On Divergence Measures and Static Index Pruning

Ruey-Cheng Chen[†*] Chia-Jung Lee W. Bruce Croft

[†]RMIT University, GPO Box 2476, Melbourne VIC 3001, Australia
University of Massachusetts, 140 Governors Drive, Amherst, MA 01003-9264
ruey-cheng.chen@rmit.edu.au, {cjlee, croft}@cs.umass.edu

ABSTRACT

We study the problem of static index pruning in a renowned divergence minimization framework, using a range of divergence measures such as f-divergence and Rényi divergence as the objective. We show that many well-known divergence measures are convex in pruning decisions, and therefore can be exactly minimized using an efficient algorithm. Our approach allows postings be prioritized according to the amount of information they contribute to the index, and through specifying a different divergence measure the contribution is modeled on a different returns curve. In our experiment on GOV2 data, Rényi divergence of order infinity appears the most effective. This divergence measure significantly outperforms many standard methods and achieves identical retrieval effectiveness as full data using only 50% of the postings. When top-k precision is of the only concern, 10% of the data is sufficient to achieve the accuracy that one would usually expect from a full index.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Static index pruning; f-divergence; Rényi divergence

1. INTRODUCTION

The study on inducing succinct data representation has a long history in the domain of information retrieval. Inspired by the need of offering search on handheld devices with limited storage space, Carmel et al. [9] motivated a technique, called *static index pruning*, that aims at creating a condensed version of an inverted index such that there is little difference from the user perspective in the top k returned results. Since an inverted index is essentially a warehouse of links between terms and documents, it makes sense to see this as a problem of choosing the most useful links, i.e., signals or features that best distinguish a relevant document from non-relevant ones, allowing the creation of an effective condensed or summarized index. Many early efforts took this perspective and have succeeded in applying relevance measures and heuristics common in information retrieval to this problem [4, 7, 9]. Some later developments used a log-based approach, exploiting session logs to uncover regularity in user queries and locate information that is likely to be reused [2, 25]. Due to this success, the notion of static index pruning as a *cache problem* has become prevalent. Although some progress has been made in exploring more sophisticated abstractions [5, 26], a dedicated theory for this task has not been described.

The problem of static index pruning appeals to theoreticians for it being a related process to relevance estimation. Instead of asking how one estimates the degree of relevance in a posting, it asks if the postings can be ordered in a way such that the least informative part can be thrown away as necessary. This perspective is valuable since it outlines an important task that people recently have started to look at, which is to create highly succinct summaries about large data in an unsupervised fashion.

In a recent paper by Chen and Lee [10], it has been shown that static index pruning has a connection to probabilistic inference, a well-motivated problem solved usually by minimizing relative entropy [17]. Their work outlined a mathematical foundation for static index pruning, but it also came with some caveats that makes exact inference difficult and hinders further implications. Several interesting research questions were left unsettled due to this difficulty. For instance, can the optimization framework generalize over a broader family of divergences, such as f-divergence or the famous Rényi divergence? Can the analysis be extended to model multiple-term queries? Does removing term posting lists entirely from the index lead to decreased performance? None of these questions can be easily answered without us first being able to develop new mathematical tools.

In this paper, we introduce a new mathematical analysis to address these issues. We built on top of the existing divergence minimization framework for static index pruning, but used a different way to describe the optimization problem. We show that both f-divergence and Rényi divergence can be exactly minimized under this formulation, and the exact solution can be efficiently computed in time complexity $O(|\mathcal{D}|L \log L)$, where $|\mathcal{D}|$ is the total number of documents and L the maximum document length in the index. The

*Part of the work was done at National Taiwan University.

ICTIR'15, September 27–30, Northampton, MA, USA.

DOI: http://dx.doi.org/10.1145/2808194.2809472.

same analysis can also be extended to model multiple-term queries. We show that, with suitable assumptions (bag-of-words), the entire Rényi divergence family can be exactly solved for up to infinite terms. This problem is readily solvable even in the presence of billions of variables, allowing us to avoid a numerical solution impractical for a problem of this scale. We also seek comparison with other strong results in the application domain. Our empirical results on the GOV2 data shows that Rényi divergence helps preserving top-k precision and document ranking. Our approach outperforms many standard methods of the task and compares well to the strong baseline in top-k precision.

The rest of paper is structured as follows. Section 2 covers some backgrounds of static index pruning. In Section 3, we develop a convex analysis that has eventually led to an exact algorithm for minimizing all the divergence measures mentioned in this work. Section 4 covers the experimental results. Our findings are discussed in Section 5 and we give out concluding remarks in Section 6.

2. BACKGROUND

Static index pruning first appeared as a theoretical problem in Carmel et al. [9] and later found application in web search [7, 14]. It is a technique that removes less important postings permanently from an index. Initially developed to mitigate efficiency issues caused by operating a large index, static index pruning later has grown into a wide range of studies that focus on reducing the seemingly inevitable performance loss. To date, many successful approaches have relied on posting importance measures, including impact [7, 9] or odds ratio in the probability ranking principle [5]. Some later efforts have based this measurement on more sophisticated techniques, such as statistical hypothesis testing [26], query-view methods [2], or information theory [10, 11].

Our approach most closely resembles the work of Chen and Lee [10] in the way the pruning problem is framed. While we use relative entropy minimization (i.e., Kullback's principle of minimum cross-entropy) much the same way as in the previous work to infer truncated models, we rely on a different generative process that eventually leads to document-centric pruning strategies. This departure helps to avoid many modeling issues previously associated with term-centric formulations. Our pruning problems are sufficiently simple and can be solved exactly using analytic techniques alone. Our approach is free of approximations such as surrogates or assumptions about uniform priors and probability renormalization, and can be easily extended to model multiple-term queries. These improvements in modeling make our results less restrictive and more practical.

3. DIVERGENCE-BASED METHOD

We start this analysis by treating inverted indexes in the context of language modeling [23]. We will view an inverted index as a joint probability measure over query terms $Q = \langle T_1, T_2, \ldots, T_n \rangle$ and document D. By saying an index is a joint probability measure, we mean that an index has the ability to produce a probability value $p(Q, D)$ for any given pair of Q and D that allows for document ranking. Here, n is said to be the *cardinality* of query.

A typical language modeling approach would suggest the following generative structure: One first chooses a document D and then makes n independent draws $T_1, T_2, \ldots, T_n$ from the discrete distribution θ_D that represents the language model for document D.

$$D \sim \text{Uniform}(1, |\mathcal{D}|),$$
$$T_k \sim \text{Discrete}(\theta_D) \quad \text{for } k = 1 \ldots n.$$

An inverted index represented this way can also be seen as a mixture of document language models. It is then straightforward to rank documents based on the joint likelihood. More advanced variants also exist that take document priors, term dependency, or proximity into account. See Zhai and Lafferty [28] for a complete treatment.

Ideally, the support of this mixture would cover the set of all possible queries and all documents. But practically this would break down to just all the postings in the index, each of the form (t, d). So if we are asked to induce a concise version of this mixture, it would be reasonable to just find a subset of given size from these postings.

