Alternatives to Bpref

Tetsuya Sakai
NewsWatch, Inc. (current affiliation) / Toshiba Corporate R&D Center
sakai@newswatch.co.jp

ABSTRACT

Recently, a number of TREC tracks have adopted a retrieval effectiveness metric called *bpref* which has been designed for evaluation environments with *incomplete* relevance data. A graded-relevance version of this metric called *rpref* has also been proposed. However, we show that the application of Q-measure, normalised Discounted Cumulative Gain (nDCG) or Average Precision (AveP) to *condensed lists*, obtained by filtering out all unjudged documents from the original ranked lists, is actually a better solution to the incompleteness problem than bpref. Furthermore, we show that the use of graded relevance boosts the robustness of IR evaluation to incompleteness and therefore that Q-measure and nDCG based on condensed lists are the best choices. To this end, we use four graded-relevance test collections from NTCIR to compare ten different IR metrics in terms of system ranking stability and pairwise discriminative power.

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

Test Collection, Evaluation Metrics, Graded Relevance

1. INTRODUCTION

Information Retrieval (IR) evaluation using *incomplete* relevance data is beginning to receive attention. Large-scale test collections constructed through *pooling* [4, 5, 16, 17], such as the TREC, CLEF and NTCIR collections, are all incomplete to some degree, in that only a small sample of the document collection has been judged for relevance for each topic. While the collection sizes tend to grow monotonically in order to mimic real-world data such as the Web, the available manpower for relevance assessments often remain more or less constant, and therefore IR researchers are expected to live with the incompleteness issue as long they adhere to the *Cranfield paradigm* [3, 13].

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

At SIGIR '04, Buckley and Voorhees [3] proposed an IR evaluation metric called *bpref* (binary preference) which is highly correlated with AveP (Average Precision) when full relevance assessments are available and is yet more robust when the relevance assessments are reduced. Recent TREC tracks have started using this metric along with AveP. Bpref penalises a system if it ranks a *judged nonrelevant* document above a judged relevant one, and is indepedendent of how the *unjudged* documents are retrieved. At the SIGIR '06 poster session, De Beer and Moens [6] proposed *rpref*, which is a graded-relevance extension of bpref. More recently at CIKM '06, Yilmaz and Aslam [16] proposed three new methods that may replace bpref, which we shall discuss in Section 2.

This paper shows that the application of Q-measure [9, 10], normalised Discounted Cumulative Gain (nDCG) [7] or AveP to *condensed lists*, obtained by filtering out all unjudged documents from the original ranked lists, is actually a better solution to the incompleteness problem than bpref. Furthermore, we show that the use of graded relevance boosts the robustness of IR evaluation to incompleteness and therefore that Q-measure and nDCG based on condensed lists are the best choices. To this end, we use four graded-relevance test collections from NTCIR to compare ten different IR metrics in terms of system ranking stability and pairwise *discriminative power* [9].

Section 2 discusses some work that are related to this study. Section 3 provides the original definitions of bpref and rpref. Section 4 redefines these metrics based on a *condensed list* of documents, which leads to some simple alternatives to bpref. Section 5 describes our NTCIR data. Section 6 compares the metrics in terms of the entire system ranking using *Kendall's rank correlation* [3, 10, 16], and Section 7 compares them in terms of discriminative power using Sakai's *bootstrap sensitivity method* [9]. Finally, Section 8 concludes this paper.

2. RELATED WORK

Several researchers have tackled the problem of reducing human effort required for relevance assessments: SIGIR '98 saw new methods for creating judgment pools efficiently [5, 17]; At SIGIR '01, Soboroff [12] proposed a method for ranking systems without any relevance assessments, but the method tends to rank them by "popularity" rather than performance [1]. More recently at SIGIR '06, Carterette, Allan and Sitaraman [4] analyzed the distribution of AveP over all possible assignments of relevance to all unjudged documents and proposed a method to construct a test collection with

minimal relevance assessments; Coincidently, Aslam, Pavlu and Yilmaz [2] proposed a method for obtaining unbiased estimates of standard metrics such as AveP based on random sampling, while Soboroff [13] used bpref to examine how "decay" of Web documents affect system evaluation.

Among existing studies, the CIKM '06 paper by Yilmaz and Aslam [16] is probably the most relevant to the present work, since they also proposed their "alternatives to bpref", namely, *Induced*, *Subcollection* and *Inferred* versions of AveP. Induced AveP is exactly what we call "AveP′", which we derived independently as shown in Section 4. The other two metrics are more complex and less widely applicable in that they require knowledge of *all* pooled documents, including *pooled but unjudged* ones. (Subcollection AveP requires even more knowledge [16].) While the aim of Yilmaz and Aslam was to estimate the "actual" AveP values accurately, the present study considers a wider range of widely applicable IR metrics, including those based on graded relevance, for the purpose of ranking systems reliably in an incomplete relevance environment. Our experiments include both AveP′ (i.e., Induced AveP) and the actual AveP. Another contribution of this paper is that we evaluate the metrics in terms of *discriminative power* [9] given incomplete relevance data.

3. BPREF AND RPREF: ORIGINAL DEFINITIONS

We begin with the original definitions of bpref and rpref, although we have replaced some of the symbols in order to maintain notational consistency throughout the paper.

3.1 Bpref

Bpref is an IR metric based on binary relevance, designed to evaluate systems using *judged* documents only. Let R denote the number of judged relevant documents, and let N denote that of judged nonrelevant documents. Let D denote a relevant retrieved document, and let $\bar{D}_R$ denote a member of the first R judged nonrelevant documents as retrieved by the system. Then, bpref in the literature [13] is known as:

$$bpref = \frac{1}{R}\sum_{D}(1 - \frac{|\bar{D}_R \text{ ranked higher than D}|}{\min(R,N)}) \quad (1)$$

where $|\bar{D}_R$ ranked higher than D| represents a *penalty* based on binary relevance. In words, the penalty is the number of judged nonrelevant documents ranked above a retrieved relevant document. To ensure that the penalty does not exceed one, it is normalised by $\min(R,N)$: Note that, by definition, both $|\bar{D}_R| \leq R$ and $|\bar{D}_R| \leq N$ hold.

