On divergence measures and static index pruning

Ruey-Cheng Chen
RMIT University

Chia-Jung Lee and **W. Bruce Croft**
University of Massachusetts Amherst

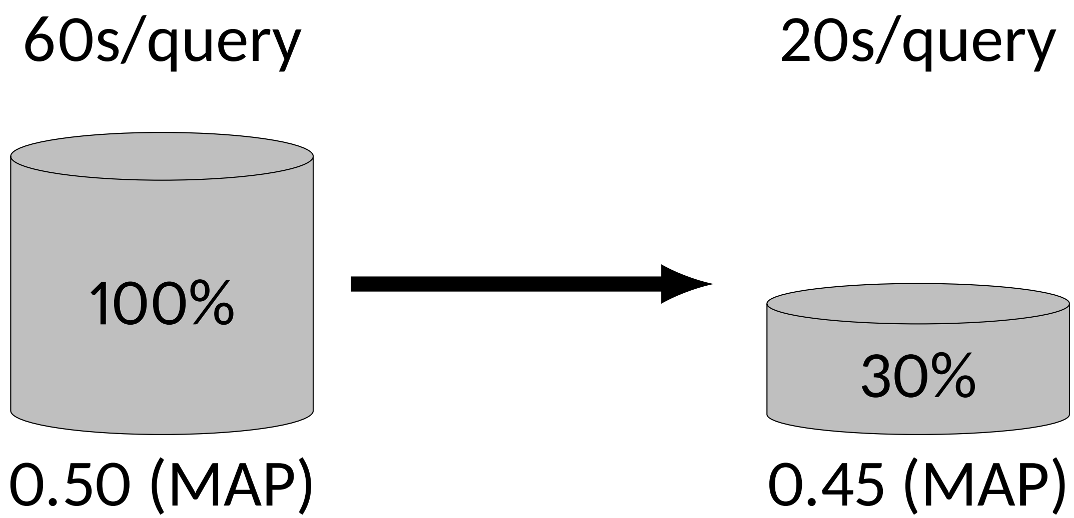
Sep 29, 2015 (ICTIR ’15)

Motivation

Why is static index pruning relevant?

Static Index Pruning

Remove a fraction of (less important) postings out of the index.



- Improved disk usage and query throughput
- Reduced retrieval performance

Why Pruning the Index?

i) **Index is too large to run.**

- Document retrieval on handheld devices.[1]

ii) **Retrieval is slow, so you use a cache to serve top results.**

- Tiered indexing for Web search.[2]

iii) **Retrieval is slow, and you can trade off some effectiveness.**

- Accelerated analysis over verbose queries or complex needs. (e.g., *question answering*, *semantic indexing*, ...)

[1] Carmel et al. (2001). "Static index pruning for information retrieval systems". SIGIR '01.

[2] Büttcher and Clarke. (2006). "A document-centric approach to static index pruning in text retrieval systems". CIKM '06.

Budget

Pruning is like running a budget over *postings*.

In my definition, a posting is of the form (t, d, n), meaning that **term** t **appears** n **times in document** d.

```
(quick, d1, 1)   (fox,  d2, 3)   (quick, d3, 3)
(brown, d1, 1)   (jump, d2, 1)   (dog,   d3, 1)
(lazy,  d1, 2)   (dog,  d2, 2)
...
```

The budget may vary from application to application, but in general you want to avoid investing on:

- Ineffective terms/documents;
- Low-impact postings.

Ideas (That Have Been Tried)

Term-based pruning[1], document-centric pruning[2], (term) informativeness and discriminative value[3], term popularity[4] and caching[5], entropy[6], probability ranking principle[7], two-sample two-proportion (2P2N)[8], information preservation[9], query view[10].

[1] Carmel et al. (2001). "Static index pruning for information retrieval systems". SIGIR '01.

[2] Büttcher and Clarke. (2006). "A document-centric approach to static index pruning in text retrieval systems". CIKM '06.

[3] Blanco and Á. Barreiro. (2007). "Static Pruning of Terms in Inverted Files". ECIR '07.

[4] Ntoulas and Cho. (2007). "Pruning policies for two-tiered inverted index with correctness guarantee". SIGIR '07.