Let us write the original index as a measure p, and say a fraction ρ of the postings needs to be removed from p in order to produce the concise version q. Then we define the domain of q as $\mathcal{Q}(\rho)$, the set of all probability measures to which only part of the original postings (as indicated by fraction $1 - \rho$) are made available. We make the usual generative assumption that a joint model has a document prior and a likelihood component, e.g., $p(t, d) = p(d)p(t|d)$ and $q(t, d) = q(d)q(t|d)$. The document prior $p(d)$ does not need to be uniform, though it has to be unique, i.e., $p(d) = q(d)$.

With this definition, it is straightforward to describe static index pruning as a search problem in the probability space. One standard approach to find the best measure is through relative entropy minimization [17]:

$$\begin{array}{ll} \text{minimize} & \mathrm{D}(q||p) \\ \text{subject to} & \mathbb{I}_{t,d} \in \{0, 1\} \text{ for all } (t, d) \\ & \sum_{t,d} \mathbb{I}_{t,d} = (1 - \rho)N \\ & q \in \mathcal{Q}(\rho) \end{array} \tag{1}$$

with the last line expands into the following:

$$q \in \mathcal{Q}(\rho) \;\Leftrightarrow\; q(t_{1:n}|d) = \frac{p(t_{1:n}|d) \prod_j \mathbb{I}_{t_j,d}}{\sum_{t'_{1:n}} p(t'_{1:n}|d) \prod_j \mathbb{I}_{t'_j,d}}. \tag{2}$$

The objective $\mathrm{D}(q||p)$ is the Kullback-Leibler (KL) divergence from q, the probability measure to be induced, to p, the original. This divergence can be replaced by other divergence measures, as we shall briefly show. Each indicator $\mathbb{I}_{t,d}$ represents a binary choice of whether posting (t, d) will be included in q or not. For any given fraction ρ, exactly $(1 - \rho)N$ such indicators have to be "turned on" (N is the total number of postings.) The binary choices $\langle \mathbb{I}_{t,d} | \forall t, d \rangle$ as a whole should create a *truncated* probability measure q out of the measure p by limiting access to part of the support. Technically, the support still covers all postings, but those not selected to enter the new index, i.e., $\mathbb{I}_{t,d} = 0$, would receive zero probability in measure q.[1] After truncation, the measure q needs to be renormalized as in (2) as its total probability mass may no longer sum to one. The denominator in (2) will hereafter be denoted as Z_d for brevity.

To sum up, comparing this formulation with that of Chen and Lee [10], it is less restrictive and free of approximation. Further exploration on divergence measures and query cardinality is thus made feasible.

[1]We disregard smoothing at the modeling stage as it would make the problem complicated.

3.1 f-Divergence and Rényi Divergence

For many decades, researchers in information theory have sought interesting ways to generalize KL divergence [13, 18, 20, 24]. Many well-known measures, such as Hellinger's distance or variational distance, are found related and can be used in place of ordinary Kullback-Leibler divergence in inference problems such as (1). In this paper, one of the focus will be on applying these results to our problem. For the choice of divergence measures, we will mainly look at two families: f-divergence and Rényi divergence of order α.

The f-divergence is independently rediscovered many times in the past for generalizing KL divergence [13, 20]. In our notation, it is written as:

$$\mathrm{D}_f(q||p) = \sum_{t_{1:n},d} p(t_{1:n},d) f\left(\frac{q(t_{1:n},d)}{p(t_{1:n},d)}\right), \quad (3)$$

where f is a convex function such that $f(1) = 0$. A broad range of divergence measures can be modeled this way via different definitions of f. Some of its special cases, such as χ^2-divergence, Hellinger's distance, and variational distance (or *total variation*), are given as follows [18].

Kullback-Leibler divergence	$f(x) = x \log x$
Variational distance	$f(x) = \lvert 1 - x \rvert$
Hellinger's distance	$f(x) = (\sqrt{x} - 1)^2$
χ^2-divergence	$f(x) = (x-1)^2$

Rényi divergence of order α comes from another independent attempt to generalize the KL divergence. This family is parametrized via a positive real number α. It was introduced in Alfred Rényi's seminal work [24], in the following form:

$$\mathrm{D}_\alpha(q||p) = \frac{1}{\alpha - 1} \log \left(\sum_{t_{1:n},d} q(t_{1:n},d)^\alpha p(t_{1:n},d)^{1-\alpha} \right). \quad (4)$$

Rényi divergence is equivalent to KL divergence when $\alpha \to 1$. Setting $\alpha = 2$ would lead to the logarithm of the χ^2-divergence. It is worth noting that one can actually take α to infinity, and by doing that we will get a special closed form for Rényi divergence of order infinity [27]:

$$\mathrm{D}_\infty(q||p) = \log \sup_{t_{1:n},d} \frac{q(t_{1:n},d)}{p(t_{1:n},d)}. \quad (5)$$

3.2 Analysis

The problem described in (1) is overwhelmingly large because the number of variables can easily exceed billions on any web retrieval system. A problem of this scale is infeasible, so further work is needed to simplify the objective. In this analysis, we basically look at two things: convexity and relations between divergence measures. We first check whether the objective in (1) is convex. The objective can be exactly minimized if it is convex [16], or otherwise efficient inference would not seem feasible. Once the convexity is established, we check if the measure is analytically related to some other measures that we have analyzed. As we shall cover later, some divergence measures are related to themselves in lower cardinality. This interesting property allows us to solve high-cardinality problems using solutions to low-cardinality ones.

Divergence	Analytic Form
$\mathrm{KL}^{(1)}$	$-\sum_d p(d) \log\left(\sum_{t'} \mathbb{I}_{t',d} p(t'\lvert d)\right)$
$\mathrm{VD}^{(1)}$	$-\sum_d p(d) \left(\sum_{t'} \mathbb{I}_{t',d} p(t'\lvert d)\right)$
$\mathrm{Hellinger}^{(1)}$	$-\sum_d p(d) \left(\sum_{t'} \mathbb{I}_{t',d} p(t'\lvert d)\right)^{1/2}$
$\chi^2\text{-div}^{(1)}$	$\sum_d p(d) \left(\sum_{t'} \mathbb{I}_{t',d} p(t'\lvert d)\right)^{-1}$
$\mathrm{Rényi}_\alpha^{(1)}$ $(1 < \alpha < \infty)$	$\sum_d p(d) \left(\sum_{t'} \mathbb{I}_{t',d} p(t'\lvert d)\right)^{1-\alpha}$
$\mathrm{Rényi}_\infty^{(1)}$	$\sup_d \left(\sum_{t'} \mathbb{I}_{t',d} p(t'\lvert d)\right)^{-1}$

Table 1: Analytic forms for cardinality $n = 1$.

Convexity for Cardinality $n = 1$. It is known that f-divergence and Rényi divergence are convex in probability measures p and q, but in pruning decisions $\langle \mathbb{I}_{t,d} | \forall t, d \rangle$ the convexity is not yet established.

