IPSJ Journal Vol. 50 No. 4 1410–1420 (Apr. 2009)

Regular Paper

On the Robustness of Information Retrieval Metrics to Biased Relevance Assessments

Tetsuya Sakai[†1]

Information Retrieval (IR) test collections are growing larger, and relevance data constructed through *pooling* are suspected of becoming more and more *incomplete* and *biased*. Several studies have used IR evaluation metrics specifically designed to handle this problem, but most of them have only examined the metrics under *incomplete but unbiased* conditions, using random samples of the original relevance data. This paper examines nine metrics in more realistic settings, by reducing the number of pooled systems and the number of pooled documents. Even though previous studies have shown that metrics based on a *condensed list*, obtained by removing all unjudged documents from the original ranked list, are effective for handling very incomplete but unbiased relevance data, we show that these results do not hold when the relevance data are biased towards particular systems or towards the top of the pools. More specifically, we show that the condensed-list versions of *Average Precision*, *Q-measure* and *normalised Discounted Cumulative Gain*, which we denote as AP′, Q′ and nDCG′, are not necessarily superior to the original metrics for handling biases. Nevertheless, AP′ and Q′ *are* generally superior to *bpref*, *Rank-Biased Precision* and its condensed-list version even in the presence of biases.

1. Introduction

Information Retrieval (IR) test collections are growing larger, and relevance data constructed through *pooling* [3),16)] are suspected of becoming more and more *incomplete* and *biased* [2),5)]. Relevance data are incomplete if some relevant documents exist among the *unjudged* documents in the test collection. Furthermore, incomplete relevance data are biased if they represent some limited aspects of the complete set of relevant documents. For example, if the number of pooled systems is small, the resultant test collection may overestimate these systems and underestimate systems that did not contribute to the pool, since these new systems are likely to retrieve relevant documents that are outside the set of known relevant documents. We will refer to this phenomenon as *system bias* [15)]. Bias may also be caused by *shallow pools*: If only documents at the very top of submitted ranked lists are judged, the resultant relevance data may contain relevant documents that are very easy to retrieve, but not those that are difficult to retrieve. For example, Buckley, et al. [3)] report that the TREC 2005 HARD/Robust test collection is biased towards documents that contain topic title words due to shallow pools. We will refer to this phenomenon as *pool depth bias* [14)].

The objective of this paper is to examine the robustness of IR effectiveness metrics in the presence of system bias and pool depth bias, with an emphasis on metrics that can handle graded relevance. Several researchers have proposed evaluation metrics specifically for handling the incompleteness of relevance data, but most of them have only examined the metrics under *incomplete but unbiased* conditions, using random samples of the original relevance data [1),2),11),16),22)]. While random sampling may mimic a situation where the number of judged documents is extremely small compared to the entire document collection, it does not address the problems due to system bias and pool depth bias. Therefore, this paper examines metrics in more realistic settings, by reducing the number of pooled systems and the number of pooled documents.

The main contributions of this paper are as follows. First, we examine as many as nine metrics for handling system bias and pool depth bias in test collections. The metrics examined are: *Average Precision* (AP), *Q-measure* (Q) [13)], *normalised Discounted Cumulative Gain* (nDCG) [7)], *Rank-Biased Precision* (RBP) [10)], *binary preference* (bpref) [2)], AP′, Q′, nDCG′ and RBP′. The latter four metrics are AP, Q, nDCG and RBP applied to a *condensed list*, obtained by removing all unjudged documents from the original ranked list [11),16)]. Thus, just like bpref, these four metrics assume that retrieved unjudged documents are *nonexistent*, while traditional metrics assume that the unjudged documents are *nonrelevant*. Even though previous studies have shown that condensed-list metrics are effective for handling very incomplete but unbiased relevance data [11),16)], we show that they are not necessarily superior to the original metrics in the presence of system bias or pool depth bias. This discrepancy suggests that the results reported in previous studies that used random sampling should be interpreted with caution. Second, our extensive experiments cover two independent

[†1] NewsWatch, Inc.

evaluation efforts, TREC and NTCIR[9], and utilise their graded relevance data. This is in contrast to most existing studies that are limited to TREC data and binary-relevance metrics[1),2),22)]. Since our results are consistent across all of our data sets, we believe that our findings are general. Our main findings are:

(1) Condensed-list metrics overestimate systems that did not contribute to the pool while traditional metrics underestimate them, and the overestimation is larger than the underestimation.

(2) When runs from a single team or a few teams are used for forming the relevance data, AP′, Q′, nDCG′ are not necessarily superior to AP, Q and nDCG in terms of *discriminative power*, i.e., the overall ability to detect pairwise statistical significance[12].

(3) Similarly, when shallow pools are used for forming relevance data, AP′, Q′, nDCG′ are not necessarily superior to AP, Q and nDCG.

(4) Nevertheless, in the presence of system bias or pool depth bias, AP′ and Q′ are generally more discriminative than bpref, RBP and RBP′.

Finding (1) substantially generalises that by Büttcher, et al.[5], who analysed a TREC Terabyte data set and observed that "Where AP underestimates the performance of a [new] system, bpref overestimates it."

The remainder of this paper is organised as follows. Section 2 discusses previous work, and Section 3 formally defines the nine metrics considered in this study. Section 4 describes the graded-relevance data and runs from TREC and NTCIR which we use for comparing the metrics. Sections 5 and 6 examine the robustness of our metrics to system bias by reducing the number of runs used to form the relevance data. Section 7 examines the robustness of our metrics to pool depth bias by reducing the number of pooled documents. Finally, Section 8 concludes this paper.

2. Related Work

A decade ago, Zobel[23] examined the effect of pool depth and that of leaving out one run for forming the TREC relevance data. As TREC test collections at that time, i.e., TRECs 3-5, were based on binary relevance, he used binary-relevance metrics such as 11-point average precision. Subsequently, TREC adopted his leave-one-out methodology for validating their test collections, but chose to leave out one participating *team* at a time since each team usually contributes multiple runs to a pool[3]. The present study also includes leave-one-*team*-out experiments as well as "take-one-team" experiments which rely on runs from a single team to form the relevance data. Sanderson and Joho[18] have examined a "take-one-*run*" approach, but they considered AP only, using data from TRECs 5-8. The present study compares nine metrics, and our analysis covers recent TREC and NTCIR data.

