The Maximum Entropy Method for Analyzing
Retrieval Measures

Javed A. Aslam  Emine Yilmaz  Virgiliu Pavlu
College of Computer and Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115
{jaa,emine,vip}@ccs.neu.edu

ABSTRACT
We present a model, based on the maximum entropy method, for analyzing various measures of retrieval performance such as average precision, R-precision, and precision-at-cutoffs. Our methodology treats the value of such a measure as a constraint on the distribution of relevant documents in an unknown list, and the maximum entropy distribution can be determined subject to these constraints. For good measures of overall performance (such as average precision), the resulting maximum entropy distributions are highly correlated with actual distributions of relevant documents in lists as demonstrated through TREC data; for poor measures of overall performance, the correlation is weaker. As such, the maximum entropy method can be used to quantify the overall quality of a retrieval measure. Furthermore, for good measures of overall performance (such as average precision), we show that the corresponding maximum entropy distributions can be used to accurately infer precision-recall curves and the values of other measures of performance, and we demonstrate that the quality of these inferences far exceeds that predicted by simple retrieval measure correlation, as demonstrated through TREC data.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and Software – Performance evaluation

General Terms
Theory, Measurement, Experimentation

Keywords
Evaluation, Maximum Entropy, Average Precision

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR ’05, August 15–19, 2005, Salvador, Brazil.
Copyright 2005 ACM 1-59593-034-5/05/0008 ...$5.00.

1. INTRODUCTION
The efficacy of retrieval systems is evaluated by a number of performance measures such as average precision, R-precision, and precisions at standard cutoffs. Broadly speaking, these measures can be classified as either system-oriented measures of overall performance (e.g., average precision and R-precision) or user-oriented measures of specific performance (e.g., precision-at-cutoff 10) [3, 12, 5]. Different measures evaluate different aspects of retrieval performance, and much thought and analysis has been devoted to analyzing the quality of various different performance measures [10, 2, 17].

We consider the problem of analyzing the quality of various measures of retrieval performance and propose a model based on the maximum entropy method for evaluating the quality of a performance measure. While measures such as average precision at relevant documents, R-precision, and 11pt average precision are known to be good measures of overall performance, other measures such as precisions at specific cutoffs are not. Our goal in this work is to develop a model within which one can numerically assess the overall quality of a given measure based on the reduction in uncertainty of a system’s performance one gains by learning the value of the measure. As such, our evaluation model is primarily concerned with assessing the relative merits of system-oriented measures, but it can be applied to other classes of measures as well.

We begin with the premise that the quality of a list of documents retrieved in response to a given query is strictly a function of the sequence of relevant and non-relevant documents retrieved within that list (as well as R, the total number of relevant documents for the given query). Most standard measures of retrieval performance satisfy this premise. Our thesis is then that given the assessed value of a “good” overall measure of performance, one’s uncertainty about the sequence of relevant and non-relevant documents in an unknown list should be greatly reduced. Suppose, for example, one were told that a list of 1,000 documents retrieved in response to a query with 200 total relevant documents contained 100 relevant documents. What could one reasonably infer about the sequence of relevant and non-relevant documents in the unknown list? From this information alone, one could only reasonably conclude that the likelihood of seeing a relevant document at any rank level is uniformly 1/10. Now suppose that one were additionally told that the average precision of the list was 0.4 (the maximum possi-
2. THE MAXIMUM ENTROPY METHOD

The concept of entropy as a measure of information was first introduced by Shannon [20], and the Principle of Maximum Entropy was introduced by Jaynes [7, 8, 9]. Since its introduction, the Maximum Entropy Method has been applied in many areas of science and technology [21] including natural language processing [1], ambiguity resolution [18], text classification [14], machine learning [15, 16], and information retrieval [6, 11], to name but a few examples. In what follows, we introduce the maximum entropy method through a classic example, and we then describe how the maximum entropy method can be used to evaluate measures of retrieval performance.

Suppose you are given an unknown and possibly biased six-sided die and were asked the probability of obtaining any particular die face in a given roll. What would your answer be? This problem is under-constrained and the most seemingly “reasonable” answer is a uniform distribution over all faces. Suppose now you are also given the information that the average die roll is 3.5. The most seemingly “reasonable” answer is still a uniform distribution. What if you are told that the average die roll is 4.5? There are many distributions over the faces such that the average die roll is 4.5; how can you find the most seemingly “reasonable” distribution? Finally, what would your answer be if you were told that the average die roll is 5.5? Clearly, the belief in getting a 6 increases as the expected value of the die rolls increases. But there are many distributions satisfying this constraint; which distribution would you choose?

