Two-stage document length normalization for Information retrieval

Seung-Hoon Na

Introduction

- Two-Stage Document Length Normalization for Information Retrieval. *ACM Transactions on Information Systems* , 33(2), 2015

- URL: http://dl.acm.org/citation.cfm?id=2699669

Why Normalization of Term Frequency?

- **Term frequency**
  - a fundamental and important component of a ranking model
- **The naïve scoring method using term frequency**
  - Intuition: The larger the term frequency of a query word in a document, the more likely the document is to be about the query topic

➔ **Problem: an excessive preference for long documents**

Normalization of term frequency is necessary!

Two Hypotheses
[Robertson and Zaragoza ’09]

- **Verbosity hypothesis**
  - Some authors are simply more verbose, using more words to say the same thing
- **Scope hypothesis**
  - Some authors have more to say: they may write a single document containing or covering more ground

Issue

- **We focus on the difference b/w the effect of the verbosity and the scope on the term frequency of a single word**

Verbosity vs. Scope

- **Verbosity**
  - Related to the **burstiness** of term frequency
  - Helps an already mentioned word in a document get a higher frequency
  - Normal verbosity vs. high verbosity
    - Even if a word has a low term frequency in normal verbosity, **its term frequency could increase significantly when the document has high verbosity**

Verbosity vs. Scope

- **Scope**
  - Mostly involves **the creation of a new word**, rather than boosting the term frequency
  - Broadening the scope of a document would help unseen words in a normal document get non-zero frequencies ➔ However, these non-zero frequencies might not be high.

Summary of the difference b/w verbosity and scope

verbosity ***leads to a significant increase in term frequency, whereas*** ***scope*** ***leads to a rather limited increase in term frequency.***

Limitation of Existing Standard Normalization

- **Length-driven approach**
  - Based only on the document length, without distinguishing between verbosity and scope
  - Limitation

insufficient penalization of a verbose document whose length is increased mainly by high verbosity

Excessive penalization of a broad document whose length is mainly derived from the broad scope

Proposal: Two-stage normalization

Verbosity and scope should be normalized separately by employing different penalization functions

1) Verbosity normalization

- For each document, **linearly divide the term frequency by the verbosity**, thus obtaining a **verbosity-normalized document representation**.

2) Scope normalization

- An existing retrieval model is applied to this verbosity-normalized document representation.

Verbosity-normalized (VN) retrieval model

Analysis on Two-stage Normalization

- We perform **comparative axiomatic analysis** of the original and the VN retrieval models, **under the setting of the axiomatic framework** introduced in [Fang et al. 2004; 2011]

➔

- the VN model indeed performs the desired separate normalizations

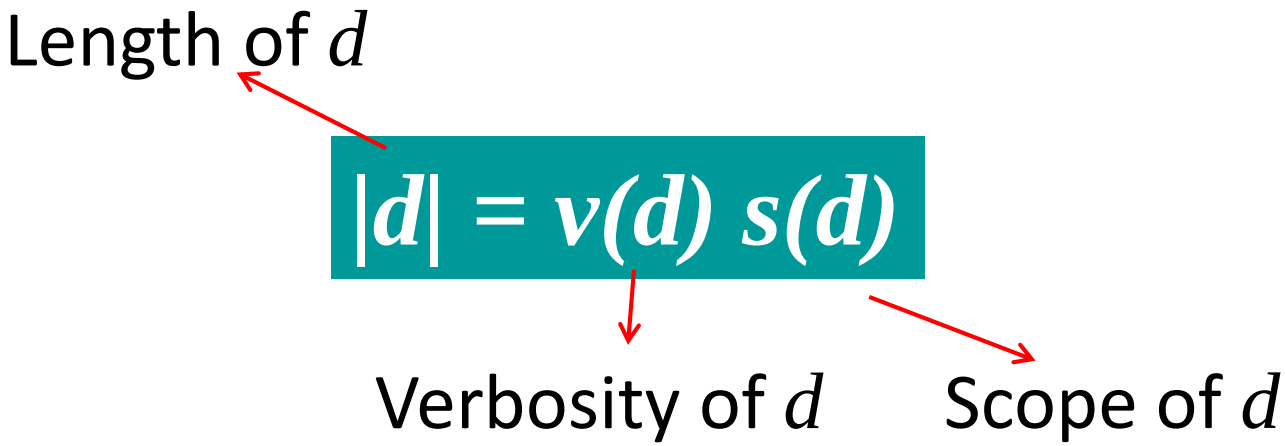
1) a strict penalization of verbosity-increased documents
2) a relaxed penalization of scope-broadened documents.