An early version of bpref, known as `old_bpref` in Chris Buckley's `trec_eval` version 8.1, is "buggy". We shall come back to this in Section 4 when we redefine bpref.

3.2 Rpref

To handle graded relevance, De Beer and Moens [6] have defined two versions of rpref, which we shall refer to as rpref_N and rpref_relative, respectively. As we shall see later, rpref_N is a generalisation of one of bpref variants called `bpref_allnonrel` in `trec_eval` version 8.1.

Let C denote a collection containing documents $D_k (1 \leq k \leq |C|)$. Let r_k denote the rank of document D_k in a system output for a topic. (For documents not retrieved, assume that $r_k = \infty$.) Let $J \subseteq C$ denote the set of *judged* documents, so that $|J| = R + N$ according to the notations we used earlier. For each $D_k \in J$, let $\phi_k \in [0,1]$ denote its *judged relevance* with respect to the topic in question, where 0 represents judged nonrelevance and 1 represents the highest possible relevance. Then, rpref_N is defined as:

$$rpref\_N = \frac{1}{R'}\sum_{D_k \in J,\ \phi_k > 0} \phi_k (1 - \frac{penalty_k}{N'}) \quad (2)$$

where

$$R' = \sum_{D_k \in J} \phi_k \quad (3)$$

$$N' = \sum_{D_k \in J} (1 - \phi_k) \quad (4)$$

and

$$penalty_k = \sum_{D_l \in J,\ r_l < r_k,\ \phi_l < \phi_k} \frac{\phi_k - \phi_l}{\phi_k} \ . \quad (5)$$

We now interpret the above equations in words. R' is the sum of all judged relevance values, which reduces to R in a binary relevance environment (i.e., if $\phi_k \in \{0,1\}$). From Eqs. 3 and 4, it is clear that

$$R' + N' = \sum_{D_k \in J} = |J| = R + N \quad (6)$$

and that N' reduces to N in a binary relevance environment. Meanwhile, instead of the binary-relevance penalty of bpref, rpref uses $penalty_k$ (Eq. 5), an analogous graded-relevance penalty: For a given retrieved relevant document D_k, we examine all judged documents ranked above it ($D_l \in J$, $r_l < r_k$), and find ones that are *less relevant* than D_k ($\phi_l < \phi_k$). For each of such illegitimately highly ranked documents D_l, we accumulate the difference between the two relevance values ($\phi_k - \phi_l$) normalised by ϕ_k, the highest possible relevance difference between D_k and any document that is less relevant. The penalty is then normalised by N': It is easy to show that $penalty_k \leq N'$. Finally, rpref_N takes the *weighted* average of $1 - penalty_k/N'$, using the relevance values ϕ_k as weights. That is, misplaced documents above a *highly* relevant document are emphasised compared to those above a *partially* relevant one. Note that, in a binary relevance environment, the weight ϕ_k in Eq. 2 reduces to a flag indicating either relevance or nonrelevance.

De Beer and Moens [6] note that N' is not a tight upper-bound of $penalty_k$, especially for highly ranked documents, since the number of documents ranked higher than D_k must be smaller than the rank of D_k, i.e., r_k. Thus, using N' for normalisation implies relatively small penalties for misplacements above highly ranked relevant documents. To put more emphasis on the higher ranks, De Beer and Moens [6] prefer "relative normalisation":

$$rpref\_relative = \frac{1}{R'}\sum_{D_k \in J,\ \phi_k > 0,\ r_k \neq 1} \phi_k (1 - \frac{penalty_k}{N_k}) \quad (7)$$

where

$$N_k = |\{D_l \in J | r_l < r_k\}| \ . \quad (8)$$

That is, N_k is the number of judged documents ranked above D_k. Note that $N_k = 0$ if $r_k = 1$, hence the "$r_k \neq 1$" condition is necessary.

4. ALTERNATIVE DEFINITIONS AND ALTERNATIVE METRICS

Section 3 showed that bpref, rpref_N and rpref_relative rely on *judged* documents only, and that they do not explicitly depend on the absolute document ranks. This section begins by redefining these preference-based metrics using a *condensed list*, obtained by removing all unjudged documents from the original system output. This suggests that "standard" metrics based explicitly on ranks may in fact be applied successfully to IR evaluation given incomplete relevance data.

4.1 Alternative Definitions of Bpref and Rpref

Suppose we need to evaluate a ranked list that contains (say) 1000 documents. The document at Rank r may or may not be a judged one. Since bpref relies on judged documents only, we can safely *remove all unjudged documents* from the list before computing bpref. As a result, we obtain a *condensed list* that contains judged documents only. Let $r'(\leq r)$ denote the new rank of a document in the condensed list. Then, based on these ranks, bpref can be rewritten as:

$$bpref = \frac{1}{R}\sum_{r'} isrel(r')(1 - \frac{\min(R, r' - count(r'))}{\min(R, N)}) \quad (9)$$

where $isrel(r')$ is one if the document at Rank r' is relevant and zero otherwise; and $count(r')$ is the number of relevant documents in top r' of the ranked list. Hence $r' - count(r')$ is the number of judged nonrelevant documents ranked above the document at Rank r'. It is easy to see that the two definitions (Eqs. 1 and 9) are equivalent. (The "buggy" bref, or `old_bpref`, uses $\min(R, N_{ret})$ instead of $\min(R, N)$ in Eq. 9, where N_{ret} is the number of *retrieved* judged nonorelevant documents. See the file `bpref_bug` in `trec_eval` version 8.1.)

Eq. 9 implies that *bpref is in fact two different metrics*: For any topic such that $R \leq N$, bpref reduces to:

$$bpref\_R = \frac{1}{R}\sum_{r'} isrel(r')(1 - \frac{\min(R, r' - count(r'))}{R}) \ . \quad (10)$$

Whereas, for any topic such that $R \geq N$, bpref reduces to:

$$bpref\_N = \frac{1}{R}\sum_{r'} isrel(r')(1 - \frac{r' - count(r')}{N}) \quad (11)$$

The latter is also known as `bref_allnonrel`. Note that $R \geq N \geq r' - count(r')$ holds in the latter case and therefore the minimum operator is not necessary.

