On the Informativeness of Cascade and Intent-Aware Effectiveness Measures

Azin Ashkan and Charles L.A. Clarke
Cheriton School of Computer Science
University of Waterloo, Canada
{aashkan, claclark}@cs.uwaterloo.ca

ABSTRACT

The Maximum Entropy Method provides one technique for validating search engine effectiveness measures. Under this method, the value of an effectiveness measure is used as a constraint to estimate the most likely distribution of relevant documents under a maximum entropy assumption. This inferred distribution may then be compared to the actual distribution to quantify the "informativeness" of the measure. The inferred distribution may also be used to estimate values for other effectiveness measures. Previous work focused on traditional effectiveness measures, such as average precision. In this paper, we extend the Maximum Entropy Method to the newer cascade and intent-aware effectiveness measures by considering the dependency of the documents ranked in a results list. These measures are intended to reflect the novelty and diversity of search results in addition to the traditional relevance. Our results indicate that intent-aware measures based on the cascade model are informative in terms of both inferring actual distribution and predicting the values of other retrieval measures.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - Retrieval Models

General Terms

Experimentation, Measurement

Keywords

effectiveness measures, evaluation, measure informativeness, novelty, diversity

1. INTRODUCTION

Retrieval effectiveness measures continue to represent a primary method for assessing the performance of search engines. These measures are computed by first executing a set of test queries against a search engine. The top results returned for each query are manually judged with respect to that query. These "editorial" judgments then form the basis for computing the various retrieval effectiveness measures, providing an overall indication of the search engine's performance.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

A large number of the retrieval effectiveness measures has been described in the research literature. The nDCG measure remains the standard measure for Web search evaluation, while other traditional measures, such as average precision, are still widely used in other areas [11]. Researchers have recently proposed a number of new measures that are intended to address some of the perceived shortcomings of nDCG and other traditional measures, particularly in a Web context.

One group of proposed measures, the *cascade measures*, are intended to better reflect user behavior [10] by considering the relationship between successive documents in a result list [4, 12, 20]. Under traditional effectiveness measures, including nDCG and average precision, the relevance of a retrieved document is considered separately from others in the results list. Even when the results list contains documents that are nearly identical, each document is given full credit in the computation of the measure. Considerations of novelty suggest a modification of this principle, in which redundancy is explicitly penalized in a results list by judging documents in the context of those already seen by the user. These considerations led to the development of effectiveness measures based on a cascade model of user behavior. Under this model, the user is assumed to scan a results list from the top downwards, eventually stopping because either their information need is satisfied or their patience is exhausted.

Another group of proposed measures, the *intent-aware measures*, are intended to better reflect the diversity of information needs underlying a query by considering different aspects and interpretations of that query [1, 7, 8, 14]. The intent-aware measures decompose a query into distinct intents, each with an associated probability. Documents may be judged with respect to each intent, allowing retrieval measures to be computed separately for each specific intent. The value of the overall intent-aware measure is computed as a linear combination of the intent-specific measures, weighted according to the associated probabilities.

As we will show later in the paper, the cascade and intent-aware measures are closely related. Many of the intent-aware measures share core features in common with the cascade measures, and all cascade measures are natural candidates for adaptation to an intent-aware framework. While evidence suggests that the cascade and intent-aware measures generally operate as intended, they are still poorly understood and are not fully validated.

Retrieval effectiveness measures may be validated in terms of properties such as their informativeness, discriminative power, robustness, and consistency with user behavior and

preferences [3, 8, 16, 18, 21]. The Maximum Entropy Method (MEM) proposed by Aslam et al. [2] analyzes the informativeness of a retrieval effectiveness measure by viewing the value of the measure as a constraint on the distribution of documents. Using MEM, Aslam et al. evaluate the informativeness of several traditional measures, including average precision. Given an effectiveness measure, called the *target* measure, the most likely distribution is inferred by MEM. By comparing the inferred distribution to the actual distribution, Aslam et al. demonstrate that the overall quality of a target measure may be quantified. The inferred distribution may also be used to estimate the value of other retrieval measures, permitting comparisons between them.

Under the traditional measures studied previously by MEM, the relevance of a retrieved document is considered separately from others in a results list. This leads to the expected value of these measures to be represented in a closed form, providing polynomial constraints accepted by MEM. However, as the cascade model of relevance and also multiple intents of queries are the bases of the measures we study in this work, the complex formulations of these target measures can not be used in MEM directly. Our computation of the expected values for these cascade and intent-aware measures produces a recursive formulation, which we solve through the application of dynamic programming and adopt it to the maximum entropy setting.

Our findings indicate that, although our extended MEM works well in predicting the value of the cascade/ traditional measures, it works even better when the intent-aware equivalents form the target measures. In other words, the intent-aware measures are found to be more informative than their non-intent-aware equivalents. Moreover, among the intent-aware measures we consider in this paper, ERR_IA and NRBP measures are found to be the most informative. Overall, the intent-aware measures are found to outperform others in terms of both inferring precision-recall curves and predicting the values of other retrieval measures.

In the next section, we review related work, with a focus on existing methods for validating retrieval effectiveness measures. Section 3 provides a detailed description of the cascade and intent-aware measures we study in this paper. In Section 4, we extend MEM to the cascade and intent-aware measures, including a description of the associated optimization problems and their solution. Finally, Section 5 reports experimental results based on runs created through the TREC 2009 Web track.

2. RELATED WORK

Many approaches to validating effectiveness measures appear in the literature. Buckley and Voorhees [3] studied the *stability* of traditional effectiveness measures, empirically examining how the behavior of a measure changes when query expressions are varied. Sakai [17] studied the *robustness* of effectiveness measures, examining the consistency of measures in the presence of system bias and pool bias.

Informativeness and discriminative power are two of the most widely considered properties for the validation of effectiveness measures [15] The *informativeness* of an effectiveness measure refers to its ability to quantify the quality of a retrieval list. The *discriminative power* of an effectiveness measure refers to its ability to differentiate the performance of retrieval systems.

