Summary of the NTCIR-10 INTENT-2 Task: Subtopic Mining and Search Result Diversification

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

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

Takehiro Yamamoto
Kyoto University, Japan
tyamamot@dl.kuis.kyoto-u.ac.jp

Yiqun Liu
Tsinghua University, P.R.C.
yiqunliu@tsinghua.edu.cn

Min Zhang
Tsinghua University, P.R.C.
z-m@tsinghua.edu.cn

Makoto P. Kato
Kyoto University, Japan
kato@dl.kuis.kyoto-u.ac.jp

ABSTRACT

The NTCIR INTENT task comprises two subtasks: *Subtopic Mining*, where systems are required to return a ranked list of *subtopic strings* for each given query; and *Document Ranking*, where systems are required to return a diversified web search result for each given query. This paper summarises the novel features of the Second INTENT task at NTCIR-10 and its main findings, and poses some questions for future diversified search evaluation.

Categories and Subject Descriptors

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

Keywords

diversity, evaluation, intents, subtopics, test collections

1. INTRODUCTION

NTCIR (NII Testbeds and Community for Information access Research)[1] is a sesquiannual evaluation forum that focusses primarily on Asian language information access. The INTENT task[2], launched at NTCIR-9 [16], is closely related to the TREC Web Diversity Task [4]. The *Second* INTENT task (INTENT-2) was concluded at the NTCIR-10 conference in June 2013 [12]. This paper summarises the novel features of the task and its main findings, and poses some questions for future diversified search evaluation.

Figure 1 outlines the INTENT task. In the *Subtopic Mining* (SM) subtask, participants are asked to return a ranked list of *subtopic strings* for each query from the topic set (Arrows 1 and 2), where a subtopic string is *a query that specialises and/or disambiguates the search intent of the original query*. The organisers create a pool of these strings for each query, and ask the assessors to manually *cluster* them, and to provide a label for each cluster. Then the organisers determine a set of important search *intents* for each query, where each intent is represented by a cluster label with its cluster of subtopics (Arrows 3 and 4). Then they ask ten assessors to vote whether each intent is important or not for a given query; and based on the votes compute the intent probabilities (Arrows 5 and 6) [16]. The SM runs are then evaluated using the intents with their associated probabilities and subtopic strings. This subtask can be regarded as a component of a search result diversification system, but other applications such as query suggestion and completion are also possible.

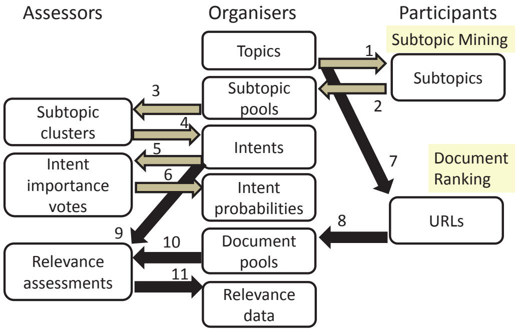

Figure 1: Structure of the INTENT task.

Table 1: Number of INTENT-2 runs (teams). E: English; C: Chinese; J: Japanese.

Subtopic Mining			Document Ranking	
E	C	J	C	J
34 (8)	23 (6)	14 (3)	12 (3)	8 (2)

The black arrows in Figure 1 show the flow of the *Document Ranking* (DR) subtask, which is similar to the TREC Diversity Task [4]. Participants are asked to return a diversified ranked list of URLs for each query from the aforementioned topic set (Arrows 7 and 8). The organisers create a pool of the URLs for each query, ask the assessors to conduct graded relevance assessments *for each intent* of each query, and consolidate the relevance assessments to form the final graded relevance data (Arrows 9, 10 and 11) [16]. The DR runs are evaluated using the intents, their probabilities and the relevance data. The aim of search result diversification is to maximise both the relevance and diversity of the first search engine result page, given a query that is *ambiguous* or *underspecified*.

Table 1 shows the number of runs submitted to and the number of teams that participated in the INTENT-2 task. The English SM subtask was introduced at INTENT-2, by using the TREC 2012 Web topics kindly provided by the Web Track coordinators.

