



Using open access literature to guide full-text query formulation

Heather Piwovar and Wendy Chapman



Literature searches, systematic reviews, and text mining require identifying articles based on full-text content. The full text of published biomedical articles contain valuable information not found in abstracts or MeSH terms.

Full-text literature is increasingly available for query. PubMed Central, Highwire Press and Google Scholar are growing fast, thanks to the NIH public access mandate.

However, it is difficult to formulate effective full-text queries manually. Prose and identifiers have large variation, and full-text portals are not designed for query evaluation.

Current full text retrieval research does not address this problem. Cutting-edge systems developed for information retrieval and extraction require complete computational access to a full-text corpora for preprocessing; publisher licenses rarely allow this.

We propose using open access literature to formulate queries for use in full-text portals. We can use open access articles to identify synonyms and lexical variants, tune performance, and generate queries compatible with full-text portal query languages.

Develop a Full-Text Query using Open Access Literature

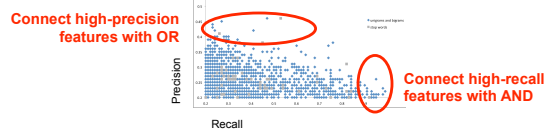
Goal: Identify studies that generate microarray data with a high-precision query

1. Formulate seed queries or provide training set:
 - Our positive seed query finds articles with links from a microarray database:
 - Positive: "microarray"[text] AND "pmc_gds"[filter]
 - Negative: "microarray"[text] NOT "pmc_gds"[filter]
2. Write Python scripts to automatically execute seed query in PubMed Central and download matching Open Access articles

There are 156,313 open access articles across 410 journals, including:

- Arthritis Research Therapy
- BMC Bioinformatics
- BMC Genomics
- BMC Medicine
- Breast Cancer Research
- Genome Biology
- Nucleic Acids Research
- PLoS Medicine
- PLoS Biology
- PLoS ONE

3. Generate unigram and bigram features from articles
4. Assess recall and precision of each feature for classification accuracy against the seed queries



5. Build queries compatible with full-text portals

PubMed Central	("gene expression"[text] AND "microarray"[text] AND "cell"[text] AND "rna"[text]) AND ("measy"[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 ma) AND (measy OR trizol OR "real-time pcr") NOT ("tissue microarray" OR "cpg island")
Google Scholar	+ "gene expression" + microarray + cell + rna + (measy OR trizol OR "real time pcr") - "cpg island" - "tissue microarray"
Scirus	Anywhere in Text, ALL: ("gene expression" AND microarray AND cell AND ma) (measy OR trizol OR "real-time pcr") ANDNOT ("cpg island" OR "tissue microarray")



Then Evaluate the Query using Full-Text Portals

1. Identify gold standard
 - Published manual review of 800 articles across 20 journals found 400 studies that generate microarray data:
 - Ochsner et al. **Much room for improvement in deposition rates of expression microarray datasets.** *Nature Methods* 2008, 5(12):991.
2. Run full-text queries in PubMed Central, Highwire Press, and Google Scholar for maximum coverage



3. Extract article citations from query results, then map to PubMed IDs using PubMed Citation Matcher
4. Calculate recall and precision of the union of the query results:

	Gold Negatives	Gold Positives	Total	
Not found by query	195	180	375	Precision: 92%
Found by query	22	250	272	
Total	217	430	647	Recall: 58%

NOTE: Google Scholar results not yet included

We hope our results will raise awareness of the constraints and opportunities within mainstream full-text information retrieval and provide a useful approach for today's researchers.

For more information, please contact Heather at hpiwovar@gmail.com