Two-Stage Normalization

- **The verbosity and the scope hypotheses**

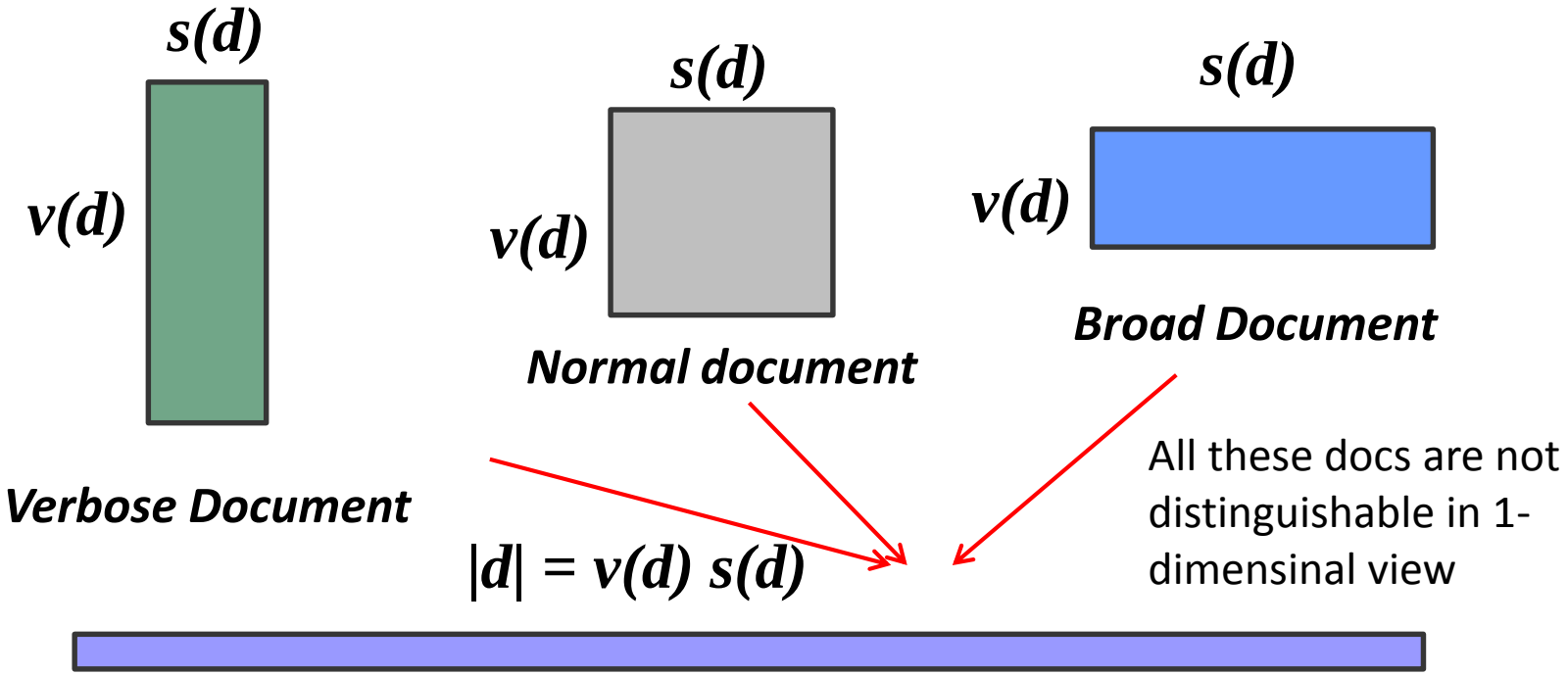
  ➔

  We assume that the document length is decomposed into the verbosity and the scope as



Two-Stage Normalization

- Document length
  - Previously, regarded as 1-dimenational object
  - In our work, regarded as 2-dimensional object



Two-Stage Normalization

- **Verbosity of d** $v(d)=\frac{|d|}{s(d)}$
- **Verbosity normalization**
  - the original term frequency is normalized by dividing it by the verbosity of the document

Original term freq.

$$c\big(w,\phi(d)\big)=k\frac{c(w,d)}{v(d)}=k\frac{c(w,d)\cdot s(d)}{|d|}$$

  - φ: *verbosity normalization operator*
  - $\varphi(d)$: the verbosity-normalized document representation of d
  - $c(w, d)$: the original term frequency of word w
  - $c(w, \varphi(d))$: the verbosity-normalized term frequency of word

