

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
 {miriam.fernandez,david.vallet,pablo.castells}@uam.es
<http://nets.ii.uam.es>



Problem and approach

The rank fusion problem

- ◆ Merge the output of several retrieval functions
 - Normalization: make outputs comparable across systems
 - Estimation: assign score to non returned items
 - Combination: merge the results into a single list
- ◆ Applications
 - Metasearch and distributed search
 - Combination of internal search engine heuristics
 - Group-based decision-making
 - Personalized search
 - Classification based on multiple evidence

Our research is here

Limitations of common normalization techniques

- ◆ Scores may have an artificial bias in their distribution
 - Even after normalization
 - Noisy biases, i.e. not corresponding to a real difference in relevance
 - The bias is different in each input system
- ◆ Scores are normalized for a single result set in isolation
 - No global comparison of a result against prior results from the same system
 - The best result of a bad run gets the same score as the best of a good run
- ◆ This may hurt the performance of the combination

Our approach: general principles

- ◆ Score-based normalization for rank fusion
- ◆ Take a wider perspective of the behavior of the input systems to be merged: use historic scoring data
- ◆ Undo artificial score biases: equalize the distributions of each system
- ◆ Shift-invariance, scale-invariance, insensitiveness to outliers
- ◆ Relevance information not required
- ◆ Improve effectiveness, retain efficiency

Common state of the art techniques: combination

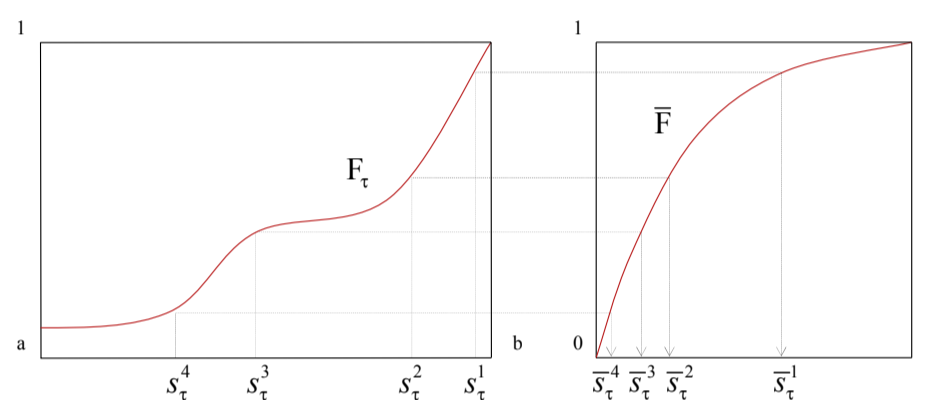
- ◆ CombSUM: $s_{\mathcal{R}}(x) = \sum_{r \in \mathcal{R}} s_r(x)$
- ◆ CombMNZ: $s_{\mathcal{R}}(x) = h(x, \mathcal{R}) \sum_{r \in \mathcal{R}} s_r(x)$
- ◆ CombMIN, CombMAX, CombMED, CombANZ, logistic regression, Markov chains...

Proposed model

- ◆ Let us assume an ideal unbiased scoring function $r(x)$ exists, ranging in $[0,1]$
- ◆ Given a scoring function $s_r(x)$, we wish to normalize $s_r(x)$ to $\bar{s}_r(x)$ so that $P(s_r(y) \leq s_r(x)) = P(r(y) \leq \bar{s}_r(x))$ (*)
- ◆ If F_r and \bar{F} are the cumulative distributions of s_r and r , it can be seen that $\bar{s}_r = \bar{F}^{-1} \circ F_r \circ s_r$ is a solution to (*)

Common state of the art techniques: normalization

- Score-based
- ◆ Standard: $\bar{s}_r(x) = \begin{cases} \frac{s_r(x) - \min_{y \in \Omega_r} s_r(y)}{\max_{y \in \Omega_r} s_r(y) - \min_{y \in \Omega_r} s_r(y)} & \text{if } x \in \Omega_r \\ 0 & \text{otherwise} \end{cases}$
 - ◆ ZMUV, 2MUV, ...
 - ◆ [Manmatha 2001]: $\bar{s}_r(x) = P(y \text{ is relevant} | s_r(y) = s_r(x))$
- Rank-based
- ◆ Rank-sim: $\bar{s}_r(x) = 1 - \frac{\tau(x) - 1}{|\Omega_r|}$
 - ◆ Borda: $\bar{s}_r(x) = \begin{cases} 1 - \frac{\tau(x) - 1}{|\Omega_r|} & \text{if } x \in \Omega_r \\ \frac{|\Omega_r| - |\Omega_r| + 1}{2|\Omega_r|} & \text{otherwise} \end{cases}$
 - ◆ Bayes: $\bar{s}_r(x) = \log \frac{P(\tau(x) | x \text{ is relevant})}{P(\tau(x) | x \text{ is not relevant})}$