[5] Skobeltsyn et al. (2008). "ResIn: a combination of results caching and index pruning for high-performance web search engines". SIGIR '08.

[6] Zheng and Cox. (2009). "Entropy-Based Static Index Pruning". ECIR '09.

[7] Blanco and A. Barreiro. (2010). "Probabilistic static pruning of inverted files". *ACM Transactions on Information Systems*.

[8] Thota and Carterette. (2011). "Within-Document Term-Based Index Pruning with Statistical Hypothesis Testing". ECIR '11.

[9] Chen et al. (2012). "Information preservation in static index pruning". CIKM '12.

[10] Altingovde et al. (2012). "Static index pruning in web search engines: Combining term and document popularities with query views". *ACM Transactions on Information Systems*.

Divergence-Based Method

Principle of Minimum Cross-Entropy[1]

Consider an initial measure p and a set of feasible measures $\mathcal{F}$. To update one's measurement about the system, choose a measure q so as to:

$$\begin{array}{ll} \text{minimize} & \mathsf{D}(q||p) \\ \text{subject to} & q \in \mathcal{F}. \end{array} \tag{1}$$

[1] Kullback. (1959). *Information Theory and Statistics*.

“Index” Version[1] of the Same Principle

Consider an index p and a set of possible index states $\mathcal{F}$ resulted from pruning p according to some space constraint. Choose **a new index** q so as to:

$$\begin{array}{ll} \text{minimize} & \mathsf{D}(q||p) \\ \text{subject to} & q \in \mathcal{F}. \end{array} \tag{2}$$

[1] Chen and Lee. (2013). “An Information-Theoretic Account of Static Index Pruning”. SIGIR ’13.

“Index” Version[1] of the Same Principle

Consider an index p and a set of possible index states $\mathcal{F}$ resulted from pruning p according to some space constraint. Choose **a new index** q so as to:

$$\begin{array}{ll} \text{minimize} & \mathsf{D}(q||p) \\ \text{subject to} & q \in \mathcal{F}. \end{array} \quad (2)$$

Optimal solution can be approximated by uniform pruning.

- Probability mass not properly renormalized;
- Objective not exactly solved;
- Multiple-term queries not modeled;
- Limited choice of divergence measure.

[1] Chen and Lee. (2013). “An Information-Theoretic Account of Static Index Pruning”. SIGIR ’13.

Research Questions

i) Is the information-theoretic framework a practical one?

- *Can we compute the exact solution?*
- *Can we generalize over the choice of divergence measures?*
- *Can this framework model multiple-term queries?*

Research Questions

i) Is the information-theoretic framework a practical one?

- *Can we compute the exact solution?*
- *Can we generalize over the choice of divergence measures?*
- *Can this framework model multiple-term queries?*

ii) What makes a good pruning strategy?

- *Is it good or bad to remove whole terms/documents entirely?*
- *How do we run the budget over multiple documents?*

Research Questions

i) Is the information-theoretic framework a practical one?

- *Can we compute the exact solution?*
- *Can we generalize over the choice of divergence measures?*
- *Can this framework model multiple-term queries?*

ii) What makes a good pruning strategy?

- *Is it good or bad to remove whole terms/documents entirely?*
- *How do we run the budget over multiple documents?*

iii) **What pruning method empirically works the best?**

Ingredient #1: Generative Story

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.

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

Then one ranks documents based on the *joint likelihood*.

Assumptions: Pruning is to induce a new set of document models.

Ingredient #2: Problem Formulation

Given an index p and a prune ratio ρ, choose an index q so as to:

$$\begin{array}{ll} \text{minimize} & \mathsf{D}(q||p) \\ \text{subject to} & \mathbb{I}_{t,d} \in \{0,1\} \text{ for all } (t,d) \\ & \sum_{t,d} \mathbb{I}_{t,d} = (1-\rho)N \\ & q \in \mathcal{Q}(p) \end{array} \tag{3}$$

The constraint $q \in \mathcal{Q}(p)$ is equivalent to:

$$q(t|d) = p(t|d)\mathbb{I}_{t,d} \qquad \text{for all } t, d. \tag{4}$$

(We originally tackled this problem.)