Most existing studies that compared metrics for evaluation with incomplete data used random sampling from the original relevance data[1),2),11),16),22)]. For example, Yilmaz and Aslam[22] used this approach to evaluate their proposed metrics, including *Induced AP* which is exactly what we call AP′, and *Inferred AP* which aims to estimate the true value of AP. An exception is the work by Büttcher, et al.[5] which included leave-one-team-out experiments to address the system bias issue. Their experiments covered a condensed-list version of precision at document cut-off 20 and *RankEff*[1]. However, precision is an unreliable metric[13], and RankEff is in fact as unreliable as bpref by definition[15].

Among the studies that used random sampling, Sakai[11] compared condensed-list metrics such as AP′, Q′, nDCG′ and bpref along with traditional metrics, using data sets from NTCIR. Sakai and Kando[16] repeated the experiments using graded-relevance data from TREC and NTCIR, and added RBP to their candidate metrics; they did not examine RBP′. The study showed that, under very incomplete but unbiased conditions, AP′, Q′, nDCG′ are superior to AP, Q, nDCG, bpref and RBP. In contrast, the present study shows that AP′, Q′, nDCG′ are not necessarily superior to AP, Q, nDCG in the presence of system bias or pool depth bias.

3. Formal Definitions of Metrics

3.1 AP, Q, nDCG and RBP

Let R denote the number of judged relevant documents. For any given ranked list of documents, let $I(r)$ be 1 if the document at rank r is relevant and 0 otherwise. Let $C(r) = \sum_{i \leq r} I(i)$. Then AP is defined as:

$$AP = \frac{1}{R}\sum_r I(r)\frac{C(r)}{r}\ . \qquad (1)$$

Let $\mathcal{L}$ denote a relevance level, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving a judged $\mathcal{L}$-relevant document. We follow the NTCIR tradition and let $\mathcal{L} \in \{S, A, B\}$[9]. As for the TREC graded relevance data, we treat "highly relevant" documents as S-relevant and "relevant" documents as B-relevant. We let $gain(S) = 3$, $gain(A) = 2$ and $gain(B) = 1$ hereafter as Q and nDCG are robust to the choice of gain values[13].

Let $g(r) = gain(\mathcal{L})$ if the document at rank r is $\mathcal{L}$-relevant and $g(r) = 0$ otherwise, i.e., if the document at rank r is either judged nonrelevant or unjudged. The *cumulative gain* at rank r is given by $cg(r) = \sum_{1\leq i\leq r} g(i)$. Consider an *ideal* ranked list of documents, which satisfies $g(r) > 0$ for $1 \leq r \leq R$ and $g(r) \leq g(r-1)$ for $r > 1$. For NTCIR, listing up all S-, A- and B-relevant documents in this order produces an ideal ranked output. Let $cg_I(r)$ denote the cumulative gain of the ideal list. Q is defined as:

$$Q\text{-}measure = \frac{1}{R}\sum_r I(r)\frac{C(r) + \beta cg(r)}{r + \beta cg_I(r)} \qquad (2)$$

where β is a parameter for reflecting the persistence of the user. Clearly, $\beta = 0$ reduces Q to AP; we let $\beta = 1$ throughout this paper.

Sakai and Robertson[17] have recently discussed a user model for AP and Q-measure.

For a given logarithm base a, let the *discounted* gain at Rank r be $dg(r) = g(r)/\log_a(r)$ for $r > a$ and $dg(r) = g(r)$ for $r \leq a$. Similarly, let $dg_I(r)$ denote the discounted gain for an ideal ranked list. nDCG at document cut-off l is defined as:

$$nDCG_l = \sum_{1\leq r\leq l} dg(r) / \sum_{1\leq r\leq l} dg_I(r)\ . \qquad (3)$$

Throughout this paper, we let $l = 1000$ as it is known that small document cut-offs hurt the stability of nDCG[13]. This original definition of nDCG is "buggy" in that a relevant document retrieved at rank 1 and one retrieved at rank a receive the same credit. We adhere to the original nDCG but let $a = 2$ to alleviate the effect of the bug. Other versions of nDCG are described elsewhere[4),8)].

Let $\mathcal{H}$ denote the highest relevance level across all topics. In all of our experiments, $\mathcal{H} = S$. Let p be the persistence parameter that represents the fixed probability that the user moves from a document at rank r to rank $(r+1)$. RBP is defined as:

$$RBP = \frac{1-p}{gain(\mathcal{H})}\sum_r g(r)p^{r-1}\ . \qquad (4)$$

Moffat and Zobel[10] explored $p = 0.5, 0.8, 0.95$, and Sakai and Kando[16] showed that $p = 0.95$ is the best choice among these three values in terms of system ranking stability and discriminative power. Hence we use $p = 0.95$ thoughout this paper. RBP is different from the other metrics considered in this paper in that it totally disregards recall. Sakai and Kando[16] have pointed out some weaknesses of this metric.

3.2 Bpref and Other Condensed-List Metrics

Sakai[11] showed that a family of metrics, which are existing metrics applied to a *condensed list* of documents obtained by removing all unjudged documents from the original list, are simpler and better solutions than bpref. Bpref itself can be expressed as a metric based on a condensed list. Let r' denote the rank of a judged document in a condensed list, whose rank in the original list was $r(\geq r')$. Let N denote the number of judged nonrelevant documents. For any topic such that $R \leq N$, bpref reduces to *bpref_R*:

$$bpref\_R = \frac{1}{R}\sum_{r'} I(r')\left(1 - \frac{\min(R, r' - C(r'))}{R}\right)\ . \qquad (5)$$

In fact, $R \leq N$ holds for every topic used in our experiments, and therefore *bpref is always bpref_R.* Whereas, for any topic such that $R \geq N$, bpref reduces to *bpref_N*:

$$bpref\_N = \frac{1}{R}\sum_{r'} I(r')\left(1 - \frac{r' - C(r')}{N}\right)\ . \qquad (6)$$

The only essential difference between bpref and AP applied to a condensed list, which we call AP′, is that bpref lacks the top-heaviness property of AP. That is, bpref is more insensitive to change in top ranked documents than AP′[11),16)]. Note that, from Eq. (1), AP′ can be expressed as:

$$AP' = \frac{1}{R}\sum_{r'} I(r')\left(1-\frac{r'-C(r')}{r'}\right) . \qquad (7)$$

Condensed-list versions of Q, nDCG and RBP will be denoted by Q′, nDCG′ and RBP′. Thus this paper considers four metrics (AP, Q, nDCG and RBP) plus five condensed-list metrics (AP′, Q′, nDCG′, RBP′ and bpref). Among these, AP, AP′ and bpref cannot handle graded relevance.

