Formulating MEDLINE queries for article retrieval based on PubMed exemplars

Alexander Garnett[1,2], Heather A. Piwowar[1,3],

Edie M. Rasmussen[2] and Judy Illes[1]

The University of British Columbia
1. National Core for Neuroethics and
2. School of Library, Information, and Archival Studies

3. The University of Pittsburgh Department of Biomedical Informatics

Bibliographic search engines allow endless possibilities for building queries based on specific words or phrases in article titles and abstracts, indexing terms, and other attributes. Unfortunately, deciding which attributes to use in a methodologically sound query is a non-trivial process. In this paper, we describe a system to help with this task, given an example set of PubMed articles to retrieve and a corresponding set of articles to exclude. The system provides the users with unigram and bigram features from the title, abstract, MeSH terms, and MeSH qualifier terms in decreasing order of precision, given a recall threshold. From this information and their knowledge of the domain, users can formulate a query and evaluate its performance. We apply the system to the task of distinguishing original research articles of functional magnetic resonance imaging (fMRI) of sensorimotor function from fMRI studies of higher cognitive functions.

Background

Although the classification of abstracts in PubMed has been studied extensively, there are few tools to help end users develop effective classification queries for use in PubMed. Several tools exist to illustrate relative recall of features, but these only provide results for a single query, rather than differential attributes between two queries. For example, the “Anne O' Tate” interface developed as part of Arrowsmith project (Smalheiser et al., 2008) allows for detailed drill-down of a single query. Plikus et al.'s PubFocus (2006) provides citation analytics and sorting by impact factor, but lacks for any means of comparison. Similarly, the PubAtlas tool maintained by UCLA's Consortium for Neuropsychiatric Phenomics is invaluable for visualizing *associations* between data sets, but does not include any means of segregating one from another.

We propose a method to suggest query components given a user-provided list of true positive and true negative PubMed identifiers. We recently developed a system to extract features from full text of open access articles, for query execution in existing full-text portals like PubMed Central, HighWire Press, and Google Scholar (Piwowar and Chapman, 2010). Here, we instead present the precision and recall of various MEDLINE features for use in a PubMed query. The current implementation evaluates unigram and bigram features of the article title and abstract, as well as medical subject heading (MeSH) indexing terms, MeSH major terms, MeSH qualifiers, and MeSH major qualifiers.

To evaluate the efficacy of this approach to query-building, we applied it to the task of identifying research articles of functional magnetic resonance imaging (fMRI) of sensorimotor function as distinct from fMRI studies of other cognitive functions.

Method

Query development features

We began with a set of fMRI research articles over the period 1991-2001 which had been manually curated based on the degree of cognitive function under observation (Illes et al., 2010).

We employed 62.5% of these features as a development corpus. Using the NCBI's Entrez Programming Utilities (eUtils) (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) in Python version 2.6, we supplied the PubMed identifiers of our positive and negative examples, then downloaded their titles, abstracts, and MeSH indexing terms. To assemble unigram and bigram features for the abstracts and titles, we split the text on whitespace and all punctuation except hyphens. We excluded any unigram or bigram that included a word less than 3 characters long, more than 30 characters long, that did not include at least one alphabetic character, or represented a PubMed stop word We also cataloged the MeSH terms, the major MeSH terms, the MeSH qualifiers, and the major MeSH qualifiers.

Query development algorithm

We immediately disqualified unigrams and bigrams that did not have at least 10% precision and 10% recall in our development corpus. We then utilized our domain knowledge to exclude

features that were sufficiently ambiguous across the positive and negative data sets (e.g. the unigram “cortex”), as well as those that were not germane to the intent of the query (e.g. the unigram “word,” despite its having a particularly high recall among a set of articles which study language in the brain), and considered generalizations of MeSH terms where appropriate. We used a simple technique to build our own binary rules, as illustrated in Figure 1.

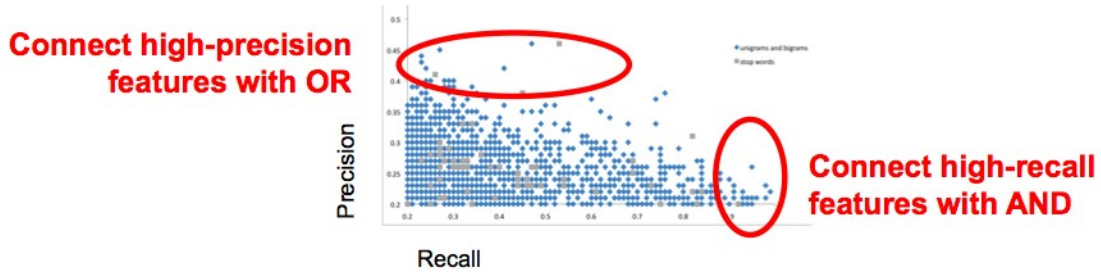

Figure 1: Method for building Boolean query from feature list. In query syntax:

(*(features with highest recall joined with* ***AND****)*
AND *(features with highest precision joined with* ***OR****)****)***

We considered NOT phrases through a manual error analysis of the false positives in the development set, and by reversing the labels of positive and negative examples to identify features that identified the negative instances with high precision and recall. The aim for our intended use-case was a query with both a precision and recall over 60%.

