Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods

Gordon V. Cormack
University of Waterloo
Waterloo, Ontario, Canada

Charles L. A. Clarke
University of Waterloo
Waterloo, Ontario, Canada

Stefan Büttcher
Google
Redmond, WA, USA

ABSTRACT

Reciprocal Rank Fusion (RRF), a simple method for combining the document rankings from multiple IR systems, consistently yields better results than any individual system, and better results than the standard method Condorcet Fuse. This result is demonstrated by using RRF to combine the results of several TREC experiments, and to build a meta-learner that ranks the LETOR 3 dataset better than any previously reported method.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]:retrieval models

General Terms: Experimentation, Measurement

Keywords: fusion, aggregation, ranking

1. RECIPROCAL RANK FUSION

While supervised learning-to-rank methods have garnered much attention of late, unsupervised methods are attractive because they require no training examples. In the search for such a method we came up with Reciprocal Rank Fusion (RRF) to serve as a baseline. We found that RRF, when used to combine the results of IR methods (including learning to rank), almost invariably improved on the best of the combined results. We also found that RRF consistently equaled or bettered other methods we tried, including established metaranking standards Condorcet Fuse and CombMNZ (cf. [4]).

RRF simply sorts the documents according to a naive scoring formula. Given a set D of documents to be ranked and a set of rankings R, each a permutation on $1..|D|$, we compute

$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}\,,$$

where $k = 60$ was fixed during a pilot investigation and not altered during subsequent validation. Our intuition in choosing this formula derived from fact that while highly-ranked documents are more important, the importance of lower-ranked documents does not vanish as it would were, say, an exponential function used. The constant k mitigates the impact of high rankings by outlier systems.

Condorcet Fuse combines rankings by sorting the documents according to the pairwise relation $r(d_1) < r(d_2)$, which is determined for each (d_1, d_2) by majority vote among the input rankings. CombMNZ requires for each r a corresponding scoring function $s_r : D \rightarrow \mathbb{R}$ and a cutoff rank c which all contribute to the CombMNZ score:

$$CMNZscore(d \in D) = |\{r \in R | r(d) \leq c\}| \cdot \sum_{\{r | r(d) \leq c\}} s_r(d)\,.$$

We conducted four pilot experiments, each combining the results of 30 configurations of Wumpus Search applied to four different TREC collections. The results of the first, shown in table 1, indicated that $k = 60$ was near-optimal, but that the choice was not critical. The results also showed, somewhat unexpectedly, that RRF bested competing approaches, as well as more sophisticated learning methods whose investigation was the original impetus for our work.

We repeated our experiment with four sets of submissions to TREC tasks; the particular sets were selected because they have been used in previous metaranking evaluation. It is worthy of note that, while our pilot runs used exactly the same set of Wumpus configurations to generate the individual rankings on different datasets, the individual rankings in these experiments were exactly those submitted by TREC participants. Table 2 shows the RRF result, as well as the best individual, Condorcet and CombMNZ results. The MAP score for RRF exceeds that of Condorcet Fuse in all cases, and CombMNZ in all but one. RRF also outperforms the best ranking in each experiment, with the exception of TREC 9, where the best ranking was derived using a human-in-the-loop. RRF outperforms the next-best ranking, which was automated.

The pilot and TREC experiments indicate that RRF outperforms Condorcet, CombMNZ and the best system by 4% to 5% on average. We use a simple sign test to establish significance. Discounting the first pilot run, RRF outperformed Condorcet all 7 times ($p \approx 0.008$), outperformed CombMNZ 6 of 7 times ($p \approx .04$), and outperformed the best individual result either 6 or 7 times ($0.008 \leq p \leq 0.04$), depending on whether or not the manual result is considered. Thus all measured differences are significant.

Our final experiment used the sample learning results supplied with the LETOR 3[1] dataset, as well as a logistic gradient descent method (LGD) which we are developing. For the purpose of analysis, we combined the seven sets of document-query pairs into one and computed an overall MAP score.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

[1] `research.microsoft.com/en-us/um/beijing/projects/letor`

k	0	10	20	30	40	50	60	70	80	90	100	500
MAP	.2072	.2123	.2134	.2139	.2138	.2144	.2145	.2146	.2147	.2145	.2142	.2098

method	Best individual	Condorcet	CombMNZ
MAP	.2016	.2074	.2039

Table 1: Pilot results. Effect of k on MAP for RR Fusion of 30 model system results on TREC topics 351-400. Results of best model system and competing fusion methods shown for comparison. Similar results were seen for the same systems applied to three other test collections.

