D-RANK: A FRAMEWORK FOR SCORE AGGREGATION IN SPECIALIZED SEARCH

Martin Veselý
LIA EPFL, CH-1015 Lausanne, Switzerland

Martin Rajman
LIA EPFL, CH-1015 Lausanne, Switzerland

Jean-Yves Le Meur
CERN, CH-1211 Geneva, Switzerland

Ludmila Marian
CERN, CH-1211 Geneva, Switzerland

Jérôme Caffaro
CERN, CH-1211 Geneva, Switzerland

Keywords: Specialized search engines, Score aggregation, Information retrieval.

Abstract: In this paper we present an approach to score aggregation for specialized search systems. In our work we focus on document ranking in scientific publication databases. We work with the collection of scientific publications of the CERN Document Server. This paper reports on work in progress and describes rank aggregation framework with score normalization. We present results that we obtained with aggregations based on logistic regression using both ranks and scores. In our experiment we concluded that score-based aggregation favored performance in terms of Average Precision and Mean Reciprocal Rank, while rank-based aggregation favored document discovery.

1 INTRODUCTION

Specialized search gains increasingly attention across scientific communities. According to a recent study, users of scientific information in the field of particle physics often turn to specialized search services such as arXiv.org [1], SPIRES [2], or the CERN Document Server (CDS) [3], rather than to general purpose search engines when accessing scientific information (Gentil-Beccot et al., 2008).

In the scope of specialized search, the traditional notion of relevance is often extended to incorporate additional attributes to score and rank documents at a search engine output. When searching for scientific documents, ranking attributes are traditionally based on citations or previous document usage such as ”reads” or document access frequency. The intuition is that citing or reading a document by peers shows evidence of document relevance within a given scientific field.

Additional attributes are sometimes used, such as the publication date. As new documents do not have a sufficient search or citation history, they might be incorrectly ranked when time is not taken in consideration.

A multitude of relevance attributes thus needs to

[1] http://arXiv.org/

[2] http://www.slac.stanford.edu/spires/

[3] http://cds.cern.ch/

Veselý M., Rajman M., Le Meur J., Marian L. and Caffaro J..
D-RANK: A FRAMEWORK FOR SCORE AGGREGATION IN SPECIALIZED SEARCH.
DOI: 10.5220/0003293404820485
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 482-485
ISBN: 978-989-8425-40-9

be aggregated within the document ranking process. In this paper we propose an aggregation mechanism that allows for aggregation of a multitude of query-independent attributes. We use two approaches, one aggregating the attribute scores and another one, aggregating ranks using weighted sum and logistic regression as the aggregation vehicle. We present the evaluation framework that targets the CDS document collection, a production database used at CERN.

In the section 2 we outline the aggregation method, in the section 3 we present the experimental data set up, in section 4 we present results that we obtained on a test data set and we conclude in section 5.

2 SCORE AGGREGATION

We divide the process of score aggregation in three step: (i) first we select relevant ranking attributes that are convenient for aggregation, (ii) in the second step, scores need to be normalized and re-scaled, and (iii) finally, scores are aggregated via a score aggregation function.

Selection of Attributes. In the first phase we select attributes that are convenient for aggregation. Attributes that are not correlated are good candidates for aggregation. On the other hand attributes that are highly correlated can be considered as substitutes and in that case we can selected only one of them.

We noticed that usage of traditional correlation coefficients such as Spearman Rank correlation or Kendal Tau coefficients do not take into account the importance of low ranks. For this reason the correlations should be adjusted so as to put more weight on changes that occur in the upper part of the ranked list. Some work in this direction has been also suggested by (Yilmaz et al., 2008).

Score Normalization. In the second step we normalize scores so that they reflect the underlying distribution of values. The idea is that a normalized score should reflect the proportion of the population of documents with lower scores as they are observed for a given ranking attribute. For example, if a score of N corresponds to a median score among all of the observed scores, it should be converted into a normalized score of 0.5.

To determine the normalization function for each of the attributes, we first calculated values at a percentile level. We then smoothed the obtained values using standard density estimation techniques to approximate the underlying densities. We then construct the cumulative distribution function summing up values over corresponding interval.

