Overview of the NTCIR-10 INTENT-2 Task

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

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

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

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

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

Ruihua Song
Microsoft Research Asia, P.R.China
Song.Ruihua@microsoft.com

ABSTRACT

This paper provides an overview of the NTCIR-10 INTENT-2 task (the second INTENT task), which comprises the Subtopic Mining and the Document Ranking subtasks. 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 received 12 Chinese runs and 8 Japanese runs. We describe the subtasks, data and evaluation methods, and then report on the *official* results, as well as the *revised* results for Subtopic Mining.

Keywords

diversity, evaluation, intents, subtopics, test collections.

1. INTRODUCTION

This paper provides an overview of the NTCIR-10 INTENT-2 task (the second INTENT task), which comprises the *Subtopic Mining* and the *Document Ranking* subtasks[1].

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*.

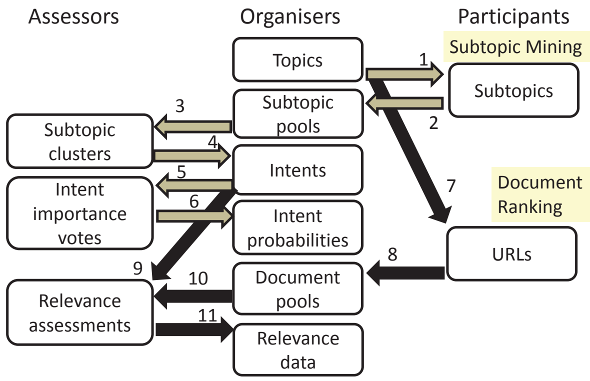

Figure 1: Structure of the INTENT task.

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 popularity of the INTENT task has declined considerably. In particular, only one team besides the organiser's team participated in the Japanse Document Ranking subtask. Whereas, the English Subtopic Mining task, which we did not have at INTENT-1, was the most popular Subtask in INTENT-2. Table 2 shows the list of INTENT-2 participants.

Table 3 shows the important dates of INTENT-2. Unlike INTENT-1 where we had the deadlines for Subtopic Mining and Document Ranking one after the other, we only had one common deadline for INTENT-2, and this deadline was the earliest among the NTCIR-10 tasks. But we feel that this only partially explains the decline in the number of participations, especially for Document Ranking. Also, unfortunately, after releasing the official results to participants, we discovered some bugs in the files that contain gold-standard subtopic strings that were used for evaluating the Subtopic

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

Table 2: INTENT participating teams. Teams with a ∗ participated in the same subtask at NTCIR-9.

team name	language	organisation
(a) Subtopic Mining		
hultech	E	University of Caen Lower-Normandy, France
ICRCS	C	Harbin Institute of Technology Shenzhen Graduate School, P.R.C.
KECIR	C	Shenyang Aerospace University, P.R.C.
KLE∗	E,J	Knowledge and Language Engineering Laboratory, POSTECH, South Korea
LIA	E	University of Avignon, France
MSINT∗	J	Microsoft Research Asia, P.R.C.
ORG∗	C,E,J	Organisers' runs using web search query suggestions/completions
SEM12	E	Toyohashi University of Technology, Japan
THCIB	E	Tsinghua University and Canon Information (Beijing) Co. Ltd, P.R.C.
THUIR∗	C,E	Tsinghua University, P.R.C.
THUIS	C	Tsinghua University, P.R.C.
TUTA1∗	C,E	University of Tokushima, Japan
(b) Document Ranking		
BASELINE	C,J	Nondiversified search results provided by organisers
KECIR	C	Shenyang Aerospace University, P.R.C.
MSINT∗	J	Microsoft Research Asia, P.R.C.
THUIR∗	C	Tsinghua University, P.R.C.

Table 3: INTENT-2 important dates.

May 31, 2012	Chinese/Japanese topics, query suggestions and non-diversified baseline Document Ranking runs released
June 18, 2012	English topics (for Subtopic Mining only) released (same as TREC 2012 web topics)
July 31, 2012	All submissions due
Aug-Dec 2012	Subtopic clustering, intent voting, per-intent relevance assessments
Dec 21, 2013	Official evaluation results released
Feb 1, 2013	Revised Subtopic Mining results released

Table 4: Official query suggestion data.

	English	Chinese	Japanese
harvested date	June, 15, 2012	March 1, 2012	
	Bing query suggestions Bing query completions Google query completions Yahoo! query completions	Baidu query suggestions Bing (Chinese) query suggestions Google (Chinese) query suggestions Sogou query suggestions	Bing (Japanese) query suggestions Bing (Japanese) query completions Google (Japanese) query completions Yahoo! (Japanese) query completions

Mining runs. We therefore released a set of *revised* Subtopic Mining results on February 1, 2013. Participants were asked to choose whether to discuss their *official* or *revised* results, and to explicitly state their choice in their papers.

The remainder of this paper is organised as follows. Sections 2 describes the details of the Subtopic Mining and Document Ranking subtasks and the test collections contructed, with an emphasis on parts that differ from INTENT-1. For more general task specifications, we refer the reader to the INTENT-1 Overview paper [16], and the aforementioned INTENT-2 homepage. Section 3 briefly describes the evaluation metrics we use. Sections 4 and 5 report on the *official* and *revised* evaluation results for Subtopic Mining, respectively. Section 6 reports on our official evaluation results for Document Ranking. Section 7 concludes this paper and the Appendix contains the details of each run as well as significance test results.

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 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. We received the English topic set from TREC on June 13, and released it to the participants on June 18 (see Table 3).

(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. Table 4 shows the harvested dates and sources of the official query suggestion 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,

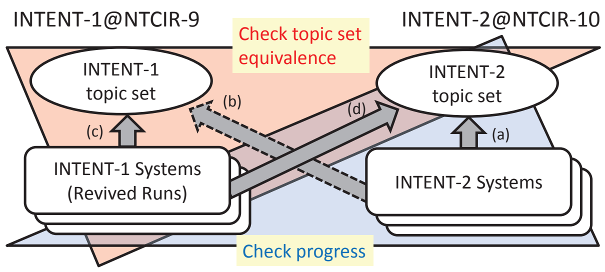

Figure 2: Comparing INTENT-1 and INTENT-2.

determine in advance which topics will (not) benefit from diversification. Moreover, to evaluate *intent type-sensitive diversification* [11], 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 [9]. 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 [12][2]. 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.

Following INTENT-1, we created 100 Chinese and 100 Japanese topics based on "torso" queries from commercial search engine logs [16]. 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 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 [11]. 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. Figure 3 shows the INTENT-2 Chinese and Japanese navigational topics. For these topics, subtopic clustering was not applied: only the relevance of each subtopic string was assessed. That is, *each navigational topic contains exactly one intent, which is navigational*. For informational topics, we tried to include both *ambiguous* and *faceted* topics.

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. The complete lists are shown in Table 5.

Table 6 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. Note that the statistics for the *revised* data are shown in parentheses. For comparison, Table 7 shows similar statistics for INTENT-1.

[2]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 [6]. 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).

(a) Chinese Navigational Topics (23)

0224	婚庆预算
0231	ip地址查询
0242	游戏鼠标评测
0244	2008年春节是哪天
0245	央视主持人周涛简历
0247	中国银监会
0248	秦时明月主题曲
0249	什么是京都议定书
0250	中国工商银行主页
0251	肝移植费用
0252	清明节的来历
0259	飞轮海mtv
0261	崔永元的博客
0262	神州租车
0266	武林外传续集播放
0267	肯德基优惠券
0270	陕西临潼农民发现秦始皇兵马俑
0271	艾滋病皮肤的症状
0273	中国政法大学
0274	新浪邮箱
0276	养老金计算办法
0279	蒋介石五大主力
0283	鉴宝节目主持人是谁

(b) Japanese Navigational Topics (33)

0344	日本感染症学会トップページ
0346	www.fujitv.co.jp
0347	パナマ国旗の画像
0349	キッザニア東京のホームページ
0352	急がば回れ 意味
0353	プラズマと液晶の違い
0355	misonoのブログ
0356	aaaの最強ベイビーの歌詞
0358	もち米の炊き方
0359	押切もえ公式ブログ
0360	中島彩オフィシャルブログ
0363	滝川クリステル 顔写真
0365	高城剛ブログ
0366	早稲田大学法科大学院トップページ
0367	京都真如堂公式ホームページ
0370	ホテルアンビア松風閣の公式トップページ
0371	日本発達心理学会の日本語トップページ
0372	東京海上日動システムズホームページ
0374	パーキンソン病 ウィキペディア
0378	円柱の体積を求める公式
0379	ハヤシライスの作り方
0380	横浜市役所トップページ
0383	丸井エポスカードトップページ
0384	かに玉の作り方
0386	大和証券グループトップページ
0388	北越銀行ホームページ
0389	寒中見舞いの文例
0390	南ヶ丘牧場トップページ
0393	中村紀洋 ウィキペディア
0394	須磨海浜水族園トップページ
0397	国際会計基準とは
0398	アンハサウェイ 写真
0399	光ファイバーとは

Figure 3: Chinese and Japanese navigational topics.

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.

e.g.
original query: "harry potter" (underspecified)
subtopic string: "harry potter philosophers stone movie"
incorrect: "harry potter hp" (does not specialise)

e.g.
original query: "office" (ambiguous)
subtopic string: "office workplace"
incorrect: "office office" (does not disambiguate; does not specialise)

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:

e.g.
original query: "avp"
subtopic string: "aliens vs predators."

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>." Figure 4 shows a screenshot of our new Subtopic Clustering Interface. This interface lets the assessor form clusters by drag and drop operations, label clusters, and put nonrelevant strings into a special cluster called NONREL.

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 [16], 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. Figure 5 shows a screenshot of our new Cluster Voting Interface.

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 [16].

The number of intents and subtopic strings obtained for Subtopic Mining are shown in Table 6. Due to the aforementioned bugs in our files, the *official* subtopic strings and the *revised* ones differ in number: the statistics for the latter are shown in parentheses. It can be observed that we missed over 1,000 subtopic strings in our *official* English SM evaluation. Similar statistics for INTENT-1 are

Table 5: INTENT-2 reused topics.

TREC TopicID	Chinese TopicID	Japanese TopicID
13	0211	0311
14	0212	0312
21	0213	0313
27	0214	0314
28	0215	0315
34	0216	0316
36	0217	0317
43	0218	0318
44	0219	0319
75	0220	0320
97	0221	0321
9	0222	–
11	0223	–
18	0224	–
32	0226	–
39	0227	–
58	0228	–
61	0229	–
99	0230	–
4	–	0322
20	–	0323
23	–	0324
24	–	0325
31	–	0326
35	–	0327
42	–	0328
45	–	0329
52	–	0330
55	–	0331
60	–	0332
64	–	0333
65	–	0334
72	–	0335
73	–	0336
74	–	0337
78	–	0338
82	–	0339
83	–	0340
92	–	0341
93	–	0342
98	–	0343

shown in Table 7. It can be observed that, despite the use of deeper pools, the the number of subtopic strings obtained at INTENT-2 is considerably smaller, due to the limited number of participants.

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.

Table 8 shows the number of runs submitted to the Subtopic Mining subtask. 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.

Table 6: Statistics of the INTENT-2 topics and intents. Those for the revised Subtopic Mining data are shown in parentheses.

		Subtopic Mining	Document Ranking
English	topics	50	–
	intents	392	–
	subtopic strings	4,157 (5,410)	–
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 (6,253)	–
	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 (2,989)	–
	unique rel docs	–	5,085

Table 7: Statistics of the INTENT-1 topics and intents.

		Subtopic Mining	Document Ranking
Chinese	topics	100	100
	intents	917	917
	subtopics	20,354	–
	unique rel docs	–	23,571
Japanese	topics	100	100
	intents	1,091	1,091
	subtopics	4,103	–
	unique rel docs	–	19,841

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*[3] for Chinese and *ClueWeb09-JA*[4] for Japanese [16]. 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. Table 9 shows the

[3] http://www.sogou.com/labs/dl/t.html
[4] http://lemurproject.org/clueweb09/

Table 8: INTENT-2 Subtopic Mining run types.

	English	Chinese	Japanese
R-runs	–	0	0
B-runs	–	4	7
A-runs	34	19	7

Table 9: INTENT-2 Document Ranking run types.

	Chinese	Japanese
R-runs	1	2
B-runs	7	6
A-runs	4	0

Table 10: 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

Table 11: INTENT-1 relevance assessment statistics.

	Chinese (100 topics)	Japanese (100 topics)
$L4$	1,436	2,201
$L3$	2,557	2,955
$L2$	7,382	6,463
$L1$	12,196	8,222
total	23,571	19,841

number of runs submitted to Document Ranking. The three Revived Runs that are usel for progress monitoring (Figure 2) are: THUIR-D-C-R1 (from Tsinghua University), MSINT-D-J-R1 and MSINT-D-J-R2 (from MSRA), which we shall discuss later.

Figure 6 shows a screenshot of the Per-intent Relevance Assessment Interface which was developed at INTENT-1. Following the reusability study by Sakai *et al.* [12], we increased the pool depth from 20 to 40 at INTENT-2, as was mentioned earlier[5]. Following INTENT-1, every document was judged independently by two assessors, and their assessments were consolidated to form five-point-scale relevance data ($L0$-$L4$). 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 6. Similar statistics for INTENT-1 are shown in Table 7. Also, Tables 10 and 11 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* [14], 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* [11], which we use as supplementary metrics for evaluating Document Ranking.

All metric values reported in this paper were computed using the *NTCIREVAL* toolkit [10][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 [12]. Recall, however, that we have increased the pool depth to 40 for both subtasks of INTENT-2.

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:

$$GG(r) = \sum_i Pr(i|q) g_i(r) \ . \tag{1}$$

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:

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

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:

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

where we let $\gamma = 0.5$ throughout this paper. The advantages of D♯-nDCG over other diversity measures are discussed elsewhere [14, 15].

D-nDCG and D♯-nDCG were originally designed for Document Ranking evaluation, as illustrated in Figure 7(a). 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, as illustrated in Figure 7(b). 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.

[5]Our pooling procedure was actually a little more complex than taking the top 40 documents from every run. First, prior to the run submission deadline, we conducted "pilot" binary relevance assessments for the top 50 documents of our non-diversified baseline runs, to identify some nonrelevant documents in advance. Then, after creating depth-40 pools from the submitted runs, we removed 1,276/1,295 (topic, nonrelevant document) pairs from the Chinese/Japanese pools, to reduce the assessment cost in the latter stage. This is because NII encouraged the organisers to "spread the money spending evenly across the task running period" due to some budget constraints. In short, we judged more than just top 40 documents of every run.