	Method			
Collection	RRF	Best individual	Condorcet	CombMNZ
TREC Robust	**.3686**	.3586	.3652	.3575
TREC 3	.4350	.4226	.4256	**.4381**
TREC 5	**.3394**	.3165	.3213	.3237
TREC 9	**.2830**	.3519 (.2801)	.2750	.2671

Table 2: MAP scores for fusion of submitted runs for TREC 3, TREC 5 and TREC 9 ad hoc tasks, plus TREC 2004 Robust track.

method	MAP_{method}	$MAP_{RRF} - MAP_{method}$	p
RRF	0.6051 (0.58 - 0.63)	–	–
Condorcet	0.5917 (0.56 - 0.62)	0.0134 (0.00 - 0.02)	.004
CombMNZ	**0.6107** (0.58 - 0.64)	-0.0056 (-0.01 - 0)	.2
ListNet [1]	0.5846 (0.56 - 0.61)	0.0205 (0.01 - 0.03)	.001
LGD	0.5837 (0.56 - 0.61)	0.0214 (0.01 - 0.04)	.003
AdaRank-MAP [6]	0.5778 (0.55 - 0.61)	0.0273 (0.01 - 0.04)	.000
RankSVM [3]	0.5737 (0.55 - 0.60)	0.0314 (0.02 - 0.04)	.000
RankBoost [2]	0.5622 (0.53 - 0.59)	0.0429 (0.03 - 0.06)	.000

Table 3: Individual rankings and fusion for 583,850 document-query pairs in LETOR 3 corpus. MAP score for each method, plus difference between fusion and individual MAP score with 95% confidence limits.

We also computed the difference between RRF and individual MAP scores, 95% confidence intervals, and p-value (likelihood under the null hypothesis that the difference is 0). Table 3 shows these results. RRF betters all individual rankings ($p < .003$), the best by a margin of 0.02 (4%); Condorcet is inferior to RRF ($p \approx .004$) while apparently bettering the individual rankings ($p \approx .2$). CombMNZ edges RRF by a small margin ($p \approx .2$). None of the measured differences among the baseline systems is significant.

2. DISCUSSION

For brevity, we report MAP as the measure of system performance. $P@k$, R-precision, and $NDCG$ yield comparable results.

RRF is simpler and more effective than Condorcet Fuse, while sharing the valuable property that it combines ranks without regard to the arbitrary scores returned by particular ranking methods [4]. RRF requires no special voting algorithm or global information; ranks may be computed and summed one system at a time, avoiding the necessity of keeping all rankings in memory. We conjecture that RRF outperforms Condorcet because it is better able to harness diversity within individual rankings. One or two systems that rank a document highly can substantially improve its rank relative to the more popular documents. With Condorcet, a simple majority of weak preferences may overrule substantially stronger ones.

CombMNZ multiplies the sum of the uncalibrated scores of individual system by the sum of a binary quantization of each rank. It is perhaps not surprising that its results have higher variance, ranging from insubstantially better than RRF to substantially worse than Condorcet. We conjecture that this effect is due to the fact that, by happenstance, some scores are more amenable than others.

To our knowledge, no reported result matches or exceeds the performance of the meta-learner formed by applying fusion to the LETOR baseline rank learning methods. So the meta-learner constitutes the best known method, and the result raises the lower bound of what is known to be learnable from the dataset. This latter question is a matter of some interest, as the MAP scores for LETOR 3 approach the 65% considered achievable with human-adjudicated relevance [5].

3. REFERENCES

[1] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. Learning to rank: from pairwise approach to listwise approach. In *ICML '07* (2007).

[2] Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *JMLR 4* (2003).

[3] Joachims, T. Optimizing search engines using clickthrough data. In *KDD '02* (2002).

[4] Montague, M., and Aslam, J. A. Condorcet fusion for improved retrieval. In *CIKM* (2002).

[5] Voorhees, E. M., and Harman, D. K., Eds. *TREC - Experiment and Evaluation in IR*. MIT Press, 2005.

[6] Xu, J., and Li, H. Adarank: a boosting algorithm for information retrieval. In *SIGIR '07* (2007).