A Formal Approach to Score Normalization for Meta-search

R. Manmatha and H. Sever
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003
[manmatha,sever]@cs.umass.edu

ABSTRACT

Meta-search, or the combination of the outputs of different search engines in response to a query, has been shown to improve performance. Since the scores produced by different search engines are not comparable, researchers have often decomposed the meta-search problem into a score normalization step followed by a combination step. Combination has been studied by many researchers. While appropriate normalization can affect performance, most of the normalization schemes suggested are ad hoc in nature.

In this paper, we propose a formal approach to normalizing scores for meta-search by taking the distributions of the scores into account. Recently, it has been shown that for search engines the score distributions for a given query may be modeled using an exponential distribution for the set of non-relevant documents and a normal distribution for the set of relevant documents. Here, it is shown that by equalizing the distributions of scores of the top non-relevant documents the best meta-search performance reported in the literature is obtained. Since relevance information is not available apriori, we discuss two different ways of obtaining a good approximation to the distribution of scores of non-relevant documents. One is obtained by looking at the distribution of scores of all documents. The second is obtained by fitting a mixture model of an exponential and a Gaussian to the scores of all documents and using the resulting exponential distribution as an estimate of the non-relevant distribution. We show with experiments on TREC-3, TREC-4 and TREC-9 data that the best combination results are obtained by averaging the parameters obtained from these approximations. These techniques work on a variety of different search engines including vector space search engines like SMART and probabilistic search engines like INQUERY.

The problem of normalization is important in many other areas including information filtering, topic detection and tracking, multilingual search and distributed retrieval. Thus, the techniques proposed here are likely to be applicable to many of these tasks.

1. INTRODUCTION

Meta-search, or the combination of the outputs of different search engines to produce a single (combined) ranked list in response to a query, has been shown to improve performance [5, 8, 9, 7, 11, 2]. Many search engines produce scores as a measure of the relevance of the document to a particular query. These scores are then used to generate document rankings. The scores produced by different search engines are usually not comparable since they are often computed using some metric (or non-metric) distance function. Researchers have often decomposed the meta-search problem into a score normalization step followed by a combination step. While the combination step has been studied by a number of researchers [5] there has been little work on the normalization techniques to be used. The normalization step can in fact be more critical for performance in many situations [13]. Existing normalization schemes [8, 9, 13] have been proposed on heuristic grounds and their performance seems to depend somewhat on the dataset used [13]. Thus, the appropriate choice of normalization is not clear.

We propose that the correct way to normalize scores for meta-search is by taking the distributions of the scores into account. Researchers have previously shown that for search engines [11] and for the outputs of filtering systems [1] the score distributions for a given query may be modeled using an exponential distribution for the set of non-relevant documents and a normal distribution for the set of relevant documents. Here, we show that by equalizing the distributions of scores of the top non-relevant documents the best meta-search performance reported in the literature is obtained. The non-relevant distributions can be equalized by mapping the minimum score to zero and equalizing the means of the different exponential distributions. Since relevance information is not available in practice, we show that there are two different ways of obtaining a good approximation to the distribution of scores of non-relevant documents (i.e. of estimating the exponential parameter).

1. The non-relevant distribution may be approximated by fitting an exponential distribution to the scores of *all* documents (rather than fitting only the non-relevant documents). This approximation is usually reasonable because the proportion of relevant documents in search engines is usually small. We show that this approximation is in fact identical with a normalization technique selected by Montague and Aslam [13] on heuristic grounds.

2. An alternative technique for estimating the non-relevant distribution involves fitting a mixture model consisting of an exponential and a Gaussian to the scores of all documents [11] and then using the exponential component as an estimate of the non-relevant distribution. The mixture model is solved using Expectation-Maximization (EM).

The best estimate turns out to be to take an average of those obtained using the above (two) approximations. We show that the combination performance obtained by taking this average is the

In HLT'02. This material is based on work supported in part by the Center for Intelligent Information Retrieval, in part by the National Science Foundation under grant numbers IRI-9619117 and in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor(s).

best reported performance for meta-search and is consistently good over all the four (TREC) datasets tested. The techniques used here have been applied to a variety of search engines operating on different principles including vector space engines like SMART and probabilistic search engines like INQUERY.

