A Preview of the NTCIR-10 INTENT-2 Results

Tetsuya Sakai[1,a)] Zhicheng Dou[1] Takehiro Yamamoto[3] Yiqun Liu[2] Min Zhang[2]
Ruihua Song[1] Makoto P. Kato[3] Mayu Iwata[4]

Abstract: The second NTCIR INTENT task (INTENT-2) will be concluded at the NTCIR-10 conference in June 2013. The task comprises two subtasks: Subtopic Mining (given a query, return a ranked list of subtopic strings) and Document Ranking (given a query, return a diversified web search result). The task attracted participating teams from China, France, Japan and South Korea: 12 teams for Subtopic Mining and 4 teams for Document Ranking. This paper provides a preview of the official results of the task, while keeping the participating teams anonymous.

Keywords: diversity, evaluation, intents, NTCIR, subtopics, web search.

1. Introduction

This paper provides a preview of the official results of the NTCIR-10 INTENT-2 task[*1], while keeping the participating teams anonymous. The team names will be disclosed at the NTCIR-10 conference in June 2013[*2].

Figure 1 shows the overall structure of our task. In Subtopic Mining, 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). The organisers then ask multiple 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). The Subtopic Mining 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.

The black arrows in Figure 1 show the flow of the Document Ranking subtask, which is similar to the TREC Web Track Diversity Task [2]. 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). The Document Ranking 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*.

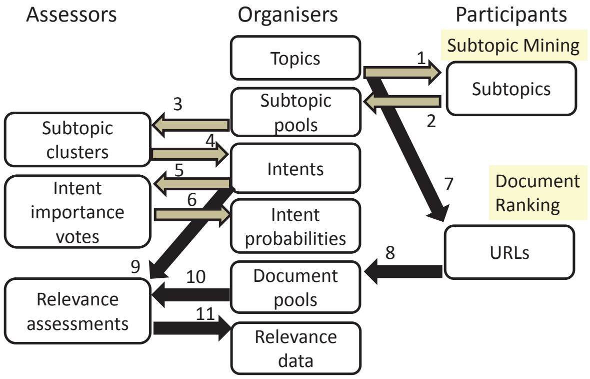

Fig. 1 Structure of the INTENT task.

Table 1 Number of INTENT-1 and INTENT-2 runs (teams).

	Subtopic Mining			Document Ranking	
	E	C	J	C	J
INTENT-1	–	42 (13)	10 (4)	24 (7)	15 (3)
INTENT-2	34 (8)	23 (6)	14 (3)	12 (3)	8 (2)

INTENT-2 attracted participating teams from China, France, Japan and South Korea. Table 1 compares the number of runs/teams for each (subtask, language) pair across INTENT-1 and INTENT-2. It can be observed that the English Subtopic Mining task, which we did not have at INTENT-1, was the most popular Subtask in INTENT-2.

2. Task and Data

2.1 What's New at INTENT-2

For both Subtopic Mining and Document Ranking, the input and output file specifications used at INTENT-2 are the same as

[1] Microsoft Research Asia, China
[2] Tsinghua University, China
[3] Kyoto University, Japan
[4] Osaka University, Japan
[a)] tetsuyasakai@acm.org

[*1] http://research.microsoft.com/en-us/people/tesakai/intent2.aspx

[*2] http://research.nii.ac.jp/ntcir/ntcir-10/

those used at INTENT-1: the run file formats are similar to the TREC run format.

New features of INTENT-2 are as follows.

(I) We introduced an *English* Subtopic Mining Subtask, using the 50 TREC 2012 Web Track topics kindly provided by its track coordinators. The diversity task of the TREC track devised their own set of "subtopics" for each topic; while we independently created the intents for each topic through our Subtopic Mining Subtask.

(II) We provided an "official" set of search engine query suggestions for each query to participants, to improve the reproducibility and fairness of experiments. Participants were asked to use these official query suggestions if their system required such data.

(III) For the Chinese and Japanese topic sets only, we provided a baseline non-diversified run and the corresponding web page contents to participants. This enables researchers to isolate the problem of diversifying a given search result from that of producing an effective initial search result. Moreover, this enables researchers to participate in the Document Ranking subtask by just reranking the baseline run, even without indexing the entire target corpus. The Chinese baseline run BASELINE-D-C-1 was provided by Tsinghua University; the Japanese one BASELINE-D-J-1 was provided by Microsoft Research Asia.

(IV) We intentionally included *navigational* queries in the INTENT-2 Chinese and Japanese topic sets. A navigational query should require one answer or one website, and therefore may not require diversification. We thereby encouraged participants to experiment with *selective diversification*: instead of uniformly applying a diversification algorithm to all topics, determine in advance which topics will (not) benefit from diversification. Moreover, to evaluate *intent type-sensitive diversification* [7], we tagged each intent with either *informational* or *navigational* based on five assessors' votes. More details will be given below.

(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 practice is useful for monitoring progress across NTCIR rounds, as we shall explain below.

Figure 2 explains Item (V) above, which is based on a proposal in a previous study which stressed the importance of comparing systems across different NTCIR rounds using the same topic set while checking the equivalence of topic sets across NTCIR rounds using the same system [5]. Because we have both INTENT-1 systems and INTENT-2 systems that process the INTENT-2 topics (Arrows (a) and (d)), we can examine if we have made any progress across the two rounds, by directly comparing the runs. In addition, although the INTENT-1 and INTENT-2 topic sets were constructed using different procedures (different contributors to the pools and different pool depths), we can investigate whether they can be regarded as comparable or "harder" than the other, using the Revived Runs from INTENT-1 that process both of these topic sets (Arrows (c) and (d)). Also, it should be noted that, although the INTENT-2 systems also processed the INTENT-1 topics (Arrow (b)), the effectiveness values obtained from the experiments are *not* reliable. This is because the INTENT-2 systems did not contribute to the INTENT-1 pools: Sakai *et al.* have actually demonstrated that the INTENT-1 Chinese Document Ranking Test Collection is *not* reusable and that runs that did not contribute to the pools are *underestimated* with this collection [8][*3]. The situation is probably even worse for the INTENT-1 Japanese Document Ranking Test Collection as only three teams contributed to the pool. Moreover, Subtopic Mining Test Collections are basically not reusable as the gold standards consist of arbitrary subtopic strings rather than document IDs. At INTENT-2, we have increased the pool depth from 20 to 40 for both subtasks.

