Relevance Weighting using Within-document Term Statistics

Kai Hui
Graduate University of Chinese Academy of Sciences
huikai10@mails.gucas.ac.cn

Ben He
Graduate University of Chinese Academy of Sciences
benhe@gucas.ac.cn

Tiejian Luo
Graduate University of Chinese Academy of Sciences
tjluo@gucas.ac.cn

Bin Wang
Institute of Computing Technology, Chinese Academy of Sciences
wangbin@ict.ac.cn

ABSTRACT

With the rapid development of the information technology, there exists the difficulty in deploying state-of-the-art retrieval models in environments such as peer-to-peer networks and pervasive computing, where it is expensive or even infeasible to maintain the global statistics. To this end, this paper presents an investigation in the validity of different statistical assumptions of term distributions. Based on the findings in this investigation, a variety of weighting models, called NG (standing for "no global statistics") models, are derived from the Divergence from Randomness framework, in which only the within-document statistics are used in the relevance weighting. Compared to the state-of-the-art weighting models in extensive experiments on various standard TREC test collections, our proposed NG models can provide acceptable retrieval performance in ad-hoc search, without the use of global statistics.

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

General Terms: Experimentation, Performance, Algorithms

Keywords

Probabilistic models, Term frequency distribution fitting, Within-document statistics

1. INTRODUCTION

The probabilistic models are among the most popular information retrieval models for their efficiency and effectiveness. The BM25 probabilistic weighting function [9], the PL2 Divergence from Randomness (DFR) model [1] and the KL-divergence language model (KLLM) [13], have been shown effective in TREC experimentation[1]. All of these three models consider the global term statistics, e.g. the document frequency and the collection-wide term frequency, for the relevance weighting. The use of global statistics in the weighting models can be beneficial to the retrieval effectiveness. However, in large-scale Web applications such as the distribution IR, peer-to-peer network and pervasive computing, it is difficult or even infeasible to maintain the global statistics [2, 11]. To this end, the aim of this paper is to explore the possibility of developing the fairly simple retrieval models that use only within-document statistics for the relevance weighting.

The main contribution of this paper is tri-fold. First, we study the term frequency distribution on various recent IR datasets, results show that other than the classical Poisson assumption proposed by Harter in 1975 [4], there are quite a few appropriate approximations of the actual term frequency distribution in recent datasets. Second, we propose a family of *NG* (No Global statistics) weighting models that use only the within-document statistics for the relevance weighting. Evaluation shows that the weighting models based on the Weibull approximation of the term frequency distribution demonstrates the best retrieval performance and robustness. Third, we develop a new weighting framework that is a simplified form of the DFR framework which can improve not only the retrieval performance, but also the robustness by reducing the parameter sensitivity.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 applies a list of distribution functions to fit the actual term frequencies in standard TREC collections. A list of new NG weighting models using only the within-document statistics for relevance weighting are proposed in Section 4. Evaluation results of these new models are then presented in Section 5. Finally, Section 6 concludes the work and suggests possible future research directions.

2. RELATED WORK

A well-known empirical description of the term frequency distribution in text collection is the so-called Zipf's law, in which a given term's collection-wide frequency is inversely proportional to its rank in the frequency table [15]. Luhn's

[1] Although these three weighting models are based on different assumptions of relevance, they are usually considered to be within the category of probabilistic models [7].

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

work in 1957 is among the earliest research on the term frequency distribution and IR. His research gave an outline for building a system based on statistical method for literature searching by machine [5]. Later in 1975, Harter proposed the 2-Poisson assumption for keyword indexing, in which the informative terms in the documents, namely specialty words, follow Poisson distribution in both the entire document collection, and the elite set [4].

The 2-Poisson assumption has been widely recognized as the standard approximation of the term frequency distribution. The BM25 and PL2 DFR models, both derived based on the 2-Poisson assumption, are currently among the most popular and effective IR models. The formulas of these two models, as well as the KLLM model, are introduced below. These three popular probabilistic models are used as the baselines in our evaluation.