Surprisingly, the performance of this model is better than one obtained by combining posterior probabilities (computed from the mixture model mentioned above) [11]. Part of the problem stems from errors in estimating the mixture model (consisting of relevant and non-relevant errors).

While our work in this paper focuses on the problem of combining the outputs of different search engines retrieving documents from a common database, score normalization is important in many different contexts. The approach here could be easily extended to the combination of different search engines indexing databases in different languages to produce a multi-lingual search engine or the combination of search engines indexing different databases (distributed retrieval). The score normalization procedure used here may also be useful for the topic detection and tracking task where scores have to be normalized across different topics.

2. PRIOR WORK

We will discuss some of the normalization and combination techniques proposed before describing some score modeling work. A recent and extensive survey of evidence combination in information retrieval is provided by Croft [5].

Fox and Shaw [7] proposed a number of combination techniques including operators like the MIN and the MAX. Other techniques included one that involved setting the score of each document in the combination to the sum of the scores obtained by the individual search engines (CombSum), while in another the score of each document was obtained by multiplying this sum by the number of engines which had non-zero scores (CombMNZ). Note that summing (CombSum) is equivalent to averaging while CombMNZ is equivalent to weighted averaging. Lee [8, 9] studied this further with six different engines. His contribution was to normalize each engine on a per query basis improving results substantially. Lee showed that CombMNZ worked best, followed by CombSum while operators like MIN and MAX were the worst. Lee also observed that the best combinations were obtained when systems retrieved similar sets of relevant documents and dissimilar sets of non-relevant documents. Vogt and Cottrell [14] also verified this observation by looking at pairwise combinations of systems. A probabilistic approach using ranks rather than scores was proposed last year by Aslam and Montague [3, 2]. This involved extensive training across about 25 queries to obtain the probability of a rank given a query. Their results for TREC-3 were close to but slightly worse than Lee's COMBMNZ technique [1]. Aslam and Montague were able to demonstrate that rank information alone can be used to produce good combination results. The main difficulty with this technique seems to be the extensive training required of every engine on a substantial number of queries.

Montague and Aslam [13] proposed three different normalization schemes for meta-search. The methods involved linearly shifting and scaling scores so that the following mappings were achieved:

Name	Method
Standard	Map min to 0 and max to 1.
Sum	Map min to 0 and the sum to 1.
ZMUV	Map mean to 0 and variance to 1.

Table 1: Normalization Methods Suggested by Montague and Aslam

They tested these with some well known combination techniques including CombSum and CombMNZ [7, 8, 9].

Recent (and independent) work by Manmatha et al [11] - for search - and Arampatzis and van Hameren [1] - in the case of filtering - has shown that the scores of non-relevant documents may be approximated by an exponential distribution and the scores of relevant documents by a Gaussian distribution. These experiments (by Manmatha et al [11]) were done by modeling the top 1000 documents for different search engines from the TREC-3 ad hoc track.

These distributions were successfully used for information filtering in [1, 15]. Manmatha et al [11] also showed that the relevant and non-relevant distributions could be recovered by solving a mixture model consisting of an exponential and a Gaussian using Expectation-Maximization (EM). They used mixture model to map scores to probabilities for each engine. The probabilities were averaged for meta-search. The results were as good as the CombMNZ technique with the Standard normalization.

3. EXISTING NORMALIZATION SCHEMES

We first start by looking at Montague and Aslam's normalization schemes. Montague and Aslam [13] reported the results of combining a set of n engines. First, a set of n engines is randomly selected from the set of all available engines and these n engines are then combined. A different set of n engines is then selected and the results combined again. The process is repeated until all possible choices for a given n have been selected. The results for a given n are then averaged and reported. We argue that this is not a useful way of reporting results in meta-search. The average precision for the average combination is often less than the average precision for the best search engine. This happens because many of the sets of n are combinations of search engines with much lower average precision.