The Maximum Entropy Method (MEM) proposed by Aslam et al [2] assesses the informativeness of traditional retrieval measures by estimating the probability of relevance for items on a results list, allowing the value of other measures for the same results list to be accurately inferred. Among the traditional measures studied by Aslam et al., average precision was found to be the most informative. Their findings indicate that one can accurately infer the distribution of relevant and non-relevant documents in a results list given only the value of average precision and the total number of relevant documents returned for the query.

The discriminative power of retrieval effectiveness measures is often assessed using significance tests. Sanderson et al. [18] make comparisons between the reliability of various retrieval measures using a number of significance tests. Sakai [16] proposes a simple method for assessing the discriminative power of effectiveness measures. The method computes a significance test between every pair of experimental runs and reports the percentage of pairs that are significant at some fixed significance level.

Most of the above work concerns only traditional effectiveness measures, such as average precision. The newer cascade and intent-aware measures are not considered. In this paper we extend the work of Aslam et al.[2] to encompass these measures. Discriminative power and other properties of these measures are examined elsewhere [6].

3. EFFECTIVENESS MEASURES

Under traditional measures, such as average precision (AP) and nDCG, the relevance of a retrieved document is treated independently of others in the results list. The cascade measures [4, 12, 20] which are based on the *cascade* model of user behavior [10], attempt to remedy this deficiency by explicitly penalizing redundancy in the results list. Moreover, traditional measures often consider relevance in terms of a single broad interpretation of a query. The intent-aware measures [1, 4, 7, 8] attempt to remedy this deficiency by explicitly considering different interpretations of an ambiguous query and/or different aspects of an underspecified query [8]. When placed into an intent-aware framework, the cascade measures provide a natural vehicle for measuring both novelty and diversity [6, 7].

In the description that follows, we treat the cascade and intent-aware measures in a unified fashion. We assume a given query has M intents. Each intent has an associated probability ρ_j, $1 \leq j \leq M$, indicating the probability that a user entering the query is seeking information related to intent j. Non-intent-aware versions of the measures correspond to the case that $M = 1$ and $\rho_1 = 1$.

3.1 Cascade Measures

Let $d_1 d_2 ...$ be the ranked list of documents returned in response to some query. Chapelle et al. [4] imagine users read the results list in order, stopping when they find the information they seek. Let q_j^i be the probability that a user who is interested in intent j will be satisfied with document d_i. The probability that a user interested in intent j will stop at document i can be expressed as a *gain value*, Q_j^i, as follows:

$$Q_j^i = q_j^i \prod_{k=1}^{i-1} (1 - q_j^k) \quad (1)$$

Values for q_j^i in this paper are estimated following the approach proposed by Clarke et al. [7], which is based on binary

relevance assessments available in our test collection [5]. Under this approach, q_j^i is defined as αg_j^i where $g \in \{0, 1\}$ is the binary relevance and $0 < \alpha \leq 1$ is a constant. With respect to this approach, $c_j^i = \sum_{k=1}^{i-1} g_j^k$ is defined as the number of documents ranked before position i that are judged relevant to intent j. The gain value Q_j^i is then calculated as:

$$Q_j^i = q_j^i \prod_{k=1}^{i-1}(1 - q_j^k) = \alpha g_j^i \prod_{k=1}^{i-1}(1 - \alpha g_j^k) = \alpha g_j^i (1-\alpha)^{c_j^i} \quad (2)$$

The gain value may be discounted by a factor D_i that depends on rank, accounting for the extra effort required to scan lower ranks and for the possibility that the user will abandon the query without finding anything satisfactory. This *discounted gain value* is thus $G_j^i = Q_j^i / D_i$. Consequently, the cascade effectiveness measure, denoted as C_j, with respect to an intent j and for the top N documents can be computed by summing the gain values for each document:

$$C_j = \sum_{i=1}^{N} G_j^i = \sum_{i=1}^{N} \frac{Q_j^i}{D_i} \quad (3)$$

In terms of a user model, α is viewed as a probability that a user would be satisfied with a judged relevant document; $1-\alpha$ could alternatively be viewed as representing the user's tolerance for redundancy. As α is decreased, the user becomes more willing to accept documents about previously seen intents.

Various forms of cascade measures that have been proposed in prior work employ different versions of the discount value D_i. The two cascade measures studied in this paper are the expected reciprocal rank measure (ERR) [4] and the rank-biased precision measure (RBP) [12]. ERR_j employs a linear discount of $D_i = i$; and RBP_j employes an exponential discount of $D_i = (1/\beta)^{i-1}$, where $0 \leq \beta \leq 1$.

$$ERR_j = \sum_{i=1}^{N} \frac{\alpha g_j^i (1-\alpha)^{c_j^i}}{i}, \quad (4)$$

$$RBP_j = \sum_{i=1}^{N} \alpha g_j^i (1-\alpha)^{c_j^i} \beta^{i-1}$$

From the perspective of a user model, the ratio D_i/D_{i+1} represents the probability that a user examining the document at rank i will continue on to examine the document at rank $i+1$. For the exponential discount, this probability is constant at all ranks $D_i/D_{i+1} = \beta$. For the linear discount, the probability increases at deeper ranks.

3.2 Intent-Aware Measures

Given a query with M intents, Clarke et al. [7, 8], Agrawal et al. [1], and Chapelle et al. [4] all model diversity by assigning a probability ρ_j, $1 \leq j \leq M$, to each intent, indicating the probability that a user entering the query is seeking information related to intent j.

