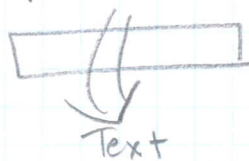


\* Croft & Belking Dec. 1992. Con of the ACM

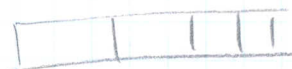
IR & IF: Two sides of the same coin

\* Con. of the ACM 2000 Feb. Special Issue on "News on Demand"

Speech



transformation



Story Segmentation

TDT: Topic Detection and Tracking

## Index Structures

### 1- Brute Force Method (Sequential Comparison)

Random Brute-Force: Search terms in document in random order.

Ranked Brute-Force: Ranked list of terms

More efficient since faster decision making

1. least frequent in docs

...

n. most frequent in docs

### 2- The Counting Method

Paul Krugman

Sample Profiles

$P_1 = (a, b)$

$P_2 = (a, d)$

$P_3 = (a, d, e)$

$P_4 = (b, f)$

$P_5 = (c, d, e, f)$

	Total	Count	a	b	c	f
$P_1$	2	0		2		
$P_2$	2	0	1			
$P_3$	3	0	1	1		
$P_4$	2	0		1		2
$P_5$	4	0			1	2

Sample Document:

Distinct Words: a b c f

Matching profiles are

$p_1$  &  $p_4$

Directory

a	→	p1	p2	p3
b	→	p1	p4	
c	→	p5		
d	→	p2	p3	p5
e	→	p3	p5	
f	→	p4	p5	

Main Memory

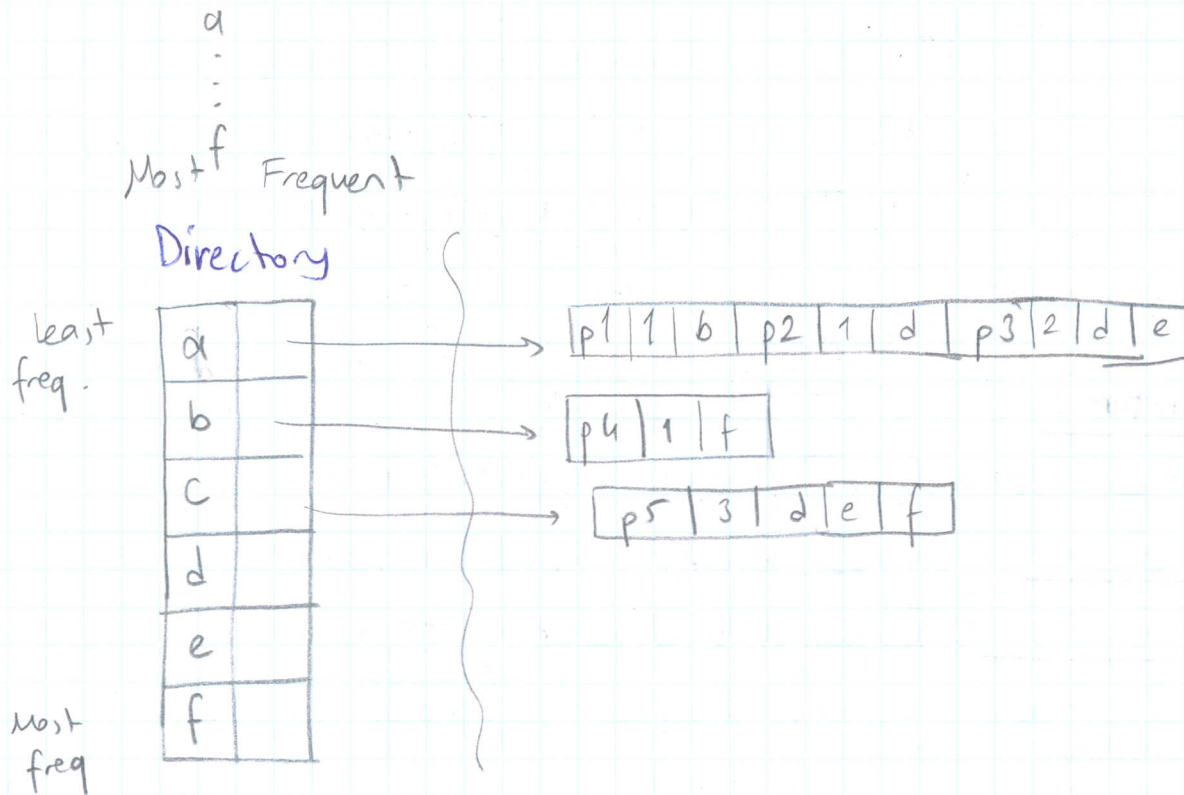
Disk

# The Key Methods

## Ranked Key Method

More frequent terms (ie. terms that appear in many different docs) are used less frequently in user profiles.

