Queen Mary
University of London

Literature Analysis

Prepared for Information Retrieval

Prepared by: Danielle Souza Da Silva, Georgia Dean, Adam Toth

February 17, 2022

Queen Mary University of London

Table of Contents

Document Retrieval Problem Formulation

Queen Mary University of London

Explanation of specific issue researchers have set out to solve

Term weighting

- Term frequency
- Inverse document frequency
- Document length normalization

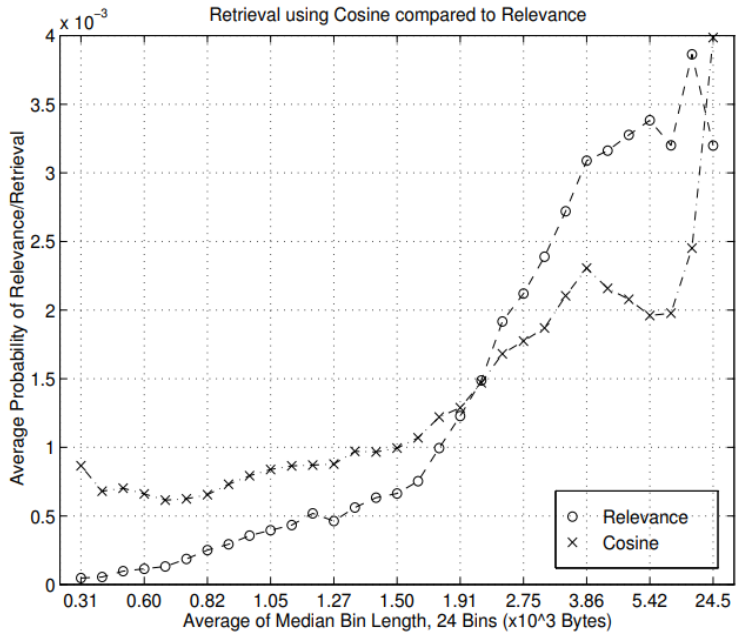
Document length normalization

- Fairly retrieve documents of all lengths
- Longer documents have an advantage
  - Higher term frequencies
  - More terms
- Normalization strategies reduce term weights based on document length
  - Cosine normalization
- Proposed normalization method
  - Pivoted normalization using cosine

Cosine Normalization

Workings of cosine normalization and results

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



(c)

Likelihood of Relevance and Retrieval

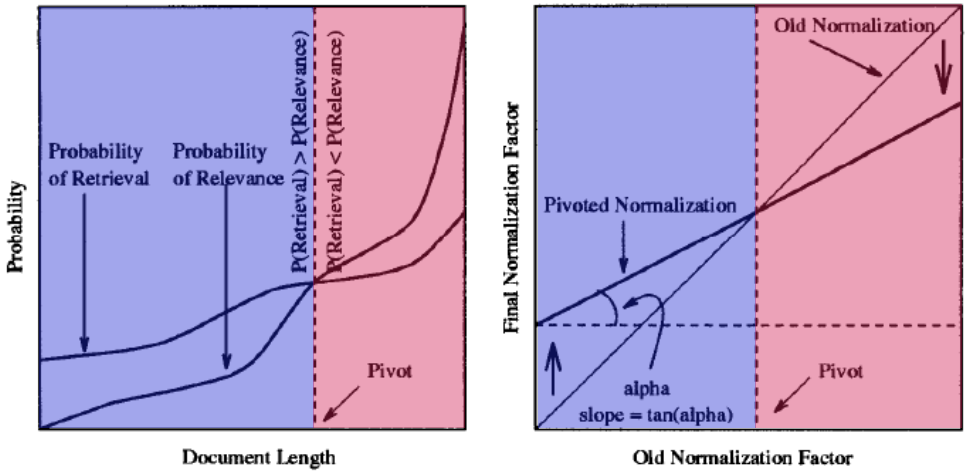
- Using <query, Relevant-document pairs> and binning method the likelihood of relevance and likelihood of retrieval are found