Agrawal et al. [1] propose *intent-aware* versions for a family of measures, each based on a traditional measure such as nDCG or AP. The traditional measure is applied to each intent independently and the results are combined to give the expected value of the measure across all users. With respect to combining novelty and diversity together, Clarke et al. [8] and Chapelle et al. [4] suggest measuring novelty independently for each intent and then combining individual intent scores into a single overall score according to the diversity underlying the query. We use this definition to produce the *intent-aware* version of either traditional or cascade measures, consistent with the definition of Agrawal et al. [1]. The overall intent-aware score I over a query with M intents may be expressed as follows:

$$I = \sum_{j=1}^{M} \rho_j S_j \quad (5)$$

where S_j represents the value of a cascade measure (i.e. C_j from Equation 3) or the value of a traditional measure (i.e. T_j such as nDCG or AP) computed over the intent j. With respect to this definition, two intent-aware measures studied in this paper, ERR_IA and NRBP, are defined as the intent-aware versions of respectively ERR and RBP by combining Equations 4 and 5 as follows:

$$ERR\_IA = \sum_{j=1}^{M} \rho_j \times ERR_j = \sum_{j=1}^{M} \rho_j \sum_{i=1}^{N} \frac{\alpha g_j^i (1-\alpha)^{c_j^i}}{i} \quad (6)$$

$$NRBP = \sum_{j=1}^{M} \rho_j \times RBP_j = \sum_{j=1}^{M} \rho_j \sum_{i=1}^{N} \alpha g_j^i (1-\alpha)^{c_j^i} \beta^{i-1}$$

The third intent-aware measure, studied in this paper, is α-DCG according to the definition proposed by Clarke et al. in [7]. In this definition, the discount function of traditional nDCG is applied, while instead of using the graded values of traditional nDCG the cascade model is used to compute the gain values as detailed above. They justify α-DCG_j as the weighted linear combination of novelty scores computed over each individual intent j, which can be incorporated to the intent-aware form presented in Equation 5 in order to compute α-DCG across all intents:

$$\alpha\text{–}DCG = \sum_{j=1}^{M} \rho_j \times \alpha\text{–}DCG_j = \sum_{j=1}^{M} \rho_j \sum_{i=1}^{N} \frac{\alpha g_j^i (1-\alpha)^{c_j^i}}{\log_2 i} \quad (7)$$

Overall, the interest of this paper is to study the informativeness of these newer measures. We compare these measures to traditional average precision (AP), as well as an intent-aware version of average precision (AP_IA). In the next section, we present the theory behind our work, followed by an experimental evaluation based on a test collection created by the TREC 2009 Web Track [5].

4. INFORMATIVENESS OF THE RETRIEVAL MEASURES

The maximum entropy method (MEM), as described by Aslam et al. [2], attempts to reconstruct a distribution of relevance values for the documents in a results list, in part by applying the value of an effectiveness measure as a constraint on the distribution. This approach uses entropy [9] as a measure of uncertainty. The approach attempts to compute the distribution that maximizes entropy, subject to a set of constraints, which Aslam et al. view as a reasonable estimate of the actual distribution. Using this approach, Aslam et al. evaluate various traditional effectiveness measures, including average precision. Noting how closely a measure can

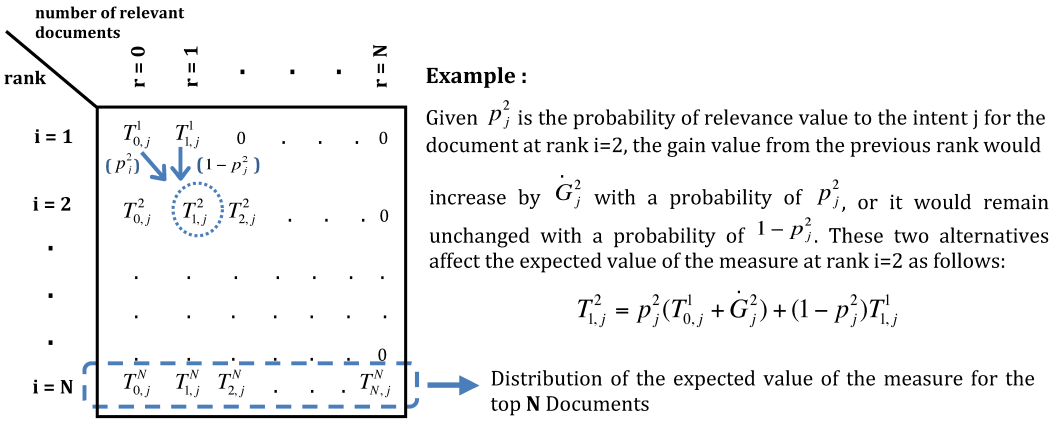


Figure 1: An example of the expected value distribution table for a list of N documents and with respect to the intent j.

predict the relevance of documents in a retrieval list and estimate other measures for that list, they attempt to quantify the informativeness of a measure according to its ability to predict the actual relevance distribution.

4.1 Adapting the Maximum Entropy Method to Cascade and Intent-Aware Measures

The maximum entropy framework is extended in this work in order to analyze the informativeness of the *cascade* and the *intent-aware* measures. Consider a list of documents corresponding to the output of a search engine for a given query. The maximum entropy method looks for the answer to this question: Can we predict the probability of seeing a document at some rank that is relevant to a given query? The first extension is applied here with respect to the cascade and intent-aware measures by changing the question to predict the probability of seeing a document at some rank that is relevant to *specific intents* of a query.

Given a list of N documents retrieved for a query with M intents, there are 2^{MN} possible patterns for the relevant documents in the retrieved list. Maximizing entropy given a set of constraints dictates the most random distribution (i.e. the most reasonable distribution) over these possible lists. Under the independence assumption, the probability distribution over the relevance values associated to a document list of depth N for a query with M intents can be written as follows:

$$P_{1,...,M}(r_1,...,r_N) = p_1(r_1)...p_M(r_1)...p_1(r_N)...p_M(r_N)$$

where $p_j(r_i)$ is the probability that the document at rank i is relevant to the j^{th} intent of the query. For notational convenience, $p_j(r_i)$ is referred to as p_j^i in the paper. Since $P_{1,...,M}(r_1,...,r_N))$ is a product distribution, then:

$$H(P_{1,...,M}(r_1,...,r_N)) = \sum_{i=1}^{N}\sum_{j=1}^{M} H(p_j^i)$$

where $H(p_j^i)$ is the binary entropy:

$$H(p_j^i) = -p_j^i \log(p_j^i) - (1-p_j^i)\log(1-p_j^i)$$

It is noted that the above formulations hold for the general case with multiple intents (i.e. for the intent-aware measures). In order to apply it to the non-intent-aware versions of the cascade measures, we set $M = 1$ and $j = 1$. Hence, in the remainder of the paper, j is omitted for non-intent-aware cascade measures.