Score Aggregation. The task of score aggregation was previously addressed in several works. Garcin et al (Garcin et al., 2009) analyze aggregation of feedback ratings into a single value. They consider different aggregations relying on informativeness, robustness and strategyproofness. On all these attributes, they show that the mean seems to be the worst way of aggregating ratings, while the median is more robust. In previous works, logistic regression was also used as a vehicle to aggregate scores (Le Calvé and Savoy, 2000) (Jacques Savoy and Vrajitoru., 1996) (Craswell et al., 1999). In our preliminary study we adopted the two mentioned aggregation models based on logistic regression and a weighted sum.

To rank documents with logistic regression we first compute the value of logit that corresponds to a particular combination of scores of individual documents. We then project the obtained result on the logistic curve and read the resulting aggregated score on the Y-axis.

A more detail about the implementation of the rank aggregation with logistic regression in d-Rank can be obtained in (Vesely and Rajman, 2009). In our study we worked with chosen regression coefficients for which we tested a variety of combinations. Eventually coefficients should be learned through an automated procedure. The way we have generated data for our experiments is in more detail described in the next section.

3 EXPERIMENTAL SETUP

Within our work we plan to perform two types of evaluation: a system evaluation using a referential that we extracted from the user access logs, and a user-centric evaluation (Voorhees, 2002).

In order to proceed with the system evaluation, we needed a referential that would allow us to compute and compare our system using standard information retrieval measures. To our knowledge the CDS collection was not used within an information retrieval evaluation in the past. One of the results of our work thus is a referential that allows for a system evaluation of various document retrieval scenarios featured by the CDS retrieval system.

To create the referential of relevance judgments, we opted for parsing the user access logs for queries that were issued by users of the CDS search system in the past. This way we have obtained a set of test queries for experimentation that is close to a real-world scenario. For this purpose, we created a tool that allows us to parse user access logs and extract information that is essential for our experimentation, including search phrases, search attributes and corresponding relevant documents. Queries in the referential are composed from all query terms that were used by a user including all parameters that were used at the search time. Document identifiers then correspond to a known relevant documents that were downloaded upon a search were then added. A typical referential entry looks as follows:

```
Query terms: Ellis, John
Field: Author
Collection: Published Articles
Action type: Search
Relevant document: 1282439
Relevant document: 1257907
...
```

For our initial experimentation we also generated a small data set constructed in the following way: we generated a collection of one thousand documents and ranked them using five artificial independent ranking attributes. Furthermore we assumed that the individual ranking attributes do perform relatively well when used separately to perform ranking. Our referential thus contains a set of documents that were ranked high, the referential documents were selected randomly with a log-normal distribution of ranks to favor good individual performance.

As far as the evaluation measure is concerned, we opted for the Average Precision and the Mean Reciprocal Rank, used previously in the TREC evaluations. These evaluation measures put more weight on better ranked relevant documents for each query in the evaluation set. The lower rank the relevant document is observed, on average, the better performance of the ranker is calculated.

4 RESULTS

In this section we present results that we obtained on a generated data set. We have conducted the following two experiments. In the first experiment we focused on performance of our aggregates and we compared them to the best-performing individual ranking attribute. In the second experiment, we focused on how the ranking aggregation allows to lift relevant documents from the bottom of the ranked list to the visible area. We selected a threshold of Rt, a rank that splits a ranked document list to two parts, the one that was visible to the user on the search output, and the ”invisible” one. We then kept only relevant documents that belong to the invisible part of the list in the referential (i.e. removed documents that ranked well enough). We again computed the AP evaluation measure for the aggregates. This way we could estimate the quantity of relevant documents that did not score well enough using the individual rankings, and were lifted into the visible ranking area after aggregation. In our experiment we selected Rt=100.

We aggregated the lists in two different ways: by score and by rank. We used the logistic regression aggregation framework and a simple weighted sum aggregation for comparison. We thus evaluated the following four ranking aggregates: (i) weighted sum of scores (AW-S), (ii) weighted sum of ranks (AW-R), (iii) score aggregate using logistic regression (LR-S), and (iv) rank aggregate using logistic regression (LR-R). We then calculated the MRR and AP@k measures for k in 5,10,20,50,100. The obtained results are shown in the Figure 1 and in the Table 1.

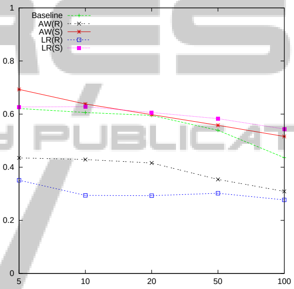

Figure 1: Average Precision at various levels for the best individual ranking attribute (baseline) and ranking aggregates.