Ingredient #2: Problem Formulation

Given an index p and a prune ratio ρ, choose an index q so as to:

$$\begin{array}{ll} \text{minimize} & \mathsf{D}(q||p) \\ \text{subject to} & \mathbb{I}_{t,d} \in \{0,1\} \text{ for all } (t,d) \\ & \sum_{t,d} \mathbb{I}_{t,d} = (1-\rho)N \\ & q \in \mathcal{Q}(p) \end{array} \tag{3}$$

The constraint $q \in \mathcal{Q}(p)$ is equivalent to:

$$q(t|d) = \frac{p(t|d)\mathbb{I}_{t,d}}{\color{red}{\sum_{t'} p(t'|d)\mathbb{I}_{t',d}}} \qquad \text{for all } t, d. \tag{4}$$

Now, the probability mass is normalized (the factor called Z_d).

Ingredient #2: Problem Formulation

Given an index p and a prune ratio ρ, choose an index q so as to:

$$\begin{array}{ll} \text{minimize} & \mathsf{D}(q||p) \\ \text{subject to} & \mathbb{I}_{t,d} \in \{0,1\} \text{ for all } (t,d) \\ & \sum_{t,d} \mathbb{I}_{t,d} = (1-\rho)N \\ & q \in \mathcal{Q}(p) \end{array} \tag{3}$$

The constraint $q \in \mathcal{Q}(p)$ is equivalent to:

$$q(t_{1:n}|d) = \frac{p(t_{1:n}|d) \prod_j \mathbb{I}_{t_j,d}}{\sum_{t'_{1:n}} p(t'_{1:n}|d) \prod_j \mathbb{I}_{t'_j,d}} \qquad \text{for all } t_{1:n}, d. \tag{4}$$

We call n the *query cardinality*.

Ingredient #3: Divergence Measures

We generalize the choice of divergence measures.

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

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

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

f-Divergence[1,2]

A family of measures parametrized by the functional f.

$$\mathsf{D}_f(q||p) = \sum_{t_{1:n},d} p(t_{1:n},d) f\left(\frac{q(t_{1:n},d)}{p(t_{1:n},d)}\right), \qquad (5)$$

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

[1] Csiszár and Shields. (2004). "Information Theory and Statistics: A Tutorial". *FnT in Communications and Information Theory*.
[2] Morimoto. (1963). "Markov Processes and the H-Theorem". *Journal of the Physical Society of Japan*.

Rényi Divergence of Order α[1]

Another well-known family parametrized by α.

$$\mathsf{D}_\alpha(q||p) = \frac{1}{\alpha - 1}\log\left(\sum_{t_{1:n},d} q(t_{1:n},d)^\alpha p(t_{1:n},d)^{1-\alpha}\right). \tag{6}$$

Kullback-Leibler divergence	$\alpha \to 1$
Logarithm of χ^2-divergence	$\alpha = 2$

[1] Rényi. (1961). "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*.

Rényi Divergence of Order Infinity[1]

One can actually take α to infinity:

$$\mathsf{D}_\infty(q||p) = \log \sup_{t_{1:n},d} \frac{q(t_{1:n}, d)}{p(t_{1:n}, d)}. \tag{7}$$

[1] Erven and Harremoes. (2014). “Rényi Divergence and Kullback-Leibler Divergence”. *IEEE Trans. Inf. Th.*

Analysis

How to solve it?

Approach

i) Work on $n = 1$:

- Use you algebra to simplify the objective;
- Check if the objective is convex;
- Form a numerical/algorithmic solution.

ii) Repeat the procedure with $n = 2, 3, \ldots$ and so on.

- Check if the problem can be reduced to smaller n.

Analytic Form: $n = 1$

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

All these objectives are convex.

Convexity: $n = 1$

It is known that both families are convex in measures p and q, but convexity in pruning decisions $\langle \mathbb{I}_{t,d} | \forall t, d \rangle$ is not yet established.