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

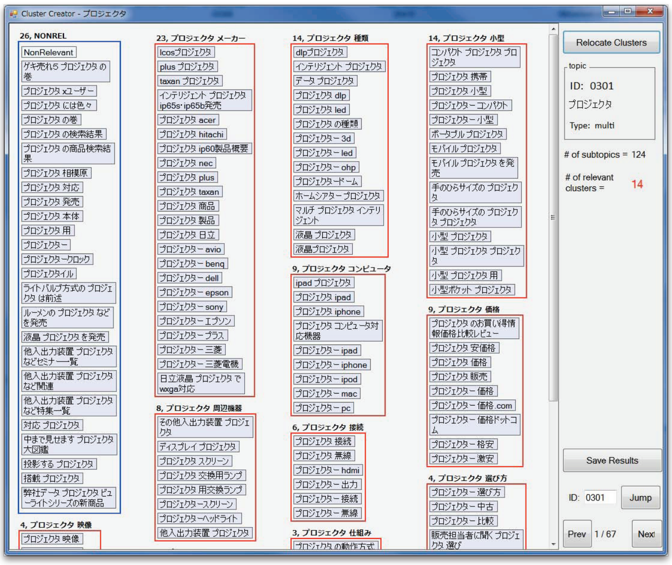

Figure 4: Subtopic Clustering Interface developed by Takehiro Yamamoto (Japanese Topic 0301 "projector").

Figure 5: Cluster Voting Interface developed by Takehiro Yamamoto (Japanese Topic 0301 "projector").

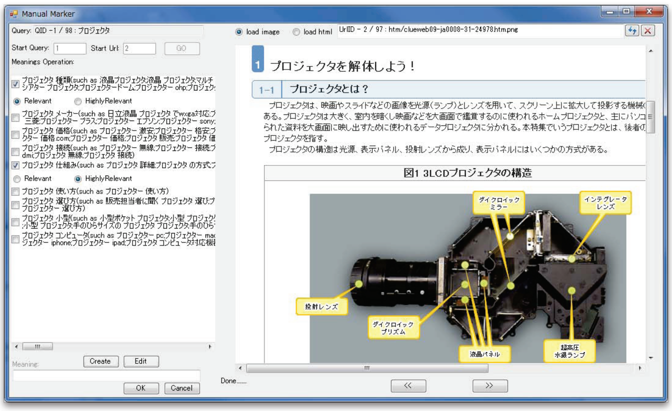

Figure 6: Per-intent Relevance Assessment Interface developed at INTENT-1 by Qinglei Wang (Japanese Topic 0301 "projector").

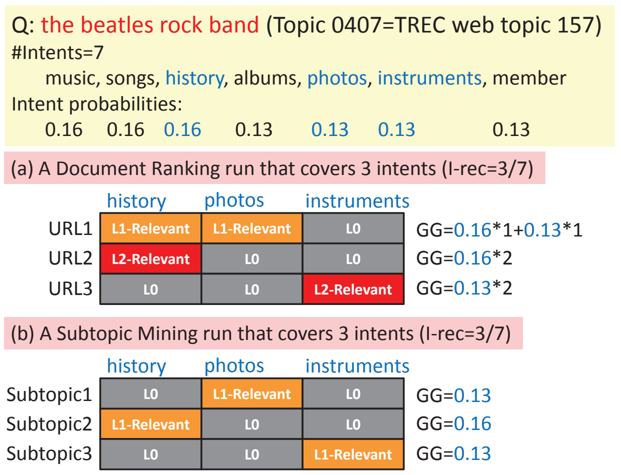

Figure 7: Computing D-nDCG for Document Ranking and Subtopic Mining: examples.

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:

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

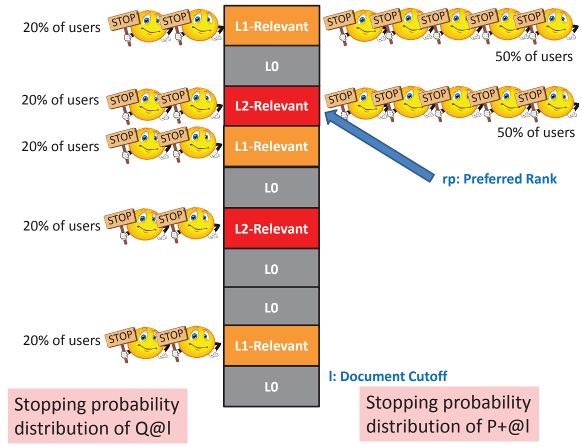

Figure 8: Stopping probability distributions for Q and P$^+$.

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 [13]: Q assumes a uniform distribution across all relevant documents retrieved above l; P$^+$ assumes a uniform distribution across all relevant documents retrieved above rp. Figure 8 illustrates this with an example ranked list with $l = 10$.

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:

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

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

While Sakai [11] 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.

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

Table 12: Discriminative power results for the Subtopic Mining evaluation (randomised two-sided Tukey's HSD test at $\alpha = 0.05$; *official*).

	disc. power	delta
(a) English (50 topics; $34 * 33/2 = 561$ run pairs)		
D♯-nDCG	184/561=32.8%	0.13
I-rec	182/561=32.4%	0.14
D-nDCG	160/561=28.5%	0.14
(b) Chinese (98 topics; $23 * 22/2 = 253$ run pairs)		
D-nDCG	50/253=19.8%	0.09
D♯-nDCG	45/253=17.8%	0.08
I-rec	34/253=13.4%	0.09
(c) Japanese (100 topics; $14 * 13/2 = 91$ run pairs)		
D-nDCG	30/91=33.0%	0.09
D♯-nDCG	26/91=28.6%	0.09
I-rec	25/91=27.5%	0.09

4. OFFICIAL SUBTOPIC MINING RESULTS

As we have mentioned earlier, our official results are based on files that contained some bugs (see Table 6). This section reports on the subtopic mining results *before* the bug fix.

First, Table 12 summarises the *discriminative power* results of I-rec, D-nDCG and D♯-nDCG for the Suptopic Mining evaluation, using a randomised version of the two-sided Tukey's Honestly Significant Differences (HSD) test at $\alpha = 0.05$, with the estimated delta in mean performances required to achieve statistical significance [8, 11]. Discriminative power counts the number of statistically significant differences between run pairs, and reflects the stability of evaluation metrics across topics. The actual significance test results are shown in the Appendix. Detailed results for the English, Chinese and Japanese Subtopic Mining runs are discussed below.

4.1 Official English Subtopic Mining Results

Table 13 shows the mean I-rec, D-nDCG and D♯-nDCG performances of the English Subtopic Mining runs, where mean D♯-nDCG is used as the sort key. Table 27 in the Appendix shows the *SYSDESC fields* [16] of these runs[8]. Figure 9 shows the corresponding I-rec/D-nDCG graph [16]. It can be observed that (a) hultech-S-E-1A is the top performer in terms of relevance (i.e. D-nDCG); (b) THUIR-S-E-1A is the top performer in terms of diversity (i.e. I-rec); and (c) THUIR-S-E-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, hultech, KLE, ORG, SEM12 and THCIB all have at least one run that is statistically indistinguishable from THUIR-S-E-1A (see Figure 33 in the Appendix). Whereas, all runs from LIA and TUTA1 significantly underperform this top run.

According to Table 27, THUIR-S-E-1A combines THUIR-S-E-2A, THUIR-S-E-3A and THUIR-S-E-4A. But the five runs from THUIR are statistically indistinguishable from one another in terms of D(♯)-nDCG and I-rec.

Figure 10 shows the per-topic Min/Max/Average D♯-nDCG performances.

4.2 Official Chinese Subtopic Mining Results

Table 14 shows the mean I-rec, D-nDCG and D♯-nDCG performances of the Chinese Subtopic Mining runs, where mean D♯-nDCG is used as the sort key. Table 28 in the Appendix shows the SYSDESC fields of these runs. Figure 11 shows the corresponding I-rec/D-nDCG graph. It can be observed that (a) THUIR-S-C-3A is the top performer in terms of relevance (i.e. D-nDCG); (b) TUTA1-S-C-1A is the top performer in terms of diversity (i.e. I-rec); and (c) TUTA1-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, ICRCS, KECIR, ORG, THUIR and THUIS (i.e. all of the other teams that participated in Chinese Subtopic Mining) all have at least one run that is statistically indistingushable from TUTA1-S-C-1A (see Figure 37 in the Appendix). In short, the six teams are statistically indistinguishable from one another.

Figure 12 shows the per-topic Min/Max/Average D♯-nDCG performances. The six topics indicated with baloons in the figure, for which the D♯-nDCG values are one, are all *navigational* topics that had exactly one intent (see Figure 3). For these topics, if a system manages to return one relevant subtopic at rank 1, then $I\text{-}rec = 1$; also, recall that D-nDCG reduces to the traditional nDCG. Thus, if the top 10 subtopics strings are all relevant to the query, then $D\text{-}nDCG = 1$ and therefore $D\sharp\text{-}nDCG = (1 + 1)/2 = 1$.

Table 13: English Subtopic Mining runs ranked by mean D♯-nDCG@10 over 50 topics (*official*). The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
THUIR-S-E-1A	**0.4107**	0.3498	**0.3803**
THUIR-S-E-3A	0.3971	0.3492	0.3732
THUIR-S-E-2A	0.3908	0.3506	0.3707
THUIR-S-E-4A	0.3842	0.3517	0.3680
THUIR-S-E-5A	0.3748	0.3550	0.3649
THCIB-S-E-2A	0.3797	0.3499	0.3648
KLE-S-E-4A	0.3951	0.3282	0.3617
THCIB-S-E-1A	0.3785	0.3384	0.3584
hultech-S-E-1A	0.3099	**0.3991**	0.3545
THCIB-S-E-3A	0.3681	0.3383	0.3532
THCIB-S-E-5A	0.3662	0.3215	0.3438
THCIB-S-E-4A	0.3502	0.3323	0.3413
KLE-S-E-2A	0.3772	0.3028	0.3400
hultech-S-E-4A	0.3141	0.3566	0.3353
ORG-S-E-4A	0.3350	0.3156	0.3253
SEM12-S-E-1A	0.3318	0.3094	0.3206
SEM12-S-E-2A	0.3380	0.3020	0.3200
SEM12-S-E-4A	0.3328	0.2994	0.3161
SEM12-S-E-5A	0.3259	0.2977	0.3118
ORG-S-E-3A	0.3366	0.2842	0.3104
KLE-S-E-3A	0.3140	0.2895	0.3018
KLE-S-E-1A	0.2954	0.2719	0.2836
ORG-S-E-2A	0.2789	0.2564	0.2677
SEM12-S-E-3A	0.2933	0.2258	0.2595
hultech-S-E-3A	0.2475	0.2498	0.2486
ORG-S-E-1A	0.2398	0.2203	0.2300
ORG-S-E-5A	0.2532	0.1976	0.2254
hultech-S-E-2A	0.2263	0.2180	0.2221
TUTA1-S-E-1A	0.1892	0.1756	0.1824
LIA-S-E-4A	0.1655	0.1740	0.1698
TUTA1-S-E-2A	0.1724	0.1569	0.1646
LIA-S-E-2A	0.0278	0.0380	0.0329
LIA-S-E-3A	0.0298	0.0261	0.0280
LIA-S-E-1A	0.0213	0.0296	0.0255

[8]INTENT run file formats are similar to TREC, except that every file is required to start with a brief system description.

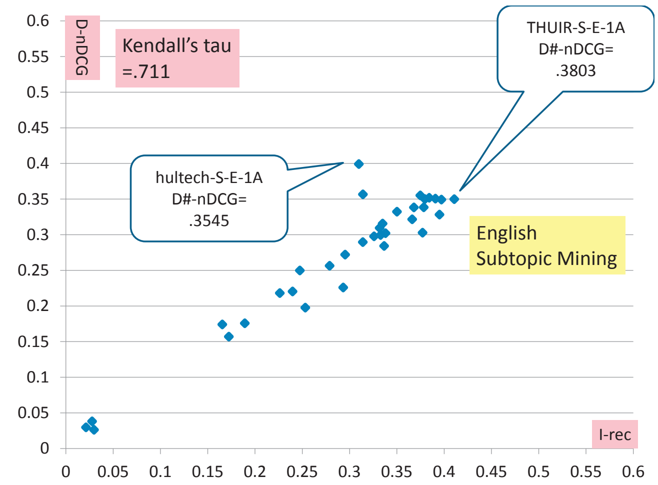

Figure 9: I-rec/D-nDCG graph for English Subtopic Mining (*official*).

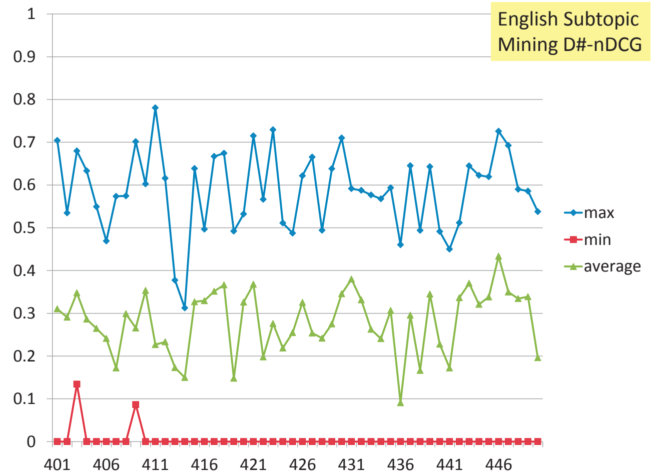

Figure 10: Per-topic D♯-nDCG performances for English Subtopic Mining (*official*).

4.3 Official Japanese Subtopic Mining Results

Table 15 shows the mean I-rec, D-nDCG and D♯-nDCG performances of the Japanese Subtopic Mining runs, where mean D♯-nDCG is used as the sort key. Table 29 in the Appendix shows the SYSDESC fields of these runs. Figure 13 shows the corresponding I-rec/D-nDCG graph. It can be observed that (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[9]. More generally, in terms of D♯-nDCG, both KLE and MSINT (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 (see Figure 41 in the Appendix). In short, the three teams are statistically indistinguishable from one another.

[9]ORG-S-J-3A uses Google query completions while ORG-S-J-5A combines query completions from multiple search engines including Google (see Table 29 in the Appendix).

