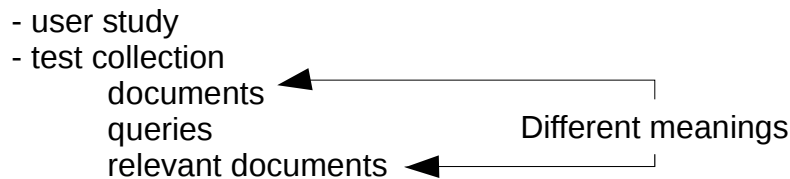


CS 533 – Information Retrieval Systems

Class Notes – Prepared by Semih Sahin

Search Result Diversification Performance Measurement



s-recall: covering as many meanings as possible in higher rank positions.

Methods are compared in terms of what percentage of meanings are mentioned in their search result first.

Rank	Document	Covered Meanings by docs. (subtopic)
1	d ₅	m ₁
2	d ₁₀	m ₅
3	d ₁	m ₁ , m ₅
4	d ₂	m ₂
5	d ₉	m ₂
6	d ₃	m ₃
7	d ₄	m ₅
8	d ₇	m ₁ , m ₃ , m ₄
9	d ₈	m ₅
10	d ₆	m ₁ , m ₂ , m ₃ , m ₄ , m ₆

$$s-recall @ k = \frac{(\prod^k subtopic(d_i))}{n_a}$$

n_a : number of different meanings

$$s-recall @ 5 = \frac{3}{6} = 0.5$$

Precision IA (Intent-Aware):

Precision → what percentage of retrieved documents are relevant

Rank	m ₁	m ₂	m ₃	m ₄	m ₅	m ₆
1	1	0	0	0	0	0
2	0	0	0	0	1	0
3	1	0	0	0	1	0
4	0	1	0	0	0	0
5	0	1	0	0	0	0
6	0	0	1	0	0	0

7	0	0	0	0	0	0
8	1	0	1	1	0	0
9	0	0	0	0	1	0
10	1	1	1	1	1	0

Precision @ 10 0.4 0.3 0.3 0.2 0.4 0

$$Precision\ IA\ @\ k = \frac{1}{m} (\sum_1^m \text{precision for meaning @ } k)$$

$$Precision\ IA\ @\ 10 = \frac{1}{6} (\frac{4}{10} + \frac{3}{10} + \frac{3}{10} + \frac{2}{10} + \frac{4}{10} + 0) = \frac{4}{15}$$

Signature Files

for each term , we have one term specific signature.

terms → term signatures (bit maps)

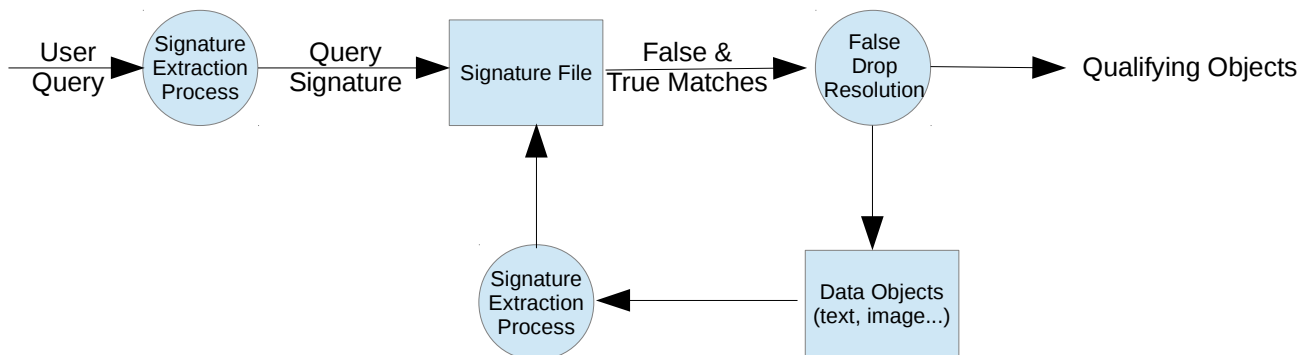
Example:	<u>term</u>	<u>term signature</u>
	“object”	1000 1000
	“signature”	0010 0100
	<u>“generation”</u>	<u>1000 0100</u>
		1010 1100 ← document signature (OR term signatures)

Weight of a signature: total number of 1s (50% of bits are 1 and 50% of bits are 0)

<u>query</u>	<u>query signature</u>	<u>result</u>
“database”	1100 0000	no match
“generation”	1000 1000	true match
“information”	1010 0000	false match

In query processing a block signature may match the query signature but in reality block may not satisfy the query. Go check the document / false drop resolution

Signature Based IR Environment



In fixed size block approach (FSB), how to make sure that half of the signature positions are 0 (1).