Lemma 1 (Convexity of f-divergence)**.** *Given $Z_d > 0$ for all d, D$_f(q||p)$ 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 of Rényi divergence)**.** *Given $Z_d > 0$ for all d, minimizing D$_\alpha(q||p)$ has an equivalent surrogate that is jointly convex in $\langle \mathbb{I}_{t,d} | \forall t, d \rangle$ for $\alpha > 1$.*

Analytic Form: $n > 1$

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

Assumption: $p(t_{1:n}|d) = \prod_j p(t_j|d)$ (bag of word)

Analytic Form: $n > 1$

Divergence	Analytic Form
$\text{KL}^{(n)}$	$\text{KL}^{(1)}$
$\text{VD}^{(n)}$	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 $1 < \alpha < \infty$
$\text{Rényi}^{(n)}_{\infty}$	$\sup_d \left(\sum_{t'} \mathbb{I}_{t',d}\, p(t'\|d)\right)^{-n}$

Assumption: $p(t_{1:n}|d) = \prod_j p(t_j|d)$ (bag of word)

KL, χ^2-div, and Rényi can be solved for arbitrary n, meaning they are more flexible in modeling multiple-term query.

Gain Functions

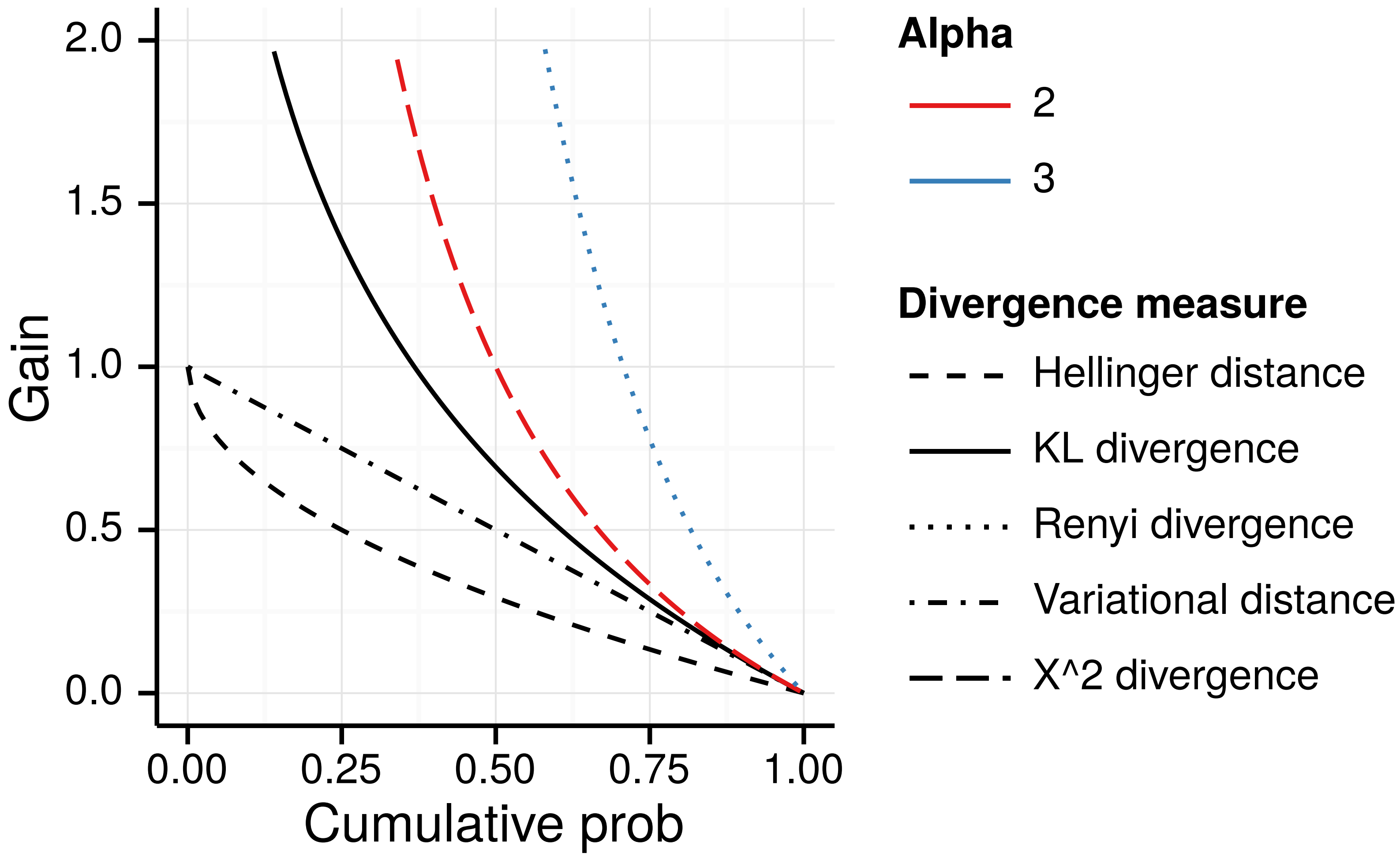
All these measures except Rényi$_\infty$ have similar analytic forms:

