The Impact of Intent Selection on Diversified Search Evaluation

Tetsuya Sakai
Microsoft Research Asia,
P.R.C.
tetsuyasakai@acm.org

Zhicheng Dou
Microsoft Research Asia,
P.R.C.
zhichdou@microsoft.com

Charles L. A. Clarke
University of Waterloo,
Canada
claclark@plg.uwaterloo.ca

ABSTRACT

To construct a diversified search test collection, a set of possible *subtopics* (or *intents*) needs to be determined for each topic, in one way or another, and per-intent relevance assessments need to be obtained. In the TREC Web Track Diversity Task, subtopics are manually developed at NIST, based on results of automatic click log analysis; in the NTCIR INTENT Task, intents are determined by manually clustering "subtopics strings" returned by participating systems. In this study, we address the following research question: *Does the choice of intents for a test collection affect relative performances of diversified search systems?* To this end, we use the TREC 2012 Web Track Diversity Task data and the NTCIR-10 INTENT-2 Task data, which share a set of 50 topics but have different intent sets. Our initial results suggest that the choice of intents may affect relative performances, and that this choice may be far more important than how many intents are selected for each topic.

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

diversity, evaluation, intents, subtopics, test collections

1. INTRODUCTION

Given an ambiguous or underspecified query, diversified search aims to cover different possible search intents with a single search engine result page, by balancing relevance and diversity. TREC[1] started a Diversity Task in the Web Track[2] in 2009, while NTCIR[3] started a related task called INTENT[4] in 2011. Unlike traditional retrieval evaluation where pooled documents are assessed in terms of relevance with respect to each *topic*, diversity evaluation requires a set of *subtopics* (or *intents*) for each topic, and pooled documents are assessed with respect to each intent. In the TREC Diversity Task, subtopics are manually developed at NIST, based on results of automatic click log analysis [4]; in the NTCIR INTENT Task, intents are determined by manually clustering "subtopics strings" returned by participating systems [6, 7]. However, it is difficult to say exactly what the most appropriate intents are for a given topic for the purpose of evaluating diversified search.

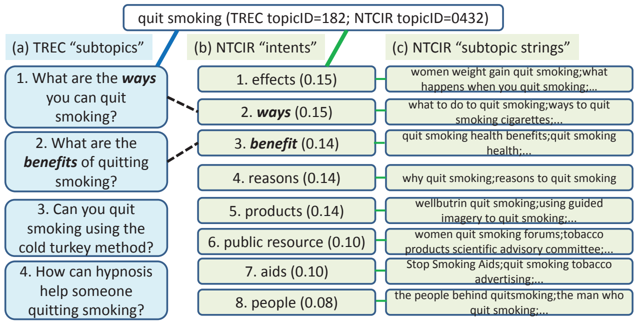
quit smoking (TREC topicID=182; NTCIR topicID=0432)
(a) TREC "subtopics"
(b) NTCIR "intents"
(c) NTCIR "subtopic strings"
1. What are the ***ways*** you can quit smoking?
2. What are the ***benefits*** of quitting smoking?
3. Can you quit smoking using the cold turkey method?
4. How can hypnosis help someone quitting smoking?
1. effects (0.15)
2. ***ways*** (0.15)
3. ***benefit*** (0.14)
4. reasons (0.14)
5. products (0.14)
6. public resource (0.10)
7. aids (0.10)
8. people (0.08)
women weight gain quit smoking;what happens when you quit smoking;...
what to do to quit smoking;ways to quit smoking cigarettes;...
quit smoking health benefits;quit smoking health;...
why quit smoking;reasons to quit smoking
wellbutrin quit smoking;using guided imagery to quit smoking;...
women quit smoking forums;tobacco products scientific advisory committee;...
Stop Smoking Aids;quit smoking tobacco advertising;...
the people behind quitsmoking;the man who quit smoking;...

Figure 1: TREC subtopics vs. NTCIR intents and subtopic strings.