In Table 2 we, therefore, compare the performance of the three normalization schemes suggested by Montague and Aslam [13] and two different combination schemes on the top five engines for data from the ad hoc track of TREC-3 data. This track provides scores for the top 1000 documents for all search engines which participated in this track. In the table Sum-CombSum means that the normalization technique used was the 'Sum' technique (see Table 1) and the combination technique was CombSum. The first row gives the average precision for the top engine (in terms of average precision). The second row reports the combination performance in terms of average precision for the top two engines. The third, fourth and fifth rows report the performance for combining the top 3 engines, the top 4 engines and the top 5 engines respectively. The results are also plotted in Figure 1.

From Table 2 it is clear that Sum-CombSum and Sum-CombMNZ perform best (with Sum-CombSum performing slightly better). This is in fact corroborated by experiments on the ad hoc tracks in TREC-4, TREC-5 data and on TREC-9 (web-track) data. From this table, the normalization used appears to be more important than the combination scheme used. That is, the Sum normalization performs best whether CombSum or CombMNZ is selected. [2].

Why should the Sum normalization scheme perform better? The paper by Montague and Aslam [13] does not give any intuition for picking one normalization scheme over another. We argue below that the appropriate normalization involves equalizing the non-

[1]The graph for Lee's technique in [3] is incorrect.

[2]Interestingly, contrary to Lee's claim [9], Standard-CombSum is in fact slightly better than Standard-CombMNZ.

Engines combined	Sum CombSum	ZMUV CombSum	Standard CombSum	Sum CombMNZ	ZMUV CombMNZ	Standard CombMNZ
inq102	0.4226	0.4226	0.4226	0.4226	0.4226	0.4226
inq102, citya1	0.4551	0.4445	0.4495	0.4503	0.4382	0.4464
inq102, citya1, brkly7	0.4807	0.4593	0.4742	0.4742	0.4553	0.4700
inq102, citya1, brkly7, inq101	0.4715	0.4471	0.4683	0.4677	0.4438	0.4648
inq102, citya1, brkly7, inq101, assctv2	0.4773	0.4500	0.4750	0.4713	0.4455	0.4692
average	0.4614	0.4447	0.4579	0.4572	0.4411	0.4546

Table 2: Non-interpolated precision of TREC-3's top 5 engines

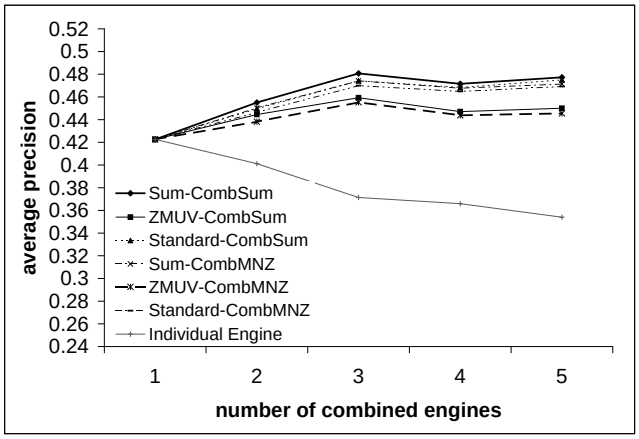

Figure 1: Average precision graphs for the different normalization schemes suggested by Montague and Aslam for combining the best five engines from TREC-3

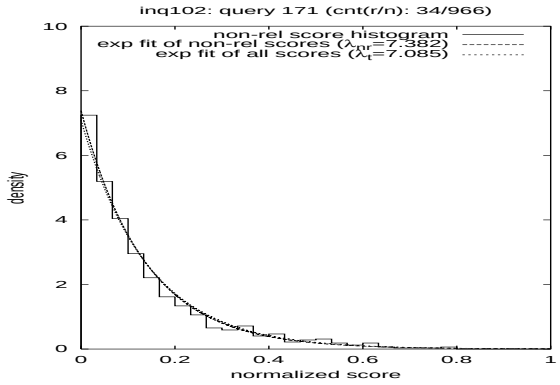

Figure 2: Histogram, exponential fit to non-relevant data and fit for entire data for query 171 INQUERY (inq102)

relevant score distributions of the outputs of different search engines. We show that the 'Sum' normalization is a good way of approximately normalizing the non-relevant score distributions of different search engines. We also show how better approximations may be obtained.

