Islamic University of Gaza

Faculty of Engineering

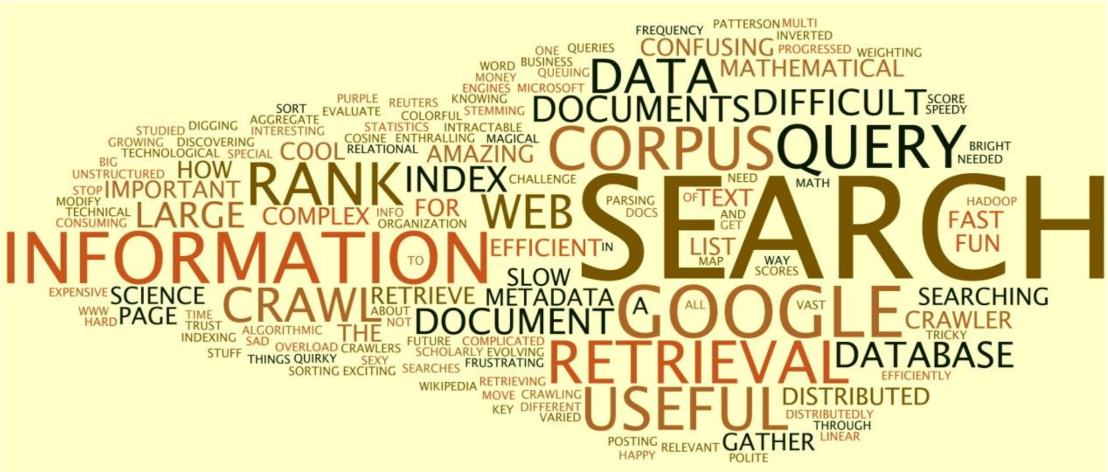
Computer Engineering Department

Information Storage and Retrieval (ECOM 5124)

IR

HW 1

Boolean Retrieval


FREQUENCY PATTERSON MULTI INVERTED
ONE QUERIES CONFUSING PROGRESSED WEIGHTING
WORD BUSINESS MONEY QUEUING DATA MATHEMATICAL
ENGINES MICROSOFT
SORT PURPLE REUTERS KNOWING
AGGREGATE EVALUATE COLORFUL STEMMING DOCUMENTS DIFFICULT SCORE SPEEDY
STUDIED DIGGING INTERESTING STATISTICS INTRACTABLE
GROWING DISCOVERING COSINE ENTHRALLING MAGICAL
BIG TECHNOLOGICAL SPECIAL COOL RELATIONAL AMAZING CORPUS QUERY BRIGHT NEEDED
UNSTRUCTURED HOW
STOP IMPORTANT RANK INDEX CHALLENGE NEED MATH
MODIFY TECHNICAL CONSUMING LARGE COMPLEX INFO FOR ORGANIZATION WEB PARSING DOCS OF TEXT AND GET HADOOP FAST
INFORMATION TO EFFICIENT IN SEARCH LIST MAP WAY SCORES FUN
EXPENSIVE SCIENCE SLOW
WWW HARD PAGE TIME TRUST CRAWL RETRIEVE ABOUT NOT METADATA A ALL VAST SEARCHING
INDEXING ALGORITHMIC THE DOCUMENT GOOGLE CRAWLER TRICKY
STUFF SAD OVERLOAD CRAWLERS FUTURE COMPLICATED
THINGS QUIRKY SEXY SCHOLARLY EVOLVING RETRIEVAL DATABASE
SORTING EXCITING SEARCHES FRUSTRATING EFFICIENTLY
WIKIPEDIA RETRIEVING MOVE CRAWLING KEY DIFFERENT VARIED USEFUL DISTRIBUTED DISTRIBUTEDLY THROUGH
POSTING HAPPY RELEVANT GATHER POLITE LINEAR

Eng. Mohammed Abdualal

February 14, 2015

Exercise 1.1

Draw the inverted index that would be built for the following document collection. (See Figure 1.3 for an example.)