One may be tempted to hypothesise that an effective diversified search system should be effective regardless of the particular choice of intents used to evaluate it. Thus, in this study, we address the following research question: *Does the choice of intents for a test collection affect relative performances of diversified search systems?* To this end, we use the TREC 2012 Web Track Diversity Task data and the NTCIR-10 INTENT-2 Task data, which share a set of 50 topics but have different intent sets [6, 7]. Figure 1 provides an actual example from the 50 topics we used in our experiments: for this topic ("quit smoking"), four subtopics were obtained for TREC at NIST as shown in Figure 1(a), by following the aforementioned click-based methodology; whereas, at NTCIR, subtopic strings were returned by the participating system of the INTENT-2 Subtopic Mining Subtask as shown in Figure 1(c), which were then manually clustered and filtered to form a set of eight *intents* as shown in Figure 1(b)[5]. As (a) and (b) were obtained completely independently using different methods, they obviously differ, although they may partially overlap as indicated by the dotted lines in the figure. For example, the first TREC subtopic for this topic is "What are the ways you can quit smoking?" which is probably similar to the second NTCIR intent "(quit smoking) ways."

[1] `http://trec.nist.gov/`

[2] `http://plg.uwaterloo.ca/~trecweb/`

[3] `http://research.nii.ac.jp/ntcir/`

[4] `http://research.microsoft.com/INTENT/`

[5] As shown in the figure, the INTENT task also estimates the probability of each intent given the query based on assessor voting. However, following the practice at TREC, we assume that the probability distribution is uniform across all intents throughout this study.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Table 1: Test collection and pseudo-qrels statistics. Eight TREC intents with no relevant documents have been removed. In Part (b), statistics for the *truncated* pseudo-qrels are shown in parentheses.

	(a) TREC 2012 Diversity	(b) English INTENT-2 and pseudo-qrels derived
topics	50 (provided from TREC to NTCIR)	
intents/topic	3.7	all: 7.8; matched: 7.7 (3.7)
subtopic strings/topic	–	82.7
pooled non-junk docs/topic	303.9	–
unique relevant/topic	111.2	287.0 (263.8)
4-relevant/topic	49.7	–
3-relevant/topic	2.6	11.8 (8.8)
2-relevant/topic	23.5	289.7 (182.7)
1-relevant/topic	111.6	881.7 (435.3)

```
foreach topic $t$ do {
    foreach NTCIR intent $i$ for $t$ do {
        foreach TREC pooled document $d$ for $t$ do {
            $matchcount = 0$;
            foreach reduced subtopic $s$ for $i$ do {
                if $d$ contains $s$ by exact match then $matchcount++$;
            }
            $relevancelevel(d) = \max(0, trunc(\log(matchount) + 1))$;
            //the function $trunc$ takes the integer part of the argument.
    }
}
```

Figure 2: Algorithm for automatically generating pseudo-qrels.

This intent was devised at NTCIR based on a cluster of subtopic strings obtained from participating systems, including "what to do quit smoking" and "ways to quit smoking cigarettes" amongst others. Hereafter, we shall also refer to TREC subtopics as "intents" to avoid confusion.

To address the above research question, we replace the TREC intents with the NTCIR intents and then re-evaluate the runs submitted to the TREC 2012 diversity task. Unfortunately, while the NTCIR-10 INTENT-2 Task had *Document Ranking* (i.e., diversified search) subtasks for Chinese and Japanese, it only had a *Subtopic Mining* subtask for English [6, 7], and therefore the English intents from NTCIR lack document relevance assessments. While it would be ideal to actually conduct relevance assessments of the TREC pooled documents with respect to each NTCIR intent, we explore a cheaper alternative in this paper, namely, to automatically construct *pseudo-qrels* by simply matching the TREC pooled documents against the NTCIR subtopic strings[6]. While the lack of *true* relevance assessments for the NTCIR intents is a limitation of this study, our pseudo-qrels do provide partial answers to our research question: our initial results suggest that the choice of intents may in fact affect relative performances, and that this choice may be more important than how many intents are selected for each topic.

2. EXPERIMENTAL SETTING

2.1 Data

Table 1(a) shows some statistics of the TREC 2012 Web Diversity topics, after having removed eight intents from the original `qrels.diversity` file as they did not have any relevant documents. At TREC, there were six relevance levels: 4 ("navigational"), 3 ("key"), 2 ("highly relevant"), 1 ("relevant"), 0 ("nonrelevant") and -2 ("junk") [4]. The table refers to the first four as 4-, 3-, 2- and 1-relevant, respectively; we treat the other two as nonrelevant. Also, as shown in the table, we had 303.9 pooled non-junk documents per topic on average, which we obtained from `qrels.diversity`[7].

