An Evaluation of Evolved Term-weighting Schemes in Information Retrieval

Ronan Cummins and Colm O'Riordan
Dept. of Information Technology
National University of Ireland
Galway, Ireland
ronan.cummins@nuigalway.ie, colmor@it.nuigalway.ie

ABSTRACT

This paper presents an evaluation of evolved term-weighting schemes on short, medium and long TREC queries. A previously evolved global (collection-wide) term-weighting scheme is evaluated on unseen TREC data and is shown to increase mean average precision over *idf*. A local (within-document) evolved term-weighting scheme is presented which is dependent on the best performing global scheme. The full evolved scheme (i.e. the combined local and global scheme) is compared to both the BM25 scheme and the Pivoted Normalisation scheme.

Our results show that the local evolved solution does not perform well on some collections due to its document normalisation properties and we conclude that *Okapi-tf* can be tuned to interact effectively with the evolved global weighting scheme presented and increase mean average precision over the standard BM25 scheme.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval, Retrieval models: Search process

General Terms: Algorithms, Measurement.

Keywords: Genetic Programming, Term-Weighting, Information Retrieval.

1. INTRODUCTION

Evolutionary computation techniques are proving to be a viable alternative to other standard analytical methods in many areas of IR. Genetic Programming (GP) [3] has been shown to be an effective approach to learning term-weighting schemes in IR [5]. These approaches, inspired by Darwin's theory of Natural Selection, are stochastic in nature and efficient for searching large complex search spaces.

This paper presents term-weighting schemes which have been evolved in a vector space framework. Our research differs from previous approaches as we break this process into two steps. Firstly, we evolve weighting schemes in a global domain which promote the best terms to use in distinguishing documents. Then, using the best global scheme, we evolve local schemes which use within-document measures to improve the mean average precision (MAP) of the system. This process eases analysis of the evolved schemes and importantly reduces the size of the search space by separating the measures into their respective domains.

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

2. GENETIC PROGRAMMING

GP [3] is an automated searching algorithm inspired by biological evolution. In the GP process, a population of solutions is created randomly. These solutions, encoded as trees, undergo generations of selection, reproduction and mutation until suitable solutions are found. In our system, term-weighting schemes are the individuals in our environment and the MAP of these solutions is used as the fitness function. Thus, the global (gw_t) and local (lw_t) schemes are evolved in the following full weighting for a term t:

$$evol_t = lw_t \times gw_t \times qtf \quad (1)$$

where qtf is the actual term-frequency of term t in the query. It can be seen that BM25 fits this model when gw_t is equal to idf_t and lw_t is equal to $Okapi\text{-}tf$.

3. EXPERIMENTAL SETUP

3.1 Benchmark Term-Weighting Approaches

We use the BM25 scheme and the pivoted normalisation scheme [4] as benchmarks. We use the default term-frequency influence value of $k_1 = 1.2$ and the normalisation influence value of $b = 0.75$. We also use the pivoted normalisation scheme (Piv) with the slope (s) set to 0.2 as a benchmark. We use the actual within-query term-frequency scheme (qtf) with both benchmark schemes as in (1).

3.2 Document Test Collections

We use collections from TREC disks 4 and 5 to test our schemes. A different set of 50 TREC topics is used for each of the collections. For each set of topics we create a short query set (s), consisting of the title field of the topics, a medium length query set (m), consisting of the title and description fields, and a long query set (l) consisting of the title, description and narrative fields. Standard stop-words from the Brown Corpus[1] are removed and remaining words are stemmed using Porter's algorithm. No additional words are removed from the narrative fields as is the case in some approaches.

[1]http://www.lextek.com/manuals/onix/stopwords1.html

Table 1: 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

3.3 Evolved Global-Weighting Scheme

A previously evolved global weighting [1], evolved on about 35,000 OHSUMED documents, is used in this research and is as follows:

$$gw_t = log(\frac{cf_t + \frac{1}{2.\sqrt{\sqrt{cf_t}}}}{df_t}) \times \sqrt{\frac{N}{df_t} \times (\frac{1}{df_t} + 1)} \quad (2)$$

where df_t is the document frequency and cf_t is the frequency of t in the entire collection of N documents.

4. RESULTS

Table 2 shows the MAP of the gw_t against the idf_t measure used in the BM25 scheme with a binary weighting on the within-document weighting ($k_1 = 0$).

