Information Preservation in Static Index Pruning

Ruey-Cheng Chen[†], Chia-Jung Lee[‡], Chiung-Min Tsai[†], and Jieh Hsiang[†]

[†]National Taiwan University
1 Roosevelt Rd. Sec. 4
Taipei 106, Taiwan

[‡]University of Massachusetts
140 Governors Drive
Amherst, MA 01003-9264

rueycheng@turing.csie.ntu.edu.tw, cjlee@cs.umass.edu
cmtsai@mail.lis.ntu.edu.tw, jhsiang@ntu.edu.tw

ABSTRACT

We develop a new static index pruning criterion based on the notion of information preservation. This idea is motivated by the fact that model degeneration, as does static index pruning, inevitably reduces the predictive power of the resulting model. We model this loss in predictive power using conditional entropy and show that the decision in static index pruning can therefore be optimized to preserve information as much as possible. We evaluated the proposed approach on three different test corpora, and the result shows that our approach is comparable in retrieval performance to state-of-the-art methods. When efficiency is of concern, our method has some advantages over the reference methods and is therefore suggested in Web retrieval settings.

Categories and Subject Descriptors

H.1.1 [**Systems and Information Theory**]: Information theory; H.3.1 [**Content Analysis and Indexing**]: Indexing methods; H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness)

General Terms

Experimentation, Performance

Keywords

Information Retrieval, Index Pruning, Conditional Entropy

1. INTRODUCTION

Static index pruning is a technique that seeks to reduce index size during or immediately after index construction. The reduction is achieved by permanently removing unwanted index entries, i.e., *postings*, from a production retrieval system. At the cost of sacrificing some degree of retrieval accuracy, this practice has been shown to enhance both disk usage and query throughput [4]. To date, static index pruning has gathered much attention for its implication to search efficiency over Web-scale text collections [2, 3].

To minimize the effect of index pruning on retrieval accuracy, many previous efforts prioritize the index entries according to their impact on the retrieval result. Carmel et al. proposed using raw retrieval scores, such as tf-idf or BM25, to measure the importance of a posting [4]. Büttcher and Clarke measured the usefulness of a posting (t, d) based on the contribution of term t to the Kullback-Leibler divergence score between document d and the entire collection [3]. Blanco and Barreiro used the odd-ratio of relevance in probability ranking principle (PRP) [5] as the decision criterion [2]. Alternative criteria other than impact have also been investigated, such as document-centric entropy-based pruning [6], and informativeness and discriminative value [1]. These measures have been shown useful in specific query scenarios.

In this paper, we discuss the idea of *information preservation* and use that to motivate a new decision measure for static index pruning. Consider that an inverted index is essentially a nonparametric predictive model $p(d|t)$, with which one estimates the likelihood of some document d being relevant to some query term t. Pruning this model permanently removes the connections between some terms and some documents, and thus causes a loss in predictive power. We propose using the conditional entropy $H(D|T)$ to quantify the predictive power and minimizing the loss with respect to the choice of pruned entries.

2. INFORMATION PRESERVATION

Information retrieval is a practice about ranking documents in response to information needs. To achieve optimal performance, documents shall be retrieved in order of the decreasing probability of relevance [5]. This notion of relevance can be realized in many different ways. For simplicity, here we consider a simple term relevance model $p(d|t)$ that assesses the probability of document d being relevant to some query term t. The model $p(d|t)$ is a nonparametric *predictive model*, in a sense that prediction is made over the choice of documents with respect to the textual input from the user. In static index pruning, the problem that we face is to preserve as much predictive power in $p(d|t)$ during the course, in spite of a considerable amount of information will be discarded afterwards.

