BM25 Extension to Field Weights

By ChungHao Ku, Yinghao Qin

BM25-CTF Improvement

By Jingye Shang, Shangyu Luo

Group E

Simple BM25 Extension to Multiple Weighted Fields - Background

- Conference paper (January 2004) by Stephen E. Robertson (UCL), Michael J. Taylor (Microsoft), Hugo Zaragoza (Websays)
- Structured-document retrieval (field weighting)
- BM25 (retrieval, weighting function)
- Early works

- Empirical studies of field weighting (by Wilkinson)

- Information aggregation (by Lalmas)

- Combining document representations from different sources in a language model (by Ogilvie and Callan)

BM25

$$\bar{d} = (d_1, ..., d_V)$$

$$\mathbf{d} = (\bar{d}[1], \bar{d}[2], ..., \bar{d}[f], ...\bar{d}[K])$$

- Weighting function

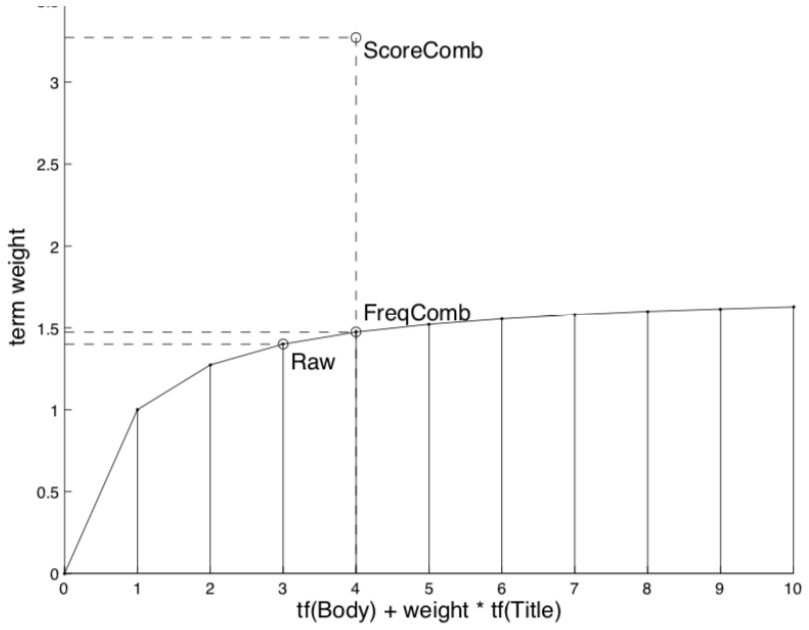
$$w_j(\bar{d}, C) := \frac{(k_1 + 1)d_j}{k_1((1 - b) + b\frac{dl}{avdl}) + d_j} log \frac{N - df_j + 0.5}{df_j + 0.5}$$

- Ranking function (scoring function)

$$W(\bar{d}, q, C) = \sum_j w_j(\bar{d}, C) \cdot q_j$$

Problems

- Linear combination of field scores is **bad** for retrieval functions like BM25
- Example: Plot of tf and term weight
- Tf scaling has to be controlled



Solutions

Okapi weights :

$$w_j(\bar{d}, C) := \frac{(k_1 + 1)d_j}{k_1((1-b) + b\frac{dl}{avdl}) + d_j} log \frac{N - df_j + 0.5}{df_j + 0.5}$$

- Unstructured-document scoring

$$W(\bar{d}, q, C) = \sum_j w_j(\bar{d}, C) \cdot q_j$$

- Structured-document scoring (each field is treated as an "unstructured document")

$$W(\bar{d}[f], q, C) = \sum_j w_j(\bar{d}[f], C) \cdot q_j$$

- Original approach: linear combination of field scores

$$W_1(\mathbf{d}, q, \mathbf{C}, \mathbf{v}) := \sum_{f=1}^{K} v_f \cdot W(\bar{d}[f], q, C)$$

