Regression Relevance Models for Data Fusion

Shengli Wu, Yaxin Bi, and Sally McClean
School of Computing and Mathematics, University of Ulster, UK
{s.wu1,y.bi,si.mcclean}ulster.ac.uk

Abstract

Data fusion has been investigated by many researchers in the information retrieval community and has become an effective technique for improving retrieval effectiveness. In this paper we investigate how to model rank-probability of relevance relationship in resultant document list for data fusion since reliable relevance scores are very often unavailable for component results. We apply statistical regression technique in our investigation. Different regression models are tried and two good models, which are cubic and logistic models, are selected from a group of candidates. Experiments with 3 groups of results submitted to TREC are carried out and experimental results demonstrate that the cubic and logistic models work better than the linear model and are as good as those methods which use scoring information.

1. Introduction

In information retrieval, many different retrieval models such as the Boolean model, the vector space model, the probabilistic model, the language model, have been proposed and used. These models are comparable in performance and there is no all-time winner. In such a situation, to use a few independent search engines to search the same document collection for the same information need, and then to merge these results from these search engines for better retrieval performance is an attractive option. This is the primary idea for data fusion. Data fusion (also known as meta-search) has been investigated by many researchers (e.g., in [1, 4, 5, and 6]) and has become a competitive option to implement an effective search engine.

There are two different situations which demand different treatment. One is to assume that the score information, which is an indicator of the estimated relevance of a document, is always available for every retrieved document. The other situation is that only ranking information is available. The first assumption fits quite well with most results submitted to Text Retrieval Conferences (TREC: http://trec.nist.gov/), which provide a good test platform to test various data fusion methods. However, there are some exceptions. Some results (e.g., ictweb10nf and ictweb10nfl in TREC 2001, NLPR04okall and uic0401 in TREC 2004) do not provide scoring information. Some other results provide very unreasonable scores. For example, in apl10wa of TREC 2001 and NLPR04semLM of TREC 2004, all the scores are negative, while all the scores in most other results are positive. In csiroOawa1 and csiroOawa2 of TREC 2001, all the scores are located in each of the three very narrow value intervals: (30050-30000), (20050-20000), and (10050-10000). In addition, most web search engines such as Google, Yahoo and alltheweb, and digital libraries such as IEEE Xplore, only provide a list of ranked documents. In statistics, it is believed that ranking is more robust than numerical scoring for distinguishing each object in a group of objects. Therefore, we may have to use ranking information for data fusion if that is the only information available; or we may prefer to use ranking information if we think that scoring information is not reliable though it is available.

In this paper we investigate data fusion methods which use ranking information. The approach is based on the modeling of relevance dependency of documents at different document ranks. Without any analysis, one simple assumption can be: the probability of relevance of a document decreases linearly with the rank position of the document in a result list. Actually, Borda fusion [1] makes such an assumption. For a set of *n* documents in a list, the top-ranked document is given *n* points, the second ranked document is given *n*-1 points, and so on. Then fusion methods such as CombSum can be used with these points. However, such a simple assumption is not very precise and further improvement is possible for a more accurate estimation. After analyzing three groups of component results submitted to TREC 9, 2001, and 2004, we find that the cubic model and logistic model fit well with these results. Using these two models for prediction, significant improvements over linear functions as in

Borda fusion can be observed for data fusion in all cases.

Fox and Shaw [4] introduced CombSum, which has been widely used by many researchers for the evaluation and comparison of data fusion methods. Suppose we have n documents $\{d_1,d_2,\ldots,d_n\}$ and m information retrieval systems $\{r_1,r_2,\ldots,r_m\}$. For the information need, each information retrieval system r_i provides a normalized relevance score s_{ij} to document d_j. CombSum uses the following formula to calculate the score for every document d_j:

$$\text{CombSum\_score}(\text{d}_\text{j}) = \sum_{i=1}^{m} s_{ij} \qquad (1)$$

The logistic model has been used by Calve and Savoy [3] to merge results in a distributed information retrieval environment, in which all component results are retrieved from totally different document collections. However, its effect on data fusion over identical databases has not been investigated. In addition, the cubic model has not been used before for data fusion or result merging in information retrieval.

