On the Evaluation of Data Fusion for Information Retrieval

David Lillis
david.lillis@ucd.ie
School of Computer Science
University College Dublin
Dublin, Ireland

ABSTRACT

Data Fusion combines document rankings from multiple systems into one, in order to improve retrieval effectiveness. Many approaches to this task have been proposed in the literature, and these have been evaluated in various ways. This paper examines a number of such evaluations, to extract commonalities between approaches. Some drawbacks of the prevailing evaluation strategies are then identified, and suggestions made for more appropriate evaluation of data fusion.

KEYWORDS

information retrieval, data fusion, rank aggregation, evaluation

ACM Reference Format:
David Lillis. 2020. On the Evaluation of Data Fusion for Information Retrieval. In *Forum for Information Retrieval Evaluation (FIRE '20), December 16–20, 2020, Hyderabad, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3441501.3441506

1 INTRODUCTION

In the context of Information Retrieval, "data fusion" (also known as "rank aggregation") is a process whereby the ranked results of several systems, searching the same corpus, are combined into one single set of results to return to a user. Numerous algorithms have been proposed to perform data fusion effectively, with most attempting to leverage one or more of the following "effects" [25]: the *skimming effect* argues that prioritising highly-ranked documents in the input results will be beneficial; the *chorus effect* argues that if several systems include a document in their results, then this is evidence of relevance; and the *dark horse effect* notes that a system may occasionally return unusually accurate (or inaccurate) results compared to others, and this can be leveraged for effective retrieval. In practice the dark horse effect is rarely leveraged due to its inherent unpredictability.

Despite the myriad approaches that have been proposed, it can be argued that the community has not converged on a consensus about what the most effective approach(es) are in practice. This paper aims to explore why that is the case, by examining how newly-proposed approaches are evaluated and makes some suggestions for how this could be better unified in the future.

2 EVALUATION OF FUSION

While the evaluation methodology employed in practice is rarely uniform, sufficient commonality exists to be able to outline some characteristics of what may be described to be a "typical" approach. The following subsections outline aspects of evaluation that can be observed in a substantial body of research.

2.1 Use of standard test collections and results

The Text REtrieval Conference (TREC)[1] has long been a rich source of experimental data for data fusion, given that the runs submitted to TREC tasks are made available afterwards. The popularity of TREC data in data fusion experiments is evidenced in many studies [1–4, 7, 9, 12, 13, 15–20, 22, 24, 26–29]. Other sources include ImageCLEF[2] [30] and the NTCIR[3] IMine track [4]. Although this may seem to represent a consistency amongst researchers, it remains the case that different tracks are used and no consistent method of choosing which and how many runs to fuse has emerged.

2.2 Use of standard evaluation metrics

A feature of most data fusion research is the use of standard IR metrics for evaluation. Popular metrics for ad hoc tasks include:

- Precision at *n* (P@n) [1, 9, 11, 12, 15–17, 19, 22, 24, 26, 28]
- Average Precision (AP) or Mean Average Precision (MAP) [1–3, 6, 7, 9, 11–13, 15–20, 22, 24, 26–28, 30]
- Precision/Recall curves [13, 20, 24]
- Binary preference (bpref) [16, 17, 24]
- Recall-level precision (R-prec) [27, 28]
- Success at *n* documents (S@n) [9]
- Normalised Discounted Cumulated Gain (NDCG) [4, 15, 28]
- Mean Reciprocal Rank (MRR) [24]

2.3 Comparison with component systems

Many fusion researchers compare the fusion output with the quality of its inputs as part of their evaluation. This is an intuitively reasonable approach, as it acknowledges that merging result sets does not guarantee an improvement in quality. Generally, the comparison is made with the input run that has achieved the highest score according to the metric being used [1–4, 6, 7, 11–13, 15, 20, 26–28, 30]. A fusion method that improves upon this justifies its development by indicating that it is not possible to achieve equivalent results simply by choosing a single high-performing IR system.

Alternative comparisons include those involving the mean evaluation scores of the component systems [2, 26] or the median run amongst those available [19].

[1] http://trec.nist.gov
[2] http://imageclef.org
[3] http://research.nii.ac.jp/ntcir

FIRE '20, December 16–20, 2020, Hyderabad, India

ACM ISBN 978-1-4503-8978-5/20/12.
https://doi.org/10.1145/3441501.3441506

2.4 Comparison with other fusion algorithms

In addition to comparing a proposed fusion algorithm with the component systems' outputs, evaluation typically also includes a comparison with one or more competing data fusion algorithms [2, 3, 11–14, 16–20, 22, 24, 26–28, 28, 30]. There is great variation in the algorithms chosen for comparison.

