An Empirical Study of Query Specificity

Avi Arampatzis[1] and Jaap Kamps[2]

[1] Electrical and Computer Engineering, Democritus University of Thrace, Greece
[2] Media Studies, University of Amsterdam, the Netherlands

Abstract. We analyse the statistical behavior of query-associated quantities in query-logs, namely, the sum and mean of IDF of query terms, otherwise known as *query specificity* and *query mean specificity*. We narrow down the possibilities for modeling their distributions to gamma, log-normal, or log-logistic, depending on query length and on whether the sum or the mean is considered. The results have applications in query performance prediction and artificial query generation.

1 Introduction and Definitions

Inverse document frequency (IDF) is a widely used and robust term weighting function capturing *term specificity* [1]. Analogously, *query specificity* (QS) or query IDF can be seen as a measure of the discriminative power of a query over a collection of documents. A query's IDF is a log estimate of the inverse probability that a random document from a collection of N documents would contain all query terms, assuming that terms occur independently. The mean IDF of query terms, which we call *query mean specificity* (QMS), is a good pre-retrieval predictor for query performance, better than QS [2]. For a query with k terms $1, \dots k$, QS and QMS are defined as

$$\mathrm{QS}_k = \log\left(\prod_{i=1}^{k} \frac{N}{\mathrm{df}_i}\right) = \sum_{i=1}^{k} \log \frac{N}{\mathrm{df}_i}\,, \qquad \mathrm{QMS}_k = \mathrm{QS}_k/k\,,$$

where df_i is the document frequency (DF), i.e. the number of collection documents in which the term i occurs.

We analyse statistical properties of QS and QMS, for all queries in a search engine's query-log and per query length, with an empirical brute-force approach. The proposed models provide insight on engine performance for given query sets. The models can also be combined with query-length models, e.g. [3], for generating artificial queries. Artificial queries have applications in areas such as score normalization for distributed retrieval or fusion [4], pseudo test collection construction [5], and efficiency testing.

2 Distributions of QS and QMS

The distribution of any of QS, QS_k, QMS, QMS_k, is a combined result of a query set and a document collection, i.e. the source of DFs, the query set is submitted to. We use two query sets: the AOL log consisting of 21M queries from AOL search (March–May 2006); and the MSN log consisting 15M queries from the MSN search engine

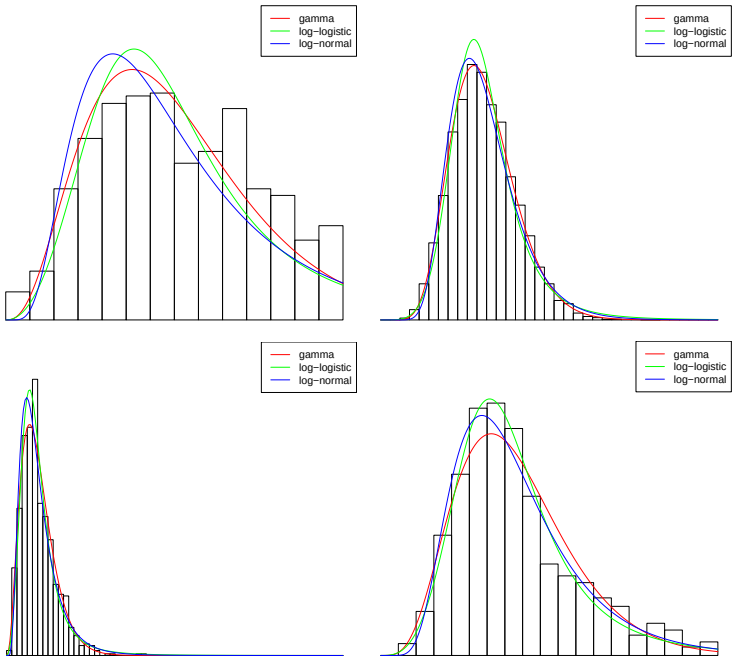

Fig. 1. Empirical distributions of QS and QMS for AOL queries using DFs from NYT, and the fitted gamma, log-logistic, and log-normal for query length $k = 1$ (top left), $k = 5$ (top right), all-lengths QS (bottom left), and all-lengths QMS (bottom right).