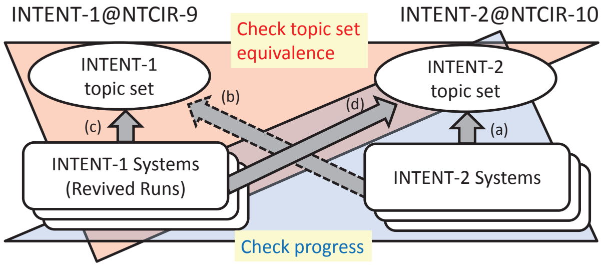

Fig. 2 Comparing INTENT-1 and INTENT-2.

Following INTENT-1, we created 100 Chinese and 100 Japanese topics based on "torso" queries from commercial search engine logs [12]. However, the INTENT-2 Chinese topic set contained two topics that overlapped with the INTENT-1 topic set (0272 and 0300), so we used only 98 topics for Chinese Subtopic Mining. Furthermore, for Document Ranking, we removed one more topic (0266) from the Chinese topic set and five topics (0356, 0363, 0367, 0370, 0371) from the Japanese topic set as they had no relevant documents in the pools.

As we have mentioned in Item (IV) above, we included navigational *topics* that probably do not require search result diversification. Moreover, we hired five assessors to individually label each *intent* with either navigational (nav) or informational (inf) using the same criteria, for the purpose of conducting intent type-sensitive search result diversification. The tests used for classifying intents into navigational and informational were as follows:

Test 1: Expected Answer Uniqueness Is the intent specific enough so that the expected relevant item (i.e. website, entity, object or answer) can be considered unique? Even if multiple relevant items exist, is it likely that there exists at least one searchable item that will completely satisfy the user and call for no additional information? If the answer is yes to either of these questions, the intent is navigational. Otherwise go to Test 2.

Test 2: Expected Answer Cohesiveness If the desired item is not unique, are these items expected to lie within a single website (which could typically be a group of mutually linked

[*3] At the NTCIR-6 Crosslingual IR Task, participants were asked to process past test collections (NTCIR-3, -4 and -5), to obtain reliable results based on multiple test collections [3]. This similar to Arrow (b) in Figure 2, but the crosslingual collections are probably more reusable than ours as they used larger pool depths (e.g. 100).

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
	amb/faceted topics	23/52	23/52
	shared topics	21	21
	reused topics	19	19
	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
	amb/faceted topics	27/40	27/40
	shared topics	21	21
	reused topics	33	33
	intents	587	582
	nav intents	–	259
	inf intents	–	323
	subtopic strings	2,979	–
	unique rel docs	–	5,085

web pages under the same domain name), so that this single website will completely satisfy the user and call for no additional information? If the answer is yes, the intent is navigational. Otherwise the intent is informational.

In the end, we classified an intent into navigational only when *four or five assessors* agreed that it is navigational. This is because, once an intent has been labelled as navigational, intent type-sensitive evaluation metrics basically ignore "redundant" information retrieved for that intent [7]. The inter-assessor agreement in terms of Fleiss' kappa was 0.4865 (confidence interval: 0.4611 to 0.5120) for Chinese and 0.2072 (confidence interval: 0.1809 to 0.2336) for Japanese. The low agreement for Japanese requires further investigation. As for the navigational *topics*, the organisers used the same criteria and labelled them ourselves through a discussion.

We also deliberately devised topics that are common across Chinese and Japanese, so that researchers can potentially conduct *cross-language search result diversification* experiments. There is in fact a one-to-one correspondence between the first 21 of the INTENT-2 Chinese and Japanese topics (0201-0221 from Chinese and 0301-0321 from Japanese). We call them *shared topics*. Moreover, some of the INTENT-2 topics were selected from past TREC Web Track topics. We call them *reused topics*. Eleven of the shared topics are also reused topics (0211-0221 and 0311-0321). In total, the Chinese topic set contains 19 reused topics, while the Japanese topic set contains 33.

Table 2 summarises the statistics of the INTENT-2 topics and intents. As the topics we lost after relevance assessments (0266 for Chinese and 0356, 0363, 0367, 0370, 0371 for Japanese) were all navigational, note that the number of navigational topics and the number of intents are accordingly smaller in the Document Ranking column.

2.2 Subtopic Mining Subtask

In this section, we provide more details on the construction of the Subtopic Mining Test Collections (the grey arrows in Figure 1).

In Subtopic Mining, participants were asked to return a ranked list of subtopic strings for each query. We provided the following instruction on the INTENT-2 home page:

A subtopic string of a given query is a query that specialises and/or disambiguates the search intent of the original query. If a string returned in response to the query does neither, it is considered incorrect... It is encouraged that participants submit subtopics of the form "<originalquery><additionalstring>" or "<originalquery>[space]<additionalstring>" whereover appropriate although we do allow subtopics that do NOT contain the original query...

As was mentioned earlier, the top 40 subtopic strings from every run were included in the pool for each topic, and the subtopic strings were manually clustered so as to form a set of intents. Each substring belongs to exactly one cluster (which could be a "nonrelevant" cluster). We hired multiple assessors for the clustering task, but each topic was entrusted to one assessor. We also asked the assessors to provide a label for each cluster in the form "<originalquery> <additionalstring>."