Cosine Normalization

- Figure C shows smoothed curves for the probability of relevance and the cosine normalized probability of retrieval
- The results show that there is a gap between the relevance and retrieval probabilities
  - Researcher aimed to use *pivoted normalization* to bring the likelihood of retrieval closer to relevance

Pivoted Normalization Function

Mechanics of the pivoted normalization



= P(Retrieval) > P(Relevance)
= P(Retrieval) < P(Relevance)

$$pivoted\ normalization =$$

$$(1.0 - slope) \times pivot + slope \times old\ normalization \quad (1)$$

Normalization factors

- Inversely related to the probability of retrieval
  - Increasing the normalization factor decreases the probability of retrieval

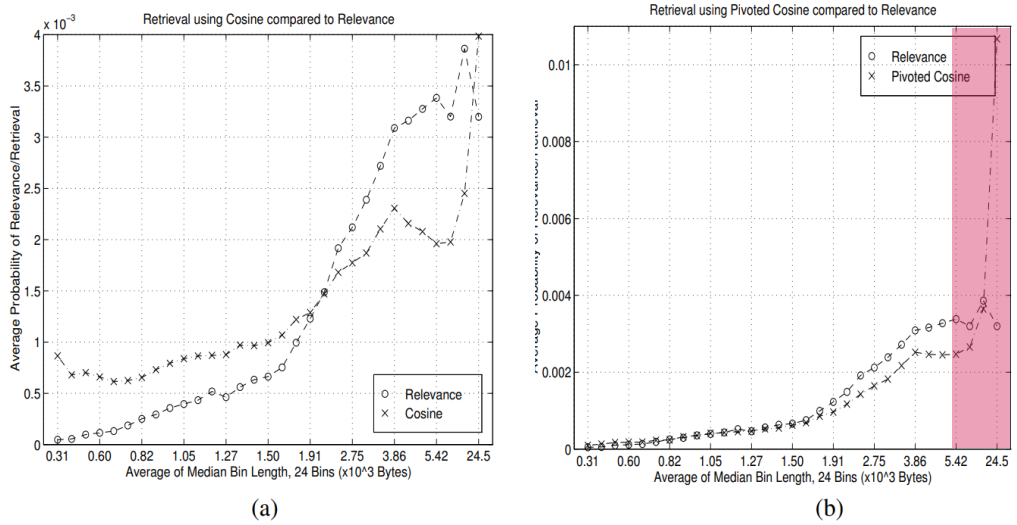
Pivoted normalization

- Pivot point is where P(Retrieval) intersects P(Relevance)
- New normalization factors are achieved by 'tilting' the slope of the old normalization factors around the pivot point
  - Red area -> lower normalization factors
  - Blue area -> higher normalization factors

Results of Pivoted Normalization

Results show cosine normalization is effective in aligning relevance and retrieval

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



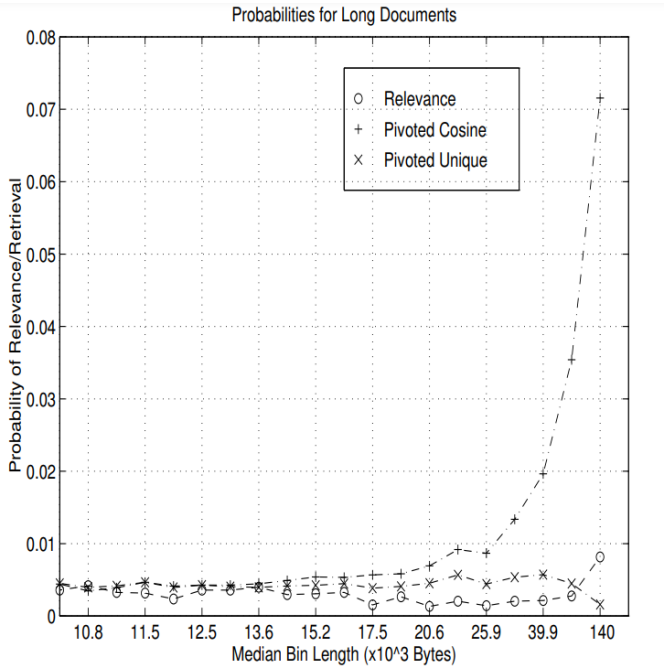
Results

- Improvement of 9-12% in average precision
- Figure displays smoothed results
  - Pivoted cosine normalization curve is closer to likelihood of relevance
- Support hypothesis: ‘schemes that retrieve different length documents with chances similar to their likelihood of relevance have higher retrieval effectiveness’

