the predicted relevant B-terms comprise <20% of all B-terms on average, and even a smaller proportion of the mixed set. To confirm that the weight parameters are not affected by the presence of predicted relevant B-terms within the mixed set, we removed the B-terms having the top 20% of B-term scores from the mixed set of each gold standard, and retrained the model using the reduced mixed set. None of the weight parameters changed significantly (data not shown).

2.4 Decomposing B-term scores into two normal distributions

When the B-term scores for each gold-standard search were plotted as a histogram, we observed that it seemed to follow a symmetric curve except for an excess of high scores (Fig. 2-I) comprising the marked relevant B-terms as well as additional B-terms from the mixed set. When a customized χ^2 goodness of fit method was employed to fit each histogram to a mixture of two normal bell-shaped curves (see Methods section), a striking pattern emerged: the rightmost bell-shape curve was centered at the mean score of the marked relevant B-terms (Fig. 2-II), whereas the leftmost bell-shape curve corresponded nicely with the mean score of the marked non-relevant B-terms (Fig. 2-III).

Thus it seemed reasonable to decompose the histogram of B-term scores into two parts, predicted relevant and predicted non-relevant:

$$f(y) = pf_{\mathrm{R}}(y) + (1-p)f_{\mathrm{N}}(y),$$

where $f(y)$ denotes the approximated probability distribution function for the B-term score y, p denotes the predicted proportion of relevant B-terms, and $f_{\mathrm{R}}(y)$ [and $f_{\mathrm{N}}(y)$] denote the normal probability distribution function for the predicted relevant [and non-relevant, respectively] B-terms with means μ_{R} [and μ_{N}] and SDs σ_{R} [and σ_{N}]. (Note that 'predicted relevance' does not imply that a given B-term will be found relevant by all users in all cases, but rather by some user conducting a two-node search involving the given pair of articles.) The decomposition also directly gives a formula for estimating the probability that a B-term with a given score y is relevant:

$$\Pr\{\mathrm{R}|\mathrm{y}\} = pf_{\mathrm{R}}(y)/(pf_{\mathrm{R}}(y) + (1-p)f_{\mathrm{N}}(y))$$

Higher ranked B-terms have a greater predicted probability of being relevant, and the relationship is sigmoidal (Supplementary Fig. S1).

2.5 Does the model adequately and robustly represent the amount of implicit information linking two sets of articles?

We compared the gold-standard searches against a variety of other searches having a wide range of topical coherence, and including literature pairs with overlap. As a negative control, we constructed 10 random literature pairs (sets of papers randomly selected from MEDLINE) to match the sizes of each of the gold-standard searches (60 random searches in all). We also conducted PubMed queries on randomly selected MeSH terms having a moderate frequency in MEDLINE, from which we manually identified six pairs of literatures that we expected would have relatively little implicit information linking them (Supplementary Table S3). Finally, as a positive control expected to contain a large amount of complementary information, we selected a set of five similarly sized searches where the literature pairs consisted of closely related topics (Supplementary Table S4) (note that overlapping papers are removed before B-terms are computed).

The gold-standard searches have a significantly greater proportion of predicted relevant B-terms (mean ± SEM = 19.3 ± 1.3%) than do the searches expected to have little linking information (10.5 ± 1.3%, two-sample *t*-test $P < 0.001$), which in turn have a higher proportion of predicted relevant B-terms than do the searches involving random sets of papers (3.2 ± 1.7%, $P = 0.003$). The closely related topical literatures gave 33.8 ± 2.0% predicted relevant B-terms, which was significantly greater than that of the gold-standard searches ($P < 0.001$). Note that all of these *t*-tests are significant at the 0.05 level after Bonferroni correction for multiple tests. Figure 3 shows the B-term score histograms for representative two-node searches from each group. These findings suggest that the model does, indeed, capture relevant aspects of complementary information. Figure S2 shows the precision/recall curves for each of the gold-standard searches.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1658/223810 by guest on 26 January 2019

