A Simple and Efficient Sampling Method for Estimating AP and NDCG

Emine Yilmaz* eminey@microsoft.com

Evangelos Kanoulas† ekanou@ccs.neu.edu

Javed A. Aslam†* jaa@ccs.neu.edu

*Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB, UK

†College of Computer and Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115

ABSTRACT

We consider the problem of large scale retrieval evaluation. Recently two methods based on *random sampling* were proposed as a solution to the extensive effort required to judge tens of thousands of documents. While the first method proposed by Aslam et al. [1] is quite accurate and efficient, it is overly complex, making it difficult to be used by the community, and while the second method proposed by Yilmaz et al., *infAP* [14], is relatively simple, it is less efficient than the former since it employs uniform random sampling from the set of complete judgments. Further, none of these methods provide confidence intervals on the estimated values.

The contribution of this paper is threefold: (1) we derive confidence intervals for infAP, (2) we extend infAP to incorporate nonrandom relevance judgments by employing stratified random sampling, hence combining the efficiency of stratification with the simplicity of random sampling, (3) we describe how this approach can be utilized to estimate nDCG from incomplete judgments. We validate the proposed methods using TREC data and demonstrate that these new methods can be used to incorporate nonrandom samples, as were available in TREC Terabyte track '06.

Categories and Subject Descriptors: H.3 Information Storage and Retrieval; H.3.4 Systems and Software: Performance Evaluation

General Terms: Experimentation, Measurement, Theory

Keywords: Evaluation, Sampling, Incomplete Judgments, Average Precision, nDCG, infAP

1. INTRODUCTION

We consider the problem of large scale retrieval evaluation, in particular, retrieval evaluation with incomplete relevance judgments.

*We gratefully acknowledge the support provided by NSF grants IIS-0533625 and IIS-0534482.

SIGIR'08, July 20–24, 2008, Singapore.

The most commonly employed methodology for assessing the quality retrieval systems is the Cranfield methodology adopted by TREC [7]. The main assumption behind this methodology is that the relevance judgments are *complete*, i.e., for each query, all documents in the collection relevant to this query are identified.

When the document collection is large, obtaining complete relevance judgments is infeasible due to the need for extensive human effort. Instead, TREC employs *depth-k pooling* (typically, depth-100 pooling) to overcome this burden, combining top k documents retrieved by the submitted systems and assuming the rest of the documents are nonrelevant. While a depth-100 pool is significantly smaller than the document collection, it still requires extensive judgment effort (e.g. 86,830 judgments for TREC 8). Furthermore, even though depth-100 pools were shown to contain most of the relevant documents [8], the size of document collections tends to increase and thus these pools may be inadequate for identifying most of the relevant documents.

Recently, evaluation with incomplete relevance judgments has gained attention as a solution to the problem of extensive judgment effort. New evaluation measures have been proposed in the literature [1, 5, 11, 12, 14] since standard evaluation measures such as average precision are not robust to incomplete judgments [3].

As a solution to this problem, Buckley and Voorhees [3] proposed *bpref*, a now commonly used measure by information retrieval community. Sakai [12] instead applied traditional measures to condensed lists of documents obtained by filtering out all unjudged documents from the original ranked lists and showed that these versions of measures are actually more robust to incompleteness than bpref.

Carterette et al. [5] and Moffat et al. [11] select a subset of documents to be judged based on the benefit documents provide in fully ranking systems or identifying the best systems, respectively.

Even though the aforementioned approaches are all shown to be more robust to incompleteness than standard evaluation measures, these methods are not guaranteed to compute or estimate the values of standard evaluation measures. Hence, the values of measures obtained by these methods are difficult to interpret.

Yilmaz and Aslam [14] and Aslam et al. [1] instead use random sampling to estimate the actual values of average precision when relevance judgments are incomplete. Both of these methods are based on treating incomplete relevance

judgments as a sample drawn from the set of complete judgments and using statistical methods to estimate the actual values of the measures. These methods are both shown to (1) produce unbiased estimates of average precision even when relevance judgments are incomplete and (2) be more robust to incomplete relevance judgments than any other measures such as *bpref* [3] or the condensed versions of the measures (referred to as induced AP in the paper)[14].

The measure proposed by Yilmaz and Aslam, *infAP* [14], became a commonly used measure by information retrieval community [2, 13] and was used in TREC VID and Terabyte tracks in 2006 [10, 4].

A limitation of *infAP* accrues from the measure's assumption that incomplete relevance judgments are a *simple random* sample drawn from the set of complete judgments. Typical evaluation measures give more weight to documents retrieved towards the top of a retrieved lists and therefore, a "top-heavy" sampling strategy would lead to more accurate results with higher efficiency in terms of judgment effort needed.

On the other hand, according to the method by Aslam et al. [1] samples are drawn according to a carefully chosen non-uniform distribution over the documents in the depth-100 pool. Even though this method is more efficient in terms of judgment effort than infAP, it is very complex both in conception and implementation and therefore less applicable.

Furthermore, although average precision estimators as proposed by both of the aforementioned methods are unbiased in expectation, in practice, when calculated using a single sample of relevance judgments, may vary in value. This necessitates the derivation and use of confidence intervals around the estimated values in order to allow confident conclusions regarding the actual value of average precision and thus the ranking of retrieval systems.