$$\sum_d p(d)\, G\left(\overbrace{\sum_t \mathbb{I}_{t,d} p(t|d)}^{\text{cumulative prob}}\right). \tag{8}$$

where $G(x)$ is non-increasing monotone, convex on $(0,1]$.

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 ($\alpha > 1$)	$x^{1-\alpha}-1$

Gain Functions



Allocation for One Document

Objective:

$$\sum_d p(d)\, G\overbrace{\left(\sum_t \mathbb{I}_{t,d}\, p(t|d)\right)}^{\text{cumulative prob}}. \quad (9)$$

Let us denote a term in some document d as $t_{[j]}$ by its rank j in descending order of $p(t|d)$.

Result: 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 (due to the property of G.)

- *The returns for each document can be seen as a step function.*
- *For some gain functions, $G(0)$ is unbounded.*

Optimal Allocation

Define $\Delta(t_{[k]}|d)$ as:

$$p(d)\left[G\left(\sum_{i=1}^{k} p(t_{[i]}|d)\right) - G\left(\sum_{i=1}^{k-1} p(t_{[i]}|d)\right)\right]. \tag{10}$$

This algorithm computes optimal allocation in $O(|\mathcal{D}| \log n)$ time.[1,2]

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 (10) ;
        Remove posting $(t_{[k]}, d)$ if $|\Delta(t_{[k]}, d)| < \epsilon$;

[1] Fox. (1966). “Discrete optimization via marginal analysis”. *Management science*.

[2] Ibaraki and Katoh. (1988). *Resource Allocation Problems: Algorithmic Approaches*.

Optimal Allocation: Variants

For $VD^{(1)}$, there is a linear-time algorithm.

for $d \in \mathcal{D}$ **do**
  for $t \in \mathrm{posting}(d)$ **do**
    Remove posting (t, d) if $p(d)p(t|d) < \epsilon$;

Optimal Allocation: Variants

For $VD^{(1)}$, there is a linear-time algorithm.

for $d \in \mathcal{D}$ **do**
    for $t \in \mathrm{posting}(d)$ **do**
        Remove posting (t, d) if $p(d)p(t|d) < \epsilon$;

For Rényi$^{(n)}_{\infty}$, run the original algorithm with (10) replaced by:

$$\left(\sum_{i=1}^{k} p(t_{[i]}|d)\right)^{-n}. \tag{11}$$

- *The document prior is disregarded.*
- *This definition is rank-invariant for* $n > 0$.

Summary

i) *The optimal solution can be exactly and efficiently computed.*

- Depends on less assumptions.
- Requires no approximation.
- Generates a set of "document-centric" approaches.

ii) *The Rényi family has the greatest flexibility in modeling queries.*

Questions:

- Relation with existing approaches
- Joint vs. conditional modeling
- $\mathrm{D}(q||p)$ vs. $\mathrm{D}(p||q)$
- Jensen-Shannon divergence
- Smoothing integrated into $\mathcal{Q}(p)$

Experiments

Caution: Bumpy road ahead

Experimental Setup

Benchmark: GOV2 collection, using both ad-hoc (topic 701-850) and efficiency topics (1-1000) from TREC Terabyte '06.

Index created using Indri with porter stemmer and standard 401 InQuery stoplist. Run Title/SD queries using BM25 in post-pruning retrieval. Three prune levels tested: 50%, 70%, and 90%.