To re-evaluate the TREC 2012 diversity runs after replacing the TREC intents with the NTCIR intents, we created pseudo-qrels as follows: The original NTCIR intents have an average of 7.8 intents and 82.7 subtopic strings per topic (recall Figure 1(b) and (c)). To obtain *pseudo-relevant* documents from the TREC pools for each NTCIR intents, we first removed the topic string from each NTCIR subtopic string automatically: for example, "women weight gain quit smoking" in Figure 1 was turned into "women weight gain." We call the resultant strings *reduced subtopics*. We then used the simple algorithm shown in Figure 2 to obtain pseudo-relevant documents for each NTCIR intent. Note that each pooled document is tested whether it matches with any of the *reduced* subtopic, under the assumption that pooled documents already contain the actual topic string (e.g., "quit smoking") or some related term. The algorithm also determines the relevance level of each document based on the number of matches with reduced subtopics: the actual number of matches ($matchcount$) varied from 0 to 19; the (natural) log-based function in the algorithm maps them to 0-3. A total of 28 pooled documents were removed during the process, as they have been detected as containing a virus.

Table 1(b) shows some statistics of the NTCIR-10 INTENT-2 intents and the pseudo-qrels we derived from them. Note that, of the 303.9 TREC pooled documents per topic, as many as 293.6 matched with at least one reduced subtopic and are treated as relevant. This strongly suggests that our pseudo-qrels contain a lot of false matches. Moreover, because of this problem, note that the average number of intents per topic in the pseudo-qrels is 7.7, which is easily twice as large as the corresponding number for TREC, namely, 3.7. Thus, if the evaluation outcome with the pseudo-qrels is different from that with the true qrels, this may be because either (i) the two intent sets contain different intents; or (ii) the two intent sets differ in size (the NTCIR intents sets are larger so may require systems to diversify more aggresively); or both.

In order to separate the above two effects, we also created another version of pseudo-qrels, called *truncated* pseudo-qrels (or simply "truncated" for short). This was done by cutting down "less popular" intents from the original pseudo-qrels to ensure that the TREC and NTCIR intent sets are equal in size for each topic. For the example shown in Figure 1, although the original pseudo-qrels has eight intents, the truncated pseudo-qrels has only the first four intents with the highest intent probabilities. The statistics for the truncated pseudo-qrels are shown in parentheses in Table 1(b).

2.2 Evaluation Metrics and Analysis Methods

We primarily consider four diversity evaluation metrics: *D-nDCG*, *D$\sharp$-nDCG* [8], *α-nDCG* and *ERR-IA* [3]. D-nDCG is a version of *normalised Cumulative Discounted Gain (nDCG)* [5] which combines per-intent graded relevance and intent probabilities to compute the gain value of each document. D$\sharp$-nDCG is a simple average of D-nDCG and *intent recall (I-rec)*, a.k.a. subtopic recall [11]. D$\sharp$-nDCG summarises a graph that plots D-nDCG (i.e. overall relevance) against I-rec (pure diversity). α-nDCG is a version of nDCG which defines graded relevance as the number of intents covered by a document, and discounts the value of a retrieved relevant document for each intent based on relevant documents already seen. This property is known as *diminishing return* [2]. ERR-IA first

[6] In TREC parlance, *qrels* means relevance assessments. When qrels are obtained automatically without involving manual relevance assessments, they are often referred to as *pseudo-qrels* [9].

[7] At TREC 2012, a common pool was created across the *ad hoc* task and the diversity task for each topic. Hence, the pooled documents obtained from `qrels.diversity` are identical to those from `qrels.adhoc`.

computes an *Expected Reciprocal Rank* value for each intent and then combines them across intents. It also possesses the diminishing return property. Both α-nDCG and ERR-IA may be expressed in terms of a common framework, differing primarily in the discounts they apply for document rank [3].

