Term-Weighting in Information Retrieval using Genetic Programming: A three stage process

Ronan Cummins and **Colm O'Riordan** [1]

1 Introduction

This paper presents term-weighting schemes that have been evolved using genetic programming in an adhoc Information Retrieval model. We create an entire term-weighting scheme by firstly assuming that term-weighting schemes contain a global part, a term-frequency influence part and a normalisation part. By separating the problem into three distinct phases we reduce the search space and ease the analysis of the schemes generated by the process.

Evolutionary computation techniques are proving to be a viable alternative to other standard analytical methods in many areas of IR. Genetic Programming (GP) [2] is an automated searching algorithm inspired by biological evolution. GP has been shown to be an effective approach to learning term-weighting schemes in IR [5]. Firstly, we evolve weighting schemes in a global domain which promote the best terms to use in distinguishing documents. Then, using a suitable global scheme, we evolve term-frequency influence schemes which uses the within-document term-frequency to correctly weight the term-frequency factor. Finally, we evolve normalisation schemes based on the best performing combined global and term-frequency scheme. This framework is an extension of work carried out in [1]. Most term-weighting schemes combine these three aspects to weight query terms and thus score a document in relation to a query.

2 Experimental Framework

The global (gw_t) and normalised term-frequency (ntf) weighting schemes are evolved in a term-weighting function which scores a document (d) in relation to a query (q) as follows:

$$score(d,q) = \sum_{t \in q \cap d} (ntf \times gw_t \times qtf) \quad (1)$$

where qtf is the actual term-frequency of term t in the query. It can be seen that both BM25 [3] and the pivoted normalisation scheme [4] fit this type of model. Tables 1, 2 and 3 show the terminals sets and some GP parameter details for the experiments. The set of functions used for all experiments is $F = \{\times, +, -, /, log, square, square\text{-}root\}$. We use an elitist GP strategy and 4% mutation for all experiments. Mean average precision (MAP) is used as the fitness function in all experiments. All solutions are limited to a depth of 6.

2.1 Document Test Collections

The training set for the global problem consisted of 35,000 OHSUMED documents from 1998 and the 63 topics. The training set

[1] University of Ireland, Galway. email: ronan.cummins@nuigalway.ie, colmor@it.nuigalway.ie

Table 1. Global Weighting Problem Terminals

Terminal	Description
N	no. of documents in the collection
df	document frequency of a term
cf	collection frequency of a term
V	vocabulary of collection (no. of unique terms)
C	size of collection (total number of terms)
0.5	*the constant* 0.5
1	*the constant* 1
10	*the constant* 10
Parameters	7 runs of Population 100 for 50 generations

Table 2. Term-Frequency Weighting Problem Terminals

Terminal	Description
tf	raw term-frequency of a term
1	*the constant* 1
10	*the constant* 10
0.5	*the constant* 0.5
Parameters	7 runs of Population 100 for 50 generations

Table 3. Normalisation Weighting Problem Terminals

Terminal	Description
l	document length (unique terms)
l_{avg}	average document length (unique terms)
l_{dev}	standard deviation of lengths (unique terms)
tl	total document length (all terms)
tl_{avg}	average total document length (all terms)
tl_{dev}	standard deviation of document lengths (all terms)
ql	query length (unique terms)
qtl	query total length (all terms)
1	*the constant* 1
10	*the constant* 10
0.5	*the constant* 0.5
Parameters	7 runs of Population 200 for 25 generations

Table 4. Document Collections

Collection	#Docs	$words/doc$	#Topics	short	medium	long
LATIMES	131,896	251.7	301-350	2.4	9.9	29.9
FBIS	130,471	249.9	351-400	2.4	7.9	21.9
FT91-93	138,668	221.8	401-450	2.4	6.5	18.7
OH90-91	148,162	81.4	0-63	-	7.9	-

