

# On divergence measures and static index pruning

**Ruey-Cheng Chen**  
RMIT University

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



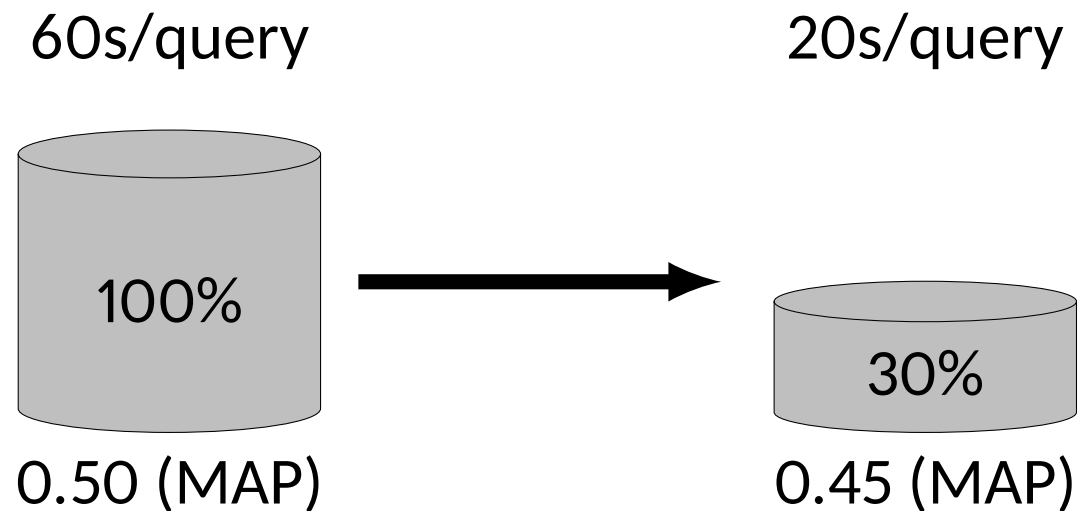
Sep 29, 2015 (ICTIR '15)

# 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.<sup>1</sup>
  
- ii) Retrieval is slow, so you use a cache to serve top results.**
  - Tiered indexing for Web search.<sup>2</sup>
  
- 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, ...*)

---

<sup>1</sup>Carmel et al. (2001). "Static index pruning for information retrieval systems". SIGIR '01.

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

---

<sup>1</sup>Carmel et al. (2001). “Static index pruning for information retrieval systems”. SIGIR '01.

<sup>2</sup>Büttcher and Clarke. (2006). “A document-centric approach to static index pruning in text retrieval systems”. CIKM '06.

<sup>3</sup>Blanco and Á. Barreiro. (2007). “Static Pruning of Terms in Inverted Files”. ECIR '07.

<sup>4</sup>Ntoulas and Cho. (2007). “Pruning policies for two-tiered inverted index with correctness guarantee”. SIGIR '07.

<sup>5</sup>Skobeltsyn et al. (2008). “ResIn: a combination of results caching and index pruning for high-performance web search engines”. SIGIR '08.

<sup>6</sup>Zheng and Cox. (2009). “Entropy-Based Static Index Pruning”. ECIR '09.

<sup>7</sup>Blanco and A. Barreiro. (2010). “Probabilistic static pruning of inverted files”. *ACM Transactions on Information Systems*.

<sup>8</sup>Thota and Carterette. (2011). “Within-Document Term-Based Index Pruning with Statistical Hypothesis Testing”. ECIR '11.

<sup>9</sup>Chen et al. (2012). “Information preservation in static index pruning”. CIKM '12.

<sup>10</sup>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<sup>1</sup>

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:

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

---

<sup>1</sup>Kullback. (1959). *Information Theory and Statistics*.



## “Index” Version<sup>1</sup> 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:

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

---

<sup>1</sup>Chen and Lee. (2013). “An Information-Theoretic Account of Static Index Pruning”. SIGIR '13.

## “Index” Version<sup>1</sup> 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:

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

*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.

---

<sup>1</sup>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  $T_1, T_2, \dots, T_n$  from the discrete distribution  $\theta_D$  that represents the language model for document  $D$ .

$$D \sim \text{Uniform}(1, |\mathcal{D}|),$$

$$T_k \sim \text{Discrete}(\theta_D) \quad \text{for } k = 1 \dots n.$$

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  $\rho$ , choose an index  $q$  so as to:

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

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

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

(We originally tackled this problem.)

