Chapter 1

Graded Relevance

Tetsuya Sakai

Abstract NTCIR was the first large-scale IR evaluation conference series to construct test collections with graded relevance assessments: the NTCIR-1 test collections from 1998 already featured *relevant* and *partially relevant* documents. In this chapter, I provide a survey on the use of graded relevance assessments and of graded relevance measures in the past NTCIR tasks which primarily tackled ranked retrieval. My survey shows that the majority of the past tasks fully utilised graded relevance by means of graded evaluation measures, but not all of them; interestingly, even a few relatively recent tasks chose to adhere to binary relevance measures. I conclude the chapter by a summary of my survey in table form.

1.1 Introduction

The evolution of NTCIR is quite different from that of TREC when it comes to how *relevance assessments* have been collected and utilised. In 1992, TREC started off with a high-recall task (i.e., the *ad hoc* track), with *binary* relevance assessments (Harman 2005). Moreover, early TREC tracks heavily relied on evaluation measures based on binary relevance such as *11-point Average Precision*, *R-precision*, and (noninterpolated) *Average Precision*. It was in the TREC 2000 (a.k.a. TREC-9) Main Web task that 3-point *graded* relevance assessments were introduced, based on feedback from web search engine companies at that time Hawking and Craswell (2005, p. 204). Accordingly, this task also Järvelin and Kekäläinen (2000) adopted Discounted Cumulative Gain (DCG), to utilise the graded relevance assessments.

NTCIR has collected graded relevance assessments from the very beginning: the NTCIR-1 test collections from 1998 already featured *relevant* and *partially relevant* documents (Kando et al. 1999). Thus, while NTCIR borrowed many ideas from TREC when it was launched in the late 1990s, its policy regarding relevance assessments seems to have followed the paths of *Cranfield II* (which had 5-point relevance

T. Sakai (✉)

Waseda University, Shinjuku-ku Okubo 3-4-1 63-05-04, Tokyo 169-8555, Japan
e-mail: tetsuyasakai@acm.org

© The Author(s) 2021

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_1

levels) Cleverdon et al. (1966, p. 21), Oregon Health Sciences University’s MEDLINE Data Collection (OHSUMED) (which had 3-point relevance levels) (Hersh et al. 1994), as well as the first Japanese IR test collections *BMIR-J1* and *BMIR-J2* (which also had 3-point relevance levels) (Sakai et al. 1999).

Interestingly, with perhaps a notable exception of the aforementioned TREC 2000 Main Web Task, it is true for both TREC and NTCIR that the introduction of graded relevance assessments did not necessarily mean immediate adoption of *evaluation measures* that can utilise graded relevance. For example, while the TREC 2003–2005 robust tracks constructed adhoc IR test collections with 3-point graded relevance assessments, they adhered to binary relevance measures such as Average Precision (AP). Similarly, as I shall discuss in this chapter,¹ while almost all of the past IR tasks of NTCIR had graded relevance assessments, not all of them fully utilised them by means of graded relevance measures. This is the case despite the fact that a graded relevance measure called the *normalised sliding ratio* (NSR)² was proposed in 1968 (Pollock 1968), and was discussed in an 1997 book by Korfhage along with another graded relevance measure (Korfhage 1997, p.209).

1.2 Graded Relevance Assessments, Binary Relevance Measures

This section provides an overview of NTCIR ranked retrieval tasks that did not use graded relevance evaluation measures even though they had graded relevance assessments.

1.2.1 Early IR and CLIR Tasks (NTCIR-1 Through -5)

The Japanese IR and (Japanese-English) crosslingual tasks of NTCIR-1 (Kando et al. 1999) constructed test collections with 3-point relevance levels, but used binary relevance measures such as AP and *R-precision* by either treating the relevant and partially relevant documents as “relevant” or treating only the relevant documents as “relevant.” However, it should be stressed at this point that using binary relevance measures with different relevance thresholds cannot serve as substitutes for a graded relevance measure that enables optimisation towards an ideal ranked list (i.e., a list of documents sorted in decreasing order of relevance levels). If partially relevant

¹A 31-page, March 2019 version of this chapter is available on arxiv.org Sakai (2019). The arxiv version contains the definitions of the main graded relevance measures used at NTCIR, as well as details on how graded relevance levels were constructed from individual assessors’ judgements for some of the tasks.

²NSR is actually what is now known as *normalised (nondiscounted) cumulative gain* (nCG): See Sakai (2019).

documents are ignored, a Search Engine Result Page (SERP) whose top l documents are all partially relevant and one whose top l documents are all nonrelevant can never be distinguished from each other; if relevant documents and partially relevant documents are all treated as relevant, a SERP whose top l documents are all relevant and one whose top l documents are all partially relevant can never be distinguished from each other.

The Japanese and English (monolingual and crosslingual) IR tasks of NTCIR-2 (Kando et al. 2001) constructed test collections with 4-point relevance levels. However, the organisers used binary relevance measures such as AP and R-precision with two different relevance thresholds. As for the Chinese monolingual and Chinese-English IR tasks of NTCIR-2 (Chen and Chen 2001), three judges independently judged each pooled document using 4-point relevance levels, and then a score was assigned to each relevance level. Finally, the scores were averaged across the three assessors. The organisers then applied two different thresholds to map the scores to binary *rigid relevance* and *relaxed relevance* data. For evaluating the runs, rigid and relaxed versions of recall-precision curves (RP curves) were used.

The NTCIR-3 CLIR (Cross-Language IR) task (Chen et al. 2002) was similar to the previous IR tasks: 4-point relevance levels were used, and two relevance thresholds were used. Finally, rigid and relaxed versions of AP were computed for each run. The NTCIR-4 and NTCIR-5 CLIR tasks (Kishida et al. 2004, 2005) adhered to the above practice.

All of the above tasks used the `trec_eval` program from TREC to compute binary relevance measures such as AP.

1.2.2 Patent (NTCIR-3 Through-6)

The NTCIR-3 Patent Retrieval task (Iwayama et al. 2003) was a news-to-patent *technical survey* search task, with 4-point relevance levels. RP curves were drawn based on *strict relevance* and *relaxed relevance*.

The main task of the NTCIR-4 Patent Retrieval task (Fujii et al. 2004) was a patent-to-patent *invalidity search* task. There were two types of relevant documents: A (a patent that can invalidate a given claim on its own) and B (a patent that can invalidate a given claim *only when used with one or more other patents*). For example, patents B_1 and B_2 may each be nonrelevant (as they cannot invalidate a claim individually), but if they are both retrieved, the pair should serve as one relevant document. At the evaluation step, rigid and relaxed APs were computed. Note that the above-relaxed evaluation has a limitation: recall the aforementioned example with B_1 and B_2 , and consider a SERP that managed to return only one of them (say B_1). Relaxed evaluation rewards the system for returning B_1 , even though this document alone does *not* invalidate the claim.

