Probabilistic Score Normalization for Rank Aggregation

Miriam Fernández, David Vallet, and Pablo Castells

Universidad Autónoma de Madrid, Escuela Politécnica Superior
Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain
{miriam.fernandez,david.vallet,pablo.castells}@uam.es

Abstract. Rank aggregation is a pervading operation in IR technology. We hypothesize that the performance of score-based aggregation may be affected by artificial, usually meaningless deviations consistently occurring in the input score distributions, which distort the combined result when the individual biases differ from each other. We propose a score-based rank aggregation model where the source scores are normalized to a common distribution before being combined. Early experiments on available data from several TREC collections are shown to support our proposal.

1 Introduction

Rank aggregation is a pervading operation in IR technology [6]. To name a few examples, rank aggregation takes place in the combination of multiple criteria for document/query similarity assessment in most search engines; in merging the outputs of different engines for meta-search; in the combination of query-based and preference-based relevance for personalized search [1]; or even in the combination of preferences from multiple users for collaborative retrieval [5]. Both rank-based and score-based aggregation techniques have been explored in prior research on this topic [7]. We hypothesize that that the performance of score-based aggregation may be affected by artificial, usually meaningless deviations consistently occurring in the input score distributions, which do not affect the performance of each ranking technique separately, but distort the combined result when the individual biases differ from each other, and therefore it should be possible to improve the results by undoing these deviations.

In order to devise a general method to merge the output of several ranking techniques, no a-priori assumption on the interpretation of the scores values should be made. The values may correspond to a degree of relevance, probability of relevance, odds of relevance, user preference, or other interpretations in a variety of retrieval models, often undergoing further mathematical transformations (scaling, dampening, logs, etc.) for practical purposes. However, in order to combine the scores, the values should be first made comparable across input systems [2], which usually involves a normalization step [6]. In this poster we propose an aggregation model where the source scores are normalized to a common *ideal* score distribution, and then merged by a linear combination. Early experiments on available data from several TREC collections are shown to support our proposal.

2 Score Normalization

In prior work, normalization typically consists of linear transformations [3], and other relatively straightforward, yet effective methods, such as normalizing the sum of scores (rather than the max) of each input system to 1, or shifting the mean of values to 0 and scaling the variance to 1 [6]. But none of these strategies takes into account the detailed distribution of the scorings, and is thus sensitive to "noise" score biases.

A work where the score distribution is taken into account is that of Manmatha et al [4], who analyze the probabilistic behavior of search engines, in order to derive a better combination of their outputs. They observe that the scoring values have an exponential distribution for the set of non-relevant documents, and a Gaussian distribution for the set of relevant ones. According to this, a score s output by a given engine for a document d is normalized to P (d is relevant | $score(d) = s$), which is computed by applying Bayes' rule, and approximating the probabilities by a mixture of an exponential and a Gaussian distribution, using the Expectation Maximization method.

Starting from Manmatha's analysis of typical score distributions, we propose an alternative approach, where input scores are mapped to an *optimal score distribution* (OSD), which we define as the score distribution of an ideal scoring function that matches the ranking by actual relevance. Of course this is a difficult concept to define, let alone to obtain, but we claim that an acceptable approximation can provide good results.

Our method works as follows. Let Ω be the universe of information objects to be ranked, and P the set of rank lists to be combined. Each rank source $\tau \in P$ can be represented as a bijection $\tau : \Omega_\tau \to \mathbf{N}^+_{|\Omega_\tau|}$ for some $\Omega_\tau \subset \Omega$, where for each $x \in \Omega_\tau$, $\tau(x)$ is the position of x in the ranking returned by τ. For each $\tau \in P$, we shall denote by $s_\tau : \Omega \to P$ the scoring function associated to τ, where we take $s_\tau(x) = 0$ if $x \notin \Omega_\tau$. Our approach consists of two phases. The first one is performed offline, as follows:

1. For each ranked list $\tau \in P$, compute the cumulative score distribution F_τ of the values s_τ returned by the ranking system that outputs τ. This can be approximated by running a significant number of calls to each system with different random inputs (e.g. queries and documents).
2. Build a strictly increasing OSD $\overline{F} : [0,1] \to [0,1]$. This step is discussed below.

In the second phase, which takes place at query-time, the outputs of the rank sources are normalized and merged:

3. Normalization: For each $x \in \Omega$ and $\tau \in P$, map the score of each rank source to the OSD: $s_\tau(x) \to \overline{s}_\tau(x) = \overline{F}^{-1} \circ F_\tau \circ s_\tau(x)$.
4. Combination: merge the normalized scores, e.g. by a linear combination or some other score-based technique.

The idea of step 3 is illustrated in figure 1. The normalization respects the order of each rank list (except in intervals where F_τ is constant, i.e. where by definition it is unlikely that any score value should fall), since $\overline{F}^{-1} \circ F_\tau$ is monotonically non-decreasing. The resulting scores $\overline{s}_\tau = \overline{F}^{-1} \circ F_\tau \circ s_\tau$ range in [0,1], and their distribution is $\overline{F}$ for all $\tau \in P$, thus undoing potential distributional biases, as intended.