Least Frequent (in docs)



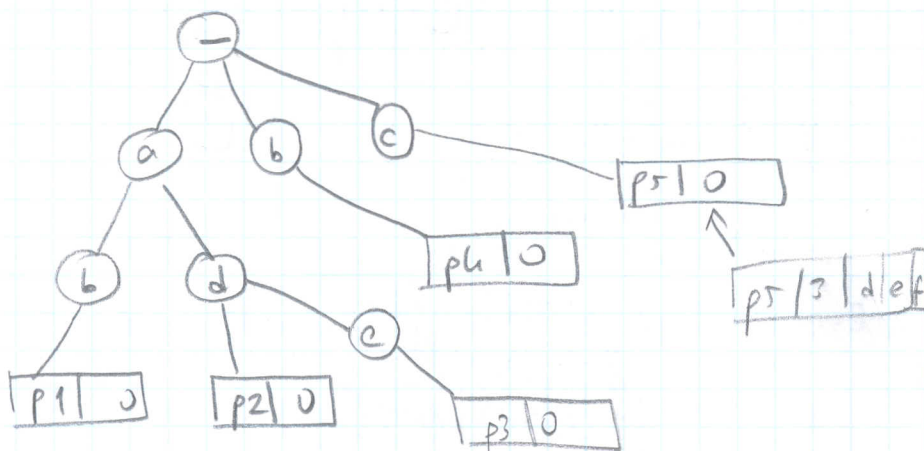
Do we need count & total array?

Incoming doc = a, b, c, f → p1

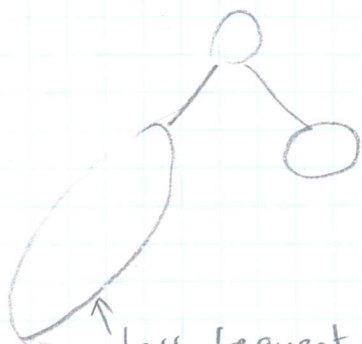
## The Tree Method

Users may have similar profiles

Use a tree structure



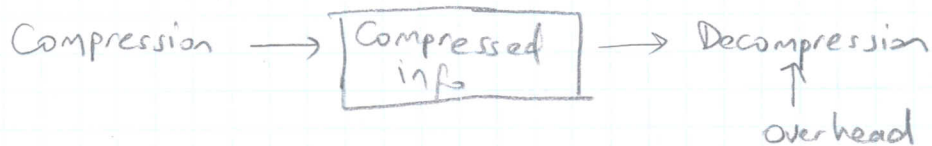
\* Save Storage  
Ranked Key Method



less frequent terms appear in more profiles.

Compression

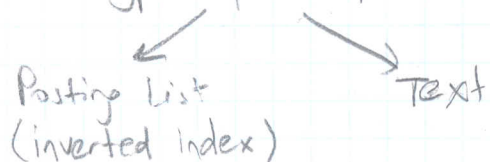
less storage



How to measure compression effectiveness

$$\text{Compression ratio} = \frac{\text{Original Length} - \text{Encoded length}}{\text{Original Length}}$$

Type of Compression



Inverted Index Compression

$t_1 \rightarrow 5, 10, 22, 30$

$t_2 \rightarrow 3, 5, 7, 10, 16$

9 integers

4 bytes/int

8 bits/byte

$$\text{Cost} = 9 \times 4 \times 8 = 288 \text{ ~~byte~~ bits}$$

Use run length encoding

$t_1 \rightarrow 5, 5, 12, 8$

$t_2 \rightarrow 3, 2, 2, 3, 6$

largest no = 12

$$12 = 1100_2$$

no. of bits needed to represent int  $n$   $\lceil \log_2 n \rceil$