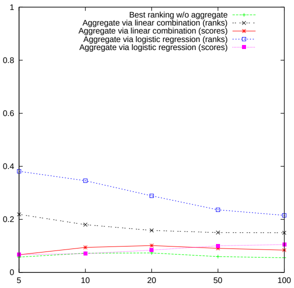

Figure 2: Potential for discovery of relevant documents using (AP@k), Rt=100.

Table 1: Results of the evaluation run on test data (AP@5,10,20 and MRR).

	AP@5	AP@10	AP@20	MRR
Baseline	0.621	0.606	0.595	0.59
AW-S	**0.693**	**0.638**	0.598	**0.632**
LR-S	0.627	0.628	**0.606**	0.547
LR-R	0.351	0.294	0.293	0.447
AW-R-	0.435	0.429	0.416	0.477

As shown in the Table 2 the performance of the ranking aggregate based on logistic regression with ranks provides best performance in terms of the mean average precision considering only relevant documents that were presumably not seen on lists ranked with the individual attributes.

We now proceed with the significance measurement for the second experiment. We worked with a 10-fold data sample. Table 3 shows values for all pairs of aggregated measures. As shown, we have found a non-significant difference between the two score-based aggregations AW-S and LR-S on 95% level of significance.

Table 2: 10-Fold validation test for measuring lift with AP@5.

Fold	Base	AW-R-	AW-S	LR-R	LR-S
1	0.077	0.231	0.080	0.382	0.053
2	0.064	0.197	0.067	0.394	0.060
3	0.062	0.210	0.067	0.381	0.054
4	0.064	0.181	0.069	0.396	0.089
5	0.062	0.248	0.070	0.412	0.069
6	0.054	0.224	0.065	0.363	0.063
7	0.044	0.243	0.061	0.345	0.069
8	0.049	0.218	0.062	0.415	0.056
9	0.056	0.199	0.069	0.310	0.083
10	0.044	0.236	0.054	0.414	0.076
Mean	0.058	0.219	0.066	0.381	0.067
StDev	0.010	0.022	0.007	0.034	0.012

Table 3: Significance test for measuring lift using AP@5.

	AW-S	AW-R	LR-S	LR-R
Baseline	2.25	21.1	**1.88**	30.0
AW-S	-	21.1	**0.18**	28.9
AW-R		-	19.1	12.8
LR-S			-	27.6

5 CONCLUSIONS AND FUTURE WORK

In this paper we proposed a framework for score aggregation in specialized search systems. In particular we focused on ranking of scientific documents in the particle physics community. We addressed the issues of score normalization and aggregation through methods of kernel density estimation and logistic regression as possible vehicles for rank aggregation. We have presented results from two experiments suggesting that score-based aggregation favored performance in terms of Mean Reciprocal Rank and Average Precision, while rank-based aggregation favored document discovery.

In the future work we plan to proceed with user-centric evaluation on real-world information retrieval system. The goal is to confirm our preliminary results obtained on a small test data collection and we plan to apply an automated procedure to learn the aggregated scoring function.

REFERENCES

Craswell, N., Hawking, D., and Thistlewaite, P. B. (1999). Merging results from isolated search engines. In *Australasian Database Conference*, pages 189–200.

Garcin, F., Faltings, B., and Jurca, R. (2009). Aggregating reputation feedback. In Paolucci, M., editor, *1st International Conference on Reputation (ICORE)*, pages 62–74, http://www.reputation09.net.

Gentil-Beccot, A., Mele, S., Holtkamp, A., O'Connell, H. B., and Brooks, T. C. (2008). Information resources in high-energy physics: Surveying the present landscape and charting the future course. *J. Am. Soc. Inf. Sci. Technol.*, 60(arXiv:0804.2701.):150–160. 27 p.

Jacques Savoy, A. L. C. and Vrajitoru., D. (1996). Report on the trec-5 experiment: Data fusion and collection fusion.

Le Calvé, A. and Savoy, J. (2000). Database merging strategy based on logistic regression. *Inf. Process. Manage.*, 36(3):341–359.

Vesely, M. and Rajman, M. (2009). Rank Aggregation in Scientific Publication Databases Based on Logistic Regression. Technical report.

Voorhees, E. (2002). The philosophy of information retrieval evaluation. In *In Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370. Springer-Verlag.

Yilmaz, E., Aslam, J. A., and Robertson, S. (2008). A new rank correlation coefficient for information retrieval. In *SIGIR*, pages 587–594.