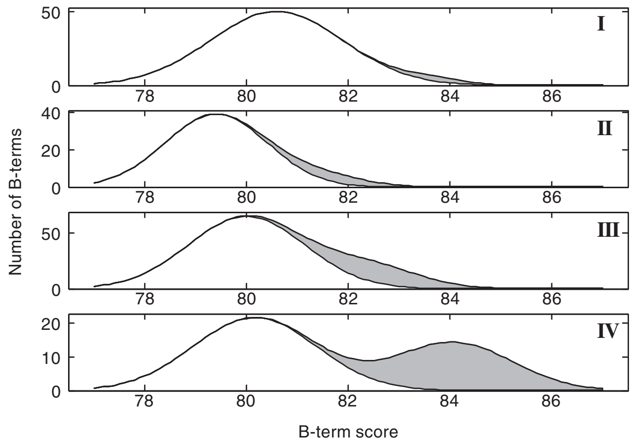

Fig. 3. Histograms of B-term scores for four different types of two-node searches (see text). (**I**) two sets of randomly selected articles, (**II**) a pair of cohesive but apparently unrelated literatures (mesothelioma versus authoritarianism), (**III**) one of the gold-standard searches (retinal detachment versus aortic aneurysm), and (IV) a pair of cohesive, closely related literatures (mesothelioma/etiology versus mesothelioma/physiology). Shaded areas indicate the proportion of predicted relevant B-terms.

2.6 Evaluating the performance of B-term ranking methods using two sets of queries

A strength of our model is that the field testers chose queries in the course of their scientific studies. However, to evaluate whether the results generalize to other types of biological queries, and to ensure that the model was not affected by subjective marking of relevant B-terms (which may differ among individuals and across different query needs), we have created an independent series of 20 gold standards based upon templated TREC 2005 Genomics Track queries (http://trec.nist.gov/data/t14_genomics.html; Hersh *et al*., 2005, 2006). These queries asked for information describing the role(s) of a gene involved in a disease, or describing the role of a gene in a specific biological process; as discussed below, these could be associated naturally with two-node searches (wherein lit A = gene and lit C = disease or biological process). Each query was searched within a biomedical text collection representing a subset of MEDLINE, and a group of judges decided which articles were relevant to the query (Hersh *et al*., 2005).

To adapt this effort to evaluating our two-node search model, we regarded the articles marked as relevant by TREC judges as 'gold standards' for each query, and extracted all terms in the titles of these papers. The terms were filtered through a stoplist (previously constructed by Don Swanson) to remove many of the 'uninteresting' terms (Swanson and Smalheiser, 1997; list available at http://arrowsmith.psych.uic.edu/arrowsmith_uic/data/stopwords_swanson), and the remaining terms were regarded as capturing some of the known, explicit information on each query. Next, we associated each query with a two-node search in which we formulated literature A = the gene name and literature C = the disease or biological process (removing any articles that mention both A and C). (A few queries which involved very large literatures were further restricted; e.g. a search on breast cancer was restricted to the PubMed query 'breast neoplasms/etiology' [majr]'.) Finally, for each of these two-node searches, the Arrowsmith system created a ranked list of the B-terms that represent the implicit information linking these two topics using the model discussed in the present article, calculated the percentage predicted as relevant to the query, and thus identified a set of B-terms that are predicted by the model as being relevant to the TREC query.

The explicit title terms taken from the gold-standard articles in the TREC queries serve the same function for evaluation as does the marked relevant B-terms in our own 6 gold-standard queries. For each set of queries, we used a standard measure of performance in information retrieval called mean average precision (MAP; Voorhees and Harman, 2005) in order to compare different methods of ranking B-terms: Given a list of B-terms, the precision is defined as the proportion of marked relevant B-terms. The average precision for a ranked list of B-terms is defined by first computing, for each marked relevant term, the precision of the B-terms ranked the same or above the marked relevant B-term, and then taking the average across these precision values. The mean average precision is defined as the mean of the average precision values across the set of all queries. Because some of the explicit terms extracted from the TREC gold-standards may represent 'noise' rather than meaningful links, and because the two-node searches are formulated in an automatic fashion, the mean average precision that is obtained for the set of TREC queries has little meaning on its own. Nevertheless, one can utilize this approach to compare different methods of ranking the B-terms against each other: comparing two proposed methods of ranking B-terms in a two-node search, the better method is one which ranks the explicit title terms more highly, as measured by having a higher mean average precision across all terms and across all queries.