The NTCIR INTENT task uses I-rec, D-nDCG and D$\sharp$-nDCG as the primary metrics for ranking the runs. Following the task's practice, we compute the values using `NTCIREVAL`[8], by using the relevance levels as the gain values. However, it should be noted that I-rec is not a good stand-alone metric for our purpose: although we measure performance at document cutoffs of 10 and 20 (denoted by "@10" and "@20"), recall that the average number of intents per topic with the true qrels is only 3.7: thus it should be fairly easy for systems to cover most intents, especially with 20 documents. Furthermore, I-rec does not work well with our pseudo-qrels: because of the aforementioned false match problem, I-rec is heavily overestimated for all of the TREC runs when the pseudo-qrels are used. Nevertheless, we include the results with I-rec for separating the effects of diversity and relevance in diversified search evaluation [8].

The TREC Web Track Diversity Task uses α-nDCG and ERR-IA along with some other metrics. Following the practice at TREC, we computed these metrics using `ndeval`[9]. It should be noted that, while `NTCIREVAL` utilises the per-intent graded relevance data to compute D($\sharp$)-nDCG, `ndeval` reduces the data to per-intent binary relevance data before computing α-nDCG and ERR-IA. Thus the computation of relevance levels in Figure 2 does not affect these two metrics.

In order to compare the relative performances of the TREC 2012 diversity runs before and after replacing the original TREC intents with the NTCIR ones, we compare the run rankings in terms Kendall's τ, and its variant called τ_{ap} [10]. These measures count the number of pairwise system swaps; τ_{ap} is more sensitive to the swaps near the top ranks than τ is. However, what is perhaps more important is whether replacing the intent sets affects statistical significance testing, which is often used for forming research conclusions in the IR community. We therefore conduct a randomised version of the two-sided *Tukey's Honestly Significantly Different (HSD) test* [1] at $\alpha = .05$ for the entire set of runs before and after replacing the intent sets. Given the entire set of runs, this kind of test is more appropriate than those that test one run pair at a time while ignoring the others. We then compare the two sets of significantly different run pairs. For example, is a significantly different run pair obtained according to the TREC intents still significantly different according to the NTCIR intents with its pseudo-qrels?

3. RESULTS AND DISCUSSIONS

Table 2 shows the τ and τ_{ap} between rankings produced by two different metrics based on the true qrels, to show how the diversity metrics behave differently. Table 3 is more important for our purpose: for each metric, the τ and τ_{ap} for the ranking with the true qrels and that with the pseudo-qrels are shown. It can be observed that the rankings with the pseudo-qrels (i.e., those based on the NTCIR intents) are quite different from those with the true qrels (i.e., those based on the TREC intents). This is true even for the truncated pseudo-qrels, as shown in Part (b) of the table, which suggests that the discrepancies between TREC and NTCIR may arise not from how many intents are used but from the actual choice of intents.

Table 2: τ/τ_{ap} between two metrics using true qrels (20 TREC 2012 diversity runs).

@10	D-nDCG	D$\sharp$-nDCG	α-nDCG	ERR-IA
I-rec	.347/.323	.642/.493	.547/.451	.568/.496
D-nDCG	-	.705/.780	.695/.719	.674/.700
D$\sharp$-nDCG	-	-	.779/.769	.695/.706
α-nDCG	-	-	-	.895/.895
@20	D-nDCG	D$\sharp$-nDCG	α-nDCG	ERR-IA
I-rec	.179/.319	.705/.661	.495/.537	.453/.522
D-nDCG	-	.474/.580	.621/.656	.600/.637
D$\sharp$-nDCG	-	-	.705/.745	.579/.635
α-nDCG	-	-	-	.853/.855

Table 3: τ/τ_{ap} between rankings by the same metric using true and (truncated) pseudo-qrels (20 TREC 2012 diversity runs).

		@10	@20
(a) true vs. pseudo	I-rec	.632/.392	.568/.302
	D-nDCG	.653/.698	.684/.692
	D$\sharp$-nDCG	.611/.651	.632/.652
	α-nDCG	.674/.673	.716/.704
	ERR-IA	.589/.611	.589/.614
(b) true vs. truncated	I-rec	.579/.317	.526/.271
	D-nDCG	.621/.660	.653/.666
	D$\sharp$-nDCG	.684/.682	.695/.668
	α-nDCG	.663/.668	.716/.706
	ERR-IA	.579/.614	.579/.616