Having clustered the subtopics, we then hired ten assessors to individually judge whether each cluster is important or not with respect to the given query. Then, in contrast to INTENT-1 where we had up to 24 intents for a single topic [12], we decided to select up to 9 intents per topic based on the votes. If there was a tie across this threshold, we removed the entire tie to ensure that it is not exceeded. This change was made because search result diversification is mainly about diversifying the *first* search engine result page, which can only accommodate around 10 URLs.

Having thus obtained the set of intents for each query, we then estimated the intent probabilities from the votes, using Eq. 2 from the INTENT-1 Overview paper [12].

The number of intents and subtopic strings obtained for Subtopic Mining are shown in Table 2.

Three types of runs were allowed in the Subtopic Mining Subtask:

R-run A Revived Run using a system from INTENT-1 (see Figure 1). Not applicable to English as INTENT-1 did not have an English Subtask.

B-run Any run that uses the organisers' Baseline non-diversified Document Ranking run in any way. Not applicable to English as there is no baseline Document Ranking run for English.

A-run Any other run.

Participants were allowed to submit up to five *new runs* (i.e. B-runs or A-runs) and two R-runs for each (subtask, language) pair. Manual runs were not allowed.

Unfortunatley, as we did not receive any Revived Runs in Subtopic Mining, the progress checking mechanism of Figure 2 does not work for this subtask.

2.3 Document Ranking Subtask

In this section, we provide more details on the construction of the Document Ranking Test Collections (the black arrows in Figure 1).

In Document Ranking, participants were asked to return a ranked list of URLs for each query. The target corpora are the same as those used at INTENT-1: *SogouT*[*4] for Chinese and *ClueWeb09-JA*[*5] for Japanese [12]. The task is similar to the TREC Web Track Diversity Task, but differs in several aspects:

- Intent probabilities and per-intent graded relevance information are utilised, as in INTENT-1;
- Participants were encouraged to *selectively diversify* search results, as some of the topics are navigational and probably do not require diversification;
- It was announced that we will also use *intent type-sensitive* evaluation metrics in addition to the primary metrics from INTENT-1, so that participants were encouraged to consider whether each intent is navigational or informational.

In the Document Ranking Subtask also, participants were allowed to submit up to five *new runs* (i.e. B-runs or A-runs) and two R-runs for each (subtask, language) pair. We received three R-Runs, TEAM09-D-C-R1 (from TEAM09), TEAM06-D-J-R1 and TEAM06-D-J-R2 (from TEAM06), which we shall discuss later for the purpose of progress monitoring (see Figure 2).

Following the reusability study by Sakai *et al.* [8], we increased the pool depth from 20 to 40 at INTENT-2, as was mentioned earlier. Following INTENT-1, every document was judged independently by two assessors, and their assessments were consolidated to form five-point-scale relevance data (*L*0-*L*4). Note that, unlike the Subtopic Mining data, a document may be relevant to multiple intents, and that these per-intent relevance assessments are graded. The maximum number of intents covered by a relevant document is six for the Chinese data and eight for the Japanese data. Recall that we have no more than nine intents for each INTENT-2 topic.

The number of unique relevant documents per topic summed across the topic set for each Document Ranking Subtask is shown in Table 2. Also, Tables 3 and 4 show the number of relevant documents by relevance level for INTENT-2 and INTENT-1, respectively. Here, note that a document is counted multiple times if it is relevant to multiple intents. It can be observed that, despite the use of deeper pools, the the number of relevant documents obtained at INTENT-2 is considerably smaller, due to the limited number of participants.

3. Evaluation Metrics

This section briefly describes the evaluation metrics used for ranking the INTENT-2 participating systems. Section 3.1 defines the intent type-agnostic *intent recall* (I-rec), *D-nDCG* and *D♯-nDCG* [10], our primary metrics which were also used at INTENT-1. These metrics were originally designed for Document Ranking, but we use them for Subtopic Mining as well. Section 3.2 defines the intent type-sensitive *DIN-nDCG* and *P+Q* [7], which we use as supplementary metrics for evaluating Document Ranking.

All metric values reported in this paper were computed using the *NTCIREVAL* toolkit [6][*6]. We use the document cutoff of $l = 10$ throughout this paper, as a post hoc analysis of the INTENT-1 runs showed that run rankings and significance test results based on $l = 30$ are not so reliable, at least when the pool depth is 20 [8]. Recall, however, that we have increased the pool depth to 40 for both subtasks of INTENT-2.

Table 3 INTENT-2 relevance assessment statistics.

	Chinese (97 topics)	Japanese (95 topics)
*L*4	224	1,596
*L*3	613	1,545
*L*2	7,265	2,779
*L*1	6,667	3,824
total	14,769	9,744

Table 4 INTENT-1 relevance assessment statistics.