In this paper, we mainly focus on inferred average precision. First, we derive confidence intervals for the measure and validate them using TREC data. We show that infAP along with the corresponding confidence intervals can allow researchers to reach confident conclusions about actual average precision, even when relevance judgments are incomplete.

We then focus on the efficiency of the measure. We employ a stratified random sampling methodology and extend the measure to incorporate relevance judgments created according to *any* such sampling distribution. This extended infAP combines the simplicity of random sampling with the efficiency of stratification and thus it is simple and easy to compute while, at the same time, it is much more efficient than infAP in terms of reducing the judgment effort. We further claim that the same methodology can be applied to other evaluation measures and demonstrate how nDCG (a commonly used measure that incorporates graded relevance judgments [9]) can be estimated using incomplete relevance judgments.

2. CONFIDENCE INTERVALS FOR INFAP

The inferred average precision, by statistical construction, is an unbiased estimator of average precision and thus it is designed to be exactly equal to average precision in expectation. However in practice, it may be low or high due to the nature of sampling (especially when the subsets of documents whose binary relevance is available is small). In other words, there is variability in the values of infAP because different samples from the collection of documents give rise to different values of infAP. The amount of the variability in infAP is measured by its *variance*.

Before computing the variance of infAP let's revisit the random experiment whose expectation is average precision [14] and identify all sources of variability in the outcome of this random experiment. Given a ranked list of documents with respect to a given topic:

1. Select a relevant document at random and let the rank of this relevant document in list be k.
2. Select a rank, j, at random from the set $\{1, ..., k\}$.
3. Output the binary relevance of the document at rank j.

In expectation, steps (2) and (3) effectively compute the *precision* at a relevant document and in combination, step (1) computes the *average* of these precisions.

The aforementioned experiment can be realized as a two-stage sampling. At the first stage — step (1) — a sample of *cut-off levels* at relevant documents is selected. The infAP value is computed as an average of the estimated precision values at the sampled cut-off levels. Even if we assume that these precision values are the actual precision values, infAP varies because different samples of cut-off levels will result in different values of infAP. Therefore, computing infAP using precision values only at a subset of cut-off levels introduces the first component of variability.

Let *rel* be the set of the judged relevant documents of size r. This first variance component can be estimated as[1]

$$\text{var. comp. } 1 = (1-p) \cdot s^2/r$$

where $p \cdot 100\%$ is the sampling percentage and s^2 the variance among the precision values at the judged relevant documents calculated as $s^2 = \left(\sum_{k \in \text{rel}} (\widehat{PC}_k - \text{infAP})\right)/r$.

At the second stage — step (2) — for each one of the selected cut-off levels, a sample of *documents* above that cut-off level document is used to estimate the corresponding to the cut-off precision value. Therefore, even for a given sample of cut-off levels, infAP has variability because different samples of documents give rise to different values of precisions and thus different values of infAP. Hence, computing the precision at some cut-off using only a subset of the documents above that cut-off introduces a second component of variability.

Assuming that precisions at different cut-off levels are independent from each other, this second variance component can be estimated as,

$$\text{var. comp. } 2 = \Big(\sum_{k \in \text{rel}} var[\widehat{PC}_k] \Big)/r^2$$

where $var[\widehat{PC}_k]$ is the variance of the estimated precision at cut-off k.

According to the Law of Total Variance, the total variance of infAP can be computed as the sum of the two aforementioned variance components; hence,

$$var[\text{infAP}] = (1-p) \cdot \frac{s^2}{r} + \frac{\sum_{k \in \text{rel}} var[\widehat{PC}_k]}{r^2}$$

[1]The complete formula of infAP variance along with the derivation can be found at the Appendix

When evaluating retrieval systems, the average of infAP values across all topics (mean infAP) is employed. The variance of the mean infAP can be computed as a function of the variance of infAP as

$$var[\text{mean infAP}] = \sum var[\text{infAP}]/(\#\text{ of queries})^2$$

According to the Central Limit Theorem one can assign 95% confidence intervals to mean infAP as a function of its variance. A 95% confidence interval centered at the mean infAP intimate that with 95% confidence the actual value of MAP is within this interval.

We used TREC 8,9 and 10 data to validate the derived variance of the mean infAP when relevance judgments are incomplete. We simulated the effect of incomplete relevance judgments as in [14]. For each TREC, we formed incomplete judgments sets by sampling from the entire depth-100 pool over all submitted runs. This is done by selecting p% of the complete judgment set uniformly at random, where $p \in \{10, 20, 30\}$. The results of our experiments led to identical conclusions over all TREC dataset and therefore, due to space limitations, we report only results for TREC 8.

Figure 1 illustrates the mean infAP values computed from a single random sample of documents per topic for each run against the actual MAP values for $p \in \{10, 20, 30\}$ for TREC 8. The 95% confidence intervals are depicted as error bars around the mean infAP values. As one can observe, the greatest majority of the confidence intervals intersect the 45^o dashed line indicating that the greatest majority of the confidence intervals cover the actual MAP values.

Furthermore, we computed the mean infAP values and the corresponding confidence intervals for 100 different sampling trials over TREC 8 data and we accumulated the deviation of the computed mean infAP values from the actual MAP values in terms of standard deviations. This way we generated a Cumulative Distribution Function of divergence of mean infAP values per system. According to the Central Limit Theorem each of these CDF's should match the CDF of the Normal Distribution. We performed a Kolmogorov-Smirnov test of fitness and for 90% of the systems the hypothesis that the two CDF's match could not be rejected ($\alpha = 0.05$) which validates our derived theoretical results.

