Study of Heuristic IR Constraints Under Function Discovery Framework

Parantapa Goswami Eric Gaussier Massih-Reza Amini

Université Grenoble Alps,
CNRS-LIG/AMA
Grenoble, France

firstname.lastname@imag.fr

ABSTRACT

In this paper we investigate the effect of the heuristic IR constraints on IR term-document scoring functions within the recently proposed function discovery framework. In the earlier study the constraints were empiricaly validated as a whole. Moreover, only the group of form constraints was utilized and the other prominent group, the adjustment constraints, was not considered. In this work we will investigate all the constraints individually and study them with two different term frequency normalization, namely normalization scheme used in DFR models and relative term count normalization used in language models.

Keywords

IR Theory, Function Discovery, Heuristic IR Constraints

1. INTRODUCTION

Fang et. al. [3] proposed a set of constraints which all "good" IR scoring functions should follow. These are divided in two categories, form and adjustment constraints (details are in Section 3). Among them Clinchant and Gaussier[1, 2] expressed the form conditions in analytical forms in terms of first and second order derivatives of the IR scoring function. They also studied these constraints under pseudo-relevavance feedback (PRF) framework and derived conditions that PRF models should satisfy. However, these studies did not consider the other important category - the adjustment constraints.

In the recently proposed function discovery approach [5] the form constraints are successfully used as a tool to prune the search space. It is ensured that the generated functions satisfy the form constraints. An experimental validation of these constraints is also provided in light of the proposed framework, which is inline with other empirical validations of these constraints [6, 4].

However, in the original study these constraints are considered together as a single module. In this paper we will investigate the effect of each individual form constraint on the scoring functions through function discovery framework. We will also investigate two adjustment constraints, which were not considered in the original work. We will do so by taking into account two different term frequency normalization techniques, namely the normalization used in divergence from randomness (DFR) models and relative term count normalization used in language models.

2. FUNCTION DISCOVERY FRAMEWORK

The general form of retrieval status value or RSV of a document d with respect to a query q can be formulated as:

$$\mathtt{RSV}(q,d) = \sum_{w \in q} a(t_w^q) \;\; g(w,d)$$

Where, t_w^q is the number of occurrences of term w within query q, $a : \mathbb{R}^+ \to \mathbb{R}^+$ is a positive real-valued function usually set to the identity function and the function $g(w,d)$ is called a scoring function which assigns a score to d for a term $w \in q$. Standard IR models like BM25, language models, information based models, DFR models etc. all fit in the form of the above equation; and depending on the model in use, the form of the scoring function varies. Table 1 summarizes notations used throughout the paper.

t_w^d	term frequency - # of occurrences of term w in document d
t_w^q	# of occurrences of term w in query q
x_w^d	normalized version of term frequency
$\mathcal{N}_w$	document frequency - # of documents in the collection containing w
y_w	normalized version of document frequency
$\mathcal{N}$	# of documents in a given collection
l_d	Length of document w in # of terms
l_{avg}	Average length of documents in a given collection

Table 1: Notations

The function discovery approach [5] deploys a context free grammar to generate closed form formulas to be used as scoring functions. Two variables are considered in that grammar, normalized term frequency denoted by x_w^d and normalized document frequency denoted by y_w. A real valued constant is also considered, but in experiments it is taken as 1 which we follow here as well. Thus scoring functions in this framework can also be written as $g(x_w^d, y_w)$.

Normalized term frequency can be expressed as a function of t_w^d and l_d, in the form $NTF(t_w^d, l_d)$. [5] considers normalization used in DFR models (DFR normalization) and in this study we also consider relative term count (RTC) normalization commonly used in language models. Thus one has:

$$NTF(t_w^d, l_d) = \begin{cases} t_w^d \log\left(1 + c\frac{l_{avg}}{l_d}\right) & \text{DFR normalization} \\ t_w^d \left(\frac{l_{avg}}{l_d}\right) & \text{RTC normalization} \end{cases}$$

Here c is a free parameter which is taken as 1, its default value, in this work.

The normalized document frequency of a term w considered in [5], as well as in this study, is the average document frequency of w with respect to the total number of documents in the collection, $y_w = \frac{\mathcal{N}_w}{\mathcal{N}}$.

3. HEURISTIC IR CONSTRAINTS

Fang et. al. [3] proposed a set of hypothetical constraints which lays a guideline of how a *good* IR scoring function should behave. The constraints are categorized into two groups, four *form constraints* and two *adjustment constraints*.