We compared the logistic regression model against two previously proposed mutual information measures (MIM)—the minimum MIM and the average MIM (Wren, 2004). Before doing so, however, we performed 6-fold cross-validation of our model as follows: average precision is computed for each of the six gold-standard two-node searches where the B-term score is based on the feature weights estimated using the remaining five gold-standards. The resulting cross-validated mean average precision represents an unbiased performance estimate, avoiding possible bias that would result if the model is both trained and tested on the same dataset.

As shown in Table 2, our model exhibited almost twice the mean average precision of either MIM model based on TREC queries (21.0% versus 12.5% and 14.0%), differences that were highly significant ($P < 0.0001$). The results were similar to that observed using the gold-standard two-node searches (25.2% versus 13.3% and 12.0%) (Table 2). Thus, the logistic regression model outperformed the MIM models as evaluated on two independent sets of queries—our six gold-standard searches and the 20 TREC queries.

2.7 Assessing the contribution of different features within the logistic regression model

As mentioned in Section 3, each individual feature within the logistic regression model showed a statistically significant association with B-term relevance. To analyze how each feature contributes to the overall performance of the model, we first

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1658/223810 by guest on 26 January 2019

Table 2. Comparison of B-term ranking methods. The performance of each ranking method is summarized by its mean average precision (MAP) for a set of two-node searches, and compared against the MAP for the B-term Score from the logistic regression model using 6-fold cross-validation. The mutual information measures (MIM) are described in the text. Each *P*-value corresponds to one-sided paired *t*-test of MAP ratios >1

	MAP (%) two-node searches	*P*-value	MAP (%) TREC Queries	*P*-value ≤
B-term Score logistic regression model	27.4		21.0	
B-term Score 6-fold cross-validation	**25.2**		**21.0**	
Average MIM model	13.3	0.0062	14.0	0.0001
Minimum MIM model	12.0	0.0056	12.5	0.0001
x_1: $N_{AB}>1$, $N_{BC}>1$	5.0	0.0007	8.2	0.0001
x_2: MeSH in common	4.7	0.0007	7.5	0.0001
x_3: mapped to UMLS	5.4	0.0014	7.9	0.0001
x_4: cohesion	16.3	0.0241	12.9	0.0001
x_5: frequency in MEDLINE	11.0	0.0168	7.8	0.0001
x_6: 1st year in MEDLINE	18.0	0.0730	10.6	0.0001
x_7: *P*-value of $N_{AB}+N_{BC}$	10.0	0.0047	16.9	0.0038
Global features (x_3, x_4, x_5, x_6)	20.8	0.0144	14.4	0.0001
Local features (x_1, x_2, x_7)	9.5	0.0043	17.5	0.0143

examined the predictive power of each individual feature as measured by its maximum odds ratio (i.e. the ratio of the greatest odds to the lowest odds of a B-term being relevant observed within the set of six gold-standard queries). The features ordered by decreasing predictive power (maximum odds ratio) are: (1) x_4, literature cohesion score (30.7); (2) x_6, year of first occurrence in MEDLINE (8.5); (3) x_7, frequency in A or C relative to frequency in MEDLINE (4.5); (4) x_5, overall count in MEDLINE (4.3); (5) x_3, mapping to a semantic category (3.7); (6) x_2, MeSH in common in AB and BC (2.7): and (7) x_1, multiple occurrences in A and C (2.1). Note that the cumulative effect of two or more features is not additive but rather measured by the product of the maximum odds ratios. Thus, no single feature measures up to the cumulative predictive power of all the features combined (e.g. the maximum odds ratio observed in the six gold-standard searches = 19 458).