4. FUSION BY EQUALIZING DISTRIBUTIONS

Figure 2 shows a histogram of scores for query 171 for INQUERY from TREC-3. Note that the data are first normalized so that the minimum and maximum score for a query are 0 and 1 respectively. A maximum-likelihood fit of an exponential curve to this data is also shown (see [11] for more details).

The probability density of the exponential distribution is described by the following equation:

$$p(x) = \lambda \exp(-\lambda x) \tag{1}$$

and its mean = $1/\lambda$.

We note that a large variety of search engines based on different principles have non-relevant score distributions described by an exponential (see [11]).

The non-relevant distribution provides information as to how the scores would be distributed even if there are no relevant documents. That is, it provides information as to how a search engine maps a random set of documents to scores for a given query. Relevant documents will tend to get picked more easily if they stand out against this baseline distribution.

Since a good normalization scheme should ensure that random documents are mapped in the same manner, it is appropriate to normalize the random distributions. For a non-relevant distribution which is exponential, this can be done by simply setting the minimum score to 0 and the means of the exponentials to be the same. The retrieved lists from different search engines are now merged based on the normalized scores and CombSum. The results obtained by this technique (top curve labeled EXPML-CombSum) are the best as seen in Figure 3. The figure shows the average precision for the best 5 individual engines from TREC-3 (the individual engine graph). Each combination technique is labeled using the normalization technique followed by the combination technique used. For example, the CombMNZ technique with Standard normalization (Table 1) is referred to as Standard-CombMNZ. This was advocated by Lee [9] as the best fusion technique. EXPML-CombSum provides an improvement of almost 3.5% over Standard-CombMNZ in terms of average precision.

In practice, relevance information is not available. Hence, the non-relevant distribution has to be estimated without relevance information. We discuss three different ways of doing this below.

4.1 Estimate Using Mixture Model Fit

Manmatha et al [11] showed that one could fit a mixture model consisting of an exponential and a Gaussian to the score distributions using Expectation Maximization. This model applies to many different search engines.

The density of a mixture model $p(x)$ can be written in terms of the densities of the individual components $p(x|j)$ as follows: [4, 12]

$$p(x) = \sum_j P(j)p(x|j) \tag{2}$$

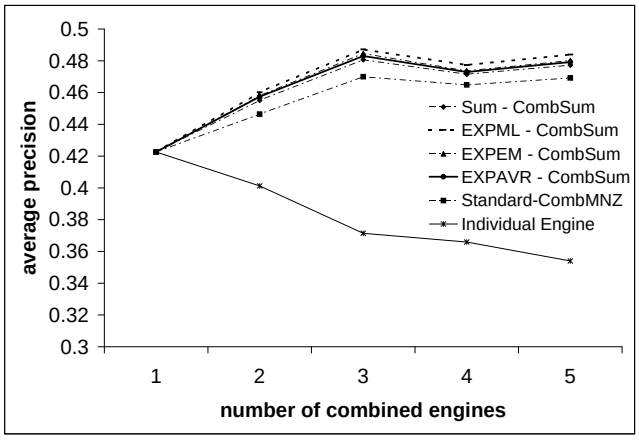

Figure 3: Average precision graphs for combining the best five engines from TREC-3

where j identifies the individual component, the $P(j)$ are known as mixing parameters and satisfy

$$\sum_{j=1}^{2} P(j) = 1, 0 \leq P(j) \leq 1. \quad (3)$$

We will denote densities with a lower case p(x) and probabilities with an uppercase P(x). In the present case, there are two components, an exponential density with mean λ

$$p(x|1) = \lambda \exp(-\lambda x) \quad (4)$$

and a Gaussian density with mean μ and variance σ^2

$$p(x|2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \quad (5)$$

The Gaussian part is an estimate of the relevant distribution while the exponential part of the mixture may be used as an estimate of the non-relevant distribution. The mixture model may be solved using Expectation-Maximization which is an iterative procedure to update the parameter values given some initial estimates. The reader is referred to [10] for details on the solution of this mixture model. The parameters of the mixture may be solved using the following update equations:

$$\mu^{new} = \frac{\sum_n P^{old}(2|x^n)x^n}{\sum_n P^{old}(2|x^n)} \quad (6)$$

$$(\sigma^{new})^2 = \frac{\sum_n P^{old}(2|x^n)||x^n - \mu^{new}||^2}{\sum_n P^{old}(2|x^n)} \quad (7)$$

$$\lambda^{new} = \frac{\sum_n P^{old}(1|x^n)}{\sum_n P^{old}(1|x^n)x^n} \quad (8)$$

$$P(1)^{new} = \frac{1}{N}\sum_n P^{old}(1|x) \quad (9)$$

Note that the updated Gaussian parameters are given by μ^{new} and σ^{new} while the updated exponential parameter is given by λ^{new}. $P(1)^{new}$ is the new estimate of the mixing parameter for the exponential ($P(2) = 1 - P(1)$). $P^{old}(j|x^n)j^n$ are estimates of the posterior for component j (j = 1,2) given the current estimates of the parameters.

For our purpose here, we use the exponential part of the mixture as an estimate of the non-relevant distribution. Equalizing the non-relevant distributions of the different search engine outputs is equivalent to equalizing the exponential components of the mixtures describing their outputs. That is, the scores of all documents are rescaled to ensure that the means of the exponential components are the same. In practice, this may be done for a given query on a search engine by by simply dividing the score of every document by the mean of the exponential component. The results obtained using their technique (labeled EXPEM-CombSum) are shown in Figure 3. EXPEM-CombSum is the second curve from the top and the closest in performance to EXPML-CombSum (our theoretical ideal). For various reasons, the exponential distribution obtained using the EM algorithm doesn't converge exactly to the non-relevant distribution.

4.2 Estimate Using Total Distribution

Figure 2 also shows an exponential fit to the scores of *all* documents. This plot is close to the exponential fit to the non-relevant documents. The distribution of all documents is often a good approximation to the distribution of non-relevant documents because the proportion of relevant documents is small in many cases (in this case 34 out of a 1000)

Figure 4 shows a plot of the difference of the means of the two distributions for all 50 queries of INQUERY (depicted as the t - nr curve). This difference is always positive and for many of the queries, but not all, it is small. Figure 4 also shows the difference between means of the exponential model obtained from the mixture model and the one obtained from the non-relevant document scores (labeled as em - nr).

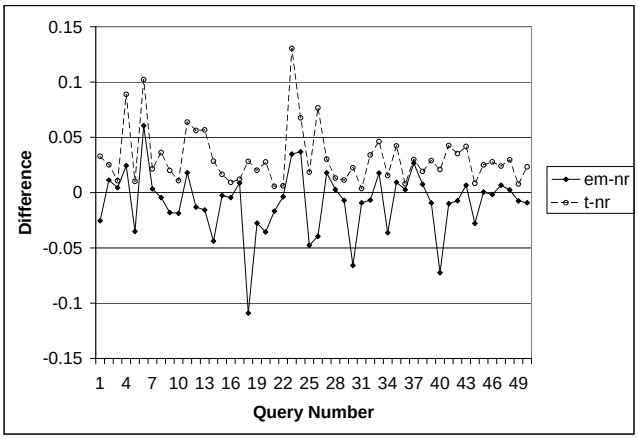

Figure 4: Differences between the means estimated using the two approximation techniques and the mean of the non-relevant distribution for INQUERY (inq102) for TREC-3. em, nr and t refer to the means obtained from the mixture model, non-relevant distribution and the total distribution respectively. Note that the t - nr curve is always positive and that the two differences are un-correlated.

Normalizing using the distribution of all scores is equivalent to setting the minimum to zero and equalizing the mean of all the scores. It is straightforward to show that this is equivalent (in terms of document ranking) to the Sum normalization scheme proposed by Montague and Aslam [13] (see Table 1). The plot labeled as Sum-CombSum in Figure 3 has the fourth best performance (but is really close to the optimal EXPML-CombSum). Our approach here provides the theoretical justification for one of Montague and Aslam's normalization schemes.

Note that for some datasets, Sum-CombSum is better than EXPEM-CombSum as shown in Figure 5 (which shows the average precision graphs for combining the best five automatic runs from the ad hoc track of TREC-4).