Similarly, let us now rewrite rpref_N (Eq. 2), which is in fact a graded-relevance version of bpref_N shown above. We begin by adopting the concept of *cumulative gain* as proposed by Järvelin and Kekäläinen [7]. Let $\mathcal{L}$ denote a *relevance level*, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving an $\mathcal{L}$-relevant document. Without loss of generality, this paper assumes that we have S-relevant (highly relevant), A-relevant (relevant) and B-relevant (partially relevant) documents as in NTCIR [8] in addition to judged nonrelevant documents. Also, let $R(\mathcal{L})$ denote the number of $\mathcal{L}$-relevant documents. Let $cg(r) = \sum_{1 \leq i \leq r} g(i)$ denote the *cumulative gain* at Rank r of the system output, where $g(i) = gain(\mathcal{L})$ if the document at Rank i is $\mathcal{L}$-relevant and $g(i) = 0$ otherwise (i.e., if the document at Rank i is either judged nonrelevant or unjudged). It is easy to see that rpref's judged relevance values $\phi_k \in [0, 1]$ (See Section 3.2) can be obtained by dividing the raw gain values by $gain(\mathcal{H})$, where $\mathcal{H}$ is the highest relevance level across all topics of the test collection, e.g., $\mathcal{H} = S$.

Although bpref and rpref do not require the notion of *ideal ranked output* [7] for their definitions, we introduce it here in order to clarify how the preference-based metrics are related to cumulative-gain-based ones: An ideal ranked output for a given topic with $R = \sum_{\mathcal{L}} R(\mathcal{L})$ relevant documents is one such that $g(r) > 0$ for $1 \leq r \leq R$ and $g(r) \leq g(r-1)$ for $r > 1$. Thus, for NTCIR, listing up all S-relevant documents, followed by all A-relevant documents, followed by all B-relevant documents produces an ideal ranked output. Note that an ideal ranked output is unaffected by the removal of unjudged documents. Let $g_I(r)$ and $cg_I(r)$ denote the (cumulative) gain of an ideal ranked output. Then:

$$cg_I(R) = \sum_{1 \leq r \leq R} g_I(r) = \sum_{\mathcal{L}} R(\mathcal{L}) gain(\mathcal{L}) \ . \quad (12)$$

That is, the ideal cumulative gain at Rank R is the sum of all available gain values for the topic. Hence, From Eqs. 3 and 12,

$$R' = cg_I(R)/gain(\mathcal{H}) \ . \quad (13)$$

Moreover, from Eqs. 6 and 13,

$$N' = R + N - cg_I(R)/gain(\mathcal{H}) \ . \quad (14)$$

We can also rewrite rpref's penalty (Eq. 5) based on the notion of gain for a condensed list. The penalty based on a relevant document at Rank r' can be expressed as:

$$penalty(r') = \sum_{i' < r',\ g(i') < g(r')} \frac{g(r') - g(i')}{g(r')} \quad (15)$$

since the $gain(\mathcal{H})$'s for relevance value transformation cancel out. Therefore, rpref_N can be rewritten as:

$$rpref\_N = \frac{gain(\mathcal{H})}{cg_I(R)} \sum_{r',\ g(r')>0} \frac{g(r')}{gain(\mathcal{H})}(1 - \frac{penalty(r')}{R + N - \frac{cg_I(R)}{gain(\mathcal{H})}})$$

$$= \frac{1}{cg_I(R)} \sum_{r',\ g(r')>0} g(r')(1 - \frac{penalty(r')}{R + N - \frac{cg_I(R)}{gain(\mathcal{H})}}) \ . \quad (16)$$

Similarly, since the number of judged documents ranked above a document at Rank r' in a condensed list is exactly $r' - 1$, rpref_relative (Eq. 7) can be rewritten as:

$$rpref\_relative = \frac{1}{cg_I(R)} \sum_{r' \neq 1,\ g(r')>0} g(r')(1 - \frac{penalty(r')}{r' - 1}) \ . \quad (17)$$

De Beer and Moens [6] also mention its binary-relevance variant, which we call bpref_relative and formalise as:

$$bpref\_relative = \frac{1}{R}\sum_{r' \neq 1} isrel(r')(1 - \frac{r' - count(r')}{r' - 1}) \ . \quad (18)$$

Note that, in a binary relevance environment, Eq. 15 reduces to $\sum_{i' < r',\ g(i')=0}(1 - 0)/1 = \sum_{i' < r',\ g(i')=0}$, or the number of judged nonrelevant documents above Rank r', i.e., $r' - count(r')$.

Thus we have successfully redefined bpref and rpref variants based on the ranks of a condensed list, which contains judged documents only.

4.2 Alternative Metrics

From our rank-based notations of Eqs. 17 and 18, it is now clear that both rpref_relative and bpref_relative have flaws. Firstly, these relative metrics ignore a relevant document at Rank 1 [6], so a system that returns only one relevant document at Rank 1 receives zero as its score. Moreover, a system that returns only one relevant document at Rank 2 also receives zero, since the penalty and the normalisation factor $r'-1$ cancel out when $r'=2$. Thus these metrics are clearly counterintuitive. Another problem is that the two metrics do not average well across topics, as the score for an ideal ranked output varies depending on the value of R: It is easy to show that rpref_relative for an ideal ranked output equals $(cg_I(R)-g_I(1))/cg_I(R)$, not one. A corollary to this is that bpref_relative for an ideal ranked output equals $(R-1)/R$: An ideal ranked list for a topic with $R=2$ relevant documents would receive 0.5 as its score, while one for a topic with $R=5$ relevant documents would receive 0.8. Clearly, more well-designed metrics are in order.

A very simple solution to all of the above problems is to use r' instead of $r'-1$ for relative normalisation. Thus,

$$rpref\_relative2 = \frac{1}{cg_I(R)} \sum_{r',\ g(r')>0} g(r')(1 - \frac{penalty(r')}{r'}) \ . \quad (19)$$

Note that now we do examine a relevant document at Rank 1. It is also easy to show that rpref_relative2 equals one for an ideal ranked output for any topic.