Table 14: Chinese Subtopic Mining runs ranked by mean D♯-nDCG@10 over 98 topics (*official*). The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
TUTA1-S-C-1A	**0.4184**	0.4686	**0.4435**
THUIS-S-C-1A	0.3881	0.4923	0.4402
THUIR-S-C-3A	0.3786	**0.4987**	0.4386
TUTA1-S-C-2A	0.4030	0.4655	0.4343
THUIS-S-C-4A	0.4036	0.4620	0.4328
THUIR-S-C-5A	0.3892	0.4757	0.4324
THUIR-S-C-1A	0.3839	0.4802	0.4321
THUIR-S-C-2A	0.3839	0.4775	0.4307
THUIR-S-C-4A	0.3792	0.4698	0.4245
ICRCS-S-C-3A	0.4046	0.4413	0.4229
THUIS-S-C-3A	0.3953	0.4504	0.4228
ICRCS-S-C-1A	0.3821	0.4219	0.4020
ORG-S-C-1A	0.3644	0.4336	0.3990
ORG-S-C-4A	0.3334	0.4516	0.3925
THUIS-S-C-2A	0.3622	0.4157	0.3890
ORG-S-C-3A	0.3366	0.4407	0.3886
ICRCS-S-C-2A	0.3704	0.4024	0.3864
KECIR-S-C-2B	0.3743	0.3941	0.3842
ORG-S-C-5A	0.3091	0.4175	0.3633
ORG-S-C-2A	0.3163	0.4098	0.3630
KECIR-S-C-1B	0.3341	0.3763	0.3552
KECIR-S-C-3B	0.3001	0.3227	0.3114
KECIR-S-C-4B	0.2917	0.3081	0.2999

Table 15: Official Japanese Subtopic Mining runs ranked by mean D♯-nDCG@10 over 100 topics (*official*). The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
ORG-S-J-3A	0.3331	**0.3150**	**0.3241**
MSINT-S-J-4A	0.2988	0.3085	0.3036
MSINT-S-J-1B	0.2969	0.3058	0.3013
ORG-S-J-5A	**0.3353**	0.2469	0.2911
MSINT-S-J-3A	0.2746	0.2980	0.2863
ORG-S-J-1A	0.2753	0.2868	0.2811
KLE-S-J-1B	0.2596	0.2639	0.2618
KLE-S-J-3B	0.2518	0.2715	0.2617
MSINT-S-J-2B	0.2659	0.2494	0.2576
ORG-S-J-2A	0.2089	0.2602	0.2345
MSINT-S-J-5B	0.2354	0.2335	0.2344
KLE-S-J-4B	0.2135	0.1658	0.1897
KLE-S-J-2B	0.2034	0.1638	0.1836
ORG-S-J-4A	0.1037	0.1071	0.1054

Figure 14 shows the per-topic Min/Max/Average D♯-nDCG performances. Again, the ten topics indicated with baloons in the figure, for which the D♯-nDCG values are one, are all *navigational* topics that had exactly one intent (See Figure 3).

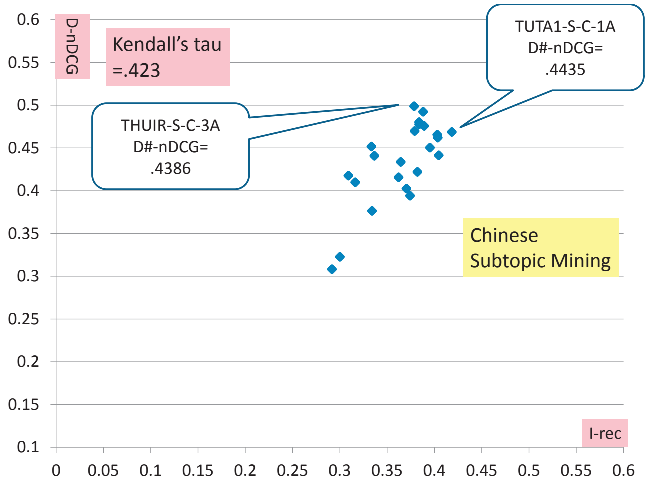

Figure 11: I-rec/D-nDCG graph for Chinese Subtopic Mining (*official*).

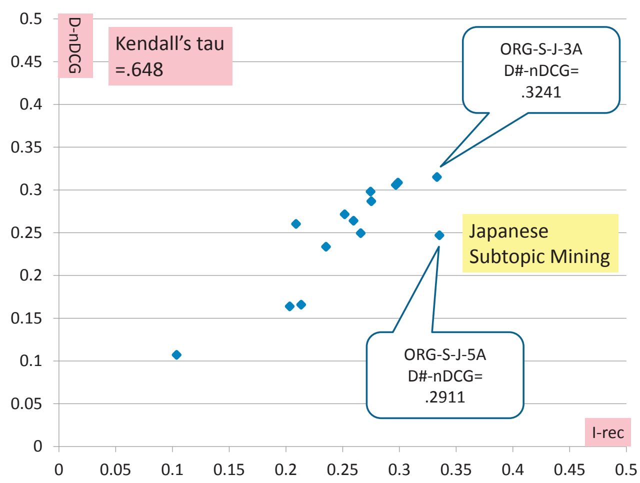

Figure 13: I-rec/D-nDCG graph for Japanese Subtopic Mining (*official*).

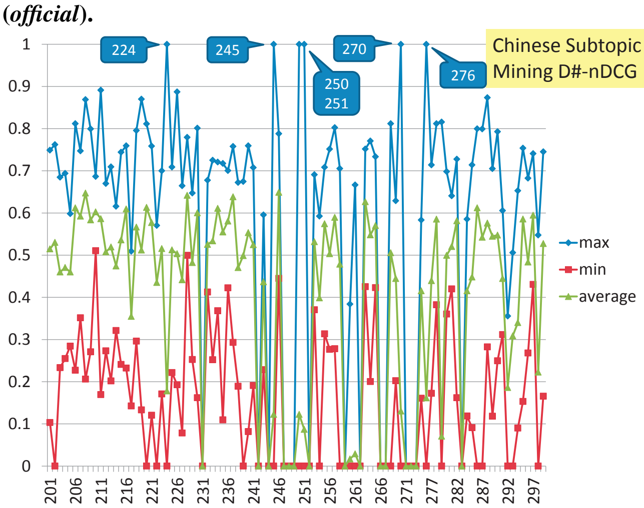

Figure 12: Per-topic D♯-nDCG performances for Chinese Subtopic Mining (*official*).

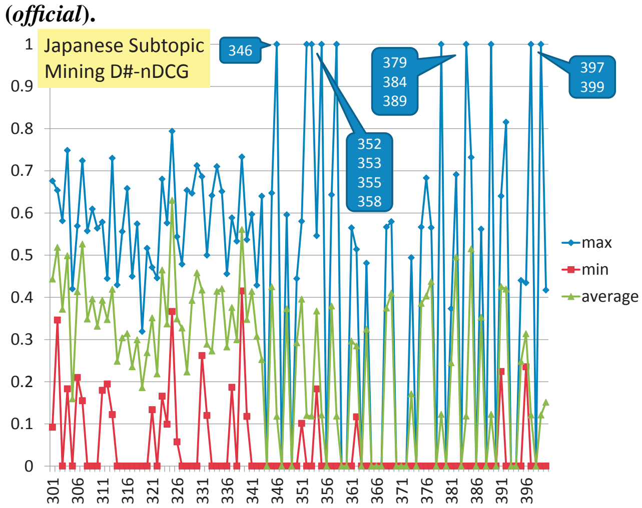

Figure 14: Per-topic D♯-nDCG performances for Japanese Subtopic Mining (*official*).

Table 16: Kendall's rank correlation between the official ranking and the revised ranking.

	I-rec@10	D-nDCG@10	D♯-nDCG
English	.943	.900	.914
Chinese	1	.992	1
Japanse	1	1/1	.978

Table 17: Discriminative power results for the Subtopic Mining evaluation (randomised two-sided Tukey's HSD test at $\alpha = 0.05$; *revised*).

	disc. power	delta
(a) English (50 topics; $34 * 33/2 = 561$ run pairs)		
I-rec	186/561=33.2%	0.14
D♯-nDCG	182/561=32.4%	0.14
D-nDCG	174/561=31.0%	0.16
(b) Chinese (98 topics; $23 * 22/2 = 253$ run pairs)		
D-nDCG	52/253=20.6%	0.09
D♯-nDCG	45/253=17.8%	0.08
I-rec	34/253=13.4%	0.09
(c) Japanese (100 topics; $14 * 13/2 = 91$ run pairs)		
D-nDCG	28/91=30.8%	0.09
D♯-nDCG	26/91=28.6%	0.09
I-rec	25/91=27.5%	0.09

Table 18: Comparison of significance test results between *official* and *revised* (randomised two-sided Tukey's HSD test at $\alpha = 0.05$).

	official−revised	official∩revised	revised−official
(a) English (50 topics; $34 * 33/2 = 561$ run pairs)			
I-rec	7	175	11
D-nDCG	15	145	29
D♯-nDCG	14	170	12
(b) Chinese (98 topics; $23 * 22/2 = 253$ run pairs)			
I-rec	0	34	0
D-nDCG	0	50	0
D♯-nDCG	0	45	0
(c) Japanese (100 topics; $14 * 13/2 = 91$ run pairs)			
I-rec	0	25	0
D-nDCG	2	28	0
D♯-nDCG	0	26	0

5. REVISED SUBTOPIC MINING RESULTS

As we have mentioned earlier, our official results are based on files that contained some bugs (see Table 6). This section reports on the subtopic mining results *after* the bug fix.

Tabl 16 compares the run rankings before and after the bug fix in terms of Kendall's rank correlation. Note that a rank correlation of one means identical rankings. Unfortunately, it can be observed that the bugs did affect the rankings. The effect on the English results are larger than that on Chinese and Japanese, reflecting the number of subtopic strings that we originally missed (see Table 6).

Table 17 summarises the *discriminative power* results of I-rec, D-nDCG and D♯-nDCG for the Suptopic Mining evaluation, in a way similar to Table 12. The actual significance test results are shown in the Appendix. Table 18 compares the significance test results before and after the bug fix. For example, "official−revised" represents run pairs that were significantly different in the *official* results but not in the *revised* results. It can be observed that there are considerable discrepancies for the English results, and that two significantly different pairs were lost after the bug fix for the Japanese results with D-nDCG; the Chinese results are not affected at all. Run pairs that showed discordant significance test results are also listed up in the Appendix.

Table 19: English Subtopic Mining runs ranked by mean D♯-nDCG@10 over 50 topics (*revised*). The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
THUIR-S-E-4A	0.4364	0.5062	**0.4713**
THUIR-S-E-1A	**0.4512**	0.4775	0.4644
THUIR-S-E-5A	0.4253	0.4893	0.4573
THUIR-S-E-2A	0.4333	0.4795	0.4564
THCIB-S-E-1A	0.4431	0.4657	0.4544
THUIR-S-E-3A	0.4346	0.4726	0.4536
THCIB-S-E-2A	0.4308	0.4744	0.4526
hultech-S-E-1A	0.3680	**0.5368**	0.4524
KLE-S-E-4A	0.4457	0.4401	0.4429
THCIB-S-E-3A	0.4248	0.4557	0.4403
THCIB-S-E-4A	0.4100	0.4521	0.4310
THCIB-S-E-5A	0.4144	0.4441	0.4292
hultech-S-E-4A	0.3688	0.4807	0.4248
KLE-S-E-2A	0.4292	0.4159	0.4225
SEM12-S-E-2A	0.3777	0.4250	0.4014
SEM12-S-E-1A	0.3780	0.4233	0.4007
ORG-S-E-4A	0.3815	0.3829	0.3822
ORG-S-E-3A	0.3841	0.3735	0.3788
KLE-S-E-3A	0.3676	0.3661	0.3668
SEM12-S-E-4A	0.3727	0.3471	0.3599
SEM12-S-E-5A	0.3659	0.3445	0.3552
KLE-S-E-1A	0.3529	0.3540	0.3535
SEM12-S-E-3A	0.3403	0.3573	0.3488
ORG-S-E-5A	0.3181	0.3365	0.3273
ORG-S-E-2A	0.3268	0.3231	0.3250
hultech-S-E-3A	0.3045	0.3345	0.3195
ORG-S-E-1A	0.2787	0.3068	0.2927
hultech-S-E-2A	0.2697	0.2986	0.2841
TUTA1-S-E-1A	0.2181	0.2577	0.2379
LIA-S-E-4A	0.2000	0.2753	0.2376
TUTA1-S-E-2A	0.1865	0.2327	0.2096
LIA-S-E-2A	0.0328	0.0474	0.0401
LIA-S-E-1A	0.0291	0.0420	0.0355
LIA-S-E-3A	0.0377	0.0329	0.0353

5.1 Revised English Subtopic Mining Results

Table 19 shows the mean I-rec, D-nDCG and D♯-nDCG performances of the English Subtopic Mining runs, where mean D♯-nDCG is used as the sort key. Table 27 in the Appendix shows the *SYSDESC fields* of these runs. Figure 15 shows the corresponding I-rec/D-nDCG graph. It can be observed that (a) hultech-S-E-1A is the top performer in terms of relevance (i.e. D-nDCG); (b) THUIR-S-E-1A is the top performer in terms of diversity (i.e. I-rec); and (c) THUIR-S-E-4A is the overall winner in terms of D♯-nDCG, whereas the *official* overall winner was THUIR-S-E-1A. However, these three runs are statistically indistinguishable in terms of D♯-nDCG. More generally, in terms of D♯-nDCG, hultech, KLE, ORG, SEM12 and THCIB all have at least one run that is statistically indistinguishable from THUIR-S-E-4A (see Figure 34 in the Appendix). Whereas, all runs from LIA and TUTA1 significantly underperform this top run.

According to Table 27, THUIR-S-E-1A combines THUIR-S-E-2A, THUIR-S-E-3A and THUIR-S-E-4A. But the five runs from THUIR are statistically indistinguishable from one another in terms of D(♯)-nDCG and I-rec.

The above main findings with the *revised* English Subtopic Mining results are the same as the ones with the *official* results, except that the top performer in terms of mean D♯-nDCG is now THUIR-S-E-4A instead of THUIR-S-E-1A. But recall that the five THUIR runs are statistically indistinguishable from one another with both *official* and *revised* results.

