Retrieving Biomedical Literature: An Open Source Search Engine Based on Open Access Resources

Hayda Almeida, Ludovic Jean-Louis, Marie-Jean Meurs
contact: meurs.marie-jean@uqam.ca

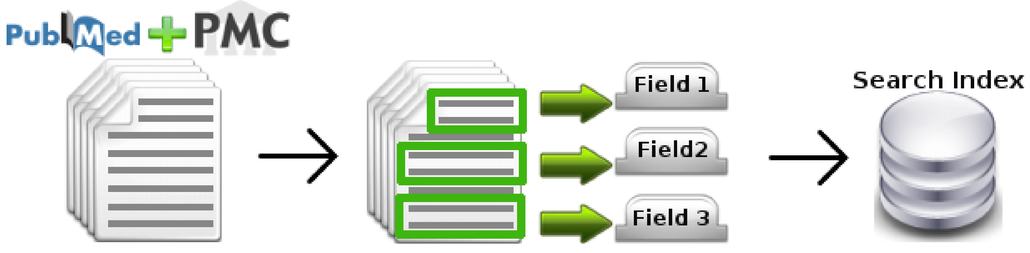
Proposed Approach

Open Access Biomedical Data
PubMed BD + PMC OA
Articles/abstracts with metadata

Search Engine
Document relevant content
Article abstract or full-text

Natural Language Query Processing
Search strategy per query type
Expansion of query terms

Document Indexing

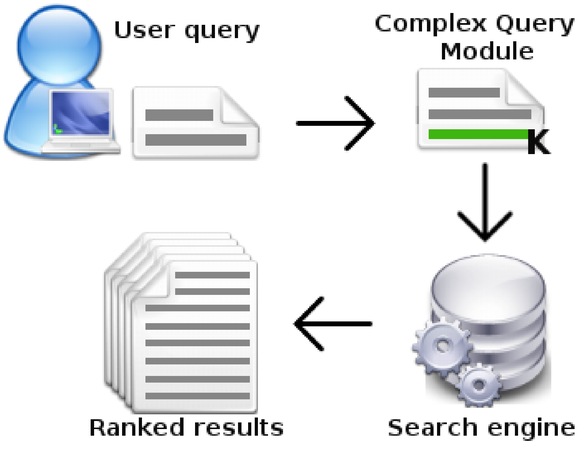


Article Data Loading
1. Load each article file
2. Extract and parse content

Document Processing
1. Select relevant fields
2. Map content to index fields

Document Indexing
1. Create document index
2. Add it to search index

Query Search



User Query & Complex Query Module
1. Input → user search request
2. Expand query → UMLS terms
3. Define: query type + search strategy
4. Generate **bioMine** query

Search engine
1. Submit **bioMine** query to search index
2. Retrieve articles → query terms

Ranked results
1. Order results → query requirements
2. Article relevance → term fields

Query Expansion

`"AIDS versus HIV"`
User query

`HIV+ [HIV Seropositivity], HIV [HIV]`
`AIDS [Acquired Immunodeficiency Syndrome]`
MetaMap [1] annotations → if no terms found in user query

`AIDS versus HIV`
`Acquired Immunodeficiency Syndrome`
Expanded query

Query Type

- Keyword K_Q → no stopwords
  `"AIDS versus HIV"`
- Open Question O_Q → ? cues
  `"what is the difference between HIV and AIDS?"`
- Statement S_Q → stopwords + no ? cues
  `"the difference between HIV and AIDS"`

Preliminary Evaluation

- 19 `{query, target article ID}` sets
- 9 PMCID articles + 10 PMID articles
- Curated for the mycoCLAP database [2]
- Mean Reciprocal Rank → MRR = $\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{position}$

Index Data

PubMed 24,350,000+ PMIDs
PMC 1,200,000+ PMCIDs

- Search index size → 25,403,053 entries
- Index schema → relevant document fields

Results

- Each query search → 20 top ranked results

$Q_\#$	PMC rank	bioMine rank	bioMine RR score	$Q_\#$	PubMed rank	bioMine rank	bioMine RR score
Q_1	3	2	0.500	Q_{10}	2	1	1.000
Q_2	1	20	0.050	Q_{11}	N/A	7	0.143
Q_3	1	2	0.500	Q_{12}	1	1	1.000
Q_4	2	8	0.125	Q_{13}	2	1	1.000
Q_5	2	13	0.077	Q_{14}	1	1	1.000
Q_6	9	1	1.000	Q_{15}	2	1	1.000
Q_7	2	5	0.200	Q_{16}	1	N/A	0.000
Q_8	1	17	0.059	Q_{17}	N/A	1	1.000
Q_9	1	10	0.100	Q_{18}	1	N/A	0.000
				Q_{19}	1	1	1.000

total # of queries = 19 **MRR = 0.513**

Conclusion

- Target articles → 1^{st} position ≈50% of the time
- `{query, PMCID}` sets: Target always in top 20
- `{query, PMID}` sets: Better overall ranking

https://github.com/BigMiners/bioMine

References
[1] Aronson A. and Lang F., *An Overview of MetaMap: Historical Perspective and Recent Advances*, JAMIA, 2010.
[2] Strasser K. et al., *mycoCLAP, the Database for Characterized Lignocellulose-active Proteins of Fungal Origin: Resource and Text Mining Curation Support*, Database, 2015

Acknowledgements
Dr. Adrian Tsang, CSFG director, Concordia University