In order to ensure that the distribution has the appropriate expectation, the expected number of relevant documents is considered as a constraint to MEM. In this paper, we assume that we are given the expected number of relevant documents (R_{ret}^j) for each intent j. In addition, the expected value of the target measure is provided to MEM as a constraint.

In order to formulate the setting, we need to formulate the expected value of each measure in terms of the p_j^i values. For the traditional measures, the expected value of a measure can be simply calculated based on the probability of relevance values with respect to a single query [2]. Once the expected value of a measure is calculated, it is passed to the optimization problem as a constraint.

Calculating the expected value of the cascade measures and the intent-aware measures such that they can be passed as polynomial constraints to an optimizer is not as straightforward as it is for the traditional measures. The reason is due to the application of the cascade model in the measures' formula for modeling novelty through considering the dependency of documents in the results list. The measures influenced by the cascade model, include ERR, RBP, ERR_IA, NRBP, and α-DCG. Other intent-aware measures, such as AP_IA, do not reflect novelty, and they only consider diversity. Therefore, they can be simply calculated by averaging over the performance of individual intents. The performance of each intent can be calculated using the formulations provided in Aslam et al. [2]. In this paper, the cascade and intent-aware measures are adjusted to MEM by calculating their expected value recursively using a dynamic programming algorithm as described below.

4.2 Expected Value of the Cascade and the Intent-Aware Measures

The calculation operates by building a *table* for each intent j (or for a single intent in case of a non-intent-aware measure). An example of this table is shown in Figure 1. The i^{th} row of the table corresponds to the expected value of the measure for the top i documents with respect to intent

j. The r^{th} value in this row, $T^i_{r,j}$, represents the expected value of the measure at rank i while r documents (among the top i documents) are relevant to the intent. The table is filled by increasing rows, top to bottom in order to take into consideration the inter-dependency of document relevance through a document gain value that depends on the relevance of documents appearing higher in the results list. At each point, the value for a cell of the table is calculated based on the value of the cells in the previous row and with respect to the probability of relevance value (i.e. p^i_j) of the document at the current rank (i.e. i), as follows:

$$T^i_{r,j} = \begin{cases} p^i_j \ (T^{i-1}_{r-1,j} + \dot{G}^i_j) + (1 - p^i_j) \ T^{i-1}_{r,j} & \text{if } i \geq r \\ 0 & \text{if } i < r \end{cases} \quad (8)$$

where $\dot{G}^i_j$ is derived from the discounted gain value G^i_j as explained in Section 3. The rationale behind this recursive formula comes from the earlier idea of measuring novelty. To achieve the aim of measuring novelty for each intent of a query, redundancy is penalized through the application of the cascade model in which the dependency of documents in a results list is considered. Given that r out of the top i documents are relevant, there are two possibilities at rank i: 1) the document at rank i is relevant (reflected by the first part of the recursive formula in Equation 8), or 2) it is not relevant (reflected by the second part of the formula).

The first part of the recursive formula considers the chance of seeing a relevant document. The expected value of the measure under this condition is the probability of seeing a relevant document (i.e. p^i_j) multiplied by the value of the measure at the previous rank with one less relevant document (i.e, $T^{i-1}_{r-1,j}$) plus the gain at this point that a relevant document is seen. From Equations 1 and 2, $Q^i_j = p^i_j(1-\alpha)^{c^i_j}$ with $c^i_j = r - 1$, since the number of relevant documents is $r - 1$ up to this point in the recursive calculation. The gain value is then discounted by D_i as described in Section 3. Therefore, $\dot{G}^i_j = (1 - \alpha)^{(r-1)}/D_i$ in Equation 8.

The second part of the formula considers the chance of seeing a non-relevant document $(1 - p^i_j)$. In this case, no gain is obtained and the effective value of the measure will remain the same as the value at the previous rank with r relevant documents (i.e. $T^{i-1}_{r,j}$).

For a given intent j, the number of relevant documents among the top N documents is given as R^j_{ret}, the expected value of the measure is estimated as $T^N_{R^j_{ret},j}$. It is noted that if the measure is a cascade measure, there will be one table for the query ($j = 1$), and therefore the expected value of the measure is $E_C = T^N_{R_{ret}}$ where R_{ret} is the given number of documents relevant to the query. For an intent-aware measure with M intents, there will be one table per intent. The expected value of the measure for the top N documents is then calculated as a linear combination of the expected values of the individual intents:

$$E_I = \frac{1}{M} \sum_{j=1}^{M} T^N_{R^j_{ret},j} \quad (9)$$

Following the conventions of the TREC 2009 Web Track, the probability assigned to each intent is assumed to be equal, and therefore $\rho_i = \frac{1}{M}$ in the equation above.

The formulations presented above all form constraints to the optimization problem. Constrained nonlinear optimization [13] is used to solve these optimization problems. The general form of the optimization problem is shown in Figure 2, where E_C and E_I are the expected values for the target cascade measure or the intent-aware measure respectively, which is passed to the optimization problem as a constraint along with the expected number of relevant documents for each intent (or for a single intent in case of the non-intent-aware measures).

Maximize $\sum_{i=1}^{N} \sum_{j=1}^{M} H(p^i_j)$

Subject to:

1. $\forall \ 1 \leq j \leq M, \quad \sum_{i=1}^{N} p^i_j = R^j_{ret}$
2. E_C or E_I and the corresponding formula

Figure 2: Maximum entropy setup for a target cascade or intent-aware measure.

Recursive calculation of E_C and E_I leads to a polynomial formulation for each measure. In order to obtain the polynomial expected value, the closed form of the recursive formulation is calculated for each measure. This is performed once for each measure by traversing the recursive formula and outputting the parameters in string form at each step until the base case is reached. The resulting string is stored as the polynomial form of the expected value for the measure. The parameters in the string are then substituted with their corresponding values at each iteration of the optimization program. The recursive form of the measure helps us to find the derivative for the second constraint recursively, passing it to the optimization setting in the same way as the original constraint.