The Document Retrieval subtask of the NTCIR-5 Patent Retrieval task (Fujii et al. 2005) was similar to its predecessor, but the relevant documents were determined purely based on whether and how they were actually used by a patent examiner to

reject a patent application; no manual relevance assessments were conducted for this subtask. The graded relevance levels were defined as follows: A (a citation that was actually used on its own to reject a given patent application) and B (a citation that was actually used along with another one to reject a given patent application). As for the evaluation measure for Document Ranking, the organisers adhered to rigid and relaxed APs. In addition, the task organisers introduced a Passage Retrieval subtask by leveraging passage-level binary relevance assessments collected as in the NTCIR-4 Patent task: given a patent, systems were required to rank the passages from that same patent. As both single passages and groups of passages can potentially be relevant to the source patent (i.e., the passage(s) can serve as evidence to determine that the entire patent is relevant to a given claim), this poses a problem similar to the one discussed above with patents B_1 and B_2 : for example, if two passages p_1, p_2 are relevant as a group but not individually, and if p_1 is ranked at i and p_2 is ranked at $i' (> i)$, how should the SERP of passage be evaluated? To address this, the task organisers introduced a binary relevance measure called the *Combinational Relevance Score* (CRS), which assumes that the user who scans the SERP must reach as far as i' to view both p_1 and p_2 .³

The Japanese Document Retrieval subtask of the NTCIR-6 Patent Retrieval task (Fujii et al. 2007) had two different sets of graded relevance assessments; the first set (“Def0” with A and B documents) was defined in the same way as in NTCIR-5; the second set (“Def1”) was automatically derived from Def0 based on the International Patent Classification (IPC) codes as follows: H (the set of IPC subclasses for this cited patent has no overlap with that of the input patent), A (the set of IPC subclasses for this cited patent has some overlap with that of the input patent), and B (the set of IPC subclasses for this cited patent is identical to that of the input patent). As for the English Document Retrieval subtask, the relevance levels were also automatically determined based on IPC codes, but only two types of relevant documents (A and B) were identified, as each USPTO patent is given only one IPC code. In both subtasks, AP was computed by considering different combinations of the above relevance levels.

1.2.3 *SpokenDoc/SpokenQuery& Doc* (*NTCIR-9 Through -12*)

The Spoken Document Retrieval (SDR) subtask of the NTCIR-9 SpokenDoc task (Akiba et al. 2011) had two “subsubtasks”: *Lecture Retrieval* and *Passage Retrieval*, where a passage is any sequence of consecutive inter-pausal units. Passage-level relevance assessments were obtained on a 3-point scale, and it appears that the lecture-

³In fact, AP, Q or any measure from the NCU family (Sakai and Robertson 2008) can easily be extended to handle *combinational relevance* for Document Retrieval (See the above example with (B_1, B_2)) and for Passage Retrieval (See the above example with (p_1, p_2)): See Sakai (2019).

level (binary) relevance was deduced from them.⁴ AP was used for evaluating Lecture Retrieval, whereas variants of AP, called *utterance-based* (M)AP, *pointwise* (M)AP, and *fractional* (M)AP were used for evaluating Passage Retrieval. These are all binary relevance measures. The NTCIR-10 SpokenDoc-2 Spoken Content Retrieval (SCR) subtask (Akiba et al. 2013) was similar to the SDR subtask at NTCIR-9, with Lecture Retrieval and Passage Retrieval subsubtasks. Lecture Retrieval used a revised version of the NTCIR-9 SpokenDoc topic set, and its gold data does not contain graded relevance assessments⁵; binary relevance AP was used for the evaluation. As for Passage Retrieval, a new topic set was devised, again with 3-point relevance levels.⁶ The AP variants from the NTCIR-9 SDR task were used for the evaluation again.

The Slide Group Segment (SGS) Retrieval subsubtask of the NTCIR-11 SpokenQuery& Doc SCR subtask involved the ranking of predefined retrieval units (i.e., SGSs), unlike the Passage Retrieval subsubtask which allows any sequence of consecutive inter-pausal units as a retrieval unit. Three-point relevance levels were used to judge the SGSs: R (relevant), P (partially relevant), and I (nonrelevant). However, binary AP was used for the evaluation after collapsing the grades to binary. As for the passage-level relevance assessments, they were derived from the SGSs labelled R or P, and were considered binary; the three AP variants were used for this subsubtask again. Segment Retrieval was continued at the NTCIR-12 SpokenQuery&Doc-2 task, again with the same 3-point relevance levels and AP as the evaluation measure.

1.2.4 Math/MathIR (NTCIR-10 Through -12)

In the Math Retrieval subtask of the NTCIR-10 Math Task, retrieved mathematical formulae were judged on a 3-point scale. Up to two assessors judged each formula, and initially 5-point relevance scores were devised based on the results. For example, for formulae judged by one assessor, they were given 4 points if the judged label was relevant; for those judged by two assessors, they were given 4 points if both of them gave them the relevant label. Finally, the scores were mapped to a 3-point scale: Documents with scores 4 or 3 were treated as relevant; those with 2 or 1 were treated as partially relevant; those with 0 were treated as nonrelevant. However, at the evaluation step, only binary relevance measures such as AP and Precision were computed using `trec_eval`, after collapsing the grades to binary. Similarly, in the Math Retrieval subtask of the NTCIR-11 Math Task (Aizawa et al. 2014), two assessors independently judged each retrieved unit on a 3-point scale, and the

⁴The official test collection data of the NTCIR-9 SDR task (`evalsdr`) contains only passage-level gold data.

⁵This was verified by examining SpokenDoc2-formalrun-SCR-LECTURE-golden-20130129.xml in the SpokenDoc-2 test collection <http://research.nii.ac.jp/ntcir/permission/ntcir-10/perm-en-SPOKENDOC.html>.

⁶This was verified by examining <http://SpokenDoc2-formalrun-SCR-PASSAGE-golden-20130215.xml> in the SpokenDoc-2 test collection <http://research.nii.ac.jp/ntcir/permission/ntcir-10/perm-en-SPOKENDOC.html>.

final relevance levels were also on a 3-point scale. If the two assessor labels were relevant/relevant or relevant/partially relevant, the final grade was relevant; if the two labels were both nonrelevant, the final grade was nonrelevant; the other combinations were considered partially relevant. As for the evaluation measures, *bpref* (Buckley and Voorhees 2004; Sakai 2007; Sakai and Kando 2008) was computed along with AP and Precision using `trec_eval`.

The NTCIR-12 MathIR task was similar to the Math Retrieval subtask of the aforementioned Math tasks. Up to four assessors judged each retrieved unit using a 3-point scale, and the individual labels were consolidated to form the final 3-point scale assessments. As for the evaluation, only Precision was computed at several cutoffs using `trec_eval`.

The NTCIR-11 Math (Aizawa et al. 2014) and NTCIR-12 MathIR (Zanibbi et al. 2016) overview papers suggest that one reason for adhering to binary relevance measures is that `trec_eval` could not handle graded relevance. On the other hand, this may not be the only reason: in the MathIR overview paper, it is reported that the organisers chose Precision because it is “*simple to understand*” (Zanibbi et al. 2016). Thus, some researchers indeed *choose* to focus on evaluation with binary relevance measures, even in the NTCIR community where we have graded relevance data by default and a tool for computing graded relevance measures is known.⁷

1.3 Graded Relevance Assessments, Graded Relevance Measures

This section provides an overview of NTCIR ranked retrieval tasks that employed graded relevance evaluation measures to fully enjoy the benefit of having graded relevance assessments.

