Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

Using open access literature to guide full-text query formulation

Heather A. Piwowar and Wendy W. Chapman

Background

Much scientific knowledge is contained in the details of the full-text biomedical literature. Most research in automated retrieval presupposes that the target literature can be downloaded and preprocessed prior to query. Unfortunately, this is not a practical or maintainable option for most users due to licensing restrictions, website terms of use, and sheer volume. Scientific article full-text is increasingly queriable through portals such as PubMed Central, Highwire Press, Scirus, and Google Scholar. However, because these portals only support very basic Boolean queries and full text is so expressive, formulating an effective query is a difficult task for users. We propose improving the formulation of full-text queries by using the open access literature as a proxy for the literature to be searched. We evaluated the feasibility of this approach by building a high-precision query for identifying studies that perform gene expression microarray experiments.

Methodology and Results

We built decision rules from unigram and bigram features of the open access literature. Minor syntax modifications were needed to translate the decision rules into the query languages of PubMed Central, Highwire Press, and Google Scholar. We mapped all retrieval results to PubMed identifiers and considered our query results as the union of retrieved articles across all portals. Compared to our reference standard, the derived full-text query found 56% (95% confidence interval, 52% to 61%) of intended studies, and 90% (86% to 93%) of studies identified by the full-text search met the reference standard criteria. Due to this relatively high precision, the derived query was better suited to the intended application than alternative baseline MeSH queries.

Significance

Using open access literature to develop queries for full-text portals is an open, flexible, and effective method for retrieval of biomedical literature articles based on article full-text. We hope our approach will raise awareness of the constraints and opportunities in mainstream full-text information retrieval and provide a useful tool for today's researchers.

Background

Much scientific information is available only in the full body of a scientific article. Full-text

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

biomedical articles contain unique and valuable information not encapsulated in titles, abstracts, or indexing terms. Literature-based hypothesis generation, systematic reviews, and day-to-day literature surveys often require retrieving documents based on information in full-text only.

Progress has been made in accurately retrieving documents and passages based on their full-text content. Research efforts, relying on advanced machine-learning techniques and features such as parts of speech, stemmed words, n-grams, semantic tags, and weighted tokens, have focused on situations in which complete full-text corpora are available for preprocessing. Unfortunately, most users do not have an extensive, local, full-text library. Establishing and maintaining a machine-readable archive involves complex issues of permissions, licenses, storage, and formats. Consequently, applying cutting-edge full-text information retrieval and extraction research is not feasible for mainstream scientists.

Several portals offer a simple alternative: PubMed Central, Highwire Press, Scirus, and Google Scholar provide full-text query interfaces to an increasingly large subset of the biomedical literature. Users can search for full-text keywords and phrases without maintaining a local archive; in fact, they need not have subscription nor access privileges for the articles they are querying. Portals return a list of articles that match the query (often with a matching snippet). Users can manually review this list and download articles subject to individual licensing agreements.

It is difficult, however, to formulate an effective query for these portals: Full-text has so much lexical variation that query terms are often too broad or too narrow. This standard information retrieval problem has been extensively researched for queries based on titles, abstracts, and indexing terms. Much less research has been done on query expansion and refinement for full-text. Today's full-text portals offer very basic Boolean query interfaces only, with little support for synonyms, stemming, n-grams, or "nearby" operations.

We suggest that open access literature can help users build better queries for use within full-text portals. An increasingly large proportion of the biomedical literature is now published in open access journals such as the BMC family, PLoS family, Nucleic Acids Research, and the Journal of Medical Internet Research . Papers published in these journals can be freely downloaded, redistributed, and preprocessed by anyone for any purpose. Furthermore, the NCBI provides a daily zipped archive of biomedical articles published by most open access publishers in a standard format, making it easy to establish and maintain a local archive of this content. If a proposed seed query has sufficient coverage, we believe that the open access literature could provide valuable information to expand and focus the query when it is applied to the general literature though established full-text portals.

We propose a method to facilitate the retrieval of biomedical literature through full-text queries run in publicly accessible interfaces. In this initial implementation, users provided a list of true positive and true negative PubMed identifiers within the open access literature. Standard text mining techniques were used to generate a query that accurately retrieved the documents based

on the provided examples. We chose text-mining techniques that resulted in query syntax that was compatible with full-text portal interfaces, such as Boolean combinations, n-grams, wildcards, stemming, and stop words. The returned query was ready to be run through the simple interfaces of existing, publicly available full-text search engines. Full-text document hits could then be manually reviewed and downloaded by the user, subject to article subscription restrictions.

To evaluate the feasibility of this query-development approach, we applied it to the task of identifying studies that use a specific biological wet-laboratory method: running gene expression microarray experiments.