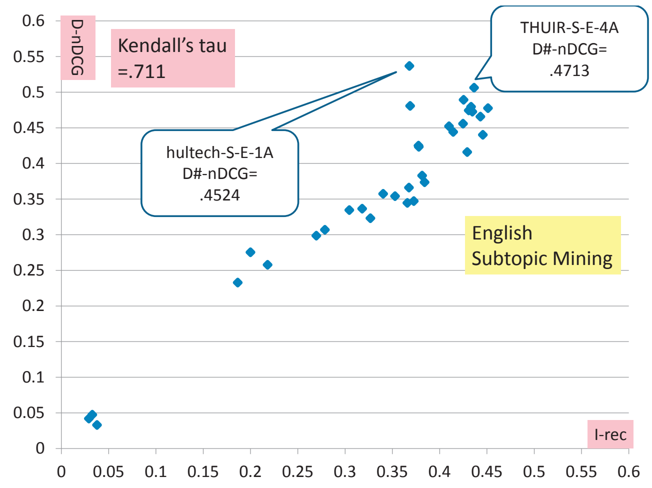

Figure 15: I-rec/D-nDCG graph for English Subtopic Mining (*revised*).

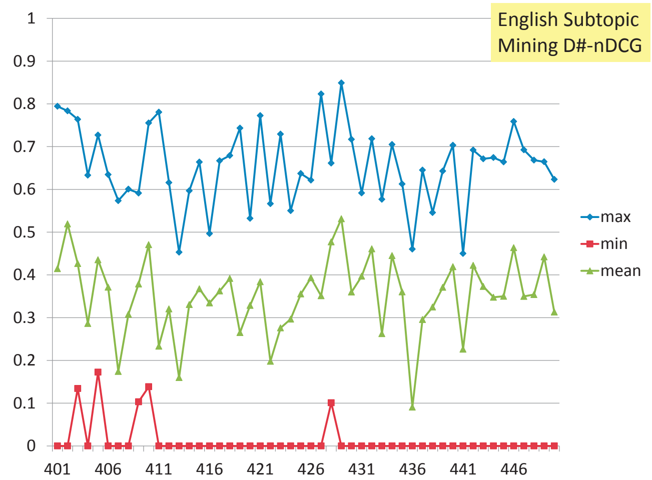

Figure 16: Per-topic D♯-nDCG performances for English Subtopic Mining (*revised*).

Table 20: Chinese Subtopic Mining runs ranked by mean D♯-nDCG@10 over 98 topics (*revised*). The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
TUTA1-S-C-1A	**0.4184**	0.4714	**0.4449**
THUIS-S-C-1A	0.3881	0.4963	0.4422
THUIR-S-C-3A	0.3786	**0.5028**	0.4407
TUTA1-S-C-2A	0.4030	0.4694	0.4362
THUIS-S-C-4A	0.4036	0.4658	0.4347
THUIR-S-C-5A	0.3892	0.4798	0.4345
THUIR-S-C-1A	0.3839	0.4843	0.4341
THUIR-S-C-2A	0.3839	0.4816	0.4327
THUIR-S-C-4A	0.3792	0.4739	0.4266
ICRCS-S-C-3A	0.4046	0.4440	0.4243
THUIS-S-C-3A	0.3953	0.4531	0.4242
ICRCS-S-C-1A	0.3821	0.4258	0.4039
ORG-S-C-1A	0.3644	0.4361	0.4003
ORG-S-C-4A	0.3334	0.4540	0.3937
THUIS-S-C-2A	0.3622	0.4194	0.3908
ORG-S-C-3A	0.3366	0.4440	0.3903
ICRCS-S-C-2A	0.3704	0.4044	0.3874
KECIR-S-C-2B	0.3743	0.3965	0.3854
ORG-S-C-5A	0.3091	0.4210	0.3650
ORG-S-C-2A	0.3163	0.4137	0.3650
KECIR-S-C-1B	0.3341	0.3799	0.3570
KECIR-S-C-3B	0.3001	0.3231	0.3116
KECIR-S-C-4B	0.2917	0.3085	0.3001

Table 21: Japanese Subtopic Mining runs ranked by mean D♯-nDCG@10 over 100 topics (*revised*). The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
ORG-S-J-3A	0.3331	**0.3150**	**0.3241**
MSINT-S-J-4A	0.2988	0.3085	0.3036
MSINT-S-J-1B	0.2969	0.3058	0.3013
ORG-S-J-5A	**0.3353**	0.2469	0.2911
MSINT-S-J-3A	0.2746	0.2980	0.2863
ORG-S-J-1A	0.2753	0.2868	0.2811
KLE-S-J-1B	0.2607	0.2656	0.2632
KLE-S-J-3B	0.2529	0.2726	0.2628
MSINT-S-J-2B	0.2659	0.2494	0.2576
MSINT-S-J-5B	0.2370	0.2341	0.2356
ORG-S-J-2A	0.2089	0.2610	0.2349
KLE-S-J-4B	0.2146	0.1687	0.1917
KLE-S-J-2B	0.2034	0.1667	0.1851
ORG-S-J-4A	0.1037	0.1071	0.1054

5.2 Revised Chinese Subtopic Mining Results

Table 20 shows the mean I-rec, D-nDCG and D♯-nDCG performances of the Chinese Subtopic Mining runs, where mean D♯-nDCG is used as the sort key. Table 28 in the Appendix shows the SYSDESC fields of these runs. Figure 17 shows the corresponding I-rec/D-nDCG graph, which is virtually indistinguishable from Figure 11 (*official* results). It can be observed that (a) THUIR-S-C-3A is the top performer in terms of relevance (i.e. D-nDCG); (b) TUTA1-S-C-1A is the top performer in terms of diversity (i.e. I-rec); and (c) TUTA1-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, ICRCS, KECIR, ORG, THUIR and THUIS (i.e. all of the other teams that participated in Chinese Subtopic Mining) all have at least one run that is statistically indistingushable from TUTA1-S-C-1A (see Figure 37 in the Appendix). In short, the six teams are statistically indistinguishable from one another.

All of the above findings are in agreement with the *official* results. Recall that the significant test results with D(♯)-nDCG and I-rec are identical before and after the bug fix for Chinese Subtopic Mining.

Figure 18 shows the per-topic Min/Max/Average D♯-nDCG performances, which is virtually indistinguishable from Figure 12 (the *official* results). The same six topics indicated with baloons in the figure, for which the D♯-nDCG values are one, are all *navigational* topics that had exactly one intent (see Figure 3).

5.3 Revised Japanese Subtopic Mining Results

Table 21 shows the mean I-rec, D-nDCG and D♯-nDCG performances of the Japanese Subtopic Mining runs, where mean D♯-nDCG is used as the sort key. Table 29 in the Appendix shows the SYSDESC fields of these runs. Figure 19 shows the corresponding I-rec/D-nDCG graph, which is virtually indistinguishable from Figure 11 (*official* results). It can be observed that (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*

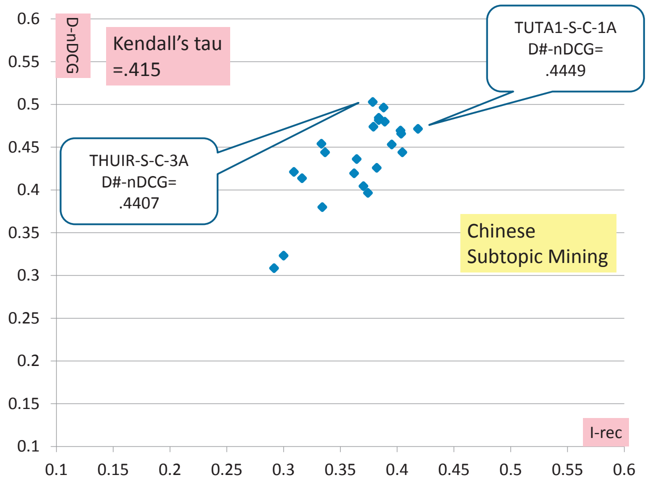

Figure 17: I-rec/D-nDCG graph for Chinese Subtopic Mining (*revised*).

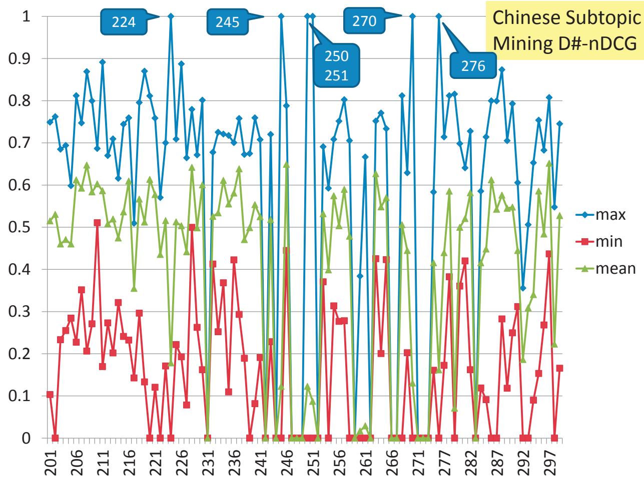

Figure 18: Per-topic D♯-nDCG performances for Chinese Subtopic Mining (*revised*).

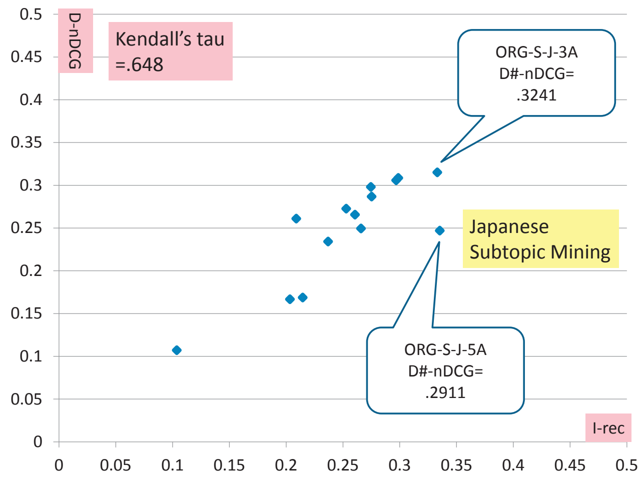

Figure 19: I-rec/D-nDCG graph for Japanese Subtopic Mining (*revised*).

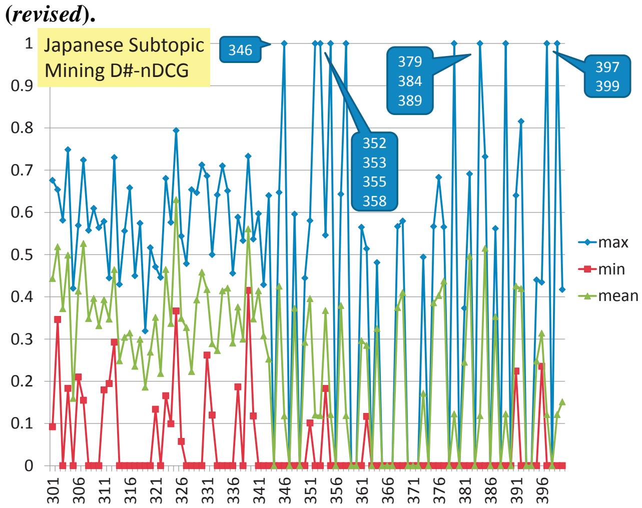

Figure 20: Per-topic D♯-nDCG performances for Japanese Subtopic Mining (*revised*).

statistically significant[10]. More generally, in terms of D♯-nDCG, both KLE and MSINT (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 (see Figure 41 in the Appendix). In short, the three teams are statistically indistinguishable from one another.

All of the above findings are in agreement with the *official* results. Recall that the significant test results with D♯-nDCG are identical before and after the bug fix for Japanese Subtopic Mining.

Figure 20 shows the per-topic Min/Max/Average D♯-nDCG performances, which is virtually indistinguishable from Figure 14 (the *official* results). The same ten topics indicated with baloons in the figure, for which the D♯-nDCG values are one, are all *navigational* topics that had exactly one intent (See Figure 3).

[10] ORG-S-J-3A uses Google query completions while ORG-S-J-5A combines query completions from multiple search engines including Google (see Table 29 in the Appendix).

Table 22: Discriminative power results for the Document Ranking evaluation (randomised two-sided Tukey's HSD test at $\alpha = 0.05$).

	disc. power	delta
(a) Chinese (97 topics; $12 * 11/2 = 66$ run pairs)		
I-rec	32/66=48.5%	0.09
D♯-nDCG	24/66=36.4%	0.08
D-nDCG	11/66=16.7%	0.09
P+Q	11/66=16.7%	0.09
DIN-nDCG	8/66=12.1%	0.07
(b) Japanese (95 topics; $8 * 7/2 = 28$ run pairs)		
DIN-nDCG	9/28=32.1%	0.04
D-nDCG	8/28=28.6%	0.04
D♯-nDCG	7/28=25.0%	0.05
P+Q	4/28=14.3%	0.04
I-rec	2/28=7.1%	0.08

6. DOCUMENT RANKING RESULTS

First, Table 22 summarises the discriminative power results of I-rec, D-nDCG, D♯-nDCG, DIN-nDCG and P+Q for the Document Ranking evaluation, using a randomised version of the two-sided Tukey's HSD test at $\alpha = 0.05$, with the estimated delta in mean performances required to achieve statistical significance. The actual significance test results are shown in the Appendix. Detailed results for the Chinese and Japanese Document Ranking runs are discussed below.

6.1 Chinese Document Ranking Results

Table 23 shows the mean I-rec, D-nDCG, D♯-nDCG performances of the Chinese Document Ranking runs, where mean D♯-nDCG is used as the sort key. In addition, the mean performances according to the intent type-sensitive metrics DIN-nDCG and P+Q are also shown. Table 30 in the Appendix shows the SYSDESC fields of these runs. Figure 21 shows the corresponding I-rec/D-nDCG graph. It can be observed that THUIR-D-C-1A is the winner in terms of all five metrics. In terms of D♯-nDCG, it significantly outperforms BASELINE-D-C-1 ($p \leq 0.001$). However, KECIR has two runs that are statistically indistinguishable from THUIR-D-C-1A in terms of D♯-nDCG (See Figure 44 in the Appendix). According to Table 30, THUIR-D-C-1A applies click-based reranking to THUIR-D-C-2A, but the gain in D♯-nDCG is not statistically significant.

Unfortunately, none of the new runs from THUIR significantly outperforms its Revived Run THUIR-D-C-R1 (see Figure 2). Therefore, we cannot conclude from these experiments that there has been substantial progress compared to INTENT-1.

Figure 22 shows the per-topic Min/Max/Average D♯-nDCG performances. The five topics indicated with baloons, for which the Maxium D♯-nDCG values were the highest, are all navigational topics with only one intent. Recall that for such topics, D-nDCG reduces to nDCG, and that it suffices to retrieve just one relevant document to achieve an I-rec of one.