	Chinese (100 topics)	Japanese (100 topics)
*L*4	1,436	2,201
*L*3	2,557	2,955
*L*2	7,382	6,463
*L*1	12,196	8,222
total	23,571	19,841

3.1 Intent Type-Agnostic Metrics

Let I be the set of known intents for a given query q, and let $I'(\subseteq I)$ be the set of intents covered by a ranked list. Then $I\text{-}rec = |I'|/|I|$. For each $i \in I$, let $Pr(i|q)$ denote its intent probability, and let $g_i(r)$ be the gain value of the item at rank r with respect to i, which we we define as x if the item is Lx-relevant to i and 0 otherwise (e.g., 4 if $L4$-relevant). The "global gain" for this item is defined as:

$$GG(r) = \sum_i Pr(i|q) g_i(r) \, . \qquad (1)$$

The "globally ideal" ranked list is obtained by sorting all relevant items by the global gain. Let $GG^*(r)$ denote the global gain in this ideal list. D-nDCG at cutoff l is defined as:

$$D\text{-}nDCG@l = \frac{\sum_{r=1}^{l} GG(r)/\log(r+1)}{\sum_{r=1}^{l} GG^*(r)/\log(r+1)} \, . \qquad (2)$$

I-rec is a pure diversity metric for set retrieval, while D-nDCG is an overall relevance metric for ranked retrieval. Hence, we plot D-nDCG against I-rec to compare participating systems. Moreover, we compute our primary metric by summarising the graph:

$$D\sharp\text{-}nDCG = \gamma I\text{-}rec + (1-\gamma) D\text{-}nDCG \qquad (3)$$

where we let $\gamma = 0.5$ throughout this paper. The advantages of D♯-nDCG over other diversity measures are discussed elsewhere [10], [11].

D-nDCG and D♯-nDCG were originally designed for Document Ranking evaluation. However, we also use it for Subtopic Mining. Note that, in the case of Subtopic Mining, each subtopic string is relevant to no more than one intent and the relevance labels are binary. Thus Eq. 1 reduces to the probability of one particular intent. That is, D-nDCG reduces to traditional nDCG where the gain value of each document is exactly the intent probability of the intent to which that document is relevant.

[*4] http://www.sogou.com/labs/dl/t.html

[*5] http://lemurproject.org/clueweb09/

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

3.2 Intent Type-Sensitive Metrics

While the aforementioned intent type-agnostic metrics aim at allocating more space in the search result page to documents that are highly relevant to popular intents, they do not consider whether each intent is informational or navigational. It is possible that exactly one URL slot in the search result page is needed for a navigational intent, while more URL slots will help for an informational intent. Intent type-sensitive metrics were designed to optimise diversification from this viewpoint.

DIN-nDCG is a type-sensitive variant of D-nDCG, which is defined as follows. Let $\{i\}$ and $\{j\}$ denote the sets of informational and navigational intents for query q, and let $isnew_j(r) = 1$ if there is no document relevant to the navigational intent j between ranks 1 and $r-1$, and $isnew_j(r) = 0$ otherwise. We redefine the global gain as:

$$GG^{DIN}(r) = \sum_i Pr(i|q) g_i(r) + \sum_j isnew_j(r) Pr(j|q) g_j(r) \, . \tag{4}$$

That is, in this formulation of the global gain, "redundant" relevant documents for informational intents are ignored. Then DIN-nDCG is defined as:

$$DIN\text{-}nDCG@l = \frac{\sum_{r=1}^{l} GG^{DIN}(r)/\log(r+1)}{\sum_{r=1}^{l} GG^{*}(r)/\log(r+1)} \, . \tag{5}$$

Clearly, $DIN\text{-}nDCG \leq D\text{-}nDCG$ holds.

The second intent type-sensitive metric we use, P+Q, is a generalisation of the *intent-aware* approach to diversity evaluation [1]. The difference is that P+Q switches between two different metrics depending on whether each intent is informational or navigational.

First, we define two existing metrics for *traditional* ranked retrieval. Let $J(r) = 0$ if a document at rank r is nonrelevant to the query and $J(r) = 1$ otherwise. Let $C(r) = \sum_{k=1}^{r} J(k)$. Let $g(r)$ denote the gain at rank r of the system output, and let $g^*(r)$ denote the gain at rank k of the ideal output (i.e., a list sorted by the gain value), respectively. Then the *blended ratio* at rank r, a graded-relevance version of precision, is defined as:

$$BR(r) = \frac{C(r) + \beta \sum_{k=1}^{r} g(k)}{r + \beta \sum_{k=1}^{r} g^*(k)} \tag{6}$$

where β (≥ 0) is a user persistence parameter which is set to 1 throughout this study. Moreover, let rp be the rank of the document that is most relevant within $1 \leq rp \leq l$ *and* is closest to the top. Then, the following metrics can be defined[*7].:

$$P^+@l = \frac{1}{C(rp)} \sum_{r=1}^{rp} J(r) BR(r) \tag{7}$$

$$Q@l = \frac{1}{\min(l, R)} \sum_{r=1}^{L} J(r) BR(r) \, . \tag{8}$$

The only difference between these two metrics is the *stopping probability distribution* over ranks [9]: Q assumes a uniform distribution across all relevant documents retrieved above l; P$^+$ assumes a uniform distribution across all relevant documents retrieved above rp.

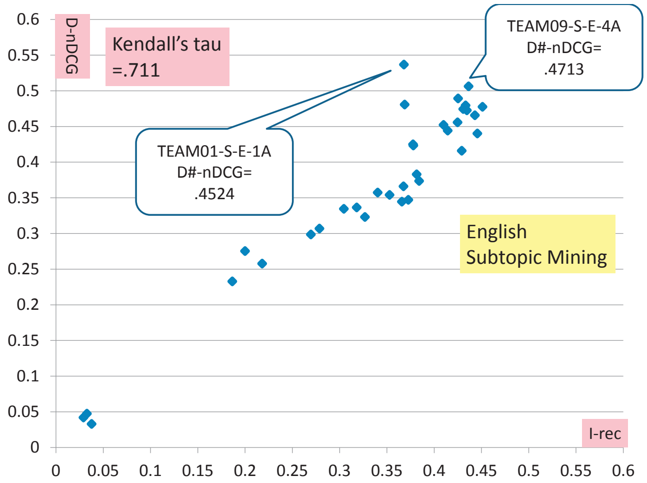

Fig. 3 I-rec/D-nDCG graph for English Subtopic Mining.

