Modeling Score Distributions for Combining the Outputs of Search Engines

R. Manmatha, T. Rath and F. Feng *
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003
[manmatha,trath,feng]@cs.umass.edu

ABSTRACT

In this paper the score distributions of a number of text search engines are modeled. It is shown empirically that the score distributions on a per query basis may be fitted using an exponential distribution for the set of non-relevant documents and a normal distribution for the set of relevant documents. Experiments show that this model fits TREC-3 and TREC-4 data for not only probabilistic search engines like INQUERY but also vector space search engines like SMART for English. We have also used this model to fit the output of other search engines like LSI search engines and search engines indexing other languages like Chinese.

It is then shown that given a query for which relevance information is not available, a mixture model consisting of an exponential and a normal distribution can be fitted to the score distribution. These distributions can be used to map the scores of a search engine to probabilities. We also discuss how the shape of the score distributions arise given certain assumptions about word distributions in documents. We hypothesize that all 'good' text search engines operating on any language have similar characteristics.

This model has many possible applications. For example, the outputs of different search engines can be combined by averaging the probabilities (optimal if the search engines are independent) or by using the probabilities to select the best engine for each query. Results show that the technique performs as well as the best current combination techniques.

*This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, in part by the National Science Foundation under grant numbers IRI-9619117 and IIS-9909073, in part by NSF Multimedia CDA-9502639 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).

submitted to ACM SIGIR'01 New Orleans, Louisiana, USA

Categories and Subject Descriptors

[**Formal Models, Fusion/Combination, Modeling of score distributions**]

1. INTRODUCTION

In this paper we model score distributions of text search engines using a novel approach. We first show that 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. We further show that when relevance information is not available, these distributions can be recovered by fitting a mixture model with a Gaussian and an exponential component to the output scores of search engines on a per query basis. This novel approach to score modeling is then used to map the scores to probabilities using Bayes' Rule. Note that no training is required for this approach and in addition no assumption is made on the kind of search engine used. The model has been shown to work for a large number of search engines on TREC-3 and TREC-4 data including probabilistic search engines like INQUERY and vector space search engines like SMART. This model has also been shown to work for other engines like the LSI search engine and the score distributions of TREC-6 INQUERY and SMART data on Chinese. To our knowledge, this is the first attempt at recovering the relevant and non-relevant distributions when no relevance information is available.

The probabilities of relevance obtained from this model have many possible applications. For example thresholds for filtering may be selected using this approach or the probabilities may be used to combine the search from many distributed databases or multi-lingual or multi-modal databases. Here, we will focus on using them to combine the outputs of different search engines (the meta-search problem).

Most combination methods proposed in the literature are ad-hoc in nature and often involve the linear combination of scores [6]. This is unsatisfactory as scores from different search engines can be very different since they are often the result of computing some metric (or non-metric) distance over sets of features. Both the distance and the features may vary from engine to engine. In fact even the distributions of scores of different search engines could vary widely - for example, the scores of relevant documents may be clumped together for one engine while for those of a second engine may be distributed in a much more uniform manner. A linear combination of results in such cases could lead to misleading results. The problem is more acute when search engines operating on different media have to be combined for then the scores really

mean different things.

The approach proposed here allows us to combine the outputs of search engines using the probabilities derived from the model of score distributions. In this paper we examine two approaches to combination. The first involves averaging the probabilities which is optimal in the sense of minimizing the Bayes' error if the search engines are treated as independent classifiers [18]. The second approach involves using the probabilities to discard "bad" engines while keeping the "good" ones. We show that the combination approaches proposed using these techniques do as well as the best combination techniques proposed in the literature. In addition, our technique is less ad-hoc and easier to justify. The technique can also be extended to multi-lingual and multi-modal combination.

The rest of the paper is divided as follows. Section 2 discusses prior work in modeling score distributions as well as in the area of combination. This is followed by Section 3 which discusses the modeling of score distributions of relevant and non-relevant documents and how these distributions may be recovered in the absence of relevance information by using a mixture model. Solving for the mixture model using Expectation-Maximization (EM) is also discussed. Finally, Bayes' Rule is used to map the scores to probabilities of relevance. Section 4 discusses the theoretical intuition behind using such models. Section 5 discusses how the model and the probabilities derived from it can be used for evidence combination. Finally, Section 6 concludes the paper.

2. PRIOR WORK

Note that it is not obvious that the non-relevant data should be fitted with an exponential and the relevant data with a Gaussian. A number of researchers in the 60's and 70's starting with Swets [17] proposed fitting both the relevant and non-relevant scores using normal distributions and then using statistical decision theory to find a threshold for deciding what was relevant. Bookstein [3] pointed out that Swets implicitly relied on an equal variance assumption. Bookstein also raised the issue of whether it might be more appropriate to model the score distributions using Poissons. This modeling does not appear to have been done. van Rijsbergen [19] commented that for search engines like SMART there was no evidence that the distributions were similarly distributed let alone normally. We observe here that the empirical data for a large number of search engines clearly shows that the two distributions are not similar. All of the previous work here assumes that relevance information is available. To our knowledge, there is no literature on recovering the relevant and non-relevant distributions when no relevance information is available and ours is the first attempt to do this.

A recent and extensive survey of evidence combination in information retrieval is provided by Croft [6]. Tumer and Ghosh [18] discuss past work in a related area - the combination of classifiers. They show that for classifiers which are statistically independent, the optimal combination is obtained by averaging the probabilities. They define optimality as equivalent to minimizing the Bayes' error.

Fox and Shaw [9] 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 [12, 13] 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 [20] 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 [1]. 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.