Now, let us modify bpref_relative similarly:

$$bpref\_relative2 = \frac{1}{R} \sum_{r'} isrel(r')(1 - \frac{r' - count(r')}{r'})$$

$$= \frac{1}{R} \sum_{r'} isrel(r') \frac{count(r')}{r'} \ . \quad (20)$$

But this is exactly the well-known noninterpolated Average Precision (AveP) based on a condensed list in place of a raw list containing both judged and unjudged documents.

We can summarise our discussions so far as follows:

- bpref is actually two metrics: bpref_R (if $R \leq N$) and bpref_N (if $R \geq N$).
- bpref_R, bpref_N and rpref_N use absolute normalisation [6], so the misplacement penalties based on highly ranked relevant documents are small;
- bpref_relative and rpref_relative use relative normalisation to emphasize misplacement penalties based on highly ranked relevant documents, but have flaws;
- bpref_relative2 and rpref_relative2 are free from all of these flaws, and more importantly, *bpref_relative2 is exactly AveP based on a condensed list.* (We shall therefore refer to bpref_relative2 as AveP$'$ hereafter; It is also known as Induced AveP [16].)

In short, *the only essential difference between bpref and AveP$'$ is that while the former uses absolute normalisation, the latter uses relative normalisation*, as noted also in [16].

The above fact poses this question: *For evaluating IR systems with incomplete relevance information, is it really necessary to introduce bpref or its variants? Why can't we just use existing metrics based explicitly on ranks instead, after filtering out all unjudged documents from the original ranked list?* Thus, let us consider, in addition to AveP, Q-measure [9, 10] and nDCG [7], both of which explicitly rely on document ranks, and can handle *graded* relevance.

Q-measure is defined as:

$$Q\text{-}measure = \frac{1}{R} \sum_r isrel(r) \frac{\beta cg(r) + count(r)}{\beta cg_I(r) + r} \quad (21)$$

where β is a parameter for controlling the penalty on late arrival of relevant documents. It is very highly correlated with AveP (Note that $\beta = 0$ reduces Q-measure to AveP) and its discriminative power (See Section 7) is known to be at least as high as that of AveP. Since Sakai [11] has shown Q-measure's robustness to the choice of β, we let $\beta = 1$.

Instead of the raw gains $g(r)$, nDCG uses *discounted* gains $dg(r) = g(r)/\log_a(r)$ for $r > a$ and $dg(r) = g(r)$ for $r \leq a$. Let $dg_I(r)$ denote a discounted gain for an ideal ranked list. Then, nDCG at document cut-off l is defined as:

$$nDCG_l = \sum_{1 \leq r \leq l} dg(r) / \sum_{1 \leq r \leq l} dg_I(r) \ . \quad (22)$$

Throughout this paper, we let $l = 1000$ as it is known that small document cut-offs hurt nDCG$'$s stability [10]. Moreover, we let $a = 2$ because it is known that using a large logarithm base makes nDCG counterintuitive and insensitive [11]. nDCG is more forgiving for low-recall topics than AveP and Q-measure, as it does not depend on R directly.

Just like AveP$'$ (i.e., bpref_relative2), we can apply Q-measure and nDCG to condensed lists, i.e., after removal of all unjudged documents. We shall refer to these metrics as Q$'$ and nDCG$'$. The remainder of this paper studies how these metrics compare to bpref and rpref variants. Since Q-measure and nDCG are robust to the choice of gain values [10], we let $gain(S) = 3, gain(A) = 2, gain(B) = 1$ throughout this paper.

5. FULL AND REDUCED DATA

We use four data sets (i.e., test collections and submitted runs) from the NTCIR CLIR task [8]: Table 1 shows some statistics. We use the top 30 runs from each data set as measured by (actual) Mean AveP for our analyses: Under this condition, Kendall's rank correlation between two system rankings, representing two IR metrics, is statistically significant at $\alpha = 0.01$ if it exceeds 0.34 (two-sided normal test) [9]. The table also shows that N is always larger than R for the NTCIR collections: For example, for NTCIR-3C, the smallest N across all topics is 911, while the largest R across all topics is only 264. Thus, since $R < N$ always holds, *bpref is equivalent to bpref_R for our data.* Hence we shall refer to bpref as bpref_R in our experiments.

To examine the effect of incompleteness on the entire system ranking and discriminative power, we created *reduced relevance data* from the full relevance data, following the original Buckley/Voorhees methodology [3]: First, for each topic, we created a randomised list of judged relevant documents of size R, and a separate randomised list of judged nonrelevant documents of size N. Then, for each *reduction rate* $j \in \{90, 70, 50, 30, 10\}$, we created a reduced set of relevance data by taking the first R_j and N_j documents from the two lists, respectively, where $R_j = \max(1, truncate(R * j/100))$ and $N_j = \max(10, truncate(N * j/100))$. The contants 1 and 10 have been copied from [3], representing the

Table 1: The NTCIR Data: $N > R$ holds for all topics in all four data sets.

		range		average over topic set					runs
	topics	N	R	N	R	$R(S)$	$R(A)$	$R(B)$	used
NTCIR-3 Chinese ("NTCIR-3C")	42	[911, 2564]	[4, 264]	1432.6	78.2	21.0	24.9	32.3	30
NTCIR-3 Japanese ("NTCIR-3J")	42	[707, 2862]	[7, 354]	1676.6	60.4	7.9	31.5	21.0	30
NTCIR-5 Chinese ("NTCIR-5C")	50	[788, 2114]	[6, 249]	1836.3	61.0	7.0	30.7	23.3	30
NTCIR-5 Japanese ("NTCIR-5J")	47	[930, 2280]	[6, 345]	1768.9	89.1	3.2	41.8	44.2	30

