Composition of TF Normalizations: New Insights on Scoring Functions for Ad Hoc IR

François Rousseau[*], Michalis Vazirgiannis[*†‡]
[*]LIX, École Polytechnique, France
[†]Department of Informatics, AUEB, Greece
[‡]Institut Mines-Télécom, Télécom ParisTech, France
rousseau@lix.polytechnique.fr, mvazirg@aueb.gr

ABSTRACT

Previous papers in ad hoc IR reported that scoring functions should satisfy a set of heuristic retrieval constraints, providing a mathematical justification for the normalizations historically applied to the term frequency (TF). In this paper, we propose a further level of abstraction, claiming that the successive normalizations are carried out through composition. Thus we introduce a principled framework that fully explains BM25 as a variant of TF-IDF with an inverse order of function composition. Our experiments over standard datasets indicate that the respective orders of composition chosen in the original papers for both TF-IDF and BM25 are the most effective ones. Moreover, since the order is different between the two models, they also demonstrated that the order is instrumental in the design of weighting models. In fact, while considering more complex scoring functions such as BM25+, we discovered a novel weighting model in terms of order of composition that consistently outperforms all the rest. Our contribution here is twofold: we provide a unifying mathematical framework for IR and a novel scoring function discovered using this framework.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Theory, Algorithms, Experimentation

Keywords

IR theory; scoring functions; TF normalizations; heuristic retrieval constraints; function composition

1. MOTIVATION

Fang *et al.* introduced in [3] a set of heuristic retrieval constraints that any scoring function used for ad hoc information retrieval (IR) should satisfy. In particular, these constraints involve term frequency, term discrimination, document length and the interactions between them. For instance, they stated that a scoring function should favor document matching more distinct query terms. It is one of the earliest works that formally defined the properties that both the TF and the IDF components of any weighting model should possess. It is a unifying theory in IR that applies to the *vector space model* (TF-IDF [13]), *probabilistic* (BM25 [10]), *language modeling* (Dirichlet prior [15]) and *information-based* (SPL [2]) approaches and the *divergence from randomness* framework (PL2 [1]).

The definition of these constraints contributed to the improvement of the overall effectiveness of most modern scoring functions. Constraints on the term frequency result in successive normalizations on the raw TF, each one satisfying one or more properties. In our work, we intended to go one step further and we propose the use of *composition* to explain how the normalizations are applied successively in the general TF×IDF weighting scheme. In section 2, we describe in details the mathematical framework we designed. In section 3, we present the experiments we conducted over standard datasets and the results obtained that indicate how important the order of composition is, along with a novel and effective weighting model, namely $\text{TF}_{lo\delta op}$×IDF. Finally, in section 4, we conclude and mention future work.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

2. MATHEMATICAL FRAMEWORK

In ad hoc IR, a scoring function associates a score to a term appearing both in a query and a document. This function consists of three components supposedly independent of one another: one at the *query level* (QF), one at the *document level* (TF) and one at the *collection level* (IDF). These components are aggregated through multiplication to obtain a final score for the term denoted hereinafter by TF×IDF. We omit voluntarily to mention QF in the name since it is a function of the term frequency in the query and it has usually a smaller impact on the score, in particular for Web queries that tend to be short. We make here a difference between the TF×IDF general weighting scheme and TF-IDF, the *pivoted normalization weighting* defined in [13]. Note that because we rank documents, these *term scores* will be aggregated through sum to obtain a *document score* but this is beyond the scope of the current paper.

2.1 A set of TF normalizations

Since the early work of Luhn [4], *term frequency* (TF) has been claimed to play an important role in information

retrieval and is at the center of all the weighting models. Intuitively, the more times a document contains a term of the query, the more relevant this document is for the query. Hence, it is commonly accepted that the scoring function must be an increasing function of the term frequency and the simplest TF component can be defined as follows:

$$TF(t,d) = tf(t,d) \tag{1}$$

where $tf(t,d)$ is the term frequency of the term t in the document d. However, as the use of the raw term frequency proved to be non-optimal in ad hoc IR, the research community started normalizing it considering multiple criteria, mainly *concavity* and *document length normalization*. Later, these normalizations were explained as functions satisfying some heuristic retrieval constraints as aforementioned [3].

Concave normalization.