Query evaluation and implementation

To evaluate the performance of the queries, we calculated the precision, recall, and harmonized f-measure of the full queries across the test samples (the PubMed positive and negative examples that were not used for development). We also tested our full queries against naïve MeSH terms which were expected to have a near-universal penetration across our test set, such as “Magnetic Resonance Imaging”. These MeSH terms were found to describe our dataset with over 95% accuracy, allowing for generalizations over the MeSH hierarchy.

Results

Queries

We applied our query-formulation approach to the task of identifying research articles of fMRI of sensorimotor function as distinct from fMRI studies of other cognitive functions. The top 20 features identified by our approach, sorted by precision, are displayed in Table 1.

Table 1: The top 20 features, sorted by precision

Query Feature	Precision	Recall	F
movements[ti]:	0.87	0.12	0.21
stimulation[ti]:	0.72	0.11	0.19
visual_cortex[ti]:	0.7	0.11	0.2
motor[ti]:	0.67	0.2	0.31
maps[abstract]:	0.65	0.13	0.22
finger[abstract]:	0.65	0.19	0.29
contralateral[abstract]:	0.64	0.18	0.28
Movement[mesh]:	0.63	0.18	0.28
Fingers[mesh]:	0.63	0.15	0.24
primary_motor[abstract]:	0.62	0.17	0.27
blood_oxygenation[abstract]:	0.62	0.11	0.18
oxygenation[abstract]:	0.61	0.11	0.19
stimulation[abstract]:	0.6	0.36	0.45
primary_visual[abstract]:	0.6	0.11	0.18
hand[abstract]:	0.59	0.19	0.29
field[abstract]:	0.58	0.15	0.24
flow[abstract]:	0.58	0.11	0.19
Hand[mesh]:	0.57	0.1	0.17
motor_cortex[abstract]:	0.56	0.19	0.28
primary[abstract]:	0.56	0.4	0.46

We derived the queries in Table 2 for the identification of basic sensorimotor functions:

Table 2: Queries

Consideration	Query
High precision, for the identification of original fMRI research	("humans"[mesh] AND "magnetic resonance imaging"[mesh] AND Journal Article[ptyp] NOT "mental disorders"[mesh] AND ("fmri"[Title/Abstract] OR "Functional MRI"[Title/Abstract] OR "Functional magnetic resonance imaging"[Title/Abstract] OR "Functional MR Imaging"[Title/Abstract])) NOT (Editorial[ptyp] OR Letter[ptyp] OR Meta-Analysis[ptyp] OR Practice Guideline[ptyp] OR Review[ptyp] OR Case Reports[ptyp] OR Comment[ptyp] OR Corrected and Republished Article[ptyp]) ("1991"[PDAT] : "2009"[PDAT])AND English[lang])
High recall, for the identification of studies of basic sensorimotor function	(Somatosensory Cortex[mesh]) OR somatosensory[Title/Abstract] OR "primary motor"[Title/Abstract] OR "primary visual"[Title/Abstract] OR sensorimotor[Title/Abstract] OR "motor area"[Title/Abstract] OR oxygenation[Title/Abstract] OR (Motor Cortex[mesh]) OR "visual cortex"[Title/Abstract] OR (Acoustic Stimulation[mesh])) NOT (memory[Title/Abstract] OR Memory[mesh] OR (Prefrontal Cortex[mesh]) OR Cognition[mesh] OR prefrontal[Title/Abstract] OR dorsolateral[Title/Abstract])

Query performance

We compare the results of the derived query to two naïve queries based on Medical Subject Heading (MeSH) terms. As seen in Table 3, the derived query had better precision than either of the MeSH queries at an acceptable recall for our intended task.

Table 3: Comparison to MeSH queries

	N	precision	recall	f-measure
("magnetic resonance imaging"[mesh] OR "	166	26%	98%	41%

somatosensory cortex"[mesh] OR "motor cortex"[mesh])				
("magnetic resonance imaging"[mesh] AND "somatosensory cortex"[mesh] AND "motor cortex"[mesh])	166	55%	7%	12%
Derived query	166	59%	61%	60%

Discussion

We described a simple mechanism for formulating effective queries for use PubMed, provided a set of example true positives and true negatives. As a proof of concept, we applied this approach to a task that was previously performed by manual annotation: identifying research articles of fMRI of sensorimotor function, as distinct from fMRI studies of other cognitive functions. The query we derived achieved 59% precision and 61% recall, making it a better fit for our intended application than lower-precision baseline MeSH queries. Although the evaluation demonstrates the usefulness of this approach only in the context of one neuroethics research task, we believe this end-user method for deriving comprehensive PubMed queries is widely applicable.