Two-Stage Normalization

- **Scope normalization**
  - we need to consider a more relaxed function than that for verbosity normalization

the scope of an original document **=**
the verbosity-normalized length of the document

$$|\phi(d)| = \sum_w c(w, \phi(d)) = \sum_w \frac{c(w,d) \cdot s(d)}{|d|} = s(d)$$

existing retrieval models perform a type of relaxed normalization by using their pivoted length or smoothed length

Two-Stage Normalization: Summary

- **$f(d, q)$: the original retrieval function that gives a score to d,**
- **Applying two-stage normalization to $f(d, q)$**
  ➔ $f(\varphi(d), q)$
- **$f(\varphi(d), q)$: Obtained by replacing $c(w, d)$ used in all terms in $f(d, q)$ with $c(w, \varphi(d))$ for all documents in the collection**

VN (verbosity-normalized) retrieval model

Examples of VN model: Dirichlet-prior (DP)

- **The VN model f(ϕ(d), q) is assumed to employ the following document-specific conjugate prior:**

$$(\mu v(d)\, p(w_1|C), \mu v(d)\, p(w_2|C), \mu v(d) p(w_{|V|}|C))$$

$$P(w|d) = \frac{c(w,d) + \mu\, v(d) p(w|C)}{|d| + \mu\, v(d)}$$

We simply use *k = 1*, because the scaling parameter *k* of *c*(*w*, (*d*)) is absorbed into the smoothing parameter μ.

$$P(w|\phi(d)) = \frac{c(w,\phi(d)) + \mu\, p(w|C)}{|\phi(d)| + \mu}$$

Examples of VN model: Dirichlet-prior (DP)

- The resulting VN scoring function (**VN-DP**)

$$\sum_{w \in q \cap d} c(w,q) \ln\left(1 + \frac{c(w,d)}{\mu \cdot p(w|C)} \frac{s(d)}{|d|}\right) + |q| \cdot \ln\left(\frac{\mu}{s(d)+\mu}\right)$$

Examples of VN model: Okapi

- Okapi's BM25 retrieval formula, as presented by [Robertson et al. 1995]

$$tf_{BM25}(w,d) = \frac{(k_1+1)c(w,d)}{k_1\left((1-b)+b\frac{|d|}{avgl}\right)+c(w,d)}$$

we assume the scale parameter *k* to be 1, because it is absorbed into k_1

$$tf_{BM25}(w,\phi(d)) = \frac{(k_1+1)c(w,d)}{k_1|d|\left((1-b)\frac{1}{s(d)}+b\frac{1}{avgs}\right)+c(w,d)}$$

Scope Measure

- The remaining problem is how to compute the scope of a document $s(d)$.

- We propose three different approaches
  - 1) Length power (**LengthPower**)
  - 2) The number of unique terms (**UniqLength**)
  - 3) Entropy power (**EntropyPower**)

Scope Measure: Length Power

- To obtain a length-based scope measure, we use Heap's law, which is given as follows [Heaps 1978]

$$l_\beta(d) = |d|^\beta$$

  - The possible range of β: $0 \leq \beta \leq 1$

Scope Measure: the Number of Unique Terms

- the number of unique terms $u(d)$, defined as

$$u(d) = |\{w|w \in d\}|$$

- a different topic is described using a domain-specific vocabulary or named entities. The more unique terms used in a document, the larger is the scope of the document

Scope Measure: Entropy Power

- the entropy power defined by the exponential of the entropy, which was initially exploited in [Kurland and Lee 2005] to construct the document structure

$$h(d) = \begin{cases} exp\left(-\sum_{w} p_{ml}(w|d)\ln(p_{ml}(w|d))\right) & if\ |d| \geq 1 \\ 0 & otherwise \end{cases}$$

Comparative Axiomatic Analysis under Standard Retrieval Constraints

- We perform a comparative axiomatic analysis performed under the standard retrieval constraints introduced

- Standard constraints (Fang et al. '04;'11)
  - Form constraints: TFC1, TFC2, TFC3, and TDC
  - Normalization constraints: LNC1, LNC2, and TF-LNC

Seven Basic Relevance Constraints

[Fang et al. 2011]

Constraints	Intuitions
TFC1	To favor a document with more occurrences of a query term
TFC2	To ensure that the amount of increase in score due to adding a query term repeatedly must decrease as more terms are added
TFC3	To favor a document matching more distinct query terms
TDC	To penalize the words popular in the collection and assign higher weights to discriminative terms
LNC1	To penalize a long document (assuming equal TF)
LNC2, TF-LNC	To avoid over-penalizing a long document
TF-LNC	To regulate the interaction of TF and document length

Slide from http://www.eecis.udel.edu/~hfang/pubs/sigir14-axiomatic.pptx

Length Normalization Constraints (LNCs)

Document length normalization heuristic:

Penalize long documents(LNC1);

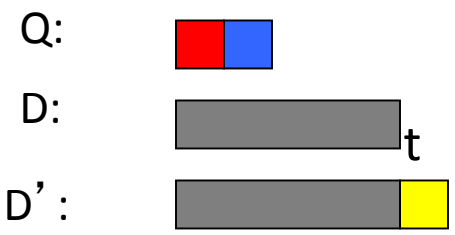
Avoid over-penalizing long documents (LNC2) .

