Overview of the NTCIR-9 INTENT Task

Ruihua Song† Min Zhang‡ Tetsuya Sakai† Makoto P. Kato∗
Yiqun Liu‡ Miho Sugimoto⋆ Qinglei Wang♣ Naoki Orii◇
†Microsoft Research Asia ‡Tsinghua University ∗Kyoto University
♣ Xi'an Jiaotong University □National Institute of Informatics ◇University of Tokyo
Song.Ruihua@microsoft.com, z-m@tsinghua.edu.cn, tetsuyasakai@acm.org

Abstract

This is an overview of the NTCIR-9 INTENT task, which comprises the Subtopic Mining and the Document Ranking subtasks. The INTENT task attracted participating teams from seven different countries/regions – 16 teams for Subtopic Mining and 8 teams for Document Ranking. The Subtopic Mining subtask received 42 Chinese runs and 14 Japanese runs; the Document Ranking subtask received 24 Chinese runs and 18 Japanese runs. We describe the subtasks, data and evaluation methods, and then report on the results.

Keywords: *test collections, intents, diversity, ambiguity, subtopics, D-measures, D♯-measures.*

1. Introduction

The INTENT task is a new NTCIR task, comprising the Subtopic Mining subtask (given a query, output a ranked list of possible subtopic strings) and the Document Ranking subtask (given a query, output a ranked list of URLs that are selectively diversified). The task attracted participating teams from seven different countries/regions (Australia, Canada, China, Japan, Korea, Taiwan, U.K.) – 16 teams for Subtopic Mining and 8 teams for Document Ranking. Table 1 shows the complete list of participating teams. The Subtopic Mining task received 42 Chinese runs and 14 Japanese runs; the Document Ranking task received 24 Chinese runs and 18 Japanese runs.

First, a word of sincere apology to NTU who participated in all subtask/language combinations: the organisers did not notice that this team had submitted Japanese Subtopic Mining and Document Ranking runs, and as a result, their Japanese runs were *not* included in the pools for the official evaluations. For the Japanese subtopic mining subtask, we re-evaluated all runs after we released the Official results to participants. We will refer to the new Japanese Subtopic Mining results that incorporate the contributions from NTU as the Revised results. (Japanese Subtopic Mining participants may choose to discuss either the Official results or the Revised results, or both.) On the other hand, we could not conduct any additional relevance assessments for the Japanese Document Ranking subtask: hence, this overview does not discuss the performances of NTU's Japanese Document Ranking runs. We will improve the submission process for the next round of the INTENT task to prevent this kind of mistake.

The important dates for the NTCIR-9 INTENT task were as follows[1]:

Jan 6	*Example topics released*
April 15	*Judgments for Chinese example topics released*
April 29	*Judgments for Japanese example topics released*
May 1	*Chinese formal run topics released*
May 17	*Japanese formal run topics released*
June 10	*Chinese runs due*
June 24	*Japanese runs due*
Aug 22	*Evaluation results released*
Sept 5	*Revised Japanese Subtopic Mining results released*

This overview paper describes the subtasks, data and evaluation methods, and then reports on the Official (and Revised) results. The overview of the NTCIR-9 One Click Access task (1CLICK) is given elsewhere [9].

2. Subtopic Mining Task Definition

2.1 Overview

In the Subtopic Mining subtask, systems were required to return a ranked list of *subtopic strings* in response to a given query. A subtopic could be a specific interpretation of an ambiguous query (e.g. "microsoft windows" or "house windows" in response to "windows") or an aspect of a faceted query (e.g. "windows 7 update" in response to "windows 7"). The subtopics collected from participants were pooled, manually clustered and thereby used as the basis for identifying the search *intents* for the query: each subtopic could belong to exactly one intent. The probability of each intent given the query was estimated through assessor voting. Finally, the relevance and diversity of the ranked subtopics were evaluated using diversified information retrieval metrics [8].

Possible practical applications of this task include search result diversification (such as our Document Ranking subtask) and generating diversified query suggestions for a given query.

NTCIR-9 Workshop Meeting, 2011, Tokyo, Japan.

[1] http://www.thuir.org/intent/ntcir9/

Table 1. INTENT participating teams. Teams with a participated in both subtasks.

team name	language	organisation
(a) Subtopic Mining		
DBIIR	C	Information School, Renmin University of China
HIT2jointNLPLab∗	C	Heilongjiang Institute of Technology / Harbin Institute of Technology
HITIR	C	Research Center for Information Retrieval, Harbin Institute of Technology
ICTIR	C	Institute of Computing Technology, Chinese Academy of Sciences
III_CYUT_NTHU∗	C	Institute for Information Industry / Chaoyang University of Technology / National Tsing Hua University
ISCAS	C	Institute of Software Chinese Academy of Sciences
KLE	C	Knowledge and Engineering Laboratory, Pohang University of Science and Technology (POSTECH)
MSINT∗	C	Microsoft Research Asia
NTU∗	C(, J)	Department of Computer Science and Information Engineering, National Taiwan University
ORG	J	Organisers (Microsoft Research Asia)
RMIT∗	J	RMIT University
THU∗	C	Tsinghua University
TUTA1	C	University of Tokushima
uogTr∗	C, J	University of Glasgow
UWat	C	University of Waterloo
WSE	J	Organisers using Web search engines (Microsoft Research Asia)
(b) Document Ranking		
HIT2jointNLPLab∗	C	Heilongjiang Institute of Technology / Harbin Institute of Technology
III_CYUT_NTHU∗	C	Institute for Information Industry / Chaoyang University of Technology / National Tsing Hua University
MSINT∗	C, J	Microsoft Research Asia
NTU∗	C(, J)	Department of Computer Science and Information Engineering, National Taiwan University
RMIT∗	J	RMIT University
SJTUBCMI	C	Center for Brain-like Computing and Machine Intelligence, Shanghai Jiao Tong University
THUIR∗	C	Tsinghua University
uogTr∗	C, J	University of Glasgow

2.2 Input

The primary input to a Subtopic Mining system is a Chinese or Japanese query string which represents a topic, encoded in UTF-8. For the Chinese Subtopic Mining subtask, participants also had access to a few additional resources such as a hyperlink graph and a query log (See Section 4).