The last normalization scheme proposed by Montague and Aslam may be viewed as normalizing the means and variances of two Gaussian distributions. Since we know that the distribution of all scores is definitely not a Gaussian this should not work very well. In fact, the ZMUV-CombMNZ and ZMUV-CombSum are the worst performing combination techniques in Figure 1 and Table 2 and in Figure 5 the performance of ZMUV-CombMNZ actually gets worse as one combines more engines. One could not have predicted this without understanding the nature of the score distributions. We note that in a related area (TDT), this is a common technique for score normalization [6] - and probably a bad choice.

4.3 Estimate by Averaging

Figure 4 shows that the two estimates of the non-relevant distribution often err in different directions. Another estimate can, therefore, be obtained by averaging the two estimates. Normalizing and combining produces the EXPAVG-CombSum plot in Figures 3 and 5. We see that this lies between EXPEM-CombSum and Sum-CombSum. This gives the best consistent performance over all datasets. Note that for Trec-4 we take the top five automatic runs from TREC-4 to show that meta search also works in such cases.

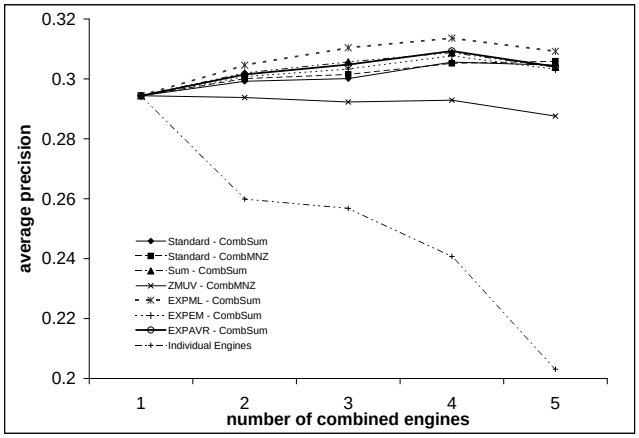

Figure 5: Average precision graphs for combining the best five automatic runs for the ad hoc track from TREC-4

Figure 6 shows the results from the Web Track of TREC-9 and again the best results are obtained by using EXP-ML.

4.4 Discussion of Results

The approach discussed here of equalizing the non-relevant score distributions of different search engines and then averaging the resulting scores clearly produces the best meta search results reported so far. Three different ways of estimating the non-relevant distribution have been discussed above but clearly the easiest technique involves estimating using the total distribution (equivalently, the Sum normalization). We have compared the results with some of the standard ad hoc normalization and combination techniques above. It is interesting to compare the results to a technique based on mapping scores to probabilites and then combining the probabilities as follows.

The mixture model may be used to compute a posterior proba-

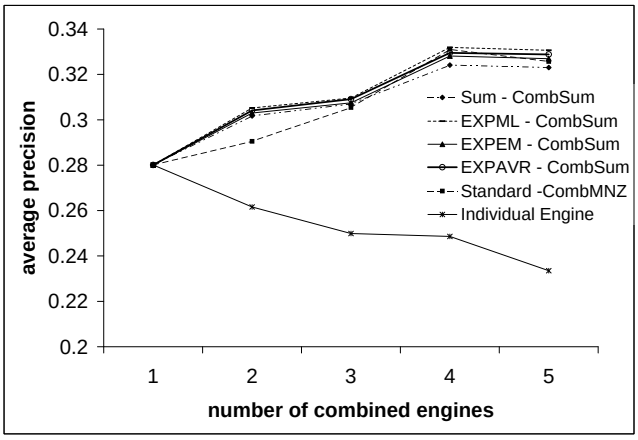

Figure 6: Average precision graphs for combining five runs from the manual interactive track from TREC-9

bility of relevance given the score (see [11]. The different search engines may then be combined by averaging the probabilities. In practice, combination by averaging posterior probabilities does not work as well as the technique proposed here involving equalizing the non-relevant distributions. We believe that the reason for this is the error involved in estimating the mixture model. For estimating the posterior probability both the relevant (Gaussian) and non-relevant (exponential) components of the mixture need to be estimated. As we have seen above, there are errors in estimating the exponential component. There are also errors in estimating the Gaussian component. The Gaussian component is also harder to estimate because the number of relevant documents could be small. Since the posterior probability of relevance given score needs a good estimate of the Gaussian component, there can be errors in it.