3. INFERRED AP ON NONRANDOM JUDGMENTS

In the previous section we derived confidence intervals for infAP in a setup where documents to be judged were a random subset of the entire document collection. Confidence intervals can be further reduced (i.e. the accuracy of the estimator can be improved) by utilizing a "top-heavy" sampling strategy. In this section we consider a setup where relevance judgments are not a random subset of complete judgments and show how infAP can be extended to produce unbiased estimates of average precision in such a setup. We denote the extended infAP measure as *xinfAP*.

Similar to the infAP paradigm, consider the case where we would like to evaluate the quality of retrieval systems with respect to a complete pool and assume that relevant judgments are incomplete. Further assume that the set of available judgments are constructed by diving the complete collection of documents into disjoint contiguous subsets (strata) and then randomly selecting (sampling) some documents from each stratum to be judged. The sampling within each stratum is performed independently, therefore, the sampling percentage can be chosen to be different for each stratum. For instance, one could choose to split the collection of documents into two strata (based on where they appear in the output of search engines), and sample 90% of the documents from the first stratum and 30% of the documents from the second stratum. In effect, one could think a large variety of sampling strategies in terms of this multi-strata strategy. For example, the sampling strategy proposed by Aslam et al. [1] can be thought as each stratum containing a single document, with different sampling probabilities assigned to different strata.

Let $\widehat{AP}$ be the random variable corresponding to the estimated average precision of a system. Now consider the first step of the random experiment whose expectation corresponds to average precision, i.e. picking a relevant document at random. Note that in the above setup, this relevant document could fall into any one of the different strata s. Since the sets of documents contained in the strata are disjoint, by definition of conditional expectation, one can write $E[\widehat{AP}]$ as:

$$E[\widehat{AP}] = \sum_{\forall s \in \text{Strata}} P_s \cdot E_s[\widehat{AP}]$$

where P_s corresponds to the probability of picking the relevant document from stratum s and $E_s[\widehat{AP}]$ corresponds to the expected value of average precision given that the relevant document was picked from stratum s.

Let R_Q be the total number of relevant documents in the complete judgment set and R_s be the total number of relevant documents in stratum s if we were to have all complete relevance judgments. Then, since selecting documents from different strata is independent for each stratum, the probability of picking a relevant document from stratum s is, $P_s = R_s/R_Q$.

Computing the actual values of R_Q and R_s is not possible, since the complete set of judgments is not available. However, we can estimate their values using the incomplete relevance judgments. Let r_s be the number of sampled relevant documents and n_s be the total number of sampled documents from stratum s. Furthermore, let N_s be the total number of documents in stratum s. Since the n_s documents were sampled uniformly from stratum s, the estimated number of relevant documents within stratum s, $\hat{R}_s$, can be computed as $\hat{R}_s = (r_s/n_s) \cdot N_s$. Then the number of relevant documents in query Q can be estimated as the sum of these estimates over all strata, i.e. $\hat{R}_Q = \sum_{\forall s} \hat{R}_s$. Given these estimates, the probability of picking a relevant document from stratum s can be estimated by, $\hat{P}_s = \hat{R}_s/\hat{R}_Q$.

Now, we need to compute the expected value of estimated average precision, $E_s[\widehat{AP}]$, if we were to pick a relevant document at random from stratum s.

Since the incomplete relevance judgments within each stratum s is a uniform random subset of the judgments in that stratum, the induced distribution over relevant documents within each stratum is also uniform, as desired. Therefore, the probability of picking any relevant document within this stratum is equal. Hence, the expected estimated average precision value within each stratum, $E_s[\widehat{AP}]$, can be computed as the average of the precisions at judged (sampled) relevant documents within that stratum.

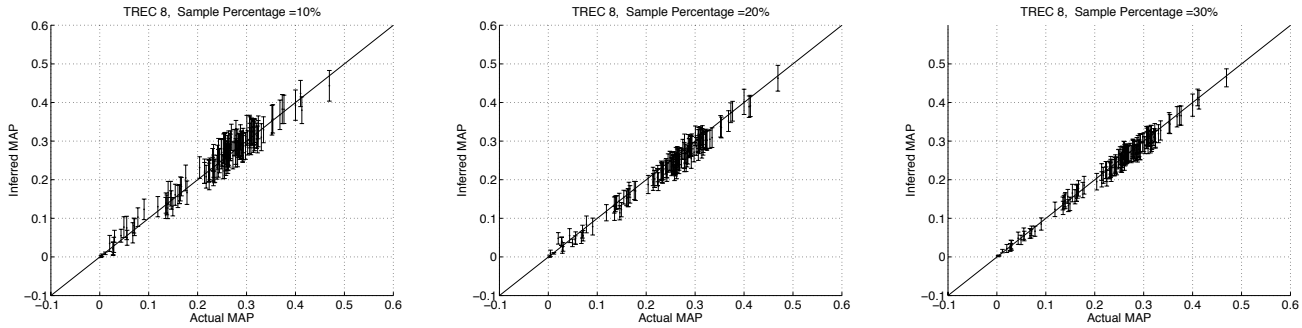

Figure 1: TREC-8 mean inferred AP along with estimated confidence intervals when relevance judgments are generated by sampling 10, 20 and 30 % of the depth-100 pool versus the mean actual AP.

