A Mutual Information-based Framework for the Analysis of Information Retrieval Systems

Peter B. Golbus Javed A. Aslam
College of Computer and Information Science
Northeastern University
Boston, MA, USA
{pgolbus,jaa}@ccs.neu.edu

ABSTRACT

We consider the problem of information retrieval evaluation and the methods and metrics used for such evaluations. We propose a *probabilistic framework* for evaluation which we use to develop new *information-theoretic* evaluation metrics. We demonstrate that these new metrics are powerful and generalizable, enabling evaluations heretofore not possible.

We introduce four preliminary uses of our framework: (1) a measure of conditional rank correlation, *information τ*, a powerful meta-evaluation tool whose use we demonstrate on understanding novelty and diversity evaluation; (2) a new evaluation measure, *relevance information correlation*, which is correlated with traditional evaluation measures and can be used to (3) evaluate a collection of systems simultaneously, which provides a natural upper bound on metasearch performance; and (4) a measure of the similarity between rankers on judged documents, *information difference*, which allows us to determine whether systems with similar performance are in fact different.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation; Theory; Measurement

Keywords

Information Retrieval, Search Evaluation

1. INTRODUCTION

In order to improve search engines, it is necessary to accurately measure their current performance. If we cannot measure performance, how can we know whether a change was beneficial? In recent years, much of the work on information retrieval evaluation has focused on user models [7, 18] and diversity measures [1, 10, 24] which attempt to accurately reflect the experience of the user of a modern internet search engine. However, these measure are not easily generalized. In this work, we introduce a probabilistic framework for evaluation that encompasses and generalizes current evaluation methods. Our probabilistic framework allows us to view evaluation using the tools of information theory [11]. While our framework is not designed to coincide with user experience, it provides immediate access to a large number of powerful tools allowing for a deeper understanding of the performance of search engines.

Our framework for evaluation is based on the observation that relevance judgments can also be interpreted as a preference between those documents with different relevance grades. This implies that relevance judgments can be treated as a retrieval system, and that evaluation can be considered as the "rank" correlation between systems and relevance judgments. To this end, we develop a probabilistic framework for rank correlation based on the expectation of random variables, which we demonstrate can also be used to compute existing evaluation metrics. However, the true value of our framework lies in its extension to new information-theoretic evaluation tools.

After a discussion of related work (Section 2), we introduce our framework in Section 3. In Section 4, we demonstrate that our framework allows for an information theoretic understanding of Kendall's τ [17], *information τ*, which we use to define a conditional version of the rank correlation between two lists conditioned on a third. In Section 5, we define a new evaluation measure based on our framework: *relevance information correlation*. We validate our measure by showing that it is highly correlated with existing measures such as average precision (AP) and normalized discounted cumulative gain (nDCG). As a demonstration of the versatility of our framework when compared to, for example, user models, we show that our measure can be used to evaluate a collection of systems simultaneously (Section 6), creating an upper bound on the performance of metasearch algorithms. Finally, in Section 7, we introduce *information difference*, a powerful new tool for evaluating the similarity of retrieval systems beyond simply comparing their performance.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2034-4/13/07 ...$15.00.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1256172. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

2. RELATED WORK

Search systems are typically evaluated against *test collections* which consist of a corpus of documents, a set of topics, and relevance assessments—whether a subset of those documents are *relevant* with respect to each topic.[1] For example, the annual, NIST-sponsored Text REtrieval Conference (TREC) creates test collections commonly used in academic research. The performance of systems is assessed with regards to a specific task. A traditional search task is to attempt to rank all relevant documents above any non-relevant documents. For this task, systems are evaluated in terms of the average trade-off between their *precision* and *recall* with respect to multiple topics. For a given topic, Let $g_i \in \{0, 1\}$ be the relevance grade of the document at rank i, and let R be the number of relevant documents in the collection. At rank k,

$$\text{precision@}k = \frac{\sum_{i=1}^{k} g_i}{k} \tag{1}$$

$$\text{recall@}k = \frac{\sum_{i=1}^{k} g_i}{R} \tag{2}$$

The trade-off between the two is measured by *average precision*, which can be interpreted as the area under the precision-recall curve.

$$AP = \frac{\sum_{i=1}^{\infty} g_i \times \text{precision@}i}{R} \tag{3}$$

Average precision does not include information about document quality and degrees of relevance, and is an inherently recall-oriented measure. It is therefore not suitable for evaluating commercial web search engines.

With the growth of the World Wide Web, test collections began to include graded, non-binary relevance judgments, e.g. $G = \{\text{non-relevant, relevant, highly relevant}\}$ or $G = \{0, \ldots, 4\}$. To make use of these graded assessments, Järvelin and Kekäläinen developed *normalized discounted cumulative gain* (nDCG) [15]. nDCG also has the advantage that it can be evaluated at arbitrary ranks, and can therefore be used for precision-oriented tasks like web search.

