An Information-Theoretic Account of Static Index Pruning

Ruey-Cheng Chen[†] and **Chia-Jung Lee**[‡]

[†]National Taiwan University
rueycheng@turing.csie.ntu.edu.tw

[‡]University of Massachusetts Amherst
cjlee@cs.umass.edu

SIGIR 2013 (Dublin, Ireland)

Outline

- Introduction
- Proposed Method
  - Minimum Cross-Entropy and Static Index Pruning
  - Uniform Pruning
- Evaluation
- Discussion

Introduction

Motivation

Static index pruning:

– Reduce the index size by removing its entries.

– Improve disk usage and query throughput.

Also a model induction problem.

– Goal: **Induce a pruned index** (a subset of the original one).

– But the predictive power varies for every possible choice.

How do we find the best pruned model?

Principle of Minimum Cross-Entropy

Suppose one has some initial hypothesis about a system and seeks to update this measurement incrementally. Kullback[1] suggested choosing a measure q that most closely resembles the previous measurement p in the sense of Kullback-Leibler divergence.

(Given a prior measure p and a set of feasible measures $\mathcal{F}$)

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

So, let us write static index pruning in this form and solve this problem. Are we done yet?

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

Result 1: Rediscovery of Uniform Pruning

Analytically solving this problem is hard because that involves combinatorial optimization. Derivation is complicated and tricky. We used weak inference techniques and surrogate modeling to tackle this beast.

End result (called *uniform pruning*):

$$\text{maximize} \quad \sum_{t,d} \mathbb{I}_{t,d} p(t|d), \tag{2}$$

where $\mathbb{I}_{t,d}$ is an indicator (1 = keep this entry, 0 = lose it).

But, uniform pruning is *not* a new invention.

Result 2: Uniform Pruning is State of the Art

A very short history of uniform pruning:

- First appearance in 2001: as a baseline method for TCP[1].
- Second in 2013: this paper.

What happened?

- Lack of experimental control (on prune ratio.)
- Then we could not employ any form of significance tests.

Within a revised experimental design, our result suggests that uniform pruning is state of the art.

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

Uniform Pruning

So, what is uniform pruning anyway?

```
Require: $\epsilon$
for all $t \in T$ and $d \in$ postings$(t)$ do
    Compute $A(t, d) = score(t, d)$
    if $A(t, d) < \epsilon$ then
        Remove $d$ from postings$(t)$
    end if
end for
```

The function $score(t, d)$ is usually called *impact*. It is the partial contribution of the retrieval score from term t to document d.

Related Work

Static index pruning:

Impact[1,2], (term) informativeness and discriminative value[3], (document) entropy[4], probability ranking principle[5], two-sample two-proportion (2P2N)[6], information preservation[7], query-view-based approach[8].

[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 Barreiro. (2007). “Static Pruning of Terms in Inverted Files”. Lecture Notes in Computer Science.

[4] Zheng and Cox. (2009). “Entropy-Based Static Index Pruning”. Lecture Notes in Computer Science.

[5] Blanco and Barreiro. (2010). “Probabilistic static pruning of inverted files”. *ACM Transactions on Information Systems*.

[6] Thota and Carterette. (2011). “Within-Document Term-Based Index Pruning with Statistical Hypothesis Testing”. Lecture Notes in Computer Science.

[7] Chen et al. (2012). “Information preservation in static index pruning”. CIKM '12.

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

Minimum Cross-Entropy and Static Index Pruning

Basics

An index entry (or *posting*) is of the form:

$$(t, d, n),$$

where $t \in T$, $d \in D$, and $n \in \mathbf{N}_+$ (positive integers). It means **term** t **appears** n **times in document** d.

An inverted index is a probabilistic model:

$$p(D|T;\theta),$$

where θ is a set of index entries. It is *nonparametric*.