Now consider computing the expected precision at a relevant document at rank k, which corresponds to the expected outcome of picking a document at or above rank k and outputting the binary relevance of the document at this rank (steps 2 and 3 of the random experiment).

When picking a document at random at or above rank k and outputting the binary relevance of that document, one of the following two cases may occur. With probability $1/k$, we pick the current document, and since this document is by definition relevant the outcome is 1. With probability $(k-1)/k$ we pick a document above the current document, in which case we need to calculate the expected precision (or expected binary relevance) with respect to the documents above rank k. Thus,

$$E[\widehat{PC}_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[\widehat{PC}_{\text{above } k}]$$

Let N_s^{k-1} be the total number of documents above rank k that belong stratum s, n_s^{k-1} be the total number of judged (sampled) documents above rank k that belong to stratum s and r_s^{k-1} be the total number of judged (sampled) relevant documents above rank k that also belong to stratum s.

When computing the expected precision within the $(k-1)$ documents above rank k, with probability $N_s^{k-1}/(k-1)$ we pick a document from stratum s. Therefore, the expected precision above rank k can be written as:

$$E[\text{prec above } k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[\widehat{PC}_{\text{above } k}]$$

where $E_s[\widehat{PC}_{\text{above } k}]$ is the expected precision above rank k within stratum s. Since we have a uniform sample of judged documents from stratum s, we can use these sampled documents to estimate the expected precision within stratum s. Since the incomplete relevance judgments from each stratum is obtained by uniform random sampling, this expected precision can be computed as r_s^{k-1}/n_s^{k-1}.

Note that in computing the expected precision in stratum s, we may face the problem of not having sampled any documents from this stratum that are above the current relevant document at rank k. Adapting the same idea used in infAP, we employ Lindstone smoothing [6] to avoid this problem. Therefore, expected precision above rank k can be computed as:

$$E[\widehat{PC}_{\text{above } k}] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot \frac{r_s^{k-1} + \epsilon}{n_s^{k-1} + 2\epsilon}$$

It is easy to see that when complete judgments are available, xinfAP is exactly equal to average precision (ignoring the smoothing effect). Further, note that infAP is a particular instantiation of this formula with a single stratum used.

Overall, the advantage and real power of the described stratified random sampling and the derived AP estimator, xinfAP, is the fact that it combines the effectiveness of the sampling method proposed by Aslam et al. [1] by employing stratification of the documents and thus better utilization of the judgment effort with the simplicity of infAP by employing random sampling within each stratum.

3.1 Inferred AP in TREC Terabyte

As mentioned, xinfAP can be used with a large variety of sampling strategy. In this section, we focus on the sampling strategy used in TREC Terabyte 2006 [4] and we show that (1) *xinfAP* is highly effective at estimating average precision and (2) it better utilizes the judgment effort compared to infAP.

First, let's briefly consider the sampling strategy used in TREC Terabyte 2006. In this track, three different sets of relevance judgments were formed, with only two of them being used for evaluation purposes. Out of these two sets, the first set of judgments, constructed by the traditional depth-50 pooling strategy, was used to obtain a rough idea of the systems average precision. The second set of judgments was constructed using random sampling in such a way that there are more documents judged from topics that are more likely to have retrieved more relevant documents. Since, in Terabyte track, the size of the document collection is very large, the systems may continue retrieving relevant documents even at high ranks (deeper in the list). This set of judgments was created to obtain an estimate of average precision if complete judgments were present.

To estimate average precision, infAP was used as the evaluation measure. Since, by design, infAP assumes that the set of relevance judgments is a random subset of complete judgments, even though the entire depth-50 pool was judged, infAP was computed only using the random sample of judgments (second set) without utilizing judgments from the depth-50 pool. Therefore, many relevance judgments were not used even though they were available.

Note that xinfAP can easily handle this setup and it could be used to utilize all the judgments, obtaining better estimates of average precision.

To test how xinfAP compares with infAP we simulate the sampling strategy used in TREC Terabyte 06 on data from

TREC 8. The TREC Terabyte data was not used due to the fact that in TREC Terabyte the actual value of average precision is not known since complete judgments are not available.

To simulate the setup used in TREC Terabyte, we first form different depth-k pools where $k \in \{1, 2, 3, 4, 5, 10, 20, 30, 40, 50\}$ and obtain judgments for all documents in each one of these pools. Then, for each value of k, we compute the total number of documents that are in the depth-k pool and we randomly sample equal number of documents from the complete judgment set[2] excluding the depth-k pool. After forming these two sets of judgments (depth-k and random) we combine them and compute xinfAP on these combined judgments.

This setup exactly corresponds to a sampling strategy where complete judgments are divided into two strata and judgments are formed by uniformly and independently sampling within each stratum.

Note that in TREC, there are some systems that were submitted but that did not contribute to the pool. To further evaluate the quality of our estimators in terms of their robustness for evaluating the quality of unseen systems (systems that did not contribute to the pool), when we form the incomplete relevance judgments, we only consider the systems that contribute to the pool but we compute the xinfAP estimates for all submitted systems.