S: number of 1s in term signature

F: signature size in terms of number of bits

D: block size in terms of number of unique terms

$$S_{opt} = F \ln 2 / D = 0.7 \times F / D$$

Signature File Structures

1. Single Level

1.1 Sequential Files

1.2 Vertical Partitioning
Bit Sliced
Frame Sliced

2. Multi Level

2.1 Signature Tree

2.2 Two Level Signature

3. Horizontal Partitioning

3.1 Linear Hashing with
Superimposed Signatures
3.2 Key Based
3.3 Gustafson's Method

Sequential Signature Files

S1: 0001 1110

S2: 1101 0001

S3: 0011 1100

S4: 1100 0011

S5: 0011 0110

S6: 1100 1001

Compare each query signature with each block signature.

Bit Sliced

Store signature files in column wise order.

C1: 010 101

C2: 010 101

C3: 001 010

C4: 111 010

C5: 101 001

C6: 101 010

C7: 100 110

C8: 010 101

Query: Q = 1000 0100

AND all C_i where $Q[i] = 1$

$C1 \& C4 = [010\ 101] \& [111\ 010] = [010\ 000]$

Questions

1. Obtain the document signature with given term signatures and indicate results ('no match', 'true match', or 'false match') of the queries in the query list.

<u>term</u>	<u>signature</u>	<u>query</u>	<u>signature</u>	<u>result</u>
information	1100 1000	semih	0100 0000	?
retrieval	1001 0010	stream	0001 1010	?
stream	0001 1010	fazli	1001 0100	?
<u>processing</u>	<u>0100 1000</u>			
doc. signature	?			

2. Using the given signature files, obtain bit sliced representation. Using bit sliced structure, list which documents should be checked for each query.

<u>document</u>	<u>signature</u>	<u>query</u>	<u>document list</u>
d1	1100 0101	0000 1110	?
d2	1010 1001	0100 0010	?
d3	0000 1111	1100 0000	?
d4	0101 1100		
d5	1111 0000		
d6	0001 1110		

3.

<u>rank</u>	<u>document</u>	<u>subtopic</u>
1	d ₁	m ₄
2	d ₅	m ₁
3	d ₆	m ₁ , m ₂
4	d ₂	m ₃ , m ₅
5	d ₈	m ₁ , m ₄ , m ₆
6	d ₃	m ₁
7	d ₁₀	m ₃
8	d ₄	m ₅
9	d ₉	m ₂ , m ₄
10	d ₇	m ₆

Find s-recall values at level 3, 7 and 10.

4. Using the table in question 3, find precision IA values at levels 3, 7 and 10.

Answers

1. doc signature: 1101 1010

query	signature	result
semih	0100 0000	false match
stream	0001 1010	true match
fazli	1001 0100	no match

2. bit sliced:

C1: 110 010
C2: 100 110
C3: 010 010
C4: 000 111
C5: 011 101
C6: 101 101
C7: 001 001
C8: 111 000

query	columns	document list
0000 1110	C5, C6, C7	d3, d6
0100 0010	C2, C7	-
1100 0000	C1, C2	d1, d5

3. $s-recall @3 = \frac{3}{6} = 0.5$ $s-recall @7 = \frac{6}{6} = 1.0$ $s-recall @10 = \frac{6}{6} = 1.0$

4.

Rank	m ₁	m ₂	m ₃	m ₄	m ₅	m ₆
1	0	0	0	1	0	0
2	1	0	0	0	0	0
3	1	1	0	0	0	0
4	0	0	1	0	1	0
5	1	0	0	1	0	1
6	1	0	0	0	0	0
7	0	0	1	0	0	0
8	0	0	0	0	1	0
9	0	1	0	1	0	0
10	0	0	0	0	0	1

$$precision\ IA@3 = \frac{1}{6} \left(\frac{2}{3} + \frac{1}{3} + 0 + \frac{1}{3} + 0 + 0 \right) = \frac{2}{9}$$

$$precision\ IA@7 = \frac{1}{6} \left(\frac{4}{7} + \frac{1}{7} + \frac{2}{7} + \frac{2}{7} + \frac{1}{7} + \frac{1}{7} \right) = \frac{11}{42}$$

$$precision\ IA@10 = \frac{1}{6} \left(\frac{4}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10} + \frac{2}{10} + \frac{2}{10} \right) = \frac{1}{6}$$