Measuring Constraint Violations in Information Retrieval

Ronan Cummins
Digital Enterprise Research Institute
National University of Ireland
Galway, Ireland
ronan.cummins@deri.org

Colm O'Riordan
Dept. of Information Technology
National University of Ireland
Galway, Ireland
colmor@it.nuigalway.ie

ABSTRACT

Recently, an inductive approach to modelling term-weighting function correctness has provided a number of axioms (constraints), to which all *good* term-weighting functions should adhere. These constraints have been shown to be theoretically and empirically sound in a number of works [2, 3, 1]. It has been shown that when a term-weighting function breaks one or more of the constraints, it typically indicates sub-optimality of that function. This elegant inductive approach may more accurately model the human process of determining the relevance a document. It is intuitive that a person's notion of relevance changes as terms that are either on or off-topic are encountered in a given document. Ultimately, it would be desirable to be able to mathematically determine the performance of term-weighting functions without the need for test collections.

Many modern term-weighting functions do not satisfy the constraints in an unconditional manner [3]. However, the degree to which these functions violate the constraints has not been investigated. A comparison between weighting functions from this perspective may shed light on the poor performance of certain functions in certain settings. Moreover, if a correlation exists between performance and the number of violations, measuring the degree of violation could help more accurately predict how a certain scheme will perform on a given collection.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Experimentation, Measurement, Performance

Keywords: Information Retrieval, Constraints, Axioms

1. MEASURING CONSTRAINT VIOLATIONS

Table 1: Characteristics of Collections

Collection	FT	FBIS	FR	OHSUMED
No. of Documents	210,158	130,471	55,630	293,856
Average Doc. Length	389	501	670	158
Standard Dev.	240	580	1380	60

Four constraints (axioms) have been postulated in order to capture the basic principles of term-weighting function correctness. These are detailed in [3] and [1]. The first constraint (C1) states that adding a new query term to a document must *always* increase the score of that document. The second constraint (C2) states that adding a non-query term to a document must *always* decrease the score of that document. The third constraint (C3) states that adding successive query terms to a document should increase the score of the document less with each successive addition. The fourth constraint (constraint 4) states that adding more non-query terms to a document should decrease the score of a document less with each occurrence [1]. These constraints (while relatively intuitive) can constrain the term-weighting function in complex ways. For example, constraint 1 cannot be guaranteed to be satisfied by modern term-weighting functions as the decrease in score due to the document increasing in length (i.e. by normalisation) cannot be guaranteed to be offset by the increase in score due to the query-term being added [1].

1.1 Approach Adopted

The approach used to measure the number of constraint violations in this work takes a stemmed query and document. The terms in the document remain in the same order in which they naturally appear. A pseudo-document is created by using the first term appearing in the document. This pseudo-document is then matched against the query using a term-weighting functions and the score is recorded. A further pseudo-document is created by including the next term appearing in the document. This is then matched against the query and the score is again recorded. This continues until the complete document is scored against the query. The violations of each constraint is measured as new terms are added to the pseudo-document.

When the score of a document does not increase when a query term is added to the pseudo-document a violation of constraint 1 (C1) is recorded. When the score of a document does not decrease when a non-query term is added a violation of constraint 2 (C2) is recorded. If the increase in score of the document when a query term is added is equal to or greater than the increase in score when the previous occurrence of that query term was added, a violation of constraint 3 (C3) is recorded. Finally, when three non-query terms appear in succession and the inverse of the score reduction is not sub-linear a violation of constraint 4 (C4) is recorded.

Due to the computational complexity of such an approach, it is infeasible to do this for an entire test collection. Therefore, we measure the number of violations of constraints on

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Table 2: Average no. of constraint violations averaged across collections for different length queries

	short				medium				long			
Functions	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
ES	0.2	0.0	0.3	0.0	27.5	0.0	23.3	0.0	114.9	0.0	85.75	0.0
DRF	2.6	0.0	2.2	0.0	150.3	0.0	125.9	0.0	359.6	0.0	291.6	0.0
$MBM25$	2.5	0.0	2.0	209.4	132.5	0.0	112.4	18.56	354.4	0.0	281.9	4.6
PIV	2.3	0.0	1.4	347.9	129.8	0.0	99.9	422.4	350.8	0.0	251.5	394.3
$BM25$	4.9	61.4	4.6	237.2	192.0	244.3	141.6	192.0	407.5	384.2	368.7	96.5

the top 1000 documents returned from a benchmark term-weighting functions. The top 1000 documents should represent a set of documents with a high number of query terms and therefore is a good sample of documents on which to measure the number of constraint violations.