Figure 23 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 6): recall that DIN-nDCG is equal to D-nDCG if all subtopics are informational.

Figure 24 compares the per-topic D-nDCG/DIN-nDCG/P+Q values for THUIR-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 $L4$-relevant document (i.e. document with the highest relevance level) is retrieved at rank 1, P+Q equals one for these topics.

Table 24 compares the performances of our Revived Run, THUIR-D-C-R1, across INTENT-1 and INTENT-2 (see Figure 2). We used a two-sample unpaired bootstrap test [8] 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.

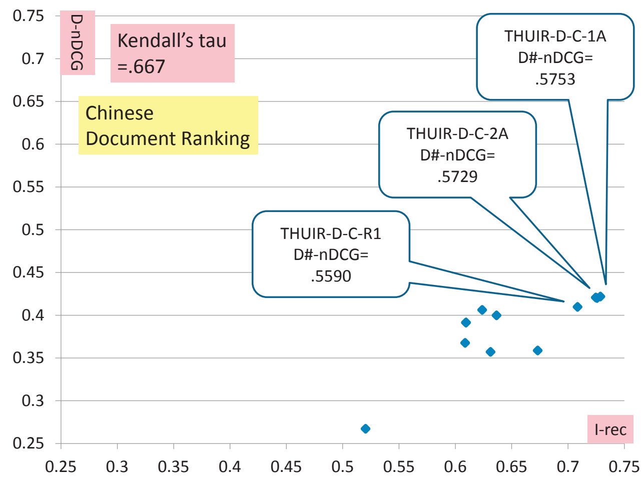

Figure 21: I-rec/D-nDCG graph for Chinese Document Ranking.

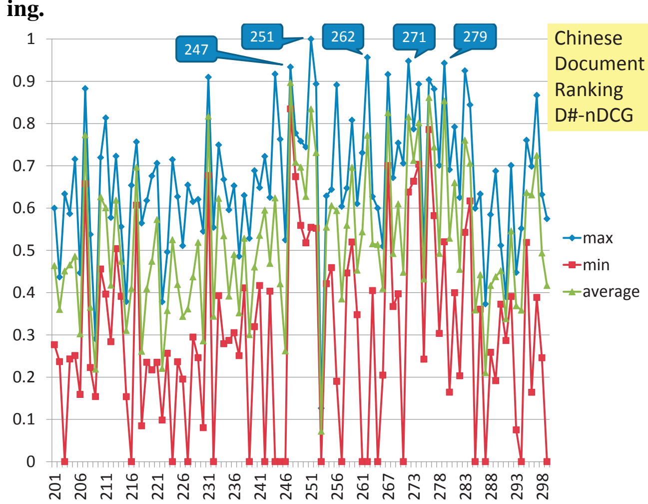

Figure 22: Per-topic D♯-nDCG performances for Chinese Document Ranking.

Table 23: Chinese Document Ranking runs ranked by mean D♯-nDCG@10 over 97 topics. The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10	DIN-nDCG@10	P+Q
THUIR-D-C-1A	**0.7288**	**0.4218**	**0.5753**	**0.2868**	**0.2667**
THUIR-D-C-2A	0.7258	0.4201	0.5729	0.2865	0.2663
THUIR-D-C-3A	0.7247	0.4207	0.5727	0.2858	0.2653
THUIR-D-C-R1	0.7085	0.4096	0.5590	0.2806	0.2569
KECIR-D-C-3B	0.6366	0.3998	0.5182	0.2789	0.2218
THUIR-D-C-4A	0.6731	0.3587	0.5159	0.2611	0.2203
KECIR-D-C-5B	0.6239	0.4062	0.5150	0.2803	0.2320
KECIR-D-C-4B	0.6095	0.3914	0.5005	0.2741	0.2134
KECIR-D-C-1B	0.6095	0.3914	0.5005	0.2741	0.2134
THUIR-D-C-5B	0.6313	0.3571	0.4942	0.2406	0.2298
BASELINE-D-C-1	0.6087	0.3676	0.4882	0.2485	0.2340
KECIR-D-C-2B	0.5204	0.2672	0.3938	0.2120	0.1331

Table 24: THUIR 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.

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

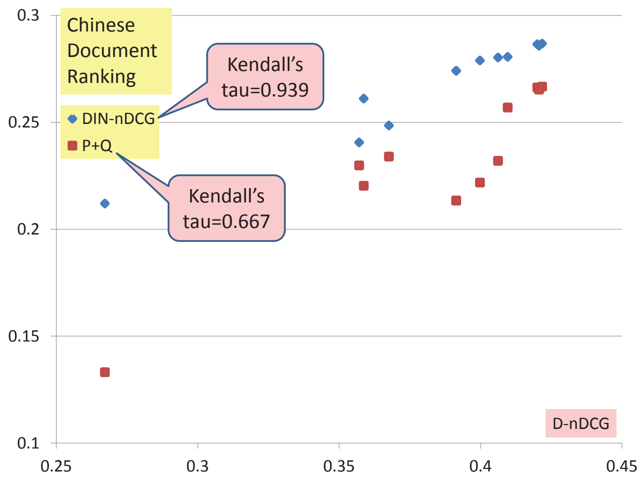

Figure 23: Correlation between D-nDCG and DIN-nDCG/P+Q for Chinese Document Ranking.

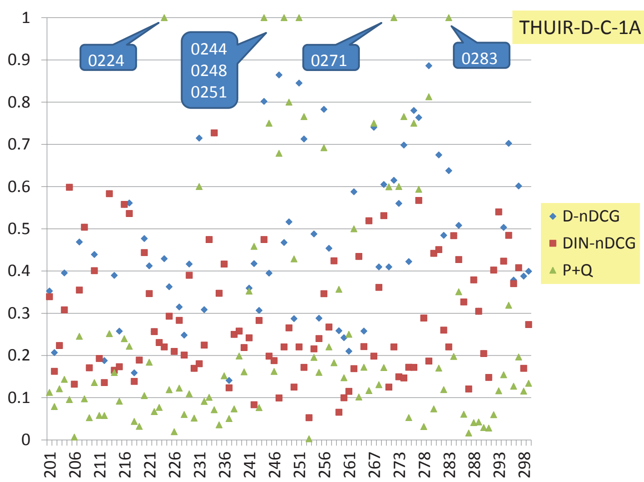

Figure 24: Per-topic D-nDCG/DIN-nDCG/P+Q performances for THUIR-D-C-1A.

6.2 Japanese Document Ranking Results

Table 25 shows the mean I-rec, D-nDCG, D♯-nDCG performances of the Japanese Document Ranking runs, where mean D♯-nDCG is used as the sort key. In addition, the mean performances according to the intent type-sensitive metrics DIN-nDCG and P+Q are also shown. Table 31 in the Appendix shows the SYSDESC fields of these runs. Figure 25 shows the corresponding I-rec/D-nDCG graph. It can be observed that MSINT-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 MSINT runs (see Figure 49). In particular, MSINT-D-J-4B significantly outperforms its Revived Runs MSINT-D-J-R1 and MSINT-D-J-R2 ($p \leq 0.001$), which suggests that the method may be substantially better than those used at INTENT-1. According to Table 31, MSINT-D-J-4B combined search results of the baseline, Yahoo! and Bing, and this seems to have been successful.

Figure 26 shows the per-topic Min/Max/Average D♯-nDCG performances. Topics 0353, 0383, 0398, 0399, indicated with baloons, for which the Maxium D♯-nDCG values were very high, are again all navigational topics with only one intent. On the other hand, Topic 0350, the second "easiest" topic, had as many as eight intents. The reason why this topic was easy is probably because it happened that none of its 257 relevant documents is relevant to multiple intents. Thus the problem is similar to traditional relevance-based retrieval, where the system is asked to return a union of eight different sets of relevant documents, which do not overlap with one another.

Figure 27 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 6).

Figure 28 compares the per-topic D-nDCG/DIN-nDCG/P+Q values for MSINT-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 $L4$-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 $L4$-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 non-relevant, it does not give a full score to the run. This is a known normalisation issue with DIN-nDCG [11].

Table 26 compares the performances of our Revived Runs, MSINT-D-J-R2 and MSINT-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.

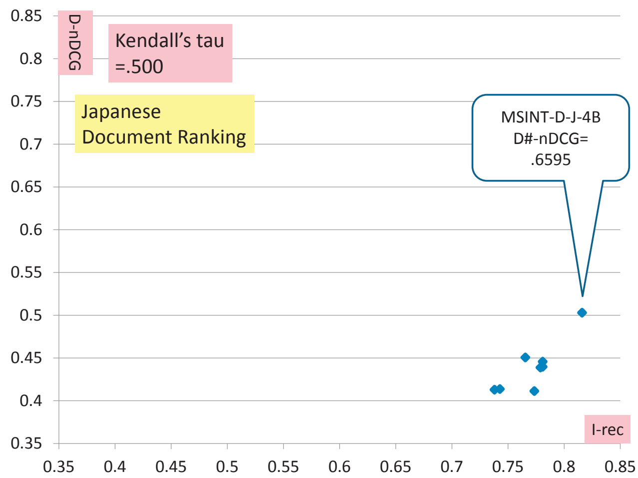

Figure 25: I-rec/D-nDCG graph for Japanese Document Ranking.

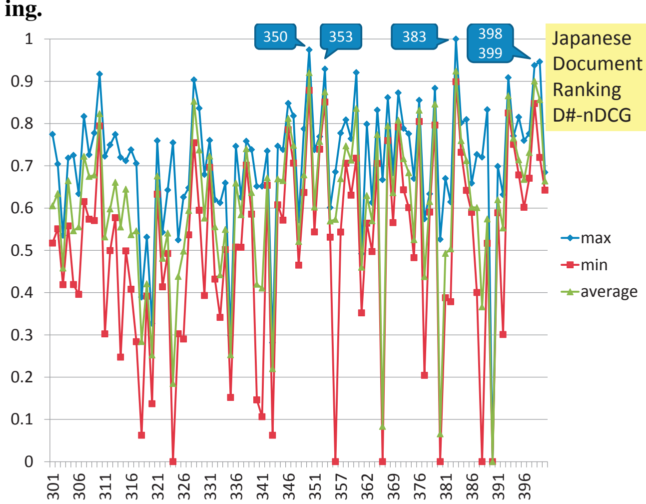

Figure 26: Per-topic D♯-nDCG performances for Japanese Document Ranking.

Table 25: Japanese Document Ranking runs ranked by mean D♯-nDCG@10 over 95 topics. The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10	DIN-nDCG@10	P+Q
MSINT-D-J-4B	**0.8160**	**0.5029**	**0.6595**	**0.3458**	**0.3666**
MSINT-D-J-3B	0.7809	0.4457	0.6133	0.3182	0.3373
MSINT-D-J-5B	0.7809	0.4397	0.6103	0.3124	0.3282
MSINT-D-J-1B	0.7789	0.4388	0.6089	0.3099	0.3248
MSINT-D-J-2B	0.7655	0.4505	0.6080	0.3159	0.3271
MSINT-D-J-R2	0.7735	0.4113	0.5924	0.2994	0.3273
BASELINE-D-J-1	0.7428	0.4136	0.5782	0.2820	0.3160
MSINT-D-J-R1	0.7380	0.4129	0.5754	0.2861	0.3154

Table 26: MSINT 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.

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

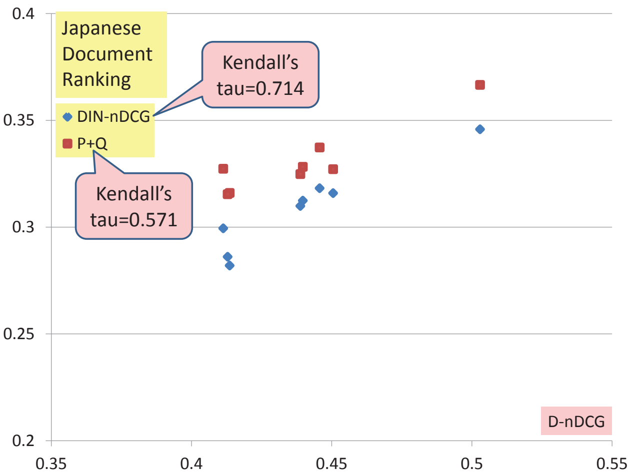

Figure 27: Correlation between D-nDCG and DIN-nDCG/P+Q for Japanese Document Ranking.

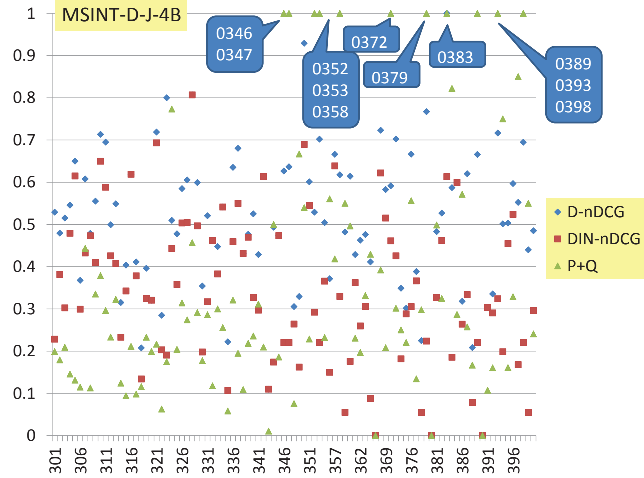

Figure 28: Per-topic D-nDCG/DIN-nDCG/P+Q performances for MSINT-D-J-4B.

7. 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 received 12 Chinese runs and 8 Japanese runs. We refer the reader to the INTENT-2 participants' papers for details of their runs[3, 4, 5, 7, 17, 18, 19, 20, 21, 22]. Our main findings are:

English Subtopic Mining In the *official* results, THUIR-S-E-1A outperformed all other runs in terms of Mean D♯-nDCG, but hultech, KLE, ORG, 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 underperform THUIR-S-E-1A. The *revised* results are the same, except that the top performer in terms of Mean D♯-nDCG is THUIR-S-E-4A instead of THUIR-S-E-1A. But the THUIR runs are statistically indistinguishable from one another with both *official* and *revised* results.

Chinese Subtopic Mining In the *official* results, 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. The *revised* results are the same.

Japanese Subtopic Mining In the *official* results, 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. The *revised* results are the same.

Chinese Document Ranking 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 outperforms its Revived Run THUIR-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 MSINT-D-J-4B outperformed all other runs in terms of Mean D♯-nDCG. In particular, it significantly outperforms its Revived Runs MSINT-D-J-R1 and MSINT-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, 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.