4. Data

Table 1 provides some statistics of the TREC and NTCIR data we used for evaluating the nine metrics. The "TREC03" and "TREC04" data are from the TREC 2003 and 2004 robust track[19)], and the "NTCIR-6J" (Japanese) and "NTCIR-6C" (Chinese) data are from the NTCIR-6 CLIR task[9)]. The NTCIR-6J and NTCIR-6C data contain a few teams that did not contribute any monolingual runs, which we have excluded from our analysis. Hence we considered only ten teams for both NTCIR-6J and NTCIR-6C.

Consider a particular topic. Let t denote a participating team, and let D_t denote the set of documents contributed to the pool by this team. For TREC03, for example, D_t is the union of the top 125 documents of each run submitted by t. The set of *unique contributions* by t is defined as $U_t = D_t - \cup_{t' \neq t} D_{t'}$ [⋆1]. Similarly, let $D_t^{rel} (\subseteq D_t)$ denote the set of judged relevant documents obtained from t. The set of *unique relevant documents* from t is defined as $U_t^{rel} = D_t^{rel} - \cup_{t' \neq t} D_{t'}^{rel}$. **Table 2** shows the participating teams that we used, along with the number of runs submitted and average unique contributions/relevant documents. For example, Table 2 (c) shows that "NICT" contributed 229.7 documents per topic and 8 *relevant* documents per topic on average, and that this was achieved by

Table 1 TREC and NTCIR data used.

	TREC03	TREC04	NTCIR-6J	NTCIR-6C
#topics	50	49	50	50
#docs	approx. 528,000		858,400	901,446
pool depth	125	100	100	100
average N	925.5	654.6	1157.9	999.4
range N	[292, 2050]	[132, 1371]	[480, 2732]	[414, 1907]
average R	33.2	41.2	95.3	88.1
range R	[4, 115]	[3, 161]	[4, 311]	[15, 400]
S-relevant	8.1	12.5	2.5	21.6
A-relevant	—	—	61.1	30.4
B-relevant	25.0	28.8	31.7	36.1
#all runs	78	110	74	46
#teams	16	14	10(12)	10(11)

Table 2 Participating teams, #runs and #unique contributions per topic, and #unique relevant documents per topic. †Not used for take-one-team experiments; ∗Used for take-three-teams experiments (See Section 6).

(a) TREC03	#runs	$\lvert U_t \rvert$	$\lvert U_t^{rel} \rvert$	(b) TREC04	#runs	$\lvert U_t \rvert$	$\lvert U_t^{rel} \rvert$
MU03rob	5	47.1	0.28	Juru	10	15.3	0.16
NLPR03∗†	5	5.1	0.06	NLPR04	11	6.2	0.04
SABIR03	3	24.1	0.18	SABIR04	6	16.2	0.76
Sel	5	16.9	0.04	apl04rs	5	15.0	0.16
THUIRr030∗	5	9.0	0.14	fub04	10	8.4	0.24
UAmsT03R	5	31.0	0.16	humR04	10	23.2	0.33
UIUC03R∗	5	11.1	0.06	icl04pos	9	42.9	0.53
VT	5	26.6	0.34	mpi04r†	10	62.7	0.41
aplrob03	5	15.0	0.24	pircRB04∗	10	6.4	0.39
fub03I†	5	12.9	0.02	polyu	6	26.0	0.31
humR03	5	18.0	0.10	uic0401†	1	12.6	0.39
oce03	5	39.5	0.16	uogRob∗	10	6.2	0.35
pircRB	5	30.6	0.54	vtum	8	9.5	0.18
rutcor03†	5	103.6	0.20	wdo	4	16.7	0.16
uic030†	5	22.6	0.26				
uwmtCR	5	17.4	0.66				
(c) NTCIR-6J	#runs	$\lvert U_t \rvert$	$\lvert U_t^{rel} \rvert$	(d) NTCIR-6C	#runs	$\lvert U_t \rvert$	$\lvert U_t^{rel} \rvert$
BRKLY	8	64.8	1.6	BRKLY	8	166.8	3.56
HUM	5	120.6	1.04	CCNU∗	2	12.9	1.22
JSCCL∗	4	12.8	0.34	HUM	5	130.9	2.26
KLE	3	28.8	1.08	I2R∗	4	22.3	0.94
NCUTW†	5	54.4	1.44	ISQUT†	3	82.8	0.92
NICT	20	229.7	8.00	NCUTW	5	25.4	0.96
OKSAT	5	65.9	1.60	NTNU†	4	32.9	0.86
TSB†	12	37.5	0.74	UniNE∗	5	13.4	1.16
UniNE∗	5	14.3	0.56	WTG	4	66.6	1.22
YLMS∗	3	7.9	0.18	pircs	4	59.6	1.56

[⋆1] For the NTCIR data, $\{t'\}$ includes the teams that did not contribute any monolingual runs. Whereas, t represents a team that contributed at least one monolingual run.

submitting 20 runs.

Let J denote the complete set of judged documents for a topic. Section 5 reports on our leave-one-team-out experiments which replace J with $J - U_t$ for each t. That is, unique contributions from t are removed from the original relevance data, so that t can be treated as a "new" team. In Section 6, we go to the other extreme and replace J with D_t. That is, runs from a single team are used for forming the relevance data. In these "take-one-team" experiments, the teams labelled with a "†" in Table 2 failed to contribute a relevant document (i.e., $D_t^{rel} = \phi$) for at least one topic, and were therefore excluded from our analysis. In addition, we chose three teams from each data set to conduct "take-three-teams" experiments, by replacing J with $\cup_{t \in T} D_t$, where T is the set of chosen teams. As indicated by "$*$"'s in Table 2, we chose three "ordinary" teams: ones with the smallest number of unique contributions.

For our pool depth bias experiments which we shall report in Section 7, we formed "shallow pool" relevance data by taking the top $pd \in \{50, 10, 1\}$ documents from every run for each data set.

