Do IR models satisfy the TDC Retrieval Constraint?

Stéphane Clinchant, Éric Gaussier

Do IR models satisfy the TDC Retrieval Constraint?

Stéphane Clinchant
Xerox Research Center Europe & Université Grenoble I, LIG
6, Chemin de Maupertuis
38240 Meylan, France
stephane.clinchant@xrce.xerox.com

Eric Gaussier
Université Grenoble I, LIG
BP 53 - 38041 Grenoble cedex 9
Grenoble, France
eric.gaussier@imag.fr

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Theory

Keywords

axiomatic constraint, TDC constraint

1. INTRODUCTION

Axiomatic methods were pioneered by Fang *et al.* [5] and used since then in several studies including [3, 2]. In a nutshell, axiomatic methods provide formal constraints that IR functions should satisfy in order to be valid, i.e. to perfom well on IR tasks. According to [2], the four main constraints for an IR function to be valid can be phrased as: the weighting function should (a) be increasing and (b) concave wrt term frequencies, (c) have an IDF effect and (d) penalize long documents. In addition to these four basic constraints, Fang *et al.* [5] introduced additional constraints to regulate the relative importance of different parameters, as TF and IDF for example.

The IDF effect mentioned above relates to the constraint referred to in [5] as the *TDC* constraint, which can be formulated as follows:

TDC: Let q be a query and $w1$, $w2$ be two query terms. Assume $l_{d1} = l_{d2}$, $c(w1, d1) + c(w2, d1) = c(w1, d2) + c(w2, d2)$. If *idf(w1)* $\geq$ *idf(w2)* and $c(w1, d1) \geq c(w1, d2)$, then $RSV(d1, q) \geq RSV(d2, q)$.

where $c(w, d)$ denotes the number of occurrences of w in d. This constraint aims at capturing the fact that, *ceteris paribus*, rarer terms (i.e. terms with a large IDF) should be preferred over more frequent ones. However, there are several ways to define the context (*ceteris paribus*) in which to place this constraint, and the study presented in [2] relies on a stricter context corresponding to a special case of the *TDC* constraint, where $w1$ only occurs in $d1$ and $w2$ only in $d2$. This constraint, referred to as *speTDC* can be formulated as:

speTDC: Let q be a query and $w1$, $w2$ two query terms. Assume $l_{d1} = l_{d2}$, $c(w1, d1) = c(w2, d2)$, $c(w2, d1) = c(w1, d2) = 0$. If *idf(w1)* $\geq$ *idf(w2)*, then $RSV(d1, q) \geq RSV(d2, q)$.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

If it has been show in previous studies (as [5, 2]) that most IR models satisfy most IR constraints, the situation of the *TDC* constraint is unclear, and the goal of this short paper is to show that several state-of-the-art IR models indeed do not comply with the general *TDC* constraint, but do satisfy the *speTDC* one. We will review here the recently introduced log-logistic model [2], as well as the Jelinek-Mercer and Dirichlet language models.

2. IR MODELS AND THE TDC CONSTRAINT

The **log-logistic model** proposed in [2] is specified by:

$$\begin{aligned} t(w,d) &= c(w,d)\log(1 + c\frac{avg(l_d)}{l_d}) \\ r_w &= \frac{N_w}{N} \\ RSV(q,d) &= \sum_{w \in q \cap d} c(w,q)(\log(r_w + t_w^d) - \log(r_w)) \end{aligned}$$

where N_w is the number of documents in the collection containing the term w and N the total number of documents in the collection; l_d is the length of document d, and $\text{avg}(l_d)$ the average document length in the collection.

Let us examine the *TDC* constraint for this model, and for that let us consider two documents d_1 and d_2 of equal length l; let $\gamma = \log(1 + c\frac{avg(l)}{l})$. For simplification, we use a to denote w_1, b to denote w_b and a_1 (resp. a_2) for $c(a, d_1)$ (resp. $c(a, d_2)$). For a query q consisting of only a and b, the difference in score between d_1 and d_2 amounts to:

$$\Delta = RSV(q, d1) - RSV(q, d2) = \log(\frac{r_a + a_1\gamma}{r_a + a_2\gamma} \times \frac{r_b + b_1\gamma}{r_b + b_2\gamma})$$

Now, let us place ourselves in the conditions specified in the *TDC* constraint and let us assume that $r_a < r_b$, $a^1 > a^2$ and $a_1 + b_1 = a_2 + b_2$ (and thus $b_2 > b_1$). The TDC constraints stipulates in that case that $\Delta \geq 0$, that is:

$$\gamma(a_1b_1 - a_2b_2) + r_b(a_1 - a_2) + r_a(b_1 - b_2)) \quad > \quad 0$$

Setting: $a_1 = 7, b_1 = 4, a_2 = 6, b_2 = 5, r_a = 0.001$ and $r_b = 0.01$ shows that the above inequality is true *iff*: $\gamma < 0.0045$. Hence, γ must be very small for the *TDC* constraint to be verified. Indeed, for documents of average length, $\gamma \approx \log(1 + c)$ and c should be chosen smaller to 0.005 for the above inequality to be satisfied.