## Ingredient #2: Problem Formulation

Given an index  $p$  and a prune ratio  $\rho$ , choose an index  $q$  so as to:

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

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

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

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



## Ingredient #2: Problem Formulation

Given an index  $p$  and a prune ratio  $\rho$ , choose an index  $q$  so as to:

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

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

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

We call  $n$  the *query cardinality*.

## Ingredient #3: Divergence Measures

We generalize the choice of divergence measures.

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

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

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

# $f$ -Divergence<sup>1,2</sup>

A family of measures parametrized by the functional  $f$ .

$$D_f(q||p) = \sum_{t_{1:n}, d} p(t_{1:n}, d) f\left(\frac{q(t_{1:n}, d)}{p(t_{1:n}, d)}\right), \quad (5)$$

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

---

<sup>1</sup>Csiszár and Shields. (2004). "Information Theory and Statistics: A Tutorial". *FnT in Communications and Information Theory*.

<sup>2</sup>Morimoto. (1963). "Markov Processes and the H-Theorem". *Journal of the Physical Society of Japan*.

# Rényi Divergence of Order $\alpha$ <sup>1</sup>

Another well-known family parametrized by  $\alpha$ .

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

Kullback-Leibler divergence  $\alpha \rightarrow 1$

Logarithm of  $\chi^2$ -divergence  $\alpha = 2$

---

<sup>1</sup>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<sup>1</sup>

One can actually take  $\alpha$  to infinity:

$$D_{\infty}(q||p) = \log \sup_{t_{1:n}, d} \frac{q(t_{1:n}, d)}{p(t_{1:n}, d)}. \quad (7)$$

---

<sup>1</sup>Erven and Harremoës. (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  $n = 2, 3, \dots$  and so on.
  - Check if the problem can be reduced to smaller  $n$ .

## Analytic Form: $n = 1$

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

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

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

## Analytic Form: $n > 1$

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

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

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

# Gain Functions

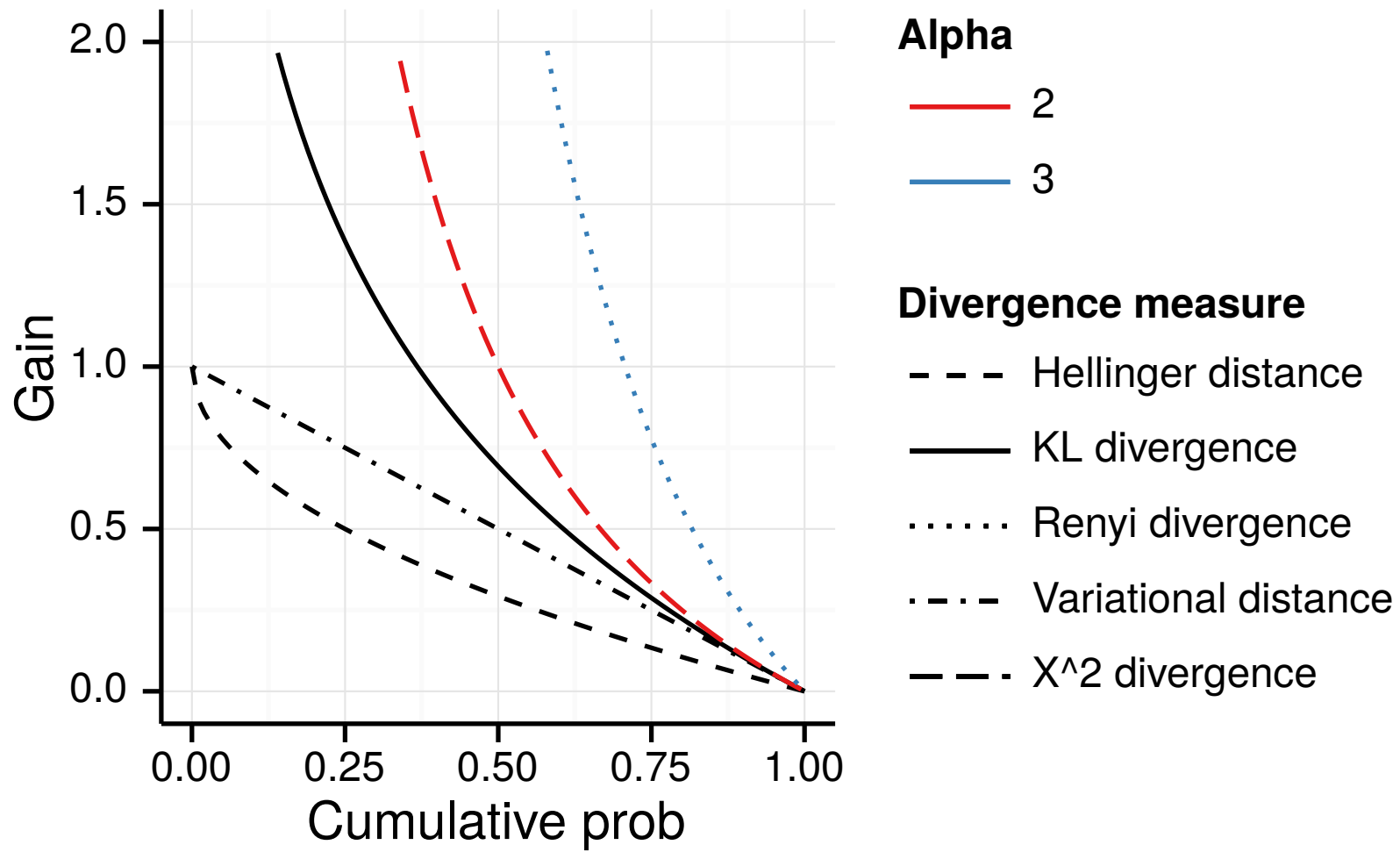
All these measures except Rényi<sub>∞</sub> have similar analytic forms:

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

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              | $1 - x^{1/2}$           |
| $\chi^2$ -divergence              | $x^{-1} - 1$            |
| Rényi divergence ( $\alpha > 1$ ) | $x^{1-\alpha} - 1$      |

# Gain Functions



# Allocation for One Document

Objective:

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

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  $(t_{[1]}, d), (t_{[2]}, d), \dots, (t_{[k-1]}, d)$  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:

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

This algorithm computes optimal allocation in  $O(|\mathcal{D}| \log n)$  time.<sup>1,2</sup>

- 1 **for**  $d \in \mathcal{D}$  **do**
- 2     Sort terms in descending order of  $p(t|d)$  ;
- 3     **for**  $k = 1, \dots, n$  **do**
- 4         Compute  $\Delta(t_{[k]}, d)$  according to (10) ;
- 5         Remove posting  $(t_{[k]}, d)$  if  $|\Delta(t_{[k]}, d)| < \epsilon$  ;

---

<sup>1</sup>Fox. (1966). "Discrete optimization via marginal analysis". *Management science*.

<sup>2</sup>Ibaraki and Katoh. (1988). *Resource Allocation Problems: Algorithmic Approaches*.

# Optimal Allocation: Variants

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

- 1 **for**  $d \in \mathcal{D}$  **do**
- 2     **for**  $t \in \text{posting}(d)$  **do**
- 3         Remove posting  $(t, d)$  if  $p(d)p(t|d) < \epsilon$ ;



# Optimal Allocation: Variants

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

- 1 **for**  $d \in \mathcal{D}$  **do**
- 2     **for**  $t \in \text{posting}(d)$  **do**
- 3         Remove posting  $(t, d)$  if  $p(d)p(t|d) < \epsilon$ ;

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

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

- *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
- $D(q||p)$  vs.  $D(p||q)$
- Jensen-Shannon divergence
- Smoothing integrated into  $Q(p)$

# 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  $p(t|d)$ :

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

- 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 =  $\rho$ )

| Title queries                      | 50%                |                    |             | 70%         |                    |             | 90%         |                    |             |
|------------------------------------|--------------------|--------------------|-------------|-------------|--------------------|-------------|-------------|--------------------|-------------|
|                                    | MAP                | P20                | J20         | MAP         | P20                | J20         | MAP         | P20                | J20         |
| Full index                         | .253               | .464               | —           | .253        | .464               | —           | .253        | .464               | —           |
| KL                                 | .234               | <u>.465</u>        | <b>.826</b> | .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        |
| $\chi^2$ -div                      | .245               | <u>.474</u>        | .799        | .232        | <u>.467</u>        | <b>.668</b> | .181        | .437               | <b>.373</b> |
| Rényi, $\alpha = 50$               | .252               | <u>.476</u>        | .743        | .244        | <u>.485</u>        | .603        | <b>.198</b> | <u>.467</u>        | .325        |
| Rényi, $\alpha \rightarrow \infty$ | <b><u>.253</u></b> | <b><u>.478</u></b> | .741        | <b>.245</b> | <b><u>.485</u></b> | .598        | <b>.198</b> | <b><u>.468</u></b> | .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  $\chi^2$ -div work better