2.3 Output

The output format for the Subtopic Mining task is as follows. The first line of each run file should contain a short textual description of the run in the form:

```
<SYSDESC>this is a dummy description.</SYSDESC>
```

and the remaining lines should be of the form:

```
[TopicID];0;[Subtopic String];[Rank];[Score];[RunTag]
```

where the "0" is a dummy field. Note that a semicolon is used as the field separator, as the subtopic string may contain white spaces. For example, Figure 1 shows a part of Chinese Subtopic Mining run:

```
0015;0;莫扎特传;1;5.0;Run1
0015;0;莫扎特的作品;2;4.4;Run1
0015;0;莫扎特简介;3;3;Run1
0015;0;莫扎特效应;4;2.5;Run1
```

Figure 1. Example run output for Chinese topic 0015 "Mozart."

Participating systems were expected to fully rank the substrings for each topic: in the evaluation phase, we ignore the Rank, Score and the RunTag fields. (See also Section 3.3 where we discuss the Document Ranking output format.) As we shall describe in the following sections, we evaluate subtopic mining by first clustering pooled subtopics to identify a set of search *intents*, and then applying diversity evaluation metrics by treating each substring as a document that is either relevant to exactly one intent or nonrelevant to all intents.

We required the participants to name each of their run files as follows:

```
<teamID>-S-<language>-<integer>.txt
```

where "S" stands for the Subtopic Mining subtask and "<language>" is either "C" (Chinese) or "J" (Japanese). For example, THU-S-C-1.txt is the first Chinese Subtopic Mining run from Tsinghua University. Each participating team was allowed to submit up to five runs; no manual runs were allowed. We required that each run file be encoded in UTF-8.

2.4 Organisers' Japanese Runs

As we had only two participating teams for the Japanese Subtopic Mining tasks[2], we produced some additional runs to enhance the coverage of possible Japanese subtopics. These runs are labelled with either ORG or WSE.

ORG-S-J-1 looks at anchor texts of retrieved Web pages to mine potential subtopics. For a given query q, we first retrieved all anchor texts containing q using Microsoft's internal Web search platform[3]. We then performed word segmentation on all of the retrieved anchor texts. Distinct terms generated from word segmentation (excluding the original query term q) form our subtopic

[2]Or so we thought – we actually had three, including NTU.

[3]http://research.microsoft.com/en-us/projects/WebStudio/

candidates. We used the following equation to evaluate the importance of subtopic candidate c:

$$\log(1 + docFreq(c)) \sum_{d \in D} (1 + \log(termFreq(c, d))) \quad (1)$$

where $docFreq(c)$ is the number of distinct Web documents pointed to by anchor text c. D is the set of domain names of the Web documents pointed to by c, and $termFreq(c, d)$ is the number of times c appears in the anchor texts that point to documents with domain name d.

ORG-S-J-2 performed clustering on titles and snippets of pages returned by the aforementioned Web search platform. We used a slightly modified version of Zeng et al's method [12]: initial experimentation indicated that the Phrase Independence feature harmed performance on clustering for the Japanese language, and was therefore removed.

The WSE-S-J runs utilized query reformulations provided by Web search engines (Bing, Google and Yahoo). Two types of query reformulation were tested: *related* queries, which are the queries shown alongside the search results in the search engine result page, and *suggested* queries, which are the queries shown in the dropdown list within the query box as the user enters a query [10].

3. Document Ranking Task Definition

3.1 Overview

The Document Ranking subtask evaluated selectively diversified Web search results. Systems were expected to (a) retrieve a set of documents that covers as many intents as possible; and (b) rank documents that are highly relevant to more popular intents higher than those that are marginally relevant to less popular intents. Two major differences between our Document Ranking subtask and the TREC Web diversity task are that we consider the *probability of each intent* given a query and *per-intent graded relevance* [8].

We encouraged participants to apply diversification algorithms selectively, as we believe that not all queries require diversification. How to automatically classify *ambiguous*, *faceted* and *clear* queries is an interesting research problem [11]. The Chinese and the Japanese topic sets for this subtask are the same as the ones we used in Subtopic Mining, and contain topics with different types of vagueness (See Section 4).

3.2 Input

The primary input to a Document Ranking system is a query string. Note that the set of possible intents for each query, identified through the Subtopic Mining subtask, is *not* provided to the participating teams. Thus, the participating systems were expected to somehow find and utilise explicit or implicit subtopics for search result diversification where necessary. Here, an explicit subtopic refers to a short string (which may possibly serve as a query), while an implicit subtopic refers to a set of texts which collectively represent a subtopic. Participants could, for example, mine implicit subtopics from some corpora, or scrape query suggestions from a Web search engine and treat them as explicit subtopics.

For the Chinese Document Ranking subtask, participants also had access to a few additional resources such as a hyperlink graph and a query log (See Section 4).

3.3 Output

The output format for the Document Ranking subtask is similar to the TREC run file format, except for the first line in the run file. The first line of each run file should contain a short textual description of the run in the form:

```
<SYSDESC>this is a dummy description.</SYSDESC>
```

and the remaining lines should be of the form:

```
[TopicID] 0 [DocumentID] [Rank] [Score] [RunTag]
```

where the "0" is a dummy field. For example, a part of a Japanese Document Ranking run may look like this:

```
0101 0 clueweb09-ja0006-97-23810 1 27.73 Run1
0101 0 clueweb09-ja0009-08-98321 2 25.15 Run1
0101 0 clueweb09-ja0003-71-19833 3 21.89 Run1
0101 0 clueweb09-ja0002-66-03897 4 13.57 Run1
```

However, unlike TREC, participating systems were expected to fully rank documents for each topic: we actually ignore the Rank and the Score fields and evaluate the ranked list of document IDs in the run file "as is." This is because, while `trec_eval`[4] reranks documents internally, our evaluation toolkit `NTCIREVAL`[5] assumes that the systems are responsible for resolving ties [5]. (Section 5 discusses the evaluation metric computed using `NTCIREVAL`.) RunTag is also ignored in our evaluation.