We propose using the conditional entropy $H(D|T)$ to quantify the predictive power in $p(d|t)$. The conditional entropy is a summary statistic regarding how difficult it is to predict the right outcome D (document) given the predictor T

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

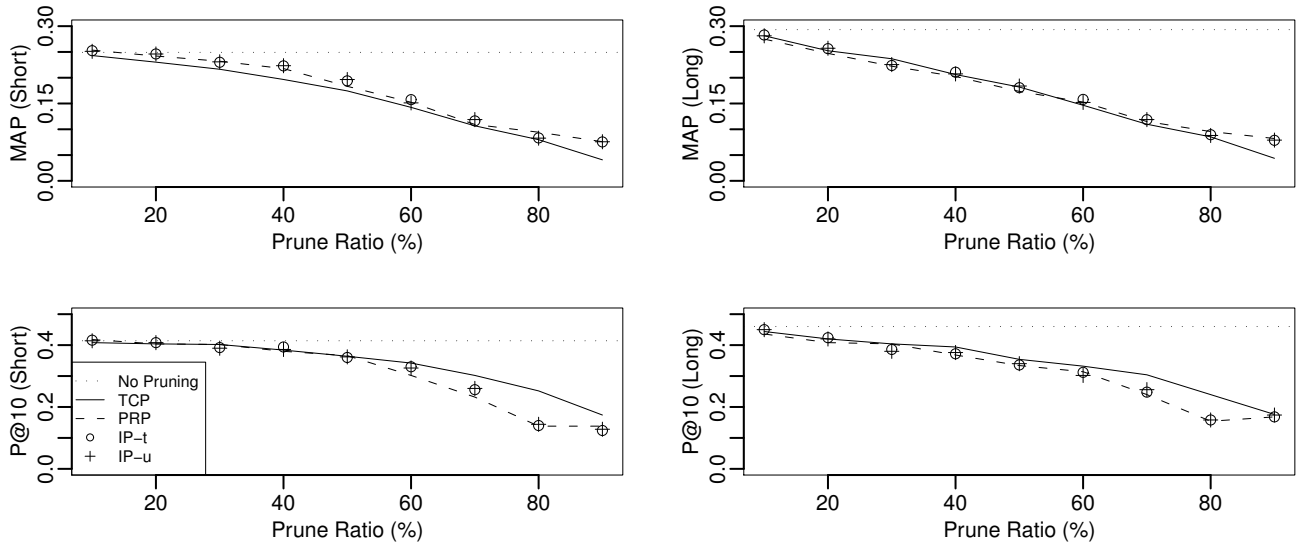

Figure 1: Performance results for all the methods on WT2G. Rows indicate different performance measures (MAP/P@10). Columns indicate different query types (short/long).

(term). Generally, it is written as:

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

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

The distribution $p(t)$ is independent of the retrieval model in use. To estimate $p(t)$, we simply assume that it is uniformly distributed. Note that this estimate can be further improved by using session logs. Now, we write $H(D|T)$ as a summation of *uncertainties* $A(t,d)$ contributed by individual term-document pairs to the model:

$$H(D|T) = \frac{1}{|T|} \sum_{t \in T} \sum_{d \in D} A(t,d), \qquad (2)$$

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

Consider any two term-document pairs (t,d) and (t',d') such that $A(t,d) < A(t',d')$. The formulation implies that predicting d from t requires less information than predicting d' from t', meaning that we are more certain about the connection for t and d. In this case, removing (t,d) from the index has less effect on the overall predictive power than removing (t',d'). By setting a cutting threshold ϵ, it is now straightforward to scan over the entire index and discard any entry whose uncertainty $A(t,d)$ is strictly lower than ϵ. This simple maneuver guarantees to retain the most predictive power with respect to a specific choice of ϵ.

3. EXPERIMENTS

3.1 Setup

We implemented two baseline approaches using the Indri API[1]: top-k term-centric pruning (denoted as TCP) and probability ranking principle (denoted as PRP). Our implementation does not update the document length values after pruning. For TCP, we set $k = 10$ to maximize the precision for the top 10 documents [4] and used BM25 as the score function. For PRP, we set $\lambda = 0.6$ for query likelihood estimation, and applied the suggested approximations to estimate the rest of the probabilities [2]. These probability estimates are summarized as follows.