The “Maximum Entropy Method” (MEM) dictates the most “reasonable” distribution satisfying the given constraints. The “Principle of Maximal Ignorance” forms the intuition behind the MEM; it states that one should choose the distribution which is least predictable (most random) subject to the given constraints. Jaynes and others have derived numerous entropy concentration theorems which show that the vast majority of all empirical frequency distributions (e.g., those corresponding to sequences of die rolls) satisfying the given constraints have associated empirical probabilities and entropies very close to those probabilities satisfying the constraints whose associated entropy is maximal [7].

Thus, the MEM dictates the most random distribution satisfying the given constraints, using the entropy of the probability distribution as a measure of randomness. The entropy of a probability distribution \( \vec{p} = \{p_1, p_2, \ldots, p_n\} \) is a measure of the uncertainty (randomness) inherent in the distribution and is defined as follows

\[
H(\vec{p}) = -\sum_{i=1}^{n} p_i \log p_i.
\]

Thus, maximum entropy distributions are probability distributions making no additional assumptions apart from the given constraints.

In addition to its mathematical justification, the MEM tends to produce solutions one often sees in nature. For example, it is known that given the temperature of a gas, the actual distribution of velocities in the gas is the maximum entropy distribution under the temperature constraint.

We can apply the MEM to our die problem as follows. Let the probability distribution over the die faces be \( \vec{p} = \{p_1, \ldots, p_6\} \). Mathematically, finding the maximum entropy distribution over die faces such that the expected die roll is
d corresponds to the following optimization problem:
Maximize: \( H(\bar{p}) \)
Subject to:
1. \( \sum_{i=1}^{N} p_i = 1 \)
2. \( \sum_{i=1}^{N} i \cdot p_i = d \)

The first constraint ensures that the solution forms a distribution over the die faces, and the second constraint ensures that this distribution has the appropriate expectation. This is a constrained optimization problem which can be solved using the method of Lagrange multipliers. Figure 1 shows three different maximum entropy distributions over the die faces such that the expected die roll is 3.5, 4.5, and 5.5, respectively.

2.1 Application of the Maximum Entropy Method to Analyzing Retrieval Measures

Suppose that you were given a list of length \( N \) corresponding to the output of a retrieval system for a given query, and suppose that you were asked to predict the probability of seeing any one of the \( 2^N \) possible patterns of relevant documents in that list. In the absence of any information about the query, any performance information for the system, or any a priori modeling of the behavior of retrieval systems, the most “reasonable” answer you could give would be that all lists of length \( N \) are equally likely. Suppose now that you are also given the information that the expected number of relevant documents over all lists of length \( N \) is \( R_{\text{ext}} \). Your “reasonable” answer might then be a uniform distribution over all \( \binom{2^N}{R_{\text{ext}}} \) different possible lists with \( R_{\text{ext}} \) relevant documents. But what if apart from the constraint on the number of relevant documents retrieved, you were also given the constraint that the expected value of average precision is \( ap \)? If the average precision value is high, then of all the \( \binom{2^N}{R_{\text{ext}}} \) lists with \( R_{\text{ext}} \) relevant documents, the lists in which the relevant documents are retrieved at low numerical ranks should have higher probabilities. But how can you determine the most “reasonable” such distribution? The maximum entropy method essentially dictates the most reasonable distribution as a solution to the following constrained optimization problem.

Let \( p(r_1, \ldots, r_N) \) be a probability distribution over the relevances associated with document lists of length \( N \), let \( rel(r_1, \ldots, r_N) \) be the number of relevant documents in a list, and let \( ap(r_1, \ldots, r_N) \) be the average precision of a list. Then the maximum entropy method can be mathematically formulated as follows:

Maximize: \( H(\bar{p}) \)
Subject to:
1. \( \sum_{r_1, \ldots, r_N} p(r_1, \ldots, r_N) = 1 \)
2. \( \sum_{r_1, \ldots, r_N} ap(r_1, \ldots, r_N) \cdot p(r_1, \ldots, r_N) = ap \)
3. \( \sum_{r_1, \ldots, r_N} rel(r_1, \ldots, r_N) \cdot p(r_1, \ldots, r_N) = R_{\text{ext}} \)

Note that the solution to this optimization problem is a distribution over possible lists, where this distribution effectively gives one’s a posteriori belief in any list given the measured constraint.