3.1 Form Constraints

Four form constraints define the general form of the scoring function g. These constraints are expressed in the following analytical forms [1]:

$$\frac{\partial g}{\partial t_w^d} > 0; \quad \frac{\partial^2 g}{\partial (t_w^d)^2} < 0; \quad \frac{\partial g}{\partial \mathcal{N}_w} < 0; \quad \frac{\partial g}{\partial l_d} < 0$$

Considering DFR term frequency normalization and $y = \frac{\mathcal{N}_w}{\mathcal{N}}$, [5] has shown that it is sufficient for a scoring function g to satisfy the following three conditions, denoted by `C1`, `C2` and `C3` respectively:

$$\underbrace{\frac{\partial g}{\partial x} > 0,}_{\text{C1}} \quad \underbrace{\frac{\partial^2 g}{\partial x^2} < 0,}_{\text{C2}} \quad \underbrace{\frac{\partial g}{\partial y} < 0}_{\text{C3}}$$

For RTC normalization it is also trivial to show that these three conditions are sufficient for any scoring function to satisfy all the form constraints.

During function generation, it is hence ensured that the generated scoring functions must satisfy `C1`, `C2` and `C3`. As these constraints are same for both DFR and RTC normalization, all the generated scoring functions will satisfy the constraints for any of the two normalization schemes.

3.2 Adjustment Constraints

Two *adjustment constraints* aim to adjust the function g satisfying the form constraints by regulating the interaction between term frequency t_w^d and document length l_d. These two constraints are:

`C4` Let q be a query. $\forall k > 1$, if d_1 and d_2 are two documents such that $l_{d_1} = k \times l_{d_2}$ and for all terms w, $t_w^{d_1} = k \times t_w^{d_2}$, then $\texttt{RSV}(q, d_1) \geq \texttt{RSV}(q, d_2)$.

`C5` Let $q = w$ be a single term query, for two documents d_1 and d_2 if $t_w^{d_1} > t_w^{d_2}$ and $l_{d_1} = l_{d_2} + (t_w^{d_1} - t_w^{d_2})$, then $\texttt{RSV}(q, d_1) \geq \texttt{RSV}(q, d_2)$.

Document length effect (fourth form constraint) penalizes longer documents, whereas the first adjustment constraint `C4` avoids over-penalizing long documents. The second adjustment constraint `C5` ensures that a longer document must not be penalized over a shorter document if the excess length is due to the occurrences of the query term.

We present here two properties which will help to study the effect of adjustment constraints over the function discovery framework.

Property 1. *If a function generated by the function discovery approach using RTC and DFR normalization satisfies* `C1`*,* `C2` *and* `C3`*, then the function also satisfies* `C4`.

Property 2. *If a function generated by the function discovery approach using RTC normalization satisfies* `C1`*,* `C2` *and* `C3`*, then the function satisfies* `C5`*. If the function is generated using DFR normalization and satisfies* `C1`*,* `C2` *and* `C3`*, then it satisfies* `C5` *when* $t_w^{d_2} \leq \frac{p.f_1(p)}{f_1(0) - f_1(p)}$ *where* $f_1(u) = \log\left(\frac{l_{d_2}+u+\beta}{l_{d_2}+u}\right)$ *and* $\beta = c.l_{avg}$*, where* $t_w^{d_2}$*,* l_{d_2} *are as explained in the definition of* `C5`.

We now proceed to prove these properties. We do so by first proving the two following lemmas.

Lemma 1. *For a term w if there are two documents d_1 and d_2 such that for any $k > 0$, their normalized term frequencies are $x_w^{d_1} = NTF(k \times t_w^{d_2}, k \times l_{d_2})$ and $x_w^{d_2} = NTF(t_w^{d_2}, l_{d_2})$ respectively, then $x_w^{d_1} \geq x_w^{d_2}$.*

Proof. Assuming RTC normalization, one has $x_w^{d_1} = x_w^{d_2}$, thus proving the property.
For DFR normalization it can be shown that:

$$x_w^{d_1} - x_w^{d_2} = t_w^{d_2} \log\left(\frac{(k+\alpha)^k}{k(1+\alpha)}\right) \quad \text{assuming } \alpha = c\frac{l_{avg}}{l_{d_2}}$$

Applying binomial expansion:

$$(k+\alpha)^k - k(1+\alpha) = (k^k - k) + (k^k - k)\alpha + \ldots + \alpha > 0$$

