Pivoted Document Length Normalization

Amit Singhal, Chris Buckley, Mandar Mitra (1996)

CS533 Presentation
Abdullah Bulbul
Omer Faruk Uzar

Introduction

- Unfair retrieval due to differing document lengths

- Long documents have advantage
    - Higher term frequency
    - More Terms
    - Therefore;
        - Easier match with queries

- Solution:
    - Normalize document lengths

Common Normalization Techniques

- Cosine Normalization

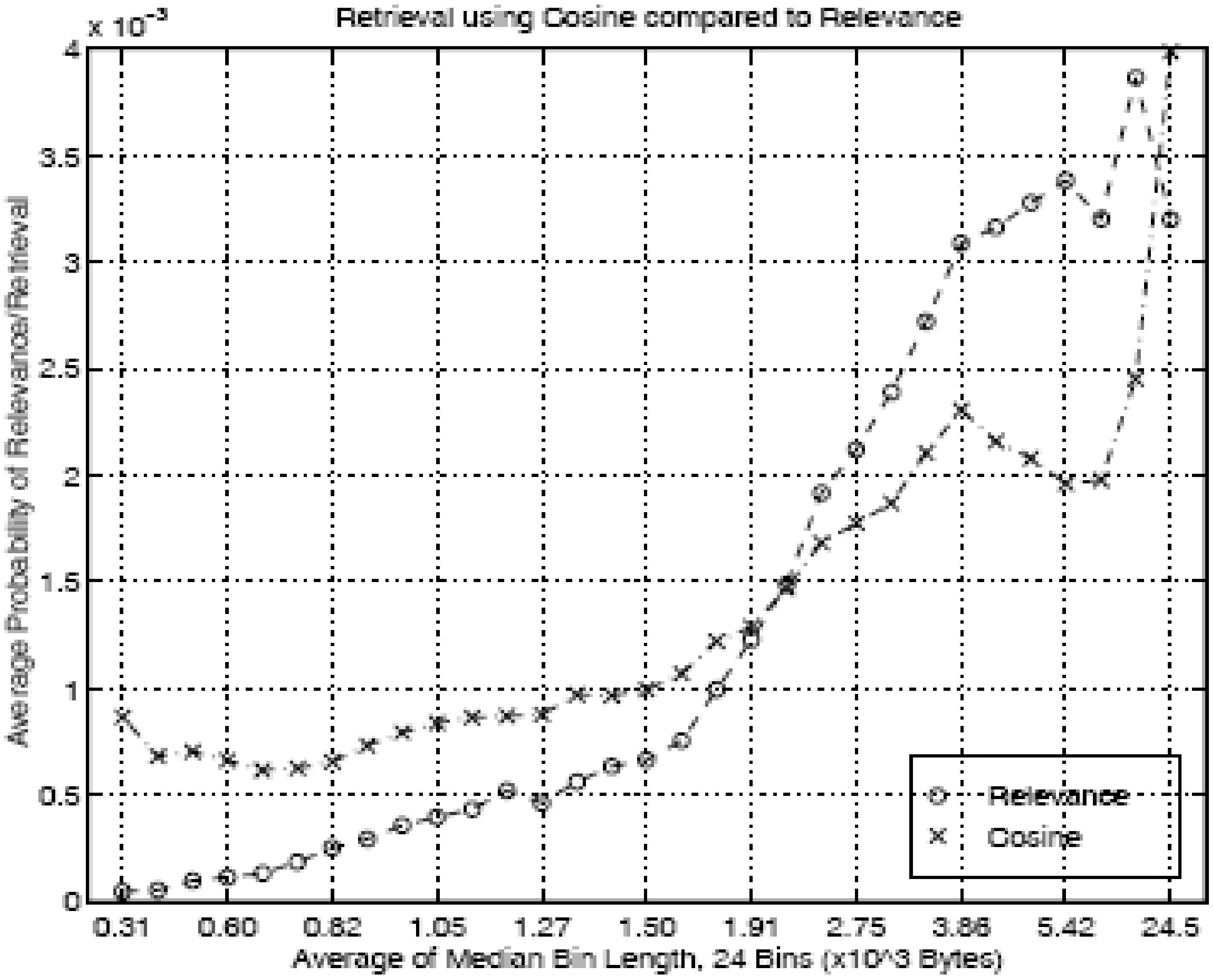
$$\sqrt{{w_1}^2 + {w_2}^2 + \ldots + {w_t}^2}$$