2. Modeling the rank-probability of relevance relationship

We used three groups of results submitted to TREC 9 (web track), 2001 (web track) and 2004 (robust track). For convenient processing and obtaining useful results, all results satisfy the following two conditions:

- They provided 1000 documents for every query;
- Their performances on mean average precision are over an arbitrarily chosen threshold 0.15. We do not include very poor results here since poor results are not useful for data fusion and we should avoid their participation.

We choose 38 results in TREC 9, 30 results in TREC 2001, and 77 results in TREC 2004. Their average performances over a group of queries (50 for TREC 9 and 2001 and 249 for TREC 2004) and standard deviations are shown in Table 1[1].

Table 1. Information about results in three groups

TREC Group	No. of results	No. of queries	MAP	Standard deviation of MAP
9	38	50	0.2107	0.0276
2001	30	50	0.1989	0.0262
2004	77	249	0.2855	0.0421

For all the results in the same year group, we checked every document involved to see if it was relevant or not according to the judgment made by human referees in TREC. Then for every year group, we calculated the probability of relevant documents at every rank position. After some observation and experimentation using SPSS[2], we found that cubic and logistic functions are likely the best two for use among a group of functions such as linear, logistic, inverse, cubic, growth, and exponential. The cubic model uses the following function

$$p = a_1 + b_1 * \ln(rank) + c_1 * \ln(rank)^2 + d_1 * \ln(rank)^3 \qquad (2)$$

to estimate curves. In (2), a_1, b_1, c_1 and d_1 are coefficients, p is the dependent variable, and ln(*rank*) is the independent variable that is the natural logarithm of rank, a participant of Calve and Savoy's functions [5]. The logistic model uses the following function

$$p = \frac{1}{\frac{1}{u} + a_2 * e^{\ln(rank)*\ln(b_2)}} \qquad (3)$$

to estimate curves. In (3), a_2 and b_2 are coefficients. u is the upper boundary value, which needs to be greater than the largest dependent variable value and therefore we assign 1 to it. As in (2), p is the dependent variable, and ln(*rank*) is the independent variable. Instead of using *rank* directly, we use ln(*rank*) as independent variable, which is in line with Calve and Savoy's usage [3]. Actually, according to our observation, this transformation can bring considerable improvement for the fitness of the estimated curves.

Tables 2 and 3 present the values of coefficients for cubic and logistic models, respectively. In all three year groups, these coefficients bear certain similarity. In Table 2, a_1 and d_1 always take positive values while b_1 and c_1 always take negative values. If only considering absolute values, we always have $|a_1|>|b_1|>|c_1|>|d_1|$. In Table 3, a_2 and especially b_2 take similar values in all three groups. In all the cases, the significance level is .0000, which represents the fact that both the cubic model and the logistic model fit well with TREC data.

Table 2. Coefficient values for the estimated cubic functions (Significance is at .0000 in all cases)

Group	a_1	b_1	c_1	d_1
9	.4137	-.0699	-.0049	.0009
2001	.4683	-.0814	-.0035	.0008
2004	.6577	-.1368	-.0019	.0012

[1] In TREC 9 and 2001 web track and TREC 2004 robust track, relevant documents were divided into highly relevant documents and (ordinary) relevant documents. In this paper we do not distinguish them.

[2] SPSS is a statistical software. The company's website is located at http://www.spss.com/

Table 3. Coefficient values for the estimated logistic functions

Group	a_2	b_2	Significance
9	.1803	2.5685	.0000
2001	.2226	2.2966	.0000
2004	.1406	2.5362	.0000