1.3.1 Web (NTCIR-3 Through-5)

The NTCIR-3 Web Retrieval task (Eguchi et al. 2003) was the first NTCIR task to use a graded relevance evaluation measure, namely, DCG.⁸ Four-point relevance levels were used. In addition, assessors chose a very small number of “best” documents from the pools. To compute DCG, two different gain value settings were used: Rigid (3 for highly relevant, 2 for fairly relevant, 0 otherwise) and Relaxed (3 for highly

⁷ NTCIREVAL has been available on the NTCIR website since 2010; its predecessor `ir4qa_eval` was released in 2008 (Sakai et al. 2008). Note also that TREC 2010 released <https://trec.nist.gov/data/web/10/gdeval.pl> for computing Normalised Discounted Cumulative Gain (nDCG) and Expected Reciprocal Rank (ERR).

⁸ This was the DCG as originally defined by Järvelin and Kekäläinen (2000) with the logarithm base $b = 2$, which means that gain discounting is not applied to documents at ranks 1 and 2. See also Sect. 1.3.3.

relevant, 2 for fairly relevant, 1 for partially relevant, 0 otherwise). The organisers of the Web Retrieval task also defined a graded relevance evaluation measure called Weighted Reciprocal Rank (WRR), designed for navigational searches. However, what was actually used in the task was the binary relevance Reciprocal Rank (RR), with two different relevance thresholds. Therefore, this measure will be denoted “(W)RR” hereafter whenever graded relevance is not utilised. Other binary relevance measures including AP and R-precision were also used in this task. For a comparison of evaluation measures designed for navigational intents including RR, WRR, and P+, see Sakai (2007).

The NTCIR-4 WEB Informational Retrieval Task (Eguchi et al. 2004) was similar to its predecessor, with 4-point relevance levels; the evaluation measures were DCG, (W)RR, Precision, etc. On the other hand, the NTCIR-4 WEB Navigational Retrieval Task (Oyama et al. 2004), used 3-point relevance levels: A (relevant), B (partially relevant), and D (nonrelevant); the evaluation measures were DCG and (W)RR, and two gain values settings for DCG were used: $(A, B, D) = (3, 0, 0)$ and $(A, B, D) = (3, 2, 0)$.

The NTCIR-5 WEB task ran the Navigational Retrieval subtask, which is basically the same as its predecessor, with 3-point relevance levels and DCG and (W)RR as the evaluation measures. For computing DCG, three gain value settings were used: $(A, B, D) = (3, 0, 0)$, $(A, B, D) = (3, 2, 0)$, and $(A, B, D) = (3, 3, 0)$. Note that the first and the third settings reduce DCG to binary relevance measures.

1.3.2 CLIR (NTCIR-6)

At the NTCIR-6 CLIR task, 4-point relevance levels (S,A,B,C) were used and rigid and relaxed AP scores were computed using `trec_eval` as before. In addition, the organisers computed “*as a trial*” (Kishida et al. 2007) the following graded relevance measures using their own script: nDCG (as defined originally by Järvelin and Kekäläinen 2002), Q-measure (Sakai 2014; Sakai and Zeng 2019) (or “Q”), and Kishida’s *generalised AP* (Kishida 2005). See Sakai (2007) for a comparison of these three graded relevance measures. The CLIR organisers developed a program to compute these graded relevance measures, with the gain value setting: $(S, A, B, C) = (3, 2, 1, 0)$.

1.3.3 ACLIA IR4QA (NTCIR-7 and -8)

At the NTCIR-7 Information Retrieval for Question Answering (IR4QA) task (Sakai et al. 2008), a predecessor of NTCIREVAL called `ir4qa_eval` was released (See Sect. 1.2.4). This tool was used to compute the Q-measure, the “Microsoft version” of nDCG (Sakai 2014), as well as the binary relevance AP. Microsoft nDCG (called MSnDCG in NTCIREVAL) fixes a problem with the original nDCG (See also

Sect. 1.3.1): in the original nDCG, if the logarithm base is set to (say) $b = 10$, then discounting is not applied from ranks 1 to 10. Hence, the ranks of the relevant documents within top 10 do not matter. Microsoft nDCG avoids this problem by using $1/\log(1+r)$ as the discount factor for *every* rank r , but thereby loses the patience parameter b (Sakai 2014).⁹ The relevance levels used were L2, L1, and L0. A *linear* gain value setting was used: $(L2, L1, L0) = (2, 1, 0)$. The NTCIR-8 IR4QA task (Sakai et al. 2010) used the same evaluation methodology as above.

1.3.4 GeoTime (NTCIR-8 and -9)

The NTCIR-8 GeoTime task (Gey et al. 2010), which dealt with adhoc IR given “when and where”-type topics, constructed test collections with the following graded relevance levels: Fully relevant (the document answers both the “when” and “where” aspects of the topic), Partially relevant—where (the document only answers the “where” aspect of the topic), and Partially relevant—when (the document only answers the “when” aspect of the topic). The evaluation tools from the IR4QA task were used to compute (Microsoft) nDCG, Q, and AP, with a gain value of 2 for each fully relevant document and a gain value of 1 for each partially relevant one (regardless of “when” or “where”) for the two graded relevance measures.¹⁰ The NTCIR-9 GeoTime task (Gey et al. 2011) used the same evaluation methodology as above.

1.3.5 CQA (NTCIR-8)

The NTCIR-8 Community Question Answering (CQA) task (Sakai et al. 2010) was an answer ranking task: given a question from Yahoo! Chiebukuro (Japanese Yahoo! Answers) and the answers posted in response to that question, rank the answers by answer quality. While the Best Answers (BAs) selected by the actual questioners were already available in the Chiebukuro data, additional *graded* relevance assessments were obtained offline to find *Good Answers* (GAs), by letting four assessors independently judge each posted answer. Each assessor labelled an answer as either A (high-quality), B (medium-quality), or C (low-quality), and hence 15 different label *patterns* were obtained: AAAA, AAAB, . . . , BCCC, CCCC. In the official evaluation at NTCIR-8, these patterns were mapped to 4-point relevance levels: for example, AAAA and AAAB were mapped to L3-relevant, and ACCC, BCCC and CCCC were mapped to L0. In a separate study, the same data were mapped to 9-point relevance levels, by giving 2 points to an A and 1 point to a B and summing

⁹D_g-nDCG implemented in NTCIREVAL also builds on the Microsoft version of nDCG, not the original nDCG.

¹⁰While the GeoTime overview paper suggests that the above relevance levels were mapped to binary relevance, this was in fact not the case: see Sakai (2019).

up the scores for each pattern. Using the graded Good Answers data, three graded relevance measures were computed: normalised gain at $l = 1$ ($nG@1$),¹¹ $nDCG$, and Q . In addition, Hit at $l = 1$ was computed for both Best Answers and Good Answers data: this is a binary relevance measure which only cares whether the top-ranked item is relevant or not.