The marginal gain of seeing an additional occurrence of a term inside a document is not constant but rather decreasing. Indeed, the change in the score caused by increasing TF from 1 to 2 should be much larger than the one caused by increasing TF from 100 to 101. Mathematically, this corresponds to applying a concave function on the raw TF. We prefer the term *concave* like in [2] to *sublinear* like in [5] since the *positive homogeneity* property is rarely respected (and actually not welcomed) and the *subadditivity* one, even though desirable, not sufficient enough to ensure a decreasing marginal gain.

There are mainly two concave functions used in practice: the one in TF-IDF [13] and the one in BM25 [10] that we respectively called log-concavity (TF_l) and k-concavity (TF_k):

$$TF_l(t,d) = 1 + \ln[1 + \ln[tf(t,d)]] \tag{2}$$

$$TF_k(t,d) = \frac{(k_1+1) \times tf(t,d)}{k_1 + tf(t,d)} \tag{3}$$

where k_1 is a constant set by default to 1.2 corresponding to the asymptotical maximal gain achievable by multiple occurrences compared to a single occurrence.

Document length normalization.

When collections consist of documents of varying lengths (like web pages for the Web), longer documents will – as a result of containing more terms – have higher TF values without necessary containing more information. For instance, a document twice as long and containing twice as more times a term should not get a score twice as large but rather a very similar score. As a consequence, it is commonly accepted that the scoring function should be an inverse function of the document length to compensate that effect. Early works in *vector space model* suggested to normalize the score by the norm of the vector, be it the L^1 norm (document length), the L^2 norm (Euclidian length) or the L^∞ norm (maximum TF value in the document) [11]. These norms still mask some subtleties about longer documents – since they contain more terms, they tend to score higher anyway. Instead, the research community has been using a more complex normalization function known as *pivoted document length normalization* and defined as follows:

$$TF_p(t,d) = \frac{tf(t,d)}{1 - b + b \times \frac{|d|}{avdl}} \tag{4}$$

where $b \in [0,1]$ is the slope parameter, $|d|$ the document length and $avdl$ the average document length across the collection of documents as defined in [12].

2.2 Composition of TF normalizations

Based on this set of properties and their associated functions, it seems natural to apply them to the raw TF successively by composing them. In the literature, the document length normalization has usually been applied to either the overall *term score* or the *document score* like one would normally normalize a vector. However, it was then hard to fully fit BM25 in the TF×IDF weighting scheme. With composition, it is just a matter of ordering the functions. We present here the two compositions behind TF-IDF and BM25, respectively TF_{pol} ($= \text{TF}_p \circ \text{TF}_l$) and TF_{kop} ($= \text{TF}_k \circ \text{TF}_p$):

$$TF_{pol}(t,d) = \frac{1 + \ln[1 + \ln[tf(t,d)]]}{1 - b + b \times \frac{|d|}{avdl}} \tag{5}$$

$$\begin{aligned} TF_{kop}(t,d) &= \frac{(k_1+1) \times \frac{tf(t,d)}{1-b+b\times\frac{|d|}{avdl}}}{k_1 + \frac{tf(t,d)}{1-b+b\times\frac{|d|}{avdl}}} \\ &= \frac{(k_1+1) \times tf(t,d)}{k_1 \times [1 - b + b \times \frac{|d|}{avdl}] + tf(t,d)} \\ &= \frac{(k_1+1) \times tf(t,d)}{K + tf(t,d)} \end{aligned} \tag{6}$$

where $K = k_1 \times (1 - b + b \times \frac{|d|}{avdl})$ as defined in [10].

Note that under this form, it is not obvious that TF_{kop} is really a composition of two functions with the same properties as the one in TF_{pol}. We think this is the main reason why composition has never been considered before. By doing so, we provide not only a way to fully explain BM25 as a TF×IDF weighting scheme but also a way to easily consider variants of a weighting model by simply changing the order of composition. As we will see in section 3, this led us to a new TF×IDF weighting scheme that outperforms BM25.

2.3 Inverse Document Frequency

While the higher the frequency of a term in a document is, the more salient this term is supposed to be, this is no longer true at the collection level. This is actually quite the inverse since these terms have a presumably lower discrimination power. Hence, the use of a function of the *term specificity*, namely the *Inverse Document Frequency* (IDF) as defined in [14] and expressed as follows:

$$IDF(t) = \log \frac{N+1}{df(t)} \tag{7}$$

where N is the number of documents in the collection and $df(t)$ the document frequency of the term t.