We present the original and estimated curves by use of cubic and logistic regression models for TREC 9 and 2001 in Figures 1 and 2. The curves for TREC 2004 are not given since they are very similar to the two figures presented. It can be seen from the figures that the cubic model is better than the logistic model for modelling the relationship between rank and relevance probability. The cubic model fits very well with the original curves in almost all ranks; while the logistic model does not fit well with the original curves in top ranks. For a more precise estimation of the fitness of these two regression models, we calculate the dissimilarity (Euclidean distance) between the original curves and the estimated curves. $E\_d(c_e, c_o) = \sqrt{\sum_{i=1}^{1000} (pe_i - po_i)^2}$. Here c_e and c_o represent the estimated and the original curves; and pe_i and po_i reperent the i-th points of the estimated and of the original curve, respectively. The results are shown in Table 4. We can clearly see that the distances for the cubic model are substantially smaller than those for the logistic model in all three year groups.

Table 4. Euclidean distance between actual and estimated curves

Group	Cubic	Logistic
TREC 9	0.3453	0.9616
TREC 2001	0.2810	0.8003
TREC 2004	0.1576	0.5867

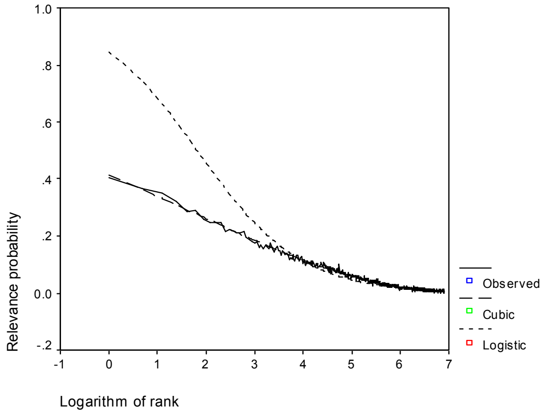

Figure 1. Original and estimated curves for a group of TREC 9 results

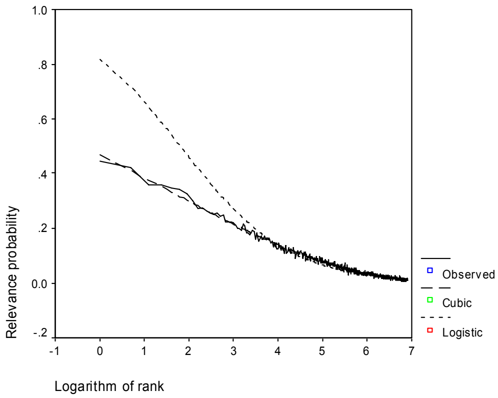

Figure 2. Original and estimated curves for a group of TREC 2001 results

3. Data fusion experimental results

In order to test the usefulness of the estimated curves for data fusion, we carry out an experiment to compare the data fusion methods, which use the cubic function and the logistic function, with three other data fusion methods CombSum, CombMNZ, and Borda. For CombSum and CombMNZ, the linear [0,1] score normalization method was applied to the raw scores associated with the documents in those component results. It normalizes the maximum raw score into 1, the minimum raw score into 0, and any other score into a value between 0 and 1 accordingly. Condorcet is not considered since we consider it less useful practically because of its time-complexity. For all the methods evaluated in this paper, their time complexity is $O(mn)$. Here m is the number of component results and n is the number of documents in each component result. However for Condorcet, the time complexity is at least $O(mn^2)$. We use the same three groups of results as in Section 3. The coefficients need to be estimated for the cubic method and the logistic method. We use the values in Tables 2 and 3 for all the selected results in that year. Then Equation 1, which is used by CombSum, is used for the fusion process. These coefficient values are far from optimistic since they are obtained from the curve estimation with all the selected component results of that year and those results are quite different from each other.

For each run in each year group, a given number of component results (3, 4, 5, 6, 7, 8, 9, 10) were selected randomly and then all 50 or 249 queries (depending on which year group) were run using all five data fusion methods. 200 runs were executed for any given number of systems. Two measures, which were mean average precision (MAP) and recall-level precision (RP), were used to evaluate the performance of these data fusion methods. These two measures have been widely used for retrieval evaluation [2].