Using BM25 to estimate $p(t|d)$:

$$\frac{\exp\left(\text{BM25}(t,d)\right)}{\sum_{t' \in d} \exp\left(\text{BM25}(t',d)\right)}. \tag{12}$$

i) Consistency with the choice of score function;
ii) Better performance.

More on Experimental Setup

Reference methods:

Term-based, uniform, document-centric, popularity-based, two-sample two-proportion test (2N2P), probability ranking principle, information preservation

Metrics:

MAP, P20, J20 (jaccard coefficient @20), Time

Result: Ad-Hoc Topics

(**Boldface** = best result; underline = better than full index; Column group = ρ)

Title queries	50%			70%			90%		
	MAP	P20	J20	MAP	P20	J20	MAP	P20	J20
Full index	.253	.464	—	.253	.464	—	.253	.464	—
KL	.234	.465	**.826**	.210	.461	.664	.143	.357	.360
Hellinger	.208	.453	.800	.162	.418	.586	.074	.238	.237
VD	.117	.382	.565	.059	.301	.275	.015	.129	.078
χ^2-div	.245	.474	.799	.232	.467	**.668**	.181	.437	**.373**
Rényi, $\alpha = 50$	.252	.476	.743	.244	**.485**	.603	**.198**	.467	.325
Rényi, $\alpha \to \infty$	**.253**	**.478**	.741	**.245**	**.485**	.598	**.198**	**.468**	.323

- The performance of Hellinger and VD is below standard.
- On MAP and P20: Rényi$_\infty >$ Rényi$_{\alpha=50} > \chi^2$-div $>$ KL.
- On J20: KL and χ^2-div work better

Result: Ad-Hoc Topic, Comparison

Title queries	50%			70%			90%		
	MAP	P20	J20	MAP	P20	J20	MAP	P20	J20
Full index	.253	.464	—	.253	.464	—	.253	.464	—
2N2P test	.239	<u>.467</u>	.714	.203	.434	.535	.076	.248	.198
Popularity-based	.223	.417	.780	.189	.365	.574	.077	.161	.199
Uniform	.231	.445	.760	.187	.376	.566	.110	.241	.273
Term-based, $k = 10$	.218	.457	**.853**	.187	.441	**.675**	.109	.311	.350
Document-centric	<u>**.253**</u>	<u>**.478**</u>	.743	.244	<u>**.485**</u>	.602	**.198**	<u>.465</u>	.325
KL	.234	<u>.465</u>	.826	.210	.461	.664	.143	.357	.360
χ^2-divergence	.245	<u>.474</u>	.799	.232	<u>.467</u>	.668	.181	.437	**.373**
Renyi, $\alpha = 50$	.252	<u>.476</u>	.743	.244	<u>**.485**</u>	.603	**.198**	<u>.467</u>	.325
Renyi, $\alpha \to \infty$	<u>**.253**</u>	<u>**.478**</u>	.741	**.245**	<u>**.485**</u>	.598	**.198**	<u>**.468**</u>	.323

- In general, document-centric $>$ term-based, 2N2P $>$ others.
- *Document-centric is competitive to our best test run.*
- Test runs works better on precision.

Result: Ad-Hoc Topic, Comparison

SD queries	50%			70%			90%		
	MAP	P20	J20	MAP	P20	J20	MAP	P20	J20
Full index	.264	.491	—	.264	.491	—	.264	.491	—
2N2P test	.242	.481	.722	.204	.442	.537	.076	.249	.188
Popularity-based	.232	.439	.781	.198	.375	.581	.080	.170	.194
Uniform	.238	.461	.755	.192	.389	.576	.111	.246	.262
Term-based, $k = 10$	.223	.474	**.852**	.188	.451	.664	.107	.312	.320
Document-centric	**.259**	**.499**	.743	.248	.507	.588	**.200**	.472	.306
KL	.240	.476	.842	.211	.470	**.678**	.137	.340	.337
χ^2-divergence	.252	.487	.824	.234	.481	.677	.180	.441	**.354**
Renyi, $\alpha = 50$	.258	.498	.750	.248	.506	.592	**.200**	.472	.306
Renyi, $\alpha \to \infty$	**.259**	.498	.740	**.249**	**.508**	.584	**.200**	**.474**	.303

- On SD queries, numbers are higher but still follow the same trend.
- Pruning can benefit P20 at level 50% and 70%.