To examine the effect of system bias and pool depth bias in terms of discriminative power and *system ranking stability* as measured by *Kendall's rank correlation* between two rankings based on two different relevance data sets [21], we randomly selected one monolingual run per team for each data set. For example, for the NTCIR-6J data, we randomly selected ten monolingual runs, each representing a team.

5. System Bias: Leave One Team Out

Table 3 shows, for each t from NTCIR-6J, how a selected monolingual run from t is affected when the original relevance data J is replaced by $J - U_t$. For

Table 3 Performance change and rank change when a run is evaluated using that team's leave-one-team-out relevance data (NTCIR-6J). A "+" indicates that a run is overestimated; a "−" indicates that it is underestimated. Rank changes are indicated in bold: For example, "**6↑5**" means going up from rank 6 to rank 5. Note that the rows represent different relevance data.

	AP′	Q′	nDCG′	RBP′	bpref	AP	Q	nDCG	RBP
BRKLY	+1.93% 4→4	+1.49% 4→4	+0.42% 5→5	+1.12% 4→4	+2.39% 4→4	−0.61% 7→7	−0.51% 7→7	−0.27% **6↓8**	−0.34% 4→4
HUM	+2.04% 8→8	+1.76% 8→8	+0.59% **6↑5**	+0.63% 8→8	+2.34% 8→8	−0.06% 9→9	+0.03% 8→8	+0.05% 5→5	−0.16% 8→8
JSCCL	+0.85% 6→6	+0.62% 6→6	+0.21% 4→4	+0.49% 5→5	+1.28% **6↑4**	−0.14% 5→5	−0.11% 5→5	−0.08% 4→4	−0.04% 5→5
KLE	+1.03% 7→7	+0.93% **7↑6**	+0.37% 7→7	+0.38% 6→6	+1.13% 7→7	0.00% 6→6	0.00% **6↓7**	−0.03% **7↓8**	−0.42% 6→6
NCUTW	+1.64% **9↑8**	+1.34% 9→9	+0.54% 9→9	+1.88% 9→9	+2.14% **9↑8**	−0.40% **8↓9**	−0.38% 9→9	−0.24% 9→9	−0.39% 9→9
NICT	+3.76% **5↑4**	+3.37% **5↑4**	+1.33% 8→8	+1.05% 7→7	+3.60% **5↑4**	+1.43% **4↓5**	+1.41% **4↓5**	+0.65% 8→8	−0.08% 7→7
OKSAT	+6.13% 10→10	+4.83% 10→10	+1.81% 10→10	+4.04% 10→10	+6.48% 10→10	−0.50% 10→10	−0.30% 10→10	−0.07% 10→10	−0.47% 10→10
TSB	+0.73% 1→1	+0.59% 1→1	+0.26% 1→1	+0.35% 1→1	+0.68% 1→1	0.00% 1→1	−0.04% 1→1	+0.01% 1→1	−0.06% 1→1
UniNE	+1.01% 3→3	+0.77% 3→3	+0.32% 2→2	+0.73% 3→3	+1.43% 3→3	+0.08% 3→3	+0.10% 3→3	+0.02% 2→2	−0.03% 3→3
YLMS	+0.12% 2→2	+0.07% 2→2	0.00% 3→3	+0.07% 2→2	+0.19% 2→2	−0.08% 2→2	−0.07% 2→2	−0.08% 3→3	0.00% 2→2

example, when a run from "BRKLY" is evaluated using nDCG with this team's leave-one-team-out relevance data, the run's score goes down by 0.27% (from .5895 to .5879), and its rank among the 10 selected runs goes down from rank 6 to rank 8. In contrast, when a run from "HUM" is evaluated using nDCG′ with this team's leave-one-team-out relevance data, the run's score goes up by 0.59% (from .5980 to .6015), and its rank a goes up from rank 6 to rank 5. It can be observed that, according to condensed-list metrics, i.e., AP′, Q′, nDCG′, RBP′ and bpref, the scores and the ranks tend to go up with the use of each leave-one-team-out relevance data, while, according to traditional metrics, i.e., AP, Q, nDCG and RBP, the scores and the ranks tend to go down. Moreover, the percentage increase of the condensed-list metrics tend to be higher than the percentage decrease of the traditional metrics. The trends are similar for TREC03, TREC04 and NTCIR-6C, but the tables are omitted due to space limitations. Hence, our first observation is that *condensed-list metrics overestimate new systems while traditional metrics underestimate them, and that the overestimation tends to be larger than the underestimation.* A new run contains many unjudged documents. Therefore, condensing its ranked list may move up the ranks of retrieved relevant documents dramatically. This is why condensed-list metrics, including bpref, overestimate new systems.

6. System Bias: Take One Team

The leave-one-team-out experiments replaced J with $J - U_t$. We now discuss a more extreme case of system bias, by replacing J with D_t, the contributions from a single team. As we have explained in Section 4, we also form relevance data using contributions from three teams with the smallest number of unique contributions.

Table 4 summarises our take-one-team results for NTCIR-6J in a way similar to Table 3. Thus, for each team t, the table shows how a particular monolingual run from t (the same run we used for the leave-one-team-out experiments) is affected when the original relevance data J is replaced by D_t. For example, when a run from "BRKLY" is evaluated using Q′ with this team's contributions only, the run goes down from rank 4 to rank 8. In contrast, when the same run is evaluated

Table 4 Performance change and rank change when a run is evaluated using that team's take-one-team relevance data (NTCIR-6J). A "+" indicates that a run is overestimated; a "−" indicates that it is underestimated. Rank changes are indicated in bold. Note that the rows represent different relevance data.