Figure 2 demonstrates how xinfAP computed using judgments generated by combining (left) depth-10, (middle) depth-5 and (right) depth-1 pools with equal number of randomly sampled judgments compares with the actual AP. Each of these depths correspond to judging 23.1%, 12.7% and 3.5% of the entire pool, respectively. The plots report the RMS error (how accurate are the estimated values?), the Kendall's τ value (how accurate are the estimated rankings of systems?) and the linear correlation coefficient, ρ, (how well do the estimated values fit in a straight line compared to the actual values?). The dot signs in the figures refer to the systems that were used to create the original pools and the plus signs refer to the systems that did not contribute to the pool.

The results illustrated in these plots reinforce our claims that xinfAP is an unbiased estimator of average precision. Furthermore, it can be seen that the measure can reliably be used to evaluate the quality of systems that were not used to create the initial samples, hence the measure is robust to evaluating the quality of unseen systems.

Figure 3 illustrates how xinfAP computed on a non-random judgment set compares with infAP computed on a random judgment set for various levels of incompleteness. In a similar manner to the experimental setup of the original infAP work, for each value of k, we generated ten different sample trials according to the procedure described in the previous paragraph, and for each one of the ten trials we computed the xinfAP for all systems. Then, all three statistics were computed for each one of the trials and the averages of these statistics over all ten trials were reported for different levels of judgment incompleteness. Using the same procedure, we also created ten different sample trials where the samples were generated by merely randomly sampling the judgment set and the infAP values were computed on them. For comparison purposes, to show how the original version of infAP behaves when this randomness assumption is violated, we also include infAP run on the same judgment set as extended infAP (marked as infAP depth+random judgments in the Figure).

It can be seen that for all levels of incompleteness, in terms of all three statistics, xinfAP is much more accurate in estimating average precision than the other two measures.

We further compared xinfAP to the sampling method proposed by Aslam et al. [1]. The robustness of xinfAP to incomplete relevance judgments is comparable to (and in some cases even better than) this method. (These results were omitted due to space limitations.)

[2]Throughout this paper, we assume that the complete judgment set corresponds to the depth-100 pool as the judgments we have are formed using depth-100 pools and assuming the remaining documents are nonrelevant.

4. ESTIMATION OF NDCG WITH INCOMPLETE JUDGMENTS

There are different versions of the nDCG metric depending on the discount factor and the gains associated with relevance grades, etc. In this paper, we adopt the version of nDCG in `trec_eval`.

Let $\Im$ denote a relevance grade and $gain(\Im)$ the gain associated with $\Im$. Also, let $g_1, g_2, \ldots g_Z$ be the gain values associated with the Z documents retrieved by a system in response to a query q, such as $g_i = gain(\Im)$ if the relevance grade of the document in rank i is $\Im$. Then, the nDCG value for this system can be computed as,

$$nDCG = \frac{DCG}{DCG_I} \quad \text{where} \quad DCG = \sum_{i=1}^{Z} g_i / \lg(i+1)$$

and DCG_I denotes the DCG value for an ideal ranked list for query q.

The estimation of nDCG with incomplete judgments can be divided into two parts: (1) Estimating DCG_I and (2) Estimating DCG. Then, the DCG and the DCG_I values can be replaced by their estimates to obtain the estimated nDCG value.[3]

4.1 Estimating DCG_I

The normalization factor, DCG_I, for a query q can be defined as the maximum possible DCG value over that query. Hence, the estimation of DCG_I can be derived in a two-step process: (1) For each relevance grade $\Im$ such as $gain(\Im) > 0$, estimate the number of documents with that relevance grade; (2) Calculate the DCG value of an optimal list by assuming that in an optimal list the estimated number of documents would be sorted (in descending order) by their relevance grades.

Using the sampling strategy described in the previous section, suppose incomplete relevance judgments were created by diving the complete pool into disjoint sets (strata) and randomly picking (sampling) documents from each stratum to be judged, possibly with different probability for each stratum.

[3]Note that this assumes that $E[nDCG] = E[DCG]/E[DCG_I]$, i.e., that DCG_I and DCG are independent of each other, which is not necessarily the case. This assumption may result in a small bias and better estimates of nDCG can be obtained by considering this dependence. However, for the sake of simplicity, throughout this paper, we will assume that these terms are independent.

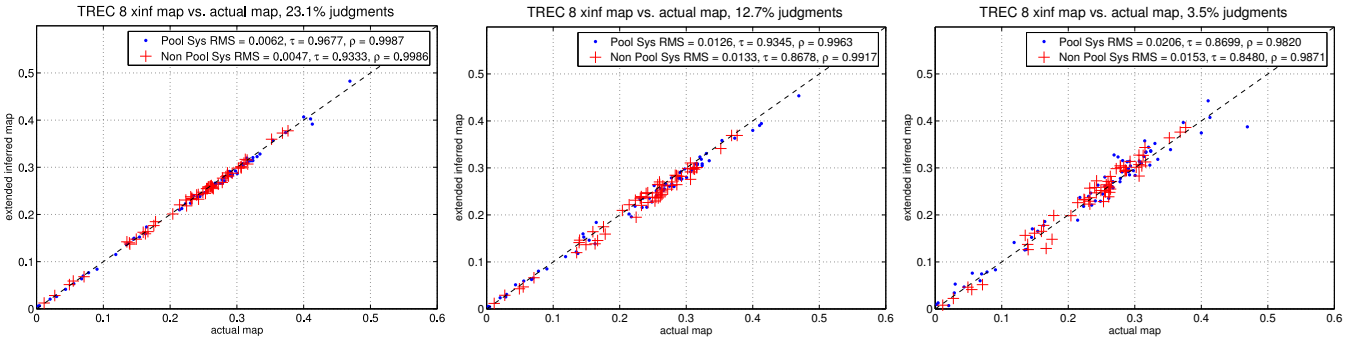

Figure 2: TREC-8 mean xinfAP when relevance judgments are generated according to depth-10, depth-5 and depth-1 pooling combined with equivalent number of randomly sampled judgments versus mean actual AP.