This is because the term $(k^k - k) > 0$ as $k > 0$, and all the remaining terms of the expression are positive. Thus we have $\left(\frac{(k+\alpha)^k}{k(1+\alpha)}\right) > 1$ giving that $x_w^{d_1} - x_w^{d_2} \geq 0$ as $t_w^{d_2} \geq 0$, which proves the property for DFR normalization. $\square$

Lemma 2. *For a term w if there are two documents d_1 and d_2 such that for any integer $p > 1$, their normalized term frequencies are $x_w^{d_1} = NTF(t_w^{d_2} + p, l_{d_2} + p)$ and $x_w^{d_2} = NTF(t_w^{d_2}, l_{d_2})$ respectively, then:*

- *for RTC normalization $x_w^{d_1} \geq x_w^{d_2}$,*
- *for DFR normalization $x_w^{d_1} \geq x_w^{d_2}$ when $t_w^{d_2} \leq \frac{p.f_1(p)}{f_1(0)-f_1(p)}$ where $f_1(u) = \log\left(\frac{l_{d_2}+u+\beta}{l_{d_2}+u}\right)$ and $\beta = c.l_{avg}$.*

Proof. For RTC normalization $x_w^{d_1} - x_w^{d_2} = \frac{p(l_{d_2} - t_w^{d_2})}{l_{d_2}(l_{d_2}+p)} \geq 0$ since $l_{d_2} \geq t_w^{d_2}$, which proves the property.
For DFR normalization it can be derived that:

$$\begin{aligned} x_w^{d_1} - x_w^{d_2} &= (t_w^{d_2} + p)\log\left(\frac{l_{d_2}+p+\beta}{l_{d_2}+p}\right) - t_w^{d_2}\log\left(\frac{l_{d_2}+\beta}{l_{d_2}}\right) \\ &\qquad (\text{assuming } \beta = c.l_{avg}) \\ &= (t_w^{d_2}+p)f_1(p) - t_w^{d_2} f_1(0) \end{aligned}$$

Let $f_2(t_w^{d_2}) = (t_w^{d_2}+p)f_1(p) - t_w^{d_2} f_1(0)$, then $f_2(t_w^{d_2})$ is a strictly decreasing function with $t_w^{d_2}$ as $f_2'(t_w^{d_2}) < 0$. We have $f_2(0) = p.f_1(p) > 0$, but $f_2(t_w^{d_2}) \to -\infty$ as $t_w^{d_2} \to +\infty$. Thus $f_2(t_w^{d_2})$ crosses zero at $t_w^{d_2} = \frac{p.f_1(p)}{f_1(0)-f_1(p)}$. So $x_w^{d_1} - x_w^{d_2} \geq 0$ when $t_w^{d_2} \leq \frac{p.f_1(p)}{f_1(0)-f_1(p)}$, thus proving the property for DFR normalization. $\square$

Since queries are considered as set of terms and the order is not considered, $\texttt{RSV}(q, d_1) \geq \texttt{RSV}(q, d_2)$ is equivalent to $g(w, d_1) \geq g(w, d_2)$ (here we used the original form of the scoring functions as in Eq. 2). As g is satisfying C1, i. e. $\frac{\partial g}{\partial x_w^d} > 0$, one has $g(x_w^{d_1}, y) \geq g(x_w^{d_2}, y)$ iff $x_w^{d_1} \geq x_w^{d_2}$. Hence the adjustment constraint C4 boils down to the Lemma 1, which is true, as shown above, for all the scoring functions generated using the function discovery approach with both DFR and RTC normalization. Thus all generated scoring functions are satisfying C4 proving Property 1.

Suppose $p > 0$ is an integer constant such that $t_w^{d_1} = t_w^{d_2} + p$. Then this constraint can be rewritten as, if $l_{d_1} = l_{d_2} + p$ then $\texttt{RSV}(q, d_1) > \texttt{RSV}(q, d_2)$. Again as g is satisfying C1, one has $g(x_w^{d_1}, y) \geq g(x_w^{d_2}, y)$ iff $x_w^{d_1} \geq x_w^{d_2}$. Thus the adjustment constraint C5 becomes Lemma 2 and is always satisfied by the generated functions if RTC normalization is used. But for DFR normalization C5 is satisfied only when $t_w^{d_2} \leq \frac{p.f_1(p)}{f_1(0) - f_1(p)}$ where $f_1(u) = \log\left(\frac{l_{d_2}+u+\beta}{l_{d_2}+u}\right)$ and $\beta = c.l_{avg}$. This proves Property 2. So for DFR normalization a generated function satisfies C5 for not so high t_w^d values which is the case in most practical scenarios.