Unlike average precision, which has a technical interpretation, nDCG can be best understood in terms of a model of a hypothetical user. In this model, a user will read the first k documents in a ranked list, deriving utility from each document. The amount of utility is proportional to the document's relevance grade and inversely proportional to the rank at which the document is encountered. We first define discounted cumulative gain (DCG).

$$DCG@k = \sum_{i=1}^{k} \frac{2^{g_i} - 1}{\log_2(i+1)} \tag{4}$$

Since the range of DCG will vary from topic to topic, it is necessary to normalize these scores so that an average can be computed. Normalization is performed with regard to an ideal ranked list. If $DCG'@k$ is the maximum possible DCG of any ranked list of documents in the collection then

$$nDCG@k = \frac{DCG@k}{DCG'@k} \tag{5}$$

However, one does not always know how many documents are relevant at each level, and therefore the ideal list used for normalization is only an approximation. Moffat and Zobel [18] introduced a measure, *rank-biased precision* (RBP), that addresses this issue. In RBP, the probability that a user will read the document at rank k is drawn from a geometric distribution, whose parameter, $\beta \in [0, 1)$, models the user's persistence. Given a utility function $u\colon G \to [0, 1]$, commonly defined as

$$u(g) = \frac{2^g - 1}{2^d} \tag{6}$$

where d is the maximum possible relevance grade, RBP is defined as the expected utility of a user who browses according to this model.

$$RBP = (1 - \beta) \sum_{i=1}^{\infty} u(g_i) \times \beta^{i-1} \tag{7}$$

Since RBP is guaranteed to be in the range [0,1) for any topic and β, it does not require normalization.

Craswell et al. [12] introduced the Cascade model of user behavior. In this model, a user is still assumed to browse documents in order, but the probability that a user will view a particular document is no longer assumed to be independent of the documents that were viewed previously, i.e. a user is not assumed to stop at a particular rank, or at each rank with some probability. Instead, the user is assumed to stop after finding a relevant document. This implies that if a user reaches rank k, then all of the $k - 1$ documents ranked before it were non-relevant. Craswell et al. demonstrated empirically that this model corresponds well to observed user behavior in terms of predicting the clickthrough data of a commercial search engine.

Chapelle et al. [7] developed an evaluation measure, *expected reciprocal rank* (ERR), based on the Cascade model. Let R_i denote the probability that a user will find the document at rank i to be relevant. Then in the Cascade model, the likelihood that a user will terminate his or her search at rank r is

$$R_r \prod_{i=1}^{r-1} (1 - R_i). \tag{8}$$

If we interpret the previously defined utility function (Equation 6) as the probability that a user will find a document relevant, i.e. $R_i = u(g_i)$, then we can computed the expected reciprocal rank at which a user will terminate his or her search as

$$ERR = \sum_{r=1}^{\infty} \frac{1}{r} R_r \prod_{i=1}^{r-1} (1 - R_i). \tag{9}$$

In this work, we propose an alternative, information-theoretic framework for evaluation. The first step is to reformulate these measures as the expected outcomes of random experiments. Computing evaluation measures in expectation is

[1] For historical reasons, the set of relevance assessments is often referred to as a *QREL*.

not uncommon in the literature, and we are not the first to suggest that reformulating an evaluation measure as an expectation allows for novel applications. For example, Yilmaz and Aslam [30] formulated average precision as the expectation of the following random experiment:

1. Pick a random relevant document,
2. Pick a random document ranked at or above the rank of the document selected in step 1.
3. Output 1 if the document from step 2 is relevant, otherwise output 0.

Their intention was to accurately estimate average precision while collecting fewer relevance judgments (a process also applied to nDCG [32]). However, this formulation led to new uses, such as defining an information retrieval-specific rank correlation measure, τ_{AP} [31], and a variation of average precision for graded relevance judgments, Graded Average Precision (GAP) [21].

Our work uses pairwise document preferences rather than absolute relevance judgments. The use of preferences is somewhat common in IR. For example, many learning-to-rank algorithms, such as LambdaMart [3] and RankBoost [13], use pairwise document preferences in their objective functions. Carterette et al. [4, 5] explored the collection of preference judgments for evaluation, showing that they are faster to collect and have lower levels of inter-assessor disagreement. More recently, Chandar and Carterette [6] crowdsourced the collection of *conditional* document preferences to evaluate the standard assumptions underlying diversity evaluation, for example that users always prefer novel documents. Relative document preferences can also be inferred from the clickthrough data collected in the logs of commercial search engines [16]. These preferences can be used for evaluation without undertaking the expense of collecting relevance judgments from assessors.

3. A PROBABILISTIC FRAMEWORK FOR EVALUATION

Mathematically, one can view the search system as providing a *total ordering* of the documents ranked and a *partial ordering* of the entire collection, where all ranked documents are preferred to unranked documents but the relative preference among the unranked documents is unknown. Similarly, one can view the relevance assessments as providing a partial ordering of the entire collection: in the case of binary relevance assessments, for example, all judged relevant documents are preferred to all judged non-relevant and unjudged documents, but the relative preferences among the relevant documents and among the non-relevant and unjudged documents is unknown. Thus, mathematically, one can view retrieval evaluation as comparing the partial ordering of the collection *induced by the search system* with the partial ordering of the collection *induced by the relevance assessments*.

To formalize and instantiate a framework for comparing such partial orderings, consider the simplest case where we have two total orderings of objects, i.e., where the entire "collection" of objects is fully ranked in both "orderings." While such a situation does not typically arise in search system evaluation (since not all *documents* are ranked by the retrieval system nor are they fully ranked by relevance assessments), it does often arise when comparing the *rankings of systems* induced by two (or more) evaluation metrics; here Kendall's τ is often the metric used to compare these (total order) rankings.

In what follows, we define a *probabilistic framework* within which to compare two total orderings, and we show how traditional metrics (such as Kendall's τ) are easily cast within this framework. The real power of such a framework is shown in subsequent sections: (1) the framework can be easily generalized to handle the comparison of two partial orderings, such as arise in search system evaluation, and (2) well-studied, powerful, and general information-theoretic metrics can be developed within this generalized framework.