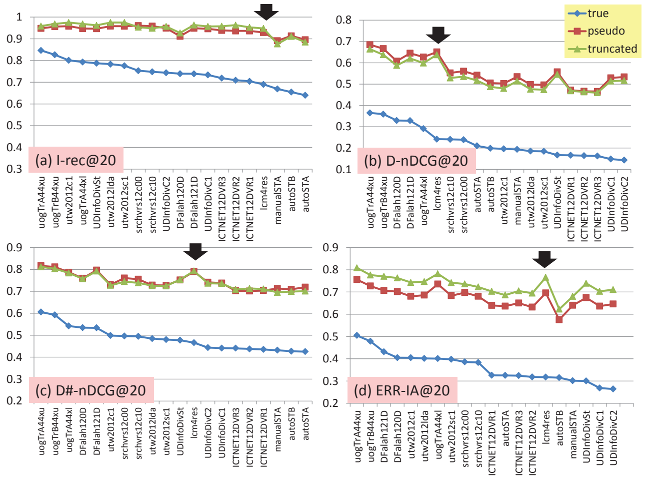

Figure 3: Run rankings: true vs. pseudo vs. truncated. The x axis represents runs sorted by a metric with true relevance data from TREC.

Figure 3 visualises the "@20" column of Table 3 for selected metrics. Recall that I-rec is a pure diversity metric; that D-nDCG is an overall relevance metric; and that D$\sharp$-nDCG and ERR-IA consider both aspects. It can be observed that I-rec with pseudo-qrels is almost completely useless for ranking runs. On the other hand, D-nDCG with pseudo-qrels does better: for example, the top two runs in terms of D-nDCG with the true qrels (uogTrA44xu and uogTrB44xu) are still the top two in terms of D-nDCG with the (truncated) pseudo-qrels. The same two runs are also top performers in terms of D$\sharp$-nDCG as well, regardless of the qrels being used. As for ERR-IA, while the same two runs are the top performer in terms of the true qrels, the second run uogTrB44xu is ranked third with the (truncated) pseudo-qrels. To sum up, while our pseudo-qrels cannot properly estimate systems's intent recall, the top run at TREC, namely, uogTrA44xu, is still the top run when evaluated with D($\sharp$)-nDCG and ERR-IA based on the NTCIR intents and the pseudo-qrels. However, the overall rankings do differ when the TREC intents are replaced with those from NTCIR. Again, since the graphs for the original and truncated pseudo-qrels behave very

[8] `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`

[9] `http://trec.nist.gov/data/web2012.html`

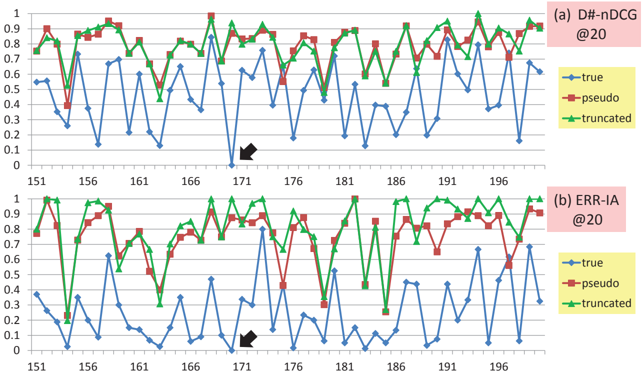

Figure 4: Per-topic performance values for lcm4res.

similarly, the discrepancies between TREC and NTCIR are probably due to the *choice* of intents.

In Figure 3, as indicated by the arrows, Run lcm4res is heavily overestimated with D(♯)-nDCG and ERR-IA based on the pseudo-qrels. Figure 4 provides a per-topic diagnosis for this run with D♯-nDCG and ERR-IA, which reveals that the pseudo-qrels overestimate the run's performance for almost all topics. Perhaps the worst case is Topic 170 ("scooters"), for which there are three TREC intents and eight NTCIR intents: even though the true D♯-nDCG and ERR-IA values are zero (as indicated by the arrows), the corresponding values with the pseudo-qrels are .8698 and .8750 (.9378 and 1 when truncated). As we have not conducted relevance assessments, we cannot rule out the possibility that this run actually retrieved many documents that are relevant to the NTCIR intents yet nonrelevant to the TREC intents for this topic. However, the overall trend across the topics strongly suggests that our pseudo-qrels do not provide accurate estimates of intent recall. We leave the analysis of the TREC runs using *true* relevance assessments for the NTCIR intents for future work.