4.3 Instances of MEM for the Target Measures

Overall, the problem of evaluating the cascade measures and the intent-aware measures using the MEM may be expressed as follows:

- **MEM for the Cascade Measures**: Given a list of N documents with respect to a single query, one is told that the expected number of relevant documents for the query is R_{ret}. The expected value of the target cascade measure is also given as E_C, and one is asked the probability of relevance of each document to each intent under the independence assumption. By applying MEM, one can determine the probability distribution (called the *probability-at-rank-query distribution*) that maximizes entropy.
- **MEM for the Intent-Aware Measure**: Given a list of N documents with respect to M intents of a query, one is told that the expected number of relevant documents for intent j is R^j_{ret}. The expected value of an intent-aware measure is also given as E_I, and one is asked the probability of relevance of each document with respect to each intent under the independence assumption. By applying MEM, one can determine the probability distribution (called the *probability-at-rank-intent distribution*) that maximizes entropy.

Maximize $\sum_{i=1}^{N} H(P^i)$

Subject to:

1. $\sum_{i=1}^{N} p^i = R_{ret}$
2. $T_r^i = p^i[T_{r-1}^{i-1} + \frac{(1-\alpha)^{r-1}}{i}] + (1-p^i)T_r^{i-1} = ERR_{ret}$

   $T_r^i = 0 \quad \forall i < r$

(a) ERR

Maximize $\sum_{i=1}^{N} H(P^i)$

Subject to:

1. $\sum_{i=1}^{N} p^i = R_{ret}$
2. $T_r^i = p^i[T_{r-1}^{i-1} + (1-\alpha)^{r-1}\beta^{i-1}] + (1-p^i)T_r^{i-1} = RBP_{ret}$

   $T_r^i = 0 \quad \forall i < r$

(b) RBP

Figure 3: Maximum entropy setup for the target cascade measures.

The cascade measures that are targeted in this work are RBP and ERR. The expected value of each cascade measure can be formulated according to the recursive formula above (Equation 8). The gain and discount values used in the recursive formula are obtained according to the general definition of the measures presented in Equation 4. Casting the expected value of measures into the general form of MEM (Figure 2) will give us the problem formulations depicted in Figure 3. ERR_{ret} and RBP_{ret} listed in Figures 3a and 3b are respectively the given values of ERR and RBP for a list of N documents that are passed to the optimizer.

The intent-aware measures targeted in this paper are the ones influenced by both novelty and diversity: ERR_IA, NRBP, and α-DCG. Applying the recursive formula from Equation 8 and averaging over intents through Equation 9 for these measures and passing each to the general form of MEM (Figure 2) give us the problem formulations depicted in Figures 4a, 4c, and 4b. Again note that the gain and discount values used in the recursive formula are obtained according to the general definition of the measures presented earlier in Equations 6 and 7. ERR_IA_{ret}, $NRBP_{ret}$, and αDCG_{ret} listed in Figures 4a to 4c are respectively the given values of ERR_IA, NRBP, and α-DCG with respect to M intents and for a list of N documents that are passed to the optimizer.

5. EXPERIMENTAL RESULTS

Using the maximum entropy probability distribution (probability -at-rank-query distribution for cascade measures and probability -at-rank-intent distribution for intent-aware measures) obtained by solving the optimization problem corresponding to a given measure, we should be able to infer the quality of a retrieval list along with the value of other cascade and intent-aware measures. As noted by Aslam et al. [2] and Yilmaz et al. [19], if a target measure can accurately predict the relevance probability of a document to a query, one should be able to estimate the actual precision-recall curve and also predict the value of other measures for the list given the value of the target measure. They use this approach to evaluate the informativeness of traditional retrieval measures. In this section, we report the results of equivalent experiments on the cascade and intent-aware measures. Specifically, we evaluate the performance of the cascade and intent-aware measures in terms of i) estimating precision-recall curves and ii) predicting the values of other cascade and intent-aware measures.

Maximize $\sum_{i=1}^{N}\sum_{j=1}^{M} H(p_j^i)$

Subject to:

1. $\forall\ 1 \le j \le M, \quad \sum_{i=1}^{N} p_j^i = R_{ret}^j$
2. $T_{r,j}^i = p_j^i[T_{r-1,j}^{i-1} + \frac{(1-\alpha)^{r-1}}{i}] + (1-p_j^i)T_{r,j}^{i-1}$

   $T_r^i = 0 \quad \forall i < r$

   $\frac{1}{M}\sum_{j=1}^{M} T_{R_{ret}^j,j}^N = ERR\_IA_{ret}$

(a) ERR_IA

Maximize $\sum_{i=1}^{N}\sum_{j=1}^{M} H(p_j^i)$

Subject to:

1. $\forall\ 1 \le j \le M, \quad \sum_{i=1}^{N} p_j^i = R_{ret}^j$
2. $T_{r,j}^i = p_j^i[T_{r-1,j}^{i-1} + (1-\alpha)^{r-1}\beta^{i-1}] + (1-p_j^i)T_{r,j}^{r-1}$

   $T_r^i = 0 \quad \forall i < r$

   $\frac{1}{M}\sum_{j=1}^{M} T_{R_{ret}^j,j}^N = NRBP_{ret}$

(b) NRBP

Maximize $\sum_{i=1}^{N}\sum_{j=1}^{M} H(p_j^i)$

Subject to:

1. $\forall\ 1 \le j \le M, \quad \sum_{i=1}^{N} p_j^i = R_{ret}^j$
2. $T_{r,j}^i = p_j^i[T_{r-1,j}^{i-1} + \frac{(1-\alpha)^{r-1}}{\log_2 i}] + (1-p_j^i)T_{r,j}^{i-1}$

   $T_{r,j}^i = 0 \quad \forall i < r$

   $\frac{1}{M}\sum_{j=1}^{M} T_{R_{ret}^j,j}^N = \alpha DCG_{ret}$

(c) α-DCG

Figure 4: Maximum entropy setup for the target intent-aware measures.