Consider two total orderings of n objects. There are $\binom{n}{2}$ (unordered) pairs of such objects, and a pair is said to be *concordant* if the two orderings agree on the relative rankings of the objects and *discordant* if the two orderings disagree. Let c and d be the number of concordant and discordant pairs, respectively. Then Kendall's τ is defined as follows:

$$\tau = \frac{c-d}{c+d}. \tag{10}$$

If we let C and D denote the *fraction* of concordant and discordant pairs then Kendall's τ is defined as

$$\tau = C - D. \tag{11}$$

Note that $c + d \neq \binom{n}{2}$ if there are ties.[2]

To define a probabilistic framework, we must specify three things: (1) a sample space of objects, (2) a distribution over this sample space, and (3) random variables over this sample space. Let our sample space Ω be all possible $2 \cdot \binom{n}{2}$ *ordered* pairs of distinct objects, and consider a *uniform distribution* over this sample space. For a given ranking R, define a random variable $X_R : \Omega \rightarrow \{-1, +1\}$ that outputs $+1$ for any ordered pair concordant with R and -1 for any ordered pair discordant with R.

$$X_R\left[(d_i, d_j)\right] = \begin{cases} 1 & \text{if } d_i \text{ appears before } d_j \text{ in } R. \\ -1 & \text{otherwise.} \end{cases} \tag{12}$$

We thus have a well-defined *random experiment*: draw an ordered pair of objects at random and output $+1$ if that ordered pair agrees with R's ranking and -1 otherwise. Since all ordered pairs of objects are considered uniformly, the *expected value* $E[X_R]$ of this random variable is zero.

Given a second ranked list S, one can similarly define an associated random variable X_S. Now consider the random experiment of *multiplying* the two random variables: the product $X_R \cdot X_S$ will be $+1$ precisely when the pair is *concordant*—i.e. both lists agree that the ordering of the objects is correct ($+1$) or incorrect (-1), and the product will be -1 when the pair is *discordant*—i.e. the lists disagree. In this probabilistic framework, Kendall's τ is the expected value

[2] Kendall defined two means by which τ can account for ties, depending on the desired behavior. Imagine comparing two ranked lists, one of which is almost completely composed of ties. τ_A, defined above, approaches 1. τ_B includes the number of ties in the denominator, and therefore approaches 0. We believe that the former approach is appropriate in this context. Since QRELs are almost exclusively composed of ties (recall that all pairs of unjudged documents in the corpus are considered to be tied), using the latter would mean that effect of the relatively rare meaningful comparisons would be negligible.

of the product of these random variables:

$$\tau = E[X_R \cdot X_S]. \quad (13)$$

The real power of this framework is in the definition of these random variables: (1) the ability to generalize them to compare partial orderings as arise in system evaluation, and (2) the ability to measure the correlation of these random variables using information-theoretic techniques.

4. INFORMATION-THEORETIC RANK CORRELATION

In Section 3, we defined Kendall's τ as the expected product of random variables. The following theorem allows us to restate Kendall's τ equivalently as the mutual information between the random variables.

Theorem 1. $I(X_R; X_S) = \frac{1+\tau}{2}\log(1+\tau) + \frac{1-\tau}{2}\log(1-\tau)$.

(For a proof of Theorem 1, see Appendix). Unlike Kendall's τ, the mutual information between ranked lists ranges from 0 on lists that are completely uncorrelated to 1 on lists that are either perfectly correlated or perfectly anti-correlated.

If we restrict our attention to pairs of lists that are not anti-correlated, then the relationship is bijective. Given this fact, we define a variant of Kendall's τ, *information* τ:

$$\tau_I(R, S) = I(X_R; X_S) \quad (14)$$

where X_R is the ranked list random variable defined in Equation 12 observed with respect to the uniform probability distribution over all pairs of distinct objects. By reframing Kendall's τ equivalently in terms of mutual information, we immediately gain access to a large number of powerful theoretical tools. For example, we can define a conditional information τ between two lists given a third. For lists R and S given T,

$$\tau_I(R, S \mid T) = I(X_R; X_S \mid X_T). \quad (15)$$

Kendall's τ can tell you whether two sets of rankings are similar, but it cannot tell you why. Information τ can be used as a meta-evaluation tool to find the underlying cause of correlation between measures. We demonstrate the use of information τ as a meta-evaluation tool by using it to analyze measures of the *diversity* of information retrieval systems. In recent years, several diversity measures (e.g. [1, 10, 24]) have been introduced to evaluate how well systems perform in response to ambiguous or underspecified queries that have multiple interpretations. These measures conflate several factors [14], including: a diversity model that rewards novelty and penalizes redundancy, and a measure of ad hoc performance that rewards systems for retrieving highly relevant documents. We wish to know not only whether two diversity measures are correlated, but also the similarity between their component diversity models. Using Kendall's τ, we can observe whether the rankings of systems by each measure are correlated. But even if they are correlated, this could still be for one of two reasons: either both the diversity and the performance components evaluate systems similarly; or else one of the components is similar, and its effect on evaluation is dominant. However, if the measures are correlated when conditioned on their underlying performance components, then this must be due to similarities in their models of diversity.

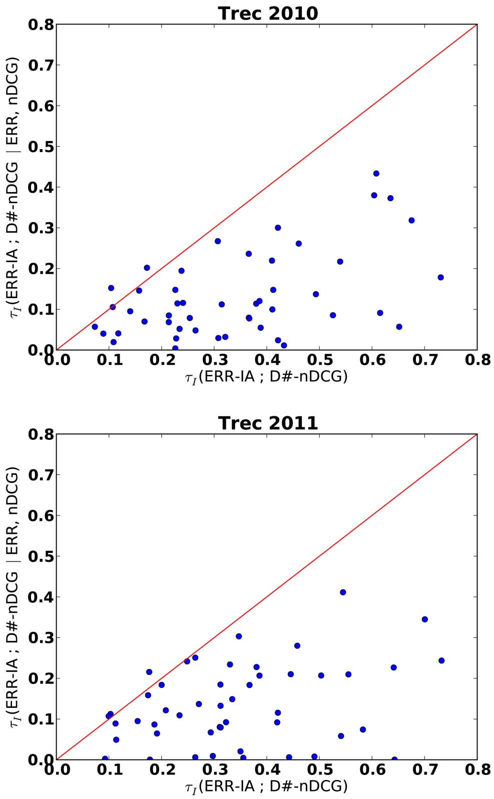

Figure 1: Per-query information τ (conditional rank correlation) between the TREC and NTCIR gold standard diversity measures conditioned on their underlying performance measures.