Second, we examined how each feature contributed to the overall mean average precision using both the 6 gold-standard queries and the 20 TREC queries (Table 2). Similar results were observed using both sets of queries: (1) The literature cohesion score alone accounts for less than 2/3 of the MAP, and in fact, all features contribute substantially so that none can be disregarded. (2) The frequency-based features of the logistic regression model (x_5 and x_7) give similar performance as the mutual information measures, again giving substantially worse performance than when using all features. (3) We hoped that the global features would achieve most, if not all, of the performance of the entire model, since this would mean that one could precompute B-term scores for each B-term regardless of the specific literatures A and C involved in a given two-node search. However, neither the subgroup of global (x_3, x_4, x_5, x_6) nor local (x_1, x_2, x_7) features taken together performed well on their own (Table 2).

2.8 Assessing the robustness of the model

To verify that small, incidental differences in the way that a PubMed query is formulated would not greatly affect the predicted proportion of relevant B-terms, we ran different variations of the sixth gold-standard search (calpain versus postsynaptic) in which the calpain literature was formulated in PubMed queries restricted to titles, MeSH terms, and/or text words. The predicted proportions showed relatively small variation across searches, ranging from 17.25% to 19.75%.

Do six gold-standard searches suffice to estimate B-term relevance probabilities reliably? We carried out a 6-fold cross-validation in which the weight parameters were estimated using five of the six gold-standard searches, and computed the average precision on the remaining search. The mean average precision was only slightly less when based on five versus six gold-standards (25.2% versus 27.4%; Table 2), suggesting that the performance values have largely converged. Also, adding more than six searches in the model is expected to produce only minor improvement in the weight estimates (Supplementary Fig. S3). This reflects the fact that the features used to characterize the B-terms are quite general (e.g. term frequency), have large individual effects, and complement each other.

3 DISCUSSION

Identifying information that implicitly links two disparate sets of articles is a fundamental and intuitive data mining strategy that can address real scientific questions. The Arrowsmith system finds title terms (so-called B-terms) that are shared across two literatures within MEDLINE and displays them in a manner that facilitates human assessment. Here, we describe a quantitative model that permits one to estimate the probability of relevance for each B-term, and to rank all B-terms among themselves according to their likely relevance. This greatly simplifies and streamlines the process of carrying out a two-node search on the Arrowsmith website (http://arrowsmith.psych.uic.edu). Each two-node search is unique insofar as it is the individual scientist who ultimately decides which, if any, B-terms indicate a meaningful or useful link between two sets of articles. Despite this, the model indicates that B-terms marked as relevant do tend to share certain generic features across diverse two-node searches.

The present formulation of the model employs seven complementary features, and outperforms other models based on frequency-based mutual information measures. We were surprised that such good performance was obtained using a model that was trained on only six gold-standard two-node searches. However, the model appears to be robust since it performed well using an independent set of 20 queries and automatically-extracted gold standard terms derived from the Genomics 2005 TREC shared task, and the results of 6-fold cross-validation suggest that performance has largely

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1658/223810 by guest on 26 January 2019

converged after six gold-standard searches. The performance is likely due to a number of factors: (1) The chosen B-term features are very general (and not specific to types of terms such as genes, diseases, etc.) and have large effects; (2) the gold-standard searches were 'well formulated' (see subsequently); (3) the total number of B-terms and the number of marked relevant B-terms were roughly balanced across the two-node searches, and B-terms were marked according to two different query needs in several searches, which helped reduce any bias towards any single two-node search; and (4) the model sought to find weights that optimally separate the set of marked relevant B-terms against a large and rich data set (the 'mixed' set), rather than against the smaller set of marked non-relevant terms.