- **New Approach**: **Linear combination of TF's (take into account all tfs)**

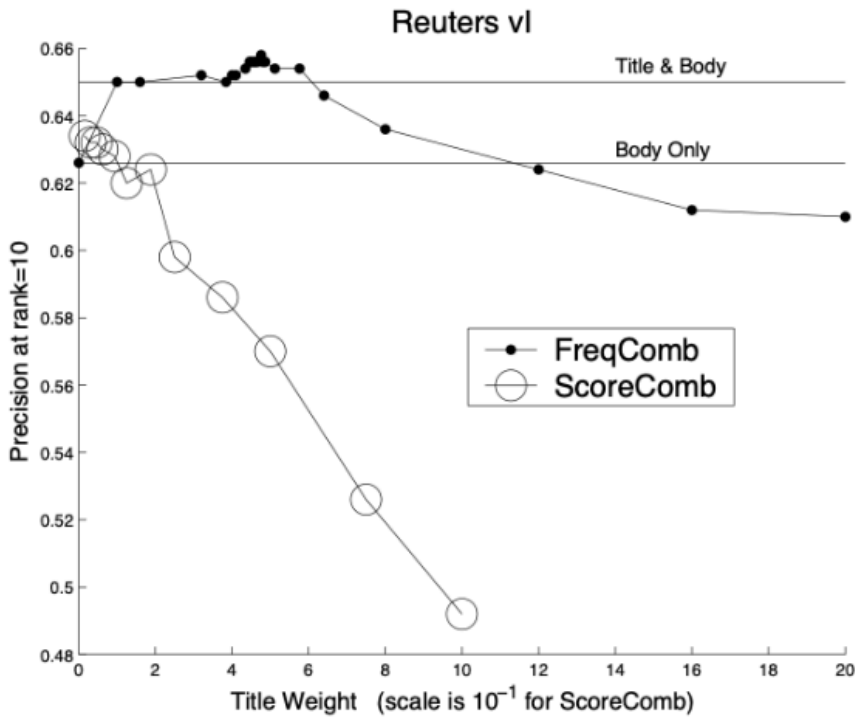
$$\mathbf{d}' := \sum_{f=1}^{K} v_f \cdot \bar{d}[f]$$

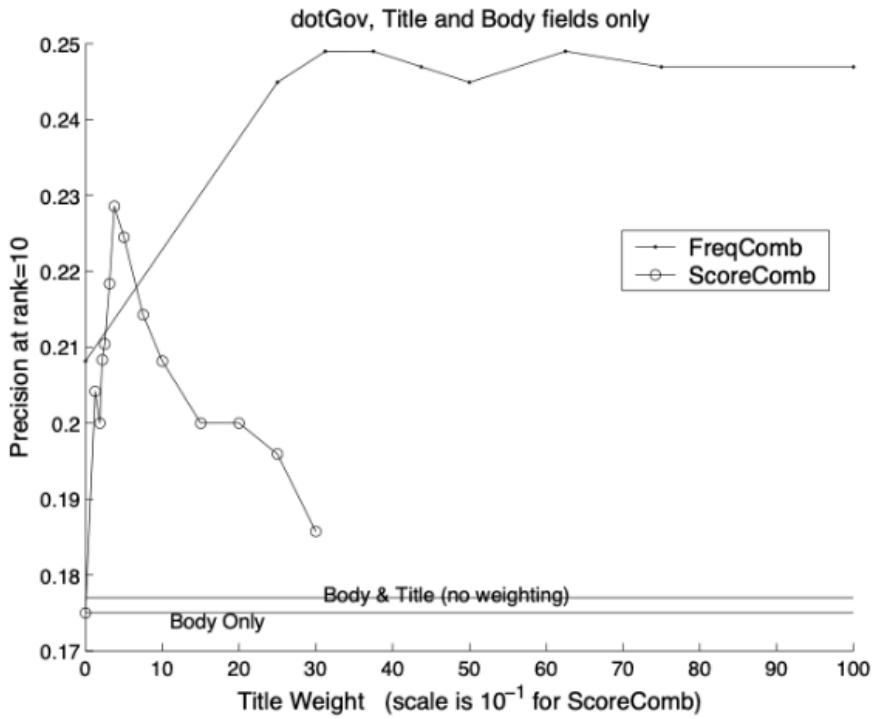
$$W_2(\mathbf{d}, q, \mathbf{C}, \mathbf{v}) := W(\mathbf{d}', q, \mathbf{C}')$$

- **Term weighting** and **scoring functions** are applied only **once** to each document

Experiment results

Figure 2: Title and Body Fields





Strengths and weaknesses

Strengths

- Resolve many of the issues
- Simplicity
- Potentially reduce the effort required to optimize the tuning parameters of a ranking function
- Intuitively understandable

Weaknesses

- Test collections were not ideal (limited fields in documents)
- Its applicability to anchor text was arguable

Another way to improve BM25

BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies

Article *in* Journal of Intelligent and Fuzzy Systems · May 2018

DOI: 10.3233/JIFS-169475

CITATIONS
2

READS
479

5 authors, including:

Sergio Jimenez
Caro y Cuervo Institute
39 PUBLICATIONS **260** CITATIONS
SEE PROFILE

Silviu Cucerzan
Microsoft
44 PUBLICATIONS **1,688** CITATIONS
SEE PROFILE

Fabio A. González
National University of Colombia
237 PUBLICATIONS **4,525** CITATIONS
SEE PROFILE

Alexander Gelbukh
Instituto Politécnico Nacional
513 PUBLICATIONS **4,935** CITATIONS
SEE PROFILE

BM25-CTF

- What are the differences between TF and CTF ?
- What is BM25-CTF ?
- What problems do BM25-CTF solve ?
- How to improve BM25 ?
- Results of comparing BM25 with BM25-CTF
- How do we think about BM25-CTF?