1.3.6 INTENT/IMine (NTCIR-9 Through 12)

The NTCIR-9 INTENT task overview paper (Song et al. 2011) was the first NTCIR overview to mention the use of the NTCIREVAL tool, which can compute various graded relevance measures for adhoc and diversified IR including Q , $nDCG$, and $D_{\#}$ -measures (Sakai and Zeng 2019). $D_{\#}$ - $nDCG$ and its components I-rec and D- $nDCG$ were used as the official evaluation measures. The Document Retrieval (DR) subtask of the INTENT task had intentwise graded relevance assessments on a 5-point scale. While the Subtopic Mining (SM) subtask of the INTENT task also used $D_{\#}$ - $nDCG$ to evaluate ranked lists of subtopic strings, no graded relevance assessments were involved in the SM subtask since each subtopic string either belongs to an intent (i.e., a cluster of subtopic strings) or not. Hence, the SM subtask may be considered to be outside the scope of the present survey; but see Sakai (2019) for a discussion.

The NTCIR-10 INTENT task was basically the same as its predecessor, with 5-point intentwise relevance levels for the DR subtask and $D_{\#}$ - $nDCG$ as the primary evaluation measure. However, as the intents came with informational/navigational tags, new measures called DIN- $nDCG$ and P+Q (Sakai 2014) were used in addition to leverage this information.

The NTCIR-11 IMine task (Liu et al. 2014) was similar to the INTENT tasks, except that its SM subtask required participating systems to return a two-level hierarchy of subtopic strings. The SM subtask was evaluated using the H-measure, which combines (a) the accuracy of the hierarchy, (b) the $D_{\#}$ - $nDCG$ score based on the ranking of the first-level subtopics, and (c) the $D_{\#}$ - $nDCG$ score based on the ranking of the second-level subtopics. However, recall the above remark on the INTENT SM subtask: intentwise graded relevance does not come into play in this subtask. On the other hand, the IMine DR subtask was evaluated in a way similar to the INTENT DR tasks, with $D_{\#}$ - $nDCG$ computed based on 4-point relevance levels: highly relevant, relevant, nonrelevant, and spam. The gain value setting used was: (2, 1, 0, 0).¹² The IMine task also introduced the TaskMine subtask, which requires systems to rank strings that represent subtasks of a given task (e.g., “take diet pills” in response to “lose weight.”). This subtask involved graded relevance assessments. Each subtask string was judged independently by two assessors from the viewpoint of whether

¹¹ $nG@1$ is often referred to as $nDCG@1$; however, note that neither discounting nor cumulation is applied at rank 1.

¹² Kindly confirmed by task organisers Yiqun Liu and Cheng Luo in a private email communication (March 2019).

the subtask is effective for achieving the input task. A 4-point per-assessor relevance scale was used,¹³ with weights (3, 2, 1, 0), and final relevance levels were given as the average of the two scores, which means that a 6-point relevance scheme was adopted. The averages were used verbatim as gain values: (3.0, 2.5, 2.0, 1.5, 1.0, 0). The evaluation measure used was nDCG, but duplicates (i.e., multiple strings representing the same subtask) were not rewarded.

The Query Understanding (QU) subtask of the NTCIR-12 IMine-2 Task (Yamamoto et al. 2016), a successor of the previous SM subtasks of INTENT/IMine, required systems to return a ranked list of (subtopic, vertical) pairs (e.g., (“iPhone 6 photo”, Image), (“iPhone 6 review”, Web)) for a given query. The official evaluation measure, called the QU-score, is a linear combination of $D\sharp$ -nDCG (computed as in the INTENT SM subtasks) and the V-score which measures the appropriateness of the named vertical for each subtopic string. Despite the binary relevance nature of the subtopic mining aspect of the QU subtask, it deserves to be discussed in the present survey because the V-score part relies on graded relevance assessments. To be more specific, the V-score relies on the probabilities $\{Pr(v|i)\}$, for intents $\{i\}$ and verticals $\{v\}$, which are derived from 3-point scale relevance assessments: 2 (highly relevant), 1 (relevant), and 0 (nonrelevant). Hence the QU-score may be regarded as a graded relevance measure. The Vertical Incorporating (VI) subtask of the NTCIR-12 IMine-2 Task (Yamamoto et al. 2016) also used a version of $D\sharp$ -nDCG to allow systems to embed verticals (e.g., Vertical-News, Vertical-Image) within a ranked list of document IDs for diversified search. More specifically, the organisers replaced the intentwise gain value $g_i(r)$ at rank r in the global gain formula (Sakai 2014) with $Pr(v(r)|i)g_i(r)$, where $v(r)$ is the *vertical type* (“Web,” Vertical-News, Vertical-Image, etc.) of the document at rank r , and the vertical probability given an intent is obtained from 3-point scale relevance assessments as described above. As for the intentwise gain value $g_i(r)$, it was also on a 3-point scale for the Web documents: 2 for highly relevant, 1 for relevant, and 0 for nonrelevant documents. Moreover, if the document at r was a vertical, the gain value was set to 2. In addition, the VI subtask collected *topicwise* relevance assessments on a 4-point scale: highly relevant, relevant, nonrelevant, and spam. The gain values used were: (2, 1, 0, 0).¹⁴ As the subtask had a set of very clear, single-intent topics among their full topic set, Microsoft nDCG (rather than $D\sharp$ -nDCG) was used for these particular topics.

1.3.7 RecipeSearch (NTCIR-11)

While the official evaluation results of Adhoc Recipe Search subtask of the NTCIR-11 RecipeSearch Task (Yasukawa et al. 2014) were based on binary relevance, the

¹³While the overview (Sect. 4.3) says that a 3-point scale was used, this was in fact not the case: kindly confirmed by task organiser Takehiro Yamamoto in a private email communication (March 2019).

¹⁴Kindly confirmed by task organisers Yiqun Liu and Cheng Luo in a private email communication (March 2019).

organisers also explored evaluation based on graded relevance: they obtained graded relevance assessments on a 3-point scale for a subset (111 topics) of the full test topic set (500 topics).¹⁵ Microsoft nDCG was used to leverage the above data with a linear gain value setting, along with the binary AP and RR.

1.3.8 *Temporalia (NTCIR-11 and -12)*

The Temporal Information Retrieval (TIR) subtask of the NTCIR-11 Temporalia Task collected relevance assessments on a 3-point scale. Each TIR topic contained a *past question*, *recency question*, *future question*, and an *atemporal question*; participating systems were required to produce a Search Engine Result Page (SERP) for each of the above four questions. This adhoc IR task used Precision and Microsoft nDCG as the official measures, and Q for reference.

While the Temporally Diversified Retrieval (TDR) subtask of the NTCIR-12 Temporalia-2 Task was similar to the above TIR subtask, it required systems to return a fifth SERP, which covers all of the above four temporal classes. That is, this fifth SERP is a diversified SERP, where the temporal classes can be regarded as different search intents for the same topic. The relevance assessment process followed the practice of the NTCIR-11 TIR task, and the SERPs for the four questions were evaluated using nDCG. As for the diversified SERPs, they were evaluated using α -nDCG (Clarke et al. 2008) and D $\ddagger$ -nDCG.