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

Method

Query development corpus

To assemble articles on the general topic of interest, we used the title and abstract filter proposed by Ocshner et al. . We limited our results to those in the open access literature by running the following PubMed query:

> **"open access"[filter]** AND (microarray[tiab] OR microarrays[tiab] OR genome-wide[tiab] OR "expression profile"[tiab] OR "expression profiles"[tiab] OR "transcription profile"[tiab] OR "transcription profiling"[tiab])

We translated the returned PubMed identifiers to PubMed Central (PMC) identifiers, then to locations on the PubMed Central server. We downloaded the full text for the first 4000 files from PubMed Central and extracted the component containing the raw text in xml format.

To automatically classify our development corpus, we used raw dataset sharing into NCBI's Gene Expression Omnibus(GEO) database as a proxy for running gene expression microarray experiments. This approach will incorrectly classify many gene-expression data articles, because either the authors did not share their gene expression data (about 50%) or they did share but did not have a link to their gene expression study in GEO (about 35%). Nonetheless, we expected the number of false negative instances to be small compared to the number of true negatives and thus sufficiently accurate for training. We implemented this filter by querying PubMed Central with the development-corpus identifiers and the filter *AND "pmc_gds"[filter],* using the NCBI's EUtils web service. We considered articles returned by this filter to be positive examples, or gene expression microarray sharing/creation articles, and articles not returned in this subset to be negative examples.

Query development features

We assembled unigram and bigram features of the article full-text. Specifically, we removed all

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

xml and split on spaces 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, or that did not include at least one alphabetic character. We excluded unigrams and bigrams that included PubMed (and PubMed Central) stop words . Due to the nature of our specific-use case for the query, we also excluded a manually derived list of bioinformatics data words, such as “geo”, “omnibus”, “accession number”, “Agilent,” and journal and formatting words, such as “bmc”, “plos”, “dtd”, and “x000b0.”

We eliminated unigrams and bigrams that did not have at least 20% precision, 20% recall, and a 35% f-measure on the entire training set.

Query development algorithm

Preliminary investigations using established rule-generation algorithms (JRip, Ridor, and others) in Weka returned queries with high f-measure but relatively low precision. Attempts to alter parameters to achieve high precision and acceptable recall were not successful, even with cost-weighted learning. Therefore, we decided to use a simple technique to build our own binary rules, as illustrated in Figure 1.

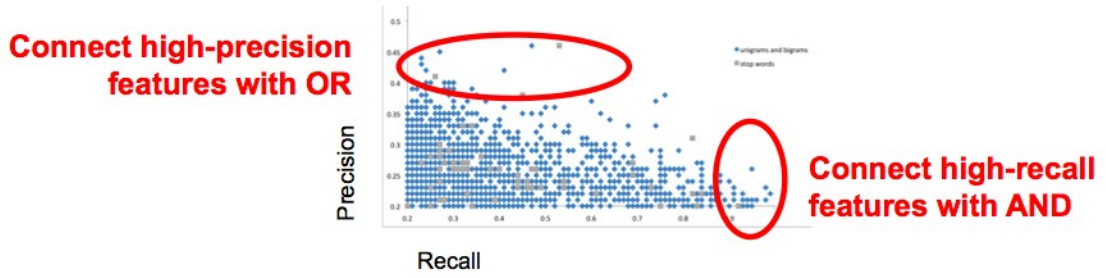

Figure 1: Method for building Boolean query from feature list. In query syntax:
(*(features with highest recall joined with* ***AND****)*
AND *(features with highest precision joined with* ***OR****)****)***

We determined NOT phrases through a manual error analysis of the false positives in the development set.

Query syntax

The search syntax supported by established full-text portals is usually not well documented. We read available help files and experimented to determine capabilities, limitations, and syntax. We then translated the derived rules into the slightly different syntaxes of each of the query engines: PubMed Central, Highwire Press, Scirus, and Google Scholar.

Query evaluation corpus

We evaluated the performance of our derived query against the reference standard established by Ochsner et al.. Although many of the reference articles have full-text freely available in PubMed Central, none are open access and thus none were in the development set.

Because the emphasis of Ochsner et al. was precision rather than recall, their analysis failed to identify a number of true positives. We searched for these misclassifications automatically by identifying whether any of the articles that were considered non-data-generating actually had linked database submissions in GEO: an indication that they did in fact generate data. We also manually examined all classification errors.

Query execution

We ran our query for all journals that included their complete content in PubMed Central first, then Highwire Press, and finally Google Scholar. This order allowed us to maximize the degree to which the query execution could be automated, as per the terms of use of the websites. We ran the queries in each location for articles published in 2007.