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

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

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

ACM 978-1-4503-2034-4/13/07 ...$15.00.

2. WHAT'S NEW AT INTENT-2

For the general task design of INTENT, we refer the reader to the *INTENT-1* overview paper [16]. New features of INTENT-2 are as follows.

(I) As we have mentioned earlier, we introduced an *English* SM subtask using the 50 TREC 2012 Web Track topics. While the TREC "subtopics" were created at NIST [4][3], we independently created a set of intents for each topic at INTENT-2 from subtopic pools, as shown in Figure 1[4].

(II) We provided an "official" set of search engine query suggestions for each query to participants, to improve the reproducibility and fairness of experiments. (At INTENT-1, different teams scraped their own versions of query suggestions from different search engines.)

(III) For the Chinese and Japanese topic sets, we provided a baseline non-diversified run and the corresponding web page contents to participants[5]. This enables researchers to isolate the problem of diversifying a given search result from that of producing an effective initial search result.

(IV) We included single-intent *navigational* topics in the Chinese and Japanese topic sets. A navigational topic should require one answer or one website, and therefore may not require diversification. We thereby encouraged participants to explore *selective diversification* [15]: instead of uniformly applying a diversification algorithm to all topics, determine in advance which topics will (not) benefit from diversification. Moreover, we tagged each intent with either *informational* or *navigational* based on five assessors' votes: each assessor judged each intent independently by referring to a specific guideline we provided, and an intent was tagged with *navigational* only if four or five assessors judged it to be navigational. This enables us to conduct *intent type-sensitive diversification* [8], whose rationale is that returning redundant results to navigational intents should not be rewarded, and that more space should be allocated to informational intents.

(V) All participants were asked to produce results not only for the INTENT-2 topics but also for the INTENT-1 topics. Moreover, participants who also participated in INTENT-1 were encouraged to submit "Revived Runs" to INTENT-2, using their systems from INTENT-1. This is to monitor progress across NTCIR rounds. Figure 2 outlines this effort. By comparing the new INTENT-2 runs with the Revived Runs (arrows (a) vs. (d)), we can discuss progress across the two rounds; by comparing the performance over the INTENT-1 topic set and that over the INTENT-2 topic set for each of the Revived Runs (arrows (c) vs. (d)), we can discuss the comparability of the two test collections [7][6]. Unfortunately, no Revived Run was submitted to the SM subtask, so we can only conduct this analysis for the DR subtask.

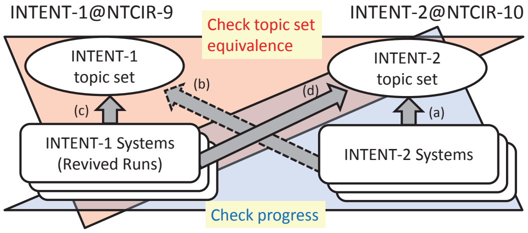

Figure 2: Comparing INTENT-1 and INTENT-2.

Table 2: Statistics of the INTENT-2 topics and intents.

		Subtopic Mining	Document Ranking
English	topics	50	–
	intents	392	–
	subtopic strings	4,157	–
Chinese	topics	98	97
	nav topics	23	22
	intents	616	615
	nav intents	–	125
	inf intents	–	490
	subtopic strings	6,251	–
	unique rel docs	–	9,295
Japanese	topics	100	95
	nav topics	33	28
	intents	587	582
	nav intents	–	259
	inf intents	–	323
	subtopic strings	2,979	–
	unique rel docs	–	5,085

Table 3: INTENT-2 relevance assessment statistics.

	Chinese (97 topics)	Japanese (95 topics)
$L4$	224	1,596
$L3$	613	1,545
$L2$	7,265	2,779
$L1$	6,667	3,824
total	14,769	9,744

Tables 2 and 3 provide some statistics of the INTENT-2 test collections that we have constructed. Our test collections contain per-intent graded relevance assessments that reflect two assessors' judgments: for example, $L4$ is the highest relevance level, which means that two assessors each assigned a score of two to the document [16]. As for the INTENT-2 document collections, they are the same as the ones used at INTENT-1 [16]: the SogouT corpus[7] and the Japanese portion of ClueWeb09[8].