A number of people have also looked at the problem of combining outputs of systems which search overlapping or disjoint databases. Voorhees et al [21] experimented with combination using a set of learned weights. Callan [4] gave a value to each database. He showed that weighting the scores by this value was substantially better than interleaving ranks. Some researchers have also investigated the notion of combining search engines over multiple media. QBIC [8] combined scores from different image techniques using linear combination. Fagin [7] used standard logical operators like MIN and MAX to combine scores in a multimedia database. However, Lee's experiments showed (at least for text) that these operators perform significantly worse than averaging).

3. MODELING SCORE DISTRIBUTIONS OF SEARCH ENGINES

In this section we describe how the outputs of different search engines were modeled using data from the text retrieval conferences (TREC). TREC data provides the scores and relevance information for the top 1000 documents for different search engines. For the experiments here data from the ad hoc track of the TREC-3 and TREC-4 for a number of different search engines was used. We will show examples of the modeling on queries from INQUERY and SMART from TREC-3. INQUERY is a probabilistic search engine from the University of Massachusetts, Amherst while Smart is a vector space engine from Cornell University.

Our modeling begins with TREC-3 data for INQUERY. There are 50 queries available with document scores and relevance information for each query. We examine the relevant and non-relevant data separately. The data are first normalized so that the minimum and maximum score for a query are 0 and 1 respectively. Figure 1 shows a histogram of scores for query 151 from TREC-3 for the relevant data. The histogram approximates a normal distribution. The plot also shows a maximum-likelihood fit using a Gaussian with mean 0.466 and variance 0.042. The maximum-likelihood fit involves setting the mean and variance of the Gaussian to the sample mean and sample variance respectively of the data [2]. The Kolmogorov-Smirnov (KS) test for significances shows that we cannot eliminate the null hypothesis that the distribution is a Gaussian. In other words, a Gaussian is not inconsistent with the data.

Figure 2 shows a histogram of scores for the set of non-relevant documents for the same query. The histogram clearly shows the rapid fall in the number of non-relevant documents with increasing score. A maximum-likelihood fit of an exponential curve to this data is also shown. For the purposes of fitting the exponential, the origin is shifted to the document with the lowest score. It can be

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

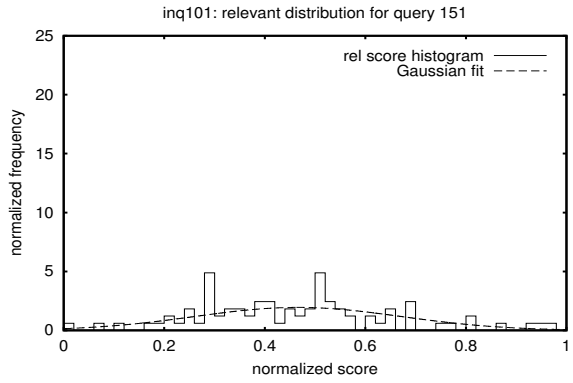

Figure 1: Histogram and Gaussian fit to relevant data for query 151 INQUERY (inq101)

shown that the maximum-likelihood for an exponential is obtained by setting the mean of the exponential to the sample mean of the data [2].

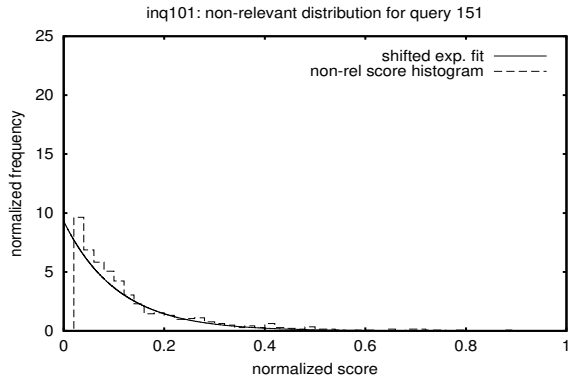

Figure 2: Histogram and shifted exponential fit to non-relevant data for query 151 INQUERY (inq101)

The same process was repeated for all 50 queries in this track and in most of those cases it was found possible to fit the non-relevant data with exponentials and the relevant data using Gaussians. The relevant data can be fitted with a Gaussian reasonably well when there is a sufficient number of relevant documents. Usually more than 60 relevant documents were needed. When the number of relevant documents was small, the fit was bad. However, we believe this is not because the Gaussian was a bad fit but because we don't have enough relevant documents to compute the statistics in these cases. The exponential was also a good fit to the non-relevant data.

We have so far been able to fit parametric forms to the score distributions given relevance information. When running a new query, however, relevance information is not available. Clearly, it would be useful to fit the score distributions of such data. A natural way to do this is to fit a mixture model of a shifted exponential and a Gaussian to the combined score distribution. This approach is discussed in the next section.

3.1 Mixture Model Fit

Consider the situation where a query is run using a search engine. The search engine returns scores but there is no relevance information available. We show below that in this situation, a mixture model consisting of an exponential and a Gaussian may be fitted to the score distributions. We can then identify the Gaussian with the distribution of the relevant information in the mixture and the exponential with the distribution of the non-relevant information in the mixture. Essentially this allows us to find the parameters of the relevant and non-relevant distributions without knowing relevance information apriori.

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: [2, 14]

$$p(x) = \sum_j P(j)p(x|j) \tag{1}$$

where j identifies the individual component, the $P(j)$ are known as mixing parameters and satisfy $\sum_{j=1}^{2} P(j) = 1, 0 \leq P(j) \leq 1$. In the present case, there are two components, an exponential density with mean λ

$$p(x|1) = \lambda \exp(-\lambda x) \tag{2}$$

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

$$p(x|2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{3}$$

A standard approach to finding the mixing parameters and the parameters of the component densities is to use Expectation Maximization (EM) [2, 14]. This is an iterative procedure where the Expectation and Maximization steps are alternated.