As one of the most established weighting models, BM25 computes the relevance score of a document d for a query Q by the following formula [9]:

$$score(d,Q) = \sum_{t \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \quad (1)$$

where qtf is the query term frequency; $w^{(1)}$ is the idf factor, which is given by:

$$w^{(1)} = \log_2 \frac{N - N_t + 0.5}{N_t + 0.5}$$

N is the number of documents in the whole collection. N_t is the document frequency of term t. K is given by $k_1((1-b) + b\frac{l}{avg\_l})$, where l and avg_l are the document length and the average document length in the collection, respectively. The document length refers to the number of tokens in a document. k_1, k_3 and b are parameters. The default setting is $k_1 = 1.2$, $k_3 = 1000$ and $b = 0.75$ [9]. qtf is the number of occurrences of a given term in the query; tf is the within document frequency of the given term.

Let $tfn = \frac{tf}{(1-b)+b\cdot\frac{l}{avg\_l}}$, where tfn denotes the normalized term frequency, we obtain:

$$score(d,Q) = \sum_{t \in Q} w^{(1)} \frac{k_1(k_1+1)tfn}{k_1 tfn + 1} \frac{(k_3+1)qtf}{k_3+qtf}$$

Hence the term frequency normalization component of the BM25 formula can be seen as:

$$tfn = \frac{tf}{(1-b) + b \cdot \frac{l}{avg\_l}} \quad (2)$$

The Kullback-Leibler divergence language model (KLLM) with Dirichlet smoothing assigns the relevance score as follows [13, 14]:

$$score(d,Q) = \sum_{t \in Q} p(t|\hat{\theta}_Q) log_2(1 + \frac{tf}{\mu P(t|\mathcal{C})}) + log_2 \frac{\mu}{\mu + l} \quad (3)$$

where $p(t|\hat{\theta}_Q)$ is the maximum likelihood of generating query term t from a query model. $p(t|\mathcal{C})$ is the generation probability from the collection model. In this paper, the free parameter μ is set by simulated annealing on training queries.

PL2 is one of the weighting models derived from the Divergence from Randomness (DFR) framework. The DFR framework assigns the relevance score of a document d for a query Q as follows [1]:

$$score(d,Q) = \sum_{t \in Q} qtf \cdot Inf_1 \cdot Inf_2 \quad (4)$$

where qtf is the query term frequency. The first measurement of the information amount Inf_1 is given by the information content $-log_2 P(t, tf|d)$. $P(t, tf|d)$ is the probability of having tf occurrences of query term t in document d. The second measurement of the information amount Inf_2 is given by Laplace succession $\frac{1}{tfn+1}$. The normalized term frequency tfn is given by Normalization 2:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg\_l}{l}), (c > 0) \quad (5)$$

where the recommended setting of free parameter c ranges from 1 to 7, depending on the search task.

Assuming the Poisson distribution of the term frequency in the collection leads to the PL2 model, where the relevance score of a document d for a query Q is given by [1]:

$$\begin{aligned} score(d,Q) = & \sum_{t \in Q} qtf \cdot \frac{1}{tfn+1} \big(tfn \cdot \log_2 \frac{tfn}{\lambda} \\ & + (\lambda + \frac{1}{12 \cdot tfn} - tfn) \cdot \log_2 e + \\ & 0.5 \cdot \log_2(2\pi \cdot tfn) \big) \end{aligned} \quad (6)$$

where λ is the mean and variance of the assumed Poisson distribution. Estimation of λ requires the global statistics of the term's distribution in the whole document collection.

In this paper, we propose to eliminate the need for global statistics in the DFR models by treating the $P(t, tf|d)$ as a function of the within-document term frequency tf, and its parameters of the term frequency distribution function. Various distribution functions are tested in our study on the distribution fitting in the following sections. In addition, as explained in Section 4.2, the average document length avg_l is obtained by averaging the document length of random samples from the collection.