Effectively querying is difficult: Synonyms, variant spellings, acronyms, and inexperience make it difficult to form effective queries (Beall, 2008). Our approach employs empirically sound information retrieval measures, yet benefits from not being fully automated. In this way, it can function as a decision support tool alongside users' domain knowledge in excluding undesired or irrelevant features. Users can also generalize appropriately by considering stemmed alternatives or using indexed terms at a higher level of the MeSH hierarchy.

We believe that this method is especially valuable for its ability to generated an automated query based on manual annotations. In this respect, it can act as an internal validation mechanism, with eventual query refinements providing a positive feedback loop. Although we have used it to monitor trends over a data set which was only annotated up to a particular time, it could be equally valuable to a cross-sectional study design. This query development method offers several advantages: It is easy to maintain, its implementation is free and open, it is extensible, and the user can be in direct control of recall/precision balance by setting recall and precision thresholds. However, it does have several limitations. While this system was built with a degree of overlap between automated and manual filtering in mind, it requires an admittedly careful eye for detail, as well as repeated testing, to ensure that no undesired elements are included in a derived query.

The system could be expanded in many ways. Its could take seed queries for input, rather than PubMed IDs. Active learning might allow for further refinement. The system could run parts-of-speech analysis or domain-specific named entity recognition on the development abstracts, if that helped to identify valuable features. The system could be enhanced to use bootstrapping to identify phrase variants (Abdalla & Teufel, 2006). Also, because some portals have some wildcard capabilities, we would like to experiment with learning regular expressions (Wu & Pottenger), though there is some evidence that this may not help (Carpenter).

Future work might expand this system's to retrieve other attributes of MEDLINE metadata, such as journal or author names. It would also be possible to use the system to evaluate a list of PubMed limits and subsets, e.g. bioethics[sb], on their precision and recall, where appropriate. Additionally, the system could be expanded to generate features appropriate for other databases like Scopus or Ovid, given their respective stemming and stopword implementations.

To better understand the relative strengths and weaknesses of this approach, it would be informative to compare its performance to other systems and algorithms on a standard task, such as the TREC Genomics corpus (Rekapalli et al., 2006), or a query that has been developed solely from article abstracts (Aphinyanaphongs et al., 2006). Still, we are confident that this new method of information retrieval marks an impressive leap forward for end-user formulation of complex biomedical search queries. The degree to which it allows a user to monitor its progress in revising a query works remarkably well in tandem with qualitative analysis of a data set, and in so doing, preserves a human element in quantitative informatics.

Acknowledgments

Generously supported by NIH/NIMH RO1 #R01MH84282-05, CIHR/INMHA CNE #85117, and the British Columbia Knowledge Development Fund.

Availability

Code will be openly available at http://ww.researchremix.org prior to formal publication of this study, and will be made available immediately (in its current, under-documented state) to anyone who contacts the authors directly.

References

Abdalla, R. M., & Teufel, S. (2006). A bootstrapping approach to unsupervised detection of cue phrase variants. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS* (Vol. 44, p. 921).

Aphinyanaphongs, Y., Statnikov, A., & Aliferis, C. F. (2006). A Comparison of Citation Metrics to Machine Learning Filters for the Identification of High Quality MEDLINE Documents. *Journal of the American Medical Informatics Association*, *13*(4), 446-455. doi:10.1197/jamia.M2031

Beall, J. (2008). The Weaknesses of Full-Text Searching. *The Journal of Academic Librarianship*, *34*(5), 438–444.

Plikus, M., Zhang, Z., & Chuong, C. (2006). PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC Bioinformatics*, *7*(1), 424. doi:10.1186/1471-2105-7-424

Piwowar, H., and Chapman, W. (2010). Using open access literature to guide full-text query formulation. Available from *Nature Precedings* http://hdl.handle.net/10101/npre.2010.4267.2.

Rekapalli, H. K., Cohen, A. M., & Hersh, W. R. (2007). A comparative analysis of retrieval features used in the TREC 2006 Genomics Track passage retrieval task, *2007*, 620-624.

Smalheiser, N. R., & Swanson, D. R. (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, *57*(3), 149–153.

Smalheiser, N., Zhou, W., & Torvik, V. (2008). Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. *Journal of Biomedical Discovery and Collaboration*, *3*(1), 2. doi:10.1186/1747-5333-3-2

Wu, T., & Pottenger, W. M. (n.d.). A semi-supervised active learning algorithm for information extraction from textual data. *Age*, *100*(94.8), 97–33.