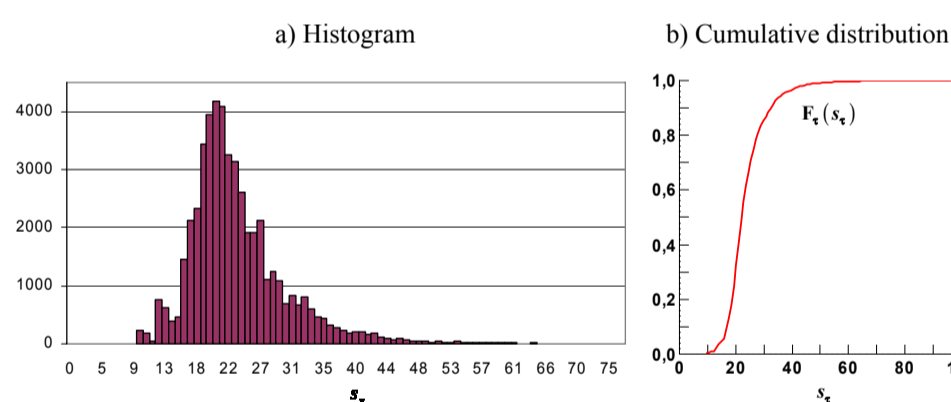


Our proposal: distribution-based score normalization

Implementation of the proposed approach

1. Compute the score distribution F_r of each input system
2. Find a good approximation to an unbiased distribution \bar{F}
 - Merge samples from several input systems so as to diffuse the “noise”
3. Normalize the scores through a mapping of the individual distributions to the common unbiased distribution
4. Combine the normalized scores \bar{s}_r using some score combination strategy

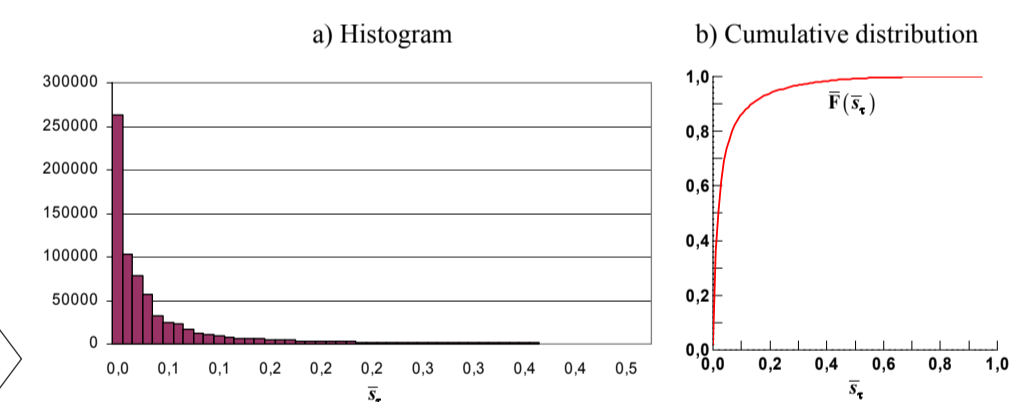
Experiments



Example: score distribution obtained for the “mds08w1” system in TREC 8 with the 1000 highest scores returned for each of 50 queries

Obtaining score distributions

1. Score distribution of each input retrieval system
 - Collect score values of the retrieval function over a period of time
 - 50-100 queries may be good enough
 - Approximate the distribution by the histogram of scores
2. Common reference “unbiased” score distribution
 - Normalize the whole historic score data series of each system to $[0,1]$ linearly
 - Build a joint histogram by merging all the series
 - Approximate the common distribution using the histogram