The previous problem can be formulated in a slightly different manner yielding another interpretation of the problem and a mathematical solution. Suppose that you were given a list of length \( N \) corresponding to output of a retrieval system for a given a query, and suppose that you were asked to predict the probability of seeing a relevant document at some rank. Since there are no constraints, all possible lists of length \( N \) are equally likely, and hence the probability of seeing a relevant document at any rank is 1/2. Suppose now that you are also given the information that the expected number of relevant documents over all lists of length \( N \) is \( R_{\text{ext}} \). The most natural answer would be a \( R_{\text{ext}}/N \) uniform probability for each rank. Finally, suppose that you are given the additional constraint that the expected average precision is \( ap \). Under the assumption that our distribution over lists is a product distribution (this is effectively a fairly standard independence assumption), we may solve this problem as follows. Let \( p(r_1, \ldots, r_N) = p(r_1) \cdot p(r_2) \cdots p(r_N) \) where \( p(r_i) \) is the probability that the document at rank \( i \) is relevant. We can then solve the problem of calculating the probability of seeing a relevant document at any rank using the MEM. For notational convenience, we will refer to this product distribution as the probability-at-rank distribution and the probability of seeing a relevant document at rank \( i \), \( p(r_i), \) as \( p_i \).

Standard results from information theory [4] dictate that if \( p(r_1, \ldots, r_N) \) is a product distribution, then

\[
H(p(r_1, \ldots, r_N)) = \sum_{i=1}^{N} H(p_i)
\]

where \( H(p_i) \) is the binary entropy

\[
H(p_i) = -p_i \log p_i - (1 - p_i) \log(1 - p_i).
\]

Furthermore, it can be shown that given a product distribution \( p(r_1, \ldots, r_N) \) over the relevances associated with docu-
Maximize: \( \sum_{i=1}^{N} H(p_i) \)

Subject to:
1. \( \frac{1}{N} \sum_{i=1}^{N} \left( \frac{p_i}{i} \left( 1 + \sum_{j=1}^{i-1} p_j \right) \right) = \text{ap} \)
2. \( \sum_{i=1}^{N} p_i = R_{\text{ret}} \)

Figure 2: Maximum entropy setup for average precision.

Maximize: \( \sum_{i=1}^{N} H(p_i) \)

Subject to:
1. \( \frac{1}{R} \sum_{i=1}^{R} p_i = \text{rp} \)
2. \( \sum_{i=1}^{N} p_i = R_{\text{ret}} \)

Figure 3: Maximum entropy setup for R-precision.

Maximize: \( \sum_{i=1}^{N} H(p_i) \)

Subject to:
1. \( \frac{1}{k} \sum_{i=1}^{k} p_i = PC(k) \)
2. \( \sum_{i=1}^{N} p_i = R_{\text{ret}} \)

Figure 4: Maximum entropy setup for precision-at-cutoff.

\[
\text{maximize } \sum_{i=1}^{N} H(p_i) \quad \text{subject to:} \quad \begin{align*}
1. & \quad \frac{1}{N} \sum_{i=1}^{N} \left( \frac{p_i}{i} \left( 1 + \sum_{j=1}^{i-1} p_j \right) \right) = \text{ap} \\
2. & \quad \sum_{i=1}^{N} p_i = R_{\text{ret}} \\
\end{align*}
\]

Figure 5: Probability-at-rank distributions.

Using the maximum entropy precision-recall curve of a measure, we can predict the maximum entropy precision-recall curve of the list for each measure. We can also predict the probability-at-rank distribution for each measure. We can then generate the maximum entropy precision-recall curve of the list for each measure. Using the maximum entropy precision-recall curve of a measure, we can infer the maximum entropy precision-recall curve of the list for each measure. Using the maximum entropy precision-recall curve of a measure, we can predict the maximum entropy precision-recall curve of the list for each measure.
constraint cannot be determined analytically; therefore, we used numerical optimization\(^1\) to find the maximum entropy distribution corresponding to average precision.

We shall refer to the execution of a retrieval system on a particular query as a run. Figure 6 shows examples of maximum entropy precision-recall curves corresponding to average precision, R-precision, and precision-at-cutoff 10 for three different runs, together with the actual precision-recall curves. We focused on these three measures since they are perhaps the most commonly cited measures in IR. We also provide results for precision-at-cutoff 100 in later plots and detailed results for all measures in a later table. As can be seen in Figure 6, using average precision as a constraint, one can generate the actual precision-recall curve of a run with relatively high accuracy.

