Synonym-based Query Expansion and Boosting-based Re-ranking: A Two-phase Approach for Genomic Information Retrieval

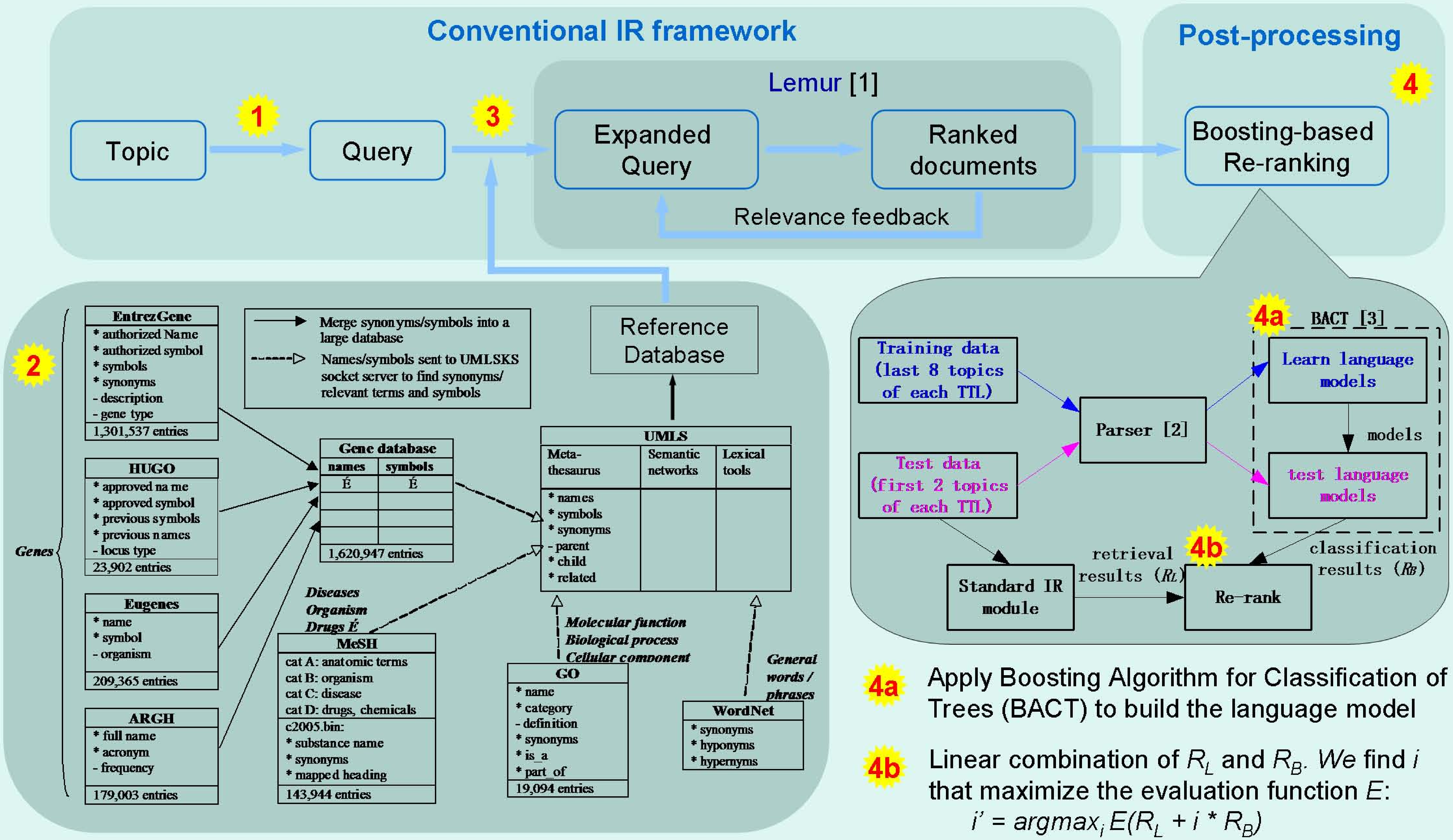
Zhongmin Shi, Baohua Gu, Fred Popowich and **Anoop Sarkar**
School of Computing Science, Simon Fraser University, Canada

1 **Query generation**: manually select keywords from official topic and build the structured query for each topic

2 **Reference database construction**: collecting synonyms from a large collection of ontology sources: EntrezGene, HUGO, Eugenes, ARGH, MeSH, GO, UMLS and WordNet.

3 **Query expansion**: look up synonyms of keywords in the reference database

4 **Re-ranking in the post-processing**: apply a boosting-based classification algorithm to re-rank the retrieved documents



4a Apply Boosting Algorithm for Classification of Trees (BACT) to build the language model

4b Linear combination of R_L and R_B. *We find i* that maximize the evaluation function *E*:

$$i' = argmax_i E(R_L + i * R_B)$$

Experimental results:

The boosting-based re-ranking does help when *bpref* of the conventional IR system is low.

topic #	metrics	*i*=0 (*i'*)	*i*=10
100	MAP	0.2221	0.1785
	bpref	0.8649	0.8649
	P10	0.4	0.3
	P100	0.28	0.22
101	MAP	0.0685	0.0195
	bpref	0.75	0.75
	P10	0	0
	P100	0.07	0.07

Table 1: Performance of re-ranking on TTL #1

topic #	metrics	*i*=0	*i*=15 (*i'*)
110	MAP	0.0012	0.0024
	bpref	0.25	0.25
	P10	0	0
	P100	0	0.01
111	MAP	0.0492	0.1602
	bpref	0.4356	0.4356
	P10	0.1	0.7
	P100	0.1	0.4

Table 2: Performance of re-ranking on TTL #2

topic #	metrics	*i*=0 (*i'*)	*i*=10
120	MAP	0.6113	0.2410
	bpref	0.8145	0.8145
	P10	1	0.3
	P100	0.88	0.29
121	MAP	0.6697	0.0328
	bpref	0.8810	0.8810
	P10	0.8	0
	P100	0.34	0

Table 3: Performance of re-ranking on TTL #3

[1] Lemur. 2005. Language Modeling Toolkit 4.1. http://www.lemurproject.org.

[2] Eugene Charniak. A maximum-entropy-inspired parser. NAACL 2000.

[3] Taku Kudo and Yuji Matsumoto. A boosting algorithm for classification of semi-structured text. EMNLP 2004.