Let us start with the simplest case where query cardinality n equals 1. In this case, (3), (4), and (5) become:

$$\mathrm{D}_f(q||p) = \sum_{t,d} p(t,d) f\left(\frac{\mathbb{I}_{t,d}}{Z_d}\right),$$

$$\mathrm{D}_\alpha(q||p) = \frac{1}{\alpha - 1} \log \sum_{t,d} p(t,d) \left(\frac{\mathbb{I}_{t,d}}{Z_d}\right)^\alpha, \quad (6)$$

$$\mathrm{D}_\infty(q||p) = \log \sup_{t,d} \frac{\mathbb{I}_{t,d}}{Z_d}.$$

Note that Z_d also falls back to a simpler form: $\sum_{t'} p(t'|d) \mathbb{I}_{t',d}$.

We shall now establish that, with the following two lemmas, that (1) is a convex programming problem under f-divergence and Rényi divergence for $n = 1$. In the first lemma, we will directly prove that f-divergence is jointly convex in pruning decisions, while in the second we will not be able to do so due to the presence of a logarithm function. Instead, we show that minimizing Rényi divergence has an equivalent surrogate that is convex. In other words, we get the same exact solution by minimizing the inside of logarithm.

LEMMA 1 (CONVEXITY). *Given $Z_d > 0$ for all d, $D_f(q||p)$ defined in (6) is jointly convex in pruning decisions $\langle \mathbb{I}_{t,d} | \forall t, d \rangle$ for any convex function f with $f(1) = 0$.*

LEMMA 2 (SURROGATE CONVEXITY). *Given $Z_d > 0$ for all d, minimizing $D_\alpha(q||p)$ in (6) has an equivalent surrogate that is jointly convex in $\langle \mathbb{I}_{t,d} | \forall t, d \rangle$ for $\alpha > 1$.*

Proofs for these two lemmas are given in the appendix. Now, with a bit of algebra, we are able to write out a simplified form for each of these divergence measures. These equations are given in Table 1. Sharp-eyed reader may notice the similarities between these equations. We shall discuss how this can be exploited to develop a general algorithmic solution in a later subsection regarding finding optimal allocation.

Convexity for Cardinality $n > 1$. When query cardinality n is greater than 1, term dependency can make the problem very hard to solve. Generally, when $n > 1$, minimizing (1) under both divergence families leads to a sophisticated geometric programming problem, for which we are not aware of any efficient solution in the billion-variable scale.

This problem can however be alleviated with the term independence ("bag-of-words") assumption, which is to let

Divergence	Analytic Form for $n > 1$
$\text{KL}^{(n)}$	$\text{KL}^{(1)}$
$\text{VD}^{(n)}$	Not convex
$\text{Hellinger}^{(n)}$	$\text{VD}^{(1)}$ for $n = 2$; Not convex otherwise
$\chi^2\text{-div}^{(n)}$	$\text{Rényi}^{(1)}_{n+1}$
$\text{Rényi}^{(n)}_{\alpha}$	$\text{Rényi}^{(1)}_{n\alpha-n+1}$ for $1 < \alpha < \infty$
$\text{Rényi}^{(n)}_{\infty}$	$\sup_d \left(\sum_{t'} \mathbb{I}_{t',d} p(t'\|d)\right)^{-n}$

Table 2: Relations between divergences of different cardinalities. KL divergence and Rényi divergence can both be solved for arbitrarily high cardinalities.

$p(t_{1:n}|d) = \prod_j p(t_j|d)$. We repeated the convex analysis as in $n = 1$ under this assumption, and found each measure for $n > 1$ fits into one of the following classes: (i) The measure is convex and can reduce to itself or other measures in cardinality 1, with examples including Kullback-Leibler divergence, Rényi divergence, and χ^2-divergence. Divergence measures in this class can be solved for arbitrary cardinality and therefore provide the greatest flexibility in modeling user queries; (ii) The measure is not convex on high cardinality, e.g., both variational distance for $n \geq 2$ and Hellinger's distance for $n > 2$ are not convex. In this case, the exact solution cannot be efficiently computed.

More detailed analyses are given in the appendix. Our full result is summarized in Table 2. Among all these measures, we find that Kullback-Leibler divergence, Rényi divergence, and χ^2-divergence the most interesting since they can be solved for arbitrary cardinality and therefore can be used to model arbitrary-length queries properly. Variational distance and Hellinger's distance fall short on this flexibility. In later section, we shall see how this theoretical limit is reflected on practical performance.

3.3 Optimal Allocation

Having established the convexity, we now turn to develop algorithmic solutions for the divergence measures. A common pattern that we observed in Table 1 is all these measures (except Rényi divergence of order infinity) are of the following format:

$$\sum_d p(d)\, G\left(\sum_t \mathbb{I}_{t,d} p(t|d)\right), \tag{7}$$

where $G(x)$ is some convex function. We call this G a *gain function*. Figure 1 (left) summarizes this for all the divergences in discussion. For divergence measures in the f-family, this function is simply $G(x) = (1-x)f(0) + xf(1/x)$, for $x > 0$. For the Rényi family, we have $G(x) = x^{1-\alpha} - 1$ for $x > 0$, $\alpha > 1$. Note that for f-divergence the gain function has a property that $G(1) = 0$; we make this consistent with Rényi divergence by adding a trailing -1.

We found that minimizing the objective (7) under the constraint (2) is equivalent to solving a multiple-resource allocation problem on convex returns [16]. It turns out that, in this problem, our objective is to minimize a mixture of document-level pay-offs, which is directly connected to the gain function $G(\cdot)$. Figure 1 (right) has a summary plot in which we print all these gain functions in different line patterns and those associated with Rényi divergence in different colors. From the plot, one can easily tell that all these measures appear to be convex, non-increasing monotone on $(0, 1]$ ("diminishing returns"). To minimize a single pay-off on any document d, it suffices to order the postings in descending order of probability $p(t|d)$ and have them enter the index consecutively until the budget runs out. The same idea also applies to a mixture of pay-offs.

General Solution. Let us denote a term t in some document d as $t_{[j]}$ by its rank j in descending order of $p(t|d)$. For any posting $(t_{[k]}, d)$ to enter the index, postings in document d with higher probabilities $(t_{[1]}, d), (t_{[2]}, d), \ldots, (t_{[k-1]}, d)$ have to be included first. Now by allowing posting $(t_{[k]}, d)$ to enter the final index, we *gain* this much in the overall objective:

$$p(d)\left[G\left(\sum_{i=1}^{k} p(t_{[i]}|d)\right) - G\left(\sum_{i=1}^{k-1} p(t_{[i]}|d)\right)\right] \tag{8}$$

Note that this value is negative. To minimize the overall gain, it suffices to go from some steady state and distribute the remaining budget to documents in a iterative fashion using the following greedy algorithm.

input: threshold ϵ
for $d \in \mathcal{D}$ **do**
  Sort terms in descending order of $p(t|d)$
  for $k = 1, \ldots, n$ **do**
    Compute $\Delta(t_{[k]}, d)$ according to (8)
    Remove posting $(t_{[k]}, d)$ if $|\Delta(t_{[k]}, d)| < \epsilon$

Algorithm 1: The general algorithm for computing optimal allocation under f-divergence and Rényi divergence.