ANOVA

	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

For testing significance, we ran a 4-way ANOVA upfront followed by a Tukey's HSD test. All effects in ANOVA come back significant for $p < 0.001$. The reported effect size is partial eta-squared.

Tukey's HSD

MAP	Mean	Grp	P20	Mean	Grp
Rényi, $\alpha \to \infty$	.2419	a....	Rényi, $\alpha \to \infty$	.4865	a...
Document-centric	.2416	a....	Document-centric	.4858	a...
Rényi, $\alpha = 50$	.2415	a....	Rényi, $\alpha = 50$	.4853	a...
χ^2-divergence	.2318	.b...	χ^2-divergence	.4709	a...
KL	.2130	..c..	KL	.4434	.b..
Popularity-based	.2073	..cd.	Term-based	.4278	.bc.
Uniform	.2034	...de	2N2P test	.4123	..cd
2N2P test	.1959	....e	Uniform	.3991	...d
Term-based	.1949	....e	Popularity-based	.3940	...d

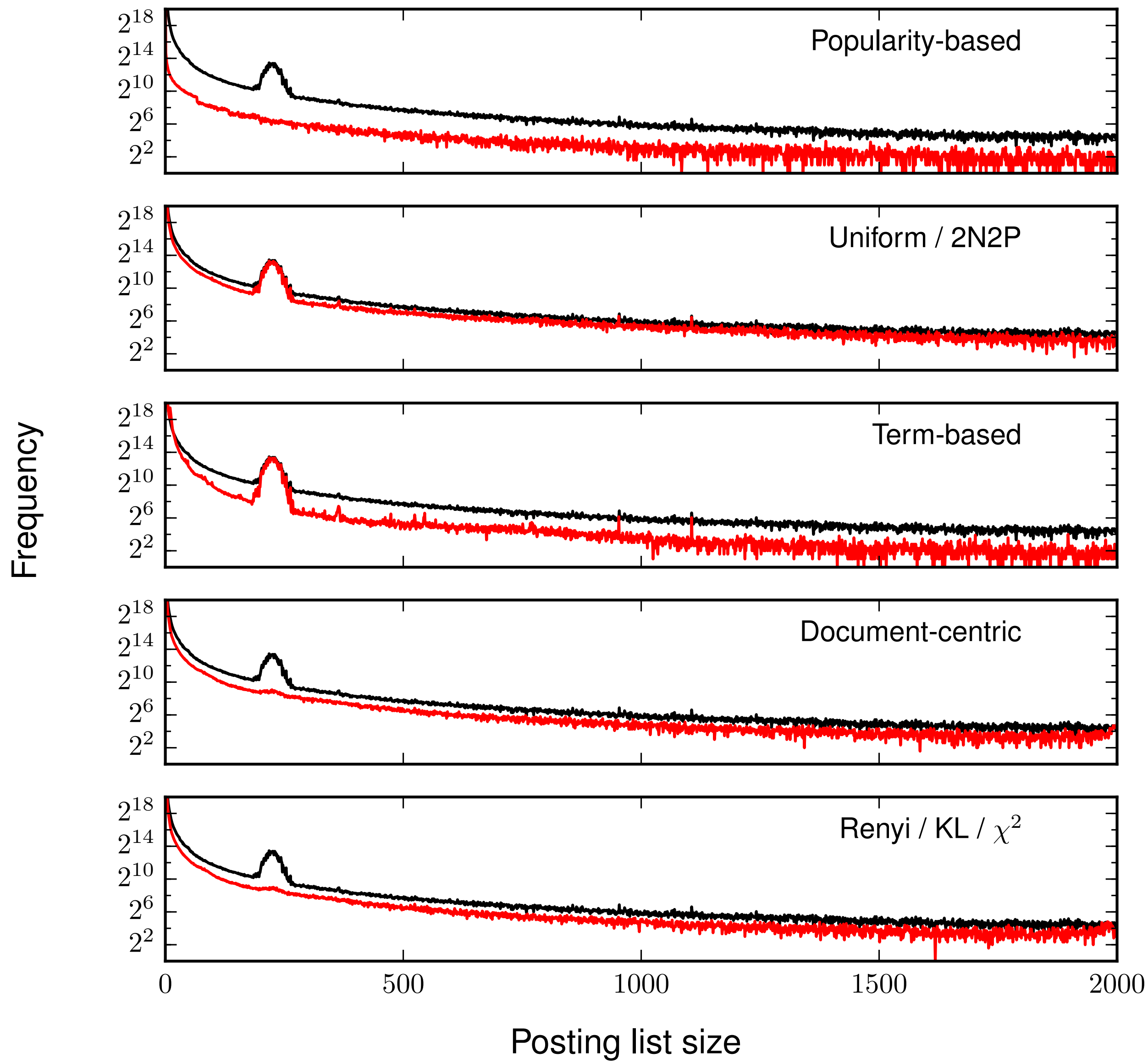
Rényi divergence appears to have a slight advantage over document-centric pruning, but the improvement is not significant.

