

Measure-based Metasearch

Javed A. Aslam*, Virgiliu Pavlu, Emine Yilmaz
College of Computer and Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115
{jaa,vip,emine}@ccs.neu.edu

ABSTRACT

We propose a simple method for converting many standard measures of retrieval performance into metasearch algorithms. Our focus is both on the analysis of retrieval measures themselves and on the development of new metasearch algorithms. Given the conversion method proposed, our experimental results using TREC data indicate that system-oriented measures of overall retrieval performance (such as average precision) yield good metasearch algorithms whose performance equals or exceeds that of benchmark techniques such as CombMNZ and Condorcet.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Retrieval models*

General Terms

Theory, Algorithms, Experimentation

Keywords

Metasearch, Retrieval Evaluation

1. INTRODUCTION

Metasearch is the well-studied process of fusing the ranked lists of documents returned by a collection of systems in response to a given user query in order to obtain a combined list whose quality equals or exceeds that of any of the underlying lists. Many metasearch techniques have been proposed and studied, and for the purposes of comparison, we consider two benchmark techniques in this work: CombMNZ and Condorcet. CombMNZ [1, 2] is based on combining the normalized scores given to each document by the underlying systems, while Condorcet [3] is based on viewing the metasearch problem as a multi-candidate election where the documents are candidates and the systems are voters expressing preferential rankings among the candidates.

Retrieval systems are evaluated using a number of standard measures of performance such as average precision, R-precision, and precisions at various cutoffs. These evaluation

*We gratefully acknowledge the support provided by NSF grant CCF-0418390.

measure implicitly assign *weights* to the relevances of documents at various ranks. For example, precision-at-cutoff 10, $PC(10)$, implicitly assigns a weight of $1/10$ to the relevances of each of the top 10 documents in a list and a weight of 0 to the relevances of all remaining documents, and as such $PC(10)$ can be calculated by multiplying the weights implicitly associated with each document by their 0-1 relevances.

We now consider two synergistic facts: (1) evaluation measures aim to assess how well a system retrieves relevant documents, as measured and evaluated by the aforementioned implicit weights and (2) retrieval systems aim to retrieve relevant documents as well as possible. As such, our hypothesis is that evaluation measures will assign “high” weights to relevant documents when applied to the lists generated by “good” retrieval systems. Thus, evaluation measures can be used to identify likely relevant documents in a list, as determined by the measure’s implicit weights. Applying such a measure to many lists and combining the weights assigned to documents appropriately, one can assign “consensus” weights to the documents collectively retrieved in multiple lists, rank these documents by their consensus weights, and thus obtain a metasearch list.

2. METHODOLOGY

We now formalize the ideas presented above to convert standard measures of retrieval performance to metasearch algorithms.

Precisions at standard cutoffs. Consider precision-at-cutoff k , $PC(k)$, for any integer k . $PC(k)$ implicitly assigns a weight of $1/k$ to each of the top k documents in a list and a weight of 0 to every remaining document. Given multiple lists, a document may be assigned multiple weights, depending on how the document is ranked in each of the underlying lists. To obtain a *consensus* weight for a document, one can simply compute the average weight assigned to the document across the underlying lists. To obtain a consensus metasearch list, one can then simply rank the documents according to these average scores (breaking ties arbitrarily).

R-precision. By definition, R-precision is $PC(R)$, where R is the total number of relevant documents for the query. As such, one can convert R-precision to a metasearch algorithm as described above. However, such a conversion does not yield a true metasearch algorithm because it demands *a priori* knowledge of R . In practice, one would need to estimate (or be given) R , a non-trivial task in a typical metasearch setting. However, we include R-precision in this discussion since it is an often cited and robust measure of overall retrieval performance.

| TREC | MNZ | COND | AP | RP | PC(5) | PC(10) | PC(15) | PC(20) | PC(30) | PC(50) | PC(100) | PC(200) | PC(500) | PC(1000) |
|------|------|------|-------------------|-------------------|-------|--------|--------|--------|--------|--------|---------|---------|---------|----------|
| 5 | .294 | .307 | .300 [•] | .308 | .254 | .265 | .275 | .280 | .284 | .288 | .294 | .285 | .258 | .237 |
| 6 | .341 | .315 | .344 [◦] | .357 [◦] | .271 | .292 | .305 | .315 | .322 | .335 | .341 | .334 | .319 | .312 |
| 7 | .320 | .308 | .333 [◦] | .334 [◦] | .283 | .295 | .304 | .311 | .319 | .327 | .331 | .321 | .305 | .301 |
| 8 | .350 | .343 | .370 [◦] | .366 [◦] | .254 | .286 | .304 | .313 | .330 | .351 | .357 | .343 | .323 | .305 |
| 9 | .351 | .348 | .345 | .353 | .266 | .309 | .314 | .303 | .316 | .320 | .323 | .310 | .294 | .259 |

Table 1: Mean average precision values for CombMNZ, Condorcet, and the metasearch algorithms corresponding to average precision, R-precision, and precisions at standard cutoff levels.

