Kullback-Leibler Divergence Revisited

Fiana Raiber
Yahoo Research, Israel
fiana@yahoo-inc.com

Oren Kurland
Technion, Israel
kurland@ie.technion.ac.il

ABSTRACT

The KL divergence is the most commonly used measure for comparing query and document language models in the language modeling framework to ad hoc retrieval. Since KL is rank equivalent to a specific weighted geometric mean, we examine alternative weighted means for language-model comparison, as well as alternative divergence measures. The study includes analysis of the inverse document frequency (IDF) effect of the language-model comparison methods. Empirical evaluation, performed with different types of queries (short and verbose) and query-model induction approaches, shows that there are methods that often outperform the KL divergence in some settings.

KEYWORDS

language models, weighted geometric mean

1 INTRODUCTION

Comparing a language model induced from the query with that induced from the document is a standard ranking approach in the language modeling framework to ad hoc document retrieval [20]. The Kullback-Leibler (KL) divergence has been the most commonly used measure for language-model comparison, as it is a natural choice for comparing probability distributions.

The KL divergence is rank equivalent to the cross entropy measure [20] which is in turn rank equivalent to a specific weighted geometric mean [19, 20]: that of the probabilities assigned to terms in the support of the query model[1] by the document language model; the probabilities assigned to these terms by the query language model serve as weights in the mean. Given the rank equivalence between the KL divergence and this weighted geometric mean, we study alternative weighted means for comparing the query and document language models; namely, the arithmetic and harmonic means, integrations of these with the geometric mean, and generalized versions of the means. In addition, we study alternative divergence measures.

Since using the KL divergence for ranking has an IDF (inverse document frequency) effect, we study this effect for the alternative language-model-comparison methods that we consider. The IDF effect holds if the impact of a term on the retrieval score becomes smaller when its corpus frequency increases.

[1]The support is the set of terms assigned a non-zero probability by the language model.

ICTIR '17, October 1–4, 2017, Amsterdam, The Netherlands
 ISBN 978-1-4503-4490-6/17/10…$15.00
DOI: http://dx.doi.org/10.1145/3121050.3121062

We perform extensive empirical evaluation of the various measures for comparing query and document language models using five TREC datasets. We vary the types of queries used (short titles vs. verbose descriptions+titles) and the query-model induction approach: unsmoothed maximum likelihood estimate (MLE) vs. pseudo-feedback-based query-model induction; specifically, we use the relevance model [21] and the mixture model [31].

We found that there are measures that can often outperform the KL divergence for certain types of queries and query-model induction methods. For example, the Power mean [6], which generalizes the harmonic and arithmetic means, and which under certain conditions converges to the geometric mean, outperforms the KL divergence when using MLEs induced from short title queries or pseudo-feedback-based query models. When using MLEs induced from verbose queries, KL is the best performing.

Our key contributions can be summarized as follows:

- We study various weighted means and divergence measures for comparing query and document language models; our study includes analysis of the IDF effect.
- We perform an extensive empirical evaluation and demonstrate that in some settings there are measures that often outperform the commonly used KL divergence.

2 RELATED WORK

Most work on using language models for ad hoc retrieval has focused on improving the language models induced from the query and a document rather than on the measure used to compare them [30]. There are studies of using various divergence measures to compare language models in natural-language-processing tasks; e.g. [23]. However, we are not aware of such in-depth studies for the ad hoc retrieval task. In contrast, document-query similarity estimates were compared in the vector space model [33].

The use of KL divergence for ad hoc retrieval was formally supported in [25]. KL was also shown to outperform a document likelihood approach for relevance-model-based ranking [21]. We show that there are measures more effective than KL for that end.

Axiomatic analysis of retrieval methods in the language modeling framework focused on the query likelihood model [14], translation models [18] and pseudo-feedback-based query models and their smoothing [9, 15]. For example, term frequency, document frequency and document length axioms have been devised and their empirical merits were demonstrated. In contrast, we evaluate different measures for comparing a query model with a document model and analyze whether they have an IDF effect.

3 COMPARING QUERY AND DOCUMENT LANGUAGE MODELS

We address the ad hoc retrieval task: ranking documents d in corpus $\mathcal{D}$ in response to query q. Let θ_q and θ_d denote the unigram

language models induced from q and d, respectively. Henceforth, these are referred to as query and document (language) models, respectively. Various language-model induction methods have been used in work on ad hoc retrieval [30].

To rank documents in the corpus, the document and query language models can be compared [20]. Since the language models are probability distributions defined over the vocabulary, a natural similarity measure, which is the most commonly used in work on ad hoc retrieval, is the Kullback-Leibler (KL) divergence. As higher KL values correspond to decreased similarity, the negative KL divergence is used:

$$-KL(\theta_q, \theta_d) \stackrel{def}{=} -\sum_w p(w|\theta_q) \log \frac{p(w|\theta_q)}{p(w|\theta_d)} = H(\theta_q) - CE(\theta_q, \theta_d). \quad (1)$$

$H(\theta_q) \stackrel{def}{=} -\sum_w p(w|\theta_q) \log p(w|\theta_q)$ is the entropy of the query language model and $CE(\theta_q, \theta_d) \stackrel{def}{=} -\sum_w p(w|\theta_q) \log p(w|\theta_d)$ is the cross entropy between the query and document models. (As is the case for KL, higher values of CE correspond to decreased similarity.) Hereinafter, summations as that in Eq. 1 are applied only over terms w for which $p(w|\theta_q) > 0$ [21]; i.e., terms in the *support* of the query language model. Since the entropy of the query model does not affect ranking, KL and CE are rank equivalent: $KL(\theta_q, \theta_d) \stackrel{rank}{=} CE(\theta_q, \theta_d)$.

If an unsmoothed maximum likelihood estimate (MLE) is used for the query language model, then ranking produced using -CE (and therefore using -KL) is equivalent to that produced using the query likelihood method [20, 29]. Formally, $p(w|\theta_x^{MLE}) \stackrel{def}{=} \frac{\text{tf}(w \in x)}{|x|}$ is the MLE of term w with respect to the text (or text collection) x where $\text{tf}(w \in x)$ is the number of occurrences of w in x; $|x| \stackrel{def}{=} \sum_{w \in x} \text{tf}(w \in x)$ is x's length. Then, the following holds:

$$-CE(\theta_q^{MLE}, \theta_d) = \sum_w p(w|\theta_q^{MLE}) \log p(w|\theta_d) \quad (2)$$
$$= \frac{1}{|q|} \log \prod_w p(w|\theta_d)^{\text{tf}(w \in q)};$$

$\prod_w p(w|\theta_d)^{\text{tf}(w \in q)}$ is q's likelihood with respect to d.

3.1 The IDF Effect

An important implication of the rank equivalence between (negative) cross entropy used with a query MLE and the query likelihood method is the *IDF effect*. That is, Zhai and Lafferty [32] showed that for query likelihood, the effect of query terms on ranking is inversely related to their corpus frequency. Hence, this IDF (inverse document frequency) effect also governs ranking using cross entropy (and KL) with a query MLE.

An interesting question which was somewhat overlooked in past literature is whether using KL or CE with query models which are not maximum likelihood estimates results in an IDF effect.[2] Indeed, while MLE is the standard choice for inducing a query model using only the query terms, there are query models with a much broader support, e.g., pseudo-feedback-based query models [21, 31].

[2]In contrast, there has been work on the importance of selecting and increasing the probability in the query language model of terms with high IDF values (e.g., [9, 15]).

We now show that regardless of the query model utilized, using CE (and hence KL) for ranking results in an IDF effect. The formal argument is similar to that used to demonstrate the IDF effect for the query likelihood model [32]. We provide the details so as to later on contrast the IDF effect entailed by using KL and CE with that entailed by using alternative measures studied below.

