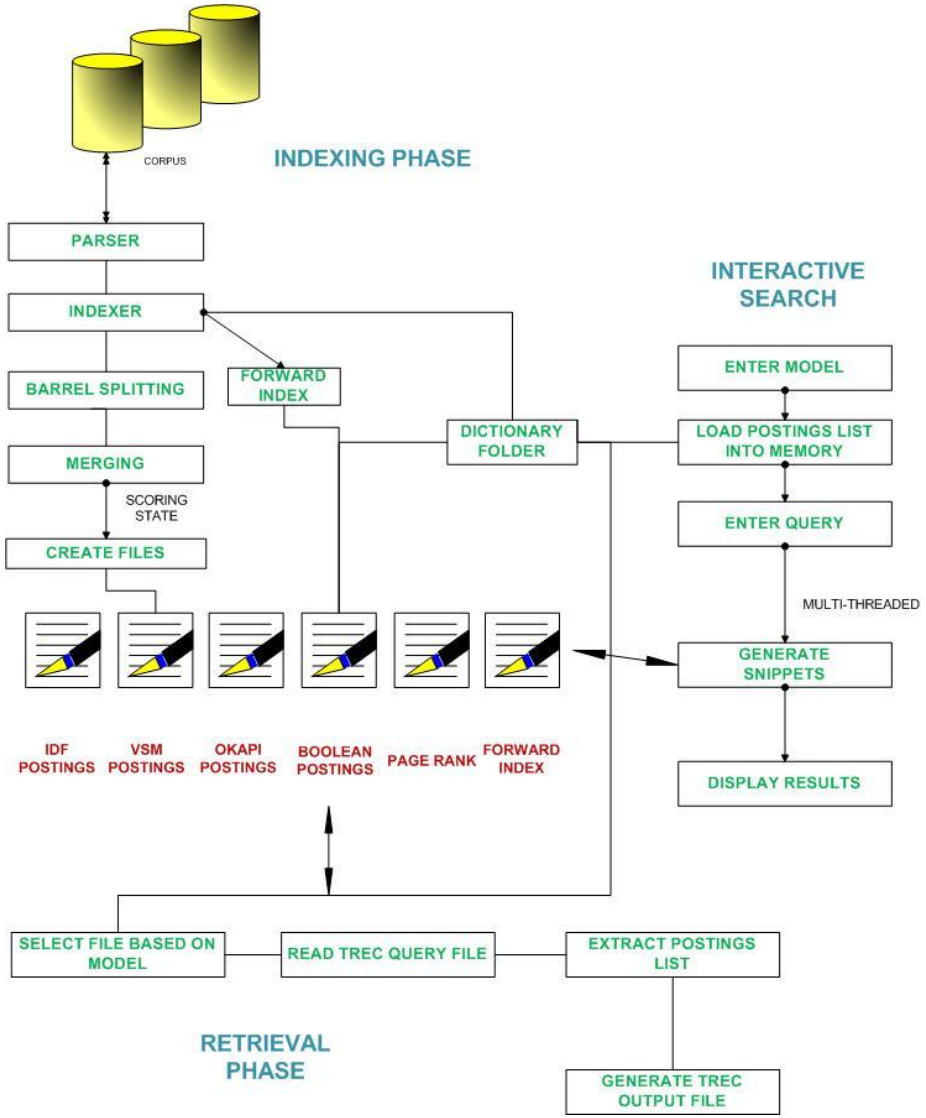
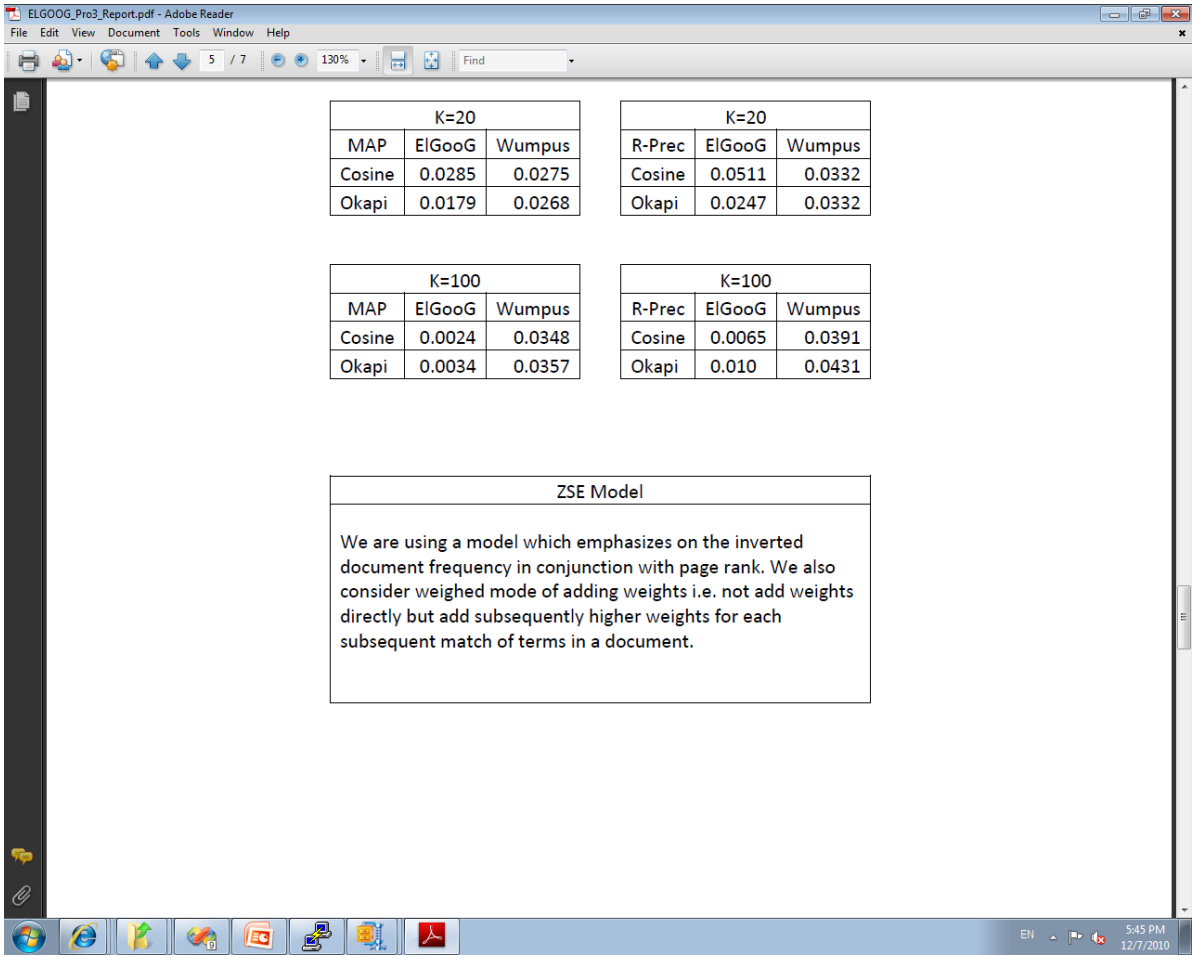
Elgoog