Table 1: RMS and MAE error values in predicting precision-recall curve w.r.t. the cascade and the intent-aware measures.

	ERR	RBP	AP	ERR_IA	NRBP	α-DCG	AP_IA
RMS	0.2283	0.3207	0.2507	0.0592	0.0727	0.0810	0.0422
MAE	0.2005	0.2937	0.2530	0.0380	0.0471	0.0496	0.0273

The experimental data collected by diversity task of TREC 2009 Web Track is used as the basis for our experiments [5]. The corpus for this track was crawled from the general Web in early 2009 and contains roughly one billion Webpages[1]. As part of the Web Track, the TREC organizers developed 50 queries, with explicitly defined intents, which were given to the track participants for execution over their systems. After executing the queries, the participants submitted ranked lists of documents to the TREC organizers, submitting a total of 49 experimental runs.Hired assessors made binary relevance judgments with respect to each intent.

For our experimental study, all measures are computed to retrieval depth $N = 10$. Following the conventions of the Web Track, a default value of $\alpha = 0.5$ is used for all measures, and a default value of $\beta = 0.8$ is adopted for NRBP and RBP (a relatively patient user). To compute non-intent-aware measures, we ignore the distinction between intents and treat a document as relevant to a query if it is relevant to any intent.

We sorted the 49 submitted runs based on the total number of relevant documents returned for each. The top 30 runs were selected for our experiments; the remaining runs did not appear to achieve sufficient performance to permit a meaningful analysis. Similarly, we sorted the 50 queries based on the total number of documents judged as relevant for each. The top 25 queries were selected for our analysis; the remaining queries did not appear to have sufficient relevant documents to permit a meaningful analysis. As seen in Figures 3 and 4, there are five optimization problems to be solved for each query of each run (one for each measure). These five optimization problems have been solved for each of the 25 selected queries and for each run in the set of 30 selected runs, giving a total of $5 \times 25 \times 30 = 3,750$ optimization problems.

5.1 Inferring the Precision-Recall Curve

For each triple <measure, query, run>, the probability distribution values (i.e. the p_j^i values) are obtained by solving the maximum entropy optimization problem corresponding to that measure and query and with respect to the measures calculated from the run. Using these values, the precision-recall curve for each triple of <measure, query, run> is estimated and compared against the corresponding actual precision-recall curve. Further details regarding this process can be found in Aslam et al. [2]. In Figure 5, we demonstrate an example of the predicted precision-recall curves with respect to the maximum entropy distribution for the intent-aware measures against the precision-recall curve of the actual results list. For any run with a total of R_{ret} relevant documents returned for a query, the recall points after R_{ret}/R are not of the interest to us (points after 0.14 for the example shown in the figure).

It can be seen in Figure 5 that AP_IA and ERR_IA estimate the actual precision-recall curve better than the other

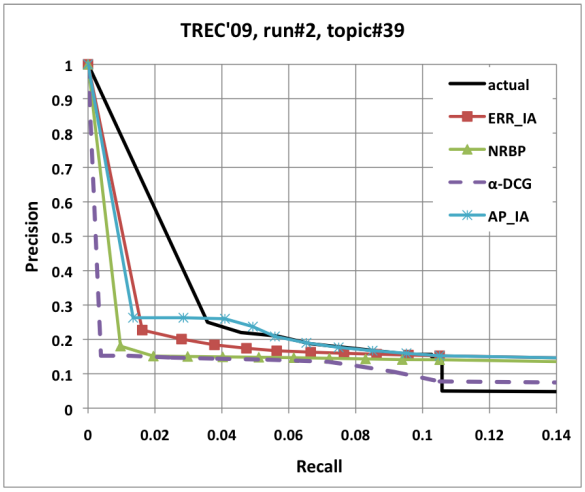


Figure 5: An example of the inferred precision-recall curve for the intent-aware measures against the actual precision-recall curve.

two measures. Of course, this observation holds only for this example. To make general observations, we need to consider the measures across the 25 queries and 30 runs. In order to evaluate how well a measure can estimate the actual precision-recall curve, we calculate the root mean square error (RMS) and the mean absolute error (MAE) between the estimated and actual precision-recall curves at the points where the recall changes and then averaged over the runs. Table 1 summarizes the errors in inferring the precision-recall curve with respect to the intent-aware and cascade measures. The results based on AP and AP_IA are also reported in the table for comparison purposes. Both errors are significantly lower for the intent-aware measures as compared to the cascade measure and the traditional measure (AP). Consistent with Figure 5, again AP_IA and ERR_AI appear to be good predictors, along with the rest of the intent-aware measures.

5.2 Inferring Other Retrieval Measures

The p_j^i values of the maximum entropy distribution based on a target measure may be used to infer the expected value of other measures. In order to evaluate the performance of such predictions, the value of each predicted measure for a run is averaged over the 25 queries. As an example, for run #1 and the target measure ERR_IA, the optimization problem shown in Figure 4a is solved 25 times, each time based on the constraints obtained for one of the 25 queries.

The results of each optimization problem are then used to estimate the value of other measures. For instance, NRBP is calculated from the result of each optimization problem and then averaged over the 25 queries. Finally, in order to evaluate how well ERR_IA can predict NRBP, for each run,

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09

Table 2: Kendall's τ correlation and the relative errors for prediction w.r.t. the intent-aware measures.

	w.r.t. ERR_IA			w.r.t. NRBP		
	AP_IA	NRBP	α-DCG	AP_IA	ERR_IA	α-DCG
Kendall's τ	0.829	0.835	0.905	0.830	0.814	0.854
RMSR	0.384	0.325	0.188	0.373	0.395	0.324
MARE	0.241	0.169	0.136	0.287	0.251	0.246
	w.r.t. α-DCG			w.r.t. AP_IA		
	AP_IA	ERR_IA	NRBP	ERR_IA	NRBP	α-DCG
Kendall's τ	0.815	0.802	0.809	0.706	0.766	0.795
RMSR	0.385	0.457	0.463	0.666	0.556	0.508
MARE	0.284	0.327	0.290	0.513	0.427	0.421