We assume a standard smoothed unigram document language model θ_d. For the two most commonly used smoothing methods in work on ad hoc retrieval, Dirichlet and Jelinek-Mercer [32], the document language model is $\theta_d \stackrel{def}{=} (1 - \alpha_d)\theta_d^{MLE} + \alpha_d \theta_{\mathcal{D}}^{MLE}$; for Dirichlet, $\alpha_d = \frac{\mu}{|d|+\mu}$ where μ is a parameter; for Jelinek-Mercer, α_d is simply a constant. For $w \in d$ we define $p_s(w|\theta_d) \stackrel{def}{=} p(w|\theta_d) = (1-\alpha_d)p(w|\theta_d^{MLE}) + \alpha_d p(w|\theta_{\mathcal{D}}^{MLE})$. For $w \notin d$, we define $p_u(w|\theta_d) \stackrel{def}{=} p(w|\theta_d) = \alpha_d p(w|\theta_{\mathcal{D}}^{MLE})$; '$s$' and '$u$' stand for "seen" and "unseen", respectively [32]. The negative CE is then:

$$-CE(\theta_q, \theta_d) = \sum_w p(w|\theta_q) \log p(w|\theta_d) \quad (3a)$$
$$= \sum_{w \in d} p(w|\theta_q) \log p_s(w|\theta_d) + \sum_{w \notin d} p(w|\theta_q) \log p_u(w|\theta_d) \quad (3b)$$
$$= \sum_{w \in d} p(w|\theta_q) \log \frac{p_s(w|\theta_d)}{p_u(w|\theta_d)} + \sum_w p(w|\theta_q) \log p_u(w|\theta_d) \quad (3c)$$
$$= \sum_{w \in d} p(w|\theta_q) \log \frac{p_s(w|\theta_d)}{\alpha_d p(w|\theta_{\mathcal{D}}^{MLE})} + \log \alpha_d - CE(\theta_q, \theta_{\mathcal{D}}^{MLE}) \quad (3d)$$
$$\stackrel{rank}{=} \sum_{w \in d} p(w|\theta_q) \log \left(1 + \frac{1 - \alpha_d}{\alpha_d} \frac{p(w|\theta_d^{MLE})}{p(w|\theta_{\mathcal{D}}^{MLE})}\right) + \log \alpha_d \quad (3e)$$

The transitions are based on separating groups of indices (Eq. 3b), ignoring document independent factors for ranking (Eq. 3e) and using definitions and arithmetic manipulations.

The summation in Eq. 3e is over terms in the query-model support that appear in d. It constitutes the "document-query match" which is based on the probability of query terms in the query ($p(w|\theta_q)$), document ($p(w|\theta_d^{MLE})$) and corpus ($p(w|\theta_{\mathcal{D}}^{MLE})$) models. The latter yields the "IDF effect" as it regularizes the impact of term occurrence in the document by occurrence in the corpus. Document length is used as a normalizer in $p(w|\theta_d^{MLE})$. If Dirichlet smoothing is applied, then α_d also (inversely) depends on document length.

3.2 Weighted Means

The negative cross entropy amounts to (cf. [19, 20]):

$$-CE(\theta_q, \theta_d) = \sum_w p(w|\theta_q) \log p(w|\theta_d) \stackrel{rank}{=} \prod_w p(w|\theta_d)^{p(w|\theta_q)}.$$

That is, negative CE is rank equivalent to a weighted geometric mean (henceforth **Geo**), $Geo(\theta_d; \theta_q)$, of the probabilities assigned to terms in θ_q's support by the document model θ_d; the probabilities assigned by θ_q serve as weights whose sum is 1.

The view of (negative) CE as a weighted geometric mean gives rise to the question, which to the best of our knowledge this paper is the first to explore, of whether using other means to compare the query and document language models can improve retrieval effectiveness. There are numerous weighted means (e.g., [4, 6]). In what follows we mainly focus on the three classical and most widely

used means: arithmetic, geometric and harmonic, their integrations and generalizations. Table 1 presents the weighted means we study.

All these means, as well as the divergence measures we discuss in Section 3.3, include a "query-document match" component[3] which relies on occurrence of terms from the query model support in the document; document length normalizes this occurrence since MLE-based estimates are used. Hence, in addition to discussing general properties of the means and divergence measures, we study whether their query-document match component has an IDF effect. Analyzing more complicated connections between document length, IDF and term frequency (cf. [14]) is left for future work.

The weighted arithmetic mean (**Ari**) is less "conservative" than the weighted geometric mean applied by CE; i.e., it is less affected by outliers. Specifically, a low probability assigned by a document model to one of the terms in the query model support incurs higher penalty in Geo than in Ari. Using manipulations similar to those in Eq. 3, Ari amounts to:

$$Ari(\theta_d;\theta_q) \stackrel{rank}{=} (1-\alpha_d)\sum_{w\in d} p(w|\theta_q)p(w|\theta_d^{MLE}) + \alpha_d \sum_w p(w|\theta_q)p(w|\theta_{\mathcal{D}}^{MLE})\,.$$

The document-query match (the first summation) has no IDF effect.

The weighted harmonic mean (**Har**), which is more conservative than Geo, is the third classical mean. Using manipulations similar to those in Eq. 3, we arrive at:

$$Har(\theta_d;\theta_q) \stackrel{rank}{=} \Big(\sum_{w\in d}\frac{(\alpha_d-1)p(w|\theta_q)p(w|\theta_d^{MLE})}{\alpha_d p_s(w|\theta_d)p(w|\theta_{\mathcal{D}}^{MLE})} + \frac{1}{\alpha_d}\sum_w \frac{p(w|\theta_q)}{p(w|\theta_{\mathcal{D}}^{MLE})}\Big)^{-1}.$$

The document-query match (first summation) has an IDF effect since $\alpha_d < 1$ and $p_s(w|\theta_d)$ increases with increasing values of corpus frequency. Note that the second summation has an "inverse" IDF effect (i.e., DF effect) which can be considered a drawback of Har: the more frequent a query term in the corpus, the higher the retrieval score.[4] However, if α_d is the same for all documents, then the second summation has no effect on ranking, and the entire retrieval score (used to rank documents) exhibits an IDF effect as was the case for the KL divergence in the transition from Eq. 3d to Eq. 3e . Indeed, using Jelinek-Mercer smoothing or assuming equal document lengths in analyzing the IDF effect yields a constant α_d. Using the equal document lengths assumption is in line with Fang and Zhai's methodology of analyzing the IDF effect [13]: assuming two documents of the same length, the document with more occurrences of query terms that are less frequent in the corpus should receive a higher retrieval score. Accordingly, hereinafter, and as mentioned above, our analysis of means and divergence measures will focus on whether the query-document match has an IDF effect.

The next two means integrate the geometric mean employed by negative CE with the arithmetic mean (**GeoAri**) and the harmonic mean (**GeoHar**). These means are computed in iterations that are guaranteed to converge. In the first iteration, $i = 1$, we initialize $g_1 = Geo(\theta_d;\theta_q)$, $a_1 = Ari(\theta_d;\theta_q)$ and $h_1 = Har(\theta_d;\theta_q)$. Then, for GeoAri, in iteration $i+1$ we set $g_{i+1} = \sqrt{a_i g_i}$ and $a_{i+1} = \frac{1}{2}(a_i + g_i)$ until g_{i+1} and a_{i+1} converge to the same value.[5] Similarly, for GeoHar, we compute $g_{i+1} = \sqrt{h_i g_i}$ and $h_{i+1} = 2(h_i^{-1} + g_i^{-1})^{-1}$ until convergence.[6] Since Geo and Har have the IDF effect, and Ari does not, GeoHar has the effect while GeoAri has it to a somewhat limited extent. The following inequality holds for the means considered thus far:

$$Ari(\theta_d;\theta_q) \geq GeoAri(\theta_d;\theta_q) \geq Geo(\theta_d;\theta_q) \geq GeoHar(\theta_d;\theta_q) \geq Har(\theta_d;\theta_q)\,.$$

Table 1: Weighted means. β and γ are free parameters.

M	$M(\theta_d;\theta_q)$	IDF Effect
Ari	$\sum_w p(w\|\theta_q)p(w\|\theta_d)$	✗
Har	$\left(\sum_w \frac{p(w\|\theta_q)}{p(w\|\theta_d)}\right)^{-1}$	✓
GeoAri	Integration of $Geo(\theta_d;\theta_q)$ and $Ari(\theta_d;\theta_q)$	✓
GeoHar	Integration of $Geo(\theta_d;\theta_q)$ and $Har(\theta_d;\theta_q)$	✓
Power	$\left(\sum_w p(w\|\theta_q)p(w\|\theta_d)^\beta\right)^{\frac{1}{\beta}}$	$\beta < 1$
Lehmer	$\frac{\sum_w p(w\|\theta_q)p(w\|\theta_d)^\gamma}{\sum_w p(w\|\theta_q)p(w\|\theta_d)^{\gamma-1}}$	$0 \leq \gamma < 1$

The weighted **Power** mean is essentially a family of means that covers a wide range of aggregates bounded by the minimal and maximal values aggregated: $\lim_{\beta\to-\infty} Power(\theta_d;\theta_q) = \min_w p(w|\theta_d)$ and $\lim_{\beta\to\infty} Power(\theta_d;\theta_q) = \max_w p(w|\theta_d)$. For $\beta = 1$ and $\beta = -1$ Power amounts to the arithmetic and harmonic means, respectively. Furthermore, $\lim_{\beta\to 0} Power(\theta_d;\theta_q) = Geo(\theta_d;\theta_q)$. (Recall that Geo and negative CE are rank equivalent.) Using a similar transition to that from Eq. 3b to Eq. 3c it can be shown that the query-document match component in Power has an IDF effect for $\beta < 1$. In Section 4.2 we show that β close to 0 yields the best retrieval performance which also often transcends that of using Geo (CE).