2.4 TF-IDF versus BM25

By TF-IDF, we refer to the TF×IDF weighting model defined in [13], often called *pivoted normalization weighting*. The weighting model corresponds to TF_{pol}×IDF:

$$TF\text{-}IDF(t,d) = \frac{1 + \ln[1 + \ln[tf(t,d)]]}{1 - b + b \times \frac{|d|}{avdl}} \times \log \frac{N+1}{df(t)} \tag{8}$$

By BM25, we refer to the scoring function defined in [10], often called *Okapi weighting*. It corresponds to TF_{kop}×IDF when we omit QF (k-concavity of parameter k_3 for $tf(t,q)$). Thus, it has an inverse order of composition between the

concavity and the document length normalization compared to TF-IDF. The within-document scoring function of BM25 is written as follows when using the IDF formula defined in subsection 2.3 to avoid negative values, following [3]:

$$BM25(t,d) = \frac{(k_1+1)\times tf(t,d)}{K + tf(t,d)} \times \log\frac{N+1}{df(t)} \quad (9)$$

2.5 Lower-bounding TF normalization

Through composition, we also allow additional constraints for the TF component to be satisfied easily. *Subadditivity* is for instance a desirable property – if two documents have the same total occurrences of all query terms, a higher score should be given to the document covering more distinct query terms. Here, it just happens that TF_l and TF_k already satisfy it as noted in [3].

Recently, Lv and Zhai introduced in [6] two new constraints to the work of Fang *et al.* to lower-bound the TF component. In particular, there should be a sufficiently large gap in the score between the presence and absence of a query term even for very long documents where TF_p tends to 0 and a fortiori the overall score too. Mathematically, this corresponds to be composing with a third function TF_δ that is always composed after TF_p since it compensates the potential null limit introduced by TF_p and it is defined as follows:

$$TF_\delta(t,d) = \begin{cases} tf(t,d)+\delta & \text{if } tf(t,d) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where δ is the gap, set to 0.5 if TF_δ is composed immediately after TF_p and 1 if concavity is applied in-between. These are the two values defined in the original papers [6, 7] and we just interpreted their context of use in terms of order of composition. We did not change nor tune the values.

The weighting models Piv+ and BM25+ defined in [6] correspond respectively to $\text{TF}_{\delta\circ p\circ l}\times$IDF and $\text{TF}_{\delta\circ k\circ p}\times$IDF while BM25L defined in [7] to $\text{TF}_{k\circ\delta\circ p}\times$IDF. We clearly see that the only difference between BM25+ and BM25L is the order of composition: this is one of the advantages of our framework – easily represent and compute multiple variants of a same general weighting model. In the experiments, we considered all the possible orders of composition between TF_k or TF_l, TF_p and TF_δ with the condition that TF_p always precedes TF_δ as explained before.

For instance, we will consider a novel model $\text{TF}_{l\circ\delta\circ p}\times$IDF with $\text{TF}_{l\circ\delta\circ p}$ defined as follows:

$$TF_{l\circ\delta\circ p}(t,d) = 1 + \ln[1 + \ln[\frac{tf(t,d)}{1-b+b\times\frac{|d|}{avdl}} + \delta]] \quad (11)$$

where b is set to 0.20 and δ to 0.5.

3. EXPERIMENTS

Following our mathematical framework that relies on composition, we wondered why the order of composition was different between two widely used scoring functions – TF-IDF and BM25. In the original papers [11, 9], there was no mention of the difference in the order and this motivated us to investigate the matter. Our initial thought was that using an inverse order of composition in BM25 could improve it or vice-versa for TF-IDF. As a consequence, we tried exhaustively the combinations among TF_k, TF_l and TF_p and report the results. Thereafter, as mentioned in subsection 2.5, we followed the same procedure considering a third function to compose with: TF_δ. Indeed, we wanted to explore the extensions considered by the research community [6, 7] in terms of composition. This led us to a novel weighting model that outperforms them (see subsection 3.4).