	AP′	Q′	nDCG′	RBP′	bpref	AP	Q	nDCG	RBP
BRKLY	+21.0% **4↓8**	+20.0% **4↓8**	+6.65% **5↓7**	0.00% **4↓8**	+18.6% **4↓8**	+21.7% **7↑4**	+19.6% **7↑4**	+6.40% **6↑4**	−0.11% 4→4
HUM	+29.6% 8→8	+28.8% 8→8	+10.8% **6↑5**	−0.04% 8→8	+24.8% **8↓9**	+31.8% **9↑4**	+30.0% **8↑4**	+11.0% **5↑4**	−0.12% **8↑5**
JSCCL	+28.6% **6↓7**	+26.9% **6↓7**	+10.2% 4→4	0.00% **5↓7**	+25.4% **6↓7**	+28.1% **5↑2**	+25.8% **5↑3**	+9.93% 4→4	−0.11% **5↑4**
KLE	+26.4% **7↓8**	+26.2% **7↓8**	+9.73% 7→7	0.00% **6↓8**	+21.6% **7↓8**	+25.5% **6↑3**	+24.5% **6↑4**	+9.30% **7↑4**	−0.15% **6↑4**
NICT	+9.07% **5↓8**	+8.97% **5↓8**	+3.24% 8→8	+0.04% **7↓8**	+7.51% **5↓8**	+6.97% 4→4	+6.79% 4→4	+2.65% **8↑4**	0.00% **7↑6**
OKSAT	+45.8% 10→10	+47.2% 10→10	+23.0% 10→10	0.00% 10→10	+36.2% 10→10	+47.3% **10↑7**	+47.1% **10↑8**	+23.2% **10↑9**	−0.05% **10↑6**
UniNE	+24.8% 3→3	+22.3% 3→3	+7.03% 2→2	0.00% 3→3	+22.9% 3→3	+24.9% **3↑2**	+21.9% **3↑2**	+6.85% 2→2	−0.10% **3↑2**
YLMS	+29.0% 2→2	+26.4% 2→2	+7.55% 3→3	−0.07% 2→2	+27.3% 2→2	+31.1% **2↑1**	+28.0% **2↑1**	+7.89% **3↑2**	−0.17% **2↑1**

Table 5 Kendall's rank correlation: the original ranking vs. take-one-team (average) / take-three-teams.

	AP′	Q′	nDCG′	RBP′	bpref	AP	Q	nDCG	RBP
TREC03									
take-three-teams	.950	.917	.900	.967	.933	.933	.933	.933	.967
take-one-team	.951	.918	.929	.958	.944	.932	.920	.935	.947
TREC04									
take-three-teams	.978	.956	.978	.934	.956	1	1	.956	1
take-one-team	.932	.936	.898	.903	.903	.906	.907	.860	.926
NTCIR-6J									
take-three-teams	.956	.911	.956	.867	.956	.822	.822	.956	.956
take-one-team	.876	.880	.925	.893	.894	.756	.782	.885	.849
NTCIR-6C									
take-three-teams	1	.956	1	.911	.956	1	1	1	1
take-one-team	.960	.929	.991	.880	.920	.853	.867	.898	.907

using Q with this team's contributions only, it goes up from rank 7 to rank 4. It can be observed that, *if a single team t is used for forming the relevance data, the run score for t goes up for all metrics (except for RBP and RBP′); however, while traditional metrics overestimate the rank of a run from t, condensed-list metrics understimate it.* Condensed-list metrics underestimate the rank of a run from t because all the other runs from $t'(\neq t)$ are substantially *overestimated*: These other runs are "new" to the take-one-team relevance data of t, and we have already observed in Section 5 that condensed-list metrics overestimate new runs. As for RBP and RBP′, replacing J with D_t does not substantially affect the run score for t, because this merely turns some relevant documents below the pool depth within that run, i.e., those that belong to $J - D_t$, into nonrelevant documents. The stability of scores for RBP and RBP′ reflects the fact that they totally disregard recall, and not necessarily that they are superior: Note that the ranks according to RBP and RBP′ are altered just like the other metrics. Similar results for TREC03, TREC04 and NTCIR-6C are omitted due to space limitations.

Table 5 compares, for each data set and metric, the ranking of the aforementioned selected runs based on the original relevance data and that based on a take-one-team / take-three-teams relevance data. The similarity between two rankings is quantified using Kendall's rank correlation, which would be 1 if the two rankings are identical and -1 if the two rankings are the exact inverse of each other. The rank correlation values for the take-one-team relevance data have been averaged across teams. It can be observed that the correlation values are generally very high. That is, it is possible to replace the original relevance data with one that is based on a single team (or three teams) and still maintain a similar system ranking. As mentioned in Section 2, this generalises a finding by Sanderson and Joho [18] who considered only AP and binary-relevance TREC data. However, obtaining a system ranking that is similar to the full relevance data is not sufficient for sound evaluation: We later show that strong system bias can introduce much noise in statistical significance tests.

Our main criterion for comparing metrics is Sakai's *discriminative power* [12]. Let C be the set of all run pairs that are being considered. For a given significance level α, let $C_*(\subseteq C)$ be the set of run pairs with a statistically significant performance difference in terms of a given metric according to a two-sided, *paired bootstrap hypothesis test* [6]. Then discriminative power is defined as $|C_*|/|C|$: It means how often a metric manages to detect a statistically significant difference for a given probability of *Type I Error*. Although $|C_*|/|C|$ can also be defined using a significance test other than the bootstrap test, one of the advantages of Sakai's method is that it can also estimate the minimum performance difference required to achieve statistical significance. More details can be found elsewhere [12].

Suppose that C_* was obtained using a given metric and the original relevance data. Now, let C'_* denote the set of pairs of runs with a statistically significant difference in terms of the same metric but with a *different* relevance data set. *Assuming* that the results based on the original relevance data are the ground truth, we can quantify the discrepancy between C_* and C'_* by reporting the number of *misses* $|C_* - C'_*|$ and that of *false alarms* $|C'_* - C_*|$.

Table 6 and **Table 7** summarise the results of our discriminative power experiments using $\alpha = 0.05$ with the take-one-team and take-three-teams relevance data. For example, given the TREC03 *full* relevance data, Q detects a statistically significant difference at $\alpha = 0.05$ for 80 run pairs out of 120 (66.7%), and this is the highest discriminative power achieved across all metrics, as indicated in bold. Moreover, given the 50 topics of TREC03, the performance difference required to reach significance is around 0.07 in Q. Whereas, the TREC03 *take-*

Table 6 Discriminative power at $\alpha = 0.05$: take one team / three teams (TREC). For each experimental condition, the highest discriminative power is indicated in bold.