3. EVALUATION METRICS

The primary evaluation metrics we use (for both SM and DR subtasks) are *intent recall* ("I-rec") and *D(♯)-nDCG*. I-rec (a.k.a. *subtopic recall* [17]) is simply the proportion of intents covered by a system output. D-nDCG is a diversity version of the well-known *normalised discounted cumulative gain* [5], where a gain value of a relevant document is computed as the "Global Gain" (GG) across all known intents. At INTENT-2, the "local" (i.e. per-intent) gain values are set to 4, 3, 2 and 1 for each $L4$-, $L3$-, $L2$- and $L1$-relevant document, respectively.

For each participating run, we plot D-nDCG (overall relevance) against I-rec (diversity), and also compute D♯-nDCG as a simple linear combination of I-rec and D-nDCG. Some advantages of the D-measure framework over *α-nDCG* [3] and *intent-aware* metrics [1, 2] have been demonstrated elsewhere [13, 14].

[3] http://trec.nist.gov/data/web/12/full-topics.xml

[4] We will report elsewhere [10] on an analysis across TREC and NTCIR using the common topic set.

[5] We did not do this for English as we had no English DR subtask.

[6] Arrow (b) in Figure 2 means evaluating new runs by reusing the INTENT-1 test collection, but it is known that diversified search test collections are highly unlikely to be reusable mainly due to shallow pooling [9, 11].

[7] http://www.sogou.com/labs/dl/t-e.html

[8] http://lemurproject.org/clueweb09/

The D-measure framework was designed for diversified search, but we used it also for evaluating the ranked list of subtopic strings in the SM subtask. In the case of the SM task, by construction, each subtopic string belongs to exactly one intent and only binary relevance is available. Thus, D-nDCG reduces to standard nDCG where each gain value is exactly the probability of the intent to which the subtopic string belongs.

In the INTENT-2 DR subtask, we additionally used *intent type-sensitive* metrics called *DIN-nDCG* and *P+Q* [8], by leveraging the informational and navigational intent tags. As was mentioned earlier, the rationale is that, since a navigational intent requires exactly one relevant document by definition, evaluation metrics should not reward systems for returning redundant information for such a topic; instead, systems should aim to allocate more space to informational intents.

DIN-nDCG is a simple variant of D-nDCG: for each navigational intent, it treats only the first retrieved relevant document as relevant, and ignores the rest. P+Q is a combination of metrics called P^+ and *Q-measure*: Q is a graded-relevance version of Average Precision and is suitable for informational intents; P^+ is a similar metric but it assumes that no user will go beyond the first most highly relevant document found in the ranked list. It is therefore suitable for navigational intents. P+Q computes a Q value for each informational intent and a P^+ value for each navigational intent, and finally combines the values using the intent-aware approach [1, 2].

In short, compared to the TREC Web Track Diversity Task, the INTENT-2 evaluation framework has the following unique features: (a) we utilise intent probabilities, to encourage systems to allocate more space in the search engine result page to popular intents than to minor ones; (b) we utilise per-intent graded relevance, to encourage retrieval of highly relevant documents over marginally relevant ones; (c) we encourage selective diversification (decide whether or not to diversify per topic) and intent type-aware diversification (do not reward systems for returning redundant information for navigational intents)[9].

4. OFFICIAL RESULTS AND DISCUSSIONS

The official I-rec/D-nDCG graphs, which show the balance between overall relevance and diversity, are shown in Figures 3-6. The Japanese DR results are omitted due to lack of space: we had only two participating teams for this subtask (see Table 1). Below, we summarise our main findings.

English Subtopic Mining (Figure 3) THUIR-S-E-4A outperformed all other runs in terms of Mean D♯-nDCG, but hultech, KLE, ORG (organisers' team), SEM12 and THCIB all have at least one run that is statistically indistinguishable from this top run. Whereas, all runs from LIA and TUTA1 significantly underperformed THUIR-S-E-4A.

Chinese Subtopic Mining (Figure 4) TUTA1-S-C-1A outperformed all other runs in terms of Mean D♯-nDCG, but the six participating teams are statistically indistinguishable from one another.