## Result: Ad-Hoc Topic, Comparison

| <i>Title queries</i>               | 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               | <b>.853</b> | .187        | .441               | <b>.675</b> | .109        | .311               | .350        |
| Document-centric                   | <b><u>.253</u></b> | <b><u>.478</u></b> | .743        | .244        | <b><u>.485</u></b> | .602        | <b>.198</b> | <u>.465</u>        | .325        |
| KL                                 | .234               | <u>.465</u>        | .826        | .210        | .461               | .664        | .143        | .357               | .360        |
| $\chi^2$ -divergence               | .245               | <u>.474</u>        | .799        | .232        | <u>.467</u>        | .668        | .181        | .437               | <b>.373</b> |
| Renyi, $\alpha = 50$               | .252               | <u>.476</u>        | .743        | .244        | <b><u>.485</u></b> | .603        | <b>.198</b> | <u>.467</u>        | .325        |
| Renyi, $\alpha \rightarrow \infty$ | <b><u>.253</u></b> | <b><u>.478</u></b> | .741        | <b>.245</b> | <b><u>.485</u></b> | .598        | <b>.198</b> | <b><u>.468</u></b> | .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

| <i>SD queries</i>                  | 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        | <b>.852</b> | .188        | .451        | .664        | .107        | .312        | .320        |
| Document-centric                   | <b>.259</b> | <u>.499</u> | .743        | .248        | <u>.507</u> | .588        | <b>.200</b> | .472        | .306        |
| KL                                 | .240        | .476        | .842        | .211        | .470        | <b>.678</b> | .137        | .340        | .337        |
| $\chi^2$ -divergence               | .252        | .487        | .824        | .234        | .481        | .677        | .180        | .441        | <b>.354</b> |
| Renyi, $\alpha = 50$               | .258        | <u>.498</u> | .750        | .248        | <u>.506</u> | .592        | <b>.200</b> | .472        | .306        |
| Renyi, $\alpha \rightarrow \infty$ | <b>.259</b> | <u>.498</u> | .740        | <b>.249</b> | <u>.508</u> | .584        | <b>.200</b> | <b>.474</b> | .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      | $\eta_p^2$ |
|-----|-------------|-----|--------|------------|
| 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, $\alpha \rightarrow \infty$ | .2419 | a . . . . | Rényi, $\alpha \rightarrow \infty$ | .4865 | a . . . |
| Document-centric                   | .2416 | a . . . . | Document-centric                   | .4858 | a . . . |
| Rényi, $\alpha = 50$               | .2415 | a . . . . | Rényi, $\alpha = 50$               | .4853 | a . . . |
| $\chi^2$ -divergence               | .2318 | . b . . . | $\chi^2$ -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 |