We required the participants to name each of their run files as follows:

```
<teamID>-D-<language>-<integer>.txt
```

where "D" stands for the Document Ranking subtask and "<language>" is either "C" (Chinese) or "J" (Japanese). Each participating team was allowed to submit up to five runs; no manual runs were allowed.

4. Data

This section describes how we created the NTCIR-9 INTENT Subtopic Mining and Document Ranking test collections for Chinese and Japanese.

4.1 Document Collections

Table 2 provides some statistics on the INTENT Chinese and Japanese document collections.

Our Chinese document collection is the SogouT corpus crawled and released in 2008 by the Tsinghua-Sohu Joint Laboratory on Search Technology[6]. (The Joint Lab also released a search performance evaluation benchmark for SogouT, known as SogouE, which includes about 10,000 query topics and corresponding answers[7]. A hyperlink graph file[8] and PageRank scores[9] for SogouT are also available.)

[4] `http://trec.nist.gov/trec_eval/`
[5] `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`
[6] `http://www.sogou.com/labs/dl/t.html`
[7] `http://www.sogou.com/labs/dl/e.html`
[8] `http://www.sogou.com/labs/dl/t-link.html`
[9] `http://www.sogou.com/labs/dl/t-rank.html`

Table 2. NTCIR-9 INTENT Chinese and Japanese Document Collections.

	Chinese	Japanese
Collection name	SogouT	ClueWeb-JA
#Pages	138 million	67.3 million
Size and Storage Medium	7z files on a 500GB hard disk	tarred/gzipped files on a 500GB hard disk

Table 3. SogouQ data record.

item	description
Date/Time	Date/time of the user click
ID	ID representing a browser cookie
Query	User's query
URL	URL clicked by the user
Rank	Rank of the clicked URL in the result list
Order	Order of the URL click within the session

Our Japanese document collection ClueWeb09-JA was extracted from the ClueWeb09 collection[10] by the Language Technologies Institute at Carnegie Mellon University. The entire ClueWeb09 collection, which consists of one billion Web pages in ten languages, was crawled between January and February 2009, and was later used in several recent TREC tracks including the Web track and the Crowdsourcing track.

4.2 Chinese Query Log

For the Chinese subtasks, a Chinese query log called SogouQ was also made available to participants. SogouQ was also constructed by the Tsinghua-Sohu Joint Laboratory on Search Technology[11]. It contains about 30 million clicks collected in June 2008. Its size is about 1.9 gigabytes and is available for download via FTP. Several other evaluation efforts have also utilised SogouQ, including the Cross-Language Evaluation Forum[12]. Participants were responsible for splitting the SogouQ data into sessions, for example, based on a 30-minute timeout heuristic.

4.3 Topics

We created a set of 100 Chinese topics (0001-0100), and another set of 100 Japanese topics (0101-0200) based on real query log data. The same topic sets were used for the Subtopic Mining and Document Ranking subtasks. The Chinese and the Japanese topics were selected from the June 2008 query log of Sogou[13] and from the January 2009 query log of Bing Japan[14], respectively. The time periods were chosen to coincide with the crawl periods of the document collections.

We considered three factors in selecting topics. First, similar to the TREC 2009 Web track [3], we sampled "torso" queries that are neither very "head" nor very "tail" from search query logs. Second, we tried to balance topics across different types of vagueness: previous research [3, 11] shows that some queries are ambiguous and may refer to different entities or concepts; some are faceted and may cover different subtopics; while others may be relatively clear and represent a specific information need. Thus, we selected queries by first examining their search results and related queries returned by several search engines, and then manually classifying them into *ambiguous*, *faceted* and *clear*. Third, as NTCIR traditionally encourages cross-language research, we chose queries so that the first twenty topics from the Chinese set and the first twenty topics from the Japanese set are equivalent to each other when translated. For example, the Chinese topic 0001 corresponds to the Japanese topic 0101. Thus these twenty topics can be utilised for cross-language information retrieval research that considers diversification, though this is outside the scope of our current task.

[10] http://lemurproject.org/clueweb09/
[11] http://www.sogou.com/labs/dl/q.html
[12] http://www.promise-noe.eu/mining-user-preference/logclef-2011/data
[13] http://www.sogou.com/
[14] http://bing.jp

4.4 Intents

Having created the topic sets, the *intents* for each topic were identified as described below.

In the Subtopic Mining subtask, participants were required to submit ranked lists of *subtopics* as was described in Section 2. Just like document IDs are pooled in traditional relevance assessments for document retrieval evaluation, a pool of subtopic strings was created for each topic based on the Subtopic Mining runs: top 20 subtopic strings from each run were pooled. The pooled subtopic strings were sorted by "popularity" as in some previous NTCIR tasks [7]. Then, using a dedicated graphical interface that we developed, assessors manually clustered subtopics with similar search *intents*: each topic was handled by exactly one assessor. The assessors also provided a short description for each intent. Subtopics that did not appear to be relevant or understandable were put into a "misc" cluster. After the manual identification of intents was completed by the assessors, the organisers reviewed the results and revised them where necessary.

Subsequently, we let 10 assessors vote whether each intent is important or not for each topic. Thus the number of votes for each intent could vary between 0 and 10. The subtopics in the "misc" cluster were given the per-intent relevance level of $L0$ (nonrelevant); all other subtopics, each corresonding to one intent, were given the per-intent relevance level of $L1$ (relevant). Moreover, all intents that did not receive any votes were further merged into a single "intent" called "others." We then assigned a positive integer as an ID for each intent (and a 0 for the "misc" cluster).