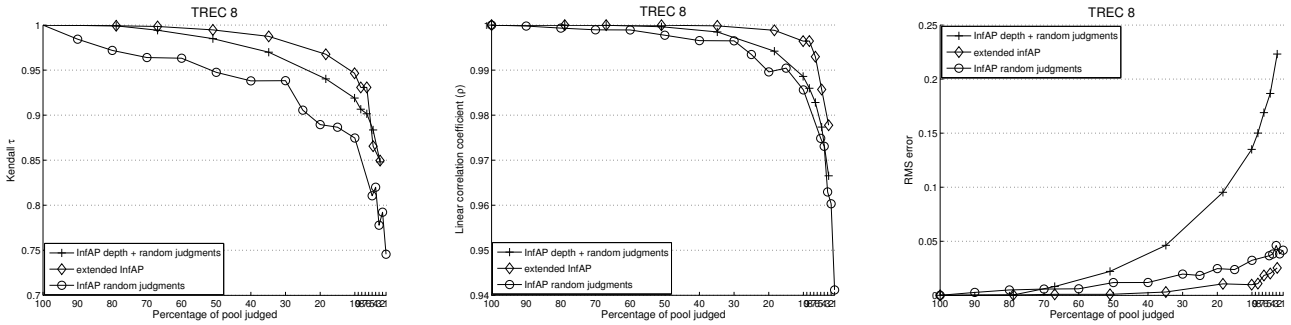

Figure 3: TREC-8 change in Kendall's τ, linear correlation coefficient (ρ), and RMS errors of xinfAP and infAP as the judgment sets are reduced when half of the judgments are generated according to depth pooling and the other half is a random subset of complete judgments and of inferred AP when the judgments are a random subset of complete judgments.

For each stratum s, let $r_s(\Im)$ be the number of sampled documents with relevance grade $\Im$, let n_s be the total number of documents sampled from strata s and N_s be the total number of documents that fall in strata s. Since the n_s documents are sampled uniformly from strata s, the estimated number of documents with relevance grade $\Im$ within this strata can be computed as

$$\hat{R}_s(\Im) = \frac{r_s(\Im)}{n_s} \cdot N$$

Then, the expected number of documents with relevance grade $\Im$ within the complete pool can be computed as

$$\hat{R}(\Im) = \sum_{\forall s} \hat{R}_s(\Im)$$

Once these estimates are obtained, one can estimate DCG_I.

4.2 Estimating DCG

Given Z documents retrieved by a search engine with relevance gain g_i for the document at rank i, for each rank i, define a new variable x_i such as $x_i = Z \cdot \frac{g_i}{lg(i+1)}$. Then, DCG can be written as the output of the following random experiment:

1. Pick a document at random from the output of the search engine, let the rank of this document be i.
2. Output the value of x_i.

It is easy to see that if we have the relevance judgments for all Z documents, the expected value of this random experiment is exactly equal to DCG.

Now consider estimating the outcome of this random experiment when relevance judgments are incomplete. Consider the first step of the random experiment, i.e. picking a document at random. Let Z_s be the number of documents in the output of a system that fall in stratum s. When picking a document at random, with probability Z_s/Z, we pick a document from stratum s.

Therefore, the expected value of the above random experiment can be written as:

$$E[DCG] = \sum_{\forall s} \frac{Z_s}{Z} \cdot E[x_i | \text{document at rank } i \in s]$$

Now consider the second step of the random experiment, computing the expected value of x_i given that the document at rank i falls in strata s. Let $sampled_s$ be the set of sampled documents from strata s and n_s be the number of documents sampled from this strata. Since documents within stratum s are uniformly sampled, the expected value of x_i can be computed as

$$E[x_i | \text{document at rank } i \in s] = \frac{1}{n_s} \sum_{\forall j \in sampled_s} x_j$$

Once E[DCG$_I$] and E[DCG] are computed, infNDCG can then be computed as infNDCG = $E[DCG]/E[DCG_I]$.

5. OVERALL RESULTS

Until now, we have shown that using a similar sampling strategy as the one used in TREC Terabyte 06 (complete

judgments divided into 2 different strata), xinfAP is highly accurate. In this section, we show that (1) this claim is consistent over different TRECs for both xinfAP and infNDCG and that (2) the two measures can be used with the complete judgments divided into more than two strata.