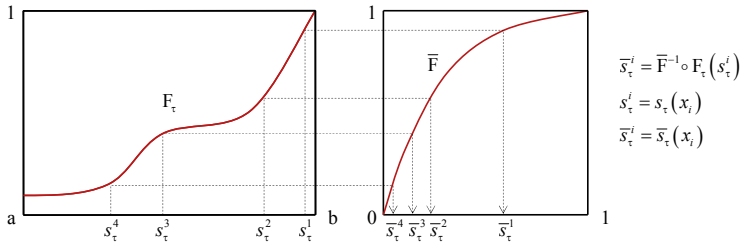

Fig. 1. Mapping scores to a common distribution

The choice of $\overline{F}$ as an appropriate OSD, in step 2 above, is critical to our method. Our proposed approach consists of computing the average distribution of several good scoring systems, as a rough approximation to an actual relevance distribution. This can be obtained empirically on a statistically significant sample of scoring systems (the ones to be merged, or different ones) and input values. In this estimation, the scores of each system are first linearly normalized to [0,1] by a variation of the standard normalization technique [3], where rather than taking the min and max scores of a single ranked list, all the scores collected from the system over several runs are included.

3 Evaluation and Results

We have tested our techniques in four different test collections from the TREC Web Results, namely TREC8, TREC9, TREC9L, and TREC2001. For the comparative evaluation we have tried our technique with two reference combination functions after the normalization step, to which we will refer as: a) DCombSUM, where the fused score is computed as $s_R(x) = \sum_{\tau \in R} \overline{s}_\tau(x)$, i.e. our score normalization step is followed by the so-called CombSUM method [6]; and b) DCombMNZ, where $s_P(x) = h(x,R) \sum_{\tau \in R} \overline{s}_\tau(x)$, and $h(x,P) = \left| \{ \tau \in R \mid s_\tau(x) > 0 \} \right|$ is the number of engines that return x, a technique named as CombMNZ in prior work [6].

We have compared these functions with other ones where the same combination step is used, but a different normalization method is applied. As a benchmark for comparison, we have taken the results published in [7], which we label as SCombSUM (CombSUM with standard score normalization), RCombSUM (CombSUM with Rank-sim normalization), and SCombMNZ (CombMNZ with standard score normalization). Table 1 shows the average results over the four collections. It can be seen that both DCombSUM and DCombMNZ are globally better that the other techniques. Although we only show the averaged results, this behavior is consistent over the four collections. DCombMNZ is only surpassed on average by SCombMNZ in TREC 2001, while the performance of DCombSUM, which could be thought of as a non-tuned version of our algorithm, performs slightly below DCombMNZ, but still globally better than any other of the benchmarks taken from [7].

Table 1. Average precision for 10 trials of the combination of 2 to 12 rank lists. The results are averaged over the four TREC collections

	2	4	6	8	10	12	Avg
SCombSUM	0.2598	0.2886	0.3084	0.3172	0.3204	0.3241	0.3031
RCombSUM	0.2567	0.2884	0.2847	0.2877	0.2971	0.2994	0.2857
SCombMNZ	0.2599	0.2884	0.3058	0.3176	0.3156	0.3231	0.3017
DCombSUM	0.2614	0.2942	0.3096	0.3184	0.3237	0.3268	0.3057
DCombMNZ	0.2637	0.2979	0.3090	0.3194	0.3228	0.3268	0.3066

5 Further Work

The possibilities for the continuation of this work are manifold. Studying score distributions is a research topic by itself. For instance, we foresee that a finer, more specialized analysis of score distributions could be achieved by identifying and separating certain conditions on which the distribution may depend, such as properties of the queries (e.g. query length), the search space, the result set, or other domain-specific factors. Also, we are currently exploring techniques where the coefficients in the linear combination are a function of application-specific variables of the ranking system, such as the uncertainty in the rankings [1].

6 Acknowledgements

This research was supported by the EC (FP6-001765 – aceMedia), and the Spanish Ministry of Science and Education (TIN2005-06885). The content expressed is the view of the authors but not necessarily the view of the aceMedia project as a whole.

References

1. Castells, P., Fernández, M., Vallet, D., Mylonas, P., Avrithis, Y. Self-Tuning Personalised Information Retrieval in an Ontology-Based Framework. 1st IFIP Intl. Workshop on Web Semantics (SWWS 2005). LNCS Vol. 3532. Agia Napa, Cyprus, 2005, pp. 455-470.
2. Croft, W. B. Combining approaches to information retrieval. In: Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval. Kluwer Academic Publishers, 2000, pp. 1-36.
3. Lee, J. H. Analyses of multiple evidence combination. 20th ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 97). New York, 1997, pp. 267-276.
4. Manmatha. R., Rath, R., Feng, F. Modelling score distributions for combining the outputs of search engines. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001). New Orleans, LA, pp. 267-275.
5. Masthoff, J. Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. User Modeling and User-Adapted Interaction 14 (1), 2004, pp. 37-85.
6. Montague, M., Aslam, J.A. Relevance score normalization for metasearch. 10th Conf. on Information and Knowledge Management (CIKM 2001). Atlanta, GA, 2001, pp. 427-433.
7. Renda, M. E., Straccia, U. Web metasearch: rank vs. score based rank aggregation methods. ACM symposium on Applied Computing. Melbourne, Florida, 2003, pp. 841-846.