We now discuss the effect of replacing the TREC intent sets with the NTCIR ones on statistical significance testing: Table 4 summarises the results. Note that, if the significance test results with true and pseudo-qrels are identical, the number of significantly different pairs in the TR, PS and TR∩PS will be the same, and that the TR−PS and PS−TR will contain zeroes. Such is not the case. That is, conclusions drawn from an experiment based on the original TREC intents and those drawn from one based on the intents derived from NTCIR can be quite different. For example, in Table 4(b), ERR-IA@20 obtains 31 significantly different run pairs with the true qrels and 18 significantly different run pairs with the pseudo-qrels; but only 9 pairs overlap. Again, truncating the pseudo-qrels (Table 4(c)(d)) does not seem to solve any problems, which again suggests that the choice of intents do matter for the purpose of comparing diversified search systems.

4. CONCLUSIONS AND FUTURE WORK

We addressed the following research question: *Does the choice of intents for a test collection affect relative performances of diversified search systems?* To this end, we used the TREC 2012 Web Track Diversity Task data and the NTCIR-10 INTENT-2 Task data, which share a set of 50 topics but have different intent sets. Our initial results suggest that the choice of intents may in fact affect relative performances, and that this choice may be more important than how many intents are selected for each topic.

One limitation of the present work is that we used automatically-generated pseudo-qrels for the NTCIR intents instead of conducting relevance assessments of TREC pooled documents for the NTCIR intents. In particular, we found that the pseudo-qrels estimate intent recall very poorly. On the other hand, we have also found that the official top performer at the TREC 2012 diversity task is still the top performer even after the intent sets have been replaced with the ones from NTCIR. In order to obtain a more clear answer to our research question, we hope to come back to it with true relevance assessments for the NTCIR intents.

Table 4: Significance test concordances and discordances between true qrels and pseudo-qrels (190 TREC 2012 diversity run pairs; randomised two-sided Tukey's HSD test at $\alpha = .05$). TR (PS): significant differences obtained with true qrels (pseudo-qrels); TR−PS (PS−TR): pairs significant with true qrels (pseudo-qrels) but not significant with pseudo-qrels (true qrels); TR∩PS: pairs significant with both true qrels and pseudo-qrels.

		TR	PS	TR−PS	TR∩PS	PS−TR
(a) true	I-rec	12	21	6	6	15
vs.	D-nDCG	51	56	9	42	14
pseudo	D♯-nDCG	25	29	9	16	13
@10	α-nDCG	26	24	16	10	14
	ERR-IA	29	19	20	9	10
(b) true	I-rec	9	11	7	2	9
vs.	D-nDCG	60	60	15	45	15
pseudo	D♯-nDCG	35	40	9	26	14
@20	α-nDCG	26	24	15	11	13
	ERR-IA	31	18	22	9	9
(c) true	I-rec	12	11	8	4	7
vs.	D-nDCG	51	47	16	35	12
truncated	D♯-nDCG	25	14	11	14	0
@10	α-nDCG	26	9	24	2	7
	ERR-IA	29	9	25	4	5
(d) true	I-rec	9	28	3	6	22
vs.	D-nDCG	60	55	20	40	15
truncated	D♯-nDCG	35	35	12	23	12
@20	α-nDCG	26	13	24	2	11
	ERR-IA	31	10	27	4	6

5. REFERENCES

[1] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, 30(1), 2012.

[2] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: Metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

[3] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of ACM WSDM 2011*, pages 75–84, 2011.

[4] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 web track. In *Proceedings of TREC 2012*, 2013.

[5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4):422–446, 2002.

[6] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, M. P. Kato, R. Song, and M. Iwata. Overview of the NTCIR-10 INTENT-2 task. In *Proceedings of NTCIR-10*, 2013.

[7] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, M. P. Kato, R. Song, and M. Iwata. Summary of the NTCIR-10 INTENT-2 task: Subtopic mining and search result diversification. In *Proceedings of ACM SIGIR 2013*, 2013.

[8] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of ACM SIGIR 2011*, pages 1043–1042, 2011.

[9] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of ACM SIGIR 2001*, pages 66–73, 2001.

[10] E. Yilmaz, J. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of ACM SIGIR 2008*, pages 587–594, 2008.

[11] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of ACM SIGIR 2003*, pages 10–17, 2003.