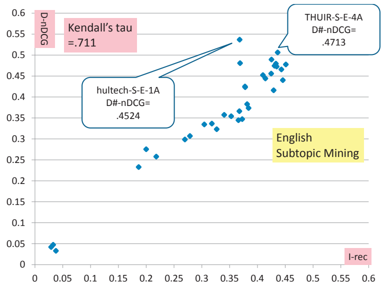

Figure 3: I-rec/D-nDCG graph for English Subtopic Mining.

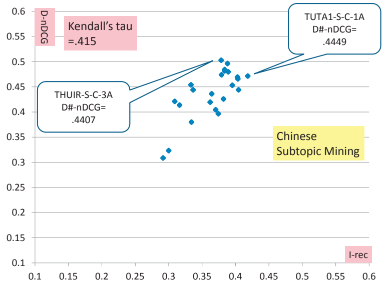

Figure 4: I-rec/D-nDCG graph for Chinese Subtopic Mining.

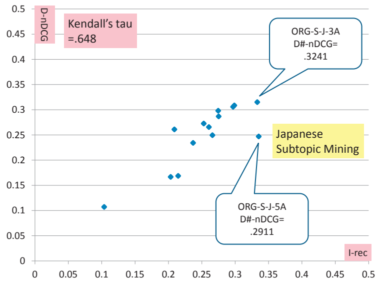

Figure 5: I-rec/D-nDCG graph for Japanese Subtopic Mining.

Japanese Subtopic Mining (Figure 5) ORG-S-J-3A outperformed all other runs in terms of Mean D♯-nDCG, but the three participating teams are statistically indistinguishable from one another.

Chinese Document Ranking (Figure 6) THUIR-D-C-1A outperformed all other runs in terms of Mean D♯-nDCG; it significantly outperformed the baseline nondiversified run. However, KECIR has two runs that are statistically indistinguishable from this top run. Moreover, none of the new runs from THUIR significantly outperformed its Revived Run THUIR-D-C-R1, and therefore this team's progress after INTENT-1 may not be substantial.

Japanese Document Ranking (figure not shown) MSINT-D-J-4B outperformed all other runs in terms of Mean D♯-nDCG. In particular, it significantly outperformed its Revived Runs MSINT-D-J-R1 and MSINT-D-J-R2, by combining multiple search engine results. Thus this marks an actual improvement over INTENT-1.

Navigational Topics The D♯-nDCG values for navigational topics tend to be high for the Chinese/Japanese SM/DR subtasks, as there is only one intent for these topics: for these topics, D-nDCG reduces to nDCG, and retrieving just one relevant document suffices for achieving an I-rec of one. Moreover, the per-topic analysis of the top Document Ranking runs suggests that navigational topics tend to receive high P+Q val-

[9] The TREC 2011 and 2012 diversity test collections have graded relevance assessments; all TREC diversity test collections (2009-2012) have the informational and navigational subtopic tags. However, they have not been utilised for evaluating the runs.

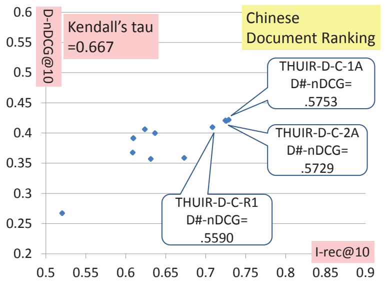

Figure 6: I-rec/D-nDCG graph for Chinese Document Ranking.

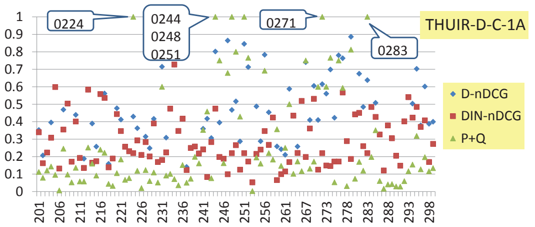

Figure 7: Per-topic D-nDCG/DIN-nDCG/P+Q performances for THUIR-D-C-1A.