(May 2006). Since we have no actual term statistics, we rely on two other sources of DF: Web term document frequencies from the Berkeley Digital Library Project (32M terms from 50M Web pages); and New York Times document frequencies based on the 1998–2000 NY Times articles in AQUAINT (506,433 terms from 314,452 articles).

Per query length k, QS_k is a linear transformation of QMS_k, thus we only need to analyse one of the two and we opt for the latter. We analyse the distributions of QS and QMS irrespective of query length, separately. The quantity QMS_k, as defined in the previous section, has a discrete distribution with a support over $\binom{N+k-1}{k}$ real numbers in $[0, \log N]$. For large N, however, it can be approximated by a continuous distribution with support in $[0, \log N]$, especially for large k where the cardinality of the support set is higher. Thus, as $N \to +\infty$, we are looking for a suitable known continuous distribution supported on the semi-infinite interval $[0, +\infty)$.

By examining histograms of empirical data, we see that the distributions are unimodal with positive skew. Three distributions which seem capable of matching well the shape of the empirical data are: gamma, log-logistic, and log-normal. They all have support in $[0, +\infty)$. Figure 1 shows the fits for $k = 1$, $k = 5$, and QS and QMS over all lengths. We eliminated the following possibilities which gave consistently worse fits: Weibull, inverse gamma, chi-square, and inverse chi-square. We also tried the beta distribution with a bounded support in $[0, 1]$ for normalized QMS and QMS_k, but it

Table 1. χ^2 goodness-of-fit test (upper one-sided at .05 significance) for observed QS and QMS against 3 fitted theoretical distributions: gamma (G), log-logistic (LL), log-normal (LN). The results are presented across all combinations of query set, collection, and query length k. First, large sets of observed data, per length or for all lengths, are uniformly down-sampled to 1,000 points. Then, each set is binned into bins of $0.3\,\sigma$ width, where σ is the standard deviation of the observed data. Bins with expected frequencies < 5 are combined; this may result to slightly different number of bins for the same dataset across candidate distributions. A plus in a cell means that the null hypothesis that the data follow the candidate cannot be rejected, while for a minus it is rejected. The leading numbers are ranks of the quality of fits according to the comparison of their χ^2 with the observed data. This is a loose (although indicative) comparison due to the possibly slightly different degrees of freedom of their χ^2 distributions, a result of the bin-combining.

		AOL/Web			AOL/NYT			MSN/Web			MSN/NYT		
	k	G	LL	LN	G	LL	LN	G	LL	LN	G	LL	LN
QS_k or QMS_k	1	1-	3-	2-	1-	2-	3-	1-	2-	3-	1-	2-	3-
	2	2+	3-	1+	1+	2-	3-	3-	2-	1+	1+	2-	3-
	3	2+	3+	1+	1+	3-	2-	1+	2+	3-	1+	2-	3-
	5	3-	2+	1+	1+	3-	2+	2+	1+	3+	1+	3+	2+
	7	3-	2-	1+	1+	3-	2+	2+	3+	1+	1+	3-	2+
	10	3-	1-	2-	3-	2+	1+	3-	2-	1-	3-	2+	1+
	15	3-	1-	2-	3-	1-	2-	3-	1-	2-	3-	1+	2-
QS		1-	2-	3-	1-	3-	2-	1-	2-	3-	1-	2-	3-
QMS		3-	2-	1-	3-	2-	1-	3-	2-	1-	2-	1-	3-

was consistently worse as well. The inverse Gaussian gave very similar shapes to the log-normal, but we eliminated it due to consistently better fits of the latter.