3. FITTING THE TERM FREQUENCY DISTRIBUTION

Harter's [4] study on the term frequency distribution was based on the analysis on a sample of the archives at the Graduate Library of University of Chicago. With the fast growth of the internet in both size and the amount of information in text documents or Web pages, an interesting research question arises: do the terms still follow the Poisson distribution in recent datasets?

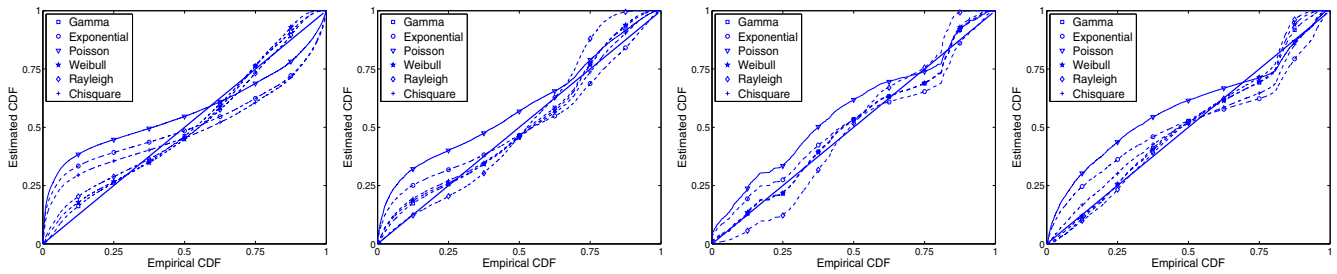
In this section, we attempt to answer the above research question by using a list of statistical distribution functions (Section 3.1) to fit the term frequency distribution in standard TREC collections, as described in Section 3.2. Detailed analysis on the distribution fitting is provided in Section 3.3.

3.1 Distribution Functions

We study a variety of distribution functions as follows:

- Poisson distribution:

$$P(t, tf|d) = \frac{e^{-\lambda}\lambda^{tf}}{tf!} \, (tf = 0, 1, 2, \cdots) \quad (7)$$



(a) disk1&2 (b) disk4&5 (c) WT10G (d) .GOV2

Figure 1: The empirical probabilities (X-axis) against the estimated probabilities (Y-axis) of the distribution fitting.

where λ is a positive real number, which is the expected number of occurrences.

- Gamma distribution:

$$P(t, tf|d) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} (tf \geq 0) \quad (8)$$

where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter.

- Exponential distribution:

$$P(t, tf|d) = \lambda e^{-\lambda tf}, (tf \geq 0) \quad (9)$$

where $\lambda > 0$ is called the rate parameter of the distribution. The distribution is supported on the interval $[0, \infty)$.

- Weibull distribution:

$$P(t, tf|d) = \frac{k}{\lambda} \left(\frac{tf}{\lambda}\right)^{k-1} e^{-(\frac{tf}{\lambda})^k} (tf \geq 0) \quad (10)$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution.

- Rayleigh distribution:

$$P(t, tf|d) = \frac{tf}{\sigma^2} e^{-\frac{tf^2}{2\sigma^2}} (tf \geq 0) \quad (11)$$

where $\sigma > 0$ is the parameter of this distribution.

- χ^2 distribution:

$$P(t, tf|d) = \frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} tf^{\frac{n}{2}-1} e^{-\frac{tf}{2}} (tf \geq 0) \quad (12)$$

where $n = 1, 2, 3 \cdots$ is the degrees of freedom.

3.2 Datasets and the Fitting Method

We use four standard TREC test collections in our study: disk1&2, disk4&5, WT10G and .GOV2. For all four test collections used, each term is stemmed using Porter's English stemmer, and standard English stopwords are removed. Only the query terms in the title field are used.

Weighted Least Squares (WLS) regression is used to determine the optimal parameters in a given distribution function by fitting the target function, namely the empirical cumulative distribution function (ECDF). The use of the CDF function instead of the probability density function (PDF) as the target of the distribution fitting is a common practise because CDF can uniquely determine a distribution, while PDF may not, for example when the PDF of a given distribution cannot be derived, or simply does not exists. The mean sum of square error (SE) between the estimated distribution and the observed distribution of all query terms in the title field is used to indicate the goodness of the fitting.