Problem Definition

Given a full index θ_0, induce a pruned model θ such that:

$$\begin{array}{ll} (1) & \theta \subseteq \theta_0, \\ (2) & |\theta|/|\theta_0| = 1 - \rho, \text{for some } 0 < \rho < 1. \end{array}$$

An implicit objective is to minimize performance loss. Write as a constrained optimization problem (the hypothetical $g(\theta)$ computes retrieval performance):

$$\begin{array}{ll} \text{maximize} & g(\theta) \\ \text{subject to} & \theta \subseteq \theta_0 \\ & |\theta|/|\theta_0| \text{ reaches } 1 - \rho. \end{array} \tag{3}$$

Principle of Minimum Cross-Entropy

Use the negative KL divergence in place of $g(\cdot)$. (Many tools in model induction can apply!)

$$\begin{array}{ll} \text{minimize} & \mathsf{D}(\theta || \theta_0) \\ \text{subject to} & \theta \subseteq \theta_0 \\ & |\theta| / |\theta_0| \text{ reaches } 1 - \rho. \end{array} \tag{4}$$

Write out the objective in full:

$$\begin{aligned} \mathsf{D}(\theta || \theta_0) &\equiv \mathsf{D}(p(D|T) || p_0(D|T)) \\ &\equiv \sum_{t,d} p(d,t) \log \frac{p(d|t)}{p_0(d|t)}. \end{aligned} \tag{5}$$

Assumptions

Assumption 1 (Query Model)**.**

$$p(d,t) = \overbrace{p(d|t)}^{index}\,\overbrace{q(t)}^{query},$$
$$p_0(d,t) = p_0(d|t)\,q(t).$$

Assumption 2 (Normalization Factor)**.** *Let* $\mathbb{I}_{t,d}$ *be an indicator for whether index entry* (t,d,n) *is retained in the induced model. Then we write the (induced) likelihood as:*

$$p(t|d) \equiv \mathbb{I}_{t,d}\,p_0(t|d)/Z_d,$$

where Z_d *is the normalization factor for document* d*.*

Key step: Let Z_d be a positive constant for all $d \in D$.

Analysis

Use Assumption 1 to dissect the joint distribution $p(d, t)$. Apply Bayes Theorem to $p(d|t)$ and $p_0(d|t)$. Now, with uniform $p(d)$ and $p_0(d)$, we have:

$$\sum_t p(t) \sum_d \frac{p(t|d)}{\sum_{d'} p(t|d')} \log \frac{p(t|d)}{p_0(t|d)} \frac{\sum_{d'} p_0(t|d')}{\sum_{d'} p(t|d')}. \qquad (6)$$

Replace $p(t|d)$ using the definition in Assumption 2. Note that all the normalization factors $(= k)$ all cancel out.

$$\sum_t p(t) \sum_d \frac{\mathbb{I}_{t,d} p_0(t|d)}{\sum_{d'} \mathbb{I}_{t,d'} p_0(t|d')} \log \mathbb{I}_{t,d} \frac{\sum_{d'} p_0(t|d')}{\sum_{d'} \mathbb{I}_{t,d'} p_0(t|d')}. \qquad (7)$$

Analysis (Cont'd)

Organize by dividing the support of the inner summation:

$$\sum_t p(t) \sum_{d:\mathbb{I}_{t,d}=1} \frac{p_0(t|d)}{\sum_{d'} \mathbb{I}_{t,d'} p_0(t|d')} \log \frac{\sum_{d'} p_0(t|d')}{\sum_{d'} \mathbb{I}_{t,d'} p_0(t|d')}. \qquad (8)$$

The innermost logarithm does not depend on d anymore. Moving it out of the summation, we find the summation cancels out:

$$\sum_t p(t) \log \frac{\sum_{d'} p_0(t|d')}{\sum_{d'} \mathbb{I}_{t,d'} p_0(t|d')}. \qquad (9)$$

When minimizing this equation, we can get rid of the numerator, i.e., $\sum_{d'} p_0(t|d')$, in the logarithm because it does not depend any combinatorial choice we make.

Surrogate Modeling

The end result is a convex integer program. But solving it exactly is not possible (i.e., too many index entries).

$$\begin{array}{ll} \text{maximize} & \sum_t p_0(t) \log \sum_d \mathbb{I}_{t,d} p_0(t|d) \\ \text{subject to} & \mathbb{I}_{t,d} \text{ is binary, for all } (t,d,\cdot) \in \theta_0, \\ & \sum_{t,d} \mathbb{I}_{t,d} = (1-\rho)|\theta_0|. \end{array} \tag{10}$$

Idea: Use Jensen's inequality to induce a surrogate objective function.