A linear gain value setting was used in both of the above subtasks.¹⁶

1.3.9 *STC (NTCIR-12 Through -14)*

The NTCIR-12 Short Text Conversation (STC) task (Shang et al. 2016) was a response retrieval task given a tweet (or a Chinese Weibo post). For both Chinese and Japanese subtasks, the response tweets were first labelled on a binary scale, for each of the following criteria: Coherence, Topical Relevance, Context Independence, and Non-repetitiveness. The final graded relevance levels were determined using the following mapping scheme:

```

if Coherent AND Topically Relevant
  if Context-independent AND Non-repetitive
    RelevanceLevel = L2
  else
    RelevanceLevel = L1
else

```

¹⁵While the overview paper says that a 4-point scale was used, this was in fact not the case: kindly confirmed by task organiser Michiko Yasukawa (March 2019) in a private email communication.

¹⁶Kindly confirmed by task organiser Hideo Joho in a private email communication (March 2019).

RelevanceLevel = $L0$.

Following the *quadratic* gain value setting often used for web search evaluation (Burgess et al. 2005) and for computing ERR (Chapelle et al. 2009), the Chinese subtask organisers mapped the $L2$, $L1$, and $L0$ relevance levels to the following gain values: $2^2 - 1 = 3$, $2^1 - 1 = 1$, $2^0 - 1 = 0$; according to the present survey of NTCIR retrieval tasks, this is the only case where a quadratic gain value setting was used instead of the linear one. The evaluation measures used for this subtask were $nG@1$, $P+$, and normalised ERR (nERR). As for the Japanese subtask which used Japanese Twitter data, the same mapping scheme was applied, but the scores $((L2, L1, L0) = (2, 1, 0))$ from 10 assessors were averaged to determine the final gain values; a binary relevance, set-retrieval accuracy measure was used instead of $P+$, along with $nG@1$ and nERR.

The NTCIR-13 STC task (Shang et al. 2017) was similar to its predecessor, although systems were allowed to *generate* responses instead of retrieving existing tweets. In the Chinese subtask, 7-point relevance levels were obtained by summing up the assessor scores, and a linear gain value setting was used to compute $nG@1$, $P+$, and nERR. In addition, an alternative approach to consolidating the assessor scores was explored, by considering the fact that some tweets receive unanimous ratings while others do not even if they are the same in terms of the sum of assessor scores (Sakai 2017). The NTCIR-13 STC Japanese subtask used Yahoo! News Comments data instead of Japanese Twitter data. The evaluation method was similar to what was used in the previous Japanese subtask; see Sakai (2019) for more details.

Although the Chinese Emotional Conversation Generation (CECG) subtask of the NTCIR-14 STC subtask (Zhang and Huang 2019) is not exactly a ranked retrieval task, we discuss it here as it is a successor of the previous Chinese STC subtasks that utilises graded relevance measures. Given an input tweet *and an emotional category* such as Happiness and Sadness, participating systems for this subtask were required to return *one* generated response. A mapping scheme similar to the previous Chinese subtasks were used to form 3-point relevance levels. As for the evaluation measures, the relevance scores $(L2, L1, L0) = (2, 1, 0)$ of the returned responses were simply summed or averaged across the test topics.

1.3.10 WWW (NTCIR-13 and -14) and CENTRE (NTCIR-14)

The NTCIR-13 We Want Web (WWW) Task (Luo et al. 2017) was an adhoc web search task. For the Chinese subtask, three assessors independently judged each pooled web page on a 4-point scale: (3, 2, 1, 0); the scores were then summed up to form the final 10-point relevance levels. For the English subtask, two assessors independently judged each pooled web page on a different 4-point scale: highly relevant (2 points), relevant (1 point), nonrelevant (0 points), and error (0 points); the scores were then summed up to form the final 5-point relevance levels. In both

subtasks, linear gain value settings were used to compute (Microsoft) nDCG, Q (the cutoff version (Sakai 2014)), and nERR.

The NTCIR-14 WWW Task (Mao et al. 2019) was similar to its predecessor. The Chinese subtask used the following judgment criteria: highly relevant (3 points), relevant (2 points), marginally relevant (1 point), nonrelevant (0 points), garbled (0 points). Although three assessors judged each topic, the final relevance levels were obtained on a majority-vote basis rather than taking the sum; hence 4-point scale relevance levels were used this time. As for the English subtask, 5-point relevance levels were obtained by following the methodology of the NTCIR-13 English subtask. Both subtasks adhered to Microsoft nDCG, (cutoff-based) Q, and nERR with linear gain value settings.

The NTCIR-14 CLEF NTCIR TREC Reproducibility (CENTRE) task (Sakai et al. 2019) encouraged participants to replicate a pair of runs from the NTCIR-13 WWW English subtask and to reproduce a pair of runs from the TREC 2013 Web Track adhoc task (Collins-Thompson et al. 2014). Additional relevance assessments were conducted on top of the official NTCIR-13 WWW English test collection, by following the relevance assessment methodology of the WWW subtask. As for the evaluation of the TREC runs with the TREC 2013 Web Track adhoc test collection, the original 6-point scale relevance levels Navigational, Key, Highly relevant, Relevant, Nonrelevant, Junk were mapped to L_4 , L_3 , L_2 , L_1 , L_0 , L_0 , respectively. All runs involved in the CENTRE task were evaluated using Microsoft nDCG, (cutoff-based) Q, and nERR, with linear gain value settings.

1.3.11 AKG (NTCIR-13)

The NTCIR-13 Actionable Knowledge Graph (AKG) task (Blanco et al. 2017) had two subtasks: *Action Mining* (AM) and *Actionable Knowledge Graph Generation* (AKGG). Both of them involved graded relevance assessments and graded relevance measures. The AM subtask required systems to rank *actions* for a given *entity type* and an *entity instance*: for example, given “*Product*” and “*Final Fantasy VIII*,” the ranked actions could contain “*play on Android*,” “*buy new weapons*,” etc. Two sets of relevance assessments were collected by means of crowd sourcing: the first set judged the verb parts of the actions (“*play*,” “*buy*,” etc.) whereas the second set judged the entire actions (verb plus modifier as exemplified above). Both sets of judgements were done based on 4-point relevance levels. The AKGG subtask required participants to rank entity *properties*: for example, given a quadruple (Query, Entity, Entity Types, Action) = (“*request funding*,” “*funding*,” “*thing*, *action*,” “*request funding*”), systems might return “*Agent*,” “*ServiceType*,” “*Result*,” etc. Relevance assessments were conducted by crowd workers on a 5-point scale. Both subtasks used nDCG and nERR for the evaluation; linear gain value settings were used.¹⁷

¹⁷Kindly confirmed by task organiser Hideo Joho in a private email communication (March 2019).

1.3.12 *OpenLiveQ (NTCIR-13 and -14)*

The NTCIR-13 OpenLiveQ task (Kato et al. 2017) required participants to rank Yahoo! Chiebukuro questions for a given query, and the offline evaluation part of this task involved ranked list evaluation with graded relevance. Five crowd workers independently judged a list of questions for query q under the following instructions: “Suppose you input q and received a set of questions as shown below. Please select all the questions that you would want to click.” Thus, while the judgement is binary for each assessor, 6-point relevance levels were obtained based on the number of votes. (Microsoft) nDCG, Q, and ERR were computed using a linear gain value setting.