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

$$p(r|D) = \frac{1}{2} + \frac{1}{10} \tanh \frac{dl - \overline{X_d}}{S_d}, \qquad (5)$$

$$p(t|\overline{r}) = p(t|C). \qquad (6)$$

To reduce the effect of retrieval method, a similar setting was adopted for the proposed method. We used Equation (4) to estimate the query likelihood $p(t|d)$ (setting $\lambda = 0.6$). For estimating document prior $p(d)$, we experimented with two approaches, which are *hyperbolic-tangent approximation* as in Equation (5) (denoted as IP-ht) and *uniform prior*, i.e., $p(d) = 1/|D|$ (denoted as IP-u).

We managed to control the prune ratio at different levels (e.g., 10%, 20%, ..., 90%.) For PRP and IP-based methods, the prune ratio depends on a global threshold ϵ. To prune the index to the right size, we sampled the decision scores from the entire index to estimate the percentiles, and then used the estimates to find the right threshold value. For TCP, we manually adjusted the parameter ϵ to approach the designated prune ratio. In our experiments, the error is controlled to roughly $\pm 0.2\%$ in prune ratio.

3.2 Retrieval Performance

We conducted a series of experiments on the LATimes, TREC-8, and WT2G corpora, using TREC topics 401-450 as queries. We tested two different query types, short (using title) and long (using title and description), using BM25 as the retrieval function. Performance is evaluated using mean average-precision (MAP) and precision-at-10 (P@10).

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

(a) LATimes

	Short Query (MAP/P@10 at 0%: 210/250)									Long Query (MAP/P@10 at 0%: 235/260)								
MAP	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
TCP	209	206	201	193	177	168	143	124	073	232	230	217	206	174	171	153	116	075
PRP	211	209	207	201$^{\blacktriangle}$	190	158	141	113	098	228	221	206	202	193	160	141	119	106
IP-ht	210	210	207	203$^{\blacktriangle}$	188	161	148	109	097	232	221	215$_{\sharp}$	207	187	160	144	116	103
IP-u	210	210	207	203$^{\blacktriangle}$	191$^{\blacktriangle}$	164	151	104	090	234	221	216	204	188	168	149	113	098

P@10	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
TCP	252	244	246	238	228	218	204	194	128	262	264	256	242	238	234	212	188	142
PRP	254	256	254	248	234	212	172	124$^{\blacktriangledown}$	130	264	256	242	252	248	228	174	132$^{\blacktriangledown}$	132
IP-ht	250	256$^{\blacktriangle}$	258	254	234	218	168	110$^{\blacktriangledown}$	130	258	254	254	244	244	222	164	122$^{\blacktriangledown}$	130
IP-u	250	256$^{\blacktriangle}$	258	250	236	224	184	116$^{\blacktriangledown}$	138	256	256	258	244	242	232	182	118$^{\blacktriangledown}$	136

(b) TREC-8

	Short Query (MAP/P@10 at 0%: 228/436)									Long Query (MAP/P@10 at 0%: 256/478)								
MAP	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
TCP	223	213	204	191	176	154	126	094	055	249	239	230	209	188	166	136	103	064
PRP	226	221	215	201	181	157	143	147$^{\blacktriangle}$	103$^{\blacktriangle}$	251	239	221	204	179	161	143	147$^{\blacktriangle}$	120$^{\blacktriangle}$
IP-ht	227$^{\blacktriangle}_{\sharp}$	223$^{\blacktriangle}$	215$^{\blacktriangle}$	202	186$_{\sharp}$	160	147	147$^{\blacktriangle}$	106$^{\blacktriangle}$	251	238	222	207	185	161	143	144$^{\blacktriangle}$	123$^{\blacktriangle}$
IP-u	227$^{\blacktriangle}$	223$^{\blacktriangle}$	216$^{\blacktriangle}$	203	187$_{\sharp}$	163	143	145$^{\blacktriangle}$	106$^{\blacktriangle}$	250	238	223	208	185	164	142	141$^{\blacktriangle}$	124$^{\blacktriangle}$