- ***LNC1***

  Let *Q* be a query and *D* be a document.

  If *t* is a non-query term,

  then $S(D \cup \{t\}, Q) < S(D, Q)$



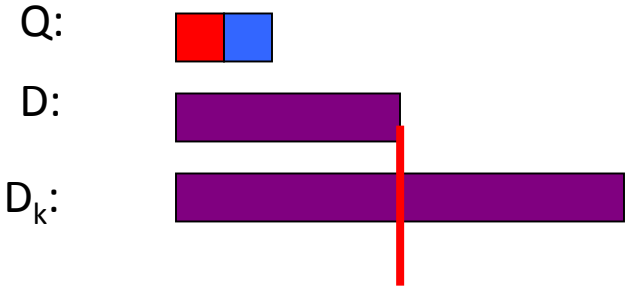
$S(Q, D') < S(Q, D)$

- ***LNC2***

  Let *Q* be a query and *D* be a document.

  If $D \cap Q \neq \phi$, and D_k is constructed by concatenating *D* with itself *k* times,

  then $S(D_k, Q) \geq S(D, Q)$



$S(Q, D_k) \geq S(Q, D)$

Slide from http://www.eecis.udel.edu/~hfang/pubs/sigir14-axiomatic.pptx

TF & Length normalization Constraint (TF-LNC)

TF-LN heuristic:

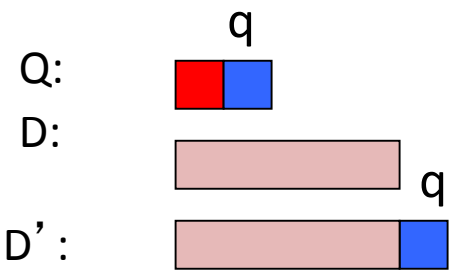
Regularize the interaction of TF and document length.

- *TF-LNC*

  Let *Q* be a query and *D* be a document.

  If *q* is a query term,

  then $S(D \cup \{q\}, Q) > S(D, Q)$.



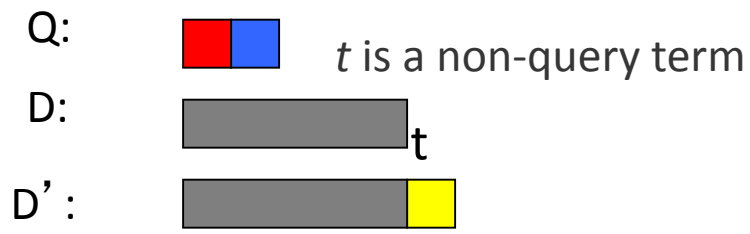
$$S(Q, D') > S(Q, D)$$

Slide from http://www.eecis.udel.edu/~hfang/pubs/sigir14-axiomatic.pptx

Analysis Results of the Original and VN Retrieval Models for Normalization Constraints

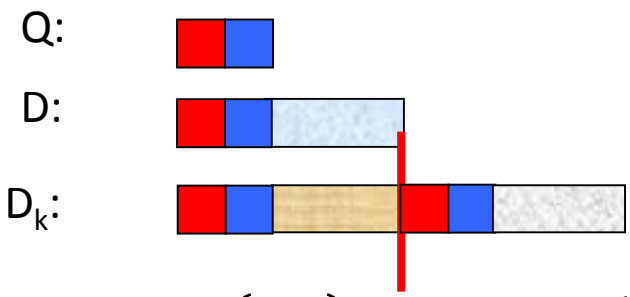
	LNC1	LNC2	TF-LNC
Original model	Yes	yes	yes
Verbosity normalized (UniqLength)	C1	C2	yes
Verbosity normalized (EntropyPower)	C1	C2	yes

- *C1*



$$Verbosity(D') \geq Verbosity(D)$$

- *C2*



$$Scope(D_k) \geq Scope(D)$$

- an original method satisfies all three constraints unconditionally
- a VN method requires additional conditions that depend on the choice of scope measure

Normalization Heuristics of VN Models

(UniqLength and EntroyPower)