The NTCIR-14 OpenLiveQ-2 task (Kato et al. 2019) is similar to its predecessor, but this time the evaluation involved *unjudged* documents, as the relevance assessments from NTCIR-13 were reused but the target questions to be ranked were not identical to the NTCIR-13 version. The organisers therefore used *condensed-list* (Sakai 2014) versions of Q, (Microsoft) nDCG, and ERR. Also, for OpenLiveQ-2, the organisers switched their primary measure from nDCG to Q, as Q substantially outperformed nDCG (at $l = 5, 10, 20$) in terms of correlation with online (i.e., click-based) evaluation in their experiments (Kato et al. 2018).

1.4 Summary

Table 1.1 summarises Sect. 1.2; Table 1.2 summarises Sect. 1.3. It can be observed that (a) the majority of the past NTCIR ranked retrieval tasks utilised graded relevance measures; and that (b) even a few relatively recent tasks, namely, SpokenQuery& Doc and MathIR from NTCIR-12 held in 2016, refrained from using graded relevance measures. As was discussed in Sect. 1.2.1, researchers should be aware that binary relevance measures with different relevance thresholds (e.g., Relaxed AP and Rigid AP) cannot serve as substitutes for good graded relevance measures. *If* the optimal ranked output for a task is defined as one that sorts all relevant documents in decreasing order of relevance levels, then by definition, graded relevance measures should be used to evaluate and optimise the retrieval systems.

One additional remark regarding Tables 1.1 and 1.2 is that the NTCIR-5 CLIR overview paper (Kishida et al. 2007) was the last to report on RP curves; the RP curves completely disappeared from the NTCIR overviews after that. This may be because (a) interpolated precisions at different recall points (Sakai 2014) do not directly reflect user experience; and (b) graded relevance measures have become more popular than before.

Over the past decade or so, some researchers have pointed out a few disadvantages of using graded relevance, especially in the context of promoting *preference judgements* (e.g., Bashir et al. 2013; Carterette et al. 2008). Carterette et al. (2008) argue that (i) it is difficult to determine relevance grades in advance and to anticipate how the decision will affect evaluation; and (ii) having more grades means more

Table 1.1 NTCIR ranked retrieval tasks with graded relevance assessments and binary relevance measures. Note that the relevance levels for the Patent Retrieval tasks of NTCIR-4 to -6 exclude the “nonrelevant” level: the actual labels are shown here because they are not simply different degrees of relevance (See Sect. 1.2.2)

Task or subtask	NTCIR round (year)	Relevance levels	Main evaluation measures discussed in overview
Japanese and JEIR	1 (1999)	3	AP, R-precision, Precision, RP curves
JEIR	2 (2001)	4	AP, R-precision, Precision, Interpolated Precision, RP curves
Chinese and CEIR	2	4 per assessor	RP curves
CLIR	3–5(2002–2005)	4	AP, RP curves
Patent retrieval	3 (2002)	4	RP curves
Patent retrieval	4 (2004)	A,B	AP, RP curves
Patent retrieval	5 (2005)	A,B	CRS (for passage retrieval), AP
Patent retrieval	6 (2007)	A,B/H,A,B (Japanese) A,B (English)	AP AP
Spoken document/content retrieval	9–11(2011–2014)	3	AP and passage-level variants
SQ-SCR (SGS)	12 (2016)	3	AP
Math retrieval	10 (2013)	5 mapped to 3	AP, Precision
Math retrieval	11 (2014)	3	AP, Precision, Bpref
MathIR	12 (2016)	3	Precision

burden on the users. Regarding (i), while it is important to always check how our use of grades affects the evaluation outcome, in many cases relevance grades can be naturally defined based on individual assessors’ labels; I argue that it is useful to preserve the raw judgements in the form of graded relevance rather than to collapse them to binary; see also the discussion below on label *distributions*. Regarding (ii), rich relevance grades can be obtained even if the individual judgements are binary or tertiary, as I have illustrated in this chapter. Moreover, while I agree that simple side-by-side preference judgements are useful (and can even be used for constructing graded relevance data), it should be pointed out that some of the approaches in the preference judgements domain require more complex judgement protocols than this, e.g., *graded* preference judgements (Carterette et al. 2008), and *contextual* preference judgements (Chandar and Carterette 2013; Golbus et al. 2014). Moreover, while I agree that utilising preference judgements is a promising avenue for future research, the *incompleteness* problem of preference judgements needs to be solved.

What lies beyond graded relevance then? Here is my personal view concerning offline evaluation (as opposed to online evaluation using click data etc.). Information

Table 1.2 NTCIR ranked retrieval tasks with graded relevance assessments and graded relevance measures. Binary relevance measures are shown in parentheses

Task or subtask	NTCIR round (year)	Relevance levels	Main evaluation measures discussed in overview
Web retrieval	3 (2003)	4 + best documents	DCG ((W)RR, AP, RP curves)
WEB informational	4 (2004)	4	DCG ((W)RR, Precision, RP curves)
WEB navigational		3	DCG, ((W)RR, UCS)
WEB navigational	5 (2005)	3	DCG, ((W)RR)
CLIR	6 (2007)	4	nDCG, Q, generalised AP (AP)
IR4QA	7–8 (2008–2010)	3	nDCG, Q (AP)
GeoTime	8–9(2010–2011)	3*	nDCG, Q (AP)
CQA	8 (2010)	4(9) + best answers	GA-{nG@1, nDCG, Q}, (GA-Hit@1, BA-Hit@1) etc.
INTENT DR	9 (2011)	5	D _# -nDCG
INTENT DR	10 (2013)	5	D _# -nDCG, DIN-nDCG, P+Q
IMine DR	11 (2014)	4 incl. Spam	D _# -nDCG
IMine TaskMine	11	6	nDCG
IMine QU	12 (2016)	3 (vertical)	QU-score
IMine VI	12	3 (vertical) 3 (intentwise) 3 + Spam (topicwise)	D _# -nDCG, nDCG
RecipeSearch	11 (2014)	3(2)	nDCG (AP, RR)
Temporalia TIR	11	3	nDCG, Q, (Precision)
Temporalia TDR	12 (2016)	3	nDCG, α -nDCG, D _# -nDCG
STC Chinese	12	3	nG@1, P+, nERR
STC Chinese	13 (2017)	7(10)	nG@1, P+, nERR
STC Japanese	12–13(2016–2017)	3 per assessor	nG@1, nERR (Accuracy)
STC CECG	14 (2019)	3	Sum/average of relevance scores
WWW English	13–14(2017–2019)	5	nDCG, Q, nERR
WWW Chinese	13 (2017)	10	nDCG, Q, nERR
WWW Chinese	14	4	nDCG, Q, nERR
AKG	13 (2017)	4 (AM) / 5 (AKGG)	nDCG, nERR
OpenLiveQ	13–14(2017–2019)	6	nDCG, Q, ERR(with condensed lists at NTCIR-14)
CENTRE	14 (2019)	5	nDCG, Q, nERR