Table 3: Kendall's τ correlation and the relative errors for prediction w.r.t. the cascade measures comparing with the ones for prediction w.r.t. AP.

	w.r.t. ERR			w.r.t. RBP			w.r.t. AP		
	AP	RBP	DCG	AP	ERR	DCG	ERR	RBP	DCG
Kendall's τ	0.767	0.824	0.851	0.785	0.797	0.849	0.639	0.745	0.790
RMSR	0.553	0.364	0.209	0.773	0.976	0.214	0.651	0.594	0.304
MARE	0.371	0.332	0.157	0.550	0.805	0.160	0.472	0.478	0.246

the obtained NRBP value is compared against the average of the actual NRBP value calculated using the 25 queries for that run. This process is repeated for each run and each target measure, in order to estimate the value of other measures. Note that each cascade measure is used to predict the rest of cascade measures plus AP (as an example of a traditional measure) while each intent-aware measure is used to predict the rest of the intent-aware measures including AP_IA.

Figures 6 to 10 plot the inferred value (corresponding to the maximum entropy distribution obtained based on each target measure) against the actual value for each run averaged over the queries. We report Kendall's τ to measure the correlation of an inferred measure and its actual value. Values range from -1 to +1, with +1 indicating perfect agreement and -1 indicating the opposite. In addition to evaluating the prediction results in terms of ranking, value-based error measures (i.e. RMSR and MARE) are used to evaluate the results. Root mean square relative error (RMSR) and mean average relative error (MARE) are chosen in order to measure the *relative* error values across the results of various retrieval measures. Tables 2 and 3 summarize the correlation and error measures for the results of the intent-aware measures and the cascade measures respectively.

There is relatively high correlation among the cascade-based intent-aware measures shown in the plots. For instance, in Figure 6c, α-DCG values inferred based on the maximum entropy distribution corresponding to ERR_IA (as the target measure) appear to be highly correlated with the actual value of α-DCG along 30 runs. They are correlated with a Kendall's τ value of 0.905 according to Table 2. On the other hand, the cascade measure ERR, for which no notion of diversity is considered, results in a maximum entropy distribution with a lower correlation reported between the inferred DCG and the actual value of DCG along the 30 runs (Kendall's $\tau = 0.851$ according to Table 3).

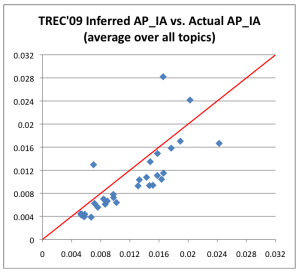
Similar results are obtained from prediction based on NRBP versus prediction based on RBP. This outcome could indicate that, although the extended MEM works well in predicting the value of the cascade/traditional measures based on a target cascade measure, it works substantially better when the intent-aware measures are used as target measures. In other words, the intent-aware measures are found to be more informative than their non-intent-aware versions (cascade metrics). Moreover, among the intent-aware measures, ERR_IA and NRBP are found to be the more informative measures in predicting the value of other measures.

Comparing this set of results with the one reported earlier in Table 1, it appears that AP_IA can compete with the intent-aware measures in terms of predicting the precision-recall curve of the actual list. This is true to a lesser extent, according to Table 2, when AP_IA is compared against the rest of the measures for inferring other retrieval measures. This could be due to the fact that average precision directly incorporates precision, and therefore it naturally should be a good predictor of the precision-recall curve. Overall, the intent-aware measures that reflect both novelty and diversity are found to be informative in terms of both inferring the precision-recall curve and predicting other retrieval measures.

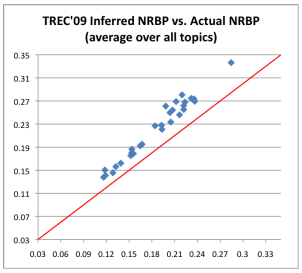
6. CONCLUDING DISCUSSION

In this paper, we extend the MEM framework to cascade and intent-aware measures. By determining the extent to which a target measure can predict an actual retrieval list and estimate the value of other measures on that list, we may quantify the informativeness of that measure. The runs created through the TREC 2009 Web track provide a vehicle for exploring and validating the cascade and intent-aware measures using the MEM framework. Our experimental comparison contains two parts: i) inferring the actual precision-recall curve based on the maximum entropy distribution obtained from a target measure, and ii) estimating the value of other measures based on the relevance probability values obtained with respect to a target measure.

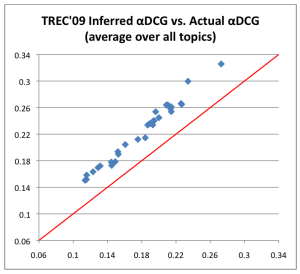
Our results indicate that the intent-aware measures that reflect both novelty (through the application of the cascade model) and diversity are informative in terms of predicting both the precision-recall and the value of other measures. They are found to be more informative than their cor-



(a) AP_IA (K's $\tau = 0.829$)

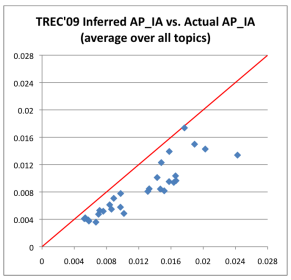


(b) NRBP (K's $\tau = 0.835$)

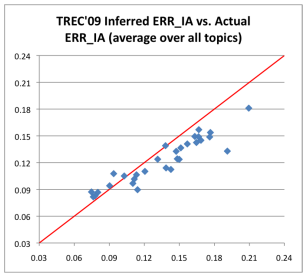


(c) α-DCG (K's $\tau = 0.905$)

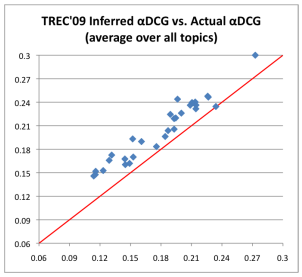
Figure 6: Prediction based on ERR_IA, TREC'09.