Table 1: Information about the test collections used.

Coll.	TREC Task	Topics	# Docs
disk1&2	1, 2, 3 ad-hoc	51-200	741,856
disk4&5	Robust 2004	301-450, 601-700	528,155
WT10G	9, 10 Web	451-550	1,692,096
GOV2	2004-2006 Terabyte Ad-hoc	701-850	25,178,548

3.3 Results of the Distribution Fitting

In Harter's work, the distribution fitting is done on the raw term frequencies [4]. We base our investigation on the distribution of the normalized term frequency (tfn) instead. For space reason, we only report the results obtained using BM25's default setting $b = 0.75$.

Figure 1 plots the P-P figures between the fitted probabilities and the empirical probabilities on the four datasets used. The average linear correlation coefficients between ECDF and CDF of the Poisson, χ^2, Exponential, Gamma, Rayleigh, Weibull distribution are 0.9908, 0.9735, 0.9743, 0.9904, 0.9874, and 0.9832 respectively. It shows that all the six distributions can fit the term frequency distribution to some extent on the four test collections being used.

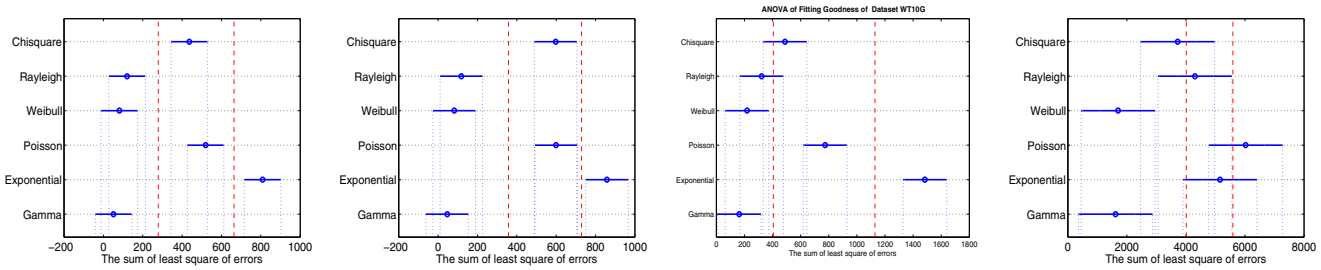
Finally, we conduct ANOVA to investigate the differences of the fitness among the distributions being used. In Figure 2), objects on the left have better fitting goodness than its right ones. Vertical dashed lines separating the objects shows the significant fitness differences in the group. According to Figure 2, Gamma and Weibull distributions demonstrate better fitness than the others.

4. PROPOSED NG MODELS

We describe our proposed NG models (no global statistics models) in Section 4.1, and estimate the average document length, which is a global variable, in Section 4.2.

4.1 Formulas

Our proposed NG models follow the DFR framework as introduced in Equation (4). Under the general probability platform of DFR framework, we can generate new NG



(a) disk1&2 (b) disk4&5 (c) WT10G (d) GOV2

Figure 2: The ANOVA test for the significance of difference in fitting goodness of different distribution functions on the four datasets used.

weighting models by replacing the probability $P(t, tf|d)$ with the proper function form which can fit the real normalized term frequency distribution well.

Table 7 explains how the proposed NG models are generated named. For example, the WL2d model assigns relevance score as follows:

$$
\begin{aligned}
score(d,Q) &= \sum_{t\in Q} qtf \cdot Inf_1 \cdot Inf_2 \qquad (13)\\
&= \sum_{t\in Q} qtf \cdot (-log_2 P(t,tf|d)) \cdot Inf_2\\
&= \sum_{t\in Q} qtf \cdot (-log_2 \frac{k}{\lambda}\left(\frac{tfn}{\lambda}\right)^{k-1} e^{-(\frac{tfn}{\lambda})^k})\\
&\quad \cdot \frac{1}{1+tfn}
\end{aligned}
$$

where the probability $P(tf, t|d)$ is estimated by Weibull (W) distribution in Equation (10), Inf_2 is given by Laplace succession (L) as explained in Table 7. The normalized term frequency *tfn* is given by Normalization 2 in Equation (5). A d at the end of the model name stands for within-document term statistics, indicating that the model is derived from the DFR framework in Equation (4). The distribution parameters λ and k are treated as free parameters that require tuning on training queries.