Table 2: % MAP for idf_t and gw_t

Collection	Topics	idf_t	gw_t
LATIMES	301-350 (s)	17.91	**19.05**
	301-350 (m)	19.04	**23.31**
	301-350 (l)	13.79	**24.24**
FBIS	351-400 (s)	11.25	**11.73**
	351-400 (m)	10.42	**14.56**
	351-400 (l)	06.97	**14.09**

The increase in MAP for the gw_t scheme over the *idf* benchmark is greater for longer queries. We can see that for the longest queries on both collections there is a considerable amount of noisy terms included in the narrative field of each topic as the MAP decreases for *idf* for the same information need (topic). The evolved global scheme is better able to correctly weight these noisy terms because of the presence of the *cf* terminal which produces a measure of density for a term [1, 2]. In general, the longer the query the better the performance of the evolved global scheme. This is because the number of terms that are assigned inappropriate weights under *idf* increases. For example, the global weighting for a short query of length one is irrelevant because only within-document features will change the score of a document.

Using a similar approach to that adopted in [1], we evolved the following local weighting scheme on the same 35,000 OHSUMED documents while keeping the gw_t global scheme constant:

$$lw_t = \sqrt{(1 + \frac{1}{log(l)}) \times (1 + \frac{log(tf)}{log(l)})} \quad (3)$$

where tf is the term-frequency and l is the number of unique terms in the document. Table 3 shows that the full evolved scheme ($evol_t$) is comparable to $BM25$ in terms of MAP on the LATIMES collections but at lower recall (P@10) is significantly poorer than BM25 on both collections. The main difference between the OHSUMED training data and the other test collections is that the document lengths of the latter are significantly longer. Thus, we incorperated the same term-influence factor as lw_t into *Okapi-tf*. It is trivial to show that this scheme's term-frequency influence is similar to that of *Okapi-tf* when $k_1 = 0.2$. As a consequence, we also evaluate a hybrid scheme which uses the gw_t evolved global weight and *Okapi-tf* when $k_1 = 0.2$. For this scheme, which we will call *ok-*gw_t, b remains 0.75. This hybrid scheme has better normalisation properties than $evol_t$ on longer documents and is consistently better than $BM25$ and Piv in terms of MAP.

Table 3: % MAP and P@10 for Schemes

Collection	Topics	Piv (P@10)	$BM25$ (P@10)	$evol_t$ (P@10)	*ok-*gw_t (P@10)
LATIMES	301-350 (s)	24.26	24.18	24.67	**25.24**
		(48.87)	**(50.37)**	(33.67)	(49.01)
	301-350 (m)	25.48	25.61	25.82	**29.68**
		(52.96)	(53.21)	(46.04)	**(53.51)**
	301-350 (l)	25.79	26.78	26.87	**31.23**
		(53.46)	(54.70)	(45.81)	**(56.89)**
FBIS	351-400 (s)	15.90	17.55	**20.96**	19.29
		(34.51)	**(36.66)**	(36.14)	(36.50)
	351-400 (m)	17.92	19.53	18.75	**22.02**
		(38.58)	(38.77)	(33.67)	**(41.60)**
	351-400 (l)	17.59	20.04	17.09	**23.22**
		(40.70)	(40.96)	(31.22)	**(42.83)**

Table 4 shows the *p-value* for a one-tailed pairwise t-test for each of the 50 topics. It also shows the 95% confidence interval (CI) for the difference in MAP between *ok-*gw_t and $BM25$. We can see that there is an increase in variability in the difference in MAP as the lengths of the topics increase.

Table 4: 95% CI for Δ MAP for *ok-*gw_t and $BM25$

Collection	Topics	*lower-limit*	*upper-limit*	*p-value*
LATIMES	301-350 (s)	-1.42	2.42	0.271
	301-350 (m)	-1.64	9.38	0.051
	301-350 (l)	-2.36	11.45	0.063
FBIS	351-400 (s)	-0.91	4.66	0.058
	351-400 (m)	-1.25	6.47	0.057
	351-400 (l)	-1.81	7.92	0.071

5. CONCLUSIONS

We have shown that global weighting schemes can be found by evolutionary techniques that outperform *idf* on general collections. We have also shown that GP can also be used to tune existing schemes to best interact with newly developed schemes. The most effective scheme identified is a combination of *Okapi-tf* and the evolved global scheme.

6. REFERENCES

[1] Ronan Cummins and Colm O'Riordan. Evolving, analysing and improving global term-weighting schemes in information retrieval. Technical Report NUIG-IT-071204, National University of Ireland, Galway, Ireland, 2004.

[2] Martin Franz and J. Scott McCarley. Word document density and relevance scoring. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference*, pages 345–347. ACM Press, 2000.

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

[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.