Jensen's Inequality

For any concave function f, we have:

$$\mathbb{E}f(X) \leq f(\mathbb{E}X).$$

So, sit the original objective at the left hand:

$$\sum_t p_0(t) \log \sum_d \mathbb{I}_{t,d} p_0(t|d) \leq \log \sum_{t,d} \mathbb{I}_{t,d} p_0(t) p_0(t|d)$$

The resulting maximization problem is written equivalently as:

$$\text{maximize} \quad \sum_{t,d} \mathbb{I}_{t,d} p_0(t) p_0(t|d). \tag{11}$$

Uniform Pruning

Keeping the top $(1-\rho)N$ term-document entries according to weighted query likelihood, i.e., $p(t)p(t|d)$, guarantees to maximize the objective.

$$\begin{array}{ll} \text{maximize} & \sum_{t,d} \mathbb{I}_{t,d} p_0(t) p_0(t|d), \\ \text{subject to} & \mathbb{I}_{t,d} \text{ is binary, for all } (t,d,\cdot) \in \theta_0, \\ & \sum_{t,d} \mathbb{I}_{t,d} = (1-\rho)|\theta_0|. \end{array} \tag{12}$$

Thus far, we have rediscovered *uniform pruning*:

– The definition matches Carmel et al. when $p(t)$ is uniform.

– Here, query likelihood loosely equals to impact.

Algorithm

```
Require: $\epsilon$
for all $t \in T$ and $d \in$ postings($t$) do
    Compute $A(t,d)$ using Equation (13)
    if $A(t,d) < \epsilon$ then
        Remove $d$ from postings($t$)
    end if
end for
```

Here:

$$A(t,d) = \begin{cases} p(t|d) & \text{for language models} \\ score(t,d) & \text{otherwise} \end{cases} . \qquad (13)$$

Evaluation

Experimental Setup

Benchmark:

Collection	# Documents	Query Topics
Disks 4 & 5	528k	401-450
WT2G	247k	401-450
WT10G	1692k	451-550

Tested two query types, title (t) and title+desc (td). Use Indri toolkit[1], with porter stemmer and the standard 401 InQuery stoplist. Use BM25 as the post-pruning retrieval method.

Proposed methods:

– UP-bm25, UP-dir ($\mu = 2500$), and UP-jm ($\lambda = 0.6$).

[1] http://www.lemurproject.org/indri.php

Experimental Setup (Cont'd)

Baseline methods:

– Top-k term-centric pruning, $k = 10$ with BM25 (TCP).

– Probability ranking principle, $\lambda = 0.6$ (PRP):

$$\frac{p(r|t,d)}{p(\overline{r}|t,d)} \equiv \frac{p(t|D)p(r|D)}{p(t|\overline{r})(1-p(r|D))}.$$

– Information preservation, $\lambda = 0.6$ with uniform document prior (IP-u):

$$-\frac{p(t|d)}{\sum_{d'} p(t|d')} \log \frac{p(t|d)}{\sum_{d'} p(t|d')}.$$

Did not implement document-length update.

Prune Ratio

Comparisons are made only at 9 prune levels at $\rho = 0.1, 0.2, \ldots, 0.9$. Here, we detail two approaches for controlling prune ratio.

– **Sample percentile**: Take a sample of index entries and calculate the prune score. Use the percentile estimates[1] to determine the right cutting threshold.

– **Bisection**: Run a binary search within the interval of feasible parameter values $[a, b]$.

We applied bisection to TCP to learn ϵ, and sample percentile to the rest of methods. All the prune ratio error is controlled to within $\pm 0.2\%$.

[1] Hyndman and Fan. (1996). "Sample Quantiles in Statistical Packages". *The American Statistician*.