Nevertheless, we intend to train the model using a larger number of more diverse gold-standard two-node searches that cover a wider range of topics, literature size and literature coherence. It is also likely that incorporating additional features related to conceptual language processing (Cohen and Hersh, 2005) may improve overall performance further: for example, it may be worthwhile to give differential weight to B-terms that are two or three word phrases (Nakagawa and Mori, 1998), noun phrases (Bennett *et al.*, 1999), abbreviations (Zhou *et al.*, 2006), that correspond to standard biomedical terminology (Bodenreider, 2004), or that correspond to particular linguistic functions such as the subject or object of a sentence (Ohta *et al.*, 2006). As well, B-terms that refer to the same thing (spelling variants, synonyms or abbreviations and their corresponding long forms) should ideally be merged and considered together. It may also be worthwhile to take B-terms not only from the titles of papers, but from their abstracts as well, because the abstract conveys a significant portion of the total information contained in a scientific paper (Kostoff *et al.*, 2004; Schuemie *et al.*, 2004; Shah *et al.*, 2003).

Besides providing a means of ranking individual B-terms, the logistic regression model also provides an estimate of the total implicit information linking two sets of articles—at least for well defined, 'well formulated' two-node searches where the two literatures are topically coherent and relatively small, where they have minimal or no overlap, and where the query need is clearly enunciated. This concept should be widely applicable to a number of other literature-based discovery applications, including the so-called one-node search which takes a single set of articles as a starting point (e.g. dealing with a scientific problem such as a disease) and searches for an unknown, disparate literature that contains information that may assist in solving the problem (e.g. describing a therapeutic agent that may potentially treat the disease but has not previously been described in this context) (Hristovski *et al.*, 2005; Srinivasan, 2004; Swanson and Smalheiser, 1997; Swanson *et al.*, 2006; Wren, 2004; Wren *et al.*, 2004; Weeber *et al.*, 2005; Yetisgen-Yildiz and Pratt, 2006). Several related gene-centric strategies also incorporate implicit links to predict how genes may be related to each other and to human diseases (Chen and Sharp, 2004; Homayouni *et al.*, 2005; Jenssen *et al.*, 2001). Before applying the logistic regression model to the one-node search problem, further research is needed to learn whether the model needs to be modified for literature pairs that are large and/or asymmetrical in size, overlapping or correlated in some manner (such as both sets consisting of papers by the same author), or non-topically organized (e.g. one set of articles might consist of papers written by authors based in California). Nevertheless, the present model should encourage much wider exploration of text mining for implicit information among the general scientific community.

4 METHODS

4.1 Selecting features and fitting the logistic regression

Each B-term was initially characterized according to eight complementary features (see Results section; Supplementary Table S1). Once the features had been normalized and scored (Supplementary Table S1), each B-term with a feature vector $\mathbf{x} = <x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8>$ was formulated as a weighted sum of the individual feature values:

$$\text{B-term score} : y(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_8 x_8,$$

where x_i denotes the value of the *i*-th feature value, and w_i denotes the corresponding weight. After the stoplist feature (x_8) was removed (see Results section), the final model contained seven features. In order to estimate the weights, the average probability of feature vector **x** being marked relevant was written as a logistic regression function as follows:

$$\Pr\{R|x\} = \frac{1}{1 + e^{-(a_1 I_1 + \cdots + a_6 I_6 + w_1 x_1 + \cdots + w_7 x_7)}},$$

where $\mathbf{x} =$ the B-term vector of the remaining feature values $<x_1, x_2, x_3, x_4, x_5, x_6, x_7>$, $I_j =$ indicator variable for search j ($= 1$ for search j, 0 otherwise) and $a_j =$ intercept for search j. To check whether any single pair of features exhibited an interactive or quadratic effect, each feature was examined separately. None of the non-linear effects were statistically significant when using a two-sided *t*-test at the 0.05 level of significance using the Bonferroni correction for multiple tests. We also confirmed that lack of statistical significance was not simply due to variance-inflation (Hosmer and Lemeshow, 2000).

4.2 Decomposing B-term scores into two normal distributions

A χ^2 goodness of fit method (Agresti, 2002) was employed to decompose the distribution of B-term scores into a mixture of two normal bell-shape distributions. The mixture can be expressed as follows: $f(y) = p f_R(y) + (1-p) f_N(y)$, where y is the B-term score, p denotes the predicted proportion of relevant B-terms, and $f_R(y)$ [and $f_N(y)$] denote the normal probability density function for the predicted relevant [and non-relevant, respectively] B-terms with a means μ_R [and μ_N] and SDs σ_R [and σ_N]. The mixture had to satisfy three conditions that reflect the observed nature of the B-term score histograms:

(1) $\sigma_N \geq \sigma_R$. The non-relevant distribution accounts for most of the variation in B-term scores.