Having established the intents, we estimated their intent probabilities ("popularity" of intents) based on the above votes, using Laplace smoothing. Thus, the probability of the j-th intent given a query was computed as:

$$p_j = \frac{\#votes_j + 0.5}{\sum_{k=1}^{n}(\#votes_k + 0.5)} \quad (2)$$

where $\#votes_j$ is the number of votes for the j-th intent and n is the total number of intents.

The intents and their probabilities were used for the evaluation in both the Subtopic Mining and Document Ranking subtasks.

Figure 2 shows the number of intents obtained for our Chinese and Japanese topics. For the Chinese topics, the number of intents per topic ranges between 4 and 16 and is 9.17 on average; each intent had 1-242 relevant substrings (23.26 on average). For Japanese, the number of intents per topic ranges between 3 and 24 and is 10.91 on average; each intent had 1-33 relevant substrings (3.45 on average) in the Official data, and 1-34 relevant substrings (3.72 on average) in the Revised data. Figure 3 visualises the probability distribution of the intents, obtained using Eq. 2.

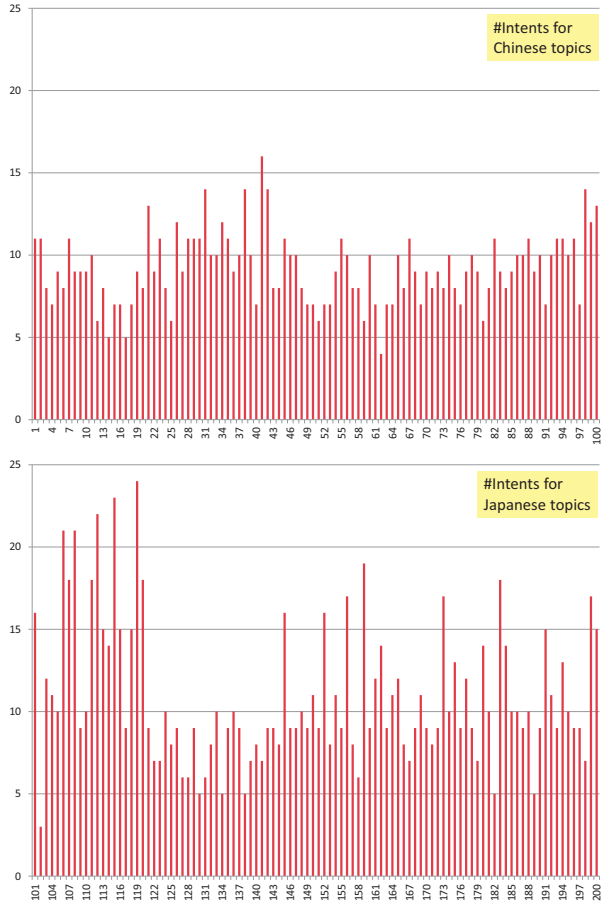

Figure 2. Number of intents for the Chinese and Japanese topics.

Figures 4 and 5 show an example topic from the Chinese topic set and the Japanese topic set: the these two topics are equivalent ("Mozart"). It can be observed that the intents identified are quite different across the two languages.

For the Japanese Subtopic Mining subtask, we used the "forgotten" four runs from NTU to revise our evaluation results. We extracted top 20 subtopic strings from the NTU runs, and obtained a set of unjudged strings for each Japanese topic by comparing the strings from NTU with the Official assessment data. Then the unjudged strings were judged and merged into *existing* intents: no new intents were devised based on the NTU runs. We will report on the Revised Japanese Subtopic Mining results as well as the Official results, showing the NTU runs in the Revised results only.

4.5 Document Relevance Assessments

As in traditional document retrieval evaluation tasks, we formed a pool of document IDs for each topic based on the submitted Document Ranking runs. We used the pool depth of 20. The pooled documents sorted by "popularity" as in some previous NTCIR tasks [7]. As was mentioned earlier, we regret to say that we did not include the three Japanese Document Ranking runs from NTU in the pools.

In contrast to traditional relevance assessments where the assessors determine the relevance of each pooled document with respect to a *topic*, we required the assessors to provide graded relevance assessments with respect to each *intent* of a given topic. A dedicated relevance assessment interface was used for this purpose[15]. Two assessors were assigned to each topic, and each relevance assessment was on a three point scale: "highly relevant" (the document fully satisfies the information need expressed in the intent), "relevant" (the document only partially satisfies the information need expressed in the intent) and "nonrelevant"[16]. Note that each document may be relevant to multiple intents, in contrast to the Subtopic Mining situation where each substring belongs to exactly one intent.

Based on the relevance assessments of the two judges, the final per-intent relevance level of each document was determined following the approach taken by Sakai *et al.* [6]: first, per-assessor assessments of "highly relevant", "relevant" and "nonrelevant" were weighted as 2, 1 and 0, respectively; then these weights were summed across the two assessors. As a result, we obtained a five-point scale graded relevance data: the highest being $L4$ (two highly relevants) and the lowest being $L0$ (two nonrelevants).

For the Chinese topics, the number of relevant documents for a topic (where each document relevant to multiple intents are counted only once) is 121 on average and ranges between 9 and 182; the number of intents covered by a relevant document is 1.94 on average and ranges between 1 and 10. For the Japanese topics, the number of relevant documents for a topic (where each document relevant to multiple intents are counted only once) is 88.96 on average and ranges between 22 and 138; the number of intents covered by a relevant document is 2.21 on average and ranges between 1 and 13. That is, there is a document that covers as many as 13 intents (out of the 21 intents for Topic 0108)!

5. Evaluation Metrics

For both of our subtasks, we used *D♯-nDCG* [8] as our primary evaluation metric. D♯-nDCG is a linear combination of *intent recall* (or "I-rec", which measures diversity) and *D-nDCG* (which measures overall relevance across intents). The advantages of D♯-nDCG over other diversity metrics such as *α-nDCG* [4] and *Intent-Aware metrics*[1] are discussed elsewhere [8].