AP′	Q′	nDCG′	RBP′	bpref	AP	Q	nDCG	RBP
TREC03 full relevance data: discriminative power and the estimated difference required								
77/120 =64.2% 0.09	77/120 =64.2% 0.07	71/120 =59.2% 0.08	55/120 =45.8% 0.04	69/120 =57.5% 0.08	77/120 =64.2% 0.07	80/120 **=66.7%** 0.07	71/120 =59.2% 0.08	55/120 =45.8% 0.04
TREC03 take-three-teams: discriminative power, misses and false alarms								
61.7% 5 2	62.5% 6 4	55.0% 8 3	42.5% 8 4	55.8% 3 1	67.5% 2 6	**68.3%** 2 4	63.3% 2 7	49.2% 1 5
TREC03 take-one-team: discriminative power, misses and false alarms (averaged)								
59.6% 8.42 2.92	59.0% 9.50 3.33	54.4% 8.92 3.17	43.1% 9.25 6.00	51.7% 8.83 1.83	66.4% 5.67 8.33	**67.6%** 5.50 6.58	61.6% 4.67 7.58	52.6% 3.17 11.25
TREC04 full relevance data: discriminative power and the estimated difference required								
61/91 =67.0% 0.07	62/91 =68.1% 0.08	58/91 =63.7% 0.09	46/91 =50.5% 0.05	57/9 =62.6% 0.09	61/91 =67.0% 0.07	63/91 **=69.2%** 0.08	58/91 =63.7% 0.08	45/91 =49.5% 0.05
TREC04 take-three-teams: discriminative power, misses and false alarms								
63.7% 3 0	65.9% 2 0	56.0% 7 0	40.7% 10 1	54.9% 8 1	69.2% 0 2	**70.3%** 0 1	61.5% 2 0	48.4% 1 0
TREC04 take-one-team: discriminative power, misses and false alarms (averaged)								
61.6% 7.92 3.00	62.3% 7.00 1.75	53.5% 11.50 2.25	45.9% 9.00 4.75	56.4% 9.67 4.00	64.4% 7.17 4.83	**67.0%** 7.00 5.00	59.7% 7.50 3.83	50.1% 3.75 4.33

Table 7 Discriminative power at $\alpha = 0.05$: take one team / three teams (NTCIR). For each experimental condition, the highest discriminative power is indicated in bold.

AP′	Q′	nDCG′	RBP′	bpref	AP	Q	nDCG	RBP
NTCIR-6J full relevance data: discriminative power and the estimated difference required								
25/45 =55.6% 0.07	28/45 =62.2% 0.09	33/45 **=73.3%** 0.08	26/45 =57.8% 0.04	23/45 =51.1% 0.08	26/45 =57.8% 0.08	28/45 =62.2% 0.07	33/45 **=73.3%** 0.08	26/45 =57.8% 0.05
NTCIR-6J take-three-teams: discriminative power, misses and false alarms								
57.8% 1 2	64.4 1 2	**71.1%** 1 0	42.2% 7 0	44.4% 3 0	66.7% 2 6	68.9% 1 4	**71.1%** 1 0	62.2% 0 2
NTCIR-6J take-one-team: discriminative power, misses and false alarms (averaged)								
61.9% 1.00 3.88	66.7% 1.13 3.13	66.1% 3.38 0.13	49.2% 5.13 1.25	50.6% 3.00 2.75	66.4% 4.13 8.00	67.2% 4.25 6.50	**67.8%** 4.25 1.75	61.4% 3.50 5.13
NTCIR-6C full relevance data: discriminative power and the estimated difference required								
36/45 =80.0% 0.07	34/45 =75.6% 0.06	34/45 =75.6% 0.07	32/45 =71.1% 0.07	34/45 =75.6% 0.07	37/45 **=82.2%** 0.08	36/45 =80.0% 0.07	34/45 =75.6% 0.08	32/45 =71.1% 0.06
NTCIR-6C take-three-teams: discriminative power, misses and false alarms								
71.1% 4 0	73.3% 1 0	75.6% 0 0	66.7% 2 0	64.4% 5 0	**82.2%** 0 0	80.0% 0 0	77.8% 0 1	77.8% 0 3
NTCIR-6C take-one-team: discriminative power, misses and false alarms (averaged)								
73.9% 3.00 0.25	75.0% 1.50 1.25	71.1% 2.13 0.13	72.5% 2.25 2.88	70.3% 4.25 1.88	**80.3%** 3.63 2.75	79.7% 3.38 3.25	75.0% 2.63 2.38	75.0% 2.00 3.75

three-teams relevance data superficially raises the discriminative power of Q to 68.3%, but this is due to 2 misses and 4 false alarms. The corresponding discriminative power of Q, averaged across the 12 take-one-team relevance data, is 67.6%.

According to these tables, take-one-team relevance data generally yield more misses and false alarms than take-three-teams relevance data. Hence we observe that, even though take-one-team relevance data may produce a system ranking that is very similar to that produced by the original relevance data, pooling runs from several teams is better than pooling runs from a single team for obtaining reliable conclusions based on statistical significance tests. The focus of this study, however, is on the comparison of different metrics under the same condition, and not on how many and what kind of teams are required to obtain reliable conclusions.

The tables also show that AP, Q and nDCG are generally more discriminative than AP′, Q′ and nDCG′, respectively, even with take-one-team or take-three-teams relevance data. For example, for TREC03, the discriminative power of Q averaged over 12 take-one-team relevance data is 67.6% while the corresponding value for Q′ is only 59.0%, even though the number of misses and that of false alarms are more or less comparable. Thus, *condensed-list metrics are not necessarily superior to traditional metrics when the relevance data are heavily biased towards one team or a few teams.* On the other hand, even with take-three-teams and take-one-team relevance data, it can be observed that AP′ and Q′ are generally more discriminative than bpref, RBP and RBP′.

7. Pool Depth Bias

We finally examine the effect of pool depth bias on IR metrics. As mentioned in Section 4, we examined pool depths 50,10 and 1. **Table 8** shows the Kendall's rank correlation values between the original ranking and a shallow-pool ranking. For example, when the original relevance data of TREC03 is replaced by that formed with a pool depth of 1, the rank correlation between the original system ranking and the new system ranking both according to AP′ is .717. It can be observed that reducing the pool depth does hurt the system ranking, but the differences across metrics are not clear from this table.