(2) $|\mu_R - \mu_N| \geq 2.576\sigma_R$. At most 0.5% of the predicted relevant distribution can occur below the mean of the non-relevant distribution. This ensures that a set of B-terms consisting almost exclusively of non-relevant terms is not decomposed into two nearby and almost identical distributions, which could lead to 50% predicted proportion of relevant terms. The number 2.576 was set by empirical observations.

(3) $f_R(y)/f_N(y)$ is non-decreasing in y for all observed y. This forces the probability of being relevant to increase with the observed B-term scores.

Given a set of n B-term scores $y_1, y_2, \ldots, y_n$ the optimal combination of parameters μ_R, σ_R, μ_N, σ_N, and p was computed as follows: first the overall mean (μ) and SD (σ) of the B-term scores were calculated. Then, the B-term scores were partitioned into 20 equally spaced bins centered

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1658/223810 by guest on 26 January 2019

at $y_1, y_2, \ldots, y_{20}$, giving an observed B-term count $g(y_i)$ for each bin i. Finally, the values for the parameters μ_R, σ_R, μ_N, σ_N and p were computed using a grid search to solve the following optimization problem:

$$\text{minimize } \chi^2 = \sum\nolimits_{i=1,2,\ldots,20} [(g(y_i) - nf(y_i))^2/nf(y_i)]$$

subject to

$$\begin{aligned}
&\mu = p\mu_R + (1-p)\mu_N \\
&\sigma^2 = p(\sigma_R^2 + (\mu_R - \mu)^2) + (1-p)(\sigma_N^2 + (\mu_N - \mu)^2) \\
&\sigma_N \geq \sigma_R \\
&|\mu_R - \mu_N| = 2.576\sigma_R \\
&f(y_i{+}1)/f(y_{i+1}) \geq f(y_i)/f(y_i), \qquad \text{for i} = 1, \ldots, 19.
\end{aligned}$$

4.3 Support for public users on the Arrowsmith website

To retrieve MEDLINE records corresponding to user queries quickly and automatically, we (1) integrated the PubMed query interface with our system and (2) set up a local customized database of MEDLINE, updated weekly. When a query is entered, the article ID numbers are downloaded from PubMed and the full records retrieved from the local database, including a tokenized and stoplisted version of each article title. Articles not found in the local database are immediately downloaded from PubMed as XML files, processed and stored in the local database. B-terms and their feature values were 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 was done. The baseline 2005 version of MEDLINE was processed to identify all terms (words and up to three word phrases) in titles and to pre-compute global features (x_3, x_4, x_5, x_6) that were the same for a B-term no matter how the A and C literatures were defined.

ACKNOWLEDGEMENTS

This research is supported by NIH Grants LM007292 and LM08364. Thanks to Ramin Homayouni, Don Kalar and Guy Perkins for assistance with the gold-standard two-node searches. We also thank Kevin Bretonnel Cohen and Larry Hunter for suggesting that we use TREC queries to evaluate and compare the performance of the ranking methods. Funding to pay the Open Access publication charges was provided by NIH grants.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (2002) *Categorical Data Analysis*. 2nd edn. Wiley, New York.

Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp*., 17–21.

Bennett,N.A. *et al*. (1999) Extracting noun phrases for all of MEDLINE. *Proc. AMIA Symp*., 671–675.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*., **32**, D267–D270.

Bowden,D.M. and Dubach,M.F. (2003) NeuroNames 2002. *Neuroinformatics*, **1**, 43–59.

Chen,H. and Sharp,B.M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, **5**, 147.

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

Hersh,W. *et al*. (2005). TREC 2005 Genomics track overview. In Voorhees,E.M. and Buckland,L.P. (eds.). *NIST Special Publication 500-266: The Fourteenth Text Retrieval Conference (TREC 2005)*. National Institute of Standards and Technology, Gaithersburg, pp. 25–50.