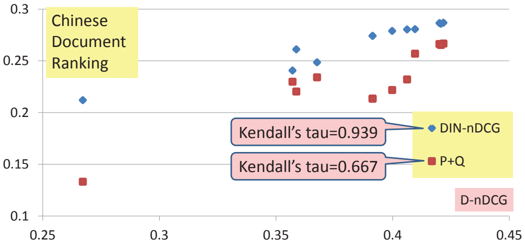

Figure 8: Correlation between D-nDCG and DIN-nDCG/P+Q for Chinese Document Ranking.

ues: for these topics that have one navigational intent and no informational intents, P+Q reduces to P$^+$, which reaches one if an $L4$-relevant document is retrieved at rank 1. Figure 7 shows per-topic performances for our Chinese DR top performer THUIR-D-C-1A: This run achieves a P+Q of one for the six navigational topics indicated with baloons.

Navigational Intents Intent type-sensitive metrics that leverage the informational and navigational intent tags produce system rankings that are somewhat different from that produced by the intent type-agnostic D-nDCG, although, by definition, DIN-nDCG approaches D-nDCG as the fraction of navigational subtopics decreases. Figure 8 visualises the correlations for the Chinese DR subtask: it can be observed that the correlation between D-nDCG and DIN-nDCG is much higher than that between D-nDCG and P+Q.

We also investigated the comparability of the INTENT-1 and INTENT-2 topic sets using the Revived Runs submitted to the Chinese and Japanese DR subtasks (one Chinese run and two Japanese runs). According to *two-sample bootstrap hypothesis tests* [6], there were no significant performance differences at $\alpha = 0.05$ for any of our primary metrics (I-rec and D($\sharp$)-nDCG), which suggests that the topics sets are more or less comparable (see Figure 1).

5. FUTURE DIRECTIONS

While it may be too demanding to expect a substantial progress between just two rounds of NTCIR, the progress monitoring practice depicted in Figure 2 is probably worth continuing and to apply it to other tasks.

The TREC Web Track has discontinued the diversity task; it is not clear if there will be an INTENT-3 task at NTCIR-11. However, it should be noted that diversity test collections are highly unlikely to be reusable [9, 11]: thus, if researchers want to continue improving diversified search[10], we do require a new diversity test collection. Note also that now a new corpus, ClueWeb12, is available [4]. Do we want a new diversity test collection based on this corpus? Do we need a new approach to evaluating diversified search? We would like to discuss these questions at the SIGIR poster session.

6. ADDITIONAL AUTHORS

Additional authors: Ruihua Song (Microsoft Research Asia, P.R.C., email: Song.Ruihua@microsoft.com) and Mayu Iwata (Osaka Unversity, Japan, email: iwata.mayu@ist.osaka-u.ac.jp).

7. REFERENCES

[1] R. Agrawal, G. Sreenivas, A. Halverson, and S. Leong. Diversifying search results. In *Proceedings of ACM WSDM 2009*, pages 5–14, 2009.

[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. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of ACM SIGIR 2006*, pages 525–532, 2006.

[7] T. Sakai. A note on progress in document retrieval technology based on the official ntcir results (in japanese). In *Proceedings of FIT 2006*, pages 67–70, 2006.

[8] T. Sakai. Evaluation with informational and navigational intents. In *Proceedings of ACM WWW 2012*, pages 499–508, 2012.

[9] T. Sakai. The unreusability of diversified search test collections. In *Proceedings of EVIA 2013*, 2013.

[10] T. Sakai, Z. Dou, and C. L. Clarke. The impact of intent selection on diversified search evaluation. In *Proceedings of ACM SIGIR 2013*, 2013.

[11] T. Sakai, Z. Dou, R. Song, and N. Kando. The reusability of a diversified search test collection. In *Proceedings of AIRS 2012*, pages 26–38, 2012.

[12] 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.

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

[14] T. Sakai and R. Song. Diversified search evaluation: Lessons from the NTCIR-9 INTENT task. *Information Retrieval*, 2013.

[15] R. L. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *Proceedings of CIKM 2010*, pages 1179–1188, 2010.

[16] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT task. In *Proceedings of NTCIR-9*, pages 82–105, 2011.

[17] 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.

[10] We note that there is a full paper session on Diversity at SIGIR 2013: diversified search is still a popular topic.