- Note: There continues to be a large gap in the likelihood of relevance and retrieval for extremely long documents

Extremely Long Documents and Unique Normalization

Use of pivoted unique normalization to manage extremely long documents



$$\frac{\frac{1+log(tf)}{1+log(average\ tf)}}{(1.0-slope)\times pivot + slope \times \#\ of\ unique\ terms}$$

Extremely long document issue

- Document that is longer than 20,000 bytes
- Pivoted cosine normalization for these documents had a very high likelihood of retrieval vs relevance

Pivoted unique normalization

- Average # of unique terms variable added to pivoted normalization to remove advantage of extremely long documents
- Average term frequency is added to initial equation to compensate for higher tfs in long documents

Results

- Figure shows results for pivoted cosine and pivoted unique cosine
- Pivoted unique cosine normalization is more closely aligned with likelihood of retrieval curve

Improving Term Frequency Normalization for Multi-topical Documents, and Application to Language Modelling Approaches

Queen Mary University of London

Research Paper 2 - [Seung-Hoon Na, In-Su Kang, Jong-Hyeok Lee, 2008]

Issue being addressed

- Multi-topical documents term frequency normalization should be differently handled compared to verbose document term frequency normalization. This is because the terms in a multi-topical document are less repeated than the ones in a verbose document.

- The paper explains that a retrieval function adopting Singhal's penalisation will make multi-topical documents unreasonably less-preferred, causing an unfair retrieval ranking.

Proposed Solution

- A novel TF normalisation method which is a type of partially-axiomatic approach.

Approach

1. Formulate two constraints that the retrieval scoring functions should satisfy for verbose and multi-topical documents.
2. Present the analysis result that previous language modelling approaches do not sufficiently satisfy these constraints.
3. Modify the language modelling approaches such that better satisfy these two constraints, derive a novel smoothing methods, and evaluate the proposed ones.

Multi-topicality vs Verbosity Problem

Impact of long documents being either multi-topical or verbose on relevance and retrieval

Original Document

Verbose Document

- Longer document covering small range of topics, so a small range of terms are *repeated frequently*

Multi-topicality Document

- Longer document covering a broad range of topics, so a broad range of terms are *repeated less*

D_1: Language modelling approach

D_2: Language modelling approach
Language modelling approach

D_3: Information retrieval model
Language modelling approach

Query: *"language modelling approach"*

In order to derive the same relevance score for all three documents for this query, **constraints need to be applied to the term frequency normalization function**

Modelling Approach and Constraint Development

Explanation

Language modelling approach

- Models the probability of a particular sequence of text by multiplying the probability associated with each term in the sequence

$$P(t|D) = P(t|\theta_D) = \frac{tf_{t,D}}{N_D}$$

Query-Likelihood Retrieval Model

- Ranks documents based on the probability they are related to the query
  - Scores each document according to the probability provided by the language model given the query

$$score(Q, D) = P(Q|\theta_D) = \prod_{i=1}^{n} P(q_i|\theta_D)$$

Constraints

VNC (Verbosity Normalization Constraint)

- Suppose a pair of D_1 and D_2. If D_2 is verbose to D_1, then *score (Q, D_1) = score (Q, D_2)*

TNC (Topicality Normalization Constraint)

- Suppose a pair of D_1 and D_2. If D_2 is N-topical to D_1, then *score (Q, D_1) = score (Q, D_2)*