We used the default setting of the `NTCIREVAL` [5] to compute the above three metrics: thus, D♯-nDCG is a simple average of I-rec and D-nDCG. The gain values for the per-intent graded relevance were set linearly: 1, 2, 3 and 4 for $L1$, $L2$, $L3$ and $L4$, respectively. (But recall that the per-topic relevance assessments for Subtopic Mining are binary, i.e., $L0$ and $L1$ only.)

We used the *measurement depths* (i.e. number of top ranked items to be evaluated) of $l = 10$, 20 and 30 for both Subtopic Mining and Document Ranking. (Recall that the pool depth was 20 for both subtasks.) Note that intent recall may not be able to reach one if l is smaller than the number of intents: this happens when $l = 10$ for the Chinese subtasks (as the maximum number of intents is 16), and when $l = 10$ and 20 for the Japanese subtasks (as the maximum number of intents is 24).

[15] Sakai and Song [8] used the same interface to add per-intent graded relevance assessments to the TREC 2009 Web track diversity task test collection.

[16] Currently inaccessible pages and spam pages were judged as nonrelevant.

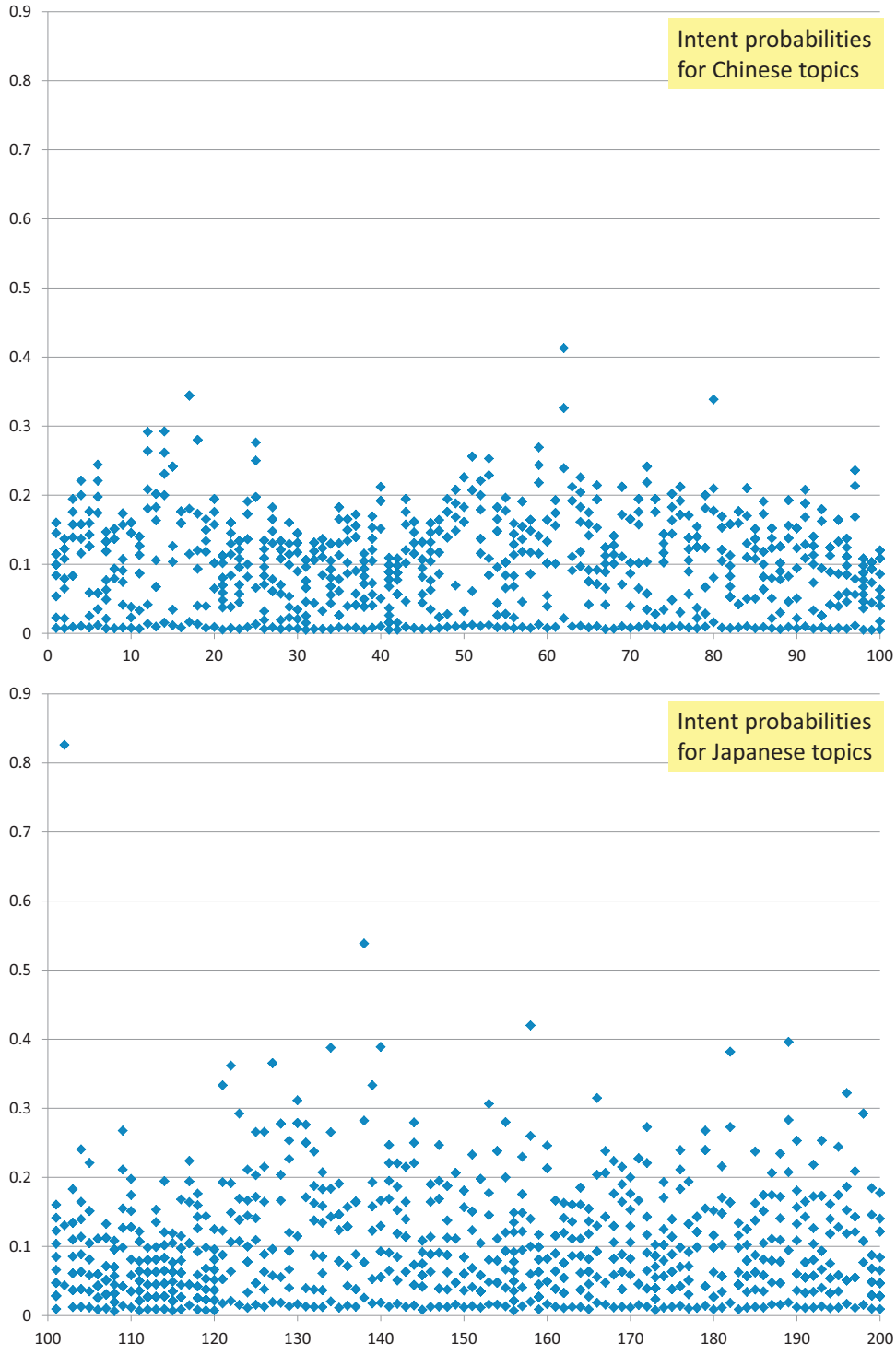

Figure 3. Estimated intent probabilities for the Chinese and Japanese topics.

```
<topic number="0015">
<query>莫扎特</query>
<intent number="1" probability="0.241379310344828">
<description>莫扎特【音乐下载、在线试听】</description>
<examples> 莫扎特音乐下载;莫扎特胎教音乐下载...</examples>
</intent>
<intent number="2" probability="0.241379310344828">
<description>莫扎特【资料】</description>
<examples> 莫扎特简介;莫扎特传...</examples>
</intent>
<intent number="3" probability="0.241379310344828">
<description>莫扎特【作品】</description>
<examples> 莫扎特歌剧;莫扎特的魔笛...</examples>
</intent>
<intent number="4" probability="0.126436781609195">
<description>莫扎特【音乐会、比赛】</description>
<examples> 维也纳莫扎特音乐会;莫扎特音乐会...</examples>
</intent>
<intent number="5" probability="0.103448275862069">
<description>莫扎特【影视小说作品】</description>
<examples> 寻找莫扎特;电影莫扎特...</examples>
</intent>
<intent number="6" probability="0.0344827586206897">
<description>莫扎特【音乐馆、学校】</description>
<examples> 小小莫扎特音乐馆;维也纳莫扎特学院...</examples>
</intent>
<intent number="7" probability="0.0114942528735632">
<description>Others</description>
<examples> 莫扎特巧克力;莫扎特巧克力球...</examples>
</intent>
</topic>
```

Listen and download

Biography

Works by Mozart

Concerts

Movies, novels etc.