As additional reference comparison we consider **Lehmer**, which is a special case of the Gini family of means [4]. For $\gamma = 1$ and $\gamma = 0$, Lehmer amounts to the arithmetic and harmonic means, respectively. It can be shown that the query-document match component in Lehmer has an IDF effect for $0 \leq \gamma < 1$, but not for $1 \leq \gamma \leq 2$. The analysis for $\gamma < 0$ and $\gamma > 2$ is quite involved and might require numerical simulation as the query-document match component and document-independent factor have mutual effects.

3.3 Divergence Measures

KL is a measure of the divergence between two probability distributions — query and document language models in our case. We therefore now turn to examine alternative divergence measures. These measures, presented in Table 2, were used in a variety of tasks, including term co-occurrence estimation [23], topic-based document segmentation [5], story link detection [7], query performance prediction [3] and static index pruning [8]. We rank documents by ascending values of the divergence measures.

[3] Other components are document-independent or α_d; sometimes, these two interact.
[4] This is also the case for the second summation in Ari.
[5] g_{i+1} is the geometric mean of g_i and a_i, while a_{i+1} is their arithmetic mean.
[6] g_{i+1} is the geometric mean of g_i and h_i, while h_{i+1} is their harmonic mean.

Table 2: Divergence measures. $p(w|\theta_{qd}^{[\eta]}) \stackrel{def}{=} \eta p(w|\theta_q) + (1-\eta)p(w|\theta_d)$**; η is a free parameter.**

D	$D(\theta_q, \theta_d)$	IDF Effect
Hellinger	$\sqrt{\sum_w \left(\sqrt{p(w\|\theta_q)} - \sqrt{p(w\|\theta_d)}\right)^2}$	✓
TotalVariation	$\sum_w \lvert p(w\|\theta_q) - p(w\|\theta_d)\rvert$	✗
JensenShannon	$KL(\theta_q, \theta_{qd}^{[\eta]}) + KL(\theta_d, \theta_{qd}^{[\eta]})$; $\eta = \frac{1}{2}$	✓
J	$KL(\theta_q, \theta_d) + KL(\theta_d, \theta_q)$	✓
ResistorAverage	$\left(KL(\theta_q, \theta_d)^{-1} + KL(\theta_d, \theta_q)^{-1}\right)^{-1}$	✓
χ^2Neyman	$\sum_w \frac{(p(w\|\theta_q)-p(w\|\theta_d))^2}{p(w\|\theta_d)}$	✓
χ^2Pearson	$\sum_w \frac{(p(w\|\theta_q)-p(w\|\theta_d))^2}{p(w\|\theta_q)}$	✓
χ^2Symmetric	$\sum_w \frac{(p(w\|\theta_q))-p(w\|\theta_d))^2}{p(w\|\theta_q)+p(w\|\theta_d)}$	✓
Skew	$KL(\theta_q, \theta_{qd}^{[\eta]})$	✓

Most of the measures we consider are special cases of the f-divergence [11][7]. Given a convex function f defined over $(0, \infty)$ such that $f(1) = 0$, the f-divergence between θ_q and θ_d is:

$$\sum_w p(w|\theta_d) f\left(\frac{p(w|\theta_q)}{p(w|\theta_d)}\right).$$

Different choices of f result in different divergence measures. For example, setting $f(x) = x \log x$ results in the KL divergence, setting $f(x) = (1-\sqrt{x})^2$ yields the **Hellinger** divergence, and setting $f(x) = |x-1|$ results in the **TotalVariation** distance.

Some of the measures we consider are not metrics as they do not satisfy at least one of the following properties: non-negativity, identity of indiscernibles[8], symmetry and triangle inequality. KL divergence, for example, is not a metric since symmetry and triangle inequality do not hold. TotalVariation, on the other hand, is a metric satisfying all four properties.

Numerous symmetric versions of the KL divergence were proposed. These include **JensenShannon** [24] (defined using the mean language model of θ_q and θ_d), **J** [16] and **ResistorAverage** [17]. The J divergence is not a metric despite being symmetric since the triangle inequality does not hold. The square root of the Jensen-Shannon divergence is a metric.[9]

Three additional instances of the f-divergence are the asymmetric χ^2**Neyman** [26] and χ^2**Pearson** [27] , where $\chi^2 Pearson(\theta_q, \theta_d) = \chi^2 Neyman(\theta_d, \theta_q)$, and their symmetric version χ^2**Symmetric**. The functions f in f-divergence that yield these three measures are $f(x) = \frac{(x-1)^2}{x}$, $f(x) = (1-x)^2$ and $f(x) = \frac{(x-1)^2}{x+1}$, respectively.

The **Skew** divergence [22, 23] was proposed to address cases where $KL(\theta_x, \theta_y)$ is not defined because the support of θ_x is not a subset of the support of θ_y. In our case, $KL(\theta_q, \theta_d)$ is defined since smoothed document language models are used.

We next make a few observations about the IDF effect in the query-document match components of the divergence measures. It can be shown that Hellinger, χ^2Neyman, χ^2Pearson and χ^2Symmetric exhibit the IDF effect. (More precisely, since we rank by ascending order of divergence values, the negative divergence measures exhibit, or not, the IDF effect.) In contrast, it is easy to verify that TotalVariation does not have an IDF effect.

For the negative J divergence, the negative $KL(\theta_q, \theta_d)$ employs an IDF effect since it is rank equivalent to -$CE(\theta_q, \theta_d)$ (see Eq. 1) and negative CE employs an IDF effect as shown in Eq. 3e. To estimate $-KL(\theta_d, \theta_q)$, we use an unsmoothed MLE for the document model and a query model smoothed via Jelinek-Mercer with the corpus (with parameter α_q), denoted θ_q^s. (See Section 4.1 for further details.) Using the fact that $KL(\theta_d^{MLE}, \theta_q^s) = -H(\theta_d^{MLE}) + CE(\theta_d^{MLE}, \theta_q^s)$, and applying manipulations as in Eq. 3, we get:

$$-KL(\theta_d^{MLE}, \theta_q^s) \stackrel{rank}{=} \sum_w p(w|\theta_d^{MLE}) \log\left(1 + \frac{1-\alpha_q}{\alpha_q} \frac{p(w|\theta_q)}{p(w|\theta_{\mathcal{D}}^{MLE})}\right) + \log \alpha_q - KL(\theta_d^{MLE}, \theta_{\mathcal{D}}^{MLE});$$

the summation is over w that are in both d and θ_q's support. High values of $p(w|\theta_{\mathcal{D}}^{MLE})$ reduce the value of the document-query match (the summation), and hence there is an IDF effect. This effect is somewhat counter balanced by the fact that documents with models similar to that of the corpus, i.e., with low $KL(\theta_d^{MLE}, \theta_{\mathcal{D}}^{MLE})$, are rewarded. Overall, since $-KL(\theta_q, \theta_d)$ and $-KL(\theta_d, \theta_q)$ have an IDF effect in their query-document match components, so does (negative) J divergence.

Using similar arguments to those used for the J divergence, we can show that ResistorAverage has an IDF effect. Along the same lines, the KL divergence factors that constitute JensenShannon and Skew have an IDF effect. However, in both measures, higher values of η make $\theta_{qd}^{[\eta]}$ more similar to θ_q. This causes loss of information about the original differences between θ_q and θ_d.

4 EMPIRICAL EXPLORATION

Our next goal is studying the retrieval effectiveness of using the different measures discussed above for comparing document and query language models.

4.1 Experimental Setup

The five datasets used for experiments are specified in Table 3. AP and ROBUST are small, mostly newswire, collections; GOV2 is a crawl of the .gov domain and CW09B is the Category B of the ClueWeb09 collection. An additional dataset, CW09BF, was created for ClueWeb09 Category B by filtering out from the initial document ranking documents assigned with a score below 50 by Waterloo's spam classifier [10]. Further details about the initial ranking are provided below. Unless stated otherwise, topic titles, which are short, served for queries. In Section 4.2.3 we also present evaluation when using verbose queries composed of both the title and description. We applied Krovetz stemming to documents and queries and removed stopwords on the INQUERY list only from queries. The Indri toolkit[10] was used for experiments.

We used a two-phase retrieval approach to compare the effectiveness of the different measures discussed in Section 3. In the first phase, an initial list of documents is retrieved using $KL(\theta_q, \theta_d)$,

[7] Also know as the Ali-Silvey distance [2].
[8] For metric D and a pair of models θ_x and θ_y, $D(\theta_x, \theta_y) = 0$ iff $\theta_x = \theta_y$.
[9] The f functions in f-divergence that yield the J and JensenShannon divergence are $f(x) = (x-1)\log x$ and $f(x) = \frac{1}{2}\log\frac{2}{1+x} + \frac{1}{2}x\log\frac{2x}{1+x}$, respectively.
[10] www.lemurproject.org/indri

Table 3: TREC data used for experiments.

corpus	# documents	data	queries
AP	242,918	Disks 1 – 3	51 – 150
ROBUST	528,155	Disks 4 – 5 (–CR)	301 – 450, 600 – 700
GOV2	25,205,179	GOV2	701 – 850
CW09B CW09BF	50,220,423	ClueWeb09 Category B	1 – 200

where an unsmoothed unigram language model, specifically, a maximum likelihood estimate (**MLE**), is used to induce θ_q as is standard [20]. In the second phase, $M(\theta_d; \theta_q)$ or $D(\theta_q, \theta_d)$ is used to re-rank the 10000 most highly ranked documents in the initial ranking, where M and D are the weighted mean and divergence measures presented in Section 3, respectively. We take a re-ranking approach since applying numerous configurations of query expansion upon large-scale corpora is computationally expensive. For a uniform experimental setting, this approach is used in all cases even if query expansion is not employed. It is important to note that re-ranking an initial list of documents was shown to yield similar performance to ranking the entire corpus for pseudo-feedback query expansion which we use here, especially for precision-oriented evaluation metrics [12]. We further note that performance patterns similar to those reported here were also observed in experiments with re-ranking a shorter list of 1000 documents. (Actual numbers are omitted as they convey no additional insight.)