In the case of multi-lingual retrieval and distributed retrieval, the problem of estimating the Gaussian component for some databases may be really severe. Imagine, for example a French news database, an English news database and an Indonesian news database. In response to a query about the mayor of Nice, the French database may have a number of relevant documents while the Indonesian database may have no relevant documents. In such a situation, the estimates of the posterior probability would be especially poor. However, the techniques described here would still work (the exponential component of the mixture would still be estimated reasonably well).

One could also ask whether equalizing the Gaussian (relevant) distributions works. There does not appear to be any obvious intuitive reason for the relevant distributions to have the same parameters (although they may all belong to the same family of distributions). In practice, this does not work very well as expected.

Combining multiple search engines usually provides substantial improvements over even the best individual search engine. There are exceptions to this rule. For example, when the performance of one search engine is much better than another search engine then combination may not improve performance. Although combining the top few search engines usually improves performance, combining more than about 5 engines does not seem to cause a substantial improvement in performance and may in fact cause degradation. As is clear from the previous figures, as one goes further down the list, the performance of the individual search engines get substantially worse. An interesting question is how many engines should

be combined before the performance shows no further improvement.

5. CONCLUSION AND FUTURE WORK

We have described a formal approach to meta search which involves equalizing the distributions of scores of the top non-relevant score distributions of different search engines on a per query basis.' Three different techniques were described for estimating the non-relevant distribution. It was shown that the techniques described here produce the best reported meta search performance in the literature. The approach described here also provides a formal basis for one of the normalization schemes previously described in the literature.

Future work could include the extension of this technique to the combination of results from search engines operating on databases in different databases to produce a multi-lingual search engine and on combining the results from different databases in distributed retrieval.

The problem of score normalization is also important in other information retrieval tasks like topic detection and tracking and the discussion in this paper could provide some insight in such areas.

6. REFERENCES

[1] A. Arampatzis and A. van Hameren. The score-distributional threshold optimization for adaptive binary classification tasks. In *In the Proc. of the 24th ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 285–293, Sept 2001.

[2] J. Aslam and M. Montague. Models for metasearch. In *In the Proc. of the 24th ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 276–284, Sept 2001.

[3] J. A. Aslam, , and M. Montague. Bayes optimal metasearch: A probabilistic model for combining the results of multiple retrieval systems. In *the Proc. of the 23rd ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 379–381, 2000.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[5] W. B. Croft. Combining approaches to information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, pages 1–36. Kluwer Academic Publishers, 2000.

[6] J. Fiscus and G. R. Doddington. Topic detection and tracking evaluation overview. In J. Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 17–31. Kluwer Academic Publishers, 2002.

[7] E. Fox and J. Shaw. Combination of multiple searches. In *the Proc. of the 2nd Text Retrieval Conference (TREC-2)*, pages 243–252. National Institute of Standards and Technology Special Publications 500-215, 1994.

[8] J. H. Lee. Combining multiple evidence form different properties of weighting schemes. In *the Proc. of the 18th Intl. Conf. on Research and Development in Information Retrieval (SIGIR'95)*, pages 180–188, 1995.

[9] J. H. Lee. Analyses of multiple evidence combination. In *the Proc. of the 20th Intl. Conf. on Research and Development in Information Retrieval (SIGIR'97)*, pages 267–276, 1997.

[10] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for meta search. *Submitted to ACM TOIS*.

[11] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *the Proc. of the 24th ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 267–275, Sept 2001.

[12] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley, 2000.

[13] M. Montague and J. Aslam. Relevance score normalization for metasearch. In *In the Proc. of the ACM Tenth International Conference on Information and Knowledge Management (CIKM)*, pages 427–433, Nov 2001.

[14] C. Vogt and G. Cottrell. Predicting the performance of linearly combined IR systems. In *the Proc. of the 21st ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 190–196, 1998.

[15] Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. In *In the Proc. of the 24th ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 294–302, Sept 2001.