Result: Efficiency Topics

T = time (sec); PT = pruning time (sec);
PL Kept = fraction of terms kept; Avg Size = average posting list size

Efficiency	50%		70%		90%		Index Status at 90%		
	J20	T	J20	T	J20	T	PT	PL Kept (%)	Avg Size
Full index	—	990	—	990	—	990	—	100.0%	128.6
2N2P	.605	366	.426	148	.128	15	2858	100.0%	12.9
PB	**.772**	815	.515	644	.182	209	**2383**	0.6%	2126.4
Uniform	.646	**272**	.450	**107**	.178	**6**	3189	55.4%	23.2
TB	.753	640	**.563**	419	**.296**	138	2695	100.0%	12.7
DC	.639	549	.487	311	.235	129	6987	40.9%	31.8
KL	.730	546	.538	325	.235	86	6541	36.0%	35.8
χ^2-div	.707	623	.546	318	.251	103	6767	37.9%	34.0
Renyi$_{\alpha=50}$	.642	511	.490	307	.236	128	8240	40.4%	31.9
Renyi$_{\infty}$	.637	551	.484	347	.233	130	6830	40.6%	31.7

Timing experiments conducted on a dedicated server with a 3.30 GHz Intel Core i5-2500 CPU (4 cores) and 16GB RAM.



Popularity-based
Uniform / 2N2P
Term-based
Document-centric
Renyi / KL / χ^2
Frequency
2^{18}
2^{14}
2^{10}
2^{6}
2^{2}
0
500
1000
1500
2000
Posting list size

Comparison

Document-centric pruning[1] (top) vs. Rényi$_\infty$ (bottom)

for $d \in \mathcal{D}$ **do**
    Sort terms in descending order of $\mathrm{BM25}(t, d)$
    for $k = 1, \ldots, n$ **do**
        Remove posting $(t_{[k]}, d)$ if $(n - k + 1)/n < \rho$

for $d \in \mathcal{D}$ **do**
    Sort terms in descending order of $p(t|d)$;
    for $k = 1, \ldots, n$ **do**
        Remove posting $(t_{[k]}, d)$ if $(\sum_{i=1}^{k} p(t_{[i]}|d))^{-1} < \epsilon$;

[1] Büttcher and Clarke. (2006). "A document-centric approach to static index pruning in text retrieval systems". CIKM '06.

Summary

i) *Experiment results are in line with theory. Generalization (e.g., choice of divergence, cardinality) does help.*

ii) *Keeping documents "accessible" can be important.*

iii) *J20 does not align well with top-k precision.*

Questions:

- Multiple test collections
- Does pruning remove stopwords?
- Estimation of $p(t|d)$
- Retrievability

Takeaway Messages

Document-centric pruning and Rényi$_\infty$ are empirically the best. Whether they are related is still an open question.

Now, we have a problem where "theory and application collides".

- If you are into application, **try my package**.

  `github.com/rueycheng/indri-pruning-toolkit`

- If you are big on theory, please stare at this for 30 seconds.

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

Give us some feedback, would you?

Any question?

Thanks for your attention.