Average precision. Average precision does not yield implicit weights associated with ranks quite as obviously as do precisions-at-cutoffs and R-precision; however, one may compute such implicit weights as follows. By definition, average precision is the average of the precisions at all relevant documents. Given a list of documents, one normally assumes that the precisions at all unretrieved relevant documents are zero. As such, one can compute average precision as follows, where N is the length of the retrieved list, $rel(i)$ is the 0-1 relevance of the document at rank i , and R is the number of relevant documents for the query.

$$\begin{aligned}
AP &= \frac{1}{R} \cdot \sum_{i:rel(i)=1} PC(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^N rel(i) \cdot PC(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^N rel(i) \sum_{j=1}^i rel(j)/i \\
&= \frac{1}{R} \cdot \sum_{1 \leq j \leq i \leq N} \frac{1}{i} \cdot rel(i) \cdot rel(j)
\end{aligned}$$

Thus, average precision effectively assigns an implicit weight $\frac{1}{R \cdot i}$ to each *pair* of ranks (i, j) , for all $1 \leq j \leq i \leq N$. To compute the implicit weight associated with each rank r , we simply sum the weights associated with all pairs involving r , yielding

$$\begin{aligned}
\sum_{j=1}^r \frac{1}{R \cdot r} + \sum_{i=r+1}^N \frac{1}{R \cdot i} &= \frac{1}{R} \cdot \left(1 + \frac{1}{r+1} + \frac{1}{r+2} + \dots + \frac{1}{N}\right) \\
&= \frac{1}{R} \cdot (1 + H_N - H_r)
\end{aligned}$$

where H_k is the k -th harmonic number. Finally, we note that R is seemingly necessary to calculate these weights, yielding similar problems for metasearch as described above for R-precision. However, unlike the situation for R-precision where knowledge of R was *necessary* to determine which documents would receive a non-zero weight, here R simply acts as a uniform scaling factor applied to all weights. As such, it is unnecessary for metasearch (i.e., ranking) purposes: we simply weight each document at rank r in a list with the value $(1 + H_N - H_r)$, compute a consensus weight for each document by averaging the weights assigned over all lists, and rank the documents according to these average scores to obtain a metasearch list.

3. EXPERIMENTAL RESULTS

We tested the metasearch algorithms associated with average precision, R-precision, and precisions at standard cutoffs using data from TRECs 5, 6, 7, 8, and 9. For each metasearch algorithm, each TREC, and each of the 50 queries in that TREC, we used the metasearch algorithm in question to combine all of the lists submitted for that query in that TREC. We evaluated these metasearch lists using average precision and averaged these AP values across the queries in a TREC to obtain the mean average precision (MAP) values reported in Table 1. The table also contains MAP values

for the benchmark CombMNZ and Condorcet algorithms for the purposes of comparison.

We first note that the metasearch algorithms associated with average precision and R-precision outperformed those algorithms obtained from precision-at-cutoff k for any k . Our hypothesis is that system-oriented measures of overall retrieval performance tend to implicitly weight the complete set of relevant documents more highly than user-oriented measures such as precisions at standard cutoffs. Second, we note that the performance of the metasearch algorithms corresponding to average precision and R-precision often equals or exceeds the performance of benchmark techniques such as CombMNZ and Condorcet. The AP and RP results shown in Table 1 constitute statistically significant improvements¹ with respect to CombMNZ and Condorcet when labeled with a \bullet and/or \circ , respectively.

4. CONCLUSIONS

We have described a generic methodology for converting a measure of retrieval performance to a metasearch algorithm, and we have demonstrated such conversions for the measures average precision, R-precision, and precisions at standard cutoffs. Our conclusion is that system-oriented measures such as average precision and R-precision tend to implicitly weight relevant documents appropriately for metasearch purposes, yielding algorithms whose performance equals or exceeds benchmark techniques such as CombMNZ and Condorcet. We intend to explore the use of this methodology to convert other measures of retrieval performance to metasearch algorithms and to further study and evaluate the quality of retrieval performance measures in general.

5. REFERENCES

- [1] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–249, Gaithersburg, MD, USA, Mar. 1994. U.S. Government Printing Office, Washington D.C.
- [2] J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–275, Philadelphia, Pennsylvania, USA, July 1997. ACM Press, New York.
- [3] M. Montague and J. A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 538–548. ACM Press, November 2002.

¹Sign test of significance; 90% confidence level.