P@10	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
TCP	436	434	428	430	432	388	338	288	188	476	478	480	456	464	436	376	322	188
PRP	438	442	456$^{\blacktriangle}$	432	414	378	296	276	202	490	486	472	440	408$^{\blacktriangledown}$	376	324	294	232
IP-ht	436	440	444$_{\flat}$	442	422	388	298	288	210	478	488	460	454	410$^{\blacktriangledown}$	388	344	300	238
IP-u	440	442	442$_{\flat}$	444$_{\sharp}$	424	388	302	288	202	482	484	462	452	410$^{\blacktriangledown}$	388	342	286	228

(c) 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$^{\blacktriangle}$	242$^{\blacktriangle}$	232	218	183	152	109	094	076$^{\blacktriangle}$	275	247	222	202	173	153	115	096	082$^{\blacktriangle}$
IP-ht	253$^{\blacktriangle}$	246$^{\blacktriangle}$	230	223$^{\blacktriangle}$	194	158	116$_{\sharp}$	083	075$^{\blacktriangle}$	283$_{\sharp}$	256$_{\sharp}$	224	211	181	158	119	089	079$^{\blacktriangle}$
IP-u	251$^{\blacktriangle}$	246$^{\blacktriangle}$	231	223$^{\blacktriangle}$	197	151	119$_{\sharp}$	083	076$^{\blacktriangle}$	281	257$_{\sharp}$	226	207	184	152	119	088$_{\flat}$	079$^{\blacktriangle}$

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$^{\blacktriangledown}$	138$^{\blacktriangledown}$	138	436	408	404	368	334	314	240$^{\blacktriangledown}$	154$^{\blacktriangledown}$	168
IP-ht	416	408	392	394	360	330$_{\sharp}$	256	140$^{\blacktriangledown}$	124$^{\blacktriangledown}$	450	424	386	372	336	312	248$^{\blacktriangledown}$	158$^{\blacktriangledown}$	168
IP-u	414	408	390	386	362	326	260$_{\sharp}$	144$^{\blacktriangledown}$	128$^{\blacktriangledown}$	450	420	380	376	340	302	256$^{\blacktriangledown}$	158$^{\blacktriangledown}$	174

Table 1: The overall performance results on three test corpora. All the reported measures are round down to the 3rd digit under the decimal point. For brevity, preceding zeroes and decimal points are ignored. Underlined entries indicate the best performance in the corresponding group. Entries that are significantly superior or inferior ($p < 0.05$) to TCP are denoted by superscripts ▲ or ▼, respectively. Analogously, entries that are significantly superior or inferior to PRP are denoted by subscripts ♯ or ♭, respectively.

The performance result is given in both figural and tabular formats. Figure 1 summarizes the evaluation result on WT2G[2] in four plots, each indicating a different combination of query type and performance measure. Each method is plotted as a curve or a series of points according to the measured performance (y-axis) at some prune ratio (x-axis). The full detail is covered in Table 1, which stresses more on performance differences between methods. Statistical significance in this respect is assessed using two-tailed paired t-test for $p < 0.05$. We use superscripts (▲ and ▼) and subscripts (♯ and ♭) in Table 1 to highlight these entries.

The result shows that the performance for IP-based methods is generally comparable to that for PRP. No consistent pattern is observed across all settings to assess one method is better than the others. Significant difference in either MAP or P@10 between IP-based methods and PRP is detected for 10 out of 54 experimental runs, among which IP-based methods are shown superior to PRP in 8 runs (denoted as ♯ in Table 1). PRP significantly outperforms only for short queries on TREC-8 at 30% and long queries on WT2G at 80% (denoted as ♭), but the latter result is inconsistent across performance measures.

It is interesting to note that TCP is generally doing slightly worse than the rest of methods in MAP but slightly better in P@10, which suggests that IP-based methods favor more on recall. This trend is observed across different corpora and

[2]Results on the other two corpora show similar trends and are therefore omitted here.