- H1: Relaxed penalization of scope-broadened documents
  - The VN retrieval method performs a relaxed penalization of a scope-broadened (from LNC1 & C1)
- H2: Strict penalization of verbosity-increased documents
  - The VN retrieval method imposes a strict penalization of a verbosity-increased document (from LNC2 & C2)

Experimentation

Experimental Setting

- Test Collections: ROBUST, WT10G, GOV2

Statistics	ROBUST	WT10G	GOV2
NumDocs	528,156	1,692,096	25,205,179
NumWords	572,180	6,346,858	40,002,579
TopicSet	Q301-450, Q601-700	Q451-550	Q701-850
Avg of |d| (CoeffVar)	233.34 (2.39)	400.25 (6.06)	690.8 (2.86)
Avg of h(d) (CoeffVar)	107.77 (0.81)	109.60 (1.45)	109.85 (0.98)
Avg of v(d) (CoeffVar)	1.77 (0.91)	2.95 (5.51)	6.11 (7.17)

DP vs. VN-DP: MAP

Type	Method	ROBUST	WT10G	GOV2
Short Keyword Queries	baseline	0.2447	0.1963	0.2907
	LengthPower(0.25)	0.2252	0.1649	0.2403
	LengthPower(0.5)	0.2401	0.1953	0.2823
	LengthPower(0.75)	0.2457	0.1969	0.2930
	LengthPower(0.9)	0.2460*	0.1963	0.2913
	UniqLength	0.2472*	0.2046	0.3055*
	EntropyPower	**0.2481***	**0.2120***	**0.3099***
Long Verbose Queries	baseline	0.2707	0.2469	0.2864
	LengthPower(0.25)	0.2697	0.2249	0.3060*
	LengthPower(0.5)	0.2765*	0.2506	0.3133*
	LengthPower(0.75)	0.2762*	0.2532	0.3005*
	LengthPower(0.9)	0.2725*	0.2501	0.2914*
	UniqLength	0.2759*	0.2553*	0.3083*
	EntropyPower	**0.2799***	**0.2614***	**0.3248***

Okapi vs. VN-Okapi: MAP

Type	Method	ROBUST	WT10G	GOV2
Short Keyword Queries	baseline	0.2447	0.1963	0.2920
	LengthPower(0.5)	0.2451	0.1957	0.2897
	LengthPower(0.75)	0.2454	0.1994*	0.2923
	LengthPower(0.9)	0.2452	0.1944	0.2923
	UniqLength	**0.2483***	0.1997	**0.3035***
	EntropyPower	0.2477*	**0.2071***	0.3004*
Long Verbose Queries	baseline	0.2419	0.2344	0.3012
	LengthPower(0.5)	0.2647*	0.2307	0.3022
	LengthPower(0.75)	0.2640*	0.2341	0.3009
	LengthPower(0.9)	0.2631	0.2366	0.3018
	UniqLength	**0.2663***	0.2368*	0.3063*
	EntropyPower	0.2659*	**0.2415***	**0.3074***

MRF vs. VN-MRF: MAP

Type	Method	ROBUST	WT10G	GOV2
Short Keyword Queries	baseline	0.2447	0.1963	0.2907
	baseline(VN-DP)	0.2481	0.2120	0.3099
	baseline(MRF)	0.2545	0.2149	0.3095
	LengthPower(0.5)	0.2506	0.2055	0.3032
	LengthPower(0.75)	0.2557*	0.2128	0.3133*
	LengthPower(0.9)	0.2545*	0.2142	0.3125*
	UniqLength	0.2572*	0.2244*	0.3270*
	EntropyPower	**0.2581***	**0.2296***	**0.3334***
Long Verbose Queries	baseline	0.2707	0.2469	0.2864
	baseline(VN-DP)	0.2799	0.2614	0.3248
	baseline(MRF)	0.2813	0.2613	0.3164
	LengthPower(0.5)	0.2866*	0.2581	0.3368*
	LengthPower(0.75)	0.2883*	0.2659	0.3280*
	LengthPower(0.9)	0.2861*	0.2617	0.3214*
	UniqLength	0.2895*	0.2687*	0.3363*
	EntropyPower	**0.2927***	**0.2757***	**0.3481***

Conclusion

- **Argument: a normalization function should use different penalizations for verbosity and scope**
- **Proposal: we propose the use of two-stage normalization.**
- **Main contributions**
  - 1) Generalize two-stage normalization such that it can be applied to any retrieval model.
  - 2) Perform comparative axiomatic analysis and capture the exact retrieval heuristics resulting from two-stage normalization and its difference from the original method.