3.1 Datasets and evaluation

We used two TREC collections to carry out our experiments: Disks 4&5 (minus the Congressional Record) and WT10G. Disks 4&5 contains 528,155 news releases while WT10G consists of 1,692,096 crawled pages from a snapshot of the Web in 1997. For each collection, we used a set of TREC topics (title only to mimic Web queries) and their associated relevance judgments: 301-450 and 601-700 for Disks 4&5 (TREC 2004 Robust Track) and 451-550 for WT10G (TREC9-10 Web Tracks).

We evaluated the scoring functions in terms of *Mean Average Precision* (MAP) and *Precision at 10* (P@10) considering only the top-ranked 1000 documents for each run. Our goal is to compare weighting models that use the same functions but with a different order of composition and select the best ones on both metrics. For example, in Table 1, $\text{TF}_{p\circ l}\times$IDF is compared with $\text{TF}_{l\circ p}\times$IDF and $\text{TF}_{k\circ p}\times$IDF with $\text{TF}_{p\circ k}\times$IDF. The statistical significance of improvement was assessed using the Student's paired t-test considering p-values less than 0.01 to reject the null hypothesis.

3.2 Platform and models

We have been using Terrier version 3.5 [8] to index, retrieve and evaluate over the TREC collections. For both datasets, the preprocessing steps involved Terrier's built-in stopword removal and Porter's stemming. We did not tune the slope parameter b of the pivoted document length normalization on each dataset. We set it to the default value suggested in the original papers: 0.20 when used with log-concavity [13] and 0.75 when used with k-concavity [10].

3.3 Results for TF-IDF versus BM25

We report in Table 1 the results we obtained on the aforementioned datasets when considering concavity and pivoted document length normalization. To the best of our knowledge, experiments regarding the same functions (TF_k, TF_l and TF_p) with a different order of composition have never been reported before. They indeed show that the original order chosen for both TF-IDF and BM25 is the most effective one: $\text{TF}_{p\circ l}\times$IDF outperforms $\text{TF}_{l\circ p}\times$IDF and $\text{TF}_{k\circ p}\times$IDF outperforms $\text{TF}_{p\circ k}\times$IDF. But since the order is different between the two, this also indicates that the order does matter depending on which function is chosen for each property.

For these two models (TF-IDF and BM25), the use of a different concave function to meet the exact same constraints requires the pivoted document length normalization to be applied before or after the function. The impact is even more significant on the Web dataset (WT10G) that corresponds the most to contemporary collections of documents.

3.4 Results for lower-bounding normalization

In Table 2, we considered in addition the lower-bounding normalization function TF_δ defined in subsection 2.5. The best-performing weighting model on both datasets is a novel one – $\text{TF}_{l\circ\delta\circ p}\times$IDF – and it even outperforms BM25+ and BM25L (significantly using the t-test and $p < 0.01$). This model has never been considered before in the literature to the best of our knowledge. In fact, the results from Table 1 establish that TF_l should apparently be applied before TF_p

Table 1: TF-IDF vs. BM25: an inverse order of composition; bold indicates significant performances

Weighting model	TREC 2004 Robust		TREC9-10 Web	
	MAP	P@10	MAP	P@10
IDF	0.1396	0.2040	0.0539	0.0729
TF	0.0480	0.0867	0.0376	0.0833
TF_p [b=0.20]	0.0596	0.1193	0.0531	0.1021
TF_p [b=0.75]	0.0640	0.1289	0.0473	0.1000
TF_l	0.1591	0.3141	0.1329	0.2063
TF_k	0.1768	0.3269	0.1522	0.2104
$TF_{p \circ k}$	0.0767	0.1932	0.0465	0.0604
$TF_{l \circ p}$	0.1645	0.3651	0.0622	0.1854
$TF_{p \circ l}$	0.1797	0.3647	0.1260	0.1875
$TF_{k \circ p}$	0.2045	0.3863	0.1702	0.2208
$TF_{p \circ k}$×IDF	0.1034	0.2293	0.0507	0.0833
$TF_{l \circ p}$×IDF	0.1939	0.3964	0.0750	0.2125
$TF_{p \circ l}$×IDF [TF-IDF]	**0.2132**	**0.4064**	**0.1430**	**0.2271**
$TF_{k \circ p}$×IDF [BM25]	**0.2368**	**0.4161**	**0.1870**	**0.2479**

like in TF-IDF. With lower-bounding normalization, it no longer holds. The formula for $TF_{l \circ \delta \circ p}$×IDF was given in equation 11. Without the use of our formal framework and composition, it would have been harder to detect and test these variants that can outperform state-of-the-art scoring functions when the order of composition is chosen carefully.