The above definitions of Q and P$^+$ suggest that they are suitable for informational and navigational needs, respectively. Hence, we define P+Q for diversity evaluation as follows:

$$P{+}Q@l = \sum_i Pr(i|q) Q_i@l + \sum_j Pr(j|q) P^+_j \tag{9}$$

where Q_i is computed for each informational intent i and P^+_j is computed for each navigational intent j.

While Sakai [7] also proposed to combine DIN-nDCG and P+Q with intent recall, we omit that particular approach here as the resultant metrics are very highly correlated with D♯-nDCG and I-rec.

Henceforth, we shall discuss statistical significance based on a randomised version of the two-sided Tukey's Honestly Significant Differences (HSD) test at $\alpha = 0.05$ unless otherwise indicated.

4. Subtopic Mining Results

4.1 English Subtopic Mining Results

Figure 3 shows the *I-rec/D-nDCG graph* [12] for the English Subtopic Mining runs. Recall that I-rec reflects pure diversity, while D-nDCG reflects the overall relevance. In the official results, (a) TEAM01-S-E-1A is the top performer in terms of relevance (i.e. D-nDCG); (b) TEAM09-S-E-1A is the top performer in terms of diversity (i.e. I-rec); and (c) TEAM09-S-E-4A is the overall winner in terms of D♯-nDCG. However, these three runs are statistically indistinguishable from one another in terms of D♯-nDCG. More generally, in terms of D♯-nDCG, TEAM01, TEAM04, ORG, TEAM07 and TEAM08 all have at least one run that is statistically indistinguishable from TEAM09-S-E-4A[*8]. Whereas, all runs from TEAM05 and TEAM11 significantly underperform this top run.

4.2 Chinese Subtopic Mining Results

Figure 4 shows the I-rec/D-nDCG graph for the Chinese Subtopic Mining runs. In the official results, (a) TEAM09-S-C-3A is the top performer in terms of relevance (i.e. D-nDCG); (b) TEAM11-S-C-1A is the top performer in terms of diversity

[*7] P^+ is defined to be 0 if there is no relevant document within $[1, l]$.

[*8] ORG is the INTENT-2 organisers's team.

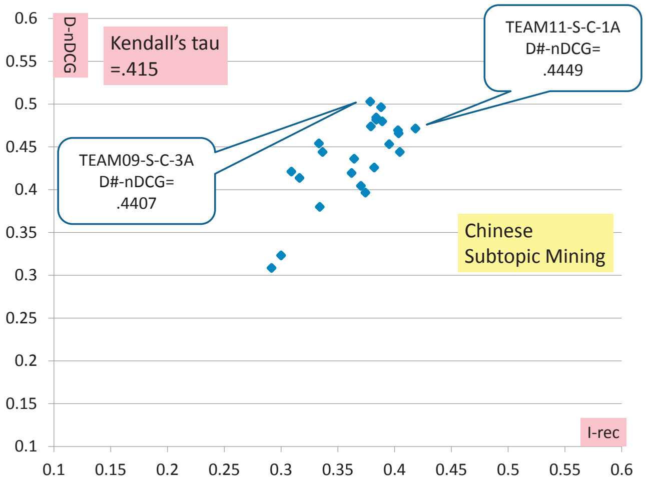

Fig. 4 I-rec/D-nDCG graph for Chinese Subtopic Mining.

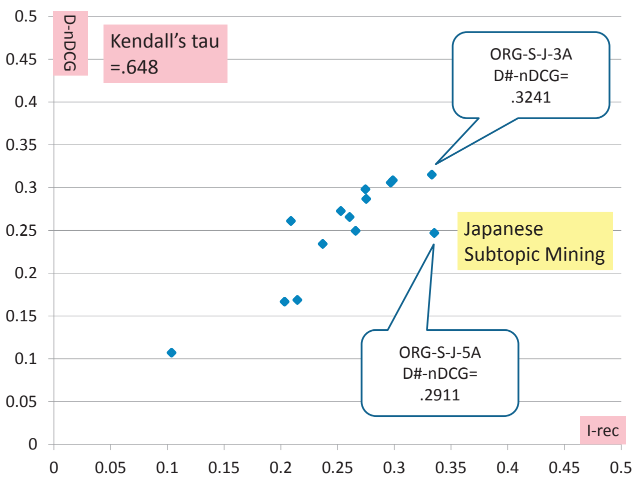

Fig. 5 I-rec/D-nDCG graph for Japanese Subtopic Mining.

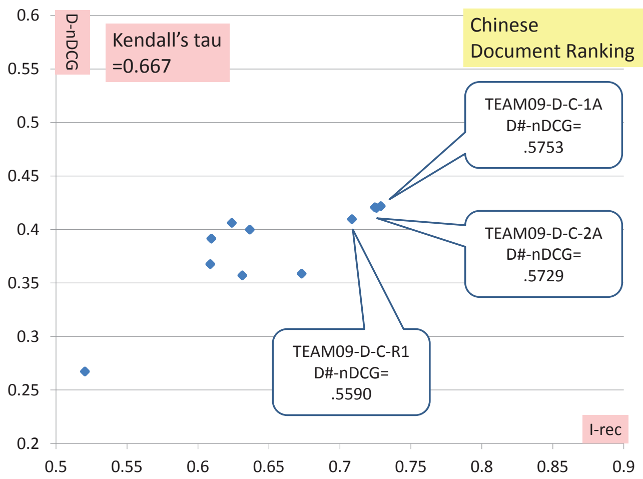

Fig. 6 I-rec/D-nDCG graph for Chinese Document Ranking.