We use three query-model induction methods in the second phase. The first utilizes unsmoothed MLE induced from the original query as was the case in the first phase.[11] The additional two models are relevance model #3 (**RM3** [1, 21]) and the mixture model (**MM** [31]). For these two models, θ_q is induced from the top 50 ranked documents in the initial list. The resultant query language models are (much) richer than— i.e., their support is much larger than that of — the one induced using MLE from a short query; these query models represent expanded queries of the short title query.

The document model θ_d is in both phases a Dirichlet-smoothed unigram language model with the smoothing parameter μ=1000 [32]. Since the document model is smoothed, $KL(\theta_q, \theta_d)$ is always well defined (i.e., the support of θ_q is a subset of the support of θ_d). In contrast, $KL(\theta_d, \theta_q)$, which is used in J and ResistorAverage, might be undefined if the support of θ_d is not a subset of the support of θ_q. We took the following two approaches to address this issue: (i) applied Jelinek-Mercer smoothing [32] with $\lambda = 0.1$ to θ_q, or (ii) considered only terms that appear in the support of both θ_q and θ_d. The document model is unsmoothed in both cases, i.e., θ_d^{MLE} is used.[12] For the J divergence, the first approach was found in our experiments to be more effective for short title queries and is therefore used in Section 4.2.1, while the second was found to be more effective for longer queries with a larger support and is used in Sections 4.2.2 and 4.2.3. For ResistorAverage, the first approach resulted in better performance in all cases. We note that the first approach entails an IDF effect for $-KL(\theta_d^{MLE}, \theta_q)$ (see Section 3.3) while the second does not.

[11] Note that the document ranking produced using $KL(\theta_q, \theta_d)$ in the second phase is in this case the same as the initial ranking.

[12] This is also the case for $KL(\theta_d, \theta_{qd}^{[\eta]})$ in JensenShannon. We note that JensenShannon is always well defined since θ_q is smoothed with θ_d in $\theta_{qd}^{[\eta]}$.

To estimate retrieval effectiveness, we use mean average precision (**MAP**), precision of the top 5 ranked documents (**p@5**), normalized discounted cumulative gain of the top 20 ranked documents (**NDCG20**) and reliability of improvement (**RI**) [28]. RI is the difference between the number of queries whose AP performance is improved and the number of queries whose AP performance is hurt compared to the initial ranking divided by the total number of queries. Ten-fold cross validation is used to set free-parameter values where MAP serves for optimization in the training phase. Folds are determined based on query IDs. The reported performance is the average over all queries in a dataset when these serve for testing. (Each query belongs to a single test fold.) Statistically significant differences of retrieval performance between two methods are computed over all the queries in a dataset when these serve for testing. The two-tailed paired t-test with $p \leq 0.05$ is used.

To construct RM3 and MM, the number of terms and the weight of the original query are set to values in $\{25, 50\}$ and $\{0.1, 0.2, \ldots, 0.9\}$, respectively. Unsmoothed document language models (MLE) are used to induce RM3. The mixture weight of the corpus in MM is in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The values of η (in Skew), β (in Power) and γ (in Lehmer) are selected from $\{0.1, 0.2, \ldots, 0.9\}$, $\{\pm 0.05, \pm 0.15, \pm 0.25, \pm 0.5, \pm 1, \pm 2, \pm 3, \pm 10\}$, and $\{0, \pm 0.05, \pm 0.15, \pm 0.25, \pm 0.5, \pm 1, \pm 2, \pm 3, \pm 10\}$, respectively.

4.2 Experimental Results

4.2.1 Short Queries. Table 4 presents the results of using the unsmoothed MLE query language model θ_q^{MLE} — i.e., only the terms in the original short query are assigned with a non-zero probability. We see that in most cases Power is the best performing measure. Power outperforms KL in 12 out of the 15 relevant comparisons for the MAP, p@5 and NDCG20 evaluation metrics (5 datasets × 3 evaluation metrics) and in 5 of these the improvement is statistically significant; Power is outperformed (but not statistically significantly) by KL in a single case (NDCG20 for ROBUST). Power is the only measure among those considered that is never statistically significantly outperformed by KL. It is also the only measure with positive RI values across all five datasets, indicating that the number of queries for which average precision performance is improved compared to the initial ranking (KL) is higher than the number of queries for which performance is hurt.

Two additional measures outperforming KL in the majority of the relevant comparisons are GeoHar (in 8 relevant comparisons; 4 are statistically significant) and Lehmer (in 9 relevant comparisons; 5 are statistically significant). All other measures are outperformed by KL in most relevant comparisons and many of the differences are statistically significant.

The descending order of the seven weighted means according to a pairwise comparison of retrieval effectiveness[13] is Power, Lehmer, GeoHar, KL (Geo), Har, GeoAri and Ari. The poor performance of Ari could be explained by the fact that it has no IDF effect as shown in Section 3.2. Har is worse than Geo potentially due to being too conservative. GeoAri improves over Ari but not over Geo while GeoHar improves over both Geo and Har, which attests to the merits of integrating these two means. Power and Lehmer rely

[13] The order is determined by counting the number of relevant comparisons (out of the 15) in which a method outperforms another method.

Table 4: Short title queries. Using unsmoothed MLE to induce query models. Because the KL divergence is used to produce the initial ranking, its RI is undefined. 'Init': initial ranking. 'k': statistically significant difference of retrieval effectiveness (MAP, p@5 and NDCG20) with the KL divergence. The best result in a column is highlighted.

	AP				ROBUST				GOV2				CW09B				CW09BF			
	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI
KL (=Init)	21.1	43.6	42.6	–	25.4	48.7	43.9	–	29.2	55.5	44.5	–	17.9	22.7	20.1	–	18.7	34.7	25.8	–
Ari	10.5_k	24.6_k	24.4_k	−69.7	6.4_k	14.5_k	12.7_k	−83.5	7.0_k	9.6_k	9.1_k	−93.2	6.6_k	5.8_k	6.0_k	−66.7	7.3_k	13.2_k	9.7_k	−67.2
Har	15.5_k	39.4_k	35.8_k	−55.6	22.7_k	47.1	41.3_k	−36.5	26.5_k	57.0	44.4	−24.3	19.5_k	28.6_k	23.8_k	27.3	19.1	35.5	26.7	8.6
GeoAri	18.0_k	41.4	39.1_k	−35.4	16.0_k	37.8_k	32.4_k	−70.7	17.1_k	34.5_k	27.4_k	−87.8	10.0_k	8.6_k	9.2_k	−66.7	12.4_k	23.9_k	16.9_k	−63.1
GeoHar	17.2_k	41.0	37.7_k	−47.5	24.0_k	47.7	42.1_k	−24.1	28.0_k	57.2	45.3	−8.1	19.6_k	27.6_k	23.3_k	35.4	19.4	35.9	26.8_k	19.7
Power	21.2	45.1	42.6	19.2	25.5	48.9	43.8	5.2	29.2	55.8	44.9_k	23.0	19.8_k	27.9_k	23.6_k	37.4	19.2	35.6	26.7_k	19.7
Lehmer	20.8	44.2	41.6_k	−9.1	25.0_k	48.8	43.7	−7.6	29.0	55.4	44.7	10.8	19.5_k	28.0_k	23.4_k	31.8	19.4	36.0	26.8_k	20.7
Hellinger	17.7_k	41.8	39.3_k	−30.3	14.2_k	34.1_k	28.9_k	−73.5	14.3_k	29.7_k	23.1_k	−89.2	9.9_k	10.8_k	10.0_k	−66.7	11.9_k	23.6_k	16.3_k	−65.2
TotalVariation	10.5_k	24.6_k	24.4_k	−70.7	6.4_k	14.5_k	12.7_k	−83.1	7.0_k	9.6_k	9.1_k	−92.6	6.6_k	5.8_k	6.0_k	−66.2	7.3_k	13.2_k	9.7_k	−67.2
JensenShannon	10.6_k	32.3_k	28.1_k	−74.7	8.3_k	22.9_k	18.9_k	−80.3	4.3_k	10.0_k	8.3_k	−94.6	4.4_k	10.1_k	7.3_k	−88.4	3.9_k	13.4_k	8.3_k	−85.9
J	19.0_k	41.0	40.7_k	−40.4	23.0_k	47.9	42.0_k	−30.9	28.5	60.4_k	46.3	−12.2	17.9	23.9	21.6	5.6	19.4	39.1_k	28.4_k	2.5
ResistorAverage	14.7_k	40.0	36.6_k	−66.7	15.5_k	40.2_k	33.7_k	−74.3	23.1_k	56.4	42.6	−51.4	16.8	21.5	19.6	4.0	18.2	37.2	26.7	−4.5
χ^2Neyman	15.5_k	39.4_k	35.8_k	−55.6	22.7_k	47.1	41.3_k	−36.5	26.5_k	56.9	44.4	−24.3	19.5_k	28.6_k	23.8_k	27.3	19.1	35.5	26.7	8.6
χ^2Pearson	10.6_k	25.3_k	24.6_k	−69.7	6.5_k	14.9_k	13.0_k	−83.9	7.1_k	10.1_k	9.4_k	−92.6	6.6_k	6.1_k	6.2_k	−66.2	7.4_k	13.5_k	9.8_k	−66.7
χ^2Symmetric	10.7_k	26.7_k	25.4_k	−67.7	6.6_k	15.3_k	13.3_k	−82.7	7.3_k	11.1_k	10.2_k	−92.6	6.8_k	6.5_k	6.4_k	−66.7	7.5_k	13.9_k	10.1_k	−67.2
Skew	9.1_k	22.4_k	21.3_k	−84.8	6.4_k	15.7_k	13.6_k	−87.6	3.1_k	6.2_k	4.5_k	−95.9	4.1_k	8.8_k	6.5_k	−86.4	3.8_k	10.7_k	7.6_k	−89.4
Cosine	7.5_k	21.8_k	19.4_k	−83.8	10.1_k	19.2_k	18.1_k	−89.6	4.1_k	13.4_k	8.7_k	−95.9	1.5_k	5.6_k	3.5_k	−88.9	1.4_k	5.8_k	3.1_k	−88.9