We used the EUtils library to automatically execute the query and obtain the results from PubMed Central. For the other query engines, we manually executed the query and manually saved the resulting html files on our computer. We parsed these html files with python scripts to extract the citations and submitted the citation lists to the PubMed Citation Matcher to obtain PubMed identifier (PMID) lists.

Query evaluation statistics

We calculated the precision and recall of the developed filters and compared this performance to that of the two most obvious baseline Medical Subject Heading (MeSH) filters:

- **"gene expression profiling"[mesh] AND "Oligonucleotide Array Sequence Analysis"[mesh]**
- **"gene expression profiling"[mesh] OR "Oligonucleotide Array Sequence Analysis"[mesh]**

We also used Fisher's exact test to verify that the filter was indeed adding value. For our use case, an eventual study of data sharing prevalence, we hoped to achieve recall of at least 50% and precision of at least 90%.

Results

Queries

We applied our query-formulation approach to the task of identifying studies that performed gene expression microarray experiments. Using the open access literature as a development corpus and links to a gene expression microarray database as a proxy endpoint, we derived the following full-text queries:

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

Portal	Query
PubMed Central	("gene expression"[text] AND "microarray"[text] AND "cell"[text] AND "rna"[text]) AND ("rneasy"[text] OR "trizol"[text] OR "real-time pcr"[text]) NOT ("tissue microarray*"[text] OR "cpg island*"[text])
HighWire Press	Anywhere in Text, ANY: ("gene expression" AND microarray AND cell AND rna) AND (rneasy OR trizol OR "real-time pcr") NOT (“tissue microarray*” OR “cpg island*”)
Google Scholar	+"gene expression” +microarray +cell +rna +(rneasy OR trizol OR "real time pcr") -"cpg island*" -"tissue microarray*"
Scirus	Anywhere in Text, ALL: ("gene expression" AND microarray AND cell AND rna) (rneasy OR trizol OR "real-time pcr") ANDNOT ("cpg island*" OR "tissue microarray*")

Evaluation portal coverage

Our evaluation corpus spanned 20 journals. We preferred to execute queries in PubMed Central when possible, since it allows automated query and results processing: three of the 20 journals have deposited all of their content in PubMed Central. HighWire Press is also easy to use, though it does require manual querying and saving of results. As seen in Table 1, eight of the non-PubMed Central journals made their articles queriable by HighWire Press. The remaining journals listed their content in Scirus. Unfortunately, we were unable to reliably query full-text through Scirus, so we queried the remaining journals through Google Scholar for this study.

Table 1: Portal coverage for the 20 journals investigated by Ochsner et al.

Portal	Journal
PubMed Central	Am J Pathol
	EMBO J
	PNAS
Highwire	Blood
	Cancer Res.
	Endocrinology
	FASEB J
	J. Biol. Chem.
	J. Endocrinol.
	J. Immunol.
	Mol. Cell. Biol.
	Mol. Endocrinol.
Scirus	Cell
	Molecular Cell
	Nature
	Nature Cell Biology
	Nature Genetics
	Nature Medicine
	Nature Methods
	Science

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

Query performance

Ochsner et al. identified 768 articles generally related to gene expression microarray data. Through a manual review, they determined that 391 of the articles documented the execution of a gene expression microarray experiment for a true positive rate of 51%. Our query replicated these results with a precision of 83%, recall of 62%, and f-measure of 69%.

Since the emphasis of the Ochsner review was precision rather than recall, we found that they were missing quite a few true positives. We searched for these misclassifications automatically by identifying whether any of the articles that were considered non-data-generating actually had linked database submissions in GEO: an indication that they did in fact generate data. Forty-four articles were reclassified based on this analysis. Our queries found seven of these reclassified articles and missed 37, resulting in a precision of 86% and recall of 57%.

We then manually examined all 41 remaining errors to see if any were due to erroneous manual classification. Based on our manual examination, we reclassified 28 articles as true positives, a true positive rate of 60%. Our query retrieved 12 of these and missed 18. Using this gold standard, the queries achieved a precision of 90% (95% confidence intervals: 86% to 93%), recall of 56% (52% to 61%), and f-measure of 69%. This performance was much improved over chance ($p<0.001$). We used the performance against this final gold standard for the remaining analyses.

To investigate if the queries would be effective in each of the full text portals, we examined the performance by portal, as shown in Table 2.

Table 2: Break down by portal source

	N	precision	recall	f-measure
PubMed Central	149	96%	50%	65%
Highwire Press	498	91%	61%	73%
Google Scholar	121	67%	30%	42%
Weighted average	**768**	**90%**	**56%**	**69%**

The performance of all of these portals was improved over chance ($p < 0.001$), indicating that even the relatively poor performance of Google Scholar was adding value.