Algorithm 1 computes the optimal allocation of $(1-\rho)N$ index entries that minimizes the divergence between the pruned and the full indexes. This algorithm has a time complexity of $O(|\mathcal{D}|L \log L)$ where $|\mathcal{D}|$ is the total number of documents in the collection and L is the maximum document length. This algorithm has a linear-time variants for variational distance, as given in Algorithm 2. These algorithms can all be linked to the result in Fox [15] by establishing the mapping between G and the function ϕ_j. Interested readers are referred to Ibaraki and Katoh [16] for more details.

input: threshold ϵ
for $d \in \mathcal{D}$ **do**
  for $t \in \text{posting}(d)$ **do**
    Remove posting (t, d) if $p(d)p(t|d) < \epsilon$

Algorithm 2: Linear-time variant for variational distance.

Rényi Divergence of Order Infinity. To compute the allocation for Rényi divergence of order infinity on arbitrary cardinality n, replace (8) in Algorithm 1 with the following:

$$\left(\sum_{i=1}^{k} p(t_{[i]}|d)\right)^{-n}. \tag{9}$$

Note that, since $p(d)$ is not involved in this equation, setting document priors would have no effect to this divergence measure. Given this condition, one can easily show that (9) is actually rank invariant for all $n > 0$. This means that Rényi divergence of order infinity can be solved for an arbitrarily high cardinality and the solution would still be the same as that of cardinality 1.

Divergence	Gain $G(x)$
f-divergence	$(1-x)f(0)+xf(1/x)$
KL divergence	$-\log x$
Variational distance	$1-x$
Hellinger's distance	$1-x^{1/2}$
χ^2-divergence	$x^{-1}-1$
Rényi divergence $(1<\alpha<\infty)$	$x^{1-\alpha}-1$

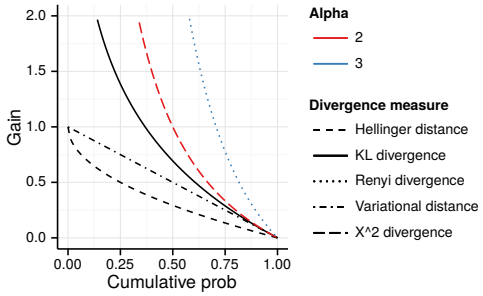

Figure 1: Gain functions (left) and plots (right) for many well-known divergence measures.

Title queries	50%			70%			90%		
	MAP	P20	J20	MAP	P20	J20	MAP	P20	J20
Full index	0.253	0.464	—	0.253	0.464	—	0.253	0.464	—
KL	0.234	0.465	**0.826**	0.210	0.461	0.664	0.143	0.357	0.360
Hellinger	0.208	0.453	0.800	0.162	0.418	0.586	0.074	0.238	0.237
Variational	0.117	0.382	0.565	0.059	0.301	0.275	0.015	0.129	0.078
χ^2-divergence	0.245	0.474	0.799	0.232	0.467	0.668	0.181	0.437	**0.373**
Renyi, $\alpha=50$	0.252	0.476	0.743	0.244	**0.485**	0.603	**0.198**	0.467	0.325
Renyi, $\alpha\to\infty$	**0.253**	**0.478**	0.741	**0.245**	**0.485**	0.598	**0.198**	**0.468**	0.323

Table 3: Retrieval performance of experimental runs on Terabyte '06 title queries, measured at prune ratios 50%, 70%, and 90%. Runs that outperform full index are underlined and best results printed in boldface.

4. EXPERIMENTS

Totally two sets of pruning experiments were conducted in this study. Our experiments were carried out on the GOV2 collection [12] using TREC 2006 Terabyte track data [8]. The GOV2 collection is a standard collection for various web-related retrieval tasks. It has 25.2 million documents and is roughly 426GB in size. We used the Indri toolkit[2] to create indexes and develop pruning algorithms. Standard preprocessing steps such as stemming and stopword removal were applied using the InQuery stoplist and porter stemmer.

We used both ad-hoc and efficiency topics as it is interesting to see how pruning algorithms respond to different types of queries. Ad-hoc topics in the Terabyte track are carefully selected questions with proper annotation, so this set is suitable for testing general retrieval performance. Efficiency topics are unannotated queries collected from session logs. Since these are real queries submitted by users, testing on top of this set gives us a better idea how pruning algorithms work "in the wild." For ad-hoc task we use all annotated topics 701-850, and for efficiency task we used the first 1,000, which are topics 1-1000. We used only title queries for both tasks.

For comparison, we chose five reference methods: term-based pruning [9], uniform pruning [9,10], document-centric pruning [7], popularity-based pruning [2,21], and two-sample two proportion (2N2P) test [26].[3] These methods implement different ideas in static index pruning and their performance have been extensively studied. Some of them such as term-based pruning and popularity-based pruning are known as standard methods of the task.

[2]`http://www.lemurproject.org/indri.php`

[3]We also tested two other methods, probability-ranking principle [5] and information preservation [11], but due to space limit these results are not included in this paper.

We use BM25 wherever applicable in post-pruning retrieval and in pruning (with the 2N2P test being the only exception) [1] to strengthen the baseline performance. Default parameters in the Indri toolkit are used: $k_1=1.25$ and $b=0.75$. For term-based pruning, we used the top-k version and set $k=10$. For document-centric pruning, we used Method 2 and set $\lambda=1-\rho$. To set up popularity-based pruning, we used term frequencies from the AOL query log [22] to compute term popularities. For the 2N2P test, we used only the Z-score version without implementing power analysis and also updated collection term frequencies. All other details were implemented based on standard settings.

We used a uniform prior $p(d)$ in our experiment. To estimate $p(t|d)$, we tested various retrieval methods, including BM25 and language modeling with both Dirichlet and Jelinek-Mercer smoothing. As BM25 is in general more effective, to prevent clutter we will not discuss the results for language modeling in this paper. Note that, since the BM25 scores are not valid probabilities, we use a softmax function to convert these scores as if they were coming out from a multinomial logistic regression model [3, p. 198]:

$$\frac{\exp\left(\text{BM25}(t,d)\right)}{\sum_{t'\in d}\exp\left(\text{BM25}(t',d)\right)}. \tag{10}$$

This estimate is not ideal since the "posterior" produced has little to do with the generative process. It is nevertheless a convenient way to incorporate non-probabilistic methods into our framework; further justification on its validity is beyond the scope of this paper.

