

Improving Search Engines using Multi-Word Indicies

Hatem Nassrat
CSCI 6403
December 2008



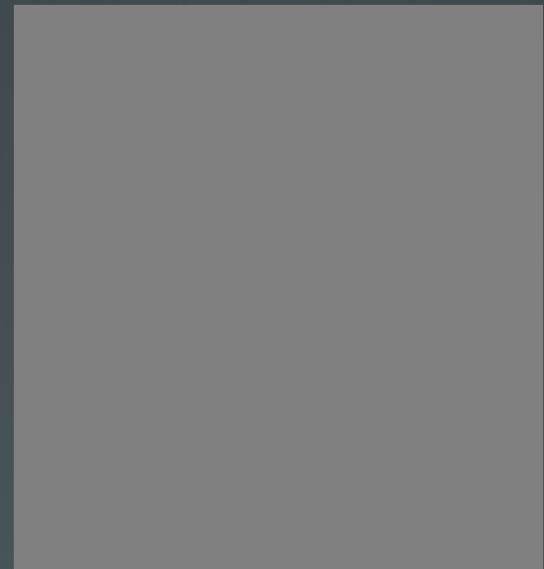
Introduction

- Multi-Word Index
- To Find
 - Advantages (Better Accuracy)
 - Disadvantages (Speed, Disk Space)



Multi Words

Memory



M = 2

N = 4

Database



Multi Words

Memory

AA (0,1)
AB (0,2)
AD (0,3)

A A B D E F G H

M = 2
N = 4

Database



Multi Words

A A B D E F G H

Memory

AB (0,2)
AD (0,3)

M = 2
N = 4

Database

Multi Words

A A B D E F G H

M = 2

N = 4

Database

Memory

AB (0,2)
AD (0,3)
AB (1,2)
AD (1,3)
AE (1,4)

Multi Words

A A B D E F G H

M = 2
N = 4

Database

Memory

AB (0,2)
AD (0,3)
AB (1,2)
AD (1,3)
AE (1,4)

Multi Words

A A B D E F G H

M = 2
N = 4

Database

Memory

AB (1,2)
AD (1,3)
AE (1,4)
BD (2,3)
BE (2,4)
BF (2,5)

Multi Words

A A B D E F G H

M = 2
N = 4

Database

Memory

AB (1,2)
AD (1,3)
AE (1,4)
BD (2,3)
BE (2,4)
BF (2,5)

Multi Words

A A B D E F G H

M = 2
N = 4

Memory

AD (1,3)
AE (1,4)
BD (2,3)
BE (2,4)
BF (2,5)

Database

AB (1,2)



Multi Words

A A B D E F G H

M = 2
N = 4

Memory

AD (1,3)
AE (1,4)
BD (2,3)
BE (2,4)
BF (2,5)
DE (3,4)
DF (3,5)
DG (3,6)

Database

AB (1,2)

Multi Words

A A B D E F G H

M = 2
N = 4

Memory

AE (1,4)
BE (2,4)
BF (2,5)
DE (3,4)
DF (3,5)
DG (3,6)

Database

AB (1,2)
AD (1,3)
BD (2,3)

Method

- Reuters 21578
- Base + Augmented System
- Inverted Files (TC B-Tree)
- Application in Vector Space
- Retrieval – TFIDF
- MW: using M=4, N=10