*two types of partially relevant (*when* and *where*) counted as one level

Retrieval (IR) and Information Access (IA) tasks have diversified, and relevance assessments require more subjective and diverse views than before. We are no longer just talking about whether a scientific article is relevant to the researcher’s question (as in Cranfield); we are also talking about whether a response of a chatbot is “relevant” response to the user’s utterance, about whether a reply to a post on social media is “relevant,” and so on. Graded relevance implies that there should be a *single* label for each item to be retrieved (e.g., “this document is highly relevant”), but these new tasks may require a *distribution* of labels reflecting different users’s points of view. Hence, instead of collapsing this distribution to form a single label, methods to preserve the distribution of labels in the test collection may be useful, as was implemented at the Dialogue Breakdown Detection Challenge (Higashinaka et al. 2017). The Dialogue Quality (DQ) and Nugget Detection (ND) subtasks of the NTCIR-14 STC task were the very first of NTCIR efforts in that direction: they compared gold label distributions with systems’ estimated distributions (Sakai 2018; Zeng et al. 2019). See also Maddalena et al. (2017) for a related idea.

Acknowledgements Many thanks to the editors (especially Doug Oard who was in charge of this chapter), reviewers (especially Damiano Spina who kindly reviewed this chapter), and all authors of this book, and to the present and past organisers and participants of the NTCIR tasks! In particular, I thank the following past task organisers for clarifications regarding their overview papers: Cheng Luo, Yiqun Liu, and Takehiro Yamamoto (IMine), Makoto P. Kato (OpenLiveQ), Michiko Yasukawa (RecipeSearch), and Hideo Joho (Temporalia and AKG tasks).

References

- Aizawa A, Kohlhase M, Ounis I (2014) NTCIR-11 Math-2 task overview. In: Proceedings of NTCIR-11, pp 88–98
- Akiba T, Nishizaki H, Aikawa K, Kawahara T, Matsui T (2011) Overview of the IR for spoken documents task in NTCIR-9 workshop. In: Proceedings of NTCIR-9, pp 223–235
- Akiba T, Nishizaki H, Aikawa K, Hu X, Ito Y, Kawahara T, Nakagawa S, Nanjo H, Yamashita Y (2013) Overview of the NTCIR-10 SpokenDoc-2 task. In: Proceedings of NTCIR-10, pp 573–587
- Bashir M, Anderton J, Wu J, Golbus PB, Pavlu V, Aslam JA (2013) A document rating system for preference judgements. In: Proceedings of ACM SIGIR 2013, pp 909–912
- Blanco R, Joho H, Jatowt A, Yu H, Yamamoto S (2017) Overview of NTCIR-13 actionable knowledge graph (AKG) task. In: Proceedings of NTCIR-13, pp 340–345
- Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Proceedings of ACM SIGIR 2004, pp 25–32
- Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: Proceedings of ICML 2005, pp 89–96
- Carterette B, Bennett PN, Chickering DM, Dumais ST (2008) Here or there: preference judgments for relevance. In: Proceedings of ECIR 2008. LNCS, vol 4956, pp 16–27
- Chandar P, Carterette B (2013) Preference based evaluation measures for novelty and diversity. In: Proceedings of ACM SIGIR 2013, pp 413–422
- Chapelle O, Metzler D, Zhang Y, Grinspan P (2009) Expected reciprocal rank for graded relevance. In: Proceedings of ACM CIKM 2009, pp 621–630
- Chen KH, Chen HH (2001) The Chinese text retrieval tasks of NTCIR workshop 2. In: Proceedings of NTCIR-2

- Chen KH, Chen HH, Kando N, Kuriyama K, Lee S, Myaeng SH, Kishida K, Eguchi K, Kim H (2002) Overview of CLIR task at the third NTCIR workshop. In: *Proceedings of NTCIR-3*
- Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: *Proceedings of ACM SIGIR 2008*, pp 659–666
- Cleverdon CW, Mills J, Keen EM (1966) Factors determining the performance of indexing systems; volume 1: Design. Technical report, The College of Aeronautics, Cranfield
- Collins-Thompson K, Bennett P, Diaz F, Clarke CL, Voorhees EM (2014) TREC 2013 web track overview. In: *Proceedings of TREC 2013*
- Eguchi K, Oyama K, Ishida E, Kando N, Kuriyama K (2003) Overview of the web retrieval task at the third NTCIR workshop. In: *Proceedings of NTCIR-3*
- Eguchi K, Oyama K, Aizawa A, Ishikawa H (2004) Overview of the information retrieval task at NTCIR-4 WEB. In: *Proceedings of NTCIR-4*
- Fujii A, Iwayama M, Kando N (2004) Overview of patent retrieval task at NTCIR-4. In: *Proceedings of NTCIR-4*
- Fujii A, Iwayama M, Kando N (2005) Overview of patent retrieval task at NTCIR-5. In: *Proceedings of NTCIR-5*
- Fujii A, Iwayama M, Kando N (2007) Overview of the patent retrieval task at the NTCIR-6 workshop. In: *Proceedings of NTCIR-6*, pp 359–365
- Gey F, Larson R, Kando N, Machado J, Sakai T (2010) NTCIR-GeoTime overview: evaluating geographic and temporal search. In: *Proceedings of NTCIR-8*, pp 147–153
- Gey F, Larson R, Machado J, Yoshioka M (2011) NTCIR9-GeoTime overview: evaluating geographic and temporal search: round 2. In: *Proceedings of NTCIR-9*, pp 9–17
- Golbus PB, Zitouni I, Kim JY, Hassan A, Diaz F (2014) Contextual and dimensional relevance judgments for reusable SERP-level evaluation. In: *Proceedings of WWW 2014*, pp 131–142
- Harman DK (2005) The TREC test collections. In: Voorhees EM, Harman DK (eds) *TREC: experiment and evaluation in information retrieval*, chap 2. The MIT Press, pp 21–52
- Hawking D, Craswell N (2005) The very large collection and web tracks. In: Voorhees EM, Harman DK (eds) *TREC: experiment and evaluation in information retrieval*, chap 2. The MIT Press, pp 200–231
- Hersh W, Buckley C, Leone T, Hickam D (1994) OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: *Proceedings of ACM SIGIR 1994*, pp 192–201
- Higashinaka R, Funakoshi K, Inaba M, Tsunomori Y, Takahashi T, Kaji N (2017) Overview of dialogue breakdown detection challenge 3. In: *Proceedings of dialog system technology challenge 6 (DSTC6) workshop*
- Iwayama M, Fujii A, Kando N, Takano A (2003) Overview of patent retrieval task at NTCIR-3. In: *Proceedings of NTCIR-3*
- Järvelin K, Kekäläinen J (2000) IR evaluation methods for retrieving highly relevant documents. In: *Proceedings of ACM SIGIR 2000*, pp 41–48
- Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. *ACM TOIS* 20(4):422–446
- Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H, Hidaka S (1999) Overview of IR tasks at the first NTCIR workshop. In: *Proceedings of NTCIR-1*, pp 11–44
- Kando N, Kuriyama K, Yoshioka M (2001) Overview of Japanese and English information retrieval tasks (JEIR) at the second NTCIR workshop. In: *Proceedings of NTCIR-2*
- Kato MP, Yamamoto T, Manabe T, Nishida A, Fujita S (2017) Overview of the NTCIR-13 OpenLiveQ task. In: *Proceedings of NTCIR-13*, pp 85–90
- Kato MP, Manabe T, Fujita S, Nishida A, Yamamoto T (2018) Challenges of multileaved comparison in practice: lessons from NTCIR-13 OpenLiveQ task. In: *Proceedings of ACM CIKM 2018*, pp 1515–1518
- Kato MP, Nishida A, Manabe T, Fujita S, Yamamoto T (2019) Overview of the NTCIR-14 OpenLiveQ-2 task. In: *Proceedings of NTCIR-14*, pp 81–89