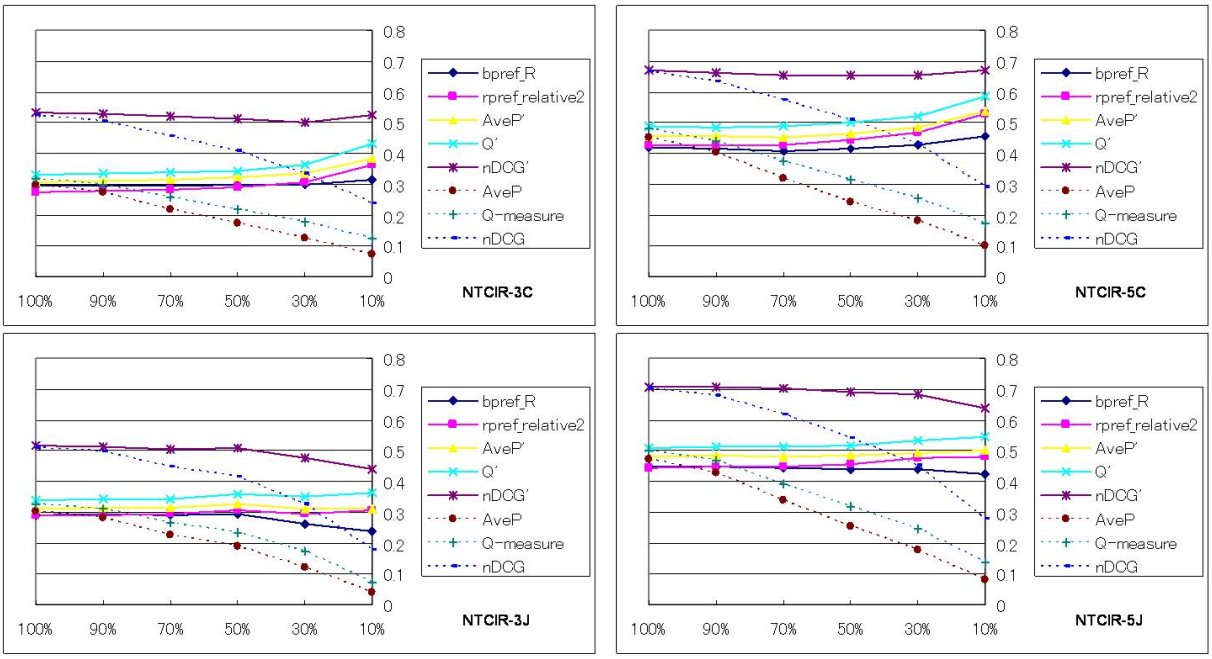

Figure 1: Reduction rate (x axis) vs. performance averaged over topics and runs (y axis).

minimum number of judged (non)relevant documents required for a topic. (In practice, the constant 10 was seldom used since N was generally very large.) This stratified sampling is essentially equivalent to random sampling from the entire set of judged documents [16].

Figure 1 shows the effect of relevance data reduction on the absolute overall performances (e.g., Mean AveP) averaged across all 30 runs for each data set. The horizontal axis represents the reduction rate j. It is clear that the "actual" AveP, Q-measure, nDCG values quickly diminish as the relevance data becomes more and more incomplete (as represented by the dotted lines), while the bpref_R (i.e., bpref) curve is relatively flat. This much supports a finding in [3]. However, it is also clear that *the other metrics (Q', $nDCG'$, rpref_relative2 and AveP',) do just as well as bpref in terms of the absolute value stability.* This further encouraged us to test whether *applying a standard metric to condensed lists works at least as well as bpref.*

6. EFFECT OF INCOMPLETENESS ON RANK CORRELATION

This section uses Kendall's rank correlation [3, 10, 16] to study the effect of incompleteness on the entire system ranking. Table 2 shows the correlation values between system rankings by two different metrics for all four data sets, given full relevance data. Among the ten metrics we consider, recall that AveP, Q-measure and nDCG are based on the original ranked lists, while AveP′, Q′ and nDCG′ are based on judged documents only just like the bpref variants. While all of the correlation values are *statistically* highly significant (See Section 5), we indicate values higher than 0.8 in bold just for convenience. Let (M_1,M_2) denote a correlation value between two rankings by metrics M_1 and M_2. We can observe that bpref_N and rpref_N are not highly correlated with "standard" metrics such as AveP: For example, (bpref_N, AveP)= .480 for NTCIR-5J. (They are not highly correlated with bpref_R either.) This reflects the fact that bpref_N and rpref_N give small penalties for misplacements above highly ranked relevant documents: Since N is generally very large for the NTCIR data, absolute normalisation virtually ignores the penalties for small values of r'. As we shall see in Section 7, bpref_N and rpref_N are also poor in terms of discriminative power. Hence we shall omit these two metrics in our reduced relevance data experiments.

As also noted in [16], the correlation between bpref_R (i.e., bpref) and the actual AveP is consistently lower than that between AveP′ and the actual AveP. For example, for NTCIR-5J, (bpref_R, AveP)= .885, while (AveP′, AveP)= .963. That is, *given the full relevance data, AveP′ is a better approximation of the actual AveP than bpref is.* Similarly, the correlations (Q′, Q-measure) and (nDCG′, nDCG) are also very high.

Figure 2 shows the effect of relevance data reduction on the system ranking for each metric. For example, the AveP curve represents the values of (AveP, AveP(j)), where AveP(j) denotes AveP computed based on the j% relevance data ($j \in \{90, 70, 50, 30, 10\}$). It can be observed that only the dotted line of the actual AveP stands apart from the others. More specifically, our observations are as follows:

Table 2: Kendall's rank correlation using full relevance data from NTCIR.