$$1 + \sum_{i=1}^M \binom{N-1}{i}$$

$$(1 + \sum_{i=1}^M \binom{N-1}{i}) \times (X + Y)$$

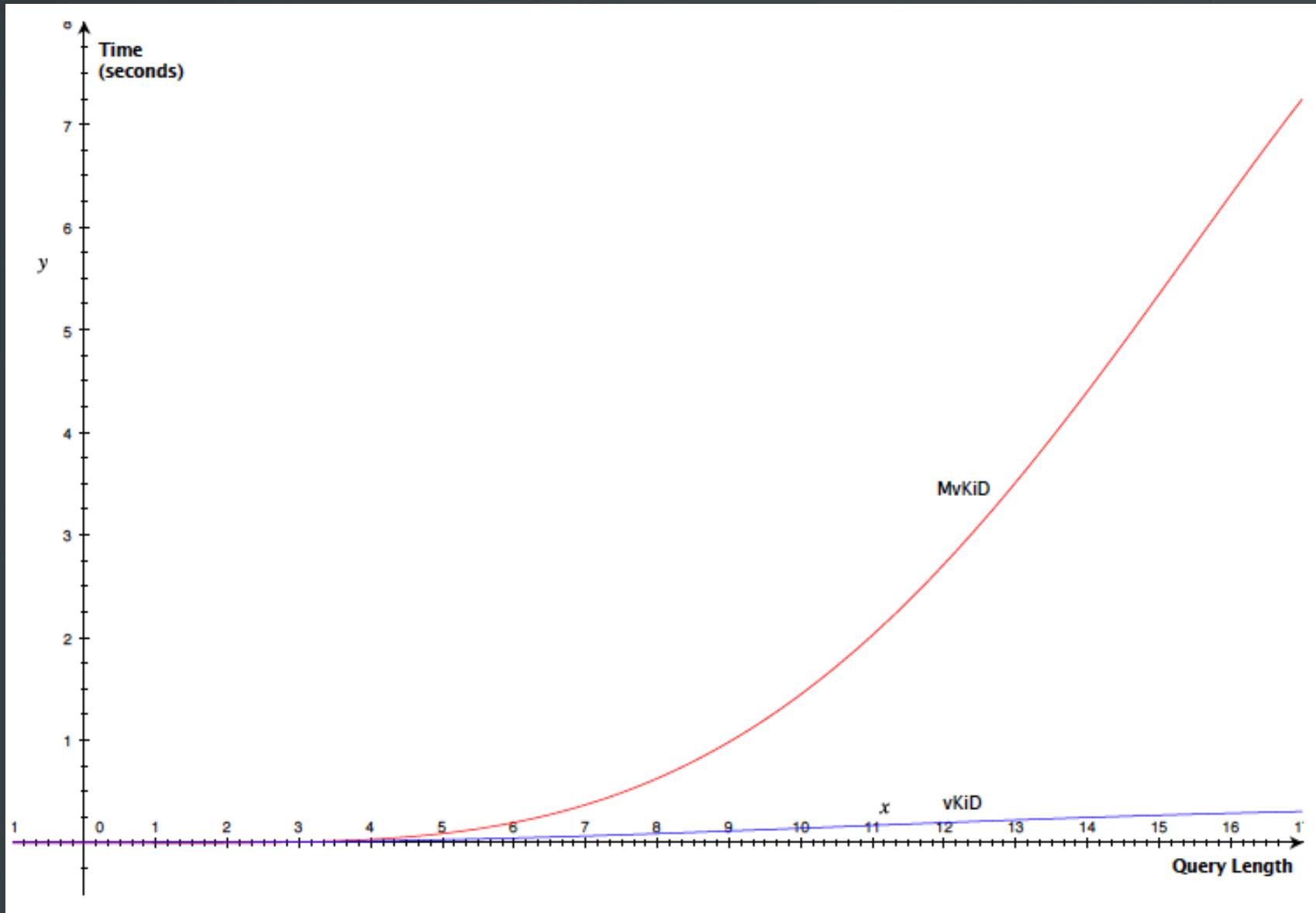


Results - Size

- Single
 - 38,067 entries
 - 9.7 MB
- Multi-Word
 - 15,178,734 entries
 - 38,067 , 2,615,008 , 8,726,517 , 3,799,142
 - 281 MB



Query Speed



Evaluation

crude china
earn italy
cocoa usa
rapeseed japan
oilseed china

coffee colombia
grain china
yen japan
carcass usa
wheat ussr

interest uk
acq uk
corn usa
interest usa
trade japan

Precision & Recall

Small pre-tagged result set					
Query	P@1	P@10	P@50	MAP	Recall
china crude	0.0%	20.0%	4.0%	8.7%	100.0%
china oilseed	0.0%	10.0%	4.0%	7.0%	100.0%
cocoa usa	0.0%	0.0%	4.0%	4.4%	90.0%
earn italy	0.0%	10.0%	2.0%	2.6%	62.5%
japan rapeseed	100.0%	60.0%	12.0%	41.6%	100.0%
Medium pre-tagged result set					
Query	P@1	P@10	P@50	MAP	Recall
carcass usa	0.0%	0.0%	0.0%	0.1%	4.5%
china grain	100.0%	90.0%	36.0%	40.5%	100.0%
coffee colombia	100.0%	100.0%	62.0%	91.0%	100.0%
japan yen	0.0%	20.0%	6.0%	9.7%	100.0%
ussr wheat	100.0%	70.0%	50.0%	51.0%	100.0%
Large pre-tagged result set					
Query	P@1	P@10	P@50	MAP	Recall
acq uk	0.0%	0.0%	8.0%	1.7%	28.9%
corn usa	0.0%	0.0%	2.0%	31.4%	94.2%
interest uk	0.0%	20.0%	18.0%	14.3%	94.1%
interest usa	0.0%	0.0%	0.0%	0.8%	37.7%
japan trade	100.0%	50.0%	26.0%	29.5%	98.8%

Table 1: Precision and Recall for the base system

Precision & Recall

Small pre-tagged result set					
Query	P@1	P@10	P@50	MAP	Recall
china crude	0.0%	20.0%	4.0%	8.9%	100.0%
china oilseed	0.0%	10.0%	4.0%	7.8%	100.0%
cocoa usa	0.0%	0.0%	2.0%	4.0%	90.0%
earn italy	0.0%	10.0%	2.0%	2.7%	62.5%
japan rapeseed	100.0%	60.0%	12.0%	52.5%	100.0%
Medium pre-tagged result set					
Query	P@1	P@10	P@50	MAP	Recall
carcass usa	0.0%	0.0%	2.0%	0.1%	4.5%
china grain	100.0%	90.0%	36.0%	38.8%	100.0%
coffee colombia	100.0%	90.0%	62.0%	84.3%	100.0%
japan yen	0.0%	10.0%	12.0%	15.1%	100.0%
ussr wheat	100.0%	90.0%	48.0%	53.7%	100.0%
Large pre-tagged result set					
Query	P@1	P@10	P@50	MAP	Recall
acq uk	0.0%	0.0%	8.0%	1.7%	28.9%
corn usa	100.0%	60.0%	40.0%	39.7%	94.2%
interest uk	0.0%	30.0%	26.0%	17.1%	94.1%
interest usa	0.0%	0.0%	0.0%	0.8%	37.7%
japan trade	100.0%	40.0%	40.0%	38.0%	98.8%

Table 2: Precision and Recall for the Phrase system

Precision & Recall

Avg(Multi-Word) - Avg(Single-Word)

map	p@1	p@10	p@50	recall
Small pre-tagged result set				
2.3%	0.0%	0.0%	-0.4%	0.0%
Medium pre-tagged result set				
-0.1%	0.0%	0.0%	1.2%	0.0%
Larger pre-tagged result set				
3.9%	20.0%	12.0%	12.0%	0.0%

Table 3: Average Precision and Recall Increase

Questions ?