Given the lack of popularity of the Document Ranking Subtask (especially for Japanese, where only one team participated), we do not have a strong reason to continue this subtask. On the other hand, it should be noted that the TREC Web Track has discontinued their diversity task. Note also that it is dangerous to assume that diversity test collections are reusable, as they are constructed using a shallow pool depth (e.g., 20-30, although INTENT-2 used 40) [12]. Thus, if researchers want to continue to have their diversified search systems evaluated fairly, the IR community probably does need to continue a diversity task/track.

Recall also that our English topic set is identical to the TREC 2012 Web Track topic set: for each topic, we have our own set of intents, while TREC has their own set of "subtopics." We will leverage the data to conduct an analysis across TREC and NTCIR elsewhere.

8. ACKNOWLEDGMENTS

We would like to thank the following people/institutions:

- INTENT-2 participants for their research efforts and cooperation;
- Charlie Clarke and Ellen Voorhees for providing the TREC 2012 web topics;
- The NTCIR general and programme chairs for their support;
- Jaime Callan for providing the ClueWeb09-JA document collection;
- Sogou.com for providing SogouT, SogouQ and other related resources.

9. ADDITIONAL AUTHORS

Additional authors: Makoto P. Kato (Kyoto University, Japan email: kato@dl.kuis.kyoto-u.ac.jp) and Mayu Iwata (Osaka University, Japan email: iwata.mayu@ist.osaka-u.ac.jp).

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

[3] R. Deveaud and E. Sanjuan. LIA at the NTCIR-10 INTENT task. In *Proceedings of NTCIR-10*, 2013.

[4] C. Guo, Y. Bai, J. Zheng, and D. Cai. KECIR at the NTCIR-10 INTENT task. In *Proceedings of NTCIR-10*, 2013.

[5] S.-J. Kim and J.-H. Lee. The KLEŠs subtopic mining system for the NTCIR-10 INTENT-2 task. In *Proceedings of NTCIR-10*, 2013.

[6] K. Kishida, K. hua Chen, S. Lee, K. Kuriyama, N. Kando, and H.-H. Chen. Overview of CLIR task at the sixth NTCIR workshop. In *Proceedings of NTCIR-6*, pages 1–19, 2007.

[7] J. Moreno and G. Dias. HULTECH at the NTCIR-10: Discovering user intents through search results clustering. In *Proceedings of NTCIR-10*, 2013.

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

[9] 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.

[10] T. Sakai. NTCIREVAL: A generic toolkit for information access evaluation. In *Proceedings of FIT 2011*, volume 2, pages 23–30, 2011.

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

[12] 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.

[13] T. Sakai and S. Robertson. Modelling a user population for designing information retrieval metrics. In *Proceedings of EVIA 2008*, pages 30–41, 2008.

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

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

[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] K. Tsukuda, Z. Dou, and T. Sakai. Microsoft research asia at the NTCIR-10 intent task. In *Proceedings of NTCIR-10*, 2013.

[18] M. Z. Ullah, M. Aono, and M. H. Seddiqui. SEM12 at the NTCIR-10 INTENT-2 english subtopic mining task. In *Proceedings of NTCIR-10*, 2013.

[19] J. Wang, G. Tang, Y. Xia, Q. Zhou, F. Zheng, Q. Hu, S. Na, and Y. Huang. Understanding the query: THCIB and THUIS at NTCIR-10 intent task. In *Proceedings of NTCIR-10*, 2013.

[20] Y. Xue, C. Fei, A. Damien, C. Luo, X. Li, S. Huo, M. Zhang, Y. Liu, and S. Ma. THUIR at NTCIR-10 INTENT task. In *Proceedings of NTCIR-10*, 2013.

[21] H. Yu and F. Ren. TUTA1 at the NTCIR-10 intent task. In *Proceedings of NTCIR-10*, 2013.

[22] X.-Q. Zhou and Y.-S. Hou. ICRCS at intent2: Applying rough set and semantic relevance for subtopic mining. In *Proceedings of NTCIR-10*, 2013.

Appendix

Table 27: SYSDESC fields of the English Subtopic Mining runs.

run name	SYSDESC field
hultech-S-E-1A	The HISGK-means algorithm is applied over a list of 50 snippets obtained from a websearch engine. The algorithm uses a second order similarity metric for calculate the similarity between words as well as the values between the cluster labels and the snippets. This particularity allows involve the cluster label task in the cluster algorithm. These labels are calculated in online time and are used as user intents.
hultech-S-E-2A	*ditto*
hultech-S-E-3A	*ditto*
hultech-S-E-4A	*ditto*
KLE-S-E-1A	We implemented the hierarchical structure with subtopic strings, and ranked them based on the related document coverage, CE, and BM25 model.
KLE-S-E-2A	We implemented the hierarchical structure with subtopic strings, and ranked them based on the related document coverage, CTF, IDF, CE, and BM25 model.
KLE-S-E-3A	We implemented the hierarchical structure with subtopic strings, and ranked them based on the related document coverage, CE, and BM25 model. Also, we used the official query suggestions as the additional related documents.
KLE-S-E-4A	We implemented the hierarchical structure with subtopic strings, and ranked them based on the related document coverage, CTF, IDF, CE, and BM25 model. Also, we used the official query suggestions as the additional related documents.
LIA-S-E-1A	We model the latent concepts for each query, using 4 different sources of information. Each concept is mapped with a Wikipedia article of which the title is used as subtopic.
LIA-S-E-2A	We model the latent concepts for each query, using 4 different sources of information. Difference with LIA-S-E-1 is that we fix the number of feedback documents to 10 (it is not fixed in LIA-S-E-1). We want to evaluate if the number of feedback documents can be fixed or if it can be automatically estimated at query time based on concept distribution. Each concept is mapped with a Wikipedia article of which the title is used as subtopic.
LIA-S-E-3A	We model the latent concepts for each query as in LIA-S-E-1, except that we use the provided commercial search engines suggestions to improve the query representation.
LIA-S-E-4A	Same run as LIA-S-E-1 with the initial query inserted before each subtopic, in order to fit to the guidelines : 'It is encouraged that participants submit subtopics of the form $<$originalquery$><$additionalstring$>$'
ORG-S-E-1A	Bing query suggestion
ORG-S-E-2A	Bing query suggestion (*actually, Bing completion*)
ORG-S-E-3A	Bing query suggestion (*actually, Google comletion*)
ORG-S-E-4A	Bing query suggestion (*actually, Yahoo completion*)
ORG-S-E-5A	Merged Bing suggestion, Bing completion, Google completion, Yahoo completion - dictionary sort
SEM12-S-E-1A	English SubTopic Mining in Knowledge Data Engineering and Information Retrieval Lab
SEM12-S-E-2A	*ditto*
SEM12-S-E-3A	*ditto*
SEM12-S-E-4A	*ditto*
SEM12-S-E-5A	*ditto*
THCIB-S-E-1A	(1) explores search recommendations (provided by NTCIR10), search completions (provided by NTCIR10), related webpages (Google), query log (ClueWeb09) and semantic descriptions (Wikipedia) to obtain concept-level subtopic candidates of each query; (2) ranks the subtopic candidates according to source weights and word frequencies in search result snippets.
THCIB-S-E-2A	(1) obtains subtopic candidates with THCIB-S-E-1A system; (2) generates expanded queries by re-positioning concepts in the query and inserting prepositional stop words between concepts within the query; (3) inputs the expanded queries to Google to obtain more recommendations and completions, which are also considered subtopic candidates; (4) ranks the subtopic candidates according to source weights and word frequencies in search result snippets.
THCIB-S-E-3A	(1) obtains subtopic candidates with THCIB-S-E-2A system; (2) generalize subtopic candidates with Freebase so as to associate named entities with the same ontology type to some ontological clusters; (3) ranks the subtopic candidates according to source weights, ontological clusters and word frequencies in search result snippets.
THCIB-S-E-4A	(1) obtains subtopic candidates with THCIB-S-E-2A system; (2) generalize subtopic candidates with Freebase so as to associate named entities with the same ontology type to some ontological clusters; (3) clusters subtopic candidates based on semantic similarity with standard AP algorithm; (4) ranks the subtopic candidates according to source weights, ontological clusters, semantic clusters and word frequencies in search result snippets.
THCIB-S-E-5A	(1) obtains subtopic candidates with THCIB-S-E-2A system; (2) generalize subtopic candidates with Freebase so as to associate named entities with the same ontology type to some ontological clusters; (3) clusters subtopic candidates based on semantic similarity with a revised AP algorithm; (4) ranks the subtopic candidates according to source weights, ontological clusters, semantic clusters and word frequencies in search result snippets.
THUIR-S-E-1A	THUIR-S-E-2A + THUIR-S-E-3A + THUIR-S-E-4A, Linear combination, Semantic similarity based re-clustering
THUIR-S-E-2A	Exraction from multiple resources (Google Insights, Google Keywords Generator, Query Suggestion/Completion, Wikipedia) + Snippet based clustering
THUIR-S-E-3A	Exraction From TMiner top results Snippet, Anchors and H1, BM25, Partition around medoid
THUIR-S-E-4A	Exraction From Search Engines top results Snippet + Query Suggestion/Completion, BM25, Partition around medoid
THUIR-S-E-5A	Extraction From TMiner top results Snippet - BM25 - Partition around medoid + Wikipedia + Official Query Suggestion/Completion; Linear combination; Semantic similarity based re-clustering
TUTA1-S-E-1A	Subtopic mining: firstly clustering the modifier graph into a number of clusters representing different subtopics;secondly selecting the subtopic instance through a linear combination of cluster recall and diversity
TUTA1-S-E-2A	Subtopic mining: firstly clustering the modifier graph into a number of clusters representing different subtopics;secondly selecting the subtopic instance according to the arriving probability

Table 28: SYSDESC fields of the Chinese Subtopic Mining runs.

run name	SYSDESC field
ICRCS-S-C-1A	clawling candidate query suggestions from 7 SE as candidate data; using semantic similairy, synonyms to filter duplicate suggestions; using semantic similairy,rank,semantic similairy to rank result.
ICRCS-S-C-2A	For the subtopic mining of Intent task, we choose the rough sets theory to design the subtopic mining algorithm, analysis the relations between the query and candidate subtopic set by mining the frequent item sets from the Baseline dataset and finish ranking for the candidate set. During the ranking, we use chinese semantic dictionary Hownet to divide the subtopic set into different groups.
ICRCS-S-C-3A	using 4 SE query suggestions applied by organizer as candidate data; using semantic similairy, synonyms to filter duplicate suggestions; using semantic similairy,rank,semantic similairy to rank result.
KECIR-S-C-1B	Run on the snippets from baseline.
KECIR-S-C-2B	Run on the similarity of the first result.
KECIR-S-C-3B	Run on the querylog and the first result.
KECIR-S-C-4B	Run on querylog,HowNet,and the first result.
ORG-S-C-1A	Bing query suggestion
ORG-S-C-2A	Baidu query suggestion
ORG-S-C-3A	Google query suggestion
ORG-S-C-4A	Sogou query suggestion
ORG-S-C-5A	Merged Bing suggestion, Baigu suggestion, Google suggestion, Sogou suggestion - dictionary sort
THUIR-S-C-1A	Subtopics from Query Suggestions, Wikipedia, Hudong; random walk based on large click log; LDA on clicked snippets; with reranking by query overlap rate.
THUIR-S-C-2A	Subtopics from Query Suggestions, Wikipedia, Hudong; random walk based on large click log; with reranking by query overlap rate.
THUIR-S-C-3A	THUIR-S-C-5A + reranking with clicked titles and snippets.
THUIR-S-C-4A	Subtopics from Query Suggestions; random walk based on SogouQ log, with reranking by query overlap rate.
THUIR-S-C-5A	Subtopics from Query Suggestions, Wikipedia, and Hudong; with reranking by query overlap rate.
THUIS-S-C-1A	THUIS subtopic mining system (THUIS-S-C-1A): (1) explores search recommendations (provided by NTCIR10), related webpages (Sogou), query log (SogouQ) and semantic descriptions (Wikipedia) to obtain concept-level subtopic candidates of each query; (2) generates expanded queries by re-positioning concepts in the query and inserting prepositional stop words between concepts within the query; (3) inputs the expanded queries to Google search engine to obtain more recommendations and completions, which are also considered subtopic candidates; (4) ranks the subtopic candidates according to source weights and word frequencies in search result snippets.
THUIS-S-C-2A	THUIS subtopic mining system (THUIS-S-C-2A): (1) obtains subtopic candidates with THCIS-S-C-1A system; (2) generates expanded queries by re-positioning concepts in the query and inserting prepositional stop words between concepts within the query; (3) clusters subtopic candidates based on semantic similarity with standard AP algorithm; (4) ranks the subtopic candidates according to source weights, semantic clusters and word frequencies in search result snippets.
THUIS-S-C-3A	THUIS subtopic mining system (THCIS-S-C-3A): (1) obtains subtopic candidates with THUIS-S-C-1A system; (2) generates expanded queries by re-positioning concepts in the query and inserting prepositional stop words between concepts within the query; (3) clusters subtopic candidates based on semantic similarity with a revised AP algorithm; (4) ranks the subtopic candidates according to source weights, semantic clusters and word frequencies in search result snippets.
THUIS-S-C-4A	THUIS subtopic mining system (THUIS-S-C-4A): (1) obtains subtopic candidates with THUIS-S-C-1A system; (2) generates expanded queries by re-positioning concepts in the query and inserting prepositional stop words between concepts within the query; (3) clusters subtopic candidates based on semantic similarity with K-means algorithm; (4) ranks the subtopic candidates according to source weights, semantic clusters and word frequencies in search result snippets.
TUTA1-S-C-1A	Subtopic mining: firstly clustering the modifier graph into a number of clusters representing different subtopics;secondly selecting the subtopic instance through a linear combination of cluster recall and diversity
TUTA1-S-C-2A	Subtopic mining: firstly clustering the modifier graph into a number of clusters representing different subtopics;secondly selecting the subtopic instance according to the arriving probability

Table 29: SYSDESC fields of the Japanese Subtopic Mining runs.