NTCIR-3C	bpref_N	rpref_N	rpref_relative2	AveP′	Q′	nDCG′	AveP	Q-measure	nDCG
bpref_R	.756	.747	**.936**	**.926**	**.931**	**.839**	**.931**	**.945**	**.839**
bpref_N	-	**.963**	.738	.720	.779	.798	.733	.756	.789
rpref_N	-	-	.729	.710	.770	**.816**	.724	.747	**.807**
rpref_relative2	-	-	-	**.945**	**.959**	**.867**	**.959**	**.972**	**.867**
AveP′	-	-	-	-	**.931**	**.821**	**.986**	**.954**	**.821**
Q′	-	-	-	-	-	**.871**	**.945**	**.968**	**.862**
nDCG′	-	-	-	-	-	-	**.834**	**.857**	**.991**
AveP	-	-	-	-	-	-	-	**.968**	**.834**
Q-measure	-	-	-	-	-	-	-	-	**.857**
NTCIR-3J	bpref_N	rpref_N	rpref_relative2	AveP′	Q′	nDCG′	AveP	Q-measure	nDCG
bpref_R	**.867**	**.862**	**.949**	**.977**	**.936**	**.885**	**.986**	**.940**	**.890**
bpref_N	-	**.986**	**.834**	**.853**	**.876**	**.880**	**.862**	**.862**	**.885**
rpref_N	-	-	**.839**	**.857**	**.880**	**.894**	**.867**	**.867**	**.899**
rpref_relative2	-	-	-	.963	.949	.890	.954	.963	.894
AveP′	-	-	-	-	**.940**	**.871**	**.991**	**.945**	**.876**
Q′	-	-	-	-	-	**.922**	**.949**	**.986**	**.926**
nDCG′	-	-	-	-	-	-	**.880**	**.908**	**.995**
AveP	-	-	-	-	-	-	-	**.954**	**.885**
Q-measure	-	-	-	-	-	-	-	-	**.913**
NTCIR-5C	bpref_N	rpref_N	rpref_relative2	AveP′	Q′	nDCG′	AveP	Q-measure	nDCG
bpref_R	.531	.508	**.839**	**.913**	**.811**	.770	**.871**	**.821**	.775
bpref_N	-	**.949**	.508	.526	.572	.586	.531	.563	.591
rpref_N	-	-	.494	.513	.559	.572	.517	.549	.577
rpref_relative2	-	-	-	**.899**	**.899**	**.867**	**.876**	**.908**	**.871**
AveP′	-	-	-	-	**.890**	**.802**	**.949**	**.899**	**.807**
Q′	-	-	-	-	-	**.857**	**.903**	**.972**	**.862**
nDCG′	-	-	-	-	-	-	.798	**.857**	**.995**
AveP	-	-	-	-	-	-	-	**.922**	**.802**
Q-measure	-	-	-	-	-	-	-	-	**.862**
NTCIR-5J	bpref_N	rpref_N	rpref_relative2	AveP′	Q′	nDCG′	AveP	Q-measure	nDCG
bpref_R	.476	.499	**.876**	**.903**	**.811**	.733	**.885**	**.821**	.738
bpref_N	-	**.949**	.416	.490	.536	.595	.480	.517	.591
rpref_N	-	-	.448	.503	.559	.628	.494	.540	.623
rpref_relative2	-	-	-	**.899**	**.853**	.766	**.890**	**.871**	.770
AveP′	-	-	-	-	**.862**	.766	**.963**	**.880**	.770
Q′	-	-	-	-	-	**.821**	**.862**	**.972**	**.825**
nDCG′	-	-	-	-	-	-	.756	**.811**	**995**
AveP	-	-	-	-	-	-	-	**.890**	.761
Q-measure	-	-	-	-	-	-	-	-	**.816**

(i) The rankings by Q′, nDCG′, rpref_relative2 and AveP′ are at least as robust to incompleteness as the ranking by bpref_R.

(ii) The rankings by the actual Q-measure and nDCG are more robust to incompleteness than the ranking by the actual AveP; In fact, they do as well as bpref_R.

For example, for NTCIR-5J, (Q′, Q′(10))=0.66, (bpref_R, bpref_R(10))=0.42 and (Q-measure, Q-measure(10))=0.41. Whereas, (AveP,AveP(10))=0.21 which is not even statistically significant.

Observation (i) supports our hypothesis that *applying a standard metric to condensed lists works at least as well as bpref.* Observation (ii) suggests that *the best graded-relevance metrics are more robust to incompleteness than the binary AveP.* Our explanation for this is as follows: Due to the distribution of $R(\mathcal{X})$ shown in Table 1, randomised relevance data reduction is expected to remove more partially relevant documents than highly relevant documents. Whereas, reflecting the same original distribution, most runs actually contain many more partially relevant documents than highly relevant ones. Hence, highly relevant documents contained within the runs are relatively unaffected by the reduction, while partially relevant documents contained within the runs may become "nonrelevant" due to the reduction. Therefore, graded-relevance metrics, which depend more on highly relevant documents than on partially relevant ones, demonstrate a higher robustness to incompleteness.

7. EFFECT OF INCOMPLETENESS ON DISCRIMINATIVE POWER

This section compares the IR metrics in terms of discriminative power using the paired test version of Sakai's *bootstrap sensitivity method* [9], which is known to agree very well with the more ad hoc Voorhees/Buckley swap method [15]. The input to this method are a test collection, a set of runs, an IR metric, and the significance level α for bootstrap hypothesis tests. Using resampled topic sets, the method conducts a bootstrap hypothesis test for every system pair, and computes the discriminative power, i.e., for how many system pairs the IR metric was able to detect a significant difference, and the overall performance difference required to achieve that significance.

Table 3 ranks the IR metrics according to discriminative power at $\alpha = 0.05$ given full relevance data. For example, for NTCIR-5C, the actual Q-measure detects a significant difference at $\alpha = 0.05$ for 174 out of 435 (30*29/2) system pairs (40%), and the estimated overall performance difference required to detect one given 50 topics is 0.11 [9]. Metrics based on judged documents only are shown in bold. It is clear that, given full relevance data, bpref_R is substantially more discriminative than bpref_N and rpref_N, but *all other metrics are more discriminative than bpref.* In particular, note that *Q′, nDCG′, rpref_relative2 and AveP′ all outperform bpref consistently, using judged documents only.* In addition, the superiority of Q′ over bpref_R, AveP′ and rpref_relative2 is consistent across all data sets.