A notable difference between our proposed NG models and the state-of-the-art probabilistic models is that the former does not involve the use of the global statistics. However, we need to apply term frequency normalization, such as the normalization 2 in Equation (5) and BM25's normalization method in Equation (2), to cope with the bias towards long documents. Both of the above mentioned normalization methods use an expected document length, given by the average document length in the entire document collection, as a normalization factor. In this paper, we propose a method to get the average documents length by estimating through random sampling in the collection and our stimulating experiments have shown good estimating results.

4.2 Estimating the Expected Document Length

We use Systematic Sampling to estimate the average document length in the tf normalization part. A unique feature of Systematic Sampling is that it can give a stable estimate of the random variable, i.e. the average document length

Table 2: The estimated (*EstL*) and the actual average document length (*avg_l*) on the four test collections used, and the error in percentage.

Coll.	*EstL*	*avg_l*	Error (%)
disk1&2	266.10	261.30	1.84
disk4&5	301.22	297.10	1.39
WT10G	406.68	399.28	1.85
GOV2	673.76	648.42	3.91

Table 3: The average error rate (Avg.), maximum positive error rate (MaxPos), minimum negative error rate (MinNeg) of the 10,000 estimated average document length. Avg. is the mean of the absolute values of MaxPos and MinNeg.

Coll.	Avg. (%)	MaxPos (%)	MinNeg (%)	CV
disk1&2	3.15	3.55	-9.23	0.8348
disk4&5	2.72	2.98	-6.8	0.7021
WT10G	3.07	3.90	-8.38	0.8306
GOV2	3.89	0.53	-8.37	0.4470

(*avg_l*) in our case, with only few samples from the dataset, which is therefore very suitable for our study.

Our stimulated tests of the estimates of the average document length on four datasets show that most of the sampled average document length *EstL* fall around the actual average document length, showing the reliability of the sampling method. Some detailed statistics of the tests are shown in Table 3. The table shows that a descent approximation result is observed in the 10,000 times of systematic sampling experiments. When sampling 2.5% of the collection's document length, the average error rate is less than 4%, and the maximum error rate is less than 10%.

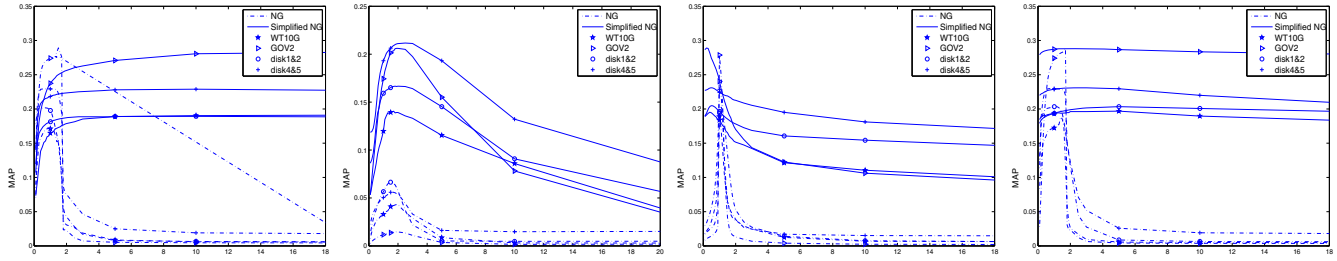
5. RETRIEVAL EVALUATION

Section 5.1 introduces the evaluation settings and Section 5.2 presents the experimental results.