Modification - Jelinek-Mercer (JM) Smoothing

Breakdown of modification made to satisfy previously set out constraints

Core idea for first modification: a pseudo document (for correcting TNC)

- Score of a document is calculated using the pseudo document model instead of original document model.
- Probability of query terms in a pseudo document is estimated by using probability of original document
- First modification produced final modified JM, which is called **JMV**, which satisfies VNC & TNC.

$$score(Q,D) = \sum_{w \in Q} \log\left(\frac{1-\lambda}{\lambda}\tau'(D)P(w|\hat{\theta}_D)\frac{l_C}{tf_C(w)} + 1\right)$$

Second modification: to handle JMV verbose-type queries issue

- To handle the problems of verbose-type queries, the TF normalization should be restricted to only document-specific terms.
- A term specificity of w in document D is defined using a probabilistic metric P(D,w)
- Using the term specificity, the pseudo document model, called **JMV2**, is modified as follows:

$$P(w|\theta_{Pseudo(D)}) = K \cdot tf_D(w) \cdot \tau'(D)^{P(D|w)} / l_D$$

JMV2 Smoothing Results

Results of smoothing compared to benchmark method

MAP	Dir			JMV2		
	SK	SV	LV	SK	SV	LV
TREC7	0.1786	0.1790	0.2209	0.1825	**0.1926**†	0.2250
TREC8	0.2481	0.2294	0.2598	**0.2505**†	**0.2354**†	0.2500
WT2G	0.3101	0.2854	0.2863	**0.3278**‡	**0.3112**‡	**0.3263**‡
TREC9	0.2038	0.1990	0.2468	**0.2068**	**0.2245**‡	0.2494
TREC10	0.1950	0.1865	0.2347	0.2091	**0.2133**†	0.2555

Pr@5	Dir			JMV2		
	SK	SV	LV	SK	SV	LV
TREC7	0.4400	0.4280	0.5240	0.4680	**0.4920**†	**0.5800**†
TREC8	0.4920	0.4320	0.5120	**0.5240**‡	0.4880	0.5280
WT2G	0.5160	0.5120	0.5280	0.5400	0.5560	**0.5920**†
TREC9	0.3000	0.3480	0.4160	0.3440	0.3720	0.3880
TREC10	0.3520	0.4040	0.4720	0.3800	0.4200	0.4880

Pr@10	Dir			JMV2		
	SK	SV	LV	SK	SV	LV
TREC7	0.3980	0.4120	0.4420	0.4100	0.4440	**0.4800**†
TREC8	0.4460	0.4120	0.4660	**0.4700**†	0.4400	0.4480
WT2G	0.4660	0.4220	0.4240	0.4920	**0.4900**‡	**0.4820**‡
TREC9	0.2560	0.2860	0.3160	0.2780	**0.3160**†	0.3220
TREC10	0.3060	0.3500	0.4040	0.3300	0.3700	0.4340

†: *test passes at 95% confidence level*
‡: *test passes at 99% confidence level*

Results for JMV2

- Dir = benchmark smoothing method used for comparison
- JMV2 significantly improved JM for all query types
- JMV2 substantially improves MAP (Mean Average Precision) for verbose queries.
- Most significant improvement occurs for the short verbose and long verbose queries

Query types:

1. **Short keyword (SK)**: using only the title of the topic description
2. **Short verbose (SV):** using only the description field (usually one sentence)
3. **Long verbose (LV)**: using the title, description and narrative

Queen Mary University of London

Our Views

Considerations for future document retrieval strategies

Pivoted normalization limitations

- Requires continued research into variable such as # of unique terms
- Content and document type could be considered when applying normalizations
  - For example, a long blog post and long legal document should not necessarily have the same normalizations applied

Multi-topical document normalization

- Partially-axiomatic approaches can be an effective method to modify certain constraints to improve found issues in a retrieval model.

Sources

Na, S., Kang, I., & Lee, J. (2008). Improving Term Frequency Normalization for Multi-topical Documents and Application to Language Modeling Approaches. *ECIR*.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *SIGIR Forum, 51*, 176-184.