(a) AP_IA (K's $\tau = 0.830$)

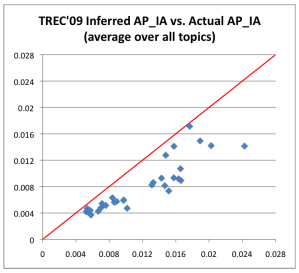


(b) ERR_IA (K's $\tau = 0.814$)

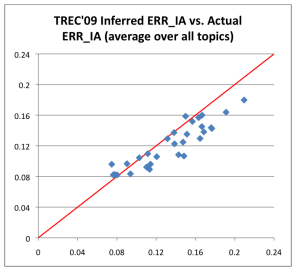


(c) α-DCG (K's $\tau = 0.854$)

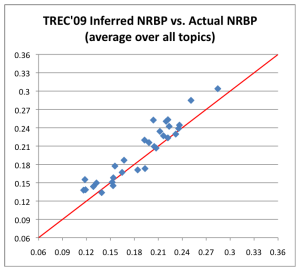
Figure 7: Prediction based on NRBP, TREC'09.



(a) AP_IA (K's $\tau = 0.815$)

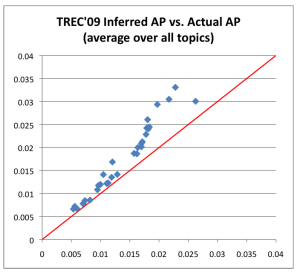


(b) ERR_IA (K's $\tau = 0.802$)

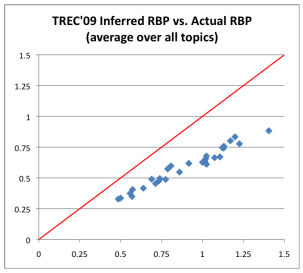


(c) NRBP (K's $\tau = 0.809$)

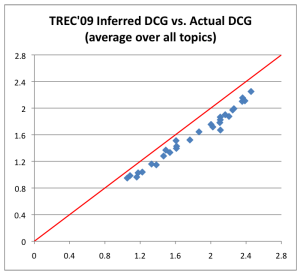
Figure 8: Prediction based on α-DCG, TREC'09.



(a) AP (K's $\tau = 0.767$)

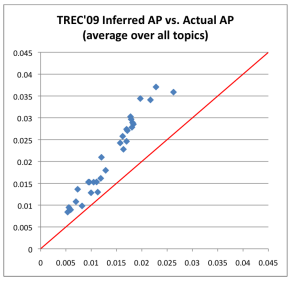


(b) RBP (K's $\tau = 0.824$)

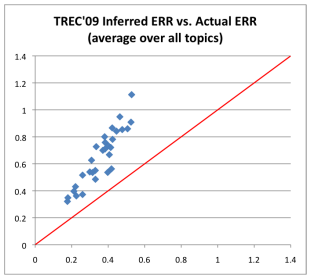


(c) DCG (K's $\tau = 0.851$)

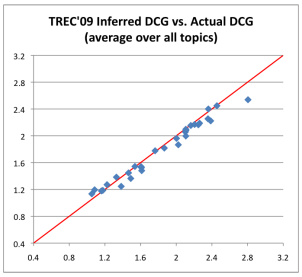
Figure 9: Prediction based on ERR, TREC'09.



(a) AP (K's $\tau = 0.785$)



(b) ERR (K's $\tau = 0.797$)



(c) DCG (K's $\tau = 0.849$)

Figure 10: Prediction based on RBP, TREC'09.

responding non-intent-aware cascade measures, which only consider novelty. As seen in the first part of our experiments, AP_IA appears to be competitive with the rest of the intent-aware measures in terms of its ability to estimate precision-recall curves, while it does not appear to be as good at predicting the values of cascade measures. We may explain this observation from the fact that AP directly incorporates precision, and therefore it naturally should be a good predictor of the precision-recall curve.

We based our default α and β parameters on those used in the TREC 2009 Web track. As future work, we plan to further explore the performance of the measures by varying the value of these parameters. Also, future TREC Web Tracks will help us to further evaluate our techniques. Finally, we hope to evaluate the cascade and intent-aware measures in terms of the user behavior seen in search logs [20, 4].

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In 2^{nd} *ACM International Conference on Web Search and Data Mining*, pages 5–14, 2009.

[2] J. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In 28^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34, 2005.

[3] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In 23^{rd} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2000.

[4] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In 18^{th} *ACM Conference on Information and Knowledge Management*, pages 621–630, 2009.

[5] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In 18^{th} *Text REtrieval Conference*, 2009.

[6] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *ACM International Conference on Web Search and Data Mining*, 2011.

[7] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkann, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In 31^{st} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–666, 2008.

[8] C. L. A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In 2^{nd} *International Conference on the Theory of Information Retrieval*, pages 188–199, 2009.

[9] T. Cover and J. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

[10] N. Craswell, O. Zoeter, M. J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *International Conference on Web Search and Web Data Mining*, pages 87–94, 2008.

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[12] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27, 2008.

[13] M. Powell. A fast algorithm for nonlinearly constrained optimization calculations. *Numerical Analysis*, pages 144–157, 1978.

[14] F. Radlinski, M. Szummer, and N. Craswell. Metrics for assessing sets of subtopics. In 33^{rd} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 853–854, 2010.

[15] S. Robertson, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. In 33^{rd} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, 2010.

[16] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In 29^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 532, 2006.

[17] T. Sakai. On the robustness of information retrieval metrics to biased relevance assessments. *Information and Media Technologies*, 4(2):547–557, 2009.

[18] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In 28^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169, 2005.

[19] E. Yilmaz and S. Robertson. On the choice of effectiveness measures for learning to rank. *Information Retrieval*, 13(3):271–290, 2010.

[20] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Incorporating user behavior information in IR evaluation. In *SIGIR 2009 Workshop on Understanding the User: Logging and Interpreting User Interactions in Information Retrieval*, 2009.

[21] Y. Zhang, L. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1):46–69, 2010.