We measured this effect on the the TREC 2011 and 2012 Web collections [8, 9]. Note that the performance measures are evaluated using graded relevance, while the diversity measures use binary judgments for each subtopic. All evaluations are performed at rank 20. Figure 1 shows the rank correlation between ERR-IA and D#-nDCG, the primary measures reported by TREC and NTCIR [26], when conditioned on their underlying performance models. Each query is computed separately, with each datapoint in the figure corresponding to a different query. Table 1 shows the results of conditioning additional pairs of diversity measures (now averaged over queries in the usual way) on their performance models. The results in Figure 1 are typical of all pairs of measures on a per-query basis.

Our results confirm that while diversity measures are very highly correlated, most of this correlation disappears when one conditions on the underlying performance model. This indicates that most of the correlation is due to the similarity between the performance components and not the diversity components. For example, in TREC 2010, ERR-IA and α-nDCG have an information τ of almost 0.9. However, when conditioned on ERR, the similarity falls to only 0.25. This means that while these two measures are mostly ranking systems for the same reason, that reason is simply ERR. However, of the 0.9 bits that are the same, 0.25 are due to some factor other than ERR. This other factor must presumably be the similarity in their diversity models.

	TREC 2010	TREC 2011
τ_I(ERR-IA ; α-nDCG)	0.8290	0.8375
τ_I(ERR-IA ; α-nDCG | nDCG)	0.4860	0.4434
τ_I(ERR-IA ; α-nDCG | ERR)	0.2499	0.3263
τ_I(ERR-IA ; α-nDCG | nDCG, ERR)	0.2451	0.2805
τ_I(ERR-IA ; D#-nDCG)	0.6390	0.5545
τ_I(ERR-IA ; D#-nDCG | nDCG)	0.3026	0.1728
τ_I(ERR-IA ; D#-nDCG | ERR)	0.1222	0.1442
τ_I(ERR-IA ; D#-nDCG | nDCG, ERR)	0.1239	0.1003

Table 1: TREC 2010 and 2011 information τ (conditional rank correlation) between diversity measures conditioned on ad hoc performance measures.

5. EVALUATION MEASURE

In this section, we demonstrate an extension of our probabilistic framework for evaluation to measuring the correlation between a system and the incomplete ranking generated by a set of relevance judgments. This allows us to define an information-theoretic evaluation measure, *relevance information correlation*. While our measure has novel applications, we will demonstrate that the evaluations produced are consistent with those of existing measures.

To compute mutual information, we must define a sample space, a probability distribution, and random variables. Let the sample space, $\Omega = \{(d_i, d_j)\}$, be the set of all ordered pairs of judged documents. This means that we are ignoring unjudged documents, rather than considering them non-relevant. This is equivalent to computing an evaluation measure on the *condensed list* [23] created by removing all non-judged documents from the list. We define the probability distribution in terms of the QREL to ensure that all ranked lists will be evaluated using the same random experiment. Let $P = U|_{I(g_i \neq g_j)}$, where g_i represents the relevance grade of document d_i, be the uniform probability distribution over all pairs of documents whose relevance grades are not equal. We define a QREL variable Q over ordered pairs of documents as

$$Q\left[(d_i, d_j)\right] = \begin{cases} 1 & \text{if } g_i > g_j \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Note that this definition can be applied to both graded and binary relevance judgments.

We now turn our attention to defining a ranked list random variable over ordered pairs of documents (d_i, d_j). If both document d_i and d_j appear in the ranked list, than our output can simply indicate whether d_i was ranked above d_j. If document d_i appears in the ranked list and d_j does not, then we will consider d_i as having been ranked above d_j, and vice versa. If neither d_i nor d_j is ranked, we will output a null value. If we were to instead restrict our attention only to judged document pairs where at least one document is ranked, then a ranked list consisting of a single relevant document followed by some number of non-relevant documents would have perfect mutual information with the QREL—all of the ranked relevant documents appear before all of the ranked non-relevant documents. However, this system must be penalized for preferring all of the ranked non-relevant documents to all of the unranked relevant documents. If we instead use a null value, our example ranked list would almost always output null. This behavior would be independent of the QREL, meaning the two variables will have almost no mutual information. In effect, the null value creates a recall component for our evaluation measure; no system can have a large mutual information with the QREL unless it retrieves most of the relevant documents.

Another problem we must consider is that mutual information is maximized when two variables are completely correlated or completely *anti*-correlated. Consider an example ranked list consisting of a few non-relevant documents followed by several relevant documents and then many more non-relevant documents. Since this example ranked list will disagree with the QREL on almost all document pairs, its random variable will have a very high mutual information with the QREL variable. The system is effectively being rewarded for finding the subset of non-relevant documents that happen to be present in the QREL. To address this, we truncate the list at the last retrieved relevant document prior to evaluation.