We now provide a more formal proof that the log-logistic model does not comply with the *TDC* constraint. Let's first

Table 1: Pair of query terms (short query) below mean corpus language model

Collection	m	μ	$diff < m$
robust	0.0003	500	62.2 %
trec1-2	0.0005	1000	62.2 %

consider the following optimization problem:

$$\text{argmax}_{t_a \geq 0, t_b \geq 0} \quad \mathcal{A} = \sum_{w \in \{a,b\}} \log(r_w + t_w) - \log(r_w)$$
$$\text{subject to} \quad \sum_{w \in \{a,b\}} t_w = s$$

where s is a pre-defined, positive value. As the log is concave, the overall objective funciton is concave, and the solution to the above optimization problem correspond to the values maximizing the following Lagrangian:

$$\Lambda = \sum_{w \in \{a,b\}} \log(r_w + t_w) - \log(r_w) - \lambda(\sum_{w \in \{a,b\}} t_w - s)$$

for which the partial derivatives are defied as:

$$\frac{\partial \Lambda}{\partial t_w} = \frac{1}{r_w + t_w} - \lambda$$

Setting these derivatives to 0 leads to the following solution[1]:

$$t_a = \frac{s + r_b - r_a}{2}, \; t_b = \frac{s + r_a - r_b}{2}$$

Now let us consider a query q with two words (a and b) occurring only once, and let d_1 a,d d_2 be two documents of equal length. Let us furthermore assume that: $\text{idf}(a) = \frac{1}{r_a} \geq \text{idf}(b) = \frac{1}{r_b}$, and:

$$t_a^{d_1} = \frac{s + r_b - r_a}{2} + \epsilon, \quad t_b^{d_1} = \frac{s + r_a - r_b}{2} - \epsilon$$
$$t_a^{d_2} = \frac{s + r_b - r_a}{2}, \qquad t_b^{d_2} = \frac{s + r_a - r_b}{2}$$

for ϵ sufficiently small for all the quantities to be positive. In this case, all the conditions of the *TDC* constraint are verified, and thus one should observe that $RSV(q, d_1) \geq RSV(q, d_2)$, which is in contradiction with the fact that the values for d_2 are the ones that maximize $\mathcal{A}$ which corresponds in this case to the retrieval status value. This shows that the log-logistic model is not compliant with the *TDC* constraint. However, as shown in [2], the log-logistic model is compliant with the *speTDC* constraint, which represents a stricter version of the *TDC* constraint.

The situation for language models wrt the *TDC* and *speTDC* constraints is identical to the one of the log-logistic model. Indeed, it has been shown in [1] that the **Jelinek-Mercer model** could be seen as a special case of the log-logistic model. All the development made above in the context of the log-logistic model applies to the Jelinek-Mercer model, which is not compliant with the *TDC* constraint (it is however compliant with the *speTDC* constraint).

As shown in [5], and using the notations introduced previously, the **Dirichlet language model** agrees with the *TDC* constraint in the following case:

$$\mu \geq \frac{a_1 - b_2}{p(b|C) - p(a|C)} \tag{1}$$

[1]As $r_a \ll t_a$ and $r_b \ll t_b$, both t_a and t_b are ≥ 0.

where $p(a|C)$ represents the collection probability. Table 1 shows for several collections the mean value of $p(w|C)$ for query terms (denoted m), the optimal values obtained for the Dirichlet smoothing parameter μ and the percentage of pairs of query terms for which the corpus language model absolute difference ($|p(w'|C) - p(w|C)|$) is below m (denoted $diff < m$). As one can note, in almost two third of the cases, the numerator of equation 1 is very small. So, for the bound given in equation 1 to hold, one needs to rely on large values for μ (larger than 2,000 when the numerator is one). As shown in table 1, we are far from these values in practice, and the Dirichlet language model is in general not compliant with the *TDC* constraint. Furthermore, using the analytical formulation of the *speTDC* constraint proposed in [2], one can show that the Dirichlet language model is compliant with the *speTDC* constraint.

3. CONCLUSION

We have shown here that several state-of-the-art IR models do not satisfy the *TDC* retrieval constraint introduced in [5]. The IR models we have considered are the recently introduced log-logistic model, and two standard versions of the language model, namely the one based on Jelinek-Mercer smoothing, and the one based on Dirichlet smoothing. Furthermore, we have seen that all these models satisfy *speTDC*, a stricter version of the *TDC* constraint introduced in [2] to directly formalize the IDF effect. Because of the good behavior of the models we have reviewed, we believe that the above development suggests that the *TDC* constraint is not valid, and should be replaced with the *speTDC* one.

Directly assessing the validity of a particular retrieval constraint is not straightforward. The work presented in [4] shows that it is possible to experimentally assess whether a particular IR model complies or not with a given constraint. It is however not clear whether all constraints can be taken into account. We have followed here a different line, based on a theoretical analysis of the behavior orf IR models wrt a particular constraint.

4. REFERENCES

[1] S. Clinchant and E. Gaussier. Bridging language modeling and divergence from randomness models: A log-logistic model for ir. In *ICTIR*, pages 54–65, 2009.

[2] S. Clinchant and E. Gaussier. Information-based models for *ad hoc* IR. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 234–241, New York, NY, USA, 2010. ACM.

[3] R. Cummins and C. O'Riordan. An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions. *Artif. Intell. Rev.*, 28:51–68, June 2007.

[4] R. Cummins and C. O'Riordan. Measuring constraint violations in information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 722–723, 2009.

[5] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004.