Table 9 and **Table 10** compare the robustness of metrics to pool depth bias in terms of discriminative power. For example, using the pool-depth-50 relevance data of TREC03, Q is the most discriminative among the nine metrics, and its discriminative power is 66.7%, which is the same as that using the full relevance data (pool-depth-125). Since there are no misses and false alarms in this case, the pool-depth-50 relevance data and the full relevance data yield identical significance test results in terms of Q. Furthermore, Q maintains relatively high discriminative power even with shallow pools. More generally, however, the discriminative power of metrics goes down as the pool depth is reduced, and the number of misses and false alarms increases. Moreover, it can be observed that, *in the presence of pool depth bias, AP′, Q′ and nDCG′ are not necessarily superior to AP, Q and nDCG in terms of discriminative power.* For example, at pool depth 10 for TREC03, the discriminative power of Q is 68.3% (with seven missses and nine false alarms), while that of Q′ is only 58.3% (with seven misses and six false alarms). Nevertheless, it can be observed that AP′ and Q′ are generally

Table 8 Kendall's rank correlation: the original vs. shallow-pool rankings

	AP′	Q′	nDCG′	RBP′	bpref	AP	Q	nDCG	RBP
TREC03 (original *pd* = 125); ranking 16 runs									
pd = 50	1	.950	.967	1	.983	.967	.967	.983	1
pd = 10	.933	.850	.900	.883	.917	.883	.867	.900	.950
pd = 1	.717	.700	.800	.800	.817	.667	.650	.750	.750
TREC04 (original *pd* = 100); ranking 14 runs									
pd = 50	1	1	1	1	1	.978	1	.978	1
pd = 10	.978	.978	.890	.956	.956	.868	.912	.846	1
pd = 1	.846	.780	.846	.692	.846	.736	.802	.626	.780
NTCIR-6J (original *pd* = 100); ranking 10 runs									
pd = 50	1	1	.956	1	.956	.911	1	.956	1
pd = 10	.911	.911	.956	1	.867	.867	.911	.956	.867
pd = 1	.644	.689	.733	.644	.689	.511	.600	.867	.689
NTCIR-6C (original *pd* = 100); ranking 10 runs									
pd = 50	1	1	1	1	1	1	1	1	1
pd = 10	1	.911	.956	.911	1	.956	.956	1	1
pd = 1	.911	.911	.956	.822	.911	.778	.867	.867	.822

Table 9 Discriminative power at $\alpha = 0.05$: shallow pools. High values are shown in bold (TREC).

AP′	Q′	nDCG′	RBP′	bpref	AP	Q	nDCG	RBP
TREC03 full relevance data (*pd* = 125): discriminative power								
64.2%	64.2%	59.2%	45.8%	57.5%	64.2%	**66.7%**	59.2%	45.8%
TREC03 *pd* = 50: discriminative power, misses and false alarms								
64.2%	64.2%	59.2%	46.7%	55.8%	64.2%	**66.7%**	60.0%	47.5%
1	1	2	0	3	1	0	2	0
1	1	2	1	1	1	0	3	2
TREC03 *pd* = 10: discriminative power, misses and false alarms								
60.8%	58.3%	58.3%	36.7%	54.2%	63.3%	**68.3%**	59.2%	46.7%
11	12	7	16	9	9	7	6	5
7	5	6	5	5	8	9	6	6
TREC03 *pd* = 1: discriminative power, misses and false alarms								
36.7%	37.5%	36.7%	10.8%	30.8%	40.8%	**41.7%**	40.0%	15.8%
43	43	39	43	39	42	46	38	37
10	11	12	1	7	14	16	15	1
TREC04 full relevance data (*pd* = 125): discriminative power								
67.0%	68.1%	63.7%	50.5%	62.6%	67.0%	**69.2%**	63.7%	49.5%
TREC04 *pd* = 50: discriminative power, misses and false alarms								
67.0%	67.0%	59.3%	49.5%	60.4%	**69.2%**	**69.2%**	60.4%	49.5%
1	1	4	1	2	0	1	3	0
1	0	0	0	0	2	1	0	0
TREC04 *pd* = 10: discriminative power, misses and false alarms								
60.4%	60.4%	56.0%	44.0%	57.1%	63.7%	**65.9%**	58.2%	46.2%
7	8	10	6	7	7	7	7	4
1	1	3	0	2	4	4	2	1
TREC04 *pd* = 1: discriminative power, misses and false alarms								
40.7%	40.7%	34.1%	0.0%	38.5%	48.4%	**49.5%**	46.2%	0.0%
27	29	29	46	28	28	29	25	45
3	4	2	0	6	11	11	9	0

Table 10 Discriminative power at $\alpha = 0.05$: shallow pools. High values are shown in bold (NTCIR).

AP′	Q′	nDCG′	RBP′	bpref	AP	Q	nDCG	RBP
NTCIR-6J full relevance data (*pd* = 125): discriminative power								
55.6%	62.2%	73.3%	57.8%	51.1%	57.8%	62.2%	**73.3%**	57.8%
NTCIR-6J *pd* = 50: discriminative power, misses and false alarms								
60.0%	64.4%	71.1%	57.8%	53.3%	60.0%	66.7%	**73.3%**	57.8%
1	1	1	0	0	0	0	0	0
3	2	0	0	1	1	2	0	0
NTCIR-6J *pd* = 10: discriminative power, misses and false alarms								
71.1%	71.1%	64.4%	57.8%	57.8%	64.4%	66.7%	**71.1%**	62.2%
0	1	4	1	1	3	2	2	1
7	5	0	1	4	6	4	1	3
NTCIR-6J *pd* = 1: discriminative power, misses and false alarms								
57.8%	62.2%	57.8%	53.3%	44.4%	60.0%	64.4%	62.2%	**66.7%**
6	5	7	5	9	9	9	9	6
7	5	0	3	6	10	10	4	10
NTCIR-6C full relevance data (*pd* = 125): discriminative power								
80.0%	75.6%	75.6%	71.1%	75.6%	**82.2%**	80.0%	75.6%	71.1%
NTCIR-6C *pd* = 50: discriminative power, misses and false alarms								
80.0%	75.6%	75.6%	71.1%	71.1%	**82.2%**	80.0%	75.6%	71.1%
0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0
NTCIR-6C *pd* = 10: discriminative power, misses and false alarms								
80.0%	75.6%	71.1%	73.3%	71.1%	77.8%	77.8%	71.1%	73.3%
1	1	2	0	3	2	2	2	1
1	1	0	1	1	0	1	0	2
NTCIR-6C *pd* = 1: discriminative power, misses and false alarms								
66.7%	**66.7%**	62.2%	55.6%	55.6%	60.0%	**66.7%**	60.0%	60.0%
7	5	7	7	9	11	6	8	6
1	1	1	0	0	1	0	1	1

more discriminative than bpref, RBP and RBP′ in the presence of pool depth bias.