Music academy

Figure 4. Chinese topic 0015 “Mozart” and its intents, with rough English translations.

```
<topic number="0115">
<query>モーツァルト</query>
<intent number="1" probability="0.118881118881119">
<description>モーツァルト 楽譜</description>
<examples> モーツァルト dscore;モーツァルト 楽譜 無料...</examples>
</intent>
<intent number="2" probability="0.104895104895105">
<description>モーツァルト 交響曲</description>
<examples> モーツァルト 交響曲;モ ツァルト 交響曲...</examples>
</intent>
<intent number="3" probability="0.0769230769230769">
<description>モーツァルト 協奏曲</description>
<examples> モーツァルト 協奏曲...</examples>
</intent>
<intent number="4" probability="0.0769230769230769">
<description>モーツァルト 音楽</description>
<examples> モーツァルト 音楽;モーツァルト クラシック音楽...</examples>
</intent>
<intent number="5" probability="0.0629370629370629">
<description>モーツァルト 音源</description>
<examples> モーツァルト mp3;モーツァルト 名盤...</examples>
</intent>
<intent number="6" probability="0.0629370629370629">
<description>モーツァルト 演奏</description>
<examples> モーツァルト 演奏...</examples>
</intent>
<intent number="7" probability="0.0629370629370629">
<description>モーツァルト ヴォルフガング アマデウス</description>
<examples> モーツァルト ヴォルフガング アマデウス;モーツァルト アマデウス...</examples>
</intent>
<intent number="8" probability="0.0629370629370629">
<description>モーツァルト ピアノ協奏曲</description>
<examples> モーツァルト ピアノ協奏曲;モーツァルトピアノ...</examples>
</intent>
<intent number="9" probability="0.048951048951049">
<description>モーツァルト 魔笛</description>
<examples> モーツァルト 魔笛;モ ツァルト 魔笛...</examples>
</intent>
<intent number="10" probability="0.034965034965035">
<description>モーツァルト 歌劇</description>
<examples> モーツァルト 歌劇;モ ツァルト オペラ...</examples>
</intent>
<intent number="11" probability="0.034965034965035">
<description>モーツァルト トルコ行進曲</description>
<examples> モーツァルト トルコ行進曲...</examples>
</intent>
<intent number="12" probability="0.034965034965035">
<description>モーツァルト 作曲</description>
<examples> モーツァルト 作曲...</examples>
</intent>
<intent number="13" probability="0.020979020979021">
<description>モーツァルト効果</description>
<examples> モーツァルト効果;モーツァルト 音楽療法...</examples>
</intent>
<intent number="14" probability="0.020979020979021">
<description>モーツァルト ドン ジョバンニ</description>
<examples> モーツァルト ドン...</examples>
</intent>
<intent number="15" probability="0.020979020979021">
<description>モーツァルト ソナタ</description>
<examples> モーツァルト ソナタ...</examples>
</intent>
<intent number="16" probability="0.020979020979021">
<description>モーツァルト フィガロ</description>
<examples> モーツァルト フィガロ...</examples>
</intent>
<intent number="17" probability="0.020979020979021">
<description>モーツァルト モーツァルト家</description>
<examples> モーツァルト モーツァルト家...</examples>
</intent>
<intent number="18" probability="0.020979020979021">
<description>モーツァルト レクイエム 歌詞</description>
<examples> モーツァルト レクイエム 歌詞...</examples>
</intent>
<intent number="19" probability="0.020979020979021">
<description>モーツァルト 帝劇</description>
<examples> モーツァルト 帝劇;モ ツァルト 帝劇...</examples>
</intent>
<intent number="20" probability="0.020979020979021">
<description>モーツァルト 販売</description>
<examples> モーツァルト 販売...</examples>
</intent>
<intent number="21" probability="0.020979020979021">
<description>モーツァルト dvd</description>
<examples> モーツァルト dvd...</examples>
</intent>
<intent number="22" probability="0.020979020979021">
<description>NHK 毎日モーツァルト</description>
<examples> 毎日モーツァルト...</examples>
</intent>
<intent number="23" probability="0.00699300699300699">
<description>Others</description>
<examples> ジュンス モーツァルト;チベットのモーツァルト;バッケンモーツァルト;モ ツァルト ケ キ;モーツァルト K;
モーツァルト ヴァイオリン;モーツァルト 岡山;モ ツァルト 岡山;モーツァルト 結婚;モーツァルト 梅田;モ ツァルト
梅田;モーツァルト 長調;モーツァルト Mozartプログラミングシステム Oz プログラミング言語;モーツァルト 映
画...</examples>
</intent>
</topic>
```

Scores
Symphony
Concerto
Music
Media files
Performances
Wolfgang Amadeus Mozart
Piano concerto
Die Zauberflöte
Opera
Turkish March
Composition
Mozart effect
Don Giovanni
Sonata
Figaro
Mozart family
Requiem lyric
Imperial Garden Theater
Sales
DVD
Mozart Every Day (TV program title)

Figure 5. Japanese topic 0115 “Mozart” and its intents, with rough English translations.

6. Subtopic Mining Results

6.1 Chinese Subtopic Mining Results

Table 4 shows the SYSDESC (system description) fields of all Chinese Subtopic Mining runs. Tables 5-7 show the mean I-rec, D-nDCG and D♯-nDCG values for $l = 10, 20, 30$ for all of these runs. The runs are sorted by mean D♯-nDCG.

Figures 6 show the I-rec / D-nDCG graphs [8] for the Chinese Subtopic Mining runs to visualise the trade offs between diversity and overall relevance.