When comparing against individual component systems or other algorithms, it is common to illustrate the difference in scores using either a chart or a table. In some cases, the degree of improvement (or disimprovement) is emphasised by displaying the increase (or decrease) in performance as a percentage of the baseline [2, 4, 6, 7, 9, 16, 17, 19, 27].

2.5 Statistical significance tests

Because most single-score evaluation metrics are averaged over a number of queries, many researchers also include statistical significance tests in order to show a significant improvement. For this purpose, a two-tailed paired t-test is most commonly employed [1, 4, 9, 12, 15–18, 22, 24, 26–30], although the Wilcoxon signed-rank test has also been used [11, 19].

2.6 Training data

Supervised data fusion approaches rely on knowledge of the past performance of the component systems. Evaluations generally divide the set of available queries into a training set and a test set [2, 6, 16–18, 24, 30].

3 DISCUSSION

The previous section illustrates that although no single evaluation framework for fusion has been agreed, certain features can be reasonably considered to be a "typical" aspect of data fusion evaluation. The following sections examine these aspects in more detail, and in particular examine the assumptions inherent in them.

3.1 Dark horse effect not considered

The dark horse effect is difficult to identify and exploit, and so it is generally not considered in the literature. Algorithms rarely attempt to leverage the effect, but also it is not considered in evaluation.

When comparing against the best component system, the most common approach is to choose a single component system based on its performance over all queries as measured by some evaluation metric. Statistical significance tests are run on a per-query basis against the outputs of that single best component system.

However, this ignores the potential for improving on single-system performance without requiring a fusion algorithm to be employed. Assuming that no component system achieves the best evaluation scores on all queries, there is potential for improvements to be achieved by choosing the best set of results on a per-query basis, thus exploiting the dark horse effect. The reason why this is important is because it represents the best performance that it is possible to achieve without merging result sets. If a data fusion algorithm can improve upon this score, then the case for fusion is clear: the results achieved by using data fusion cannot possibly be achieved using non-merging methods.

Examples of this being taken into account are rare. In [20], the comparison with the best component system includes the results of choosing the best component system on a per-query basis. This is done with a view to examining the limits of what is possible. Result set selection has also been attempted as an alternative to fusion [5]. This attempts to improve upon an individual component system by trying to identify the best result set for each query and return that to the user, without fusion being performed.

3.2 Fusion is different to classic IR

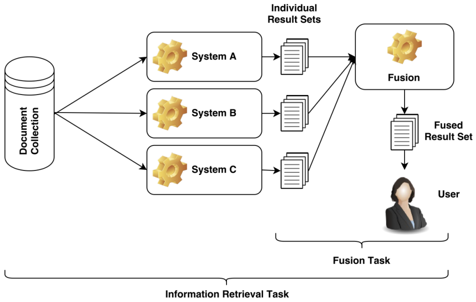

Figure 1: The Data Fusion Process

The use of standard evaluation metrics and standard data sets (e.g. from TREC) shows that data fusion is typically treated in the same way as standard ad hoc IR. However, it can be argued that fusion is a different challenge with its own characteristics.

Figure 1 illustrates the way in which the IR process can incorporate a data fusion algorithm. It operates as follows:

(1) A number of component IR systems search a document collection when they are provided with a query.
(2) Each component system produces a ranked list of results.
(3) A single, combined set of results is created and returned.

From this illustration, it can be seen that there are actually two contexts within which fusion can be applied. Firstly, when engaging in an ad hoc task that is based on a document corpus, a data fusion component can play a role in the overall architecture to combine the results of several rankers. In this situation, all documents in the corpus are available to the system as a whole, and traditional ad hoc retrieval evaluation is entirely appropriate.

However, many fusion researchers use pre-existing runs from separate systems in order to make comparisons with other fusion algorithms, as opposed to comparing with entire systems. In this scenario, the process that is actually being evaluated begins with the individual result sets (marked as "Fusion Task" in Figure 1).