In order to quantify how good an evaluation measure is in generating the precision-recall curve of an actual list, we consider two different error measures: the root mean squared error (RMS) and the mean absolute error (MAE). Let \(\{\pi_1, \pi_2, \ldots, \pi_{R_\text{ret}}\}\) be the precisions at the recall levels \(\{1/R, 2/R, \ldots, R_\text{ret}/R\}\) where \(R_\text{ret}\) is the number of relevant documents retrieved by a system and \(R\) is the number of documents relevant to the query, and let \(\{m_1, m_2, \ldots, m_{R_\text{ret}}\}\) be the estimated precisions at the corresponding recall levels for a maximum entropy distribution corresponding to a measure. Then the MAE and RMS errors are calculated as follows.

\[
\text{RMS} = \sqrt{\frac{1}{R_\text{ret}} \sum_{i=1}^{R_\text{ret}} (\pi_i - m_i)^2}
\]

\[
\text{MAE} = \frac{1}{R_\text{ret}} \sum_{i=1}^{R_\text{ret}} |\pi_i - m_i|
\]

The points after recall \(R_\text{ret}/R\) on the precision-recall curve are not considered in the evaluation of the MAE and RMS errors since, by TREC convention, the precisions at these recall levels are assumed to be 0.

In order to evaluate how good a measure is at inferring actual precision-recall curves, we calculated the MAE and RMS errors of the maximum entropy precision-recall curves corresponding to the measures in question, averaged over all runs for each TREC. Figure 7 shows how the MAE and RMS errors for average precision, R-precision, precision-at-cutoff 10, and precision-at-cutoff 100 compare with each other for each TREC. The MAE and RMS errors follow the same pattern over all TREC’s. Both errors are consistently and significantly lower for average precision than for the other measures in question, while the errors for R-precision are consistently lower than for precision-at-cutoffs 10 and 100.

Table 1 shows the actual values of the RMS errors for all measures over all TREC’s. In our experiments, MAE and RMS errors follow a very similar pattern, and we therefore omit MAE results due to space considerations. From this table, it can be seen that average precision has consistently lower RMS errors when compared to the other measures. The penultimate column of the table shows the average RMS errors per measure averaged over all TREC’s. On average, R-precision has the second lowest RMS error after average precision, and precision-at-cutoff 30 is the third best measure in terms of RMS error. The last column of the table shows the percent increase in the average RMS error of a measure when compared to the RMS error of average precision. As can be seen, the average RMS errors for the other measures are substantially greater than the average RMS error for average precision.

We now consider a second method for evaluating how informative a measure is. A highly informative measure should properly reduce one’s uncertainty about the distribution of relevant and non-relevant documents in a list; thus, in our maximum entropy formulation, the probability-at-rank distribution should closely correspond to the pattern of relevant and non-relevant documents present in the list. One should then be able to accurately predict the values of other measures from this probability-at-rank distribution.

Given a probability-at-rank distribution \(p_1, p_2, \ldots, p_N\), we can predict average precision, R-precision and precision-at-cutoff \(k\) values as follows:

\[
\text{ap} = \frac{1}{R} \sum_{i=1}^{N} p_i \frac{1}{1 + \sum_{j=1}^{i-1} p_j}
\]

\[
\text{rp} = \frac{1}{R} \sum_{i=1}^{R} p_i
\]

\[
\text{PC}(k) = \frac{1}{k} \sum_{i=1}^{k} p_i
\]

The plots in the top row of Figures 8 and 9 show how average precision is actually correlated with R-precision, precision-at-cutoff 10, and precision-at-cutoff 100 for TREC’s 6 and 8, respectively. Each point in the plot corresponds to a system and the values of the measures are averaged over all queries. Using these plots as a baseline for comparison, the plots in the bottom row of the figures show the correlation between the actual measures and the measures predicted using the average precision maximum entropy probability-at-rank distribution. Considering precision-at-cutoff 10 values using the average precision maximum entropy distributions in TREC 6. Without applying the maximum entropy method, Figure 8 shows that the two measures are correlated with a Kendall’s \(\tau\) value of 0.671. However, the precision-at-cutoff 10 values inferred from the average precision maximum entropy distribution have a Kendall’s \(\tau\) value of 0.871 when compared to actual precisions-at-cutoff 10. Hence, the predicted precision-at-cutoff 10 and actual precision-at-cutoff 10 values are much more correlated than the actual average precision and actual precision-at-cutoff 10 values. Using a similar approach for predicting R-precision and precision-at-cutoff 100, it can be seen in Figures 8 and 9 that the measured values predicted by using average precision maximum entropy distributions are highly correlated with actual measured values.