The EM procedure leads to the following update equations for the parameters:

$$\mu^{new} = \frac{\sum_n P^{old}(2|x^n)x^n}{\sum_n P^{old}(2|x^n)} \tag{4}$$

$$(\sigma^{new})^2 = \frac{\sum_n P^{old}(2|x^n)||x^n - \mu^{new}||^2}{\sum_n P^{old}(2|x^n)} \tag{5}$$

$$\lambda^{new} = \frac{\sum_n P^{old}(1|x^n)}{\sum_n P^{old}(1|x^n)x^n} \tag{6}$$

$$P(j)^{new} = \frac{1}{N}\sum_n P^{old}(j|x) \tag{7}$$

The procedure needs an initial estimate of the component densities and mixing parameters. Given that, it rapidly converges to a solution. Using EM to fit the data gives the mixture fit shown in Figure 4. The figure plots the mixture density as well as the component densities for the exponential and Gaussian fits. The adjacent figure (Figure 3) shows the exponential and Gaussian fits to the non-relevant and relevant data. Comparing the two figures, it appears that the strategy of interpreting the Gaussian component of the mixture with the relevant distribution and the exponential component of the mixture with the non-relevant distribution is a reasonable one. We should note that the correspondence between the mixture components and the fits to the relevant/non-relevant data is not always as good as that shown here but in general it is a reasonable fit.

The same technique was used to model the result of query 151 for the Cornell Smart vector space engine. Similar results were obtained as shown in Figures 5 and 6.

This model has been fitted to a large number of search engines on TREC-3 and TREC-4 data including probabilistic engines like INQUERY and CITY and a vector space engine (SMART) as well as Bellcore's LSI engine. The fit appears to be better for "good"

search engines (engines with a higher average precision in TREC-3) and worse for those with a lower average precision. The model has also been able fitted to document scores for searches on INQUERY and SMART indexing a Chinese database.

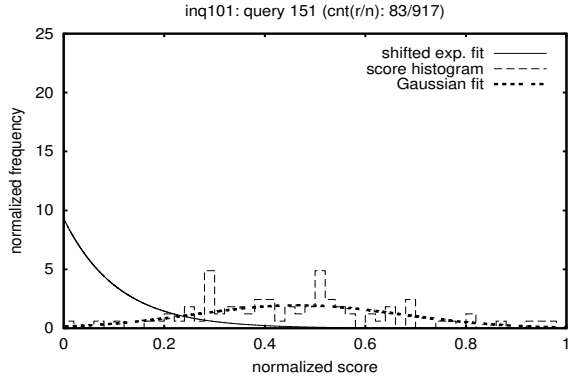

Figure 3: Exponential fit to non-relevant data and Gaussian fit to relevant data for query 151 INQUERY (inq101)

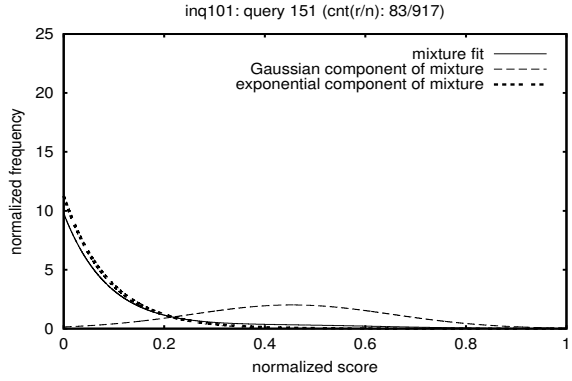

Figure 4: Mixture model fit showing exponential component, Gaussian component and the combined mixture for query 151 INQUERY (inq101). Compare with Figure3

3.2 Computing Posterior Probabilities

Using Bayes' Rule one can compute the probability of relevance given the score as

$$P(rel|score) = \frac{P(score|rel)P(rel)}{P(score|rel)P(rel) + P(score|nonrel)P(nonrel)} \quad (8)$$

where $P(rel|score)$ is the probability of relevance of the document given its score, $P(score|rel)$ and $P(score|nonrel)$ are the probabilities of score given that the document is relevant and score given that the document is non-relevant respectively. P(rel) and P(nonrel) are the prior probabilities of relevance and non-relevance.

In our model, $P(score|rel)$ is given by the Gaussian component of the mixture while $P(score|nonrel)$ is given by the exponential part of the mixture. P(rel) and P(nonrel) may be obtained by using the mixing parameters. Thus, $P(rel|score)$ can be computed in a simple manner.

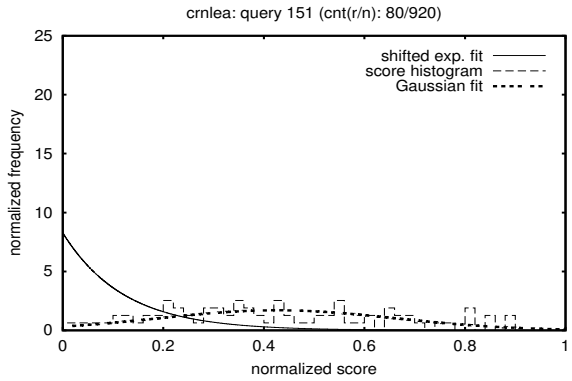

Figure 5: Exponential fit to non-relevant data and Gaussian fit to relevant data for query 151 SMART (crnlea).