Table 5. % MAP of benchmark and evolved schemes on unseen test collections

		short			medium			long		
Collection	Topics	idf_{piv}	idf_{rsj}	gw_t	idf_{piv}	idf_{rsj}	gw_t	idf_{piv}	idf_{rsj}	gw_t
LATIMES	301-350 (m)	17.83	17.91	18.12	19.11	19.16	22.49	13.57	13.79	24.27
FBIS	351-400 (m)	11.19	11.24	11.72	10.30	10.41	15.68	06.76	06.97	13.32
FT91-93	401-450 (m)	21.69	21.69	21.79	27.38	28.15	27.86	23.11	23.13	28.28
OH90-91	0-63 (m)	-	-	-	21.68	21.72	25.69	-	-	-
Collection	Topics	$Piv_{s=0}$	$BM25_{b=0}$	$tf.gw_t$	$Piv_{s=0}$	$BM25_{b=0}$	$tf.gw_t$	$Piv_{s=0}$	$BM25_{b=0}$	$tf.gw_t$
LATIMES	301-350 (m)	20.95	24.75	24.89	13.80	20.55	24.38	10.94	13.98	25.87
FBIS	351-400 (m)	16.30	19.98	20.27	13.40	13.47	19.06	08.45	08.35	16.25
FT91-93	401-450 (m)	22.50	31.38	31.35	23.62	33.03	32.37	19.36	26.59	30.72
OH90-91	0-63 (m)	-	-	-	18.40	25.36	28.80	-	-	-
Collection	Topics	Piv	$BM25$	$ntf.gw_t$	Piv	$BM25$	$ntf.gw_t$	Piv	$BM25$	$ntf.gw_t$
LATIMES	301-350 (m)	24.26	24.17	23.87	25.48	25.61	28.64	25.79	26.77	30.80
FBIS	351-400 (m)	15.90	17.55	19.89	17.92	19.53	24.26	17.59	20.03	24.21
FT91-93	401-450 (m)	30.38	31.27	33.98	34.47	35.33	36.57	34.49	35.35	36.86
OH90-91	0-63 (m)	-	-	-	26.76	28.08	29.84	-	-	-

for the term-frequency influence problem consisted of 32,000 document from the LATIMES collection and 37 medium topics. The training set for the normalisation problem consisted of the same 32,000 documents from the LATIMES collection but we used 12 short, 13 medium and 12 long topics for this problem as it has been suggested that query length may have an impact on normalisation. We tested the solutions for generality on collections from TREC disks 4 and 5 to test our schemes. Table 4 details the collections and lengths of short (title), medium (title and description) and long (title and description and narrative) queries. Standard stop-words are removed and remaining words are stemmed.

2.2 Benchmark Term-Weighting

The full BM25 and pivoted normalisation scheme with default values are used as benchmarks for the entire schemes. The default term-frequency influence value of $k_1 = 1.2$ and normalisation influence value of $b = 0.75$ is used for BM25 while the slope (s) set to 0.2 is used for the pivoted normalisation scheme (Piv). We use the BM25 ($BM25_{b=0}$) and pivoted normalisation scheme ($Piv_{s=0}$) with no normalisation (i.e. assuming all documents are of equal length) as benchmarks for the global schemes combined a term-frequency influence factor. We use the idf scheme as found in the BM25 (idf_{rsj}) and pivoted normalisation scheme (idf_{piv}) as benchmarks for the global part of the scheme. We use the actual within-query term-frequency scheme (qtf) with all schemes as in (1).

2.3 Term-Weighting Scheme

One of the best evolved global schemes is as follows:

$$gw_t = \frac{cf^2\sqrt{cf}}{df^3} \tag{2}$$

The best evolved term-frequency factor, based on the global scheme, is as follows:

$$log(\frac{10}{\sqrt{(0.5/tf)+0.5}}) = log(\sqrt{\frac{200.tf}{1+tf}}) \tag{3}$$

We assume the term-frequency factor is normalised (ntf) as follows:

$$ntf = log(\sqrt{\frac{200.\frac{tf}{n}}{1+\frac{tf}{n}}}) \tag{4}$$

where n is some normalisation factor. One of the best evolved normalisation schemes is as follows:

$$n = \sqrt{log(qtl)} \times log(qtl) \times \frac{l}{l_{avg}} \tag{5}$$

Queries of length one were given the same normalisation as queries of length two during testing as $n = 0$ when $qtl = 1$. This occurred as there was no query of length one in the training set for the normalisation problem.

3 Discussion and Conclusions

It is worth noting that none of the randomly created solutions were as good as the best solution from the final generation. We can see for the global weighting problem that the evolved solution presented has a higher MAP on all topic lengths and collections. The increase over the idf type schemes is quite large for medium and long queries. For the term-frequency influence problem (assuming no normalisation), we can see that the MAP of the evolved solution ($tf.gw_t$) is higher than the benchmarks on most collections at this point especially for longer queries. We can see that the term-frequency part of the Piv scheme is a lot poorer than the default term-influence setting for BM25 at this stage. We can see that normalisation is beneficial to all schemes for medium and long queries but slightly degrades some short queries. We can see that our full evolved scheme is the best performing scheme on the collections for all but the short queries on the LATIMES collection. We have shown that term-weighting schemes can be found by evolutionary techniques that fit certain known aspects of weighting schemes and also contain new features such as query length.

REFERENCES

[1] Ronan Cummins and Colm O'Riordan, 'Evolving general term-weighting schemes for information retrieval: Tests on larger collections.', *Artif. Intell. Rev.*, **24**(3-4), 277–299, (2005).

[2] John R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA, USA, 1992.

[3] S. E. Robertson and S. Walker, 'On relevance weights with little relevance information', in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 16–24. ACM Press, (1997).

[4] A. Singhal, 'Modern information retrieval: A brief overview', *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, **24**(4), 35–43, (2001).

[5] Andrew Trotman, 'Learning to rank', *Information Retrieval*, **8**, 359 – 381, (2005).