on a free parameter in contrast to the other means — some of which they generalize as noted in Section 3.2. We further study the effect of β on the retrieval performance of Power in Section 4.2.5.

The descending order of the ten divergence measures according to a pairwise performance comparison is KL, χ^2Neyman, J, ResistorAverage, Hellinger, χ^2Symmetric, JensenShannon, χ^2Pearson, TotalVariation and Skew. TotalVariation is low ranked potentially because it does not have an IDF effect. Skew is low ranked (and to some extent also JensenShannon) potentially due to smoothing the document model with the query model. KL, χ^2Neyman, J and ResistorAverage are presumably among the top 4 as they all have an IDF effect.

4.2.2 Pseudo-Feedback-Based Query Models. Table 5 presents the results of using RM3 and MM as the query language models. (The initial ranking was attained, as described in Section 4.1, using an unsmoothed MLE induced from the title queries.) These query language models are much richer than the unsmoothed MLE query models explored in Section 4.2.1 ; i.e., their support is much larger. Two measures that stand out are the J divergence and Power. The former outperforms KL in 8 relevant comparisons for RM3 and in 13 for MM, while the latter outperforms KL in 12 relevant comparisons for RM3 and 10 for MM; some of these improvements are statistically significant. Both measures are statistically significantly outperformed by KL in at most two cases for RM3 and never for MM. The RI of both measures is always positive, whereas for the KL divergence it is positive in all but a single case: MM for CW09B.

Apart from Geo (KL), which is promoted to the second position after Power, the performance order of the weighted means remains the same as in Section 4.2.1. The top 4 divergence measures in descending order of pairwise comparisons for RM3 are J, KL, ResistorAverage and χ^2Neyman. For MM, χ^2Neyman and ResistorAverage switch places. These measures were also the top 4 in Section 4.2.1. The bottom 4 measures for both RM3 and MM are χ^2Pearson, JensenShannon, Skew and TotalVariation. These measures also appeared at the bottom of the list in Section 4.2.1.

4.2.3 Verbose Queries. To further study the effect of the query-model support size on retrieval performance, we repeated the experiment from Section 4.2.1 but now using verbose queries composed of the topic's title and description. As a result, the support of the query model is (much) larger than that used in Section 4.2.1 where MLE of short title queries is used. The results are presented in Table 6. As was the case thus far, the two measures outperforming the KL divergence are the J divergence and Power. However, unlike the case in Sections 4.2.1 and 4.2.2, both measures are outperformed by KL in at least the same number of relevant comparisons in which they outperform it. Most performance differences between these measures and KL are not statistically significant. The performance-based ordering of the weighted means is as in Section 4.2.1, except that now Geo (KL) is promoted to the first rank. The top 4 divergence measures are KL, J, ResistorAverage and Hellinger, while the bottom 4 are JensenShannon, Skew, χ^2Pearson and TotalVariation as was the case in Sections 4.2.1 and 4.2.2.

4.2.4 Power vs. J divergence. We saw that Power is the most effective weighted mean while J is the most effective divergence measure among the alternatives we considered. Table 7 contrasts their performance for: MLE query models induced from short (title) or verbose (title+description) queries, and pseudo-feedback-based query models (RM3 or MM). Evidently, Power is more effective than J for short title queries and for RM3, while the reverse holds for verbose queries and MM.

4.2.5 Further Analysis. We showed that Power often outperforms KL for query models induced from short titles using unsmoothed MLE and those induced using pseudo feedback. In Figure 1 we study the effect of β on the MAP performance of Power. Short title queries and the MLE query model are used. We note that similar trends were also observed for RM3, MM and verbose queries. These results are omitted due to space considerations and as they convey no further insight. We see that the performance of Power is the highest for values of β close to 0. (Recall from Section 3.2 that Power converges to Geo as β approaches zero and that it has an IDF effect for $\beta < 1$.) A case in point, we found that

Table 5: Using RM3 and MM to induce query models. 'Init': initial ranking. 'i' and 'k' mark statistically significant differences of retrieval effectiveness (MAP, p@5 and NDCG20) with Init and KL divergence, respectively. The best result in a column per query model is highlighted.