Among the top performers, it can be observed that THU-S-C-2, THU-S-C-3 and ICTIR-S-C-1 are ranked within top five in terms of D♯-nDCG for all three document cutoffs. Moreover, the graphs show that while THU-S-C-2 and THU-S-C-3 are relatively relevance-oriented (i.e., their D-nDCG values are high), ICTIR-S-C-1 is relatively diversity-oriented (i.e., their I-rec values are high).

The set of 42 runs were tested for statistical significance with D♯-nDCG using a randomised version of two-sided Tukey's Honestly Significant Differences (HSD) test with 1000 trials [2]. Run pairs significantly different at $\alpha = 0.05$ are listed up in Figures 11-13 (**Appendix**). Of the 861 run pairs, 450 pairs were significantly different with $l = 10$; 490 pairs were significantly different with $l = 20$; and 489 pairs were significantly different with $l = 30$. The aforementioned three top performers are not significantly different from one another.

It is also clear that the correlations between D-nDCG and I-rec are very high: .596, .654, .700 in Kendall's τ and .936, .929, .930 in Pearson correlation for $l = 10, 20, 30$.

6.2 Japanese Subtopic Mining Results

Table 8 shows the SYSDESC (system description) fields of all Japanese Subtopic Mining runs. Note that NTU's approaches appear to be very similar to our WSE runs: in fact, NTU-S-J-2 is *identical* to WSE-S-J-3, except that the ranked list size differ for exactly two topics (out of one hundred). Hence the performances of these two runs are identical.

Tables 9-11 show the *Official* mean I-rec, D-nDCG and D♯-nDCG values for $l = 10, 20, 30$ for all of these runs. The runs are sorted by mean D♯-nDCG. Similarly, Tables 12-14 show the *Revised* results, after incorporating the subtopic contributions from NTU.

Figure 7 shows the Official I-rec / D-nDCG graphs for the Japanese Subtopic Mining runs to visualise the trade offs between diversity and overall relevance. Similarly, Figure 8 shows the Revised results.

It can be observed that, while the top performers perform similary at $l = 10$, ORG-S-J-1 and ORG-S-J-2 clearly stand out at $l = 20, 30$, both in the Official and in the Revised results. These two runs are followed by WSE-S-J-3/NTU-S-J-2 (these runs perform identically, as was mentioned earlier), WSE-S-J-1 and uogTr-S-J-2, in both sets of results.

For the revised results, the set of 14 runs were tested for statistical significance with D♯-nDCG using randomised two-sided Tukey's HSD with 1000 trials. Run pairs significantly different at $\alpha = 0.05$ are listed up in Figures 14-16 (**Appendix**). Of the 91 run pairs, 34 pairs were significantly different with $l = 10$; and 52 pairs were significantly different with $l = 20, 30$. Among the top performers mentioned above, both ORG-S-J-1 and ORG-S-J-2 are significantly better than WSE-S-J-3/NTU-S-J-2, WSE-S-J-1 and uogTr-S-J-2 with $l = 20, 30$.

Again, the correlations between D-nDCG and I-rec are high: .556, .778, .822 (Official) and .538, .780, .846 (Revised) in Kendall's τ; .985, .990, .992 (Official) and .976, .983, .986 (Revised) in Pearson correlation, for $l = 10, 20, 30$.

Table 4. Chinese subtopic mining runs with their SYSDESC fields.