Although this may appear at first to be a minor distinction, it has the consequence that when a fusion algorithm is being compared against others, the set of documents that can be included in the final output is not the entire document collection, and is likely not the entire set of relevant documents. This is in contrast to standard ad hoc IR. Instead, only the subset of documents that have been retrieved by at least one component system that is available. When considering evaluation purely from the point of view of making comparisons, this will not have a major effect, as a superior algorithm should achieve higher evaluation scores regardless. However, as discussed below, this has consequences for how the results are interpreted and is contrary to readers' expectations with regard to the intuitive understanding of evaluation metrics.

3.3 Reduction in recall

All mainstream ad hoc evaluation metrics operate within a range between 0 (no relevant documents) and 1 (perfect retrieval). While raw evaluation scores are not useful in isolation, the theory of these metrics is based on intuitive upper and lower bounds.

Since the documents available to a data fusion algorithm are only those returned by the component systems, recall will be reduced in most cases. Any relevant document from the corpus that no component system has returned cannot be included in the fused results. This has consequences for the popular metrics, as follows:

- Precision: Omitting some relevant documents will not affect precision. Perfect precision is possible if the fuser filters all non-relevant documents.
- P@10: In most practical situations there will be no effect. An exception is if fewer than 10 relevant documents are available due to poor system performance or a difficult query.
- MAP: A system that fails to return all relevant documents will achieve MAP below 1.0. This happens if any relevant document is not included in some input.
- NDCG: In a similar way to MAP, NDCG will be less than 1.0 with imperfect recall, because it will include documents that are not available to the fuser.
- NDCG@10: As with P@10, the effect on this metric is likely to be less pronounced. It will be unaffected if the input systems collectively find at least 10 documents at the maximum relevance level.
- bpref: For a perfect bpref score, all relevant documents must be returned before any judged non-relevant document. Again, any relevant documents that are not available to the fuser will adversely affect the score in the same way as for MAP.

This analysis indicates that when considering the fusion task in isolation, the theoretical upper bound for several metrics is now unknown. This is illustrated in Table 1, which shows the practical upper bounds for a particular set of inputs from the ad hoc task of the TREC 2014 Web Track. The first 1,000 results in each run are considered for this illustrative experiment. Each metric is the result of perfectly fusing four runs (i.e. choosing only the judged relevant documents and ranking them based on the grade of relevance). First the perfect fuser fused the 4 runs with the highest NDCG for the ad hoc task. This was repeated for the bottom 4 runs.

Table 1: Illustration of upper bounds for fusion.

Metric	Top 4	Bottom 4
Recall	0.8102	0.6224
Precision	1.0000	1.0000
P@10	0.9720	0.9760
MAP	0.8360	0.6602
NDCG	0.8920	0.7387
NDCG@10	0.9748	0.9152
bpref	0.8360	0.6602

The results demonstrate the highest evaluation stores that are theoretically possible with fusion for these inputs. When the collective recall of the component systems is imperfect, the maximum scores achievable for almost all metrics fall below 1.0.

P@10 and NDCG@10 fall below 1.0 as a result of a small number of topics where there are fewer than 10 relevant documents. The other metrics (MAP, NDCG and bpref) all suffer a reduction due to the absence of some relevant documents in the input result sets.

This issue has occasionally been taken into account in the fusion literature. For example, in [3], the fusion output was compared to two bounds, chosen to reflect the possible performance level. The "naïve bound" referred to the best results that were possible if only documents that had been returned by the component systems were considered. The "ordered pairs bound" additionally assumed that a fusion algorithm would act reasonably, so that a document that was ranked above another in all input result sets could not be ranked below it in the final output. Similarly, [23] compared fusion results against an "oracle" that always exhibited optimal selection.

To restore the intuitive theoretical 0-to-1 range, the evaluation metrics could be normalised when only the fusion task is being evaluated. A simple suggestion would be to only use judgments for documents that were available to fuser through its input result sets.

3.4 Percentage increases

Several studies report their percentage increase in certain metrics over their baselines. This has similar problems in that this presentation is unintuitive and difficult to interpret beyond a simple comparison between approaches. This happens for two reasons. Firstly, lower baseline scores have a greater potential for percentage increases. Secondly, the maximum percentage increase that it is possible to achieve will be unknown for the reasons outlined above in the previous section. Thus the raw percentage increase will be very difficult to contextualise. Although normalised metrics help address the latter issue, the first problem remains unaddressed. As such, percentage increases are not recommended for fusion evaluation.