		AP				ROBUST				GOV2				CW09B				CW09BF			
		MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI
	Init	21.1	43.6	42.6	–	25.4	48.7	43.9	–	29.2	55.5	44.5	–	17.9	22.7	20.1	–	18.7	34.7	25.8	–
RM3	KL	27.1^{i}	48.7^{i}	46.5^{i}	43.4	28.8^{i}	48.8	44.6	23.7	32.8^{i}	58.0	47.3^{i}	43.9	19.0^{i}	26.1^{i}	22.5^{i}	11.6	19.9^{i}	37.2	27.1^{i}	13.6
	Ari	11.0$^{i}_{k}$	27.5$^{i}_{k}$	26.1$^{i}_{k}$	−64.6	6.4$^{i}_{k}$	14.6$^{i}_{k}$	13.1$^{i}_{k}$	−84.3	8.2$^{i}_{k}$	15.7$^{i}_{k}$	13.7$^{i}_{k}$	−91.9	6.6$^{i}_{k}$	5.7$^{i}_{k}$	6.1$^{i}_{k}$	−63.1	7.9$^{i}_{k}$	15.3$^{i}_{k}$	11.3$^{i}_{k}$	−63.1
	Har	17.8$^{i}_{k}$	40.8$_{k}$	40.0$_{k}$	−41.4	22.3$^{i}_{k}$	47.2	40.6$^{i}_{k}$	−37.3	23.8$^{i}_{k}$	53.5$_{k}$	42.3$_{k}$	−47.3	14.0$^{i}_{k}$	28.0^{i}	19.8$_{k}$	−28.8	14.9$^{i}_{k}$	32.2$_{k}$	22.3$^{i}_{k}$	−43.4
	GeoAri	24.1$^{i}_{k}$	45.1$_{k}$	45.0$_{k}$	20.2	24.0$^{i}_{k}$	45.1$^{i}_{k}$	41.1$^{i}_{k}$	−24.9	25.7$^{i}_{k}$	50.3$^{i}_{k}$	40.6$^{i}_{k}$	−35.1	12.3$^{i}_{k}$	14.4$^{i}_{k}$	13.6$^{i}_{k}$	−48.5	14.8$^{i}_{k}$	30.8$^{i}_{k}$	20.5$^{i}_{k}$	−42.4
	GeoHar	18.7$^{i}_{k}$	41.0$_{k}$	41.0$_{k}$	−26.3	23.9$^{i}_{k}$	47.6	42.3$^{i}_{k}$	−28.1	26.0$^{i}_{k}$	54.1$_{k}$	43.6$_{k}$	−32.4	15.7$^{i}_{k}$	29.0^{i}	20.9	−21.2	16.7$^{i}_{k}$	34.8	24.2$_{k}$	−25.3
	Power	26.8$^{i}_{k}$	48.9^{i}	46.5^{i}	42.4	28.9^{i}	49.3	44.7	24.1	33.2$^{i}_{k}$	58.8	47.5^{i}	39.9	19.4^{i}	28.3$^{i}_{k}$	23.3$^{i}_{k}$	21.2	19.9^{i}	37.4	27.2^{i}	6.6
	Lehmer	24.4$^{i}_{k}$	46.5	46.0^{i}	32.3	26.4$^{i}_{k}$	49.0	43.9	2.4	30.2$_{k}$	57.2	46.0	10.1	17.9$_{k}$	22.9$_{k}$	20.8$_{k}$	−5.1	19.0$_{k}$	35.8	26.1$_{k}$	−3.0
	Hellinger	25.1$^{i}_{k}$	47.1	46.0^{i}	22.2	22.8$^{i}_{k}$	46.1$^{i}_{k}$	40.8$^{i}_{k}$	−24.9	23.8$^{i}_{k}$	54.1$_{k}$	42.0$_{k}$	−39.2	15.5$^{i}_{k}$	27.0^{i}	20.1$_{k}$	−35.9	16.7$^{i}_{k}$	36.2	24.5$_{k}$	−28.8
	TotalVariation	13.8$^{i}_{k}$	43.2$_{k}$	36.6$^{i}_{k}$	−60.6	11.3$^{i}_{k}$	33.8$^{i}_{k}$	26.3$^{i}_{k}$	−82.7	11.7$^{i}_{k}$	41.1$^{i}_{k}$	30.3$^{i}_{k}$	−86.5	7.7$^{i}_{k}$	19.2$_{k}$	13.2$^{i}_{k}$	−78.8	8.8$^{i}_{k}$	26.3$^{i}_{k}$	16.7$^{i}_{k}$	−76.3
	JensenShannon	18.3$^{i}_{k}$	43.6	40.2$_{k}$	−34.3	13.9$^{i}_{k}$	36.5$^{i}_{k}$	30.6$^{i}_{k}$	−75.9	12.7$^{i}_{k}$	41.8$^{i}_{k}$	31.0$^{i}_{k}$	−85.1	9.8$^{i}_{k}$	23.5	16.1$^{i}_{k}$	−71.7	10.1$^{i}_{k}$	26.8$^{i}_{k}$	17.6$^{i}_{k}$	−69.7
	J	26.3$^{i}_{k}$	47.7	46.5^{i}	42.4	28.1$^{i}_{k}$	50.4$_{k}$	45.2^{i}	29.7	33.2$^{i}_{k}$	60.7$^{i}_{k}$	48.6$^{i}_{k}$	40.5	18.7	27.0^{i}	21.8^{i}	13.6	19.9^{i}	39.9$^{i}_{k}$	27.4^{i}	8.1
	ResistorAverage	23.5$^{i}_{k}$	47.5	45.1	10.1	22.0$^{i}_{k}$	44.8$^{i}_{k}$	40.0$^{i}_{k}$	−41.4	26.6$^{i}_{k}$	58.0	44.5$_{k}$	−29.7	18.5	28.3^{i}	21.9^{i}	9.1	19.1	40.0^{i}	26.7	6.1
	χ^2Neyman	19.3$^{i}_{k}$	41.8$_{k}$	40.9$_{k}$	−16.2	23.5$^{i}_{k}$	47.6	41.8$^{i}_{k}$	−30.5	26.0$^{i}_{k}$	56.1	44.1$_{k}$	−35.1	18.7	29.0^{i}	23.2^{i}	13.6	18.6$_{k}$	35.5	26.4	−7.6
	χ^2Pearson	19.4$_{k}$	44.4$_{k}$	42.1$_{k}$	−14.1	15.9$^{i}_{k}$	39.4$^{i}_{k}$	33.2$^{i}_{k}$	−64.3	15.5$^{i}_{k}$	46.2$^{i}_{k}$	34.3$^{i}_{k}$	−74.3	9.7$^{i}_{k}$	21.9$_{k}$	15.7$^{i}_{k}$	−72.7	11.1$^{i}_{k}$	30.6$_{k}$	19.6$^{i}_{k}$	−69.7
	χ^2Symmetric	21.2$_{k}$	45.5	43.5$_{k}$	0.0	16.8$^{i}_{k}$	40.6$^{i}_{k}$	33.9$^{i}_{k}$	−60.6	17.4$^{i}_{k}$	47.4$^{i}_{k}$	35.4$^{i}_{k}$	−67.6	11.7$^{i}_{k}$	23.6	17.5$^{i}_{k}$	−58.1	12.8$^{i}_{k}$	32.9$_{k}$	21.2$^{i}_{k}$	−60.6
	Skew	17.4$^{i}_{k}$	42.4$_{k}$	39.3$_{k}$	−42.4	14.1$^{i}_{k}$	35.3$^{i}_{k}$	30.0$^{i}_{k}$	−75.5	12.3$^{i}_{k}$	41.1$^{i}_{k}$	29.8$^{i}_{k}$	−87.8	9.1$^{i}_{k}$	21.4$_{k}$	14.3$^{i}_{k}$	−76.3	9.7$^{i}_{k}$	25.2$^{i}_{k}$	16.8$^{i}_{k}$	−72.2
MM	KL	27.8^{i}	50.1^{i}	46.9^{i}	22.2	27.1^{i}	47.5	43.2	24.1	32.1^{i}	58.0	45.4	31.1	18.9	24.8	21.8	−5.1	20.4^{i}	38.7^{i}	27.4	10.6
	Ari	18.7$_{k}$	38.0$_{k}$	36.6$^{i}_{k}$	−24.2	10.8$^{i}_{k}$	22.0$^{i}_{k}$	19.9$^{i}_{k}$	−68.7	10.7$^{i}_{k}$	21.1$^{i}_{k}$	16.5$^{i}_{k}$	−89.2	6.8$^{i}_{k}$	5.4$^{i}_{k}$	6.2$^{i}_{k}$	−65.7	8.0$^{i}_{k}$	14.7$^{i}_{k}$	11.9$^{i}_{k}$	−76.3
	Har	19.4$_{k}$	44.8$_{k}$	41.0$_{k}$	−23.2	22.5$^{i}_{k}$	46.4^{i}	40.5$^{i}_{k}$	−32.5	24.9$^{i}_{k}$	54.7	42.1$_{k}$	−44.6	13.5$^{i}_{k}$	26.0	18.1$_{k}$	−42.9	16.0$^{i}_{k}$	34.4$_{k}$	24.0$_{k}$	−33.3
	GeoAri	25.2$^{i}_{k}$	49.1^{i}	45.5	16.2	22.8$^{i}_{k}$	42.7$^{i}_{k}$	39.3$^{i}_{k}$	−28.5	24.8$^{i}_{k}$	47.4$^{i}_{k}$	39.0$^{i}_{k}$	−33.8	12.4$^{i}_{k}$	13.0$^{i}_{k}$	13.7$^{i}_{k}$	−53.5	15.4$^{i}_{k}$	31.7$_{k}$	21.9$^{i}_{k}$	−36.9
	GeoHar	22.1$_{k}$	43.8$_{k}$	41.4$_{k}$	−11.1	23.6$^{i}_{k}$	47.2	41.8^{i}	−30.1	26.9$^{i}_{k}$	56.6	43.7	−20.9	15.6$^{i}_{k}$	27.3^{i}	19.9	−27.3	17.6$_{k}$	36.6	25.4$_{k}$	−18.7
	Power	27.9^{i}	49.3^{i}	46.7^{i}	20.2	27.3^{i}	47.6	43.5	24.9	32.1^{i}	58.1	45.6	27.0	20.1$^{i}_{k}$	29.9$^{i}_{k}$	24.1$^{i}_{k}$	27.8	20.3^{i}	38.6^{i}	27.7^{i}	17.2
	Lehmer	27.6^{i}	49.9^{i}	46.9^{i}	22.2	26.0$_{k}$	45.8^{i}	42.5	8.8	30.1$_{k}$	57.3	46.4	21.6	17.8$_{k}$	23.3	20.3$_{k}$	−21.2	19.5$_{k}$	38.8^{i}	26.9	−6.6
	Hellinger	26.4$^{i}_{k}$	47.3	46.4	22.2	20.3$^{i}_{k}$	40.5$^{i}_{k}$	36.2$^{i}_{k}$	−24.1	23.2$^{i}_{k}$	50.7$_{k}$	37.8$^{i}_{k}$	−40.5	15.6$^{i}_{k}$	24.0	19.4$_{k}$	−40.9	17.0$^{i}_{k}$	37.4	24.2$_{k}$	−34.3
	TotalVariation	22.1$_{k}$	45.1$_{k}$	43.2$_{k}$	0.0	14.2$^{i}_{k}$	33.3$^{i}_{k}$	28.1$^{i}_{k}$	−54.2	15.2$^{i}_{k}$	38.4$^{i}_{k}$	28.0$^{i}_{k}$	−79.7	7.5$^{i}_{k}$	12.0$^{i}_{k}$	11.0$^{i}_{k}$	−78.8	10.0$^{i}_{k}$	24.5$^{i}_{k}$	17.3$^{i}_{k}$	−69.2
	JensenShannon	24.5$^{i}_{k}$	46.7	44.5	14.1	16.9$^{i}_{k}$	34.7$^{i}_{k}$	31.3$^{i}_{k}$	−47.8	13.2$^{i}_{k}$	32.4$^{i}_{k}$	25.0$^{i}_{k}$	−86.5	10.2$^{i}_{k}$	21.4	14.9$^{i}_{k}$	−74.7	10.7$^{i}_{k}$	25.6$^{i}_{k}$	16.8$^{i}_{k}$	−72.2
	J	28.2$^{i}_{k}$	49.7^{i}	47.3^{i}	21.2	27.3^{i}	48.5$_{k}$	43.8$_{k}$	24.9	32.2^{i}	58.2	46.5$_{k}$	39.2	19.2^{i}	26.1^{i}	21.8^{i}	12.6	20.8^{i}	41.8$^{i}_{k}$	29.0$^{i}_{k}$	27.8
	ResistorAverage	25.1$^{i}_{k}$	47.3	45.3	8.1	20.7$^{i}_{k}$	42.7$^{i}_{k}$	38.7$^{i}_{k}$	−32.5	26.9$^{i}_{k}$	54.7	42.5$_{k}$	−20.3	18.4	23.3	21.0	8.1	19.6	38.8	26.6	−1.0
	χ^2Neyman	21.6$_{k}$	46.5	43.4$_{k}$	−6.1	22.8$^{i}_{k}$	44.7$^{i}_{k}$	40.3$^{i}_{k}$	−31.3	26.3$^{i}_{k}$	56.2	44.3	−27.0	18.5	29.5$^{i}_{k}$	23.0^{i}	4.0	18.7$_{k}$	36.9	26.1	−8.1
	χ^2Pearson	22.9$_{k}$	46.3$_{k}$	43.5$_{k}$	4.0	14.6$^{i}_{k}$	33.7$^{i}_{k}$	28.5$^{i}_{k}$	−51.0	15.8$^{i}_{k}$	40.7$^{i}_{k}$	29.2$^{i}_{k}$	−71.6	10.5$^{i}_{k}$	22.9	16.7$^{i}_{k}$	−64.6	11.6$^{i}_{k}$	33.4$_{k}$	20.6$^{i}_{k}$	−58.1
	χ^2Symmetric	24.0$_{k}$	47.5	45.0	8.1	15.2$^{i}_{k}$	33.6$^{i}_{k}$	29.1$^{i}_{k}$	−45.4	16.6$^{i}_{k}$	43.0$^{i}_{k}$	31.1$^{i}_{k}$	−70.3	12.3$^{i}_{k}$	23.5	17.8$_{k}$	−55.6	13.1$^{i}_{k}$	34.4$_{k}$	21.7$^{i}_{k}$	−48.0
	Skew	23.8$_{k}$	46.7	43.7$_{k}$	14.1	16.8$^{i}_{k}$	32.6$^{i}_{k}$	29.1$^{i}_{k}$	−47.8	14.2$^{i}_{k}$	33.5$^{i}_{k}$	25.6$^{i}_{k}$	−86.5	9.5$^{i}_{k}$	18.2$^{i}_{k}$	13.3$^{i}_{k}$	−78.8	11.0$^{i}_{k}$	26.0$^{i}_{k}$	17.3$^{i}_{k}$	−73.7