Design

Fixed-effect, 4-way no interaction, repeated measure design:

$$Y_{i,j,k,l} = a_i + b_j + c_k + d_l + \epsilon_{i,j,k,l},$$

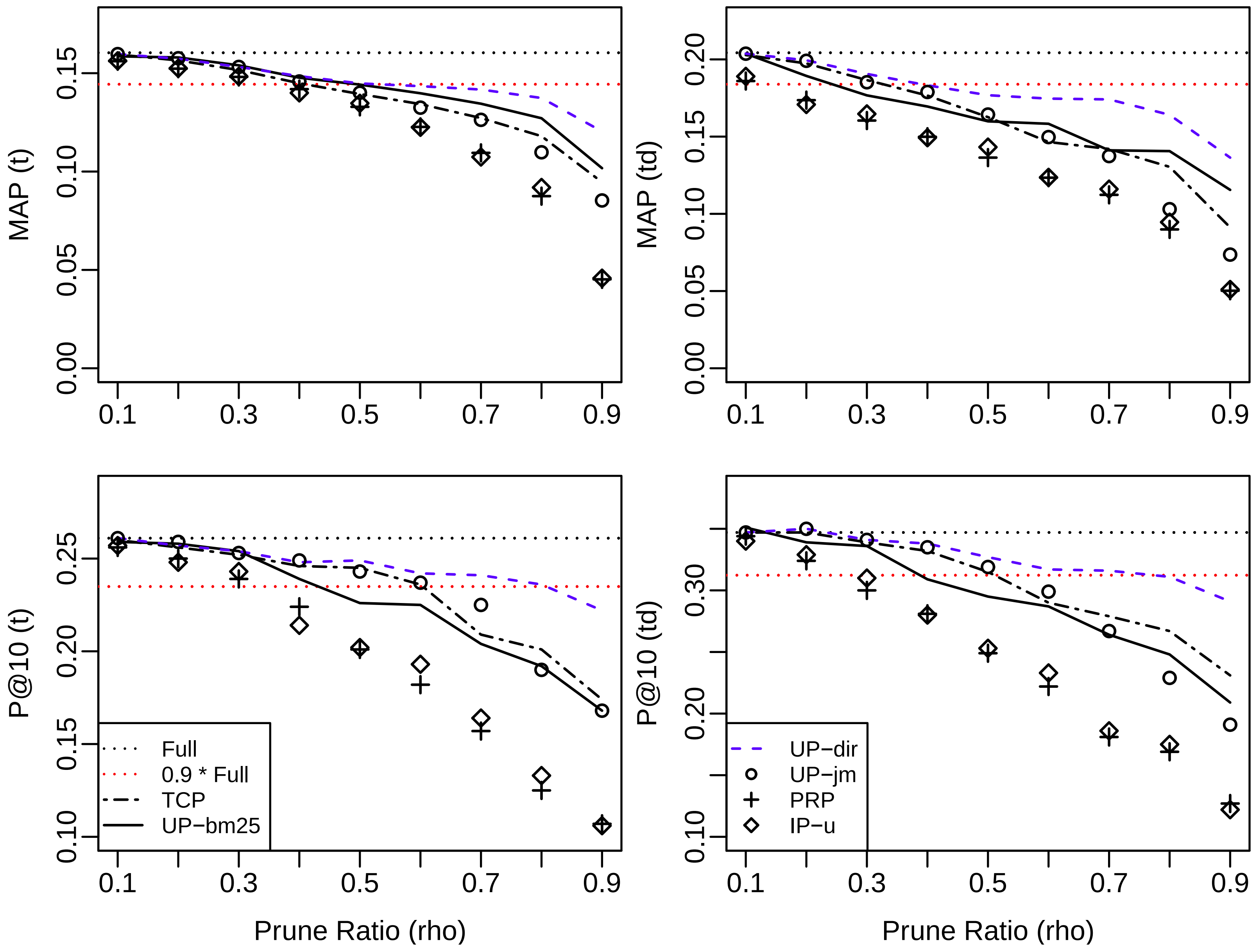
where $Y_{i,j,k,l}$ is the measured performance, a_i, b_j, c_k, and d_l are the four main effects (query type, prune ratio, method, and topic), and $\epsilon_{i,j,k,l}$ is the error.

Easy to incorporate many data points. Robust to non-normality.

Procedures:

– Conduct the omnibus test.

– If significant, run post-hoc tests on the method effect.