	TCP	PRP	IP-ht	IP-u
TCP	–	0.332	0.665	0.661
PRP	0.332	–	0.282	0.281
IP-ht	0.665	0.282	–	0.998
IP-u	0.661	0.281	0.998	–

Table 2: Correlation analysis for the decision measures on the LATimes corpus. The correlation is estimated using Pearson's product-moment correlation coefficient, weighted using term frequencies of index entries.

experimental settings, and is more amplified in the short query case. Comparing IP-based methods with TCP in all 54 runs, we find that IP-based methods significantly outperform TCP in 14 (denoted as ▲), and TCP significantly outperforms IP-based methods in 6 (denoted as ▼). The case we have observed for long queries on WT2G at 90% prune ratio is difficult to interpret: TCP performs significantly worse in MAP but does better in P@10.

3.3 Correlation Analysis

Our experimental result gives rise to an interesting question that whether different pruning methods lead to different prioritization over index entries. To investigate the effect of pruning methods in this respect, we conducted a simple correlation analysis on the LATimes corpus. For each index entry, we retrieved the decision scores produced by all four algorithms and compiled them into a tuple. We collected totally 36,497,224 such tuples. For each pair of methods, we computed Pearson's product-moment correlation coefficient, weighted using term frequencies of index entries.

The result, which is summarized in Table 2, shows that the decision scores produced by two IP-based methods are strongly correlated (0.998). In this case, we conclude that uniform prior is more favorable than hyperbolic-tangent approximation in real-world settings, since the former is easier to compute. IP-based methods also show medium correlation (0.661 and 0.665) with TCP, slightly stronger than that (0.332) with PRP. We want to point out here that, since the decision score used in TCP corresponds to BM25, IP-based scores lean more toward BM25 in terms of the effect on index entry prioritization. This can be useful in some other information retrieval applications.

4. CONCLUDING REMARKS

In this paper, we develop the notion of information preservation in the context of static index pruning, and use this idea to motivate a new decision criterion. In the experiments conducted on three different test corpora, the proposed method shows consistent, competitive performance to state-of-the-art methods. So far, there is only minor evidence to interpret the performance differences between the proposed approach and the reference methods for specific cases. We expect this to be made clear with further experimentation.

Our approach has a few advantages in terms of efficiency. First, term-centric pruning has an overhead in computing the cutting threshold for each term, since a sorting algorithm is involved to order the postings in terms of their impact values. Our decision measures do not suffer from this issue. Second, computation for the PRP measure depends on three different probability estimates, while the proposed IP-u measure (uniform prior) relies on only the query likelihood. In our experiments, it took roughly 146 seconds for PRP to scan through all the entries in the TREC-8 corpus; IP-u did the same thing in only 126 seconds. We believe that this 20-second difference can be far more amplified on a Web-scale setting.

There are many ways to extend this work. One possible direction that we have in mind is to combine weakly-correlated measures, such as ours and PRP. Given the correlation analysis result, we believe that doing so is feasible and can be beneficial. Moreover, this study also provides an alternative viewpoint toward prioritization of index entries. Impact and uncertainty are intrinsically two different concepts, while in this very application our result somehow closes the gap in between. This connection may lead to a new postulation for information retrieval. We hope that our efforts will invite further investigation into these interesting issues.

5. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments. The research effort described in this paper is supported under the National Taiwan University Digital Archives Project (Project No. NSC-98-2631-H-002-005), which is sponsored by National Science Council, Taiwan.

6. REFERENCES

[1] R. Blanco and A. Barreiro. Static pruning of terms in inverted files advances in information retrieval. In G. Amati, C. Carpineto, and G. Romano, editors, *Advances in Information Retrieval*, volume 4425 of *Lecture Notes in Computer Science*, chapter 9, pages 64–75. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2007.

[2] R. Blanco and A. Barreiro. Probabilistic static pruning of inverted files. *ACM Transactions on Information Systems*, 28(1), Jan. 2010.

[3] S. Büttcher and C. L. A. Clarke. A document-centric approach to static index pruning in text retrieval systems. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 182–189, New York, NY, USA, 2006. ACM.

[4] 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.

[5] S. Robertson. The probability ranking principle in IR. In K. S. Jones and P. Willett, editors, *Reading in Information Retrieval*, chapter The probability ranking principle in IR, pages 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[6] L. Zheng and I. J. Cox. Entropy-Based static index pruning. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, chapter 72, pages 713–718. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2009.