Table 6: Verbose (title+description) queries. Using unsmoothed MLE to induce query models. Because the KL divergence is used to produce the initial ranking, its RI is undefined. 'Init': initial ranking. 'k': statistically significant difference of retrieval effectiveness (MAP, p@5 and NDCG20) with the KL divergence. The best result in a column is highlighted.

	AP				ROBUST				GOV2				CW09B				CW09BF			
	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI	MAP	p@5	NDCG20	RI
KL (=Init)	23.8	49.1	46.0	–	28.4	54.8	47.3	–	30.0	61.4	47.1	–	18.1	29.2	24.2	–	18.1	36.1	26.6	–
Ari	9.8$_{k}$	29.3$_{k}$	26.6$_{k}$	−77.8	5.6$_{k}$	13.3$_{k}$	12.5$_{k}$	−93.6	7.2$_{k}$	10.9$_{k}$	9.8$_{k}$	−91.9	4.8$_{k}$	3.1$_{k}$	4.2$_{k}$	−63.1	6.2$_{k}$	12.0$_{k}$	8.7$_{k}$	−50.0
Har	10.2$_{k}$	25.5$_{k}$	24.2$_{k}$	−87.9	17.6$_{k}$	36.9$_{k}$	32.0$_{k}$	−82.7	15.5$_{k}$	45.5$_{k}$	31.9$_{k}$	−78.4	9.2$_{k}$	19.3$_{k}$	14.4$_{k}$	−66.7	8.8$_{k}$	20.4$_{k}$	14.1$_{k}$	−74.2
GeoAri	17.3$_{k}$	43.0$_{k}$	40.4$_{k}$	−57.6	14.8$_{k}$	35.9$_{k}$	29.7$_{k}$	−82.3	16.5$_{k}$	33.5$_{k}$	27.3$_{k}$	−83.8	8.6$_{k}$	7.3$_{k}$	8.6$_{k}$	−51.0	12.0$_{k}$	25.2$_{k}$	17.7$_{k}$	−30.8
GeoHar	15.3$_{k}$	33.5$_{k}$	32.2$_{k}$	−75.8	22.1$_{k}$	42.2$_{k}$	37.9$_{k}$	−69.9	20.0$_{k}$	51.1$_{k}$	36.7$_{k}$	−71.6	12.0$_{k}$	24.1$_{k}$	17.9$_{k}$	−59.6	11.4$_{k}$	25.1$_{k}$	17.9$_{k}$	−66.2
Power	23.8	48.7	46.7$_{k}$	9.1	28.0$_{k}$	53.9	47.2	−6.0	29.9	60.1	47.0	11.5	17.2$_{k}$	29.1	23.3$_{k}$	−35.4	17.8	37.0	26.5	20.7
Lehmer	23.2$_{k}$	48.1	45.4	−13.1	27.0$_{k}$	51.4$_{k}$	45.2$_{k}$	−34.1	27.3$_{k}$	57.8$_{k}$	44.6$_{k}$	−49.3	15.1$_{k}$	24.2$_{k}$	20.5$_{k}$	−51.0	15.8$_{k}$	32.5$_{k}$	24.0$_{k}$	−44.4
Hellinger	18.1$_{k}$	45.1	42.5$_{k}$	−43.4	16.6$_{k}$	39.4$_{k}$	32.7$_{k}$	−71.1	16.7$_{k}$	33.9$_{k}$	28.4$_{k}$	−85.1	11.5$_{k}$	18.7$_{k}$	15.0$_{k}$	−42.9	13.1$_{k}$	30.6$_{k}$	20.8$_{k}$	−30.8
TotalVariation	7.9$_{k}$	25.1$_{k}$	23.3$_{k}$	−73.7	5.0$_{k}$	12.9$_{k}$	11.5$_{k}$	−92.8	6.5$_{k}$	10.4$_{k}$	8.7$_{k}$	−95.9	3.6$_{k}$	3.0$_{k}$	3.3$_{k}$	−70.2	4.5$_{k}$	9.3$_{k}$	6.5$_{k}$	−67.2
JensenShannon	9.7$_{k}$	34.5$_{k}$	30.2$_{k}$	−83.8	9.0$_{k}$	26.9$_{k}$	22.0$_{k}$	−88.0	4.5$_{k}$	13.2$_{k}$	10.4$_{k}$	−97.3	4.3$_{k}$	12.1$_{k}$	8.5$_{k}$	−78.3	3.8$_{k}$	13.7$_{k}$	8.9$_{k}$	−75.8
J	23.8	49.7	46.0	−11.1	28.4	54.5	47.3	−3.2	29.8$_{k}$	59.6$_{k}$	47.0	−29.7	18.3$_{k}$	30.4	24.3	10.1	18.1	36.3	26.4	−3.5
ResistorAverage	12.7$_{k}$	42.0$_{k}$	36.2$_{k}$	−83.8	13.3$_{k}$	35.8$_{k}$	30.1$_{k}$	−86.7	19.9$_{k}$	53.5$_{k}$	39.6$_{k}$	−68.9	15.8$_{k}$	26.8	22.3	−10.1	16.3$_{k}$	36.1	25.5	−10.6
χ^2Neyman	12.1$_{k}$	30.9$_{k}$	28.0$_{k}$	−85.9	19.1$_{k}$	39.2$_{k}$	34.7$_{k}$	−78.7	17.8$_{k}$	49.3$_{k}$	35.2$_{k}$	−75.7	10.9$_{k}$	22.2$_{k}$	16.7$_{k}$	−60.6	10.4$_{k}$	23.4$_{k}$	16.9$_{k}$	−68.2
χ^2Pearson	8.2$_{k}$	26.9$_{k}$	24.7$_{k}$	−74.7	5.3$_{k}$	14.1$_{k}$	12.4$_{k}$	−92.0	6.8$_{k}$	10.8$_{k}$	9.9$_{k}$	−95.3	4.0$_{k}$	4.8$_{k}$	4.2$_{k}$	−67.2	5.1$_{k}$	10.8$_{k}$	8.1$_{k}$	−62.1
χ^2Symmetric	8.8$_{k}$	29.5$_{k}$	26.4$_{k}$	−75.8	5.8$_{k}$	16.1$_{k}$	13.6$_{k}$	−92.8	7.3$_{k}$	12.7$_{k}$	11.5$_{k}$	−94.6	4.7$_{k}$	6.8$_{k}$	5.5$_{k}$	−64.1	5.9$_{k}$	13.9$_{k}$	9.9$_{k}$	−58.1
Skew	8.3$_{k}$	29.3$_{k}$	24.6$_{k}$	−85.9	6.9$_{k}$	20.0$_{k}$	16.7$_{k}$	−88.8	3.4$_{k}$	8.9$_{k}$	6.4$_{k}$	−97.3	4.0$_{k}$	10.5$_{k}$	7.5$_{k}$	−77.3	3.6$_{k}$	11.6$_{k}$	8.0$_{k}$	−74.7