MAP (t)
0.00
0.05
0.10
0.15
MAP (td)
0.00
0.05
0.10
0.15
0.20
P@10 (t)
0.10
0.15
0.20
0.25
P@10 (td)
0.10
0.20
0.30
0.1
0.3
0.5
0.7
0.9
Prune Ratio (rho)
Prune Ratio (rho)
Full
0.9 * Full
TCP
UP−bm25
UP−dir
UP−jm
PRP
IP−u

Analysis of Variance

		Disks 4 & 5			WT2G			WT10G		
Resp.	Effect	DF	F	η_p^2	DF	F	η_p^2	DF	F	η_p^2
MAP	QT	F(1,5336)	74.10	.01	F(1,5336)	42.57	.01	F(1,10686)	192.25	.02
	PR	F(8,5336)	240.30	.26	F(8,5336)	306.17	.31	F(8,10686)	193.26	.13
	M	F(5,5336)	11.00	.01	F(5,5336)	40.20	.04	F(5,10686)	61.47	.03
	T	F(49,5336)	885.35	.89	F(49,5336)	335.89	.76	F(49,10686)	422.46	.80
P@10	QT	F(1,5336)	66.16	.01	F(1,5336)	10.89	.00	F(1,10686)	622.34	.06
	PR	F(8,5336)	105.00	.14	F(8,5336)	133.98	.17	F(8,10686)	122.43	.08
	M	F(5,5336)	20.34	.02	F(5,5336)	44.06	.04	F(5,10686)	71.01	.03
	T	F(49,5336)	484.06	.82	F(49,5336)	296.88	.73	F(49,10686)	226.31	.68

The 4-way no-interaction ANOVA result. Each cell indicates a combination of performance measure (row) and test collection (column). For each effect, we report the degrees of freedom, F-value, and the effect size (measured using η_p^2.)

Here, all the effects are significant for $p < 0.001$ in every combination.

Method Effect

	Disks 4 & 5			WT2G			WT10G		
	Method	Mean	Group	Method	Mean	Group	Method	Mean	Group
MAP	UP-bm25	.204	a..	UP-dir	.223	a...	UP-dir	.162	a..
	UP-dir	.200	a..	UP-bm25	.211	.b..	UP-bm25	.151	.b.
	TCP	.196	ab.	TCP	.204	.b..	TCP	.148	.b.
	UP-jm	.191	.bc	UP-jm	.192	..c.	UP-jm	.145	.b.
	PRP	.187	..c	IP-u	.181	...d	IP-u	.129	..c
	IP-u	.187	..c	PRP	.179	...d	PRP	.127	..c
P@10	UP-dir	.433	a..	UP-dir	.404	a...	UP-dir	.286	a..
	TCP	.433	a..	TCP	.385	ab..	TCP	.268	.b.
	UP-jm	.424	a..	UP-jm	.367	.bc.	UP-jm	.265	.b.
	UP-bm25	.417	a..	UP-bm25	.359	..c.	UP-bm25	.259	.b.
	PRP	.392	.b.	IP-u	.322	...d	IP-u	.222	..c
	IP-u	.389	.b.	PRP	.319	...d	PRP	.219	..c

The result of Tukey's HSD test on the method effect. In each combination, the pruning methods are sorted based on their means and tested for group difference.

Common group labels means insignificant performance difference.

Discussion

Impact is a good approximate to per-entry information.

- Impact-sorted indexes with early termination heuristics[1].
- Impact-based dynamic pruning[2].

Why does Dirichlet smoothing work better?

1. BM25 might be a poor approximation to $p(t|d)$.
2. Parameter optimization was lacking.

No-depletion constraint: “avoid draining any term posting list.” But does it matter practically?

[1] Anh et al. (2001). “Vector-space ranking with effective early termination”. SIGIR ’01.

[2] Anh and Moffat. (2006). “Pruned query evaluation using pre-computed impacts”. SIGIR ’06.

Conclusion

We proposed a model-based induction framework to static index pruning. Under suitable assumptions, we can write static index pruning as a convex program. This program has a simple surrogate model—uniform pruning.

We proposed a controlled experiment design for static index pruning.

Uniform pruning is state of the art.

- Significantly superior than all the others in Web-scale settings
- Robust to large prune ratio
- Efficient
- UP-dir retains $\geq 85\%$ of baseline performance at 80%

Thanks for your attention

Any question?

Sponsored by ACM SIGIR and Singhal Foundation from Google