2. EXPERIMENTS

2.1 Experimental Setup

We use the FBIS, FT, FR collections from TREC disks 4 and 5 as test collections. For topics 251 to 450 we create a short query set (title field only), a medium length query set (title and description), and a long query set (title, description and narrative). We also use the OHSUMED collection and its topics. Table 1 shows some of the characteristics of the collections used in this research. As per the original axiomatic study [3], we performed stemming but did *not* remove stopwords. A term-weighting function which correctly models relevance should be able to correctly weight all terms.

We use five term-weighting functions in these experiments. We use the default $BM25$ function, the pivoted normalisation function (PIV), the $I(n)L2$ function from the divergence from randomness model (DFR), a modified $BM25$ function ($MBM25$) in which the *idf* part is replaced with the pivoted document length normalisation *idf* function and a learned term-weighting function (ES) [1].

2.2 Experimental Results

Table 2 shows the number of constraint violations per document per query averaged over all the test collections for the top 1000 documents of the best retrieval run. For example, the original $BM25$ function violates all the constraints and, on average, violates constraint 1 an average of 407.5 times for each document for long queries. Thus, this table gives a general view of the constraint violations across the collections. It is intuitive that there are more constraint violations for longer queries as there are more matching query-terms for the average document and therefore more complex interactions.

Tables 3 show the performance of the schemes on the individual collections. The best scheme is in bold and statistical significance (0.05% level) using a one-tailed t-test compared to the next best scheme (DFR) is denoted by an asterisks (*). The best performing scheme across the collections is the scheme which breaks constraints less often on the test collections (i.e. the ES scheme).

The ρ measure is Spearman's correlation between the total number of constraint violations of a function on that particular collection and the MAP of the scheme on that collection. We can see that although the sample size is quite small, the data indicates that there is a consistent inverse correlation between the ranking of the schemes by performance and the ranking of schemes by the total number of constraint violations. The large number of violations of constraints on medium and long queries for the original $BM25$ schemes explains the very poor performance of this scheme on these types of queries (as indicated in the original work [3]) due to the non-removal of stopwords.

Table 3: MAP on test collections

short queries				
Functions	FT	FBIS	FR	OHSUMED
Topics	251-450	300-450	251-450	1-63
ES	**0.2426**	**0.2674***	**0.2653**	-
DRF	0.2353	0.2351	0.2525	-
$MBM25$	0.2322	0.2305	0.2503	-
PIV	0.2243	0.2164	0.2132	-
$BM25$	0.2261	0.2273	0.2473	-
ρ	-0.9	-1.0	-1.0	
medium queries				
ES	0.2545	**0.2687***	**0.2909***	**0.3333***
DRF	**0.2618**	0.2447	0.2622	0.3164
$MBM25$	0.2564	0.2420	0.2581	0.3142
PIV	0.2479	0.2253	0.2243	0.3184
$BM25$	0.1673	0.1207	0.1530	0.2804
ρ	-0.7	-0.9	-1.0	-0.7
long queries				
ES	0.2589	**0.2400**	**0.3151**	-
DRF	0.2715	0.2397	0.2872	-
$MBM25$	**0.2736**	0.2395	0.2893	-
PIV	0.2565	0.2213	0.2553	-
$BM25$	0.0963	0.0445	0.0499	-
ρ	-0.7	-0.9	-1.0	

3. CONCLUSION

We have outlined an approach that counts the number of actual constraint violations using a inductive framework and shown that the total number of constraint violations is inversely correlated with performance.

4. REFERENCES

[1] Ronan Cummins and Colm O'Riordan. An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions. *Artif. Intell. Rev.*, 28(1):51–68, 2007.

[2] Hui Fang, Tao Tao, and ChengXiang 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*, pages 49–56. ACM Press, 2004.

[3] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 480–487. ACM Press, 2005.