Doc 1 new home sales top forecasts
Doc 2 home sales rise in july
Doc 3 increase in home sales in july
Doc 4 july new home sales rise

Solution :

First list each unique term - new, home, sales, top, forecasts, rise, in, july, increase.

Then, arrange the terms in alphabetical order - forecasts, home, in, increase, july, new, rise, sales, top.

forecasts -> Doc 1
home -> Doc 1, Doc 2, Doc 3, Doc 4
in -> Doc 2, Doc 3
increase -> Doc 3
july -> Doc 2, Doc 3, Doc 4
new -> Doc 1, Doc 4
rise -> Doc 4
sales -> Doc 1, Doc 2, Doc 3, Doc 4
top -> Doc 1

Exercise 1.2

Consider these documents:

Doc 1 breakthrough drug for schizophrenia
Doc 2 new schizophrenia drug
Doc 3 new approach for treatment of schizophrenia
Doc 4 new hopes for schizophrenia patients

a. Draw the term-document incidence matrix for this document collection.

The **term document incidence matrix** has the list of terms as rows and the list of documents as columns. Each cell in the matrix represents whether the term is present in the document (value 1 if present, else value 0).
The term document incidence matrix is created as below

	Doc 1	Doc 2	Doc 3	Doc 4
approach	0	0	1	0
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
hopes	0	0	0	1
new	0	0	1	1
Of	0	0	1	0
patients	0	0	0	1
schizophrenia	1	1	1	1
treatment	0	0	1	0

b. Draw the inverted index representation for this collection, as in Figure 1.3

The **inverted index** for the above collection is as below

approach	Doc 3			
breakthrough	Doc 1			
drug	Doc 1	Doc 2		
for	Doc 1	Doc 3	Doc 4	
hopes	Doc 4			
new	Doc 3	Doc 4		
of	Doc 3			
patients	Doc 4			
schizophrenia	Doc 1	Doc 2	Doc 3	Doc 4
treatment	Doc 3			

Exercise 1.3

For the document collection shown in Exercise 1.2, what are the returned results for these queries:

a. **schizophrenia AND drug**
Solution
Here we use the term-document incidence matrix to perform a boolean retrieval for the given query

For the terms schizophrenia and drug, we take the row (or vector) which indicate the document the term appears in,

schizophrenia - 1 1 1 1
drug - 1 1 0 0
Doing a bitwise AND operation for each of the term vectors gives,
1 1 1 1 AND 1 1 0 0 = 1 1 0 0

The result vector 1 1 0 0 gives Doc 1 and Doc 2 as the documents in which the terms schizophrenia AND drug both are present.

b. for AND NOT(drug OR approach)

for AND NOT (drug OR approach)

Term vectors

for - 1 0 1 1

drug - 1 1 0 0

approach - 0 0 1 0

First we do a boolean bit wise OR for drug, approach, which gives
1 1 0 0 OR 0 0 1 0 = 1 1 1 0

The we do a NOT operation on 1 1 1 0 (i.e. on drug OR approach), which gives 0 0 0 1

Then we do an AND operation on 1 0 1 1 (i.e. for) AND 0 0 0 1 (i.e. NOT(drug OR approach)), which gives 0 0 0 1

Thus the document that contains for AND NOT (drug OR approach) is Doc 4.

These exercise illustrate the Boolean Retrieval model for search of query terms in given list of documents.

Exercise 1.7
Recommend a query processing order for

(tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes) given the following postings list sizes:

Term	Postings size
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

Solution:

Using the conservative estimate of the length of the union of postings lists, the recommended order is:
(kaleidoscope OR eyes) (300,321) AND (tangerine OR trees) (363,465) AND (marmalade OR skies)(379,571)
However, depending on the actual distribution of postings, (tangerine OR trees) may well be longer than (marmalade OR skies), because the two components of the former are more asymmetric. For example, the union of 11 and 9990 is expected to be longer than the union of 5000 and 5000 even though the conservative estimate predicts otherwise