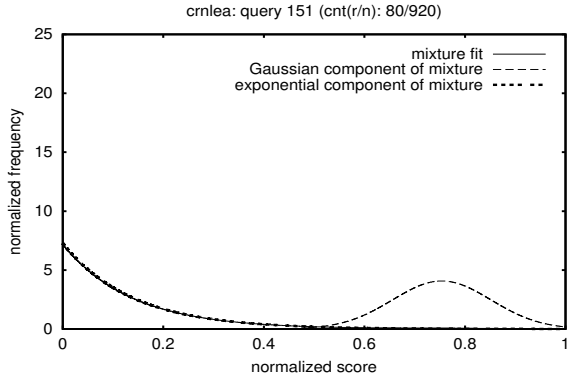

Figure 6: Mixture model fit showing exponential component, Gaussian component and the combined mixture for query 151 SMART (crnlea). Compare with Figure 5

Figure 7 shows the posterior probability obtained for SMART for query 164 using the fits shown in Figures 5 and 6. The figure shows the posterior probabilities obtained from the separate Gaussian and exponential fits when relevance information is available and also the posterior probabilities obtained from the Gaussian and exponential part of the mixture. P(rel) and P(nonrel) are taken to be the mixing parameters in this case. Note that the differences in the two curves reflect fitting errors both for the mixture fit as well as the separate Gaussian and exponential fits obtained when relevance information is available.

In general we expect the posterior probabilities to be a monotonic function of the score. In other words as the score increases so should the posterior probability. In some cases we may have the situation depicted in Figure 8 where the posterior seems to decrease with increasing scores. The figure depicts the posterior probabilities for INQUERY for query 154 using TREC-3 data. This situation arises because the Gaussian density falls much more rapidly than the exponential and hence the two densities intersect twice. Note that in this case the posterior probabilities obtained both from the mixture fit (no relevance information available) as well as that obtained using relevance data show this problem. One solution would be not to use a Gaussian density but to use another function which has the same form (like a Gamma distribution) but decreases less

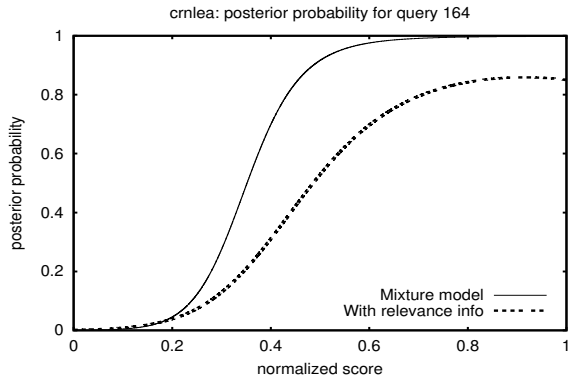

Figure 7: Posterior probability for query 164 for the SMART engine for TREC-3 data. The dotted line is obtained from the separate Gaussian and exponential fits computed using relevance information. The solid line is obtained from the mixture fits.

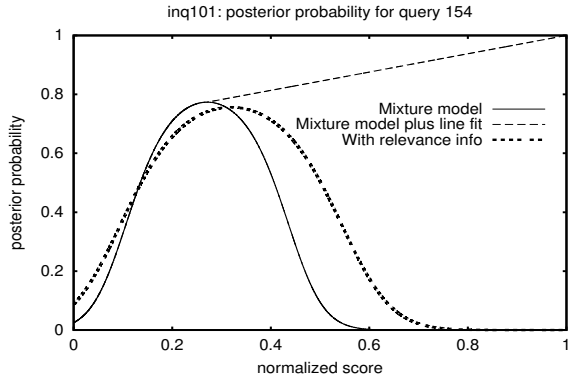

Figure 8: Posterior probability for query 154 for the INQUERY engine for TREC-3 data. The bold dotted line is obtained from the separate Gaussian and exponential fits computed using relevance information. The solid line is obtained from the mixture fits. The dotted line joins the maximum point of the mixture to the point(1,1). The final posterior mapping follows the solid line up to the maximum point and then the straight line curve thus preserving monotonicity

rapidly. As we discuss below the problem with this approach is that the mixture model does not converge to a reasonable solution. Instead we force the posterior probability to be monotonic by drawing a straight line from the point where the posterior is maximum to the point (1,1). The final posterior probability curve is given by the portion of the posterior probability computed using Bayes' rule up to the maximum portion of the curve and the straight line thereafter.

We have assumed that the priors P(rel) and P(nonrel) may be estimated using the mixing parameters. When there are few relevant documents the mixing parameters provide a poorer estimate of the priors. In a normal retrieval the number of relevant documents is small and hence estimates of the mixing parameters are less accurate. Extensive experiments have shown that P(nonrel) is best estimated using the following procedure. Let P(1) be the mixing parameter corresponding to the exponential. Then

$$P(nonrel) = \begin{cases} P(1) & if\ P(1) \leq 0.8 \\ 0.8 & otherwise \end{cases} \quad (9)$$

and P(rel) = 1 - P(nonrel). This approach to estimating the priors improves the average precision results slightly when we combine results.

3.3 Comments on Fitting Distributions and Mixture Models

There is a large family of densities which could possibly have fit the data. For example, the Poisson and Gamma distributions approximate the Gaussian for appropriate parameter choices. However, using a Poisson/Poisson (non-relevant/relevant) or an exponential/ Poisson combination did not fit the data well. On the other hand, while an exponential/Gamma fit the non-relevant and relevant data when separately fitted, a mixture fit with exponential and Gamma components did not converge to the right answer. In this case the Gamma component also converged to an exponential (the exponential density is a special case of the Gamma function). Thus our choice of distributions - exponential for the non-relevant and Gaussian for the relevant - is dictated by the consideration that the functions fit the data well and by the consideration that they can be recovered using a mixture model when relevance information is not available.

Like any non-linear equation solver, EM may find solutions arising from local maxima and is sometimes sensitive to initial conditions [14]. Different approaches to picking initial conditions were tried.