run name	SYSDESC field
KLE-S-J-1B	We implemented the hierarchical structure with subtopic strings, and ranked them based on the related document coverage, CE, and BM25 model.
KLE-S-J-2B	We implemented the hierarchical structure with subtopic strings, and ranked them based on the related document coverage, CTF, IDF, CE, and BM25 model.
KLE-S-J-3B	We implemented the hierarchical structure with subtopic strings, and ranked them based on the related document coverage, CE, and BM25 model. Also, we used the official query suggestions as the additional related documents.
KLE-S-J-4B	We implemented the hierarchical structure with subtopic strings, and ranked them based on the related document coverage, CTF, IDF, CE, and BM25 model. Also, we used the official query suggestions as the additional related documents.
MSINT-S-J-1B	Use query log, query suggestions and search result cluster. Rank subtopics based on the weighted overlap of search results
MSINT-S-J-2B	Use query log, query suggestions and search result cluster. Rank subtopics based on hit count
MSINT-S-J-3A	Use query log and query suggestion. Rank subtopics based on the weighted overlap of search results
MSINT-S-J-4A	Use query suggestion. Rank subtopics based on the weighted overlap of search results
MSINT-S-J-5B	Use search result cluster. Rank subtopics based on the weighted overlap of search results
ORG-S-J-1A	Bing query suggestion
ORG-S-J-2A	Bing query completion
ORG-S-J-3A	Google query completion
ORG-S-J-4A	Yahoo query completion
ORG-S-J-5A	Merged Bing suggestion, Bing completion, Google completion, Yahoo completion - dictionary sort

Table 30: SYSDESC fields of the Chinese Document Ranking runs.

run name	SYSDESC field
BASELINE-D-C-1	baseline ranking without diversificaton, 1000 results
KECIR-D-C-1B	Based on the baseline result and appearances of subtopics in the snnipets.
KECIR-D-C-2B	Based on the similarity result and appearances of subtopics in the htmls.
KECIR-D-C-3B	Based on the similarity result and appearances of subtopics in the snnipets.
KECIR-D-C-4B	Based on the querylog result and appearances of subtopics in the snippets.
KECIR-D-C-5B	Based on the querylog and HowNet result,also cumulative gain of subtopics in the snnipets.
THUIR-D-C-1A	THUIR-D-C-2A + click-based rerank (large click logs).
THUIR-D-C-2A	THUIR-D-C-3A + novelty-based rerank.
THUIR-D-C-3A	Retrieve on full text, anchor and click text documents (baseline of 1A and 2A).
THUIR-D-C-4A	Subtopic mining + retrieve on multiple subtopics + decay global gain based diverse results selection.
THUIR-D-C-5B	Official baseline + HITS-based rerank + novelty-based rerank + click-based rerank (SogouQ).
THUIR-D-C-R1	THUIR-D-C-5 retrieval on full text, anchor text and click text, + HITS-based rerank.

Table 31: SYSDESC fields of the Japanese Document Ranking runs.

run name	SYSDESC field
BASELINE-D-J-1	baseline ranking without diversificaton, 1000 results
MSINT-D-J-1B	Use Dou's search result diversification model, considering intent type probability.
MSINT-D-J-2B	Use Dou's search result diversification model, considering intent type probability. Combine search results of baseline, Yahoo and Bing
MSINT-D-J-3B	Use Dou's search result diversification model, considering intent type probability. Not diversify search result when topic has a navigational intent
MSINT-D-J-4B	Use Dou's search result diversification model. Combine search results of baseline, Yahoo and Bing
MSINT-D-J-5B	Use Dou's search result diversification model.
MSINT-D-J-R1	MSINT-D-J-3 MSINT:WSE suggestion
MSINT-D-J-R2	MSINT-D-J-2 MSINT: WSE suggestion and site domain

```
hultech-S-E-1A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-2A
hultech-S-E-2A with KLE-S-E-2A,KLE-S-E-4A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,
                    THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
hultech-S-E-3A with KLE-S-E-4A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A
hultech-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-2A
KLE-S-E-1A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A
KLE-S-E-2A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-2A
KLE-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-5A,TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-1A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-2A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-3A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-4A with ORG-S-E-3A,ORG-S-E-4A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-4A,SEM12-S-E-5A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,
                THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-1A with THCIB-S-E-1A,THCIB-S-E-2A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
ORG-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
ORG-S-E-5A with THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A
SEM12-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
```

Figure 29: English Subtopic Mining: significantly different run pairs in terms of I-rec@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official*).

```
hultech-S-E-1A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
hultech-S-E-2A with KLE-S-E-2A,KLE-S-E-4A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,
                    THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
hultech-S-E-3A with KLE-S-E-4A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THCIB-S-E-1A,THUIR-S-E-1A
hultech-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-1A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-2A
KLE-S-E-2A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-1A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-2A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-3A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-4A with ORG-S-E-3A,ORG-S-E-4A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,
                THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-1A with THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-2A with TUTA1-S-E-2A
ORG-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
ORG-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-3A with TUTA1-S-E-2A
SEM12-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
```

Figure 30: English Subtopic Mining: significantly different run pairs in terms of I-rec@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *revised*).

Table 32: English Subtopic Mining: discrepancies between *official* and *revised* in terms of I-rec@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

significant with *official* only	significant with *revised* only
hultech-S-E-3A with THUIR-S-E-2A	hultech-S-E-1A with TUTA1-S-E-1A
hultech-S-E-3A with THUIR-S-E-3A	hultech-S-E-2A with THCIB-S-E-4A
hultech-S-E-3A with THUIR-S-E-4A	hultech-S-E-3A with THCIB-S-E-1A
KLE-S-E-4A with ORG-S-E-5A	hultech-S-E-4A with TUTA1-S-E-1A
ORG-S-E-5A with THUIR-S-E-1A	KLE-S-E-1A with LIA-S-E-4A
ORG-S-E-5A with THUIR-S-E-2A	KLE-S-E-1A with TUTA1-S-E-2A
ORG-S-E-5A with THUIR-S-E-3A	KLE-S-E-3A with TUTA1-S-E-1A
	LIA-S-E-4A with SEM12-S-E-3A
	ORG-S-E-1A with THCIB-S-E-3A
	ORG-S-E-2A with TUTA1-S-E-2A
	SEM12-S-E-3A with TUTA1-S-E-2A

```
hultech-S-E-1A with hultech-S-E-2A,hultech-S-E-3A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-5A,SEM12-S-E-3A,TUTA1-S-E-1A,
                    TUTA1-S-E-2A
hultech-S-E-2A with hultech-S-E-4A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
hultech-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A
hultech-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-5A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-1A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A
KLE-S-E-2A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,TUTA1-S-E-2A
KLE-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A
KLE-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-1A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A
LIA-S-E-2A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A
LIA-S-E-3A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A
LIA-S-E-4A with ORG-S-E-4A,SEM12-S-E-1A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,
                THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-1A with THUIR-S-E-5A
ORG-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
ORG-S-E-5A with THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
SEM12-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-2A with TUTA1-S-E-2A
SEM12-S-E-4A with TUTA1-S-E-2A
SEM12-S-E-5A with TUTA1-S-E-2A
THCIB-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
```

Figure 31: English Subtopic Mining: significantly different run pairs in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official*).

```
hultech-S-E-1A with hultech-S-E-2A,hultech-S-E-3A,KLE-S-E-1A,KLE-S-E-3A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,
               ORG-S-E-5A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,TUTA1-S-E-1A,TUTA1-S-E-2A
hultech-S-E-2A with hultech-S-E-4A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,
               THUIR-S-E-4A,THUIR-S-E-5A
hultech-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THUIR-S-E-4A
hultech-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-1A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A
KLE-S-E-2A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A
KLE-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-1A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
           THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
           TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-2A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
           THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
           TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-3A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
           THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-4A with THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-1A with THCIB-S-E-1A,THCIB-S-E-2A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-2A with THUIR-S-E-2A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-5A with THUIR-S-E-4A
SEM12-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-4A with THUIR-S-E-4A
SEM12-S-E-5A with THUIR-S-E-4A
THCIB-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
```

Figure 32: English Subtopic Mining: significantly different run pairs in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *revised*).

Table 33: English Subtopic Mining: discrepancies between *official* and *revised* in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

significant with *official* only	significant with *revised* only
hultech-S-E-4A with ORG-S-E-5A	hultech-S-E-1A with KLE-S-E-1A
LIA-S-E-4A with ORG-S-E-4A	hultech-S-E-1A with KLE-S-E-3A
LIA-S-E-4A with SEM12-S-E-1A	hultech-S-E-1A with ORG-S-E-3A
ORG-S-E-4A with TUTA1-S-E-1A	hultech-S-E-1A with SEM12-S-E-4A
ORG-S-E-4A with TUTA1-S-E-2A	hultech-S-E-1A with SEM12-S-E-5A
ORG-S-E-5A with THCIB-S-E-1A	hultech-S-E-2A with THCIB-S-E-1A
ORG-S-E-5A with THCIB-S-E-2A	hultech-S-E-2A with THCIB-S-E-2A
ORG-S-E-5A with THCIB-S-E-3A	hultech-S-E-2A with THCIB-S-E-3A
ORG-S-E-5A with THCIB-S-E-4A	hultech-S-E-2A with THUIR-S-E-1A
ORG-S-E-5A with THUIR-S-E-1A	hultech-S-E-2A with THUIR-S-E-2A
ORG-S-E-5A with THUIR-S-E-2A	hultech-S-E-2A with THUIR-S-E-3A
ORG-S-E-5A with THUIR-S-E-3A	hultech-S-E-3A with THUIR-S-E-4A
ORG-S-E-5A with THUIR-S-E-5A	hultech-S-E-4A with ORG-S-E-2A
SEM12-S-E-4A with TUTA1-S-E-2A	KLE-S-E-2A with TUTA1-S-E-1A
SEM12-S-E-5A with TUTA1-S-E-2A	LIA-S-E-1A with TUTA1-S-E-2A
	LIA-S-E-2A with TUTA1-S-E-2A
	LIA-S-E-3A with TUTA1-S-E-2A
	ORG-S-E-1A with THCIB-S-E-1A
	ORG-S-E-1A with THCIB-S-E-2A
	ORG-S-E-1A with THUIR-S-E-1A
	ORG-S-E-1A with THUIR-S-E-2A
	ORG-S-E-1A with THUIR-S-E-3A
	ORG-S-E-1A with THUIR-S-E-4A
	ORG-S-E-2A with THUIR-S-E-2A
	ORG-S-E-2A with THUIR-S-E-4A
	ORG-S-E-2A with THUIR-S-E-5A
	SEM12-S-E-2A with TUTA1-S-E-1A
	SEM12-S-E-4A with THUIR-S-E-4A
	SEM12-S-E-5A with THUIR-S-E-4A

```
hultech-S-E-1A with hultech-S-E-2A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-5A,TUTA1-S-E-1A,TUTA1-S-E-2A
hultech-S-E-2A with KLE-S-E-4A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,
                    THUIR-S-E-5A
hultech-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THUIR-S-E-1A
hultech-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-1A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A
KLE-S-E-2A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-2A
KLE-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-5A,TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-1A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-2A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-3A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-4A with ORG-S-E-3A,ORG-S-E-4A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-4A,SEM12-S-E-5A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,
                THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-1A with THCIB-S-E-1A,THCIB-S-E-2A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
ORG-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
ORG-S-E-5A with THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
SEM12-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
```

Figure 33: English Subtopic Mining: significantly different run pairs in terms of D♯-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official*).

```
hultech-S-E-1A with hultech-S-E-2A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,TUTA1-S-E-1A,TUTA1-S-E-2A
hultech-S-E-2A with hultech-S-E-4A,KLE-S-E-2A,KLE-S-E-4A,LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,
                    THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
hultech-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,THUIR-S-E-1A,THUIR-S-E-4A
hultech-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-1A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,TUTA1-S-E-2A
KLE-S-E-2A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,TUTA1-S-E-1A,TUTA1-S-E-2A
KLE-S-E-3A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,TUTA1-S-E-2A
KLE-S-E-4A with LIA-S-E-1A,LIA-S-E-2A,LIA-S-E-3A,LIA-S-E-4A,ORG-S-E-1A,TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-1A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
           THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
           TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-2A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-3A with LIA-S-E-4A,ORG-S-E-1A,ORG-S-E-2A,ORG-S-E-3A,ORG-S-E-4A,ORG-S-E-5A,SEM12-S-E-1A,SEM12-S-E-2A,SEM12-S-E-3A,SEM12-S-E-4A,SEM12-S-E-5A,
                THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A,
                TUTA1-S-E-1A,TUTA1-S-E-2A
LIA-S-E-4A with ORG-S-E-3A,ORG-S-E-4A,SEM12-S-E-1A,SEM12-S-E-2A,THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THCIB-S-E-5A,THUIR-S-E-1A,
                THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-1A with THCIB-S-E-1A,THCIB-S-E-2A,THCIB-S-E-3A,THCIB-S-E-4A,THUIR-S-E-1A,THUIR-S-E-2A,THUIR-S-E-3A,THUIR-S-E-4A,THUIR-S-E-5A
ORG-S-E-2A with THUIR-S-E-1A,THUIR-S-E-4A
ORG-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
ORG-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
ORG-S-E-5A with THUIR-S-E-4A
SEM12-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
SEM12-S-E-3A with TUTA1-S-E-2A
SEM12-S-E-4A with TUTA1-S-E-2A
SEM12-S-E-5A with TUTA1-S-E-2A
THCIB-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THCIB-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-1A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-2A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-3A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-4A with TUTA1-S-E-1A,TUTA1-S-E-2A
THUIR-S-E-5A with TUTA1-S-E-1A,TUTA1-S-E-2A
```

Figure 34: English Subtopic Mining: significantly different run pairs in terms of D♯-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *revised*).