Retrieval performance is measured at three prune ratios, 50%, 70%, and 90%. In all the experimental runs, prune ratio is controlled by using sample quantile [10] with reservoir sampling to find the right threshold ϵ. This estimation error was empirically bounded to within 0.005%. Given the same original index to start with, a pruning method is deemed

Title queries	50%				70%				90%			
	MAP	P20	J20	T (s)	MAP	P20	J20	T (s)	MAP	P20	J20	T (s)
Full index	0.253	0.464	—	101.7	0.253	0.464	—	101.7	0.253	0.464	—	101.7
2N2P test	0.239	0.467	0.714	40.3	0.203	0.434	0.535	18.3	0.076	0.248	0.198	2.2
Popularity-based	0.223	0.417	0.780	89.4	0.189	0.365	0.574	65.4	0.077	0.161	0.199	16.3
Uniform	0.231	0.445	0.760	**33.5**	0.187	0.376	0.566	**14.4**	0.110	0.241	0.273	**1.8**
Term-based, $k = 10$	0.218	0.457	**0.853**	67.9	0.187	0.441	**0.675**	46.8	0.109	0.311	0.350	14.7
Document-centric	**0.253**	**0.478**	0.743	54.7	0.244	**0.485**	0.602	38.8	**0.198**	0.465	0.325	16.1
KL	0.234	0.465	0.826	64.9	0.210	0.461	0.664	41.7	0.143	0.357	0.360	10.6
χ^2-divergence	0.245	0.474	0.799	59.9	0.232	0.467	0.668	38.2	0.181	0.437	**0.373**	13.2
Renyi, $\alpha = 50$	0.252	0.476	0.743	54.4	0.244	**0.485**	0.603	37.4	**0.198**	0.467	0.325	15.5
Renyi, $\alpha \to \infty$	**0.253**	**0.478**	0.741	54.9	**0.245**	**0.485**	0.598	37.0	**0.198**	**0.468**	0.323	15.6

SD queries	50%				70%				90%			
	MAP	P20	J20	T (s)	MAP	P20	J20	T (s)	MAP	P20	J20	T (s)
Full index	0.264	0.491	—	516.9	0.264	0.491	—	516.9	0.264	0.491	—	516.9
2N2P test	0.242	0.481	0.722	190.7	0.204	0.442	0.537	81.3	0.076	0.249	0.188	7.6
Popularity-based	0.232	0.439	0.781	389.6	0.198	0.375	0.581	277.4	0.080	0.170	0.194	56.0
Uniform	0.238	0.461	0.755	**141.3**	0.192	0.389	0.576	**50.4**	0.111	0.246	0.262	**3.5**
Term-based, $k = 10$	0.223	0.474	**0.852**	330.3	0.188	0.451	0.664	212.9	0.107	0.312	0.320	61.7
Document-centric	**0.259**	**0.499**	0.743	269.3	0.248	0.507	0.588	192.2	**0.200**	0.472	0.306	73.2
KL	0.240	0.476	0.842	313.1	0.211	0.470	**0.678**	181.2	0.137	0.340	0.337	39.3
χ^2-divergence	0.252	0.487	0.824	296.9	0.234	0.481	0.677	177.5	0.180	0.441	**0.354**	51.9
Renyi, $\alpha = 50$	0.258	0.498	0.750	269.1	0.248	0.506	0.592	183.4	**0.200**	0.472	0.306	71.3
Renyi, $\alpha \to \infty$	**0.259**	0.498	0.740	264.9	**0.249**	**0.508**	0.584	182.2	**0.200**	**0.474**	0.303	71.9

Table 4: Overall comparison with reference methods on Terabyte '06 title queries (top) and SD queries (bottom). Performance is measured at prune ratios 50%, 70%, and 90% where query execution is also timed. Runs do better than or equally well to full index are underlined; boldface indicates the best result.

better if the produced index delivers better result. To evaluate retrieval performance, we used the following measures: mean average-precision (MAP), precision-at-20 (P20) and top-20 Jaccard coefficient (J20). MAP and P20 measure how well one algorithm does in preserving postings that are relevant. J20 measures the degree of overlap in top 20 documents between retrieval results before and after pruning, commonly used as a proxy of precision-based measures when relevance judgments are not available. Note that other rank coefficient measures such as Kendall's tau may also be used in place of J20. These measures were selected mainly for consistency with the existing work.

4.1 Ad-Hoc Task

In the first task, we used 150 ad-hoc topics in the test. Besides title queries, we also managed to perform pruning experiments on sequential dependence (SD) queries [19]. Our purpose is to see how pruning algorithms respond to the change in term dependencies. In our case, the change is from full independence to sequential dependence. Although setting up SD queries on top of BM25 is unusual, it did achieve better performance as we had expected.

The result for our experimental runs on title queries is given in Table 3. Two variants of Rényi divergence are reported here: $\alpha = 50$ and $\alpha \to \infty$(order infinity). Among all the experimental runs, we found that Rényi divergence generally does the best, χ^2-divergence the second, and KL divergence the third. This seems to suggest that, for Rényi divergence, larger α tends to provide a better returns curve.

An overall comparison with reference methods is summarized in Table 4 with title queries on the top and SD queries on the bottom. Hellinger's distance and variational distance were not included in this comparison as the performance is below standard. Among the reference methods we tested, document-centric pruning does the best on MAP and P20. It outperforms all other reference runs by a large margin. The runner-up is term-based pruning, followed by 2N2P test, uniform pruning, and popularity-based pruning. The 2N2P test performed well on MAP at low prune ratio. In our test, the performance of uniform pruning is on par with term-based pruning on MAP, which is consistent with the previous results [10]. Popularity-based pruning also appeared comparable, but at high prune ratio its performance is just disappointing. Our result suggests that popularity-based method has no advantage in the ad-hoc task, although its parameters was trained on a very sizable source.

From Table 4, all four proposed methods do fairly well with precision-based measures. Rényi divergence of order infinity outperforms all the baseline methods on MAP and P20; the other three measures also achieve good performance but do not appear to surpass the strong baseline. On J20, KL divergence and χ^2-divergence both appear comparable to term-based method. We found that, in the Rényi family, the ones with small alpha (KL divergence and χ^2-divergence) tend to do better on J20 and the ones with large alpha ($\alpha = 50$ and $\alpha \to \infty$) do better on MAP and P20. Among all the proposed methods, only Rényi divergence managed to achieve comparable performance on MAP and P20 to the strong baseline. For testing statistical significance, we ran a 4-way ANOVA upfront followed by a Tukey's HSD test, whose result is given in Table 5. All effects in ANOVA come back significant for $p < 0.001$. The Tukey's test suggests that Rényi divergence and document-centric pruning significantly outperform all the other meth-

	Effect	DF	F	η_p^2
MAP	Query Type	1	15.1	.0015
	Method	8	96.6	.0693
	Prune Ratio	3	1262.0	.2673
	Topic	147	306.9	.8129
P20	Query Type	1	30.8	.0030
	Method	8	82.2	.0596
	Prune Ratio	3	355.4	.0931
	Topic	147	197.9	.7371

MAP	Mean	Grp
Rényi, $\alpha \to \infty$	.2419	a....
Document-centric	.2416	a....
Rényi, $\alpha = 50$	.2415	a....
χ^2-divergence	.2318	.b...
KL	.2130	..c..
Popularity-based	.2073	..cd.
Uniform	.2034	...de
2N2P test	.1959	....e
Term-based	.1949	....e

P20	Mean	Grp
Rényi, $\alpha \to \infty$	.4865	a...
Document-centric	.4858	a...
Rényi, $\alpha = 50$	.4853	a...
χ^2-divergence	.4709	a...
KL	.4434	.b..
Term-based	.4278	.bc.
2N2P test	.4123	..cd
Uniform	.3991	...d
Popularity-based	.3940	...d

Table 5: 4-way ANOVA (left) and Tukey's HSD result (right).