Let r_i represent the rank of document d_i in the list S. Then the ranked list variable R_S is defined as

$$R_S\left[(d_i, d_j)\right] = \begin{cases} 1 & \text{if } r_i < r_j \\ 0 & \text{if neither } d_i \text{ nor } d_j \text{were retrieved} \\ -1 & \text{otherwise.} \end{cases} \tag{17}$$

We define our new measure, *Relevance Information Correlation*, as the mutual information between the QREL variable Q and the truncated ranked list variable R

$$RIC(System) = I(R_{System}; Q). \tag{18}$$

RIC is computed separately for each query, and then averaged, as with mean average precision.

In order to compute RIC we must estimate the joint probability distribution of document preferences over Q and R. This could be done in various ways. In this work, we use the maximum likelihood estimate computed separately for each query. Since the MLE requires a large number of observations, RIC is only accurate for recall-oriented evaluation. In future work, we intend to explore other means of estimating $P(Q, R)$ that will allow RIC to be used for precision-oriented evaluation as well.

We also note that RIC has no explicit rank component, and would therefore seem to treat all relevant documents equally independent of the rank at which they were observed. However, there is an *implicit* rank component in that a relevant document that is not retrieved early in the list must be incorrectly ranked below many non-relevant documents. This argument is similar in spirit to Bpref [2].

Our measure is quite novel in its formulation, and makes many non-standard assumptions about information retrieval evaluation. Therefore it is necessary to validate experimentally that our measure prefers the same retrieval systems as existing measures. Note that for two evaluation measures to be considered compatible, it is sufficient that they rank systems in the same relative order; it is not necessary that they always assign systems similar absolute scores. For example, a system's nDCG is often higher than its average precision.

To show that RIC is consistent with AP and nDCG, we computed the RIC, AP, and nDCG of all systems submitted to TRECs 8 and 9. Figure 2 shows the output of RIC plotted against AP (top) and nDCG (bottom) on TRECs 8 (left) and 9 (right) [28, 29]. TREC 8 uses binary relevance

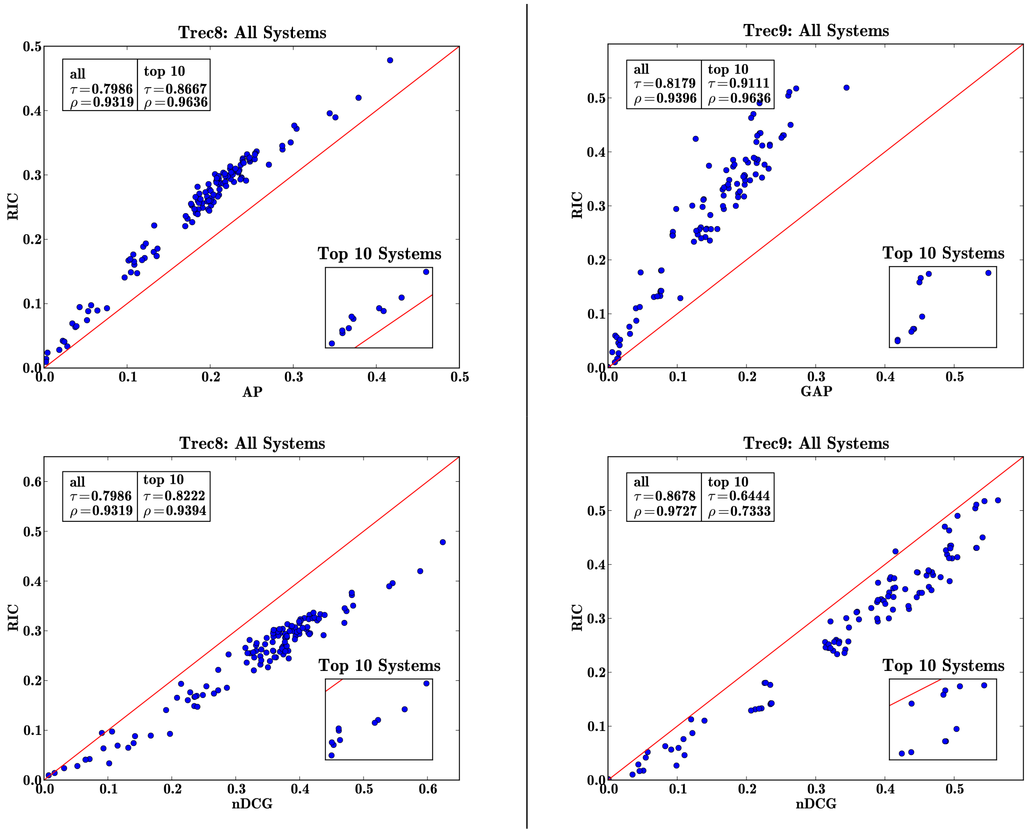

Figure 2: Correlation between RIC and AP (top) and nDCG (bottom). TREC 8 (left) uses binary relevance judgments. TREC 9 (right) uses graded relevance judgments.

	TREC 8	TREC9
(G)AP	0.716	0.648
nDCG	0.713	0.757
MI	0.719	0.744

Table 2: Discriminative power of (graded) AP and nDCG vs. RIC

judgments. TREC 9 uses graded relevance judgments, requiring the use of graded average precision. Inset into each plot is the output of the measures on the top ten systems. For each experiment, we report the Kendall's τ and Spearman's ρ [27] rank correlations for all systems, and for the top ten systems. With Kendall's τ values of at least 0.799 on all systems and 0.644 on top ten systems, the ranking of systems by RIC is still highly correlated with those of both AP and nDCG. However, RIC is not as highly correlated with either AP or nDCG as AP and nDCG are with each other. Note that the correlation between RIC and GAP on TREC 9 is highly monotonic, even if is not particularly linear. This implies that the two measures do rank systems in a consistent relative order, even if RIC is a biased estimator of GAP.