- Maximum tf Normalization

  e.g. $0.5 + 0.5 \times \frac{tf}{max\_tf}$

- Byte Length Normalization

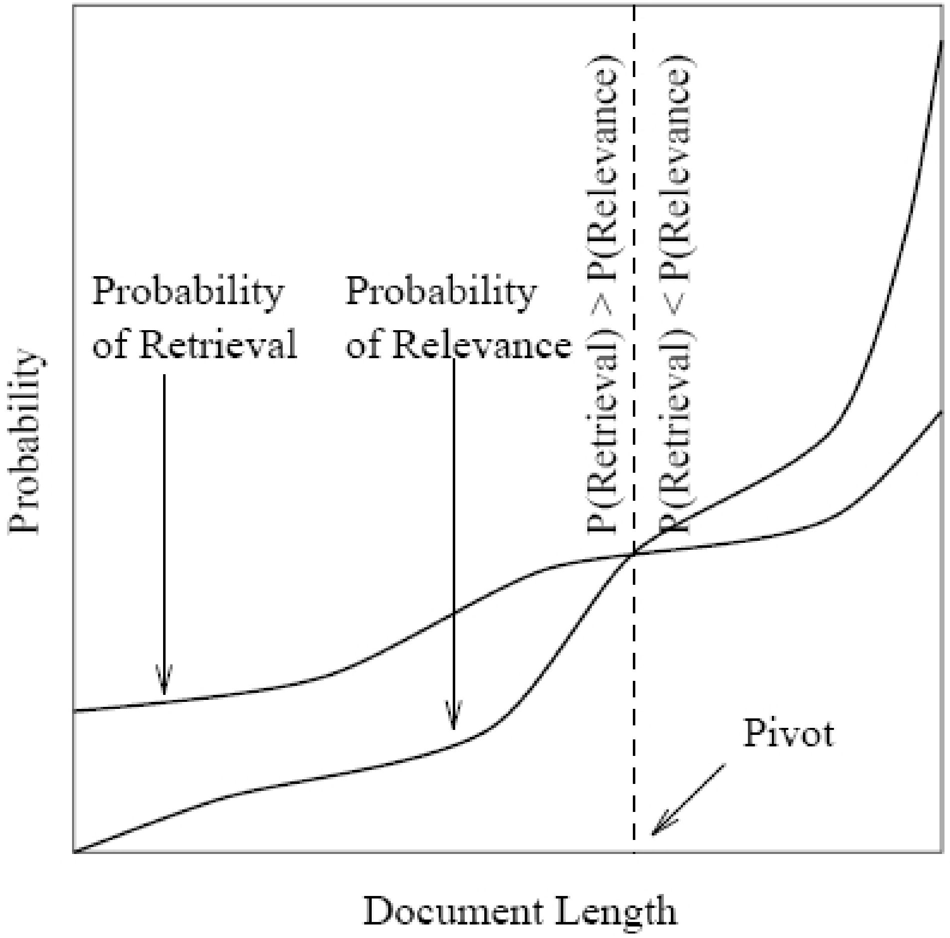
Pivoted Normalization

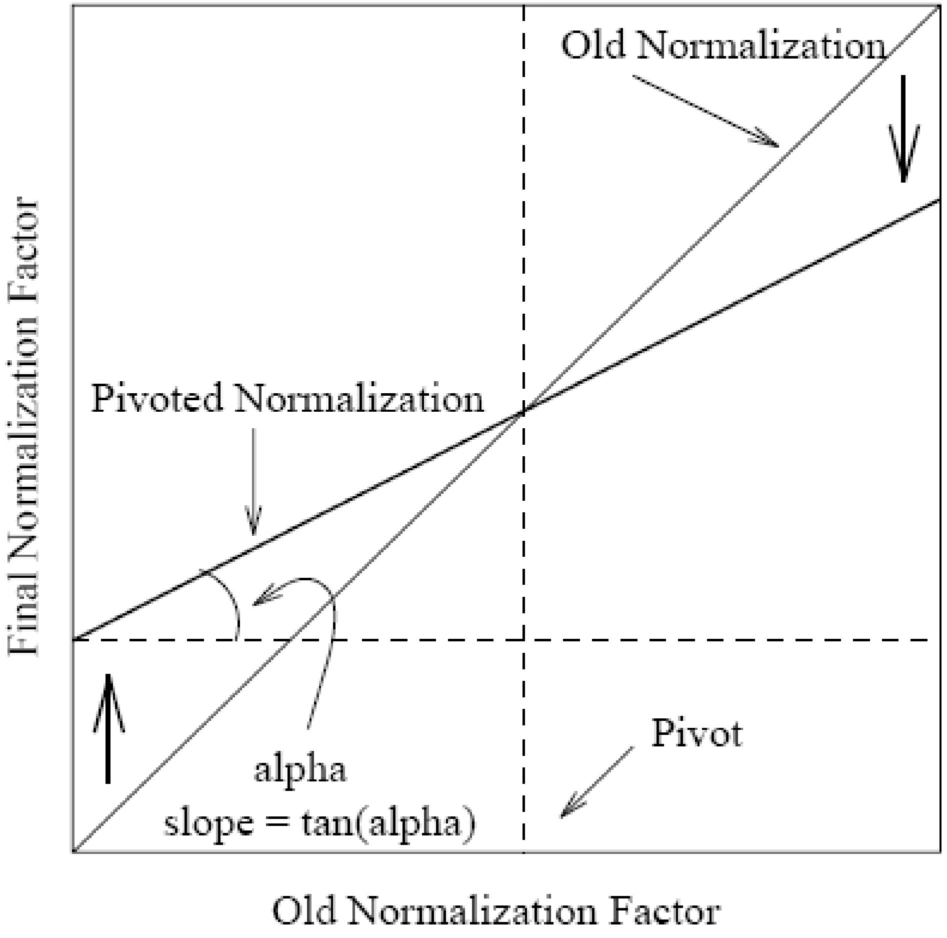


(c)