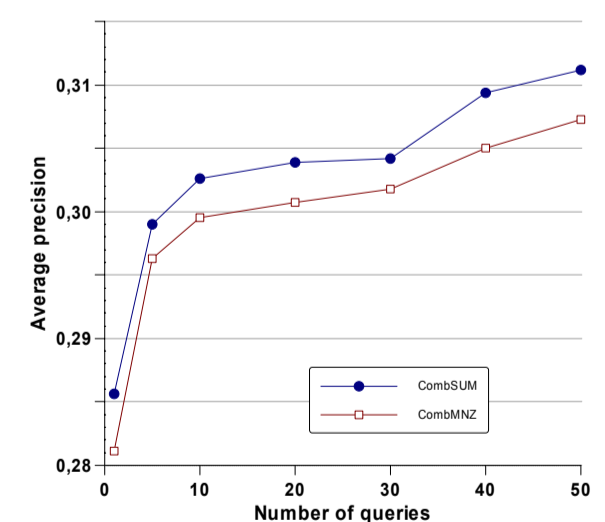
Joint score distribution obtained for the twelve best input systems from TREC 8, 9, 9L and 2001

Experimental setting

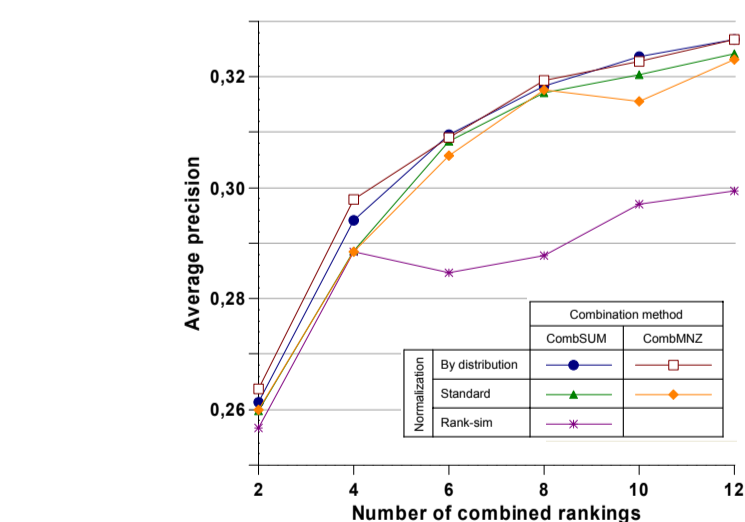
- ◆ Our normalization technique was compared against standard normalization and Rank-sim normalization
- ◆ The normalization steps were combined with CombSUM and CombMNZ for testing
- ◆ We reproduce the experiments in [Renda 2003], and we take the data for standard normalization and Rank-sim from that experiment
- ◆ The combinations were tested with data from TREC8, 9, 9L, and 2001, with 50 queries each, taking the 12 best engines, and the 1000 top results from each ranked list
- ◆ The experiment consisted of merging 2, 4, 6, 8, 10, and 12 lists, randomly selected, repeat 10 times per query, and take the average precision for each group
- ◆ A low-volume test was conducted also, by using several score sampling sets of different size (1 to 50 queries)

Application in the aceMedia project

- ◆ Combination of query-based (visual, NL) and preference-based relevance for personalized retrieval in multimedia corpora
- ◆ Combination of ontology-based semantic query relevance scores and keyword-based query relevance scores



How much historic data is enough? Average performance of our approach (with CombSUM and CombMNZ) over the four TREC collections using different size of historic data (result set scores), averaging the precision over the 2 to 12 list combinations



Average performance over the four TREC collections: our normalization approach is compared to standard normalization and Rank-sim, followed by CombSUM and CombMNZ

References

- [Croft 2000] 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, 1-36.
- [Lee 1997] Lee, J. H. Analysis of multiple evidence combination. 20th ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 97). New York, 1997, 267-276.
- [Manmatha 2001] Manmatha, R., Rath, R., Feng, F. Modeling score distributions for combining the outputs of search engines. 24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001). New Orleans, LA, 267-275.
- [Montague 2001] Montague, M., Aslam, J.A. Relevance score normalization for metasearch. 10th Conf. on Information and Knowledge Management (CIKM 2001). Atlanta, GA, 2001, 427-433.
- [Renda 2003] Renda, M. E., Straccia, U. Web metasearch: rank vs. score based rank aggregation methods. ACM symposium on Applied Computing. Melbourne, Florida, 2003, 841-8.