- Kishida K (2005) Property of average precision and its generalization: an examination of evaluation indicator for information retrieval. Technical report NII-2005-014E. National Institute of Informatics
- Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH, Myaeng SH, Eguchi K (2004) Overview of CLIR task at the fourth NTCIR workshop. In: Proceedings of NTCIR-4
- Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH, Myaeng SH (2005) Overview of the CLIR task at the fifth NTCIR workshop (revised version). In: Proceedings of NTCIR-5
- Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH (2007) Overview of the CLIR task at the sixth NTCIR workshop. In: Proceedings of NTCIR-6, pp 1–19
- Korfhaage RR (1997) Information storage and retrieval. Wiley, New Jersey
- Liu Y, Song R, Zhang M, Dou Z, Yamamoto T, Kato M, Ohshima H, Zhou K (2014) Overview of the NTCIR-11 IMine task. In: Proceedings of NTCIR-11, pp 8–23
- Luo C, Sakai T, Liu Y, Dou Z, Xiong C, Xu J (2017) Overview of the NTCIR-13 we want web task. In: Proceedings of NTCIR-13, pp 394–401
- Maddalena E, Roitero K, Demartini G, Mizzaro S (2017) Considering assessor agreement in IR evaluation. In: Proceedings of ACM ICTIR 2017, pp 75–82
- Mao J, Sakai T, Luo C, Xiao P, Liu Y, Dou Z (2019) Overview of the NTCIR-14 we want web task. In: Proceedings of NTCIR-14, pp 455–467
- Oyama K, Eguchi K, Ishikawa H, Aizawa A (2004) Overview of the NTCIR-4 WEB navigational retrieval task 1. In: Proceedings of NTCIR-4
- Pollock SM (1968) Measures for the comparison of information retrieval systems. *Am Docum* 19(4):387–397
- Sakai T (2007a) Alternatives to bpref. In: Proceedings of ACM SIGIR 2007, pp 71–78
- Sakai T (2007b) On penalising late arrival of relevant documents in information retrieval evaluation with graded relevance. In: Proceedings of EVIA 2007, pp 32–43
- Sakai T (2007c) On the properties of evaluation metrics for finding one highly relevant document. *IPJSJ Digital Courier* 3:643–660. <https://doi.org/10.2197/ipsjdc.3.643>
- Sakai T (2014) Metrics, statistics, tests. In: PROMISE winter school 2013: bridging between information retrieval and databases. LNCS, vol 8173, pp 116–163
- Sakai T (2017) Unanimity-aware gain for highly subjective assessments. In: Proceedings of EVIA 2017, pp 39–42
- Sakai T (2018) Comparing two binned probability distributions for information access evaluation. In: Proceedings of ACM SIGIR 2018, pp 1073–1076
- Sakai T (2019) Graded relevance assessments and graded relevance measures of NTCIR: a survey of the first twenty years. Technical report, <https://arxiv.org/pdf/1903.11272>
- Sakai T, Kando N (2008) On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf Retrieval* 11:447–470
- Sakai T, Robertson S (2008) Modelling a user population for designing information retrieval metrics. In: Proceedings of EVIA 2018, pp 30–41
- Sakai T, Zeng Z (2019) Which diversity evaluation measures are “good”? In: Proceedings of ACM SIGIR 2019, pp 595–604
- Sakai T, Kitani T, Ogawa Y, Ishikawa T, Kimoto H, Keshi I, Toyoura J, Fukushima T, Matsui K, Ueda Y, Tokunaga T, Tsuruoka H, Nakawatase H, Agata T, Kando N (1999) BMIR-J2: a test collection for evaluation of Japanese information retrieval systems 33(1):13–17
- Sakai T, Kando N, Lin CJ, Mitamura T, Shima H, Ji D, Chen KH, Nyberg E (2008) Overview of the NTCIR-7 ACLIA IR4QA task. In: Proceedings of NTCIR-7, pp 77–114
- Sakai T, Ishikawa D, Kando N (2010a) Overview of the NTCIR-8 community QA pilot task (part II): system evaluation. In: Proceedings of NTCIR-8, pp 433–457
- Sakai T, Shima H, Kando N, Song R, Lin CJ, Mitamura T, Sugimoto M, Lee CW (2010b) Overview of NTCIR-8 ACLIA IR4QA. In: Proceedings of NTCIR-8, pp 63–93
- Sakai T, Ferro N, Soboroff I, Zeng Z, Xiao P, Maistro M (2019) Overview of the NTCIR-14 CENTRE task. In: Proceedings of NTCIR-14, pp 494–509

- Shang L, Sakai T, Lu Z, Li H, Higashinaka R, Miyao Y (2016) Overview of the NTCIR-12 short text conversation task. In: Proceedings of NTCIR-12, pp 473–484
- Shang L, Sakai T, Li H, Higashinaka R, Miyao Y, Arase Y, Nomoto M (2017) Overview of the NTCIR-13 short text conversation task. In: Proceedings of NTCIR-13, pp 194–210
- Song R, Zhang M, Sakai T, Kato MP, Liu Y, Sugimoto M, Wang Q, Orii N (2011) Overview of the NTCIR-9 INTENT task. In: Proceedings of NTCIR-9, pp 82–105
- Yamamoto T, Liu Y, Zhang M, Dou Z, Zhou K, Markov I, Kato MP, Ohshima H, Fujita S (2016) Overview of the NTCIR-12 IMine-2 task. In: Proceedings of NTCIR-12, pp 8–26
- Yasukawa M, Diaz F, Druck G, Tsukada N (2014) Overview of the NTCIR-11 Cooking recipe search task. In: Proceedings of NTCIR-11, pp 483–496
- Zanibbi R, Aizawa A, Kohlhase M, Ounis I, Topić G, Davila K (2016) NTCIR-12 MathIR task overview. In: Proceedings of NTCIR-12, pp 299–308
- Zeng Z, Kato S, Sakai T (2019) Overview of the NTCIR-14 short text conversation task: dialogue quality and nugget detection subtasks. In: Proceedings of NTCIR-14, pp 289–315
- Zhang Y, Huang M (2019) Overview of the NTCIR-14 short text generation subtask: emotion generation challenge. In: Proceedings of NTCIR-14, pp 316–327

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.