5.1 Evaluation Settings

Our proposed NG models are evaluated against the KLLM, BM25, and PL2 weighting models, which are among the most effective weighting models as shown by the evaluation results in the literature such as the TREC experimentation [12]. The experiments are conducted using an in-house version of the Terrier toolkit [8].

We again use the same TREC collections and their associated title-only queries as in Section 3.2. On each collection,



(a) Gamma distribution's parameter λ (see Eq. 8) (b) Rayleigh distribution's parameter σ (see Eq. 11) (c) Weibull distribution's parameter k (see Eq. 10) (d) Weibull distribution's parameter λ (see Eq. 10)

Figure 3: The parameter value of the distribution functions (X-axis) against the MAP obtained (Y-axis) using the NG models (dashed lines) and its simplified forms (solid lines).

we evaluate the NG models by a two-fold cross-validation. The test topics associated to each collection are split into two equal-size subsets by parity. We use one subset of topics for training, and use the remaining subset for testing. The overall retrieval performance is averaged over the two test subsets of topics. We use the TREC official evaluation measures in our experiments, namely the Mean Average Precision (MAP) [12]. All statistical tests are based on Wilcoxon matched-pairs signed-rank test at the 0.05 level.

Table 4: Experimental results of the baselines.

Coll.	KLLM	PL2	BM25
disk1&2	.2351	.2336	.2404
disk4&5	.2565	.2570	.2535
WT10G	.2153	.2126	.2080
GOV2	.3028	.3042	.2997

5.2 Evaluation Results

The evaluation results of the NG models are presented in this section. Tables 4 and 5 contain the evaluation results of the three state-of-the-art probabilistic models and our proposed NG models, respectively. According the results, we have the observations as follows:

First, the NG models based on Weibull and Poisson distributions provide the best retrieval performance, while those based on the Rayleigh and χ^2 distributions are not as good as expected. This indicates that the effectiveness of the models in the term frequency distribution fitting and the retrieval is not necessarily completely related.

Second, the use of two different normalization methods leads to in general comparable retrieval performance of our proposed models, which conforms to our previous finding in Section 3.3 that these two different normalization methods lead to similar fitting effectiveness.

Third, our proposed NG models based on Weibull and Poisson distribution can lead to comparable retrieval performance with the baseline models as the performance is sometimes not significant different between our models and the BM25 model 5. As a matter of fact, without using global statistics, we indeed have less information than the baseline models, for instance the collection information, e.g. the document frequency and the term frequency in the collection. So our models' performances mainly depend on what distribution function to use and how to use them to describe the real situation.

Finally, Figure 3 examines how the parameter setting affects the NG models' retrieval performance. From the experiments, we find that the retrieval performance of the NG models are highly sensitive to their parameters of their underlying distribution functions, which can potentially hurt the robustness of the models. In the next section, we improve the robustness by a simplified DFR framework.

5.3 Improving the Robustness

In this section, we propose to improve the robustness of the NG models. The underlying idea of our method is to fit the relevance scores produced by a given NG model using a function that is more simple than the information content components in the original DFR framework (i.e. $Inf_1 \cdot Inf_2$ in Equation (4)). In IR applications, it is a common practise to simplify the mathematical forms that are relatively complicated, in order to achieve an easier implementation or a better robustness. For example, Robertson & Walker simplified the 2-Poisson model by fitting the actual term frequency distribution, which eventually led to the proposal of the BM25 model [9, 10].

With an idea similar to [10], we propose a simple form of the DFR framework as given by the following function:

$$score(d, Q) \propto \sum_{t \in Q} qtf \cdot (1 - \beta \cdot P(tf, t|d))$$

Following the similar steps of the distribution function fitting in Section 3, we obtain $\beta = 1$. Thus, a simplified DFR framework is given as follows:

$$score(d, Q) \propto \sum_{t \in Q} qtf \cdot (1 - P(tf, t|d)) \qquad (14)$$

In Table 6, the retrieval performance of the simplified NG models is compared to the state-of-the-art models, and the NG models. The Poisson distribution is not suitable to the simplified DFR framework because of the difficulty caused by expanding the factorials using Stirling formula [1]. From Table 6, we can see that the simplified NG models achieve statistically significant improvement over the NG models on most cases, showing that the simplified DFR framework can indeed enhance the retrieval performance of the NG models.

