Information Retrieval System based on Phrase Indices

Hatem Nassrat
CSCI 6403
December 2008

Introduction

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

Method

- Reuters 21578
- Base + Augmented System
- Inverted Files (TC B-Tree)
- Application in Vector Space
- Retreival - TFIDF

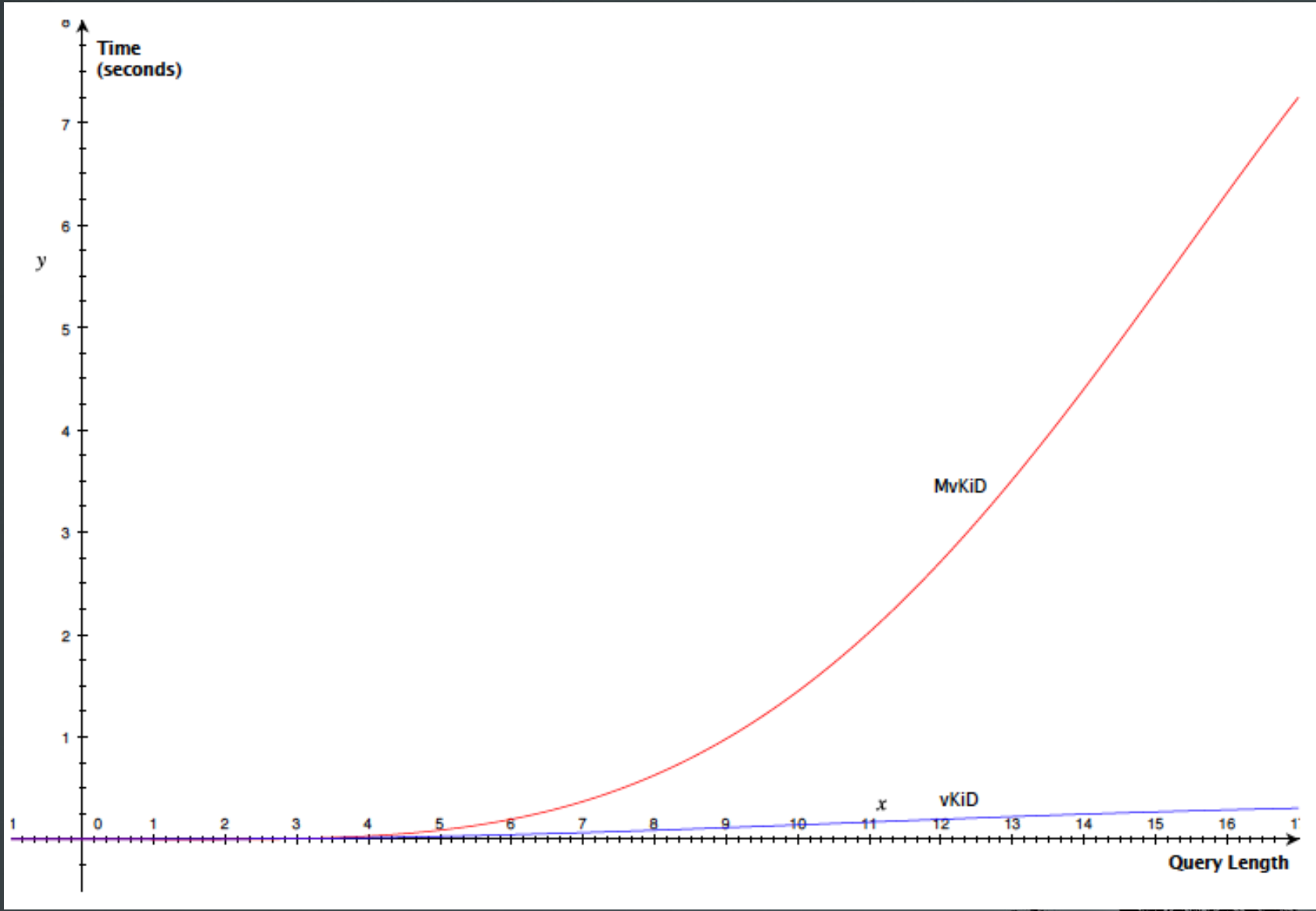
$$1 + \sum_{i=1}^{M} \binom{N-1}{i}$$

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

Results - Size

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

Speed



Evaluation

crude china	coffee colombia	interest uk
earn italy	grain china	acq uk
cocoa usa	yen japan	corn usa
rapeseed japan	carcass usa	interest usa
oilseed china	wheat ussr	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

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 ?