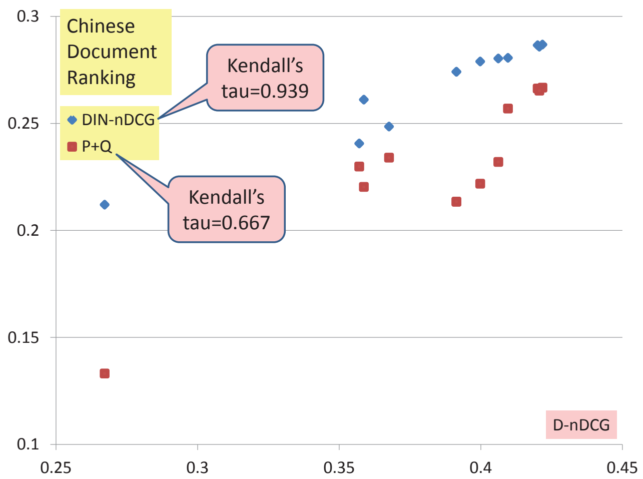

Fig. 7 Correlation between D-nDCG and DIN-nDCG/P+Q for Chinese Document Ranking.

(i.e. I-rec); and (c) TEAM11-S-C-1A is the overall winner in terms of D♯-nDCG. However, the difference between these two runs in D♯-nDCG is *not* statistically significant. More generally, in terms of D♯-nDCG, TEAM02, TEAM03, ORG, TEAM09 and TEAM10 (i.e. all of the other teams that participated in Chinese Subtopic Mining) all have at least one run that is statistically indistinguishable from TEAM11-S-C-1A. In short, the six teams are statistically indistinguishable from one another.

4.3 Japanese Subtopic Mining Results

Figure 5 shows the I-rec/D-nDCG graph for the Japanese Subtopic Mining runs. In the official results, (a) ORG-S-J-3A is the top performer in terms of relevance (i.e. D-nDCG); (b) ORG-S-J-5A is the top performer in terms of diversity (i.e. I-rec); and (c) ORG-S-J-3A is the overall winner in terms of D♯-nDCG. However, the difference between these two runs in D♯-nDCG is *not* statistically significant. More generally, in terms of D♯-nDCG, both TEAM04 and TEAM06 (i.e. all of the other teams that participated in Japanese Subtopic Mining) have at least one run that is statistically indistinguishable from ORG-S-J-3A. In short, the three teams are statistically indistinguishable from one another.

5. Document Ranking Results

5.1 Chinese Document Ranking Results

Figure 6 shows the I-rec/D-nDCG graph for the Chinese Document Ranking runs. In the official results, TEAM09-D-C-1A is the winner in terms of in terms of all five metrics. In terms of D♯-nDCG, it significantly outperforms BASELINE-D-C-1 ($p \leq 0.001$). However, TEAM03 has two runs that are statistically indistinguishable from TEAM09-D-C-1A in terms of D♯-nDCG.

Unfortunately, none of the new runs from TEAM09 significantly outperforms its Revived Run TEAM09-D-C-R1. Therefore, we cannot conclude from these experiments that there has been substantial progress compared to INTENT-1.

Figure 7 shows the correlation between the type-agnostic D-nDCG and the type-sensitive DIN-nDCG/P+Q when ranking the Chinese Document Ranking runs. It can be observed that the correlation between D-nDCG and DIN-nDCG is higher than that between D-nDCG and P+Q. The correlation between D-nDCG and DIN-nDCG is particularly high for this test collection as only a small fraction of the subtopics is navigational (125 out of 615= 20%, as shown in Table 2): recall that DIN-nDCG is equal to D-nDCG if all subtopics are informational.

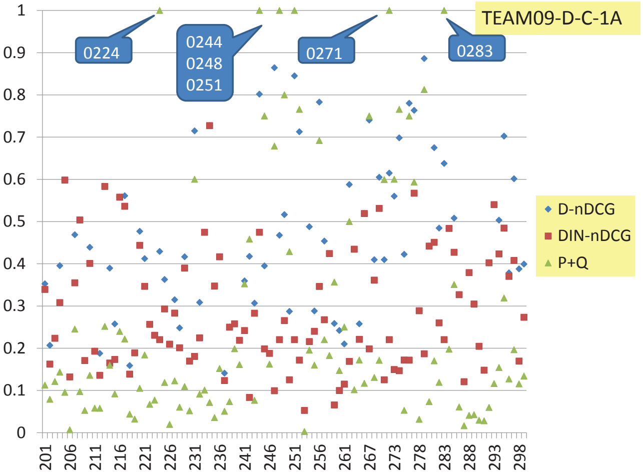

Fig. 8 Per-topic D-nDCG/DIN-nDCG/P+Q performances for TEAM09-D-C-1A.

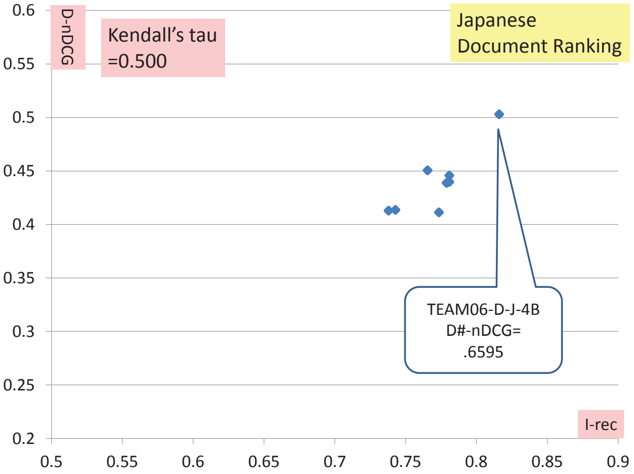

Fig. 9 I-rec/D-nDCG graph for Japanese Document Ranking.