To further validate our measure, we also compute the *discriminative power* [22] of the various measures. Discriminative power is a widely used tool for evaluating a measure's *sensitivity* i.e. how often differences between systems can be detected with high confidence. A high sensitivity can be seen as a necessary, though not sufficient, condition for a good evaluation measure. Discriminative power is defined as the percentage of pairs of runs that are found to be statistically significantly different by some significance test. As per Sakai, we use a two-tailed paired bootstrap test with 1000 bootstrap samples per pair of systems. Our results are displayed in Table 2. As measured by discriminatory power, we see that RIC is at least as sensitive, if not more so, than AP and nDCG.

6. UPPER BOUND ON METASEARCH

In Section 5, we defined an evaluation measure in terms of mutual information. One advantage of this approach is that collections of systems can be evaluated directly by considering the output of their random variables jointly, without their needing to be combined. For a collection of systems, denoted S_1 through S_n, the relevance information correlation can be defined as

$$RIC(S_1, \ldots, S_n) = I(R_{S_1}, \ldots, R_{S_n}; Q) \qquad (19)$$

In this section, we will show that this produces a natural upper bound on metasearch performance that is consistent with other upper bounds appearing in the literature.

We compare our upper bound against those of Montague [19]. Montague describes metasearch algorithms as sorting functions whose comparators, as well as the documents to be sorted, are defined in terms of collections of input systems.

By also using the QREL as input, these algorithms can estimate upper bounds on metasearch performance. These bounds range from the ideal performance that cannot possibly be exceeded by any metasearch algorithm, to descriptions of reasonable metasearch behavior that should be similar to the performance of any quality metasearch algorithm.

Montague defines the following upper bounds on metasearch:

1. Naive: Documents are sorted by comparison of relevance judgments, i.e. the naive upper bound is created by returning all relevant documents returned by any system in the collection above any non-relevant document. Relevant documents not retrieved by any system are not ranked.

2. Pareto: If document A is ranked above document B by all systems, then document A is considered "greater" than document B. Otherwise, the documents are sorted by comparison of relevance judgments.

3. Majoritarian: If document A is ranked above document B by at least half of the systems, then document A is considered "greater" than document B. Otherwise, the documents are sorted by comparison of relevance judgments.

We will compare our direct joint evaluation with these upper bounds, and several metasearch algorithms commonly used as baselines in the IR literature: the CondorcetFuse metasearch algorithm [20], and the comb family of metasearch algorithms [25].

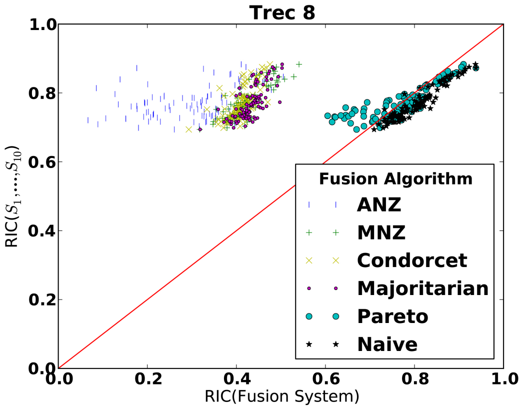

Figure 3: RIC of systems output by metasearch algorithms (Fusion System) versus RIC of systems computed directly ($S_1, \ldots, S_{10}$) without combining.

We examined the direct evaluation and metasearch performance of collections of ten randomly selected systems. Experiments were performed on TREC 8 and 9, with both binary and graded relevance judgments. To conserve space, we only show the results from TREC 8. The results from TREC 9 were highly similar, both when using binary and graded relevance judgments.

Figure 3 shows the RIC of the system output by a metasearch algorithm plotted against the joint RIC of the input systems, and Table 3 shows various measures of their correlation. Montague found that combANZ is inferior to CondorcetFuse and combMNZ, CondorcetFuse and combMNZ perform comparably to the Majoritarian bound, and the Naive bound is not appreciably better than the Pareto bound. If direct evaluation and the Naive bound are both reasonable estimates of the actual upper bound, then these results should be confirmed by Figure 3 and Table 3, as indeed they are. Note that there is almost no correlation between the joint evalution and the weakest metasearch algorithm, combANZ: combANZ does not approximate the upper bound on metasearch. The correlation improves as the quality of the metasearch algorithm improves, and it does so in a manner consistent with Montague. The correlations between the joint evaluation and the output of combMNZ, CondorcetFuse, and the Majoritarian bound are similar; while they are still biased as estimators, the correlation is beginning to approach monotonicity. Finally, with a root mean square error of 0.039, the joint evaluation estimation of the upper bound is essentially identical to that of the Naive upper bound. If the Naive upper bound is a reasonable estimate of the upper bound on metasearch performance, then so is the joint evaluation of the input systems.

TREC 8	τ	ρ	RMSE
ANZ	0.221	0.330	0.481
MNZ	0.587	0.764	0.351
Condorcet	0.519	0.689	0.362
Majoritarian	0.552	0.735	0.340
Pareto	0.657	0.836	0.044
Naive	0.788	0.931	0.039

Table 3: Correlation between joint distribution and metasearch algorithms (Kendall's τ, Spearman's ρ, root mean square error).

7. INFORMATION DIFFERENCE

In this section, we introduce a novel application of our probabilistic framework. Imagine that you are attempting to improve an existing ranker. On what basis do you decide whether or not your changes are beneficial? One typically evaluates both systems on a number of queries, and measures the difference in average performance. If one system outperforms the other, whether you have made an improvement is clear. But what happens when the systems perform similarly? It could be that your new system is essentially unchanged from your old system, but it is also possible that the two systems chose highly different document and just happened to have very similar evaluation scores. In the latter case, it may be possible to create a new, better system based on a combination of the two existing systems.