8. Conclusions and Future Work

Several recent studies discussed the effect of incomplete relevance data in retrieval evaluation using random samples of the original relevance data[1),2),11),16),22)]. They discussed neither system bias nor pool depth bias. However, in reality, relevance data formed through pooling are never a random sample of the full relevance data. In light of this, we examined the effect of system bias and that of pool depth bias on IR metrics. Even though previous studies[11),16)] showed that AP′, Q′ and nDCG′ are effective for handling very incomplete but unbiased data, we showed that they are not necessarily effective in the presence of these biases. Using data from both TREC and NTCIR, we first showed that condensed-list metrics overestimate new systems while traditional metrics underestimate them, and that the overestimation tends to be larger than the underestimation. We then showed that, when relevance data are heavily biased towards a single team or a few teams, AP′, Q′ and nDCG′ are not necessarily superior to AP, Q and nDCG in terms of discriminative power. Moreover, we showed that AP′, Q′ and nDCG′ are not advantageous in the presence of pool depth bias either. Hence previous studies that used random sampling should be interpreted with caution. Nevertheless, AP′ and Q′ *are* generally more discriminative than bpref, RBP and RBP′ in the presence of system bias and pool depth bias. Hence we maintain that AP′ and Q′ are better solutions than bpref, RBP and RBP′ to the problem of incompleteness and bias.

Traditional metrics assume that retrieved unjudged documents are nonrelevant, while condensed-list metrics, including bpref, assume that they are nonexistent. In essence, the present study showed that the latter assumption is no better than the former when the number of pooled systems or the number of pooled documents is small. In our future work, we would like to couple efficient and reliable test construction methods with reliable graded-relevance metrics. We also plan to establish quantitative criteria for choosing good evaluation metrics: Although discriminative power is probably one important criterion, there are probably other aspects that need to be examined, including the ability to *predict* the system ranking according to an intuitive metric such as precision-at-ten, given a set of new topics[20)].

References

1) Ahlgren, P. and Grönqvist, L.: Evaluation of Retrieval Effectiveness with Incomplete Relevance Data: Theoretical and Experimental Comparison of Three Measures, *Information Processing and Management*, Vol.44, No.1, pp.212–225 (2008).

2) Buckley, C. and Voorhees, E.M.: Retrieval Evaluation with Incomplete Information, *Proc. ACM SIGIR 2004*, pp.25–32 (2004).

3) Buckley, C., et al.: Bias and the Limits of Pooling for Large Collections, *Information Retrieval*, Vol.10, No.6, pp.491–508 (2007).
4) Burges, C., et al.: Learning to Rank using Gradient Descent, *Proc. ACM ICML 2005*, pp.89–96 (2005).
5) Büttcher, S., et al.: Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements, *Proc. ACM SIGIR 2007*, pp.63–70 (2007).
6) Efron, B. and Tibshirani, R.: *An Introduction to the Bootstrap*, Chapman & Hall/CRC (1993).
7) Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM TOIS*, Vol.20, No.4, pp.422–446 (2002).
8) Järvelin, K., Price, S., Delcambre, L.M.L. and Nielsen, M.L.: Discounted Cumulative Gain Based Evaluation of Multiple-Query IR Sessions, *Proc. ECIR 2008* (*LNCS 4956*), pp.4–15 (2008).
9) Kando, N.: Overview of the Sixth NTCIR Workshop, *Proc. NTCIR-6*, pp.i–ix (2007).
10) Moffat, A. and Zobel, J.: Rank-Biased Precision for Measurement of Retrieval Effectiveness, *ACM TOIS*, Vol.7, No.1, p.Article No.2 (2008).
11) Sakai, T.: Alternatives to Bpref, *Proc. ACM SIGIR 2007*, pp.71–78 (2007).
12) Sakai, T.: Evaluating Information Retrieval Metrics based on Bootstrap Hypothesis Tests, *IPSJ Trans. Databases*, Vol.48, No.SIG 9 (TOD35), pp.11–28 (2007).
13) Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, Vol.43, No.2, pp.531–548 (2007).
14) Sakai, T.: Comparing Metrics across TREC and NTCIR: The Robustness to Pool Depth Bias, *Proc. ACM SIGIR 2008*, pp.691–692 (2008).
15) Sakai, T.: Comparing Metrics across TREC and NTCIR: The Robustness to System Depth Bias, *Proc. ACM CIKM 2008*, pp.581–590 (2008).
16) Sakai, T. and Kando, N.: On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments, *Information Retrieval*, Vol.11, No.5, pp.447–470 (2008).
17) Sakai, T. and Robertson, S.: Modelling A User Population for Designing Information Retrieval Metrics, *Proc. EVIA 2008*, pp.30–41 (2008).
18) Sanderson, M. and Joho, H.: Forming Test Collections with No System Pooling, *Proc. ACM SIGIR 2004*, pp.33–40 (2004).
19) Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track, *Proc. TREC 2004* (2005).
20) Webber, W., Moffat, A., Zobel, J. and Sakai, T.: Precision-At-Ten Considered Redundant, *Proc. ACM SIGIR 2008*, pp.695–696 (2008).
21) Yilmaz, E., Aslam, J. and Robertson, S.: A New Rank Correlation Coefficient for Information Retrieval, *Proc. ACM SIGIR 2008* (2008).
22) Yilmaz, E. and Aslam, J.A.: Estimating Average Precision with Incomplete and Imperfect Judgments, *Proc. ACM CIKM 2006* (2006).
23) Zobel, J.: How Reliable are the Results of Large-Scale Information Retrieval Experiments?, *Proc. ACM SIGIR '98*, pp.307–314 (1998).

(Received September 26, 2008)
(Accepted January 7, 2009)

(Original version of this article can be found in the Journal of Information Processing Vol.17, pp.156–166.)

Tetsuya Sakai was born in 1968. He received his Master's degree and Ph.D. from Waseda University in 1993 and 2000, respectively. From 1993 to 2007, he worked at the Toshiba R&D Center. From 2000 to 2001, he was a visiting scholar at the University of Cambridge. In 2007, he became the director of the Natural Language Processing Laboratory at NewsWatch, Inc. He has received awards from IPSJ, IEICE Engineering Sciences Society (ESS) and at Forum on Information Technology (FIT) conferences. He is a member of ACM, BCS-IRSG, IPSJ-SIGFI and IEICE.