Figure 8 compares the per-topic D-nDCG/DIN-nDCG/P+Q values for TEAM09-D-C-1A, our top performer. Five instances where the P+Q values are one are indicated with baloons. These topics are all navigational, so P+Q reduces to P$^+$. Thus, if an *L*4-relevant document (i.e. document with the highest relevance level) is retrieved at rank 1, P+Q equals one for these topics.

Table 5 compares the performances of our Revived Run, TEAM09-D-C-R1, across INTENT-1 and INTENT-2 (see Figure 2). We used a two-sample unpaired bootstrap test [4] to see whether the two topic sets are statistically significantly different. As indicated in the table, Only the difference in D-nDCG was statistically significant at $\alpha = 0.10$ ($p = 0.087$). Judging from these limited results alone, it appears that the two topic sets are more or less comparable.

5.2 Japanese Document Ranking Results

Figure 9 shows the I-rec/D-nDCG graph for the Japanese Document Ranking runs. It can be observed that TEAM06-D-J-4B is the winner in terms of all five metrics. In terms of D♯-nDCG, it outperforms all other runs, i.e. BASELINE-D-J-1 and other TEAM06 runs. In particular, TEAM06-D-J-4B significantly outperforms its Revived Runs TEAM06-D-J-R1 and TEAM06-D-J-R2 ($p \leq 0.001$), which suggests that the method may be substantially better than those used at INTENT-1. TEAM06-D-J-4B combined search results of the baseline, Yahoo! and Bing, and this seems to have been successful.

Figure 10 shows the correlation between the type-agnostic D-nDCG and the type-sensitive DIN-nDCG/P+Q when ranking the Japanese Document Ranking runs. Again, it can be observed that the correlation between D-nDCG and DIN-nDCG is higher than that between D-nDCG and P+Q. Moreover, the correlation between D-nDCG and DIN-nDCG is lower than the Chinese case, reflecting the fact that the Japanese topic set contains a considerably higher fraction of navigational subtopics (259 out of 582= 45%, as shown in Table 2).

Figure 11 compares the per-topic D-nDCG/DIN-nDCG/P+Q values for TEAM06-D-J-4B, our top performer. Eleven instances where the P+Q values are one are indicated with baloons. Again, these topics are all navigational topics, so P+Q reduces to P$^+$. Thus, if a *L*4-relevant document is retrieved at rank 1, P+Q equals one for these topics. In particular, for Topic 0383, D-nDCG is also one, while DIN-nDCG is only 0.6131. There are only two relevant documents (both of which are *L*4-relevant) for this topic, and the run managed to retrieve these two documents at ranks 1 and 2. However, as DIN-nDCG treats the second relevant document as nonrelevant, it does not give a full score to the run. This is a known normalisation issue with DIN-nDCG [7].

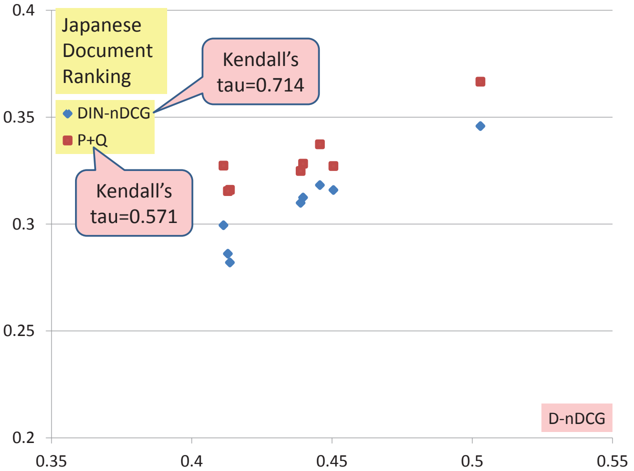

Fig. 10 Correlation between D-nDCG and DIN-nDCG/P+Q for Japanese Document Ranking.

Table 6 compares the performances of our Revived Runs, TEAM06-D-J-R2 and TEAM06-D-J-R1 across INTENT-1 and INTENT-2 (see Figure 2). Again, we used a two-sample unpaired bootstrap test to see whether the two topic sets are statistically significantly different, but did not obtain any significant differences. Judging from these limited results alone, it appears that the two topic sets are more or less comparable.

6. Conclusions and Future Work

INTENT-2 attracted participating teams from China, France, Japan and South Korea – 12 teams for Subtopic Mining and 4 teams for Document Ranking (including an organisers' team). The Subtopic Mining subtask received 34 English runs, 23 Chinese runs and 14 Japanese runs; the Document Ranking subtask

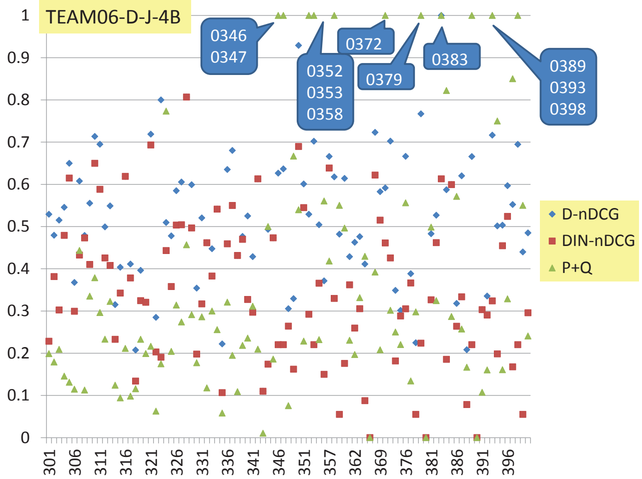

Fig. 11 Per-topic D-nDCG/DIN-nDCG/P+Q performances for TEAM06-D-J-4B.

