

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Sep, 2019
Chennai Mathematical Institute



அட பாடல் போல தேடல் கூட ஒரு சுகமே
Search, like a song, is also a joy.

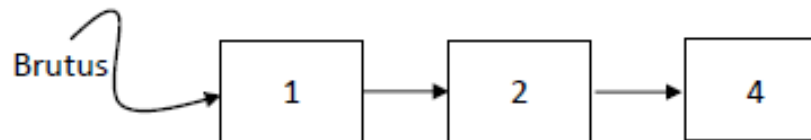
- From the movie, Thulladha Manamum Thullum. Lyrics by Vaali.



Query Processing Order

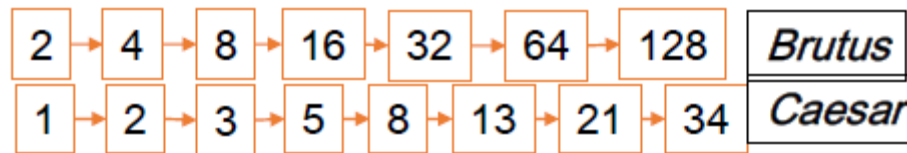
Term Document Matrix & Postings List

		Documents					
		Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Terms	Antony	1	1	0	0	0	1
	Brutus	1	1	0	1	0	0
	Caesar	1	1	0	1	1	1
	Calpurnia	0	1	0	0	0	0
	Cleopatra	1	0	0	0	0	0
	mercy	1	0	1	1	1	1
	worser	1	0	1	1	1	0



Query Processing

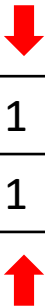
- Brutus AND Caesar



- Which document(s) should result?
- How many comparisons did you do?

Query Processing

- X AND Y



X	1	2	3	4	5	6	7	8
Y	1	2	3					

- Which document(s) should result?
- How many comparisons did you do?
- Is there any way we could get:
 - $|\text{result}| > \min(|x|, |y|)$
 - No!

Query Processing

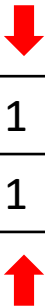
- X AND Y

X								
Y								

- If $|X| = 3$, $|Y| = 5$, Can you fill the boxes in two different ways such that the number of comparisons are different?

Query Processing

- X AND Y

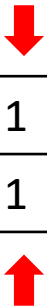


X	1	3	5					
Y	1	2	3	4	5			

- No. of Comparisons = $|(1,1), (3,2), (3,3), (5,4), (5,5)| = 5$.

Query Processing

- X AND Y



X	1	2	3					
Y	1	2	3	4	5			

- No. of Comparisons = $|(1,1), (2,2), (3,3)| = 3.$

Query Processing

- X AND Y

X								
Y								

- If $|X| = 3$, $|Y| = 5$, Can you fill the boxes in two different ways such that the number of comparisons are ~~different~~ **maximum**?

Query Processing

- X AND Y

X	1	2	8					
Y	4	5	6	7	8			

- No. of Comparisons =
 $| (1,4), (2,4), (8,4), (8,5), (8,6), (8,7), (8,8) | = 7$
- Can you do any better?

Query Processing

- X AND Y
- Min. No. of Comparisons = 3

X	1	2	3					
Y	1	2	3	4	5			

- Max. No. of Comparisons = 7

X	1	2	8					
Y	4	5	6	7	8			

- Is there a better answer?

Query Processing

- Query: Brutus AND Caesar AND Calpurnia
- Assumption:
 - Brutus appears in 10 documents.
 - Caesar appears in 5 documents.
 - Calpurnia appears in 3 documents.
- How many comparisons?
 - Option 1: Merge Brutus AND Caesar first. Merge the result with Calpurnia.
 - Option 2: Merge Caesar AND Calpurnia first. Merge the result with Brutus.

Query Processing Order

- Option 1: Merge Brutus AND Caesar first. Merge the result with Calpurnia.
 - Brutus AND Caesar: In worst case, requires $(10 + 5) = 15$ comparisons.
 - Result AND Calpurnia: In worst case, requires $(5 + 3) = 8$ comparisons.
 - Therefore, requires $15 + 8 = 23$ comparisons.
- Option 2: Merge Caesar AND Calpurnia first. Merge the result with Brutus.
 - Caesar AND Calpurnia: In worst case, requires $5 + 3 = 8$ comparisons.
 - Result AND Brutus: In worst case, requires $3 + 10 = 13$ comparisons.
 - Therefore, $13 + 8 = 21$ comparisons.

*We approximate worst case comparisons to $x+y$ for convenience.

Process in increasing order by frequency.

Do you now see why we store frequency
with our Dictionary terms?

term	doc.	freq.	→	postings lists
ambitious	1	1	→	2
be	1		→	2
brutus	2		→	1 → 2
capitol	1		→	1
caesar	2		→	1 → 2
did	1		→	1
enact	1		→	1
hath	1		→	2



Quiz

What is the best order of processing “eyes and skies and trees”?

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

What about “(eyes or skies) and (trees or tangerine) and (marmalade or kaleidoscope)”?

**Find the query processing order for
(A OR B) AND (C OR D) AND (E OR F)**



$(A \text{ OR } B) \text{ AND } (C \text{ OR } D) \text{ AND } (E \text{ OR } F)$

- Let $x = \text{Freq}(A) + \text{Freq}(B)$
- Let $y = \text{Freq}(C) + \text{Freq}(D)$
- Let $z = \text{Freq}(E) + \text{Freq}(F)$
- A OR B leaves us with A union B items. In worst case, we have $\text{freq}(A) + \text{freq}(B)$ items.
- We know how to solve $(x \text{ AND } y \text{ AND } z)$

Quiz

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

What about (tangerine OR trees) AND (marmalade OR skies) AND
(kaleidoscope OR eyes)?

Answer: ((kaleidoscope OR eyes) AND (tangerine OR trees)) AND (marmalade or skies)

Thank You