1. The first involved picking arbitrary initial conditions.

2. A second approach involved fitting an initial exponential to all the document scores (relevant and non-relevant). This is reasonable since there are far more non-relevant documents than relevant documents. Thus, the distribution of the scores of the combined documents essentially resembles that of the of scores of the non-relevant documents i.e. its an exponential. Scores of documents which do not fit the exponential are removed and fitted with a separate Gaussian. The exponential and Gaussian provide initial estimates of the parameters.

Some sensitivity to initial conditions was noticed but usually for the poorer search engines (search engines much lower down in a TREC-3 ranking by average precision).

4. SHAPE OF DISTRIBUTIONS

We will now attempt to give some insight into the shape of the score distributions.

Given a term (or word) assume that the distribution of this term in the set of relevant documents is given by a Poisson distribution with parameter λ_r. That is,

$$P_r(x) = \frac{\lambda_r^x \exp(-\lambda_r)}{x!} \quad (10)$$

where x is the number of times that the term occurs in a particular document and $P_r(x)$ is the probability of x occurrences of the term in the set of relevant documents. Also assume that its distribution in the set of non-relevant documents is given by another Poisson distribution with the parameter λ_n. That is,

$$P_n(x) = \frac{\lambda_n^x \exp(-\lambda_n)}{x!} \quad (11)$$

where $P_n(x)$ is the probability of x occurrences of the term in the set of non-relevant documents. In general, λ_n will be much smaller than λ_r.

Numerous attempts have been made to model word distributions in the past. Harter [11] used a mixture of 2 Poissons to model the distributions of words in a document. Our model in this section is closely related to his model. It has been argued by some researchers that the 2 Poisson model is not a good approximation and that other distributions like the negative binomial are better models of the distributions of words in documents [15]. A mixture of a large number of Poissons has also been used to fit the data [5]. Since we would like to fit a distribution to the relevant and another to the non-relevant, it is much more convenient for us to assume the 2-Poisson model here. Additionally, the main purpose of this section is to provide some insight and not a rigorous derivation.

Given a query consisting of 1 term and assuming that the score given to a document is proportional to the number of matching words in the document, the distribution of the scores of relevant documents is then given by the Poisson distribution:

$$P_r(x) = \frac{\lambda_r^x \exp(-\lambda_r)}{x!} \tag{12}$$

and the distribution of the scores of non-relevant documents is given by the Poisson distribution:

$$P_n(x) = \frac{\lambda_n^x \exp(-\lambda_n)}{x!} \tag{13}$$

The actual scores for many search engines is weighted by some function of the term frequency and the inverse document frequency. However, empirical evidence [10] shows that the score may be reasonably approximated as being proportional to the number of matching words.

For the set of relevant documents, λ_r will usually be large. For large values of λ, the Poisson distribution tends to a normal distribution (see Figure 9). On the other hand for small values of λ, the Poisson distribution will tend towards a distribution which is falling rapidly (see Figure 9). The shape of these curves is consistent with the experimental modeling of scores for TREC-3 and TREC-4 data (see the previous section). The experiments showed us that the normal distribution is a good fit for the score distributions of the relevant data. For non-relevant data, the experiments show that the exponential distribution is a good fit. For small λ_n, the Poisson distribution shows a decreasing distribution. Although, this is not the same as an exponential distribution, it does have the same general shape as an exponential (rapid monotonic decrease).

It is much harder to derive the score distributions if the query consists of two or more terms. This is because the actual scores of search engines are complicated functions. However, there is empirical evidence that the major contribution to the scores is provided by the number of matching terms [10]. We also note that Robertson and Walker [16] motivated a $tf - idf$ scoring function from the 2-Poisson model. We assume first that the score is proportional to the number of matching words and provide an intuitive argument for queries with two or more terms. For simplicity we will consider the case where the query has just two terms but the argument applies in general. In this case we can assume that the two terms say "oil" and "spill" can be clubbed together to create a single term - "oil spill". Then the λ_r (the average frequency of a term over relevant documents) for joint occurrences of this term "oil spill" is much lower than the λ_r for either "oil" or "spill". In other words the chances that the terms "oil spill" occur together is much less than that of finding "oil" or "spill". When the query contains two terms, it is reasonable to assume that the λ_n (i.e. the average frequency over non-relevant documents) does not change much as it essentially reflects the background probabilities of the word.

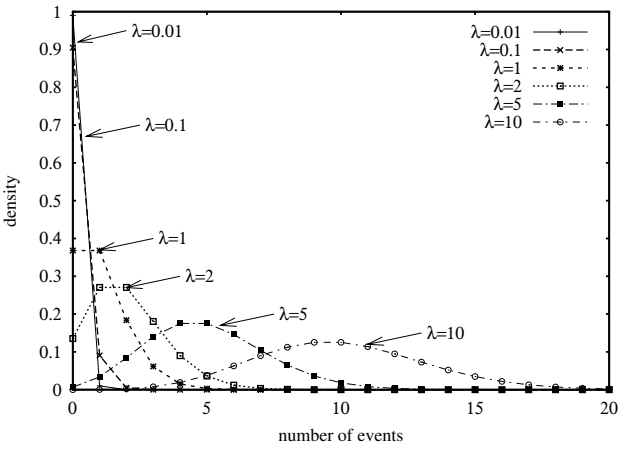

Figure 9: The Poisson distribution for different values of λ.

The Poisson model for the shape of the relevant and non-relevant distributions that we have derived applies to both probabilistic engines like INQUERY and vector space engines like SMART. For vector space engines the number of matching terms is given by the dot product of two vectors - one representing the query and the other the document. Further, this model is language independent (as long as word frequencies in any language have an approximately 2-Poisson like distribution). Thus, we predict that a mixture of exponential and Gaussian distributions will fit a much larger class of text search engines operating on different languages.