4. EXPERIMENTAL EVALUATION

Here we examine the effect of each constraint separately. Experiments are performed on six IR collections (Table 2), five from TREC (trec.nist.gov) and one from CLEF (www.clef-campaign.org) campaigns. These collections are indexed using Terrier IR Platform v3.5 (terrier.org). Preprocessing steps in creating an index include stemming using Porter stemmer and removing stop-words using the stopword list provided by Terrier. Generated functions are also implemented in Terrier. We specify by $\mathcal{C}_V$, respectively by

Collection	$\mathcal{N}$	l_{avg}	Index size	#queries
TREC-3	741,856	261	427.7 MB	50
TREC-5	524,929	339	378.0 MB	50
TREC-6,7,8	528,155	296	373.0 MB	50
CLEF-3	169,477	301	126.2 MB	60

Table 2: Statistics of various collections used in our experiments, sorted by size.

$\mathcal{C}_N$, the set of functions which satisfy all the constraints, respectively none of the constraints of a given length, and by $\mathcal{C}_N^i$ the set of functions which only satisfy constraint Ci. Performances of $\mathcal{C}_V$ and $\mathcal{C}_N$ are compared to empirically justify the usefulness of the heuristic IR constraints as a whole.

An initial intuition can be made by the sizes of the sets $\mathcal{C}_N^1$, $\mathcal{C}_N^2$ and $\mathcal{C}_N^3$. Figure 1 shows the number of functions in each of the sets till length 8. Clearly the number of functions satisfying C2 is the minimum, whereas the number of functions satisfying C1 is the maximum. Thus the constraint C2 is the harshest one, whereas C1 is the loosest one. Another trivial yet interesting observation is that $\mathcal{C}_N$ is the biggest set and $\mathcal{C}_V$ is the smallest one among all the five sets.

From each of the sets $\mathcal{C}_V$, $\mathcal{C}_N^1$, $\mathcal{C}_N^2$, $\mathcal{C}_N^3$ and $\mathcal{C}_N$, 10 subsets are created. Each subset contains 100 randomly selected sample functions chosen from the initial set. When creating a subset, 100 functions are selected without replacement. When creating another different subset, again all functions are considered for selection. Thus a function may be repeated in different subsets but never within the same subset. These samples are tested on CLEF-3 and TREC-3,5,6,7,8. For each function MAP is noted and it is averaged over all 100 functions within a single sample set. Finally, average

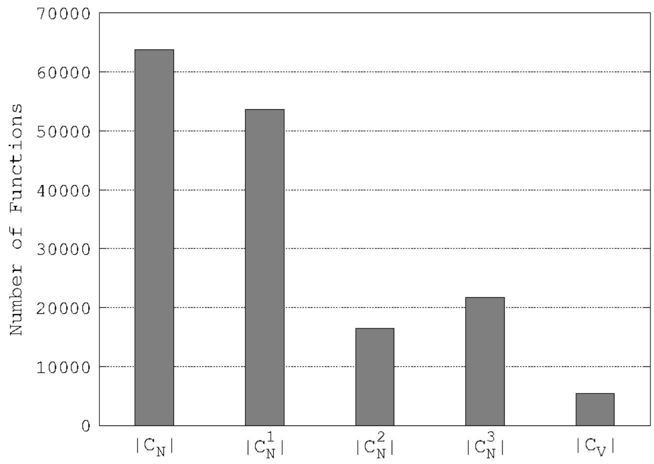
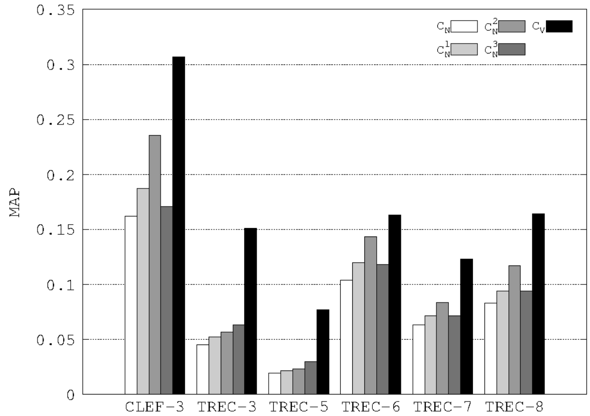
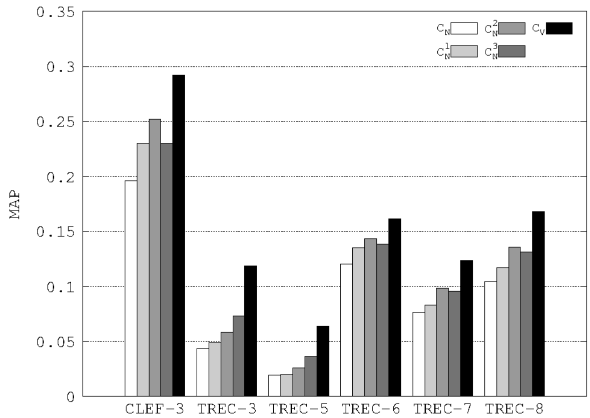