Finally, 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
"gene expression profiling"[mesh] OR "Oligonucleotide Array Sequence Analysis"[mesh]	768	81%	66%	73%
"gene expression profiling"[mesh] AND "Oligonucleotide Array Sequence Analysis"[mesh]	768	88%	24%	38%
Derived query	**768**	**90%**	**56%**	**69%**

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

Discussion

We described a mechanism for formulating effective queries for use in publicly available, established full-text search portals, using the open access literature as training material. As a proof of concept, we applied this approach to a task that requires searching the full text of research articles: identifying studies that ran gene expression microarray experiments. The query we derived achieved 90% precision and 56% 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 in only one situation, we believe the method for deriving full-text queries could have widespread potential.

Effectively querying full-text is difficult: Synonyms, variant spellings, acronyms, and inexperience make it difficult to form effective queries . Although difficult, searching full-text is often the only way to identify methods / detect harm , extract detailed data, or identify all of the biomedical concepts or genes explored in the study . There is also evidence that searching full-text is more effective than searching meta-data or abstracts for identifying articles of overall relevance .

Domain-specific biomedical NLP and data integration systems, such as Textpresso ,

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010

Pharmspresso , BioText , and BioLit , illustrate the potential value of accessing, exploring, and analyzing full-text, though none of these tools is designed to facilitate searching across domain-independent open-access and closed-access biomedical literature. Other systems have been built to take a preassembled corpus of positive and negative examples to build a filter query for execution in PubMed , but to our knowledge, none suggest an easily accessed open-source training set nor result in a full-text query for use in domain-independent, publicly accessible online portals.

Existing full-text search portals, such as Google Scholar, Scirus, Highwire Press, and PubMed Central,differ in their features and performance , though we believe their full-text searching capabilities have not yet been compared. We found differences in retrieval performance, but because our dataset was relatively small, it was not clear if any differences between portals were due to the portal or the subset of journals we searched.

While portals provide a source of articles, many prohibit systematic downloads . Furthermore, it is unclear whether standard licensing agreements and fair use allow text mining, “a question on which informed people continue to disagree . Luckily, open access articles are available for download and all kinds of reuse. Evidence suggests that these articles have similar textual characteristics to traditional journal articles , and so we used them as a proxy for all articles.

Our method offers several advantages over alternatives: It is easy to maintain, it is free and open to query both open- and subscription-based content, and the user can be in direct control of recall/precision balance by setting recall and precision thresholds. It does have several limitations, however. This technique can only identify articles with full-text available for query in full-text portals, although we estimate that this is a sizeable amount of the total literature when results from PubMed Central, Highwire Press, Scirus, and Google Scholar are aggregated. A related limitation is that the distribution of articles in full-text portals could influence the distribution of retrieved articles. Articles published within the last year are unlikely to be retrieved, since many journals take full advantage of the NIH Public Access embargo period . Furthermore, while a few journals have made their entire back archives digitally queriable, we suspect that recall of articles more than 10-years old would be relatively poor.

We also recognize that since this technique uses open access articles as a proxy for all articles, our queries would be most refined in areas that are well represented in open access articles. To the extent that there are topics poorly covered by open access articles, this technique could have difficulty deriving keywords to find them.

The system could be expanded in many ways. Its input could instead involve a seed query and a list of "true positive" passages. Other publicly available resources could also be consulted, including the UMLS, WordNet, MEDLINE fields, and MeSH terms. Active learning might allow for further refinement. The system could run parts of speech analysis or domain-specific named entity recognition on the open access training set, if that helped to identify valuable features. It could extract features only from a certain subsection of manuscripts, if there were reason to believe that all relevant information would be in the Methods section, for example. The system

could be enhanced to use bootstrapping to identify phrase variants . Since some portals have some wildcard capabilities, we would like to experiment with learning regular expressions , though there is some evidence that this may not help . Finally, more sophisticated natural language processing algorithms would become easier if this method were implemented within a system like LingPipe .

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 , or a query that has been developed just on abstracts .

While our system will undoubtedly underperform compared with those at the cutting edge of research, we believe it will raise awareness of the constraints in mainstream full-text information retrieval and provide a useful tool for today's researchers.

Acknowledgments
Funding from National Library of Medicine (5T15-LM007059-19 to HAP, 1R01-LM009427-01 to WWC) and the Department of Biomedical Informatics at the University of Pittsburgh.

Availability
Code and data will be publicly 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

Nature Precedings : hdl:10101/npre.2010.4267.2 : Posted 9 Mar 2010