The model in this section although intuitive can be used to make a prediction. The model predicts that on a statistical basis as the number of query terms is increased the overall λ should decrease and hence the mean and variance of the Gaussian fitting the relevant documents should also decrease. Note that for any particular query the mean score can be arbitrary. However when a large enough sample of queries is considered, the mean query should decrease with the number of query terms.

It is a well known fact in information retrieval that with query expansion the score of the relevant documents decreases and the range also decreases which is consistent with this prediction. For the 50 queries from TREC-3 for INQUERY (inq101) and SMART (crnlea), we plotted the mean scores of the relevant data versus the number of query terms (including expanded queries). A small statistical decrease with the number of query terms was observed for INQUERY and SMART. The figures are not produced here because of a lack of space.

5. COMBINING SEARCH ENGINES

The posterior probabilities obtained by using the model discussed above has many possible applications. For example the probabilities could be used to select a threshold for filtering documents or for combining the outputs in distributed retrieval. Here we discuss one possible application which involves combining the outputs of different search engines on a common database to improve results.

It would be of considerable use to combine the output of different search engines. In this section we discuss how the score of search engines may be combined while taking into account the actual score distributions.

In general the range of search engine scores may vary quite a bit -

for example, one engine may have scores ranging from 0 to 1 while another can have scores ranging from -20 to 150. Other approaches to combining score distributions have focused on normalizing the range of the scores and then combining them by simple techniques like linear combination or by taking the minimum and maximum scores. However, range normalization does not take into account the actual distribution of the scores. Consider, for example, the model of the scores discussed previously where the scores of the relevant documents follow a normal distribution and the scores of the non-relevant follow an exponential distribution. Also consider two different search engines which have different parameter values for these distributions. A simple (linear) range normalization and combination does not clearly suffice.

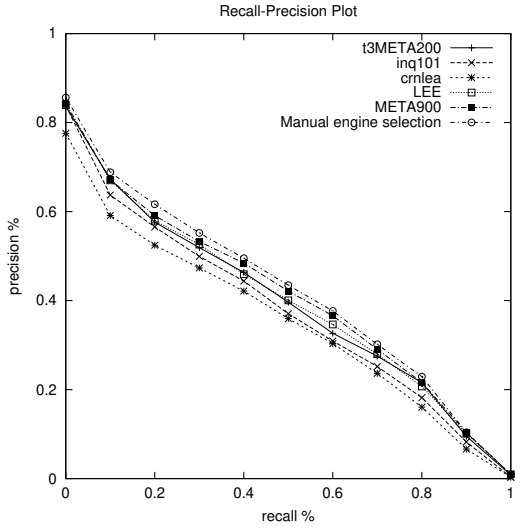

Figure 10: Recall precision graphs for combining inq101 and crnlea using different techniques (see text). Data from TREC-3

There are a number of possible ways the probabilities can be used to combine the search engines. We propose two alternative schemes for combination. The first involves averaging the probabilities. This is optimal in the sense of minimizing the Bayes' error if the search engines are independent [18]. Of course the outputs of search engines are not necessarily independent. In the following figures and discussions, data are taken from TREC-3. inq101 and crnlea denote runs of the INQUERY and SMART engines, META200 denotes combination by averaging the posteriors obtained using the mixture model, while META900 denotes the combination by averaging the posterior probabilities using the Gaussian and exponential fits assuming relevance is given. Thus, any difference between META200 and META900 is caused by the errors in performing a mixture fit to the model. LEE denotes Lee's COMBMNZ technique while the manual engine selection technique involves selecting and discarding the best engine (or engines) on a per query basis using the average precision for that query. Manual engine selection provides an indication of the best combination result we can achieve. Note that both META900 and manual engine selection require relevance information and are only plotted to provide a baseline for understanding the limits of combination.

Figure 10 shows recall-precision plots for combining INQUERY and SMART on TREC-3 data. Precision is defined as the fraction of retrieved documents which are relevant while recall is the fraction of relevant documents which have been retrieved. The recall-precision graph is usually created by averaging over a certain number of queries - in this case 50. As the figure shows META200 performs considerably better than either INQUERY and SMART - in fact about 6% better than INQUERY and 13% better than SMART. LEE is slightly better (about 1%) than META200 although the difference is not significant. META900 has an average precision about 10% better than INQUERY and clearly performs better than either META200 or LEE's implying that if the mixture fit could be improved the technique would perform even better. Finally, the plot for manual engine selection clearly indicates that both META200 and LEE's are close to obtaining the best performance possible from combination.

Figure 11 describes combination results for the top five engines in TREC-3. The x-axis is the number of engines combined while the y-axis is the average precision. As the plot clearly shows combination clearly improves the results. There are four graphs in the figure. The first curve is the average precision of the individual search engines. The second plot META200 shows the combination method applied to 1, 2, 3, 4 or 5 engines. As can be clearly seen there is a considerable improvement over using even the best search engine and overall the improvement seems to increase with the number of search engines combined. With the top 2 engines, META200 shows an improvement of 6% over the best single engine and using the top 3 engines, META200 shows an improvement of almost 12%. LEE's COMBMNZ technique is also shown in the same graph. It's average precision is seen to be slightly worse than META200 but the difference is not really significant. The performance obtained using META900 (i.e. combination with the posterior probabilities obtained with relevance information) is 15% better than the best single engine. Again this indicates that if the mixture fit were improved we could do even better.