Efficiency queries	50%		70%		90%		Index Status at 90%		
	J20	T (s)	J20	T (s)	J20	T (s)	PruneT (s)	PL Kept (%)	Avg Size
Full index	—	990.49	—	990.49	—	990.49	—	100.0%	128.6
2N2P test	0.605	365.94	0.426	148.29	0.128	15.07	2858.09	100.0%	12.9
Popularity-based	**0.772**	815.23	0.515	643.52	0.182	209.14	**2382.56**	0.6%	2126.4
Uniform	0.646	**272.07**	0.450	**106.51**	0.178	**6.12**	3188.70	55.4%	23.2
Term-based	0.753	639.65	**0.563**	419.35	**0.296**	138.49	2694.58	100.0%	12.7
Document-centric	0.639	548.52	0.487	311.22	0.235	128.71	6987.10	40.9%	31.8
KL	0.730	545.84	0.538	324.64	0.235	85.96	6541.08	36.0%	35.8
χ^2-divergence	0.707	622.85	0.546	317.95	0.251	102.88	6767.16	37.9%	34.0
Renyi, $\alpha = 50$	0.642	511.48	0.490	306.96	0.236	128.26	8240.08	40.4%	31.9
Renyi, $\alpha \to \infty$	0.637	551.26	0.484	347.12	0.233	130.40	6830.29	40.6%	31.7

Table 6: Overall comparison with reference methods on 1,000 Terabyte '06 efficiency queries. Retrieval performance is evaluated using J20 as the sole indicator since relevance judgments are not available; pruning (PruneT) and query execution (T) are both timed and reported. Runs do better than or equally well to full index are underlined; boldface indicates the best result. Note that pruning time for popularity-based method does not include the time needed to preprocess query logs and is only indicative.

ods both on MAP and P20 (on P20 χ^2-divergence is also in the leading group). Rényi divergence appears to have a slight advantage over document-centric pruning, but the improvement is not significant.

The performance of document-centric pruning has raised some concerns. We believe that its effectiveness has been previously overlooked, since many studies either compared with the version with the KLD score function, which is inferior to our implementation, or did not replicate the result at all. According to our experiments, which cover many recent approaches, document-centric pruning may currently be the best pruning strategy for preserving top-k precision.

Based on all these findings, we conclude that the proposed divergence-based methods are effective in producing quality pruned indexes, and their performance is among the best on the GOV2 data. One thing worth noting is that on title queries with Rényi divergence, we delivered better P20 scores than on the full index at all prune ratios up to 90%. On SD queries the same method delivered better P20 results for prune ratios up to 70%.

4.2 Efficiency Task

We conducted the second experiment on the efficiency task data. This is to see how pruning algorithms react to more realistic query topics. In the experiment, we tested each pruning method against the first 1,000 efficiency topics. As relevance judgments are not available, we had to rely on J20 as the sole performance indicator. Note that one caveat with this experimental setting is that J20 can be optimistic and does not reliably reflect true retrieval performance, as can be seen in the result of our first experiment. Nevertheless, without relevance judgments it is perhaps the best proxy measure to the true performance.

We also conducted timing experiments on a dedicated server with a 3.30 GHz Intel Core i5-2500 CPU (4 cores) and 16GB RAM. We report time needed to produce the pruned index (PruneT), and query execution time (T) over the entire set of 1,000 topics. Note that PruneT is only indicative because it was an one-off measurement; we did not make a second timing pass because pruning is very costly. Query execution time is however properly measured in a two-pass timing procedure to isolate possible caching effects.

The result is given in Table 6. We only report PruneT for 90% prune ratio for simplicity. On J20, popularity-based method works the best at 50% prune ratio, but as prune ratio increases term-based pruning tends to deliver better performance. Other reference methods are not effective on J20. Among the experimental runs, J20 favors more towards KL divergence but at high prune ratio Rényi divergence also does equally well. Overall, term-based pruning delivers the best performance on J20.

The timing result on PruneT (pruning time, in seconds) confirms that term-centric methods, e.g., the 2N2P test, popularity-based, uniform, and term-based, are more efficient to run. Among all these methods, Rényi divergence (the one of finite order) would be the most expensive, since its gain function has a power component that takes more time to compute. The others in the document-centric camp appear roughly comparable on pruning time. On query execution time T (in seconds), uniform pruning is the fastest in

most cases, although in our previous experiment this speed gain did not translate into precision. The 2N2P test is also fast and performs better than uniform pruning at low prune ratio. Divergence measures and document-centric pruning are not known to be fast; in our test, they all seem equally slow. Nevertheless popularity-based and term-based took even more time to evaluate queries. Although these methods are known to produce more matching against user input, the result is still surprising. Also, we noticed something unusual that, on 90% prune ratio, uniform pruning ran through 1,000 query topics in only 6.12 seconds. These anomalies suggest that the final index produced by these methods may have been seriously degraded.

4.3 Distribution of Posting List Size

To investigate why some reference algorithms had been acting strangely in the timing test, we came back to the indexes on 90% prune ratio and did more analysis. We started by looking at two indicators, which are percentage of non-empty posting lists (PL Kept), and average posting list size (Avg Size), both included in Table 6. We also find it informative to look at the frequency distribution of posting list size, which is covered in Figure 2. Table 6 shows that divergence-based methods and document-centric pruning keep only 36–40% of the posting lists and maintained an average size of posting list at 31–36 postings. The frequency distributions (Figure 2) also look fairly smooth and normal, suggesting that these pruning algorithms are well-behaved.

While uniform pruning and the 2N2P test both have very similar frequency distributions and are more or less comparable in retrieval performance, on PL Kept and Avg Size it is term-based pruning that agrees more with the 2N2P test. Uniform pruning does nothing unusual as well; it did not make large changes to long/short posting lists nor eliminate more postings than the others do. From the distribution plot, it appears to fit the original distribution fairly well. The number of posting lists in the index (55.4%) can be a little bit more, but this should not be an issue because it takes *less* time processing queries. The real problem with uniform pruning, we suspect, is that meaningless contents such as HTML fragments or overlong words found their way into the index. BM25 incorrectly assigned high scores to these strings, producing a large set of high-probability garbage that uniform pruning provides no safe-guard mechanism against. Evaluating queries on the index would result in far less execution time because the underlying index contains many useless terms and nothing is left in there to match.

