Information Preservation in Static Index Pruning

Ruey-Cheng Chen[†], Chia-Jung Lee[*], Chiung-Min Tsai[†], and Jieh Hsiang[†]
[†]National Taiwan University, Taiwan [*]University of Massachusetts Amherst, MA

An Information-Theoretic Argument

The inverted index $p(d|t)$ is essentially a nonparametric *predictive model*. Removing index entries from this model may inevitably incur a loss in its predictive power. Here, we seek to answer the following two questions:

- How do we estimate the predictive power of an inverted index?
- Under a given prune ratio, how do we minimize the loss in predictive power?

Predictive Power

We estimate the predictive power of $p(d|t)$ using the conditional entropy $H(D|T)$:

$$\sum_{t \in T} p(t) \left(- \sum_{d \in D} p(d|t) \log p(d|t) \right), \qquad (1)$$

where $p(t)$ denotes the probability of term t being used in a query, and $p(d|t)$ assesses the relevance between document d and term t.

Approximation

Assuming that $p(t)$ is uniform, we estimate the contribution of a term-document pair to the overall predictive power (denoted as $A(t, d)$) using this equation:

$$- \frac{p(t|d)p(d)}{\sum_{d'} p(t|d')p(d')} \log \frac{p(t|d)p(d)}{\sum_{d'} p(t|d')p(d')}. \qquad (3)$$

Here, $p(t|d)$ and $p(d)$ denote the query likelihood and the document prior. The following algorithm computes an approximate solution to Equation (2). This simple maneuver guarantees to retain the most predictive power with respect to a specific choice of ϵ.

Require: ϵ: a threshold value
1: **for all** $t \in T$ **do**
2: **for all** $d \in$ postings(t) **do**
3: Compute $A(t, d)$ using Equation (3)
4: **if** $A(t, d) < \epsilon$ **then**
5: Remove d from postings(t)
6: **end if**
7: **end for**
8: **end for**

Optimization Problem

Let θ_0 denote the set of term-document pairs in the index. Under a given prune ratio ρ, we maximize the predictive power of the model:

$$\begin{array}{ll} \text{maximize} & H_\theta(D|T) \\ \text{subject to} & \theta \subseteq \theta_0 \\ & |\theta|/|\theta_0| \text{ reaches } \rho. \end{array} \qquad (2)$$

Nevertheless, this combinatorial optimization is intractable in general.

Probability Estimation

We follow Blanco and Barreiro [1] in estimating these probabilities:

$$p(t|d) = (1 - \lambda) p_{\mathrm{ML}}(t|D) + \lambda p(t|C), \qquad (4)$$

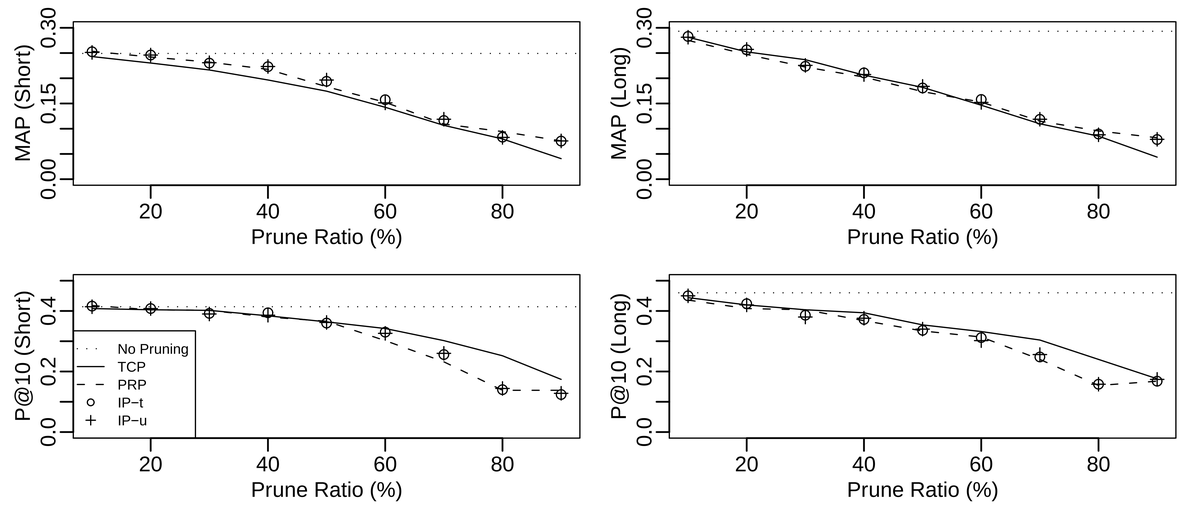
$$p(d) = \frac{1}{2} + \frac{1}{10} \tanh \frac{dl - \overline{X_d}}{S_d}. \qquad (5)$$

We set $\lambda = 0.6$ in Equation (4). For the document prior $p(d)$, we experimented with two approaches: (i) *hyperbolic-tangent approximation* (denoted as IP-ht, as in Equation (5)) and (ii) *uniform prior*, i.e., $p(d) = 1/|D|$ (denoted as IP-u).

Experimental Setup

Experiments were conducted on LATimes, TREC-8, and WT2g. We used TREC topics 401-450 as queries, and used BM25 in post-pruning retrieval. Two baseline approaches, TCP [2] ($k = 10$) and PRP [1], were implemented using the Indri API. We do not update document length values after pruning.

Result on WT2g



Short Query (MAP/P@10 at 0%: 249/414) Long Query (MAP/P@10 at 0%: 293/460)

MAP	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
TCP	243	230	216	197	174	142	107	080	041	281	252	237	206	182	147	110	085	044
PRP	254▲	242▲	232	218	183	152	109	094	076▲	275	247	222	202	173	153	115	096	082▲
IP-ht	253▲	246▲	230	223▲	194	158	116♯	083	075▲	283♯	256♯	224	211	181	158	119	089	079▲
IP-u	251▲	246▲	231	223▲	197	151	119♯	083	076▲	281	257♯	226	207	184	152	119	088♭	079▲

P@10	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
TCP	408	404	402	384	364	342	302	252	174	444	420	404	394	354	332	304	240	176
PRP	418	404	402	380	366	302	232▼	138▼	138	436	408	404	368	334	314	240▼	154▼	168
IP-ht	416	408	392	394	360	330♯	256	140▼	124▼	450	424	386	372	336	312	248▼	158▼	168
IP-u	414	408	390	386	362	326	260♯	144▼	128▼	450	420	380	376	340	302	256▼	158▼	174

The result is given in both figural and tabular formats for different settings of query types and measures. Measured performance (y-axis) for each method is plotted against prune ratio (x-axis); numbers are given in the corresponding table. Statistical significance is assessed using two-tailed paired t-test for $p < 0.05$, denoted using superscripts ▲ and ▼ (against TCP) and subscripts ♯ and ♭ (against PRP.)

Retrieval Performance The performance for IP-based methods is comparable to that for PRP and TCP. In general, we find it difficult to assert that any of these methods is better than the others. What is worth noting is that TCP does slightly worse in MAP but better in P@10, which suggests that IP-based methods favor more on recall. This trend is more pronounced in short queries.

Efficiency TCP has an overhead in sorting term postings. IP-u relies only on query likelihood estimates and is therefore more efficient to compute than PRP.

References

[1] R. Blanco and A. Barreiro. Probabilistic static pruning of inverted files. *ACM Transactions on Information Systems*, 28(1), Jan. 2010.
[2] 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 the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 43–50, New York, NY, USA, 2001. ACM.