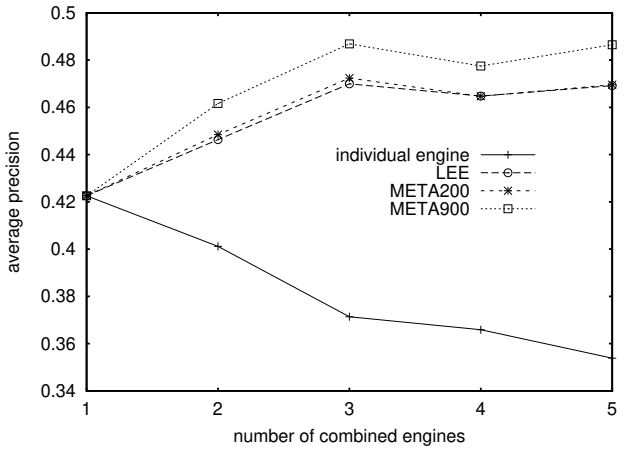

Figure 11: Recall precision graphs for combining the best five techniques from TREC-3.

Figure 12 demonstrates that this approach also works for other languages. The figure shows the combination results for INQUERY and SMART when indexing a Chinese database. The data in this case is from TREC-6. As can be clearly seen, combination using both META200 and LEE's COMBMNZ show an improvement over either engine. However, in this particular case the improvement is much less than that for English. Also the difference between META900 and META200 is small indicating that perhaps we are close to the limit of what can be achieved.

Combination of "good" search engines usually improves the scores.

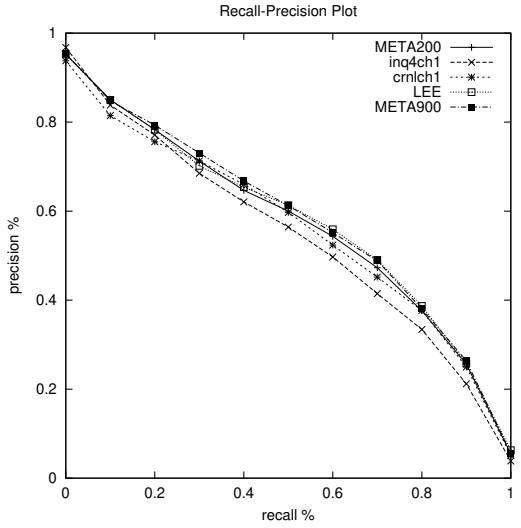

Figure 12: Recall precision graphs for combining inq4ch1 and crnlch1 with Chinese queries and Chinese databases. Data from TREC-6

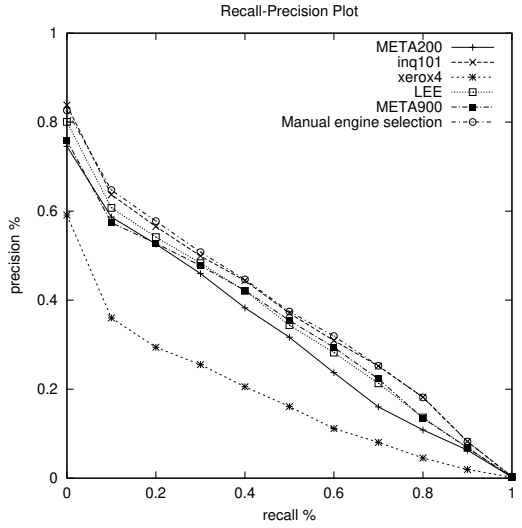

Figure 13: Recall precision graphs for combining inq101 and xerox4 using different techniques. Data from TREC-3.

Partly this reflects the fact that the score distribution models fit "good" search engines better than "poor" search engines. However, the combination of two search engines when the performance of one is substantially worse than the others leads to a result which can be worse than that of the best engine. This partly reflects the well known observation that combining a bad classifier with a good classifier can lead to a result which may not be better than the best individual classifier. Two search engines INQUERY (inq101) and XEROX (xerox4) were picked. On the basis of average precision, inq101 is ranked 4th while xerox4 is ranked 35th among 40 engines in TREC-3. The average precision of inq101 is more than twice that of xerox4. Figure 13 clearly shows that INQUERY (inq101) performs much better than the XEROX engine (xerox4). The combination META200 is much better than XEROX but worse than INQUERY. LEE's is slightly better than META200 is still worse than INQUERY. Clearly the best option in such cases is to avoid combination.

5.1 Automatic Engine Selection

The previous example shows that if we could have somehow figured out that we need to pick INQUERY as the best possible engine for every query then the performance would improve considerably. The ability to model and compute the relevant and non-relevant distributions allows us to develop techniques to automatically selecting engines on a per query basis. Here, we examine two such approaches.

The first approach essentially tries to ensure that the distance between the mean of the normal distribution and the point at which the densities intersect is large (all distributions are obtained using the mixture model). The idea is that if this distance is large then it will be easier to separate the relevant and non-relevant documents. If the distance is less than a threshold, the engine is discarded for that query. The posterior probability of all engines selected (i.e. not discarded) for a particular query are averaged to obtain document probabilities as before. If all engines for a particular query are below the threshold, then the one with the highest posterior probability is selected. The threshold is selected based on empirical data to be 0.16. The results for this technique are labelled as META206 in Figure 14

The second approach uses the posterior probabilities obtained using Bayes' rule from the mixture components. In some situations the maximum of the posterior probability is quite small. A posterior of 0.5 indicates that the relevant and non-relevant distributions weighted by the priors are of equal magnitude. In other words a posterior of 0.5 indicates the point at which the exponential and Gaussian densities intersect after weighting by the prior. It is clear that a "good" engine should preferably have a higher posterior. Empirically if the posterior for a particular engine and a particular query was less than 0.7 then that engine was regarded as poor and discarded for that particular query. If both engines had maximum posteriors greater than 0.7 then they were averaged. If neither engine had a maximum posterior greater than 0.7 both were again averaged and combined. The results for this technique are plotted as META207.