In order to check (2), we use a different sampling strategy than the one in Terabyte; we divide the complete judgment set (assuming depth-100 pool is the complete judgment set) into 4 different strata. The first stratum is the regular depth-k pool, fully judged. Instead of randomly sampling equal to the depth-k pool number of judgments from the remainder of the collection, we now divide the rest of the documents into three other strata and distribute the remaining judgments with a ratio of 3:1.5:1 (judge 55% of the documents in the top depth stratum, 27% of the documents in the middle depth stratum and 18% in the lowest depth stratum). This way, more weight is given to judging documents retrieved towards the top of the ranked lists of the search engines. Note, however, that as the number of strata increase, there values of the estimates may slightly deviate from the actual values since the effect of smoothing also increase (smoothing is needed for each stratum).

Figure 4 shows the quality of xinfAP and infNDCG (referred as extended infNDCG to avoid confusion) computed on these samples according to Kendall's τ and RMS Error statistics, for TRECs 8, 9 and 10. For comparison purposes, the plots also contain infAP and nDCG (the standard formula computed on random judgments, assuming unjudged documents are nonrelevant).

Looking at all plots, it can be seen that according to both statistics, using the same number of judgments, the extended infAP (xinfAP) and infNDCG consistently outperform infAP and nDCG on random judgments, respectively. The high RMS error of nDCG on random judgments is due to the fact that nDCG is computed on these judgments as it is, without aiming at estimating the value of the measure.

6. CONCLUSIONS

In this work, we extended inferred AP in two different ways. First, we derived confidence intervals for infAP to capture the variability in infAP values. Employing confidence intervals enables comparisons and eventually ranking of systems according to their quality measured by AP with high confidence. Second, we utilized a stratified random sampling strategy to select documents to be judged and extended infAP to handle the non-random samples of judgments. We applied the same methodology for estimating nDCG in the presence of incomplete non-random judgments. Stratified random sampling combines the effectiveness of stratification and thus better utilization of the relevance judgments with the simplicity of random sampling. We showed that xinfAP and infNDCG are more accurate than infAP and nDCG on equal number of random samples.

Note that the sampling strategy (i.e. the number of strata, the size of each stratum and the sampling percentage from each stratum) used here is rather arbitrary. The confidence intervals as described in the first part of this paper could be used as an objective function to determine an optimal sampling strategy. The sampling strategy is highly important for the quality of the estimates and identifying an optimal strategy is a point of future research.

Furthermore, confidence intervals as a function of the sample size could be used to determine the appropriate number of documents to be judged for an accurate MAP estimation which is a point we plan to investigate.

7. REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 541–548. ACM Press, August 2006.

[2] T. Bompada, C.-C. Chang, J. Chen, R. Kumar, and R. Shenoy. On the robustness of relevance measures with incomplete judgments. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 359–366, New York, NY, USA, 2007. ACM Press.

[3] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, 2004.

[4] S. Buttcher, C. Clarke, and I. Soboroff. The TREC 2006 terabyte track. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, 2006.

[5] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268–275, 2006.

[6] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, 1996. Morgan Kaufmann Publishers.

[7] C. Cleverdon. The cranfield tests on index language devices. *Readings in information retrieval*, pages 47–59, 1997.

[8] D. Harman. Overview of the third text REtreival conference (TREC-3). In D. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 1–19. U.S. Government Printing Office, Apr. 1995.

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

[10] W. Kraaij, P. Over, and A. Smeaton. TRECVID 2006 - an introduction. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

[11] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 375–382, New York, NY, USA, 2007. ACM.

[12] T. Sakai. Alternatives to bpref. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 71–78, New York, NY, USA, 2007. ACM Press.

[13] I. Soboroff. A comparison of pooled and sampled relevance judgments. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 785–786, New York, NY, USA, 2007. ACM Press.

[14] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management*. ACM Press, November 2006.

APPENDIX

Let s_d be a sample of cut-off levels at relevant documents. According to the Law of Total Variance, the variance in infAP can be calculated as,

$$var[\text{infAP}] = var[E[\text{infAP}|s_d]] + E[var[\text{infAP}|s_d]]$$

Let's consider the first term of the right-hand side of the above equation, which corresponds to the variance due to sampling cut-off levels.