Popularity-based pruning is found to overfit the query log data in an amusing way. Table 6 shows that it threw away more than 99% of term posting lists and kept only 0.6% in the final index. In Figure 2, we can see there is a noticeable gap between the size distribution and the original, a sign that indicates *under-fitting*. This extreme strategy would not have succeeded unless carried out in an environment where the majority of query topics are covered in the training data. This explains why popularity-based pruning is more successful in log-based experiments and why it was not working as expected on Terabyte '06 ad-hoc topics.

5. DISCUSSIONS

We have gone through a series of analyses in light of answering the research questions we raised in Section 1. Our

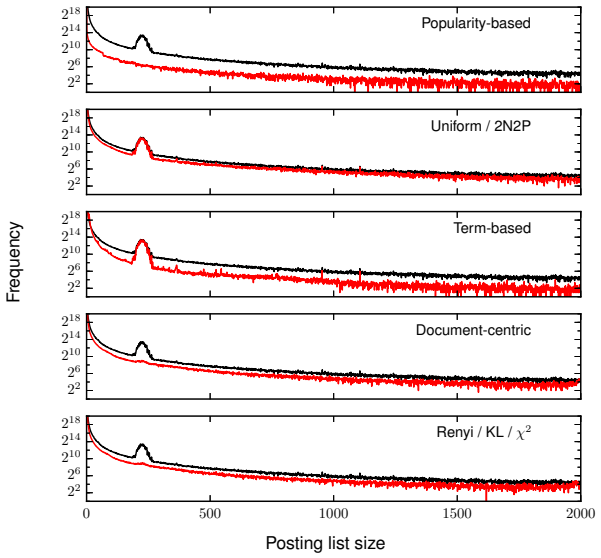

Figure 2: Semi-log plots on posting list size and frequency. Each plot has a black curve for the full index and a red curve for the 90% pruned index using the named method(s).

first finding is that generalizing the divergence measure in problem (1) is not only feasible but advantageous. Through experimentation, we show that using Rényi divergence of order infinity in place of ordinary KL divergence gives the best result. This is in line with our theoretical analysis in Section 3.2 which states that Rényi divergence has the greatest flexibility in modeling multiple-term queries.

We also found that modeling multiple-term queries does not make optimization any harder. Many divergence measures that we considered can be solved for arbitrarily high cardinality, and therefore the solutions can be empirically achieved and tested.[4]

Term-based pruning is known for retaining top k postings for each term, effectively keeping all posting lists active in the final index [9]. In our study, we found that Rényi divergence shares this interesting trait that it retains at least one posting on each document, ensuring access to every document is possible. This is due to the nature of its document pay-offs. As the gain functions for Rényi-compatible measures are unbounded at the point where the cumulative probability is zero, by allowing the first posting in each document to enter the index we would gain minus infinity in the objective value. This "keeping everything accessible" strategy is actually a side effect of the returns curves.

Our experiments also showed that the J20 measure does not align well with precision, and methods with a strong presence in J20 can be weak on both precision and recall, which is evidenced by low MAP values at higher prune ratios. Optimizing top-k similarity in web search to produce quality data summary may now seem like an unfounded idea.

[4]For measures that are not convex at higher cardinality, their actual performance remains unknown because the true solutions cannot be computed using our algorithm.

6. CONCLUSIONS

In this paper, we provide a thorough study on a wide range of divergence measures and their use on static index pruning. We developed a set of theoretical analyses on the improved divergence minimization framework. Our work has paved the way for practical implementation of optimal pruning strategies for large-scale nonparametric models such as inverted indexes. We have also uncovered interesting effects that different divergence measures and cardinality settings may have on the solution quality. The analysis of cardinality suggests that using Rényi divergence of order infinity in static index pruning delivers the best performance across different query cardinality settings, which is confirmed empirically with extensive experiments.

For future work, one possible direction would be to use the returns curves of divergence measures to assign term weights. This technique may be directly applied to other problems such reranking or summarization. As static index pruning relies heavily on term weighting schemes (i.e., score functions) to estimate posting importance, it is worthwhile to explore the relationship between the two.

7. ACKNOWLEDGMENTS

This work was supported in part by ARC Discovery Grant DP140102655, in part by the Center for Intelligent Information Retrieval, and in part by NSF IIS-1160894. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] I. S. Altingovde, R. Ozcan, and O. Ulusoy. A practitioner's guide for static index pruning. In *Proceedings of ECIR '09*, pages 675–679. Springer Berlin / Heidelberg, 2009.

[2] I. S. Altingovde, R. Ozcan, and O. Ulusoy. Static index pruning in web search engines: Combining term and document popularities with query views. *ACM Trans. Inf. Syst.*, 30(1), Mar. 2012.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[4] R. Blanco and A. Barreiro. Static pruning of terms in inverted files. In *Proceedings of ECIR '07*, pages 64–75. Springer Berlin Heidelberg, 2007.

[5] R. Blanco and A. Barreiro. Probabilistic static pruning of inverted files. *ACM Trans. Inf. Syst.*, 28(1), Jan. 2010.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

[7] S. Büttcher and C. L. A. Clarke. A document-centric approach to static index pruning in text retrieval systems. In *Proceedings of CIKM '06*, pages 182–189. ACM, 2006.

[8] S. Büttcher, C. L. A. Clarke, and I. Soboroff. The TREC 2006 terabyte track. In *TREC*, volume 6, page 39, 2006.

[9] D. Carmel, D. Cohen, R. Fagin, E. Farchi, M. Herscovici, Y. S. Maarek, and A. Soffer. Static index pruning for information retrieval systems. In *Proceedings of SIGIR '01*, pages 43–50. ACM, 2001.

[10] R.-C. Chen and C.-J. Lee. An information-theoretic account of static index pruning. In *Proceedings of SIGIR '13*, pages 163–172. ACM, 2013.

[11] R.-C. Chen, C.-J. Lee, C.-M. Tsai, and J. Hsiang. Information preservation in static index pruning. In *Proceedings of CIKM '12*, pages 2487–2490. ACM, 2012.

[12] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC-2004 terabyte track. In *Proceedings of TREC-2004*, 2004.

[13] I. Csiszár and P. C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends®in Communications and Information Theory*, 1(4):417–528, 2004.

[14] E. S. de Moura, C. F. dos Santos, D. R. Fernandes, A. S. Silva, P. Calado, and M. A. Nascimento. Improving web search efficiency via a locality based static pruning method. In *Proceedings of WWW '05*, pages 235–244. ACM, 2005.

[15] B. Fox. Discrete optimization via marginal analysis. *Management science*, 13(3):210–216, 1966.

[16] T. Ibaraki and N. Katoh. *Resource Allocation Problems: Algorithmic Approaches*. MIT Press, 1988.

[17] S. Kullback. *Information Theory and Statistics*. Wiley, 1959.

[18] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Th.*, 52(10):4394–4412, Oct. 2006.

[19] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proceedings of SIGIR '05*, pages 472–479. ACM, 2005.

[20] T. Morimoto. Markov processes and the H-Theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, Mar. 1963.

[21] A. Ntoulas and J. Cho. Pruning policies for two-tiered inverted index with correctness guarantee. In *Proceedings of SIGIR '07*, pages 191–198. ACM, 2007.