Table 2: $TF_{l \circ \delta \circ p}$×IDF vs. BM25+ and BM25L; bold indicates significant performances.

Weighting model	TREC 2004 Robust		TREC9-10 Web	
	MAP	P@10	MAP	P@10
$TF_{\delta \circ p \circ k}$	0.1056	0.2349	0.0556	0.0771
$TF_{\delta \circ l \circ p}$	0.1807	0.3751	0.0668	0.2021
$TF_{l \circ \delta \circ p}$	0.2130	0.4064	0.1907	0.2625
$TF_{\delta \circ p \circ l}$	0.2002	0.3876	0.1436	0.2021
$TF_{k \circ \delta \circ p}$	0.2155	0.3936	0.1806	0.2292
$TF_{\delta \circ k \circ p}$	0.2165	0.3956	0.1835	0.2354
$TF_{\delta \circ p \circ k}$×IDF	0.1466	0.2723	0.0715	0.1000
$TF_{\delta \circ l \circ p}$×IDF	0.2096	0.4048	0.0806	0.2292
$TF_{l \circ \delta \circ p}$×IDF	**0.2495**	**0.4305**	**0.2084**	**0.2771**
$TF_{\delta \circ p \circ l}$×IDF [Piv+]	0.2368	0.4157	0.1643	0.2438
$TF_{k \circ \delta \circ p}$×IDF [BM25L]	**0.2472**	**0.4217**	**0.2000**	**0.2563**
$TF_{\delta \circ k \circ p}$×IDF [BM25+]	**0.2466**	**0.4145**	**0.2026**	**0.2521**

4. CONCLUSIONS AND FUTURE WORK

Scoring function design is a cornerstone issue in information retrieval. In this short paper, we intended to provide new insights on scoring functions for ad hoc IR. In particular, we proposed a unifying mathematical framework that explains how weighting models articulate around a set of heuristic retrieval constraints introduced in related work.

Using composition to combine the successive normalizations historically applied to the term frequency, we were able to fully explain BM25 as a TF×IDF weighting scheme with just an inverse order of composition between the concavity and the document length normalization compared to TF-IDF. Besides, the framework also allowed us to discover and report a novel weighting model – $TF_{l \circ \delta \circ p}$×IDF – that consistently and significantly outperformed BM25 and its extensions on two standard datasets in MAP and P@10.

Future work might involve the design of novel retrieval constraints and their compositions with existing ones. We are confident that refining the mathematical properties behind scoring functions will continue to improve the effectiveness of these models in ad hoc IR.

5. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their useful feedbacks. This material is based upon work supported by the French DIGITEO Chair grant LEVETONE.

6. REFERENCES

[1] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, Oct. 2002.

[2] S. Clinchant and E. Gaussier. Information-based models for ad hoc IR. In *Proceedings of SIGIR'10*, pages 234–241, 2010.

[3] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of SIGIR'04*, pages 49–56, 2004.

[4] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, Oct. 1957.

[5] Y. Lv and C. Zhai. Adaptive term frequency normalization for BM25. In *Proceedings of CIKM'11*, pages 1985–1988, 2011.

[6] Y. Lv and C. Zhai. Lower-bounding term frequency normalization. In *Proceedings of CIKM'11*, pages 7–16, 2011.

[7] Y. Lv and C. Zhai. When documents are very long, BM25 fails! In *Proceedings of SIGIR'11*, pages 1103–1104, 2011.

[8] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of SIGIR'06*, 2006.

[9] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR'94*, pages 232–241, 1994.

[10] S. E. Robertson, S. Walker, K. Spärck Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC-3*, pages 109–126, 1994.

[11] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[12] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of SIGIR'96*, pages 21–29, 1996.

[13] A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira. AT&T at TREC-7. In *Proceedings of TREC-7*, pages 239–252, 1999.

[14] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20, 1972.

[15] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'01*, pages 334–342, 2001.