3.5 Model complexity

The most commonly-used baselines tend to be simple techniques like CombMNZ [10], Borda fuse [3] or Reciprocal Rank Fusion [8]: straightforward models that are easy to implement. More complex models are much less commonly used, and as such a detailed comparison of these from the literature is difficult. Traditionally, implementations of proposed models have often not been made available to other researchers for comparison, and so there is a burden on researchers to re-implement more complex models.

4 CONCLUSIONS AND FUTURE WORK

This paper has presented an examination of how data fusion is evaluated. Evaluations tend to use established datasets from TREC and mainstream IR evaluation metrics. Comparisons are generally made with the best individual system as well as with other baseline fusion techniques. The paired t-test is the most common significance test. Although not all papers surveyed had all these characteristics, this is a reasonable characterisation of a "typical" setup.

However, this paper also argues that the fusion task in isolation is different to traditional ad hoc retrieval. Using ad hoc metrics as-is removes the intuitive 0-1 range that it normally found in evaluation. It is suggested that normalised metrics should be used for fusion.

The recent TrecTools project [21] includes implementations of some fusion techniques. This may provide a platform upon which future fusion work can be based if implementations of new and existing models are shared amongst the community.

REFERENCES

[1] Yael Anava, Anna Shtok, Oren Kurland, and Ella Rabinovich. 2016. A Probabilistic Fusion Framework. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, Indianapolis Indiana USA, 1463–1472. https://doi.org/10.1145/2983323.2983739

[2] Javed A. Aslam and Mark Montague. 2000. Bayes Optimal Metasearch: A Probabilistic Model for Combining the Results of Multiple Retrieval Systems. In *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA, 379–381. https://doi.org/10.1145/345508.345665

[3] Javed A. Aslam and Mark Montague. 2001. Models for Metasearch. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA, 276–284. https://doi.org/10.1145/383952.384007

[4] Ashraf Bah and Ben Carterette. 2016. PDF: A Probabilistic Data Fusion Framework for Retrieval and Ranking. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ACM, Newark Delaware USA, 31–39. https://doi.org/10.1145/2970398.2970419

[5] Niranjan Balasubramanian and James Allan. 2010. Learning to Select Rankers. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10*. ACM Press, New York, New York, USA, 855. https://doi.org/10.1145/1835449.1835650

[6] Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew. 1994. Automatic Combination of Multiple Ranked Retrieval Systems. In *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA, 173–181.

[7] Steven M Beitzel, Ophir Frieder, Eric C Jensen, David Grossman, Abdur Chowdhury, and Nazli Goharian. 2003. Disproving the Fusion Hypothesis: An Analysis of Data Fusion via Effective Information Retrieval Strategies. In *SAC '03: Proceedings of the 2003 ACM Symposium on Applied Computing*. New York, NY, USA, 823–827. https://doi.org/10.1145/952532.952695

[8] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '09*. ACM Press, Boston, MA, USA, 758. https://doi.org/10.1145/1571941.1572114

[9] Mohamed Farah and Daniel Vanderpooten. 2007. An Outranking Approach for Rank Aggregation in Information Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. ACM Press, New York, New York, USA, 591–598. https://doi.org/10.1145/1277741.1277843

[10] Edward A Fox and Joseph A Shaw. 1994. Combination of Multiple Searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215*. 243–252.

[11] Kripabandhu Ghosh, Swapan Kumar Parui, and Prasenjit Majumder. 2015. Learning Combination Weights in Data Fusion Using Genetic Algorithms. *Information Processing & Management* 51, 3 (May 2015), 306–328. https://doi.org/10.1016/j.ipm.2014.12.002

[12] Anna Khudyak Kozorovitsky and Oren Kurland. 2011. Cluster-Based Fusion of Retrieved Lists. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 893–902. https://doi.org/10.1145/2009916.2010035

[13] Alexandre Klementiev, Dan Roth, and Kevin Small. 2007. An Unsupervised Learning Algorithm for Rank Aggregation. In *Machine Learning: ECML 2007*, Joost N. Kok, Jacek Koronacki, Raomon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron (Eds.). Lecture Notes in Computer Science, Vol. 4701. Springer Berlin Heidelberg, Berlin, Heidelberg, 616–623. https://doi.org/10.1007/978-3-540-74958-5