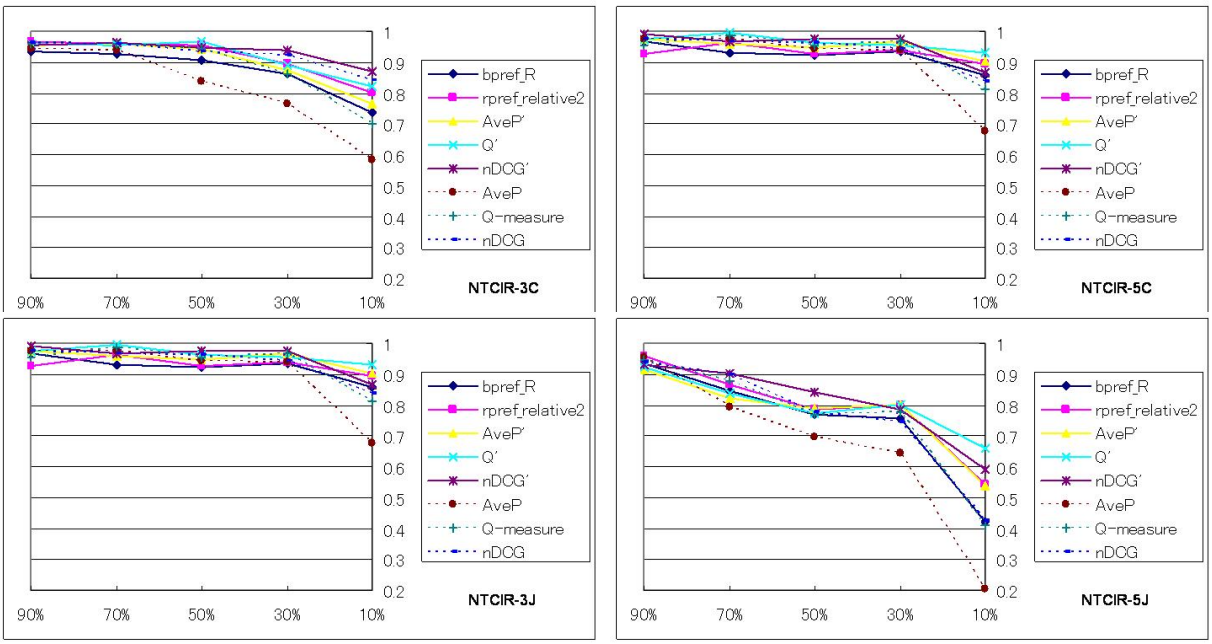

Figure 2: Reduction rate (x axis) vs. Kendall's rank correlation (y axis).

Table 3: Discriminative power at $\alpha = 0.05$ using full relevance data from NTCIR.

NTCIR-3C	disc. power	diff. required	NTCIR-5C	disc. power	diff. required
Q-measure	242/435=55.6%	0.10	Q-measure	174/435=40.0%	0.11
AveP	240/435=55.2%	0.11	**Q′**	**167/435=38.4%**	**0.11**
Q′	**238/435=54.7%**	**0.10**	nDCG	163/435=37.5%	0.10
nDCG′	**236/435=54.3%**	**0.11**	**rpref_relative2**	**163/435=37.5%**	**0.10**
rpref_relative2	**236/435=54.3%**	**0.10**	AveP	159/435=36.6%	0.11
nDCG	235/435=54.0%	0.13	**AveP′**	**158/435=36.3%**	**0.11**
AveP′	**234/435=53.8%**	**0.12**	**nDCG′**	**156/435=35.9%**	**0.10**
bpref_R	**232/435=51.5%**	**0.12**	**bpref_R**	**132/435=30.3%**	**0.10**
rpref_N	**203/435=46.7%**	**0.13**	**bpref_N**	**109/435=25.1%**	**0.11**
bpref_N	**198/435=45.6%**	**0.14**	**rpref_N**	**105/435=24.1%**	**0.12**
NTCIR-3J	disc. power	diff. required	NTCIR-5J	disc. power	diff. required
nDCG′	**317/435=72.9%**	**0.12**	Q-measure	136/435=31.3%	0.09
nDCG	316/435=72.6%	0.14	**Q′**	**131/435=30.1%**	**0.09**
Q′	**308/435=70.8%**	**0.11**	nDCG	120/435=27.6%	0.13
Q-measure	305/435=70.1%	0.13	**nDCG′**	**119/435=27.4%**	**0.12**
AveP	298/435=68.5%	0.11	**rpref_relative2**	**115/435=26.4%**	**0.10**
rpref_relative2	**298/435=68.5%**	**0.11**	AveP	113/435=26.0%	0.10
AveP′	**296/435=68.0%**	**0.11**	**AveP′**	**113/435=26.0%**	**0.10**
bpref_R	**283/435=65.1%**	**0.11**	**bpref_R**	**100/435=23.0%**	**0.10**
bpref_N	**272/435=62.5%**	**0.16**	**bpref_N**	**68/435=15.6%**	**0.11**
rpref_N	**272/435=62.5%**	**0.16**	**rpref_N**	**56/435=12.9%**	**0.12**

Figure 3 shows the effect of relevance data reduction ($j \in \{70, 50, 30, 10\}$) on discriminative power at $\alpha = 0.05$. For example, for NTCIR-5C with only 10% relevance data, the actual AveP detects a significant difference for only about 2.5% of all run pairs. (Note that its discriminative power is 36.6% with 100% relevance data, as shown in Table 3.) Whereas, under the same condition, the actual Q-measure and nDCG are comparable to bpref (a little over 10% in discriminative power); Q′, nDCG′, rpref_relative2 and AveP′ are the most discriminative (over 20%). Although the actual Q-measure does not do well with NTCIR-3J at $j = 10$, and nDCG(′) appears to do worse at $j = 50$ than at $j = 30$ with NTCIR-5C and 5J, the trends consistent across all four data sets are as follows:

(i′) Q′, nDCG′, rpref_relative2 and AveP′ are more discriminative than bpref_R in a very incomplete relevance environment; Q′ and nDCG′ are probably the most discriminative.

(ii′) Given incomplete relevance data, bpref_R is more discriminative than the actual AveP; but so are the actual Q-measure and nDCG.

Thus, the benefit of introducing bpref is not clear in terms of discriminative power either.