for $\beta = -0.05$ Power outperforms KL in terms of MAP for all five datasets. The performance for $\beta = -1$ (which amounts to Har), $\beta = 1$ (Ari) and especially for $\beta = 2$ (weighted quadratic mean) and $\beta = 3$ (weighted cubic mean) is often much lower than that attained for $\beta = -0.05$. The highest improvement over KL is attained for CW09B, the noisiest Web collection among those considered.

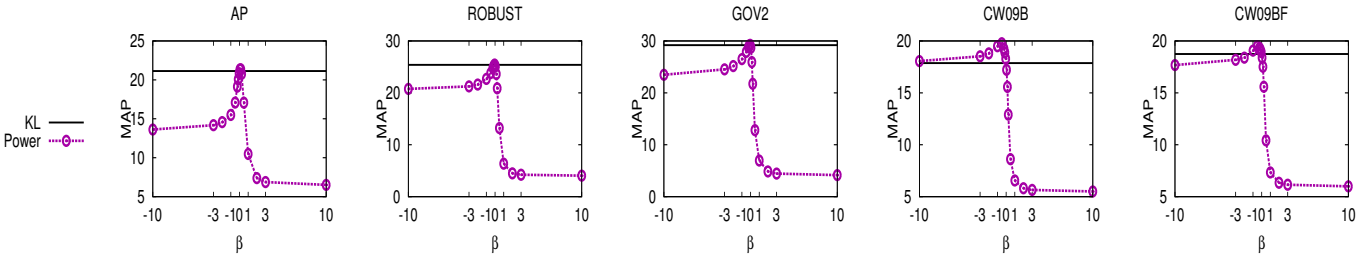

Figure 1: The effect of β on the MAP performance of Power. Short title queries and the MLE query model are used.

Table 7: Power vs. J divergence. The number of relevant comparisons out of 15 (5 datasets × 3 evaluation measures: MAP, p@5 and NDCG20) in which the method in the row (statistically significantly) outperforms the method in the column.

	Short Query MLE			RM3			MM			Verbose Query MLE		
	KL	Power	J	KL	Power	J	KL	Power	J	KL	Power	J
KL	0 (0)	1 (0)	7 (4)	0 (0)	1 (1)	5 (2)	0 (0)	4 (0)	1 (0)	0 (0)	12 (3)	5 (2)
Power	12 (5)	0 (0)	10 (6)	12 (3)	0 (0)	6 (4)	10 (3)	0 (0)	3 (3)	2 (1)	0 (0)	5 (1)
J	7 (3)	5 (1)	0 (0)	8 (5)	6 (2)	0 (0)	13 (6)	11 (4)	0 (0)	5 (1)	8 (3)	0 (0)

5 CONCLUSIONS

Motivated by the fact that comparing query and document language models using the KL divergence is rank equivalent to using a specific weighted geometric mean, we studied alternative weighted means as well as divergence measures; specifically, we analyzed the inverse document frequency (IDF) effect of the methods. Empirical evaluation showed that KL can be often outperformed in several settings by some alternatives.

Acknowledgments We thank the reviewers for their helpful comments. This paper is based upon work supported in part by the Israel Science Foundation under grant no. 433/12.

REFERENCES

[1] Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Marck D., and Courtney Wade. 2004. UMASS at TREC 2004 — Novelty and HARD. In *Proc. of TREC-13*.

[2] S. M. Ali and S. D. Silvey. 1966. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)* 28, 1 (1966), 131–142.

[3] Javed A. Aslam and Virgiliu Pavlu. 2007. Query Hardness Estimation Using Jensen-Shannon Divergence Among Multiple Scoring Functions. In *Proc. of ECIR*. 198–209.

[4] Gleb Beliakov, Humberto Bustince Sola, and Tomasa Calvo. 2016. *A Practical Guide to Averaging Functions*. Studies in Fuzziness and Soft Computing, Vol. 329. Springer.

[5] Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proc. of CIKM*. 211–218.

[6] B.S. Bullen. 2003. *Handbook of means and their inequalities*. Springer-Science+Business Media, B.V.

[7] Francine Chen, Ayman Farahat, and Thorsten Brants. 2004. Multiple Similarity Measures and Source-Pair Information in Story Link Detection. In *Proc. of HLT-NAACL*. 313–320.

[8] Ruey-Cheng Chen, Chia-Jung Lee, and W. Bruce Croft. 2015. On Divergence Measures and Static Index Pruning. In *Proc. of ICTIR*. 151–160.

[9] Stéphane Clinchant and Éric Gaussier. 2013. A Theoretical Analysis of Pseudo-Relevance Feedback Models. In *Proc. of ICTIR*. 6.

[10] Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval* 14, 5 (2011), 441–465.

[11] Imre Csiszár. 1967. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* 2 (1967), 299–318.

[12] Fernando Diaz. 2015. Condensed List Relevance Models. In *Proc. of ICTIR*. 313–316.

[13] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *Proc. of SIGIR*. 49–56.

[14] Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *Proc. of SIGIR*. 480–487.

[15] Hussein Hazimeh and ChengXiang Zhai. 2015. Axiomatic Analysis of Smoothing Methods in Language Models for Pseudo-Relevance Feedback. In *Proc. of ICTIR*. 141–150.

[16] Harold Jeffreys. 1939. *Theory of Probability*. Oxford University Press.

[17] Don Johnson and Sinan Sinanovic. 2001. Symmetrizing the Kullback-Leibler Distance. *IEEE Transactions on Information Theory* (2001).

[18] Maryam Karimzadehgan and ChengXiang Zhai. 2012. Axiomatic Analysis of Translation Language Model for Information Retrieval. In *Proc. of ECIR*. 268–280.

[19] Oren Kurland and Lillian Lee. 2010. PageRank without hyperlinks: Structural reranking using links induced by language models. *ACM Transactions on information systems* 28, 4 (2010), 18.

[20] John D. Lafferty and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proc. of SIGIR*. 111–119.

[21] Victor Lavrenko and W. Bruce Croft. Relevance Models in Information Retrieval. In *Language Modeling for Information Retrieval*. Number 13 in Information Retrieval Book Series. Kluwer, Chapter 2, 11–56.

[22] Lillian Lee. 1999. Measures of Distributional Similarity. In *Proc. of ACL*.

[23] Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Proc. of AISTATS*. 65–72.

[24] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Information Theory* 37, 1 (1991), 145–151.

[25] Ramesh Nallapati. 2006. *The smoothed dirichlet distribution: Understanding cross-entropy ranking in information retrieval*. Ph.D. Dissertation. University of Massachusetts.

[26] Jerzy Neyman. 1949. Contribution to the theory of the χ^2 test. In *Proc. of the Berkeley symposium on mathematical statistics and probability*, Vol. 1. University of California Press Berkeley, 239–273.

[27] Karl Pearson. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5* 50, 302 (1900), 157–175.

[28] Tetsuya Sakai, Toshihiko Manabe, and Makoto Koyama. 2005. Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing* 4, 2 (2005), 111–135.

[29] Fei Song and W. Bruce Croft. 1999. A general language model for information retrieval. In *Proc. of SIGIR*. 279–280.

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

[31] Chengxiang Zhai and John D. Lafferty. 2001. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *Proc. of CIKM*. 403–410.

[32] Chengxiang Zhai and John D. Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proc. of SIGIR*. 334–342.

[33] Justin Zobel and Alistair Moffat. 1998. Exploring the similarity space. *ACM SIGIR forum* 18, 1 (1998), 18–34.