Moreover, from Figure 3, we can see that the simplified NG models markedly reduce the parameter sensitivity. Overall, the simplified NG models based on the Weibull esti-

Table 5: Experimental results of the comparison between our proposed models and the baselines in Table 4. Statistically significant difference with KLLM, PL2, and BM25 are marked with $*$, †, and ‡, respectively.

Coll.	PL2d	PLBd	CL2d	CLBd	EL2d	ELBd	GL2d	GLBd	RL2d	RLBd	WL2d	WLBd
disk1&2	.2044∗ † ‡	.2032∗ † ‡	.1630∗ † ‡	.1190∗ † ‡	.2004∗ † ‡	.2034∗ † ‡	.2004∗ † ‡	.1988∗ † ‡	.0664∗ † ‡	.0678∗ † ‡	.2024∗ † ‡	.2048∗ † ‡
disk4&5	.2301∗ † ‡	.2178∗ † ‡	.1936∗ † ‡	.1388∗ † ‡	.2294∗ † ‡	.2298∗ † ‡	.2289∗ † ‡	.2132∗ † ‡	.0541∗ † ‡	.0532∗ † ‡	.2300∗ † ‡	.2300∗ † ‡
WT10G	.1934∗†	.1808∗ † ‡	.1055∗ † ‡	.0739∗ † ‡	.1760∗ † ‡	.1926∗ † ‡	.1702∗ † ‡	.1286∗ † ‡	.0436∗ † ‡	.0486∗ † ‡	.1774∗ † ‡	.1878∗ † ‡
GOV2	.2855∗†	.2705∗ † ‡	.1538∗ † ‡	.0715∗ † ‡	.2778∗ † ‡	.2844∗ † ‡	.2635∗ † ‡	.2580∗ † ‡	.0305∗ † ‡	.0200∗ † ‡	.2890∗†	.2890∗†

Table 6: Experimental results using the *simplified* NG models, whose model names end with an *S* for *simplified*. Statistically significant difference with KLLM, PL2, and BM25 (see in Table 4) are marked with $*$, †, and ‡, respectively. A $\star$ indicates a statistically significant improvement over the corresponding model in Table 5, e.g. on disk1&2, R2dS leads to a statistically significant difference over RL2d, which is marked by a $\star$.

Coll.	C2dS	CBdS	E2dS	EBdS	G2dS	GBdS	R2dS	RBdS	W2dS	WBdS
disk1&2	.1924∗ † ‡⋆	.1974∗ † ‡⋆	.1967∗ † ‡	.1966∗ † ‡	.1898∗ † ‡	.1918∗ † ‡⋆	.1664∗ † ‡⋆	.1656∗ † ‡⋆	.2029∗ † ‡	.2048∗ † ‡
disk4&5	.2245∗ † ‡⋆	.2284∗ † ‡⋆	.2258∗ † ‡	.2247∗ † ‡	.2283∗ † ‡	.2280∗ † ‡⋆	.2104∗ † ‡⋆	.1930∗ † ‡⋆	.2304∗ † ‡	.2284∗ † ‡
WT10G	.1857∗ † ‡⋆	.1946∗ † ⋆	.1844∗ † ‡⋆	.1871∗ † ‡⋆	.1904∗ † ‡⋆	.1934∗ † ⋆	.1365∗ † ‡⋆	.1276∗ † ‡⋆	.1934∗†	.1920∗†
GOV2	.2590∗ † ‡⋆	0.2586∗ † ‡⋆	.2664∗ † ‡	.2630∗ † ‡	.2804∗ † ‡⋆	.2866∗ † ⋆	.2012∗ † ‡⋆	.1938∗ † ‡⋆	.2884∗†	.2878∗†

mation of the term frequency distribution, especially W2dS, has demonstrated the best retrieval effectiveness and robustness out of all the NG models proposed. Therefore, it is recommended to apply the W2dS model in large-scale IR applications where it is difficult to maintain the global statistics.