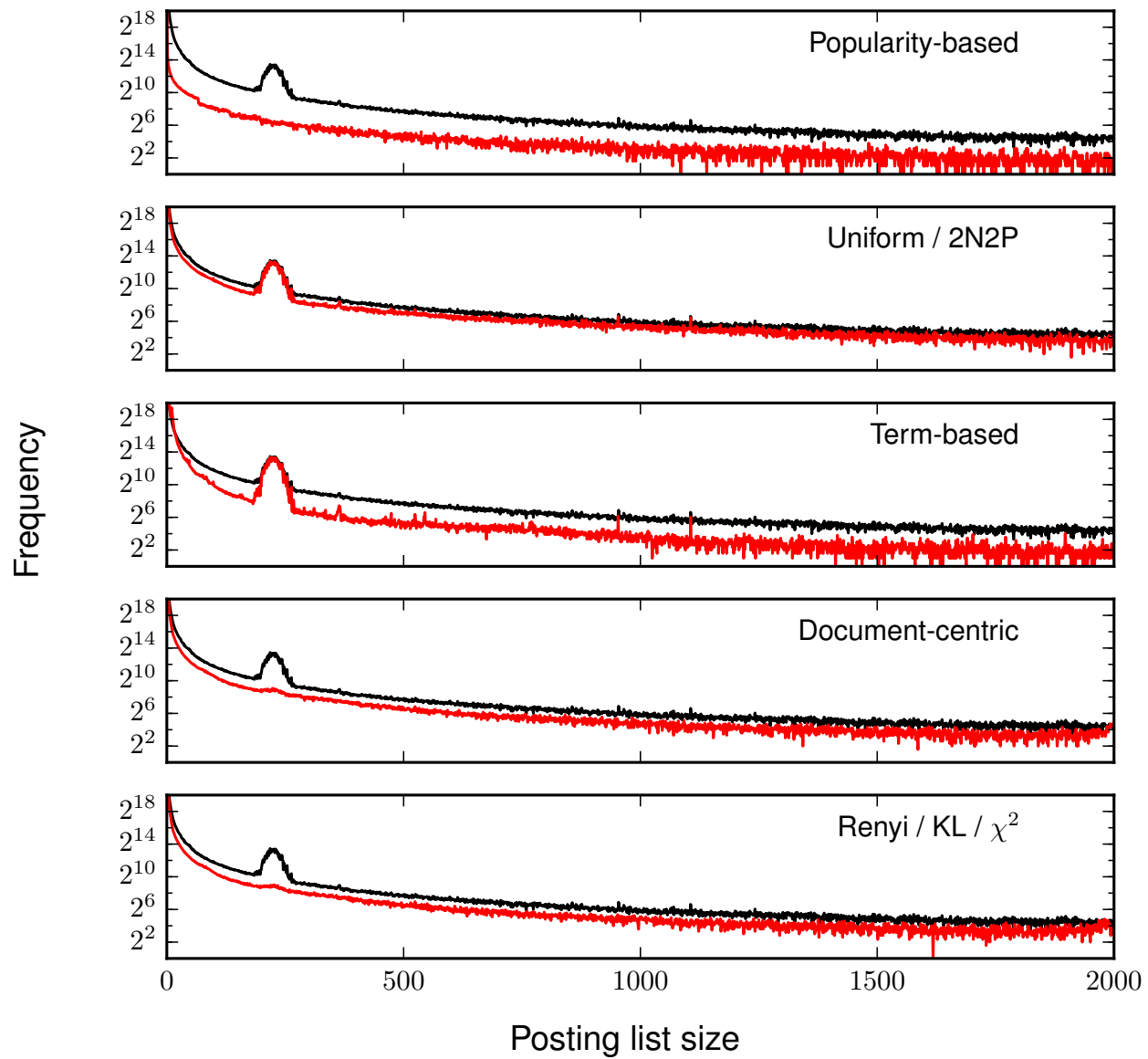
Rényi divergence appears to have a slight advantage over document-centric pruning, but the improvement is not significant.

## 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                                      | <b>.772</b> | 815        | .515        | 644        | .182        | 209      | <b>2383</b>         | 0.6%        | 2126.4   |
| Uniform                                 | .646        | <b>272</b> | .450        | <b>107</b> | .178        | <b>6</b> | 3189                | 55.4%       | 23.2     |
| TB                                      | .753        | 640        | <b>.563</b> | 419        | <b>.296</b> | 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     |
| $\chi^2$ -div                           | .707        | 623        | .546        | 318        | .251        | 103      | 6767                | 37.9%       | 34.0     |
| Renyi <sub><math>\alpha=50</math></sub> | .642        | 511        | .490        | 307        | .236        | 128      | 8240                | 40.4%       | 31.9     |
| Renyi <sub><math>\infty</math></sub>    | .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.



# Comparison

Document-centric pruning<sup>1</sup> (top) vs. Rényi<sub>∞</sub> (bottom)

```
1 for  $d \in \mathcal{D}$  do  
2   Sort terms in descending order of  $\text{BM25}(t, d)$   
3   for  $k = 1, \dots, n$  do  
4     Remove posting  $(t_{[k]}, d)$  if  $(n - k + 1)/n < \rho$ 
```

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

---

<sup>1</sup>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<sub>∞</sub> 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](https://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.