8. CONCLUSIONS

This paper showed that the application of Q-measure, nDCG or AveP to condensed lists, obtained by filtering out all unjudged documents from the original ranked lists, is actually a better solution to the incompleteness problem than bpref. Furthermore, we showed that the use of graded relevance boosts the robustness of IR evaluation to incompleteness and therefore that Q-measure and nDCG based on condensed lists are the best choices. In terms of the entire system ranking, Q-measure, nDCG and AveP based on condensed lists (Q′, nDCG′ and AveP′) are more robust to incompleteness than bpref_R (i.e., bpref); even the actual

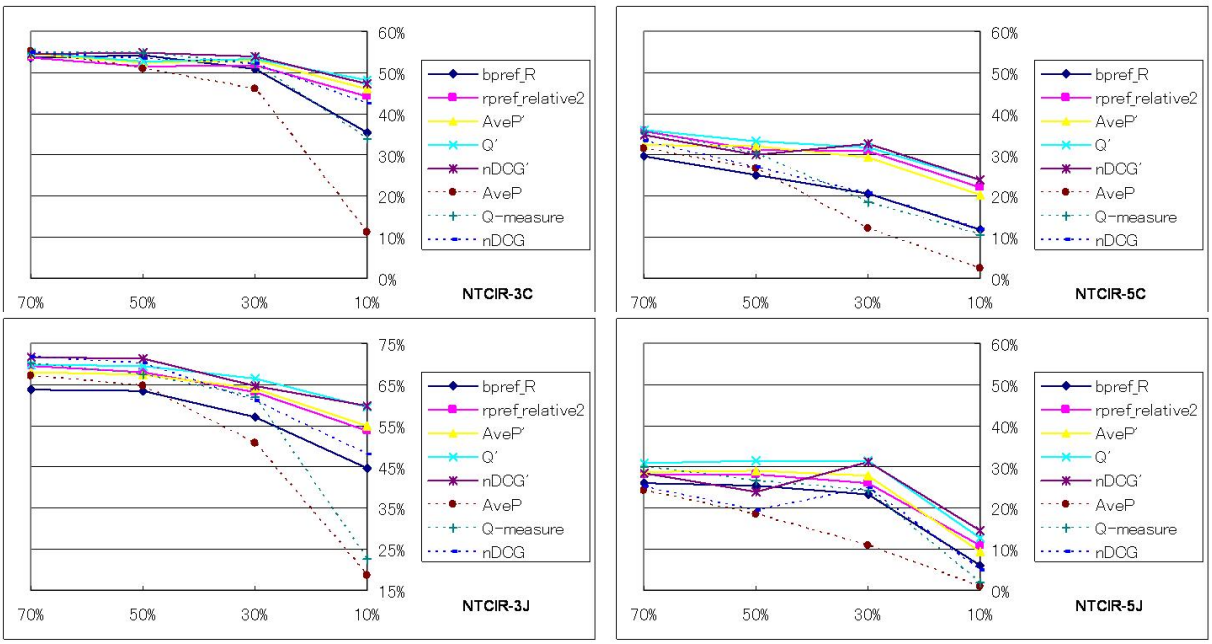

Figure 3: Reduction rate (x axis) vs. discriminative power at $\alpha = 0.05$ (y axis).

Q-measure and nDCG do as well as bpref_R. In terms of pairwise discriminative power, Q′, nDCG′ and AveP′ are more discriminative than bpref_R whether full or reduced relevance data is used; Q′ is slightly but consistently more discriminative than AveP′; even the actual Q-measure and nDCG can often be as discriminative as bpref_R in a very incomplete relevance environment. A variant of bpref/rpref (rpref_relative2) does well also, but the benefit of introducing this relatively complex metric is not clear either.

Finally, it should be noted that we have not examined IR metrics from the viewpoint of user satisfaction: It remains to be seen, for example, how recent criticisms of AveP [14] extend to different experimental environments and different metrics. Properties such as rank correlation with an established metric and discriminative power provide *necessary* conditions for a good IR metric, not *sufficient* conditions.

9. REFERENCES

[1] Aslam, J. A. and Savell, R.: On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments, *ACM SIGIR 2003 Proceedings*, pp. 361-362, 2003.

[2] Aslam, J. A., Pavlu, V. and Yilmaz, E.: A Statistical Method for System Evaluation Using Incomplete Judgments, *ACM SIGIR 2006 Proceedings*, pp. 541-548, 2006.

[3] Buckley, C. and Voorhees, E. M.: Retrieval Evaluation with Incomplete Information, *ACM SIGIR 2004 Proceedings*, pp. 25-32, 2004.

[4] Carterette, B., Allan, J. and Sitaraman, R.: Minimal Test Collections for Retrieval Evaluation, *ACM SIGIR 2006 Proceedings*, pp. 268-275, 2006.

[5] Cormack, G. V., Palmer, C. R. and Clarke, C. L. A. Efficient Construction of Large Test Collections, *ACM SIGIR '98 Proceedings*, pp. 282-289 (1998).

[6] De Beer, J. and Moens, M.-F.: Rpref - A Generalization of Bpref towards Graded Relevance Judgments, *ACM SIGIR 2006 Proceedings*, pp. 637-638, 2006.

[7] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp. 422-446, 2002.

[8] Kando, N.: Overview of the Fifth NTCIR Workshop, *NTCIR-5 Proceedings*, 2005.

[9] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *ACM SIGIR 2006 Proceedings*, pp. 525-532, 2006.

[10] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, 43(2), pp. 531-548, 2007.

[11] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proceedings of the First International Workshop on Evaluating Information Acess (EVIA 2007)*, to appear, 2007.

[12] Soboroff, I., Nicholas, C. and Cahan, P.: Ranking Retrieval Systems without Relevance Judgments, *ACM SIGIR 2001 Proceedings*, pp. 66-73, 2001.

[13] Soboroff, I.: Dynamic Test Collections: Measuring Search Effectiveness on the Live Web, *ACM SIGIR 2006 Proceedings*, pp. 276-283, 2006.

[14] Turpin, A. and Scholer, F.: User Performance versus Precision Measures for Simple Search Tasks, *ACM SIGIR 2006 Proceedings*, pp. 11-18, 2006.

[15] Voorhees, E. M. and Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error, *ACM SIGIR 2002 Proceedings*, pp. 316-323, 2002.

[16] Yilmaz, E. and Aslam, J. A.: Estimating Average Precision with Incomplete and Imperfect Judgments, *CIKM 2006 Proceedings*, 2006.

[17] Zobel, J.: How Reliable are the Results of Large-Scale Information Retrieval Experiments? *ACM SIGIR '98 Proceedings*, pp. 307-314, 1998.