We propose to measure the magnitude of the difference between systems in their ranking of documents for which we have relevance information, rather than the magnitude of the difference between their performance. We denote this new quantity as the *information difference* between systems. Our definition of information difference is inspired by the Boolean Algebra symmetric difference operator as applied to information space (see Figure 4).

$$id(S_1, S_2) = I(S_1; Q \mid S_2) + I(S_2; Q \mid S_1) \qquad (20)$$

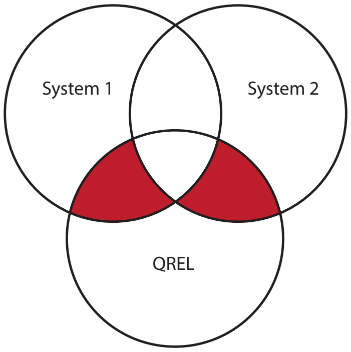

Figure 4: Information difference corresponds to the symmetric difference between the intersections of the systems with the QREL in information space (red portion of the Venn diagram).

As a preliminary validation of information difference, we analyzed the change in AP and information difference between pairs of systems submitted to TREC 8, selected at random. We expect the two to be somewhat directly correlated, since, in general, if two systems rank documents similarly, we would expect them to have similar AP. However, we expect that they will not be highly correlated, since we believe that information difference is much more informative. Our intuition is supported by Figure 5, which shows the magnitude of the change in AP on the horizontal axis, and the information difference on the vertical axis.

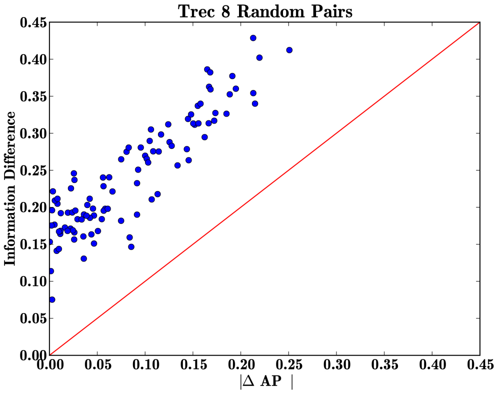

Figure 5: Scatter plot of information difference and the magnitude of change in AP of random pairs of TREC 8 systems.

To demonstrate the utility of information difference, we sorted all the systems submitted to TREC 8 by AP and separated them into twenty equal-sized bins. By construction, each bin contained systems with small differences in performance. Our goal is to distinguish between similar and dissimilar systems within each bin. To this end, all systems within each bin were compared with one other (see Table 4). When the system pairs were sorted by their information difference, both systems in the first 27 pairs were submitted by the same group, whereas sorting by |Δ AP| produced no discernible pattern. It is reasonable to assume that these systems were different instantiations of the same underlying technology. We can therefore conclude that information difference is able to determine whether systems with the same underlying performance are in fact similar, as desired.

Rank	System 1	System 2	*id*	|Δ AP|
1	UB99T	UB99SW	0.010	0.005
2	unc8al32	unc8al42	0.012	0.002
3	fub99tt	fub99tf	0.017	0.000
4	nttd8al	nttd8alx	0.023	0.002
5	ibmg99a	ibmg99b	0.027	0.012
		⋮		
28	isa25t	cirtrc82	0.084	0.004
29	CL99SD	CL99SDopt2	0.086	0.000
30	ok8amxc	ok8alx	0.086	0.006
31	tno8d4	MITSLStd	0.088	0.016
32	uwmt8a2	uwmt8a1	0.089	0.002

Table 4: The systems from TREC 8 were binned by average precision. Information difference and Δ AP were computed for all system pairs within each bin. Sorting by information difference, both systems in the first 27 pairs were submitted by the same group.

8. CONCLUSION

In this work, we developed a probabilistic framework for the analysis of information retrieval systems based on the correlation between a ranked list and the preferences induced by relevance judgments. Using this framework, we developed powerful information theoretic tools for better understanding information retrieval systems. We introduced four preliminary uses of our framework: (1) a measure of conditional rank correlation, *information* τ, which is a powerful meta-evaluation tool whose use we demonstrated on understanding novelty and diversity evalution; (2) a new evaluation measure, *relevance information correlation*, which is correlated with traditional evaluation measures and can be used to (3) evaluate a collection of systems simultaneously, which provides a natural upper bound on metasearch performance; and (4) a measure of the similarity between rankers on judged documents, *information difference*, which allows us to determine whether systems with similar performance are actually different.

Our framework is based on the choice of sample space, probability distribution, and random variables. Throughout this work, we only used a uniform distribution on appropriate pairs of documents. However, not all document pairs are equal. The use of additional distributions is an immediate avenue for improvement that we intend to explore in future work. For example, a geometric distribution may be employed to force our evaluation tools to concentrate their attention at the top of a ranked list.

The primary limitation of our evaluation measure as implemented in this work is that it is only applicable to recall-oriented retrieval tasks. In future work, we intend to develop a precision-oriented version that is applicable to web search. Given such a measure, judgments can be combined in the way systems were in our upper bound on metasearch. In that way, a small number of expensive to produce nominal relevance judgments, a somewhat larger number of somewhat less expensive preference judgments, and a gold-stander ranker could all be used simultaneously to evaluate systems.

Finally, we intend to explore the application of information difference to the understanding of information retrieval models. For example, BM25 and Language Models have long been used as baselines in information retrieval experiments. On the surface, these two models appear to be completely different. And yet, the two share deep theoretical connections [33]. Using information difference, we can determine whether their theoretical similarities outweigh their superficial differences in terms of how they rank documents.

9. REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[2] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, 2004.