Figure 1: Number of functions in the sets $\mathcal{C}_N$, $\mathcal{C}_N^1$, $\mathcal{C}_N^2$, $\mathcal{C}_N^3$ and $\mathcal{C}_V$ till length 8.



(a) DFR



(b) RTC

Figure 2: Average MAP of the sets $\mathcal{C}_N$ (□), $\mathcal{C}_N^1$ (■), $\mathcal{C}_N^2$ (■), $\mathcal{C}_N^3$ (■) and $\mathcal{C}_V$ (■) till length 8 with (a) DFR and (b) RTC normalization.

performance over 10 sample sets is reported.

Figure 2 shows a plot of average MAP of 10 sample sets from all five sets $\mathcal{C}_N$, $\mathcal{C}_N^1$, $\mathcal{C}_N^2$, $\mathcal{C}_N^3$ and $\mathcal{C}_V$ with DFR and RTC normalization. As expected $\mathcal{C}_V$ is always best and $\mathcal{C}_N$ is always worst among the five sets. Performance of other three sets $\mathcal{C}_N^1$, $\mathcal{C}_N^2$ and $\mathcal{C}_N^3$ are in between $\mathcal{C}_V$ and $\mathcal{C}_N$. Both for DFR and RTC, C2 is best performing on 4 out of 6 collections. But for TREC-3 and TREC-5 C3 is slightly better than C2. There is no deterministic comparative pattern between C1 and C3. All possible relative orders in terms of performance between C1 and C3 are visible. As for example in case of DFR (Figure 2(a)) C1>C3 on CLEF-3, C1<C3 on TREC-3,5 and C1≈C3 on TREC-6,7,8. Though for RTC C1<C3 for 4 out of 6 collections (Figure 2(b)). In summary the general trend is that C2 is the most effective among three constraints although the plots display an inconclusive pattern. Thus it can be said that the relative effectiveness of the constraints is highly dependent on the collection in hand.

Above experiments are performed to study the effects of each constraint. But these experiments also revealed that combination of all the constraints (i.e. set $\mathcal{C}_V$) always performs best. Hence for all practical purposes it is always better to utilize all the constraints together.

5. CONCLUSION

In this paper we showed that the first adjustment constraint is satisfied by all the functions generated using the approach proposed in [5] with both DFR and RTC normalization. However, the second adjustment constraint is always satisfied by all the generated functions for RTC normalization, but it is satisfied only for not so high t_w^d values for DFR normalization. We experimentally studied the effects of each form constraint separately and found that C2 is the harshest among the three as it allows minimum number of functions. According to performances, for both DFR and RTC normalization, on most collections C2 is more effective than C1 and C3 and there is no deterministic pattern between C1 and C3.

Here we have studied the constraints for DFR and RTC normalization, as three constraints C1, C2 and C3 takes the same form with these two normalization schemes. For Okapi, the other popular normalization scheme, the forms of these constraints changes thus generating entirely different sets of valid functions.

6. REFERENCES

[1] S. Clinchant and E. Gaussier. Information-based models for ad hoc ir. In *Proceedings of the* 33^{rd} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 234–241. ACM, 2010.

[2] S. Clinchant and E. Gaussier. Retrieval constraints and word frequency distributions a log-logistic model for ir. *Information Retrieval*, 14(1):5–25, 2011.

[3] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of the* 27^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56. ACM, 2004.

[4] H. Fang and T. T. C. Zhai. Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems (TOIS)*, 29(2):7:1–7:42, 2011.

[5] P. Goswami, S. Moura, E. Gaussier, M.-R. Amini, and F. Maes. Exploring the space of ir functions. In *Proceedings of the* 36^{th} *European Conference on Information Retrieval (ECIR)*, pages 372–384. Springer, 2014.

[6] W. Zheng and H. Fang. Axiomatic approaches to information retrieval – university of delaware at trec 2009 million query and web tracks. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009*. National Institute of Standards and Technology (NIST), 2009.