Figures 3-8 present the experimental results. Figures 3-5 present mean average precision of all methods over a total of 200 runs for every given number of results, while Figures 6-8 present recall-level precision of these data fusion methods. In Figures 3-4 and 6-7, each data point is the average of 50 queries *200 runs. In Figures 5 and 8, each data point is the average of 249 queries *200 runs.

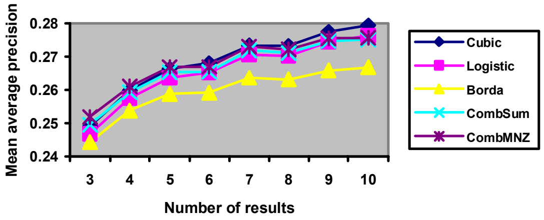

Figure 3. Mean average precision of several data fusion methods in TREC 9

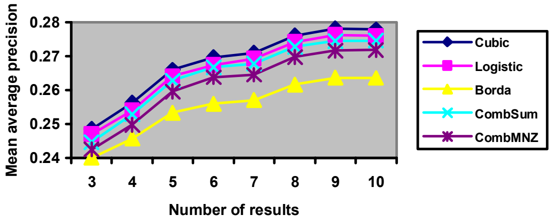

Figure 4. Mean average precision of several data fusion methods in TREC 2001

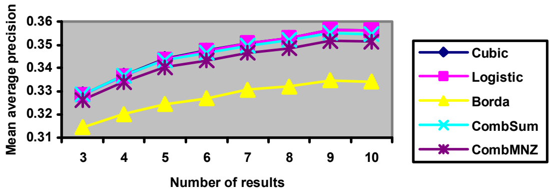

Figure 5. Mean average precision of several data fusion methods in TREC 2004 (the curve of Cubic is covered by the curve of Logistic)

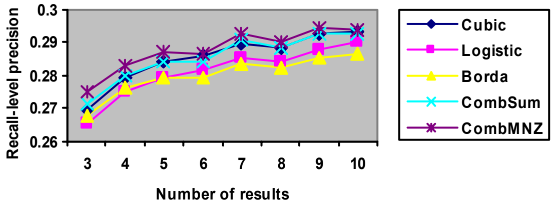

Figure 6. Recall-level precision of several data fusion methods in TREC 9

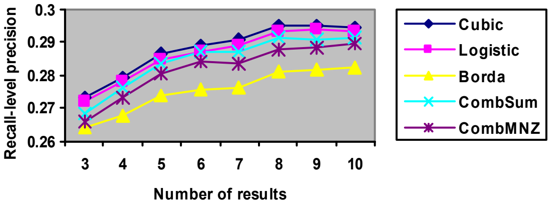

Figure 7. Recall-level precision of several data fusion methods in TREC 2001

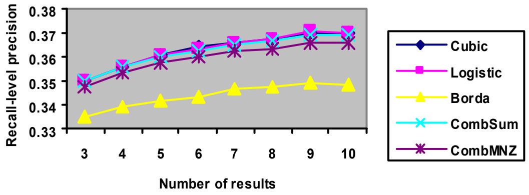

Figure 8. Recall-level precision of several data fusion methods in TREC 2004 (the curve of Cubic is covered by the curve of Logistic)

From these figures we can see that Borda fusion is almost always the worst among all the methods. A two-tailed *t* test is carried out to compare their means and it shows that the differences between Borda fusion and all other methods are significant at a level of .000, or the probability is over 99.95% that Borda fusion is worse than all other methods. On average, the cubic method, the logistic method, CombSum, and CombMNZ outperform Borda by 4.67%, 4.33%, 4.20%, and 3.64% respectively on MAP; they outperform Borda by 3.94%, 3.46%, 3.68%, and 3.29% respectively on RP. This demonstrates that the linear relevance model used by Borda fusion can be improved by using non-linear functions such as cubic or logistic functions. On the other hand, it demonstrates that the fused results using rank information (the cubic method and the logistic method) can be as good as those using scoring information (CombSum and CombMNZ).