BM25

$$\sum_{w \in Q} idf(w) \times \frac{(k_1+1) \times tf(w,d)}{k_1 \times K(d) + tf(w,d)} \times \frac{(k_3+1) \times tf(w,q)}{k_3 + tf(w,q)}$$

BM25-CTF

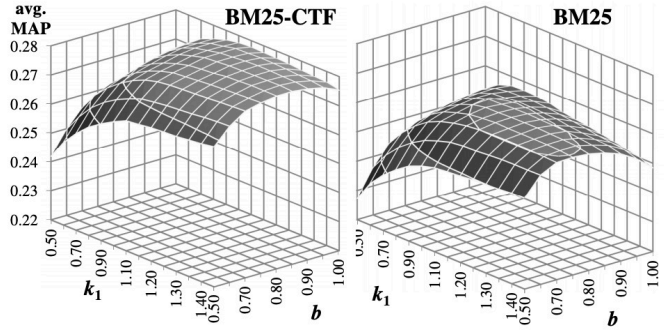
$$\sum_{w \in q} bidf(w) \times \frac{(k_1+1) \times btf(w,d)}{k_1 \times K(d) + btf(w,d)} \times \frac{(k_3+1) \times btf(w,q)}{k_3 + btf(w,q)}$$

$$bidf(w) = ictf(w) \times pidf(w) \times idf(w)$$

$$btf(w,d) = C(d) \times \frac{tf(w,d)}{\hat{tf}(w,d)}$$

$$ictf(w) = \log\left(\frac{M}{ctf(w)}\right), \qquad pidf(w) = \log\left(\frac{\hat{df}(w)}{\hat{ctf}(w)} + 1\right)$$

Compare BM25 with BM25 CTF



Compare MAP Matrix

There are 8 collections
Compared the optimal k1 and b throughout the tested collections between BM25 and BM25-CTF(table 5)

Table 5
Optimal parameters for BM25 and BM25-CTF using MAP across collections

model	param.	TREC-8	TREC-7	TREC-6	TREC-5	TREC-4	TREC-3	TREC-2	TREC-1	Average	Average
BM25	k_1	0.80	1.20	0.90	0.70	1.50	1.50	1.25	1.40	1.16(0.30)	1.02(0.33)
	b	0.95	0.75	0.85	1.00	0.60	0.75	0.70	0.75	0.79(0.12)	0.83(0.16)
BM25-CTF	k_1	1.30	1.10	0.90	1.25	1.50	1.50	1.50	1.50	1.32(0.21)	1.21(0.22)
	b	0.85	0.95	0.95	0.95	0.70	0.80	0.75	0.80	0.84(0.09)	0.88(0.11)

Compare BM25 with BM25 CTF

Table 7
Comparison for each collection using its default parameters

	MAP			P@10		
TREC	BM25	BM25-CTF	Improv.	BM25	BM25-CTF	Improv.
1	0.323	0.3305	2.32%	0.5158	0.5248	1.75%
2	0.3404	0.3401	-0.09%	0.4416	0.4576	3.62%
3	0.3287	0.3316	0.88%	0.68	0.674	-0.88%
4	0.2028	0.2204*	8.68%	0.582	0.582	0.00%
5	0.2141	0.2373*	10.84%	0.656	0.654	-0.30%
6	0.2282	0.2611*	14.42%	0.454	0.458	0.88%
7	0.2146	0.2427*	13.09%	0.414	0.416	0.48%
8	0.2332	0.2744*	**17.67%**	0.428	0.4520*	**5.61%**

[*] significantly better using Wilcoxon's test (p-value<0.05)

- Conclusion :
- In the 7 performance measures considered, BM25-CTF was significantly better 20 times
- The improvement of BM25-CTF over BM25 can be only observed when BM25-CTF is used in collections containing relatively large documents.

References

- [1] Robertson, Stephen & Zaragoza, Hugo & Taylor, Michael. (2004). Simple BM25 extension to multiple weighted fields. International Conference on Information and Knowledge Management, Proceedings.
- [2] Jimenez, Sergio & Cucerzan, Silviu & González, Fabio & Gelbukh, Alexander & Dueñas, George. (2018). BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies. Journal of Intelligent & Fuzzy Systems.