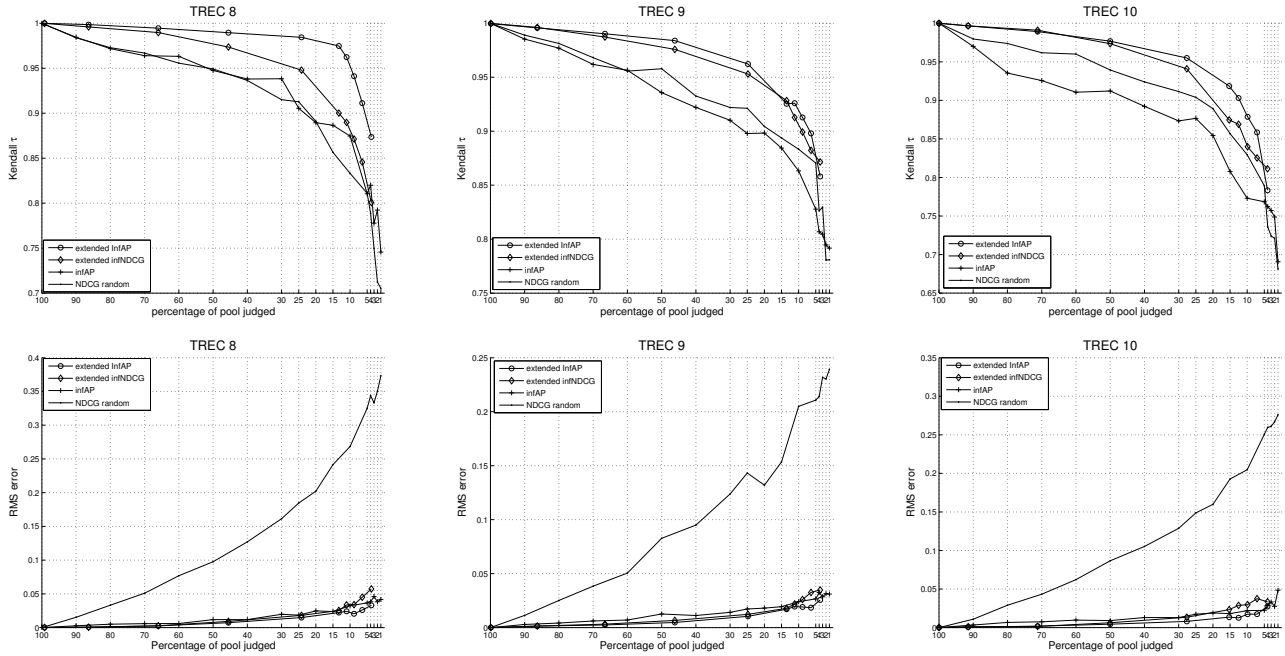

Figure 4: Comparison of extended inferred map, (extended) mean inferred ndcg, inferred map and mean ndcg on random judgments, using Kendall's τ (first row) and RMS error (last row) for TREC 8, 9 and 10.

Let r the number of relevant documents in s_d. Then, the conditional expectation of *infAP* is,

$$E[\text{infAP}|s_d] = \frac{1}{r}\sum_{k\in s_d} E[\widehat{PC}_k|s_d] = \frac{1}{r}\sum_{k\in s_d} PC_k$$

where $\widehat{PC}_k$ and PC_k denote the estimated and actual precision at cut-off k, respectively. Thus,

$$var[E[\text{infAP}|s_d]] = var\left[\frac{1}{r}\sum_{k\in s_d} PC_k\right] = (1-p)\,\frac{\sigma^2}{r}$$

where $p100\%$ is the sampling percentage of documents from the entire depth-100 pool and σ^2 is the actual variance among the precision values at all cut-off's of relevant documents and it can be estimated by, $\left(\sum_{k\in s_d}(\widehat{PC}_k - \text{infAP})^2\right)/(r-1)$.

Now, let's consider the second term of the right-hand side of the equation deduced by the Law of Total Variance, that is the variance due to sampling documents above a cut-off level in order to estimate the precision at that cut-off level,

$$var[\text{infAP}|s_d] = var\left[\frac{1}{r}\sum_{k\in s_d}\widehat{PC}_k\right] = \frac{1}{r^2}var\left[\sum_{k\in s_d}\widehat{PC}_k\right]$$

Considering $\widehat{PC}_k$ independent from each other

If we consider precisions at different cut-off levels independent from each other the variance of infAP for a given set of sampled cut-off levels depends on the summation of the precision variances at each individual cut-off level,

$$var[\text{infAP}|s_d] = \frac{1}{r^2}\sum_{k\in s_d} var[\widehat{PC}_k|s_d]$$

The precision at cut-off 1 is always 1 and therefore the variance is 0. Moreover, the precision at relevant documents not in the retrieved list is always assumed to be 0 and therefore, the variance at those cut-off levels is also 0. In all other case $\widehat{PC}_k$ is calculated as, $\widehat{PC}_k = 1/k + ((k-1)/k)\cdot\widehat{PC}_{\text{above k}}$ and therefore,

$$var[\widehat{PC}_k|s_d] = \left(\frac{k-1}{k}\right)^2 var[\widehat{PC}_{\text{above k}}]$$

Let r_{k-1} and n_{k-1} be the number of relevant documents and total number of documents sampled above cut-off k, respectively and let $|\text{d100}|_{k-1}$ be the number of documents in the depth-100 pool above cut-off k. The precision above cut-off k is estimated by [4], $\widehat{PC}_{k-1} = \frac{|\text{d100}|_{k-1}}{k-1}\cdot\frac{r_{k-1}}{n_{k-1}}$, which follows a hypergeometric distribution and its variance can be calculated as,

$$var[\widehat{PC}_{k-1}|s_d] = \left(\frac{p(1-p)}{n_{k-1}}\right)\cdot\left(1 - \frac{n_{k-1}-1}{|\text{d100}|_{k-1}-1}\right)$$

By considering the expected value of $var[\text{infAP}|s_d]$ over all samples of cut-off levels we get,

$$E[var[\text{infAP}|s_d]] = \frac{\sum_{k\in s_d} var[\widehat{PC}_k|s_d]}{r^2}$$

Considering $\widehat{PC}_k$ dependent to each other

If we do not consider precisions at different cut-off levels independent from each other the covariance between precisions can be calculated as,

$$cov[\widehat{PC}_k, \widehat{PC}_m] = \frac{k}{m}var[\widehat{PC}_k] \quad \text{where} \quad k < m$$

[4]For simplicity reasons we ignore the effect of smoothing that is introduced in the formula of infAP. Smoothing was considered in all experiments ran and it was observed that the effect of smoothing in variance is negligible.