Table 34: English Subtopic Mining: discrepancies between *official* and *revised* in terms of D♯-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

significant with *official* only	significant with *revised* only
hultech-S-E-1A with ORG-S-E-5A	hultech-S-E-1A with ORG-S-E-1A
KLE-S-E-3A with LIA-S-E-4A	hultech-S-E-2A with hultech-S-E-4A
KLE-S-E-4A with ORG-S-E-5A	hultech-S-E-2A with KLE-S-E-2A
LIA-S-E-4A with SEM12-S-E-4A	hultech-S-E-2A with THCIB-S-E-4A
LIA-S-E-4A with SEM12-S-E-5A	hultech-S-E-2A with THCIB-S-E-5A
ORG-S-E-5A with THCIB-S-E-1A	hultech-S-E-3A with THUIR-S-E-4A
ORG-S-E-5A with THCIB-S-E-2A	KLE-S-E-1A with TUTA1-S-E-2A
ORG-S-E-5A with THCIB-S-E-3A	ORG-S-E-1A with THCIB-S-E-3A
ORG-S-E-5A with THUIR-S-E-1A	ORG-S-E-1A with THCIB-S-E-4A
ORG-S-E-5A with THUIR-S-E-2A	ORG-S-E-2A with THUIR-S-E-1A
ORG-S-E-5A with THUIR-S-E-3A	ORG-S-E-2A with THUIR-S-E-4A
ORG-S-E-5A with THUIR-S-E-5A	SEM12-S-E-3A with TUTA1-S-E-2A
SEM12-S-E-4A with TUTA1-S-E-1A	
SEM12-S-E-5A with TUTA1-S-E-1A	

```
ICRCS-S-C-1A with KECIR-S-C-3B,KECIR-S-C-4B
ICRCS-S-C-3A with KECIR-S-C-3B,KECIR-S-C-4B,ORG-S-C-2A,ORG-S-C-5A
KECIR-S-C-1B with TUTA1-S-C-1A
KECIR-S-C-2B with KECIR-S-C-4B
KECIR-S-C-3B with THUIR-S-C-1A,THUIR-S-C-2A,THUIR-S-C-5A,THUIS-S-C-1A,THUIS-S-C-3A,THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
KECIR-S-C-4B with THUIR-S-C-1A,THUIR-S-C-2A,THUIR-S-C-3A,THUIR-S-C-4A,THUIR-S-C-5A,THUIS-S-C-1A,THUIS-S-C-3A,THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
ORG-S-C-2A with THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
ORG-S-C-4A with TUTA1-S-C-1A
ORG-S-C-5A with THUIS-S-C-3A,THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
```

Figure 35: Chinese Subtopic Mining: significantly different run pairs in terms of I-rec@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official* and *revised*).

```
ICRCS-S-C-1A with KECIR-S-C-3B,KECIR-S-C-4B
ICRCS-S-C-2A with KECIR-S-C-4B,THUIR-S-C-3A,THUIS-S-C-1A
ICRCS-S-C-3A with KECIR-S-C-3B,KECIR-S-C-4B
KECIR-S-C-1B with THUIR-S-C-1A,THUIR-S-C-2A,THUIR-S-C-3A,THUIR-S-C-4A,THUIR-S-C-5A,THUIS-S-C-1A,TUTA1-S-C-1A,TUTA1-S-C-2A
KECIR-S-C-2B with THUIR-S-C-3A,THUIS-S-C-1A
KECIR-S-C-3B with ORG-S-C-1A,ORG-S-C-2A,ORG-S-C-3A,ORG-S-C-4A,ORG-S-C-5A,THUIR-S-C-1A,THUIR-S-C-2A,THUIR-S-C-3A,THUIR-S-C-4A,THUIR-S-C-5A,THUIS-S-C-1A,
                  THUIS-S-C-2A,THUIS-S-C-3A,THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
KECIR-S-C-4B with ORG-S-C-1A,ORG-S-C-2A,ORG-S-C-3A,ORG-S-C-4A,ORG-S-C-5A,THUIR-S-C-1A,THUIR-S-C-2A,THUIR-S-C-3A,THUIR-S-C-4A,THUIR-S-C-5A,THUIS-S-C-1A,
                  THUIS-S-C-2A,THUIS-S-C-3A,THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
ORG-S-C-2A with THUIR-S-C-3A
```

Figure 36: Chinese Subtopic Mining: significantly different run pairs in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official* and *revised*).

```
ICRCS-S-C-1A with KECIR-S-C-3B,KECIR-S-C-4B
ICRCS-S-C-2A with KECIR-S-C-4B
ICRCS-S-C-3A with KECIR-S-C-3B,KECIR-S-C-4B
KECIR-S-C-1B with THUIR-S-C-1A,THUIR-S-C-3A,THUIR-S-C-5A,THUIS-S-C-1A,THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
KECIR-S-C-2B with KECIR-S-C-4B
KECIR-S-C-3B with ORG-S-C-1A,ORG-S-C-3A,ORG-S-C-4A,THUIR-S-C-1A,THUIR-S-C-2A,THUIR-S-C-3A,THUIR-S-C-4A,THUIR-S-C-5A,THUIS-S-C-1A,THUIS-S-C-2A,THUIS-S-C-3A,
                  THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
KECIR-S-C-4B with ORG-S-C-1A,ORG-S-C-3A,ORG-S-C-4A,THUIR-S-C-1A,THUIR-S-C-2A,THUIR-S-C-3A,THUIR-S-C-4A,THUIR-S-C-5A,THUIS-S-C-1A,THUIS-S-C-2A,THUIS-S-C-3A,
                  THUIS-S-C-4A,TUTA1-S-C-1A,TUTA1-S-C-2A
ORG-S-C-2A with THUIS-S-C-1A,TUTA1-S-C-1A
ORG-S-C-5A with THUIS-S-C-1A,TUTA1-S-C-1A
```

Figure 37: Chinese Subtopic Mining: significantly different run pairs in terms of D♯-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official* and *revised*).

```
KLE-S-J-1B with ORG-S-J-4A
KLE-S-J-2B with MSINT-S-J-1B,MSINT-S-J-4A,ORG-S-J-3A,ORG-S-J-4A,ORG-S-J-5A
KLE-S-J-3B with ORG-S-J-4A
KLE-S-J-4B with ORG-S-J-3A,ORG-S-J-4A,ORG-S-J-5A
MSINT-S-J-1B with ORG-S-J-2A,ORG-S-J-4A
MSINT-S-J-2B with ORG-S-J-4A
MSINT-S-J-3A with ORG-S-J-4A
MSINT-S-J-4A with ORG-S-J-2A,ORG-S-J-4A
MSINT-S-J-5B with ORG-S-J-3A,ORG-S-J-4A,ORG-S-J-5A
ORG-S-J-1A with ORG-S-J-4A
ORG-S-J-2A with ORG-S-J-3A,ORG-S-J-4A,ORG-S-J-5A
ORG-S-J-3A with ORG-S-J-4A
ORG-S-J-4A with ORG-S-J-5A
```

Figure 38: Japanese Subtopic Mining: significantly different run pairs in terms of I-rec@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official* and *revised*).

```
KLE-S-J-1B with KLE-S-J-2B,KLE-S-J-4B,ORG-S-J-4A
KLE-S-J-2B with KLE-S-J-3B,MSINT-S-J-1B,MSINT-S-J-2B,MSINT-S-J-3A,MSINT-S-J-4A,ORG-S-J-1A,ORG-S-J-2A,ORG-S-J-3A,ORG-S-J-5A
KLE-S-J-3B with KLE-S-J-4B,ORG-S-J-4A
KLE-S-J-4B with MSINT-S-J-1B,MSINT-S-J-2B,MSINT-S-J-3A,MSINT-S-J-4A,ORG-S-J-1A,ORG-S-J-2A,ORG-S-J-3A
MSINT-S-J-1B with ORG-S-J-4A
MSINT-S-J-2B with ORG-S-J-4A
MSINT-S-J-3A with ORG-S-J-4A
MSINT-S-J-4A with ORG-S-J-4A
MSINT-S-J-5B with ORG-S-J-4A
ORG-S-J-1A with ORG-S-J-4A
ORG-S-J-2A with ORG-S-J-4A
ORG-S-J-3A with ORG-S-J-4A
ORG-S-J-4A with ORG-S-J-5A
```

Figure 39: Japanese Subtopic Mining: significantly different run pairs in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official*).

```
KLE-S-J-1B with KLE-S-J-2B,KLE-S-J-4B,ORG-S-J-4A
KLE-S-J-2B with KLE-S-J-3B,MSINT-S-J-1B,MSINT-S-J-2B,MSINT-S-J-3A,MSINT-S-J-4A,ORG-S-J-1A,ORG-S-J-2A,ORG-S-J-3A
KLE-S-J-3B with KLE-S-J-4B,ORG-S-J-4A
KLE-S-J-4B with MSINT-S-J-1B,MSINT-S-J-3A,MSINT-S-J-4A,ORG-S-J-1A,ORG-S-J-2A,ORG-S-J-3A
MSINT-S-J-1B with ORG-S-J-4A
MSINT-S-J-2B with ORG-S-J-4A
MSINT-S-J-3A with ORG-S-J-4A
MSINT-S-J-4A with ORG-S-J-4A
MSINT-S-J-5B with ORG-S-J-4A
ORG-S-J-1A with ORG-S-J-4A
ORG-S-J-2A with ORG-S-J-4A
ORG-S-J-3A with ORG-S-J-4A
ORG-S-J-4A with ORG-S-J-5A
```

Figure 40: Japanese Subtopic Mining: significantly different run pairs in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *revised*).

Table 35: Japanese Subtopic Mining: discrepancies between *official* and *revised* in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

significant with *official* only	significant with *revised* only
KLE-S-J-2B with ORG-S-J-5A KLE-S-J-4B with MSINT-S-J-2B	

```
KLE-S-J-1B with ORG-S-J-4A
KLE-S-J-2B with MSINT-S-J-1B,MSINT-S-J-3A,MSINT-S-J-4A,ORG-S-J-1A,ORG-S-J-3A,ORG-S-J-5A
KLE-S-J-3B with ORG-S-J-4A
KLE-S-J-4B with MSINT-S-J-1B,MSINT-S-J-3A,MSINT-S-J-4A,ORG-S-J-1A,ORG-S-J-3A,ORG-S-J-4A,ORG-S-J-5A
MSINT-S-J-1B with ORG-S-J-4A
MSINT-S-J-2B with ORG-S-J-4A
MSINT-S-J-3A with ORG-S-J-4A
MSINT-S-J-4A with ORG-S-J-4A
MSINT-S-J-5B with ORG-S-J-3A,ORG-S-J-4A
ORG-S-J-1A with ORG-S-J-4A
ORG-S-J-2A with ORG-S-J-3A,ORG-S-J-4A
ORG-S-J-3A with ORG-S-J-4A
ORG-S-J-4A with ORG-S-J-5A
```

Figure 41: Japanese Subtopic Mining: significantly different run pairs in terms of D♯-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$; *official* and *revised*).

```
BASELINE-D-C-1 with KECIR-D-C-2B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-R1
KECIR-D-C-1B with KECIR-D-C-2B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-R1
KECIR-D-C-2B with KECIR-D-C-3B,KECIR-D-C-4B,KECIR-D-C-5B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-4A,THUIR-D-C-5B,THUIR-D-C-R1
KECIR-D-C-3B with THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A
KECIR-D-C-4B with THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-R1
KECIR-D-C-5B with THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A
THUIR-D-C-1A with THUIR-D-C-5B
THUIR-D-C-2A with THUIR-D-C-5B
THUIR-D-C-3A with THUIR-D-C-5B
```

Figure 42: Chinese Document Ranking: significantly different run pairs in terms of I-rec@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
BASELINE-D-C-1 with KECIR-D-C-2B
KECIR-D-C-1B with KECIR-D-C-2B
KECIR-D-C-2B with KECIR-D-C-3B,KECIR-D-C-4B,KECIR-D-C-5B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-4A,THUIR-D-C-5B,THUIR-D-C-R1
```

Figure 43: Chinese Document Ranking: significantly different run pairs in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
BASELINE-D-C-1 with KECIR-D-C-2B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-R1
KECIR-D-C-1B with KECIR-D-C-2B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A
KECIR-D-C-2B with KECIR-D-C-3B,KECIR-D-C-4B,KECIR-D-C-5B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-4A,THUIR-D-C-5B,THUIR-D-C-R1
KECIR-D-C-4B with THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A
THUIR-D-C-1A with THUIR-D-C-5B
THUIR-D-C-2A with THUIR-D-C-5B
THUIR-D-C-3A with THUIR-D-C-5B
```

Figure 44: Chinese Document Ranking: significantly different run pairs in terms of D♯-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
KECIR-D-C-1B with KECIR-D-C-2B
KECIR-D-C-2B with KECIR-D-C-3B,KECIR-D-C-4B,KECIR-D-C-5B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-R1
```

Figure 45: Chinese Document Ranking: significantly different run pairs in terms of DIN-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
BASELINE-D-C-1 with KECIR-D-C-2B
KECIR-D-C-1B with KECIR-D-C-2B
KECIR-D-C-2B with KECIR-D-C-3B,KECIR-D-C-4B,KECIR-D-C-5B,THUIR-D-C-1A,THUIR-D-C-2A,THUIR-D-C-3A,THUIR-D-C-4A,THUIR-D-C-5B,THUIR-D-C-R1
```

Figure 46: Chinese Document Ranking: significantly different run pairs in terms of P+Q@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
BASELINE-D-J-1 with MSINT-D-J-4B
MSINT-D-J-4B with MSINT-D-J-R1
```

Figure 47: Japanese Document Ranking: significantly different run pairs in terms of I-rec@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
BASELINE-D-J-1 with MSINT-D-J-4B
MSINT-D-J-1B with MSINT-D-J-4B
MSINT-D-J-2B with MSINT-D-J-4B,MSINT-D-J-R2
MSINT-D-J-3B with MSINT-D-J-4B
MSINT-D-J-4B with MSINT-D-J-5B,MSINT-D-J-R1,MSINT-D-J-R2
```

Figure 48: Japanese Document Ranking: significantly different run pairs in terms of D-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
BASELINE-D-J-1 with MSINT-D-J-4B
MSINT-D-J-1B with MSINT-D-J-4B
MSINT-D-J-2B with MSINT-D-J-4B
MSINT-D-J-3B with MSINT-D-J-4B
MSINT-D-J-4B with MSINT-D-J-5B,MSINT-D-J-R1,MSINT-D-J-R2
```

Figure 49: Japanese Document Ranking: significantly different run pairs in terms of D$\sharp$-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
BASELINE-D-J-1 with MSINT-D-J-2B,MSINT-D-J-3B,MSINT-D-J-4B,MSINT-D-J-5B
MSINT-D-J-1B with MSINT-D-J-4B
MSINT-D-J-3B with MSINT-D-J-R1
MSINT-D-J-4B with MSINT-D-J-5B,MSINT-D-J-R1,MSINT-D-J-R2
```

Figure 50: Japanese Document Ranking: significantly different run pairs in terms of DIN-nDCG@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).

```
BASELINE-D-J-1 with MSINT-D-J-4B
MSINT-D-J-1B with MSINT-D-J-4B
MSINT-D-J-2B with MSINT-D-J-4B
MSINT-D-J-4B with MSINT-D-J-R1
```

Figure 51: Japanese Document Ranking: significantly different run pairs in terms of P+Q@10 (two-sided randomised Tukey's HSD at $\alpha = 0.05$).