Table 5 TEAM09 Revived Run performances for the INTENT-1 and INTENT-2 topic sets. Only the difference in D-nDCG is statistically significant at $\alpha = 0.10$ according to an unpaired bootstrap test: the p-value is shown below.

Run: TEAM09-D-C-R1	INTENT-1 topics	INTENT-2 topics
I-rec@10	0.6861	0.7085
D-nDCG@10	0.4573 (p = .087)	0.4096
D♯-nDCG@10	0.5717	0.5590

Table 6 TEAM06 Revived Run performances for the INTENT-1 and INTENT-2 topic sets. None of the differences is statistically significant according to an unpaired bootstrap test.

Run: TEAM06-D-J-R2	INTENT-1 topics	INTENT-2 topics
I-rec@10	0.7307	0.7735
D-nDCG@10	0.4101	0.4113
D♯-nDCG@10	0.5704	0.5924
Run: TEAM06-D-J-R1	INTENT-1 topics	INTENT-2 topics
I-rec@10	0.7369	0.7380
D-nDCG@10	0.4352	0.4129
D♯-nDCG@10	0.5861	0.5754

received 12 Chinese runs and 8 Japanese runs. Some preliminary findings are:

English Subtopic Mining TEAM09-S-E-4A outperformed all other runs in terms of Mean D♯-nDCG, but TEAM01, TEAM04, ORG, TEAM07 and TEAM08 all have at least one run that is statistically indistinguishable from this top run. Whereas, all runs from TEAM05 and TEAM11 significantly underperform TEAM09-S-E-1A.

Chinese Subtopic Mining TEAM11-S-C-1A outperformed all other runs in terms of Mean D♯-nDCG, but the six participating teams are statistically indistinguishable from one another.

Japanese Subtopic Mining 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 TEAM09-D-C-1A outperformed all other runs in terms of Mean D♯-nDCG; it significantly outperformed the baseline nondiversified run. However, TEAM03 has two runs that are statistically indistinguishable from this top run. Moreover, none of the new runs from TEAM09 significantly outperforms its Revived Run TEAM09-D-C-R1, and therefore it is not clear whether there has been a substantial improvement between INTENT-1 and INTENT-2.

Japanese Document Ranking TEAM06-D-J-4B outperformed all other runs in terms of Mean D♯-nDCG. In particular, it significantly outperforms its Revived Runs TEAM06-D-J-R1 and TEAM06-D-J-R2. It appears that the gain over these systems from INTENT-1 comes from combination of multiple search engine results.

Navigational Topics The D♯-nDCG values for navigational topics tend to be high for the Chinese/Japanese Subtopic Mining/Document Ranking subtasks[*9], as there is only one intent for these topics. Moreover, the per-topic analysis of the top Document Ranking runs suggests that navigational topics tend to receive high P+Q values (which reduce to P$^+$ for these topics). The effectiveness of selective diversification (e.g. switching off diversification for seemingly navigational topics) remains to be investigated.

Navigational Intents As the rank correlation values between D-nDCG and DIN-nDCG/P+Q show, intent type-agnostic and type-sensitive evaluation metrics produce somewhat different rankings, although by definition DIN-nDCG approaches D-nDCG as the fraction of navigational subtopics decreases. The effectiveness of intent type-sensitive diversification (e.g. allocating more space in the search engine result page to informational intents compared to navigational intents) remains to be investigated.

The actual team names will be disclosed at the NTCIR-10 conference in June 2013. More importantly, the future of the INTENT task will be discussed there.

[*9] This particular observation is not discussed in the results section of this paper due to lack of space.

References

[1] Agrawal, R., Sreenivas, G., Halverson, A. and Leong, S.: Diversifying Search Results, *Proceedings of ACM WSDM 2009*, pp. 5–14 (2009).

[2] Clarke, C. L. A., Craswell, N. and Voorhees, E. M.: Overview of the TREC 2012 Web Track, *Proceedings of TREC 2012* (2013).

[3] Kishida, K., hua Chen, K., Lee, S., Kuriyama, K., Kando, N. and Chen, H.-H.: Overview of CLIR Task at the Sixth NTCIR Workshop, *Proceedings of NTCIR-6*, pp. 1–19 (2007).

[4] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *Proceedings of ACM SIGIR 2006*, pp. 525–532 (2006).

[5] Sakai, T.: A Note on Progress in Document Retrieval Technology based on the Official NTCIR Results (in Japanese), *Proceedings of FIT 2006*, pp. 67–70 (2006).

[6] Sakai, T.: NTCIREVAL: A Generic Toolkit for Information Access Evaluation, *Proceedings of FIT 2011*, Vol. 2, pp. 23–30 (2011).

[7] Sakai, T.: Evaluation with Informational and Navigational Intents, *Proceedings of ACM WWW 2012*, pp. 499–508 (2012).

[8] Sakai, T., Dou, Z., Song, R. and Kando, N.: The Reusability of a Diversified Search Test Collection, *Proceedings of AIRS 2012*, pp. 26–38 (2012).

[9] Sakai, T. and Robertson, S.: Modelling A User Population for Designing Information Retrieval Metrics, *Proceedings of EVIA 2008*, pp. 30–41 (2008).

[10] Sakai, T. and Song, R.: Evaluating Diversified Search Results Using Per-Intent Graded Relevance, *Proceedings of ACM SIGIR 2011*, pp. 1043–1042 (2011).

[11] Sakai, T. and Song, R.: Diversified Search Evaluation: Lessons from the NTCIR-9 INTENT Task, *Information Retrieval* (2013).

[12] Song, R., Zhang, M., Sakai, T., Kato, M. P., Liu, Y., Sugimoto, M., Wang, Q. and Orii, N.: Overview of the NTCIR-9 INTENT Task, *Proceedings of NTCIR-9*, pp. 82–105 (2011).