Pivoted Normalization

a

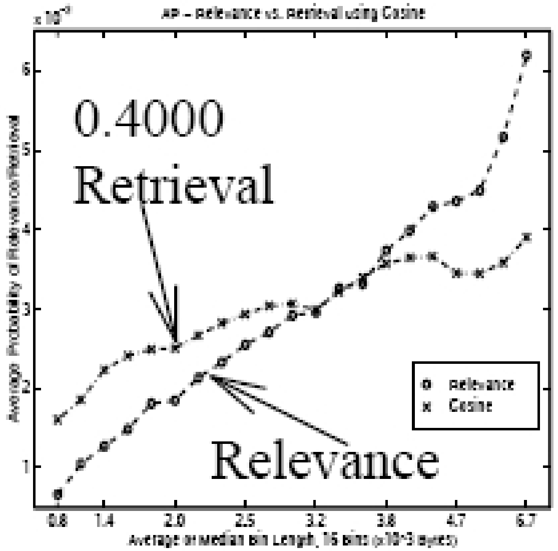




- Pivot
- Tilt

Pivoted Norm. = (1-slope) x pivot + slope x old normalization

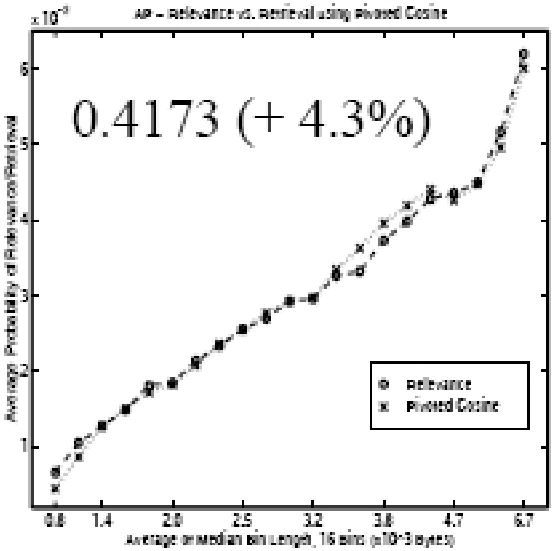
Cosine Norm. vs Pivoted Cosine Norm.



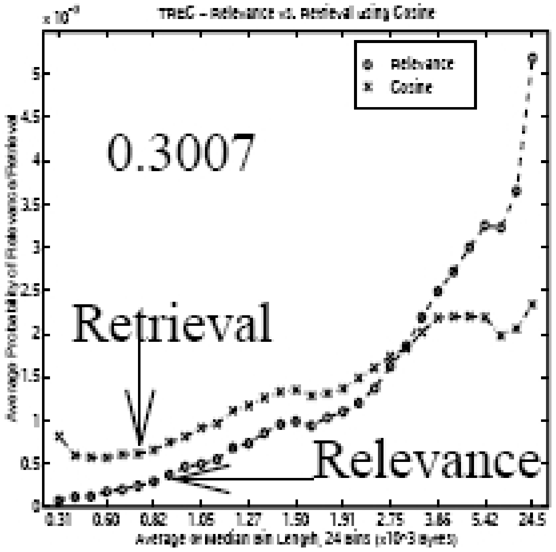
0.4000

Retrieval

Relevance



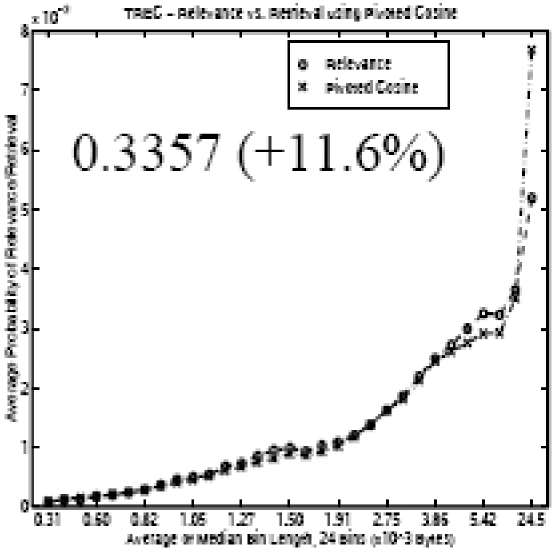
0.4173 (+ 4.3%)