Contributors:
Suchismit Mahapatra
Zlatan Sadibasic
Eric Lee


CORPUS
INDEXING PHASE
PARSER
INDEXER
BARREL SPLITTING
FORWARD INDEX
MERGING
SCORING STATE
CREATE FILES
DICTIONARY FOLDER
INTERACTIVE SEARCH
ENTER MODEL
LOAD POSTINGS LIST INTO MEMORY
ENTER QUERY
MULTI-THREADED
GENERATE SNIPPETS
DISPLAY RESULTS
IDF POSTINGS
VSM POSTINGS
OKAPI POSTINGS
BOOLEAN POSTINGS
PAGE RANK
FORWARD INDEX
SELECT FILE BASED ON MODEL
READ TREC QUERY FILE
EXTRACT POSTINGS LIST
RETRIEVAL PHASE
GENERATE TREC OUTPUT FILE

PERFORMANCE



K=20		
MAP	ElGooG	Wumpus
Cosine	0.0285	0.0275
Okapi	0.0179	0.0268

K=20		
R-Prec	ElGooG	Wumpus
Cosine	0.0511	0.0332
Okapi	0.0247	0.0332

K=100		
MAP	ElGooG	Wumpus
Cosine	0.0024	0.0348
Okapi	0.0034	0.0357

K=100		
R-Prec	ElGooG	Wumpus
Cosine	0.0065	0.0391
Okapi	0.010	0.0431

ZSE Model
We are using a model which emphasizes on the inverted document frequency in conjunction with page rank. We also consider weighed mode of adding weights i.e. not add weights directly but add subsequently higher weights for each subsequent match of terms in a document.

SYSTEM PERFORMANCE

- Queries with 5 terms or less --→ less than 1 s
- High density queries (description based queries) have been observed to execute no greater than 3 secs.

DEMO

DEMO TIME…

ENGINEERING ISSUES/CHALLENGES

- Understanding the corpus data
- Expressing IR concepts via programming language
- Synchronizing and merging the code
- Ways to improve performance of the system
- Issues with data storage, in the beginning