We conducted similar experiments using the maximum entropy distributions corresponding to other measures, but since these measures are less informative, we obtained much smaller increases (and sometimes even decreases) in inferred correlations. (These results are omitted due to space considerations.) Table 2 summarizes the correlation improvements possible using the maximum entropy distribution corresponding to average precision. The row labeled \(\tau_{\text{act}}\) gives the actual Kendall’s \(\tau\) correlation between average precision and the measure in the corresponding column. The row labeled \(\tau_{\text{inf}}\) gives the Kendall’s \(\tau\) correlation between the

\(^1\)We used the TOMLAB Optimization Environment for Matlab.
TREC8 System fub99a Query 435 AP = 0.1433
actual prec−recall
ap maxent prec−recall
rp maxent prec−recall
pc−10 maxent prec−recall

TREC8 System MITSLStd Query 404 AP = 0.2305
actual prec−recall
ap maxent prec−recall
rp maxent prec−recall
pc−10 maxent prec−recall

TREC8 System pir9At0 Query 446 AP = 0.4754
actual prec−recall
ap maxent prec−recall
rp maxent prec−recall
pc−10 maxent prec−recall

Figure 6: Inferred precision-recall curves and actual precision-recall curve for three runs in TREC8.

Figure 7: MAE and RMS errors for inferred precision-recall curves over all TREC.

<table>
<thead>
<tr>
<th>TREC3</th>
<th>TREC5</th>
<th>TREC6</th>
<th>TREC7</th>
<th>TREC8</th>
<th>TREC9</th>
<th>AVERAGE</th>
<th>%INC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.1185</td>
<td>0.1220</td>
<td>0.1191</td>
<td>0.1299</td>
<td>0.1390</td>
<td>0.1505</td>
<td>0.1298</td>
</tr>
<tr>
<td>RP</td>
<td>0.1767</td>
<td>0.1711</td>
<td>0.1877</td>
<td>0.2016</td>
<td>0.1878</td>
<td>0.1630</td>
<td>0.1813</td>
</tr>
<tr>
<td>PC-5</td>
<td>0.2724</td>
<td>0.2242</td>
<td>0.2451</td>
<td>0.2639</td>
<td>0.2651</td>
<td>0.2029</td>
<td>0.2456</td>
</tr>
<tr>
<td>PC-10</td>
<td>0.2474</td>
<td>0.2029</td>
<td>0.2183</td>
<td>0.2321</td>
<td>0.2318</td>
<td>0.1851</td>
<td>0.2196</td>
</tr>
<tr>
<td>PC-15</td>
<td>0.2320</td>
<td>0.1806</td>
<td>0.2063</td>
<td>0.2132</td>
<td>0.2137</td>
<td>0.1747</td>
<td>0.2048</td>
</tr>
<tr>
<td>PC-20</td>
<td>0.2210</td>
<td>0.1806</td>
<td>0.2005</td>
<td>0.2020</td>
<td>0.2068</td>
<td>0.1701</td>
<td>0.1968</td>
</tr>
<tr>
<td>PC-30</td>
<td>0.2051</td>
<td>0.1711</td>
<td>0.1950</td>
<td>0.1946</td>
<td>0.2032</td>
<td>0.1694</td>
<td>0.1897</td>
</tr>
<tr>
<td>PC-100</td>
<td>0.1787</td>
<td>0.1777</td>
<td>0.2084</td>
<td>0.2239</td>
<td>0.2222</td>
<td>0.1849</td>
<td>0.1993</td>
</tr>
<tr>
<td>PC-200</td>
<td>0.1976</td>
<td>0.2055</td>
<td>0.2435</td>
<td>0.2576</td>
<td>0.2548</td>
<td>0.2057</td>
<td>0.2274</td>
</tr>
<tr>
<td>PC-500</td>
<td>0.2641</td>
<td>0.2488</td>
<td>0.2884</td>
<td>0.3042</td>
<td>0.3027</td>
<td>0.2400</td>
<td>0.2747</td>
</tr>
<tr>
<td>PC-1000</td>
<td>0.3164</td>
<td>0.2763</td>
<td>0.3134</td>
<td>0.3313</td>
<td>0.3323</td>
<td>0.2608</td>
<td>0.3051</td>
</tr>
</tbody>
</table>

Table 1: RMS error values for each TREC.