0.3007

Retrieval

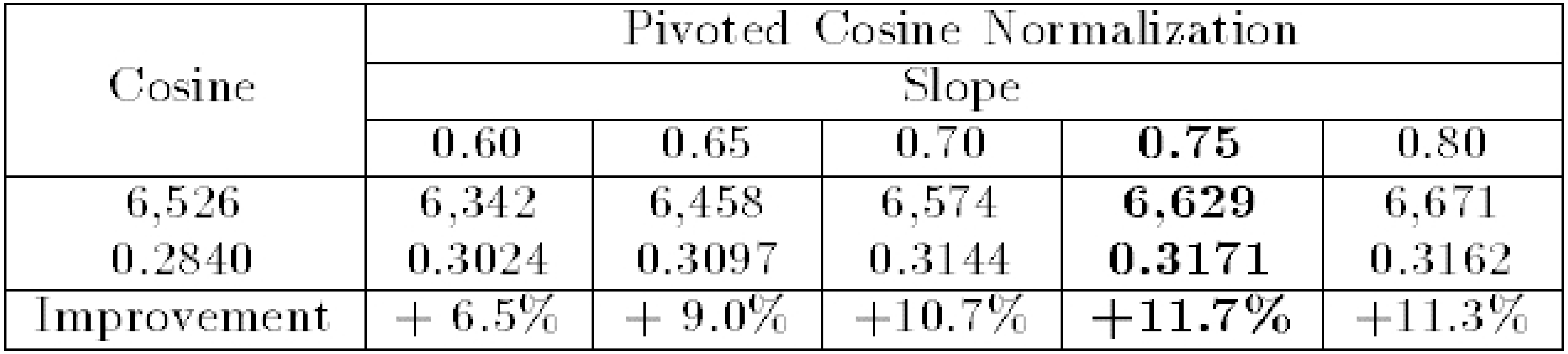
Relevance



0.3357 (+11.6%)

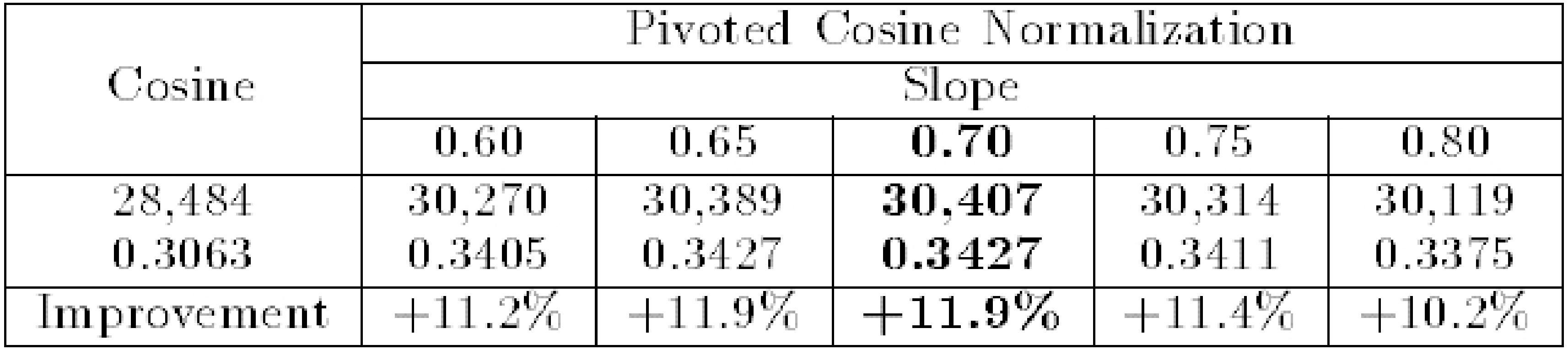
Results

Trec Queries 151-200



Cosine	Pivoted Cosine Normalization				
	Slope				
	0.60	0.65	0.70	**0.75**	0.80
6,526 0.2840	6,342 0.3024	6,458 0.3097	6,574 0.3144	**6,629** **0.3171**	6,671 0.3162
Improvement	+ 6.5%	+ 9.0%	+10.7%	**+11.7%**	+11.3%

Trec Queries 1-150



Cosine	Pivoted Cosine Normalization				
	Slope				
	0.60	0.65	**0.70**	0.75	0.80
28,484 0.3063	30,270 0.3405	30,389 0.3427	**30,407** **0.3427**	30,314 0.3411	30,119 0.3375
Improvement	+11.2%	+11.9%	**+11.9%**	+11.4%	+10.2%

End of Presentation

Thank you