Figure 14 shows the results of combining two engines whose performance is very different. We again use inq101 and xerox4. As is clear from Figure 14, META206 and META207 perform about equally well and and both are better than META200 (straight averaging of posterior probabilities). The average precision of META206 and META207 are essentially the same as LEE's. We have also carried out other experiments with other engines all of which demonstrate that engine selection can be done using the models of score distributions discussed here. We note that both META206 and META207 are still worse than using INQUERY alone indicating that there is further scope for improving the engine selection procedure. Of course, this also implies that when one search engine performs much worse than another it may be best not to use the "poor" search engine.

5.2 Discussion of Combination Results

The results above show that the mixture modeling performs as well as the best current techniques (Lee's) available for combination. There is scope for a slight improvement in estimating the mixture parameters as well using that for obtaining better combination. Of course it is also clear that we are approaching the limits

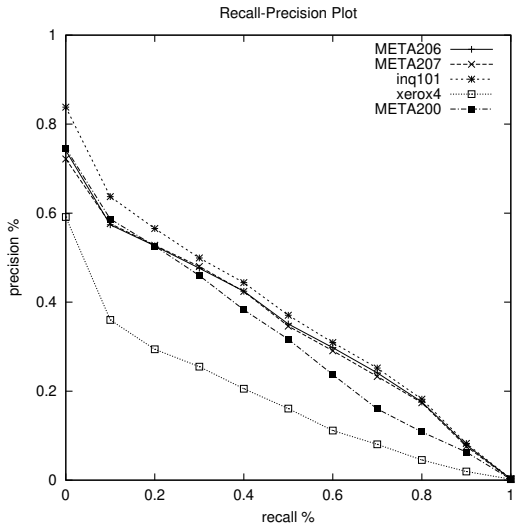

Figure 14: Recall precision graphs for combining inq101 and xerox4 by automatically selecting the best engine using probabilities and distributions.

of the best performance we can achieve.

Lee's COMBMNZ technique performs surprisingly well. In the case where INQUERY and SMART are combined we note that for many queries INQUERY and SMART have distributions which are very similar. In such a situation, their posterior distributions will also look remarkably similar and hence averaging is a good strategy. Since COMBSUM involves computing a document score by just adding the scores for all engines which find that document, it will produce the same ranking as averaging and hence it will also be good. COMBMNZ involves multiplying COMBSUM by the number of engines which found that document and hence it will also produce good results. In this particular situation COMBMNZ involves essentially combining the posterior probabilities without having to do the mixture modeling. However, in the more general case, the good performance of COMBMNZ is hard to explain.

The model for combination proposed here is more intuitively satisfying for a number of reasons. First, it combines engines in a natural way using probabilities and is therefore easier to explain. Second, it indicates where improvements can be made for better performance. Third, the same technique may be used for combining multi-lingual engines. It will also extend to multi-modal engines provided the distributions of scores behave in a similar way for search engines indexing other media.

6. CONCLUSION

We have demonstrated how to model the score distributions of a number of text search engines. Specifically, it was shown empirically that the score distributions on a per query basis may be fitted using an exponential distribution for the set of non-relevant documents and a normal distribution for the set of relevant documents.

It was then shown that given a query for which relevance information is not available, a mixture model consisting of an exponential and a normal distribution may be fitted to the score distribution. These distributions were used to map the scores of a search engine to probabilities.

The model of score distributions was used to combine the results from different search engines to produce a meta-search engine. The results were substantially better than either search engine provided no "search engine" performed really poorly. Different combination techniques were proposed including averaging the posterior probabilities of the different engines as well as using the probabilities and distributions to selectively discard some engines on a per query basis.

Future work will include attempts to further improve the modeling for better performance. Other possible applications of modeling score distributions like filtering will also be examined. Finally we will also examine the possibility that search engines indexing other media like images can also be modeled in the same way.

7. ACKNOWLEDGEMENTS

The origins of this work can be traced back to a discussion at the University of Glasgow. The first author would like to thank Bruce Croft for extensive discussions and also for pointing out relevant literature especially previous literature on score modeling in information retrieval.

8. REFERENCES

[1] 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.

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

[3] A. Bookstein. When the most "pertinet" document should not be retrieved - an analysis of the swets model. *Information Processing and Management*, 13:377–383, 1977.

[4] J. Callan, Z. Lu, and W. B. Croft. Trec and tipster experiments with inquery. In *the Proc. of the 18th ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 21–28, 1995.

[5] K. W. Church and W. A. Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

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

[7] R. Fagin. Fuzzy queries in multimedia database systems. In *the Proc. of the 17th ACM Conference on Prnciples of Database Systems (PODS)*, pages 1–10, 1998.

[8] M. Flickner, H. S. Sawhney, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer Magazine*, 28(9):23–30, Sept. 1995.

[9] 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.

[10] W. Greiff. The use of exploratory data analysis in information retrieval research. In W. B. Croft, editor, *Advances in Information Retrieval*, pages 37–72. Kluwer Academic Publishers, 2000.

[11] S. P. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 20:197–206, 1975.

[12] 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.

[13] 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.

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

[15] F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison Weseley, 1964.

[16] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *the Proc. of the 17th ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 232–241, 1994.

[17] J. A. Swets. Information retrieval systems. *Science*, 141:245–250, 1963.

[18] K. Tumer and J. Ghosh. Linear and order statistics combiners for pattern clasification. In A. Sharkey, editor, *Combining Artificial Neural Networks*, pages 127–162. Springer-Verlag, 1999.

[19] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[20] 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.

[21] E. Voorhees, N. Gupta, and B. Johnson-Laird. Learning collection fusion strategies. In *the Proc. of the 18th ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 172–179, 1995.