The goodness-of-fit results are summarized in Table 1. For $k = 1$, the data are messy and difficult to model. This may be due to their discrete nature that comes more into effect for small k, or due to unusual terms like full URLs. However, the gamma seems more flexible than the alternatives. The good fits come at lengths 2 to 7, where the gamma and the log-normal provide better approximations than the log-logistic. At larger k, the log-normal and log-logistic provide better fits than the gamma. Since short queries are more frequent, we are inclined to suggest modeling QS_k and QMS_k with a gamma. The gamma shape of the short lengths and the fact that short queries dominate the aggregate, influence strongly QS, where the gamma is the best fit throughout, but not QMS, where the log-normal fits best and the log-logistic is not bad either.

Since we have not arrived to a single model distribution, we analyse statistics of the datasets, shown in Table 2, rather than a specific distribution's parameters. Using the median as central tendency is more suitable than the mean, since the data are skewed. Given that QMS is correlated with query performance, the fact that the median and standard deviation are declining with increasing k suggests that performance may be declining with query length. But this may not be the case, since past research has found that such correlations may be weakening with increasing k [6], According to the median QMS of the aggregates, AOL and MSN queries would perform better on Web than on NYT. This a multiplicative result of having larger normalized QMS on the Web than on the NYT (as expected for Web query sets), and N being larger for Web than for NYT. The expected result that larger collections improve performance is apparent. Comparing

Table 2. Median and standard deviation of observed *normalized* QS and QMS, across all combinations of query set, collection, and query length k. The median and std. dev. of QS_k, which are not shown, are k times those of QMS_k. In order to enable comparisons across collections of different size, we scale the data by dividing them by $\log N$ per collection. N equals 49,602,191 for Web and 314,452 for NYT. This procedure normalizes QMS, QMS_k in $[0, 1]$, and QS, QS_k in $[0, k_{\max}]$, where $k_{\max}$ is the maximum observed query length.

		AOL/Web		AOL/NYT		MSN/Web		MSN/NYT	
	k	median	std.dev.	median	std.dev.	median	std.dev.	median	std.dev.
QMS_k	1	0.557	0.234	0.476	0.235	0.501	0.241	0.448	0.241
	2	0.402	0.129	0.394	0.151	0.376	0.122	0.375	0.155
	3	0.360	0.102	0.346	0.116	0.343	0.099	0.344	0.121
	5	0.316	0.082	0.291	0.096	0.300	0.067	0.279	0.092
	7	0.283	0.065	0.258	0.086	0.269	0.058	0.244	0.083
	10	0.253	0.068	0.215	0.072	0.244	0.055	0.205	0.074
	15	0.237	0.060	0.189	0.070	0.224	0.047	0.177	0.061
QS		0.901	0.506	0.837	0.511	0.788	0.464	0.807	0.452
QMS		0.395	0.184	0.365	0.180	0.384	0.186	0.362	0.186

the two sets of queries with each other, the QMS indicates a similar performance. This is also expected; we do not see why one query-set would be better than the other.

3 Conclusions

We empirically investigated the distributions of query specificity and mean specificity for query-logs. We have not arrived to a single model, but narrowed down the possibilities considerably. Per query length, both specificity and mean specificity are well approximated with a gamma distribution for short to medium queries, and with a log-normal or log-logistic distribution for long queries. Irrespective of query length, specificity can be approximated with a gamma, and mean specificity by either a log-normal or log-logistic. For all practical purposes, these distributions provide good approximations of all queries in a query-log or per length. We have interpreted the results from a query performance perspective, which may suggest ways to improve performance by a directed expansion of collection coverage or support in query formulation. Further, the proposed models can be applied for artificial query generation.

References

1. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **28** (1972) 11–21
2. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: SIGIR, ACM (2002) 299–306
3. Arampatzis, A., Kamps, J.: A study of query length. In: SIGIR, ACM (2008) 811–812
4. Arampatzis, A., Kamps, J.: A signal-to-noise approach to score normalization. In: CIKM, ACM (2009) 797–806
5. Tague, J., Nelson, M., Wu, H.: Problems in the simulation of bibliographic retrieval systems. In: SIGIR. (1980) 236–255
6. He, B., Ounis, I.: Query performance prediction. Inf. Syst. **31**(7) (2006) 585–594