[14] Joon Ho Lee. 1997. Analyses of Multiple Evidence Combination. *SIGIR Forum* 31 (1997), 267–276. https://doi.org/10.1145/278459.258587

[15] Shangsong Liang, Ilya Markov, Zhaochun Ren, and Maarten de Rijke. 2018. Manifold Learning for Rank Aggregation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press, Lyon, France, 1735–1744. https://doi.org/10.1145/3178876.3186085

[16] David Lillis, Fergus Toolan, Rem Collier, and John Dunnion. 2006. ProbFuse: A Probabilistic Approach to Data Fusion. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, Seattle, WA, USA, 139–146. https://doi.org/10.1145/1148170.1148197

[17] David Lillis, Fergus Toolan, Rem Collier, and John Dunnion. 2008. Extending Probabilistic Data Fusion Using Sliding Windows. In *Advances in Information Retrieval. Proceedings of the 30th European Conference on Information Retrieval Research (ECIR 2008)*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Lecture Notes in Computer Science, Vol. 4956. Springer Berlin Heidelberg, Berlin, 358–369. https://doi.org/10.1007/978-3-540-78646-7_33

[18] David Lillis, Lusheng Zhang, Fergus Toolan, Rem W. Collier, David Leonard, and John Dunnion. 2010. Estimating Probabilities for Effective Data Fusion. In *Proceedings of the 33rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Geneva, Switzerland, 347–354. https://doi.org/10.1145/1835449.1835508

[19] Craig Macdonald and Iadh Ounis. 2006. Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management - CIKM '06*. ACM Press, New York, New York, USA, 387–396. https://doi.org/10.1145/1183614.1183671

[20] R Manmatha, T Rath, and F Feng. 2001. Modeling Score Distributions for Combining the Outputs of Search Engines. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA, 267–275. https://doi.org/10.1145/383952.384005

[21] João Palotti, Harrisen Scells, and Guido Zuccon. 2019. TrecTools: An Open-Source Python Library for Information Retrieval Practitioners Involved in TREC-like Campaigns. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Paris France, 1325–1328. https://doi.org/10.1145/3331184.3331399

[22] Ella Rabinovich, Ofri Rom, and Oren Kurland. 2014. Utilizing Relevance Feedback in Fusion-Based Retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '14*. ACM Press, New York, New York, USA, 313–322. https://doi.org/10.1145/2600428.2609573

[23] Daniel Sheldon, Milad Shokouhi, Martin Szummer, and Nick Craswell. 2011. LambdaMerge. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining - WSDM '11*. ACM Press, New York, New York, USA, 795. https://doi.org/10.1145/1935826.1935930

[24] Milad Shokouhi. 2007. Segmentation of Search Engine Results for Effective Data-Fusion. In *Proceedings of the 29th European Conference on Information Retrieval Research (ECIR '07)*. Rome, Italy, 185–197.

[25] Christopher C Vogt and Garrison W Cottrell. 1999. Fusion via a Linear Combination of Scores. *Information Retrieval* 1, 3 (1999), 151–173. https://doi.org/10.1023/A:1009980820262

[26] Shengli Wu. 2013. The Weighted Condorcet Fusion in Information Retrieval. *Information Processing & Management* 49, 1 (Jan. 2013), 108–122. https://doi.org/10.1016/j.ipm.2012.02.007

[27] Shengli Wu, Yaxin Bi, Xiaoqin Zeng, and Lixin Han. 2009. Assigning Appropriate Weights for the Linear Combination Data Fusion Method in Information Retrieval. *Information Processing & Management* 45, 4 (July 2009), 413–426. https://doi.org/10.1016/j.ipm.2009.02.003

[28] Shengli Wu and Fabio Crestani. 2015. A Geometric Framework for Data Fusion in Information Retrieval. *Information Systems* 50 (June 2015), 20–35. https://doi.org/10.1016/j.is.2015.01.001

[29] Shengli Wu, Chunlan Huang, Liang Li, and Fabio Crestani. 2019. Fusion-Based Methods for Result Diversification in Web Search. *Information Fusion* 45 (Jan. 2019), 16–26. https://doi.org/10.1016/j.inffus.2018.01.006

[30] Xin Zhou, Adrien Depeursinge, and Henning Muller. 2010. Information Fusion for Combining Visual and Textual Image Retrieval. In *2010 20th International Conference on Pattern Recognition*. IEEE, 1590–1593. https://doi.org/10.1109/ICPR.2010.393