6. CONCLUSIONS AND FUTURE WORK

We have conducted a thorough study of the term frequency distribution on recent TREC collections. Six different distribution functions are used to fit the actual frequency distribution of query terms from TREC collections. Our experimental results show that apart from Poisson distribution there are other probabilistic models of term occurrences are suitable for describing the term frequency distribution in document collections.

Based on the above finding, we have proposed a list of the NG models generated from the DFR framework. A unique feature of the NG models is the exclusion of global statistics. Extensive experiments on four TREC test collections show that our proposed NG models can provide acceptable retrieval performance for ad-hoc search. In addition, we have improved the robustness of the NG models by fitting relevance scoring fitting using simplified NG models. Finally, it is recommended to apply th W2dS model based on the Weibull distribution in large-scale IR applications, for its effectiveness and robustness.

While our proposed models' evaluation results are not good enough, and as to we have conducted a wide investigation about which distributions can be deployed, we plan to refine the form of the distributions appeared in the weighting formula in the future to improve the performance.Besides. Devising a method that automatically estimates the parameters on a per-term basis without the use of global statistics is also scheming. Another future research direction is to investigate the application of query expansion and the term proximity models [6] on top of the NG models. Finally, we plan to investigate the effectiveness of the proposed NG models in large-scale distributed IR applications such as in peer-to-peer networks or sensor networks in practise.

7. REFERENCES

[1] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20(4): 357-389, 2002.

[2] M. Bender, S. Michel, P. Triantafillou, and G. Weikum. Global Document Frequency Estimation in Peer-to-Peer Web Search. In *Proc. WebDB*, 2006.

[3] Witschel, H. F. *Global term weights in distributed environments*. Inf. Process. Manage. 44(3): 1049-1061, 2008.

[4] S. Harter. A probabilistic approach to automatic keyword indexing (part I & II). *Journal of the American Society for Information Science*, 2(6): 197-206 (Part I), 280-289 (Part II), 1975.

[5] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4), 309-317, 1957.

[6] C. Macdonald and I. Ounis. Global Statistics in Proximity Weighting Models. In *Proc. Web N-gram Workshop of ACM SIGIR*, 2010.

[7] C. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2009.

[8] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform information retrieval. In *Proc. ACM OSIR*, 2006.

[9] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *Proc. TREC-4*, 1995.

[10] S. E. Robertson, and S. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In Proc. ACM SIGIR: 232-241, 1994.

[11] C. Viles and J. French. Dissemination of collection wide information in distributed information retrieval systems, In *Proc. ACM SIGIR*, 12-20, 1995.

[12] E. Voorhees. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

[13] C. Zhai. Statistical Language Models for Information Retrieval A Critical Review. *Foundations and Trends in Information Retrieval*, 2(3): 137-213, 2008.

[14] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. ACM SIGIR*, 334-342, 2001.

[15] G. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley. 1949.

Appendix: Table 7 explains the notations in naming the NG models.

Table 7: Naming of the proposed NG models.

Distribution function for $P(tf, t\|d)$			
P:	Poisson, Eq. 7	C:	χ^2, Eq. 12
E:	Exponential, Eq. 9	RL:	Rayleigh, Eq. 11
WB:	Weibull, Eq. 10	G:	Gamma, Eq. 8
Inf_2 in Eq. 4			
L:	Laplace succession, $\frac{1}{tf+1}$		
Term frequency normalization			
2:	Normalization 2, Eq. 5	B:	BM25's normalization, Eq. 2
Suffix of the model name			
d:	NG model derived from the original DFR framework in Eq. 4		
dS:	NG model derived from the simplified DFR framework in Eq. 14		

Acknowledgements: This work is supported in part by CAS e-Learning Fund (Y129017EA2) and the President Fund of GUCAS (Y15101FY00).