Hersh,W.R. *et al*. (2006). Enhancing access to the Bibliome: the TREC 2004 Genomics Track. *J. Biomed. Discovery Collaboration*, **1**, 3.

Homayouni,R. *et al*. (2005) Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics*, **21**, 104–115.

Hosmer,D.W. and Lemeshow,S. (2000) *Applied Logistic Regression*. 2nd edn. Wiley, New York.

Hristovski,D. *et al*. (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform*., **74**, 289–298.

Hunter,L. and Cohen,K.B. (2006) Biomedical language processing: what's beyond PubMed? *Mol. Cell*, **21**, 589–594.

Jensen,L.J. *et al*. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet*., **7**, 119–129.

Jenssen,T.K. *et al*. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet*., **28**, 21–28.

Kostoff,R.N. *et al*. (2004) Information content in Medline record fields. *Int. J. Med. Inform*., **73**, 515–527.

Krallinger,M. and Valencia,A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol*., **6**, 224.

Li,X. and Liu,B. (2005) Learning from positive and unlabeled examples with different data distributions. In Gama,J., *et al*. (eds.) *Proceedings of the 16th European Conference Machine Learning (ECML 2005)*. *Lecture Notes in Artificial Intelligence*. Vol. 3720. Springer-Verlag Press, Berlin. pp. 218–229.

McCray,A.T. *et al*. (2001) Aggregating UMLS semantic types for reducing conceptual complexity. *Medinfo*., **10**, 216–220.

Nakagawa,H. and Mori,T. (1998) Nested collocation and compound noun for term recognition. *Proceedings of Workshop on Computational Terminology*, Montreal, Canada. pp. 64–70.

Nigam,K. *et al*. (2000) Text classification from labeled and unlabeled documents using EM. *Mach. Learn*., **39**, 103–134.

Ohta,T. *et al*. (2006) An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australia. pp. 17–20.

Schuemie,M.J. *et al*. (2004) Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, **20**, 2597–2604.

Shah,P.K. *et al*. (2003) Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, **4**, 20.

Smalheiser,N.R. (2005) The Arrowsmith project: 2005 status report. In Hoffmann,A. *et al*. (ed.) *Discovery Science 2005*. *Lecture Notes in Artificial Intelligence*. Vol. 3735, Springer-Verlag Press, Berlin, pp. 26–43.

Smalheiser,N.R. and Swanson,D.R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed*., **57**, 149–153.

Smalheiser,N.R. *et al*. (2006) Collaborative development of the Arrowsmith two-node search interface designed for laboratory investigators. *J. Biomed. Discovery Collaboration*, **1**, 8.

Srinivasan,P. (2004) Text mining: generating hypotheses from MEDLINE. *J. Am. Soc. Information Sci. Technol*., **55**, 396–413.

Swanson,D.R. and Smalheiser,N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Int*., **91**, 183–203.

Swanson,D.R. *et al*. (2006) Ranking indirect connections in literature-based discovery: The role of Medical Subject Headings (MeSH). *J. Am. Soc. Information Sci. Technol*., **57**, 1427–1439.

Tanabe,L. and Wilbur,W.J. (2004) Generation of a large gene/protein lexicon by morphological pattern analysis. *J. Bioinform. Comput. Biol*., **1**, 611–626.

Voorhees,E.M. and Harman,D.K. (2005) *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA.

Weeber,M. *et al*. (2005) Online tools to support literature-based discovery in the life sciences. *Brief Bioinform*., **6**, 277–286.

Wren,J.D. (2004) Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, **5**, 145.

Wren,J.D. *et al*. (2004) Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, **20**, 389–398.

Yetisgen-Yildiz,M. and Pratt,W. (2006) Using statistical and knowledge-based approaches for literature-based discovery. *J. Biomed. Inform*., **39**, 600–611.

Zhou,W. *et al*. (2006) ADAM: Another database of abbreviations in MEDLINE. *Bioinformatics*, **22**, 2813–1818.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/1658/223810 by guest on 26 January 2019