[22] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*, page 1. ACM, 2006.

[23] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR '98*, pages 275–281. ACM, 1998.

[24] A. Rényi. On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, 1961.

[25] G. Skobeltsyn, F. Junqueira, V. Plachouras, and R. B. Yates. ResIn: a combination of results caching and index pruning for high-performance web search engines. In *Proceedings of SIGIR '08*, pages 131–138. ACM, 2008.

[26] S. Thota and B. Carterette. Within-document term-based index pruning with statistical hypothesis testing. In *Proceedings of ECIR '11*, pages 543–554. Springer Berlin Heidelberg, 2011.

[27] T. van Erven and P. Harremoes. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Th.*, 60(7):3797–3820, July 2014.

[28] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, Apr. 2004.

APPENDIX

Convexity

In this section, we outline the proofs of Lemma 1 on convexity of f-divergence and Lemma 2 on convexity of a surrogate of Rényi divergence.

LEMMA 1 (CONVEXITY). *Given $Z_d > 0$ for all d, $D_f(q||p)$ defined in (6) is jointly convex in pruning decisions $\langle \mathbb{I}_{t,d} | \forall t, d\rangle$ for any convex function f with $f(1) = 0$.*

PROOF. Let us split the support of the summation and organize. We have:

$$\begin{aligned} \mathrm{D}_f(q||p) &= \sum_{t,d} p(t,d)\left[(1-\mathbb{I}_{t,d})f(0) + \mathbb{I}_{t,d} f(1/Z_d)\right] \\ &= \sum_{d} p(d)\left[(1-Z_d)f(0) + Z_d f(1/Z_d)\right]. \end{aligned} \tag{11}$$

The term $Z_d f(1/Z_d)$ is convex in Z_d because it is a special type of perspective function [6]. Since Z_d is affine in pruning decisions, this proof follows. □

LEMMA 2 (SURROGATE CONVEXITY). *Given $Z_d > 0$ for all d, minimizing $D_\alpha(q||p)$ in (6) has an equivalent surrogate that is jointly convex in $\langle \mathbb{I}_{t,d} | \forall t, d\rangle$ for $\alpha > 1$.*

PROOF. Since logarithm function is monotone, when $\alpha > 1$, minimizing $\mathrm{D}_\alpha(q||p)$ as in (6) is equivalent to minimizing

$$\sum_{t,d} p(t,d)\left(\frac{\mathbb{I}_{t,d}}{Z_d}\right)^\alpha = \sum_d p(d) Z_d^{1-\alpha}, \tag{12}$$

which is jointly convex in pruning decisions $\langle \mathbb{I}_{t,d} | \forall t, d\rangle$. □

Analysis for Cardinality $n > 1$

In the following paragraphs, we describe how to simplify and analyze each divergence measures discussed in the paper.

Kullback-Leibler Divergence. It can be shown that minimizing KL divergence for cardinality $n > 1$ is equivalent to minimizing for $n = 1$. The key idea is to split

$$\log[q(t_{1:n}, d)/p(t_{1:n}, d)]$$

into a summation $\sum_j \log[q(t_j|d)/p(t_j|d)]$. Then we have

$$\begin{aligned} &\arg\min\, n\left(\sum_d p(d)\sum_t q(t|d)\log\frac{\mathbb{I}_{t,d}}{Z_d}\right) \\ =&\arg\min -\sum_d p(d)\log Z_d. \end{aligned} \tag{13}$$

Variational Distance. This divergence measure is not convex even under the bag-of-words assumption. To see how, we first write out the definition and split support:

$$\begin{aligned} &\sum_{t_{1:n},d} p(t_{1:n},d)\left|1 - \prod_j \frac{\mathbb{I}_{t_j,d}}{Z_d}\right| \\ =&\sum_{t_{1:n},d} p(t_{1:n},d)\left(\prod_j \mathbb{I}_{t_j,d}(Z_d^{-n}-1) + (1-\prod_j \mathbb{I}_{t_j,d}).\right) \end{aligned} \tag{14}$$

Then this would lead to $2(1 - \sum_d p(d) Z_d^n)$, which is not convex for all integer $n > 1$.

Hellinger's Distance. When $n = 2$ Hellinger's distance has the same analytic form as variational distance with cardinality 1. For $n > 2$, the divergence no longer remains convex. At some point in the derivation, we have Hellinger's distance in the following form:

$$\begin{aligned} &2\left(1 - \sum_{t_{1:n},d} p(d) Z_d^{-n/2} \prod_j p(t_j|d)\mathbb{I}_{t_j,d}\right) \\ =&2\left(1 - \sum_d p(d) Z_d^{n/2}\right). \end{aligned} \tag{15}$$

This would fall back to variational distance with cardinality 1 (cf. Table 2) when $n = 2$.

χ^2-Divergence. There is a one-one mapping between χ^2-divergence of cardinality n to Rényi divergence of cardinality 1 with $\alpha = n + 1$. Let us start by plugging the bag-of-words assumption into the definition and replace all the $q(t_j|d)/p(t_j|d)$ with $\mathbb{I}_{t_j,d}/Z_d$. Organize a bit, the divergence is written as follows:

$$-1 + \sum_d p(d)\left(\sum_t p(t|d)\frac{\mathbb{I}_{t,d}}{Z_d^2}\right)^n = -1 + \sum_d p(d) Z_d^{-n}. \tag{16}$$

This has the same analytic form as Rényi divergence. Therefore, minimizing χ^2-divergence for arbitrary cardinality n is equivalent to minimizing Rényi divergence of $\alpha = n + 1$ in cardinality 1.

Rényi Divergence of Order α. It turns out Rényi divergence has an interesting property that, under the term independence assumption, minimizing Rényi divergence of order α in cardinality n is equivalent to doing Rényi divergence of order $n\alpha - n + 1$ in cardinality 1. Following the same maneuver, we have:

$$\begin{aligned} &\frac{1}{\alpha - 1}\log\sum_d p(d)\left(\sum_t p(t,d)\mathbb{I}_{t,d} Z_d^{-\alpha}\right)^n \\ =&\frac{1}{\alpha-1}\log\sum_d p(d) Z_d^{n(1-\alpha)}. \end{aligned} \tag{17}$$

It is clear that since $n(1-\alpha) = 1 - (n\alpha - n + 1)$, this is equivalent to minimizing a cardinality-1 Rényi divergence. This is a bijection as $\alpha > 1 \Leftrightarrow n\alpha - n + 1 > 1$ for all integer $n \geq 0$.

Rényi Divergence of Order Infinity. This measure is perhaps the most curious one in the study. Despite being a special case of Rényi divergence of order α, when evaluated in high cardinality n the divergence does *not* automatically fall back to itself in cardinality 1. Nevertheless, this divergence is actually rank invariant for all cardinalities $n > 0$ (cf. Section 3.3). Its analytic form can be derived as follows:

$$\log \sup_{t_{1:n},d} Z_d^{-n} \prod_j \mathbb{I}_{t_j,d} = \log \sup_d Z_d^{-n}. \tag{18}$$