[3] Christopher J.C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

[4] Ben Carterette and Paul N. Bennett. Evaluation measures for preference judgments. In *SIGIR*, 2008.

[5] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there: preference judgments for relevance. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, 2008.

[6] Praveen Chandar and Ben Carterette. Using preference judgments for novel document retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, 2012.

[7] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[8] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Gordon V. Cormack. Overview of the TREC 2010 Web Track. In *19th Text REtrieval Conference*, Gaithersburg, Maryland, 2010.

[9] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. Overview of the TREC 2011 Web Track. In *20th Text REtrieval Conference*, Gaithersburg, Maryland, 2011.

[10] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[12] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 87–94, New York, NY, USA, 2008. ACM.

[13] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4, December 2003.

[14] Peter B. Golbus, Javed A. Aslam, and Charles L.A. Clarke. Increasing evaluation sensitivity to diversity. In *Journal of Information Retrieval*, To Appear.

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

[16] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, 2002.

[17] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, June 1938.

[18] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, December 2008.

[19] Mark Montague. *Metasearch: Data Fusion for Document Retrieval*. PhD thesis, Dartmouth College. Dept. of Computer Science, 2002.

[20] Mark Montague and Javed A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, 2002.

[21] Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. Extending average precision to graded relevance judgments. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, 2010.

[22] Tetsuya Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, 2006.

[23] Tetsuya Sakai. Alternatives to Bpref. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, 2007.

[24] Tetsuya Sakai and Ruihua Song. Evaluating diversified search results using per-intent graded relevance. In *SIGIR*, pages 1043–1052, 2011.

[25] Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.

[26] Ruihua Song, Min Zhang, Tetsuya Sakai, Makoto P. Kato, Yiqun Liu, Miho Sugimoto, Qinglei Wang, and Naoki Orii. Overview of the ntcir-9 intent task. In *Proceedings of the 9th NTCIR Workshop*, Tokyo, Japan, 2011.

[27] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 1904.

[28] E. M. Voorhees and D. Harman. Overview of the eighth text retrieval conference (TREC-8). In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, 2000.

[29] E. M. Voorhees and D. Harman. Overview of the ninth text retrieval conference (TREC-9). In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, 2001.

[30] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, 2006.

[31] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, 2008.

[32] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and nDCG. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, 2008.

[33] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004.

10. APPENDIX

THEOREM 1. *$I(X_R; X_S) = \frac{1+\tau}{2}\log(1+\tau) + \frac{1-\tau}{2}\log(1-\tau)$.*

PROOF. Denote X_R and X_S as X and Y. Consider the following joint probability distribution table.

		Y	
		-1	1
X	-1	a	b
	1	c	d

Observe that: $a+b+c+d = 1$; $C = a+d$, $D = b+c$, and therefore $\tau = a+d-b-c$; and since document pairs appear in both orders, $a = d$ and $b = c$.

The joint probability distribution can be rewritten as follows.

		Y	
		-1	1
X	-1	$\frac{C}{2}$	$\frac{D}{2}$
	1	$\frac{D}{2}$	$\frac{C}{2}$

Observe that the marginal probability $P(X) = P(Y) = \left(\frac{C}{2}+\frac{D}{2}, \frac{C}{2}+\frac{D}{2}\right) = \left(\frac{1}{2}, \frac{1}{2}\right)$.

$$\begin{aligned} I(X;Y) &= KL(P(X,Y)||P(X)P(Y)) \\ &= \sum_{x,y} p(x,y) \lg \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \lg p(x,y) + \sum_{x,y} p(x,y) \lg \frac{1}{p(x)p(y)}. \end{aligned}$$

Since $P(X,Y) = \left(\frac{C}{2}, \frac{D}{2}, \frac{C}{2}, \frac{D}{2}\right)$ and $P(X)P(Y) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$,

$$\begin{aligned} I(X,Y) &= 2 \cdot \frac{C}{2} \lg \frac{C}{2} + 2 \cdot \frac{D}{2} \lg \frac{D}{2} + 2 \cdot \frac{C}{2} \lg 4 + 2 \cdot \frac{D}{2} \lg 4 \\ &= C \lg \frac{C}{2} + D \lg \frac{D}{2} + 2C + 2D \\ &= C \lg C - C + D \lg D - D + 2C + 2D \\ &= C \lg C + D \lg D + 1 \\ &= C \lg C + (1-C) \lg(1-C) + 1 \end{aligned}$$

Since $C + D = 1$ and $\tau = C - D$, we have that $\tau = 2C - 1$, $C = \frac{1+\tau}{2}$ and $D = 1 - C = \frac{1-\tau}{2}$.

In terms of C, if H_2 represents the entropy of a Bernoulli random variable ,[3]

$$\begin{aligned} I(X;Y) &= -H_2(C) + 1 \\ &= -H_2\left(\frac{1+\tau}{2}\right) + 1 \\ &= \frac{1+\tau}{2} \lg \frac{1+\tau}{2} + \frac{1-\tau}{2} \lg \frac{1-\tau}{2} + 1 \\ &= \frac{1+\tau}{2} \lg(1+\tau) - \frac{1+\tau}{2} + \frac{1-\tau}{2} \lg(1-\tau) \\ &\quad - \frac{1-\tau}{2} + 1 \\ &= \frac{1+\tau}{2} \lg(1+\tau) + \frac{1-\tau}{2} \lg(1-\tau) \end{aligned}$$

□

COROLLARY 1. *For two ranked lists R and S, $I(X_R; X_S) = 1 - H_2(K)$ where $K = \frac{1-\tau}{2}$ is the normalized Kendall's τ distance between R and S.*

[3] $H_2(p) = -p \lg p - (1-p) \lg(1-p)$. Note that $H_2(p) = H_2(1-p)$.