Comparing the cubic method with CombSum, the former is better than the latter at a significant level of .000 on MAP and a significant level of .006 on RP. For the cubic method, MAP is 0.2937 and RP is 0.3116; for CombSum, MAP is 0.2924 and RP is 0.3109. Comparing the cubic method with CombMNZ, the former is better than the latter at a significant level of .000 on both MAP and RP. For CombSum, MAP is 0.2908 and RP is 0.3097. The difference between the logistic method and CombSum is not significant (0.155 on MAP and 0.290 on RP). The difference between the logistic method and CombMNZ is significant on MAP (0.008) but not significant on RP (0.645).

Finally let us compare the cubic method and the logistic method. Over three year groups, their average performances are very close. For the cubic method, MAP is 0.2937 and RP is 0.3116; for the logistic method, MAP is 0.2927 and RP is 0.3102. Interestingly, these small differences (0.33% on MAP and 0.46% on RP) are still statistically significant at a significance level of .002 (*t* test), which suggests that the cubic method is slightly better than the logistic method. According to the observations in Section 3, which demonstrate that the cubic function can fit much better than the logistic function with TREC data, we expect that the difference between the cubic method and the logistic method should be bigger than this. Therefore, we take a more careful look at the logistic curves. Actually, they do not fit well with TREC data on the top 30-50 ranked documents, which are the most important documents for data fusion. However, we find that the logistic curves always overestimate the relevance probability of the document in those ranks, and in a consistent way! That is to say, for those top-ranked documents, the rates of overestimation decrease with ranks. Such an overestimation does not affect the effectiveness of data fusion very much because of two reasons. First, those documents which are top-ranked in any component result are very likely top-ranked in the fused results whether their relevance probabilities are overestimated or not. Second, the relative rankings of those top-ranked documents are not very much affected because of the pattern of overestimation. A linear relationship exists between the estimated logistic curve and the original data curve for those top-ranked documents. In such a situation, if we only consider the top 30-50 ranked documents, the estimated logistic curve will bring the same fused result as the original curve by using Equation 1 for data fusion. This can explain why the logistic curve is almost as good as the cubic curve for data fusion even though it is not as good as the cubic curve for the estimation of the relationship of rank-probability of relevance.

Author names and affiliations are to be centered beneath the title and printed in Times 12-point, non-boldface type. Multiple authors may be shown in a two- or three-column format, with their affiliations italicized and centered below their respective names. Include e-mail addresses if possible. Author information should be followed by two 12-point blank lines.

5. Conclusion

In this paper we have investigated applying regression relevance models for data fusion, in which only ranking information is used. We find that the cubic models and the logistic models are two good models for this. Compared with Borda fusion, the cubic method and the logistic method are more effective by 4%. Compared with CombSum and CombMNZ, the cubic method is slightly better than them and the logistic method is as good as them in performance but no score information is needed for the cubic method and the logistic method.

Comparing the logistic models with the cubic models, the latter can fit TREC data more precisely than the former. However, when used for data fusion, the cubic method is slightly better than the logistic method.

6. References

[1] Aslam, J. and Montague, M. Models for Metasearch. In *Proceedings of of 24th Annual ACM SIGIR Conference*, 2001, pages 275-284, New Orleans, USA.

[2] Buckley, C. and Voorhees, E. M. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual ACM SIGIR Conference*, 2000, pages 33-40, Athens, Greece.

[3] Calve, A. L. and Savoy, J. Database Merging Strategy Based on Logistic Regression. *Information Processing & Management*, 2000, 36(3): 341-359.

[4] Fox, E. A. and Shaw, J. A. Combining of Multiple Searches. In *Proceedings of the 2nd Annual Text Retrieval Conference (TREC-2)*, 1994, NIST, Gaithersburg, USA, November.

[5] Wu, S., & McClean, S. Performance prediction of data fusion for information retrieval. *Information Processing & Management*, 2006, 42(4): 899-915.

[6] Wu, S., & McClean, S. Improving High Accuracy Retrieval by Eliminating the Uneven Correlation Effect in Data Fusion. Journal of the American Society for Information Science and Technology, 57(14): 1962-1973, 2006.