<table>
<thead>
<tr>
<th>TREC3</th>
<th>TREC5</th>
<th>TREC6</th>
<th>TREC7</th>
<th>TREC8</th>
<th>TREC9</th>
<th>AVERAGE</th>
<th>%INC</th>
</tr>
</thead>
<tbody>
<tr>
<td>τact</td>
<td>0.921</td>
<td>0.815</td>
<td>0.833</td>
<td>0.939</td>
<td>0.762</td>
<td>0.868</td>
<td>0.913</td>
</tr>
<tr>
<td>τinf</td>
<td>0.941</td>
<td>0.863</td>
<td>0.954</td>
<td>0.948</td>
<td>0.870</td>
<td>0.941</td>
<td>0.927</td>
</tr>
<tr>
<td>%Inc</td>
<td>2.2</td>
<td>5.9</td>
<td>14.5</td>
<td>1.0</td>
<td>14.2</td>
<td>8.4</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Table 2: Kendall’s τ correlations and percent improvements for all TREC.

<table>
<thead>
<tr>
<th>TREC5</th>
<th>TREC8</th>
<th>TREC9</th>
</tr>
</thead>
<tbody>
<tr>
<td>τact</td>
<td>0.917</td>
<td>0.745</td>
</tr>
<tr>
<td>τinf</td>
<td>0.934</td>
<td>0.877</td>
</tr>
<tr>
<td>%Inc</td>
<td>1.9</td>
<td>17.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TREC6</th>
<th>TREC7</th>
<th>TREC8</th>
<th>TREC9</th>
</tr>
</thead>
<tbody>
<tr>
<td>τact</td>
<td>0.925</td>
<td>0.818</td>
<td>0.873</td>
</tr>
<tr>
<td>τinf</td>
<td>0.932</td>
<td>0.859</td>
<td>0.944</td>
</tr>
<tr>
<td>%Inc</td>
<td>0.6</td>
<td>21.7</td>
<td>5.4</td>
</tr>
</tbody>
</table>
measure inferred from the maximum entropy distribution corresponding to average precision and the measure in the corresponding column. The row labeled %Inc gives the percent increase in correlation due to maximum entropy modeling. As can be seen, maximum entropy modeling yields great improvements in the predictions of precision-at-cutoff values. The improvements in predicting R-precision are noticeably smaller, though this is largely due to the fact that average precision and R-precision are quite correlated to begin with.

4. CONCLUSIONS AND FUTURE WORK

We have described a methodology for analyzing measures of retrieval performance based on the maximum entropy method, and we have demonstrated that the maximum entropy models corresponding to “good” measures of overall performance such as average precision accurately reflect underlying retrieval performance (as measured by precision-recall curves) and can be used to accurately predict the values of other measures of performance, well beyond the levels dictated by simple correlations.

The maximum entropy method can be used to analyze other measures of retrieval performance, and we are presently conducting such studies. More interestingly, the maximum entropy method could perhaps be used to help develop and gain insight into potential new measures of retrieval performance. Finally, the predictive quality of maximum entropy models corresponding to average precision suggest that if one were to estimate some measure of performance using an incomplete judgment set, that measure should be average precision—from the maximum entropy model corresponding to that measure alone, one could accurately infer other measures of performance.

Note that the concept of a “good” measure depends on the purpose of evaluation. In this paper, we evaluate measures based on how much information they provide about the overall performance of a system (a system-oriented evaluation). However, in different contexts, different measures may be more valuable and useful, such as precision-at-cutoff 10 in web search (a user-oriented evaluation). R-precision and average precision are system-oriented measures, whereas precision-at-cutoff $k$ is typically a user-oriented measure. Another important conclusion of our work is that one can accurately infer user-oriented measures from system-oriented measures, but the opposite is not true.

Apart from evaluating the information captured by a single measure, we could use the MEM to evaluate the information contained in combinations of measures. How much does knowing the value of precision-at-cutoff 10 increase one’s knowledge of a system’s performance beyond simply knowing the system’s average precision? Which is more informative: knowing R-precision and precision-at-cutoff 30, or knowing average precision and precision-at-cutoff 100? Such questions can be answered, in principle, using the MEM. Adding the values of one or more measures simply adds one or more constraints to the maximum entropy model, and one can then assess the informativeness of the combination. Note that TREC reports many different measures. Using the MEM, one might reasonably be able to conclude which are the most informative combinations of measures.

5. REFERENCES


Figure 9: Correlation improvements, TREC8.