run name	SYSDESC field
DBIIR-S-C-1	We use related queries to a given query and urls that users clicked in the query log, and build a bipartite graph. Then SimRank is used to calculate the similarities between queries. At last, these queries are clustered into several clusters which represent different intents. And for those queries we can get little information from query log, we use the results of search engines.
HIT2jointNLPLab-S-C-1	We used SogouQ query log to find potential user intent or sub-queries. First of all, we counted up how many times a given topic has appeared in the log. Then we counted up those log entries in which certain topic has appeared, and calculated the top 32 most frequently appeared bi-grams both before and after it. Using these bi-grams, we linked them and formed possible sub-queries. Finally, we calculated the similarity between sub-queries for each topic and roughly eliminated the redundent ones.
HIT2jointNLPLab-S-C-2	*ditto*
HITIR-S-C-1	We extracted query candidates from multiple sources including "sogou query log", "sogou T corpus" and "Baidu Baike". Futher we applied Affility propagation algorithm to cluster these candidates. One examplar is selected for every cluster and is used to represented a sub topic of the query.
HITIR-S-C-2	We first extracted query candidates from Sogou Query Log and Sogou T corpus. Then we selected a set of intent words and finally ranked these intent words
HITIR-S-C-3	We extracted query candidates from multiple sources including "sogou query log", "sogou T corpus" and "Baidu Baike". Futher we apnbnsplied Affility propagation algorithm (asymmetric similarity)to cluster these candidates. One examplar is selected for every cluster and is used to represented a sub topic of the query.
HITIR-S-C-4	We extracted query candidates from multiple sources including "sogou query log", "sogou T corpus". Futher we applied Affility propagation algorithm(using symmetric similarity) to cluster these candidates. One examplar is selected for every cluster and is used to represented a sub topic of the query.
HITIR-S-C-5	We extracted query candidates from multiple sources including "sogou query log", "sogou T corpus" and "Baidu Baike". Futher we applied Affility propagation algorithm (not choose different preference) to cluster these candidates. One examplar is selected for every cluster and is used to represented a sub topic of the query.
ICTIR-S-C-1	For every topic, we first collect related queries from query logs. To solve the sparseness problem, external resources such as real search engines and online encycopedias are also used. Then, several clustering algorithms are conducted to find subtopics. Finally, each cluster forms one subtopic and one query is selected to represent the subtopic. (Other resources we used are Sina query logs, Baidu search, Sogou search, SoSo search, Wiki, and Hudongbaike.)
ICTIR-S-C-2	*ditto*
ICTIR-S-C-3	*ditto*
ICTIR-S-C-4	*ditto*
ICTIR-S-C-5	*ditto*
III_CYUT_NTHU-S-C-1	Constructing a subtopic list using query log
ISCAS-S-C-1	RUNS ON RESULTS OF BAIDU BAIKE / GOOGLE / BAIDU, 10 CLUSTERS
ISCAS-S-C-2	RUN ON SOGOU LOG, 10 CLUSTERS
ISCAS-S-C-3	RUNS ON RESULTS OF BAIDU BAIKE / GOOGLE / BAIDU, 5 CLUSTERS
ISCAS-S-C-4	RUN ON SOGOU LOG, 5 CLUSTERS
KLE-S-C-1	Only query-log : freq 70% and url 30%
KLE-S-C-2	Only query-log : freq 50% and url 50%
KLE-S-C-3	Only query-log : freq 30% and url 70%
MSINT-S-C-1	Combine Log, D1_filter and D2 by method 1 and diversify the list by both string similarity and search results similarity. No external resource is used.
MSINT-S-C-2	Combine Log, D1_filter and D2 by method 2 and diversify the list by string similarity. No external resource is used.
MSINT-S-C-3	D1_filter:clustering query-dependent fragments from top 1000 documents, and then generating descriptions for clusters. Furthermore, we apply a model trained from example topics to filter noisy subtopics. No external resource is used.
MSINT-S-C-4	D2:clustering query-dependent fragments from top 1000 documents, and then generating descriptions for clusters. Furthermore, we rank subtopics by using n-grams trained on SogouQ. No external resource is used.
MSINT-S-C-5	D1:clustering query-dependent fragments from top 1000 documents, and then generating descriptions for clusters. No external resource is used.
NTU-S-C-1	document clustering
NTU-S-C-2	using google's related search
NTU-S-C-3	using bing's related search
THU-S-C-1	Hints from Search Engines with user needs rerank, removing duplicate ones with Qurey-Url graph model
THU-S-C-2	Hints from Search Engines with user needs rerank, removing duplicate ones with Qurey-Url graph model, reranking based on snippets and titles of pages
THU-S-C-3	Hints from Search Engines with user needs rerank
THU-S-C-4	Topics generated based on the log, using query-url model. Appended with anchor text according to retrieved documents.
THU-S-C-5	Topics generated based on large logs, using query-url model. Appended with anchor text according to retrieved documents.
TUTA1-S-C-1	Intent mining through traversing the graph constructed with query segments, using query log of SougouLab and snippets of google as resources
uogTr-S-C-1	SogouQ subtopic sample re-ranked by a data-driven approach
uogTr-S-C-2	SogouQ subtopic sample re-ranked by reformulation likelihood
uogTr-S-C-3	SogouQ subtopic sample comprising partial matches with common sessions or clicks to each test query
uogTr-S-C-4	SogouQ subtopic sample re-ranked by a data-driven approach and diversified by clicked hosts
uogTr-S-C-5	Query suggestions provided by a web search engine
UWat-S-C-1	We explore query intents discovery from anchor text and linkage information in the corpus.
UWat-S-C-2	*ditto*

Table 5. Chinese Subtopic Mining runs ranked by mean D♯-nDCG@10. The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
THU-S-C-2	0.4801	**0.7186**	**0.5993**
MSINT-S-C-2	0.5130	0.6806	0.5968
THU-S-C-3	0.4828	0.7107	0.5967
THU-S-C-1	0.4946	0.6896	0.5921
ICTIR-S-C-1	**0.5161**	0.6434	0.5797
uogTr-S-C-5	0.4947	0.6598	0.5772
MSINT-S-C-4	0.4864	0.6604	0.5734
ICTIR-S-C-4	0.5035	0.6417	0.5726
ICTIR-S-C-2	0.4826	0.6576	0.5701
HITIR-S-C-5	0.4936	0.6449	0.5693
ISCAS-S-C-1	0.5022	0.6336	0.5679
ICTIR-S-C-3	0.4808	0.6530	0.5669
HITIR-S-C-1	0.4854	0.6453	0.5653
ISCAS-S-C-3	0.4910	0.6386	0.5648
MSINT-S-C-1	0.5002	0.6240	0.5621
NTU-S-C-2	0.4683	0.6546	0.5615
MSINT-S-C-5	0.4578	0.6543	0.5560
NTU-S-C-3	0.4807	0.6308	0.5558
HITIR-S-C-4	0.4738	0.6291	0.5514
HITIR-S-C-3	0.4738	0.6291	0.5514
HIT2jointNLPLab -S-C-2	0.4596	0.6407	0.5501
MSINT-S-C-3	0.4587	0.6256	0.5422
ICTIR-S-C-5	0.4714	0.5832	0.5273
DBIIR-S-C-1	0.4694	0.5620	0.5157
HIT2jointNLPLab -S-C-1	0.4240	0.5946	0.5093
TUTA1-S-C-1	0.3405	0.6762	0.5084
NTU-S-C-1	0.4335	0.4836	0.4586
ISCAS-S-C-4	0.3062	0.4810	0.3936
KLE-S-C-3	0.3185	0.4461	0.3823
KLE-S-C-1	0.3162	0.4466	0.3814
KLE-S-C-2	0.3162	0.4464	0.3813
ISCAS-S-C-2	0.3019	0.4491	0.3755
THU-S-C-5	0.2888	0.4455	0.3672
III_CYUT_NTHU -S-C-1	0.3085	0.4099	0.3592
THU-S-C-4	0.2654	0.4040	0.3347
UWat-S-C-2	0.3324	0.3355	0.3340
uogTr-S-C-1	0.3210	0.3385	0.3297
uogTr-S-C-4	0.3176	0.3364	0.3270
UWat-S-C-1	0.2388	0.2492	0.2440
uogTr-S-C-2	0.1753	0.1772	0.1763
uogTr-S-C-3	0.1682	0.1698	0.1690
HITIR-S-C-2	0.0393	0.0242	0.0317

Table 6. Chinese Subtopic Mining runs ranked by mean D♯-nDCG@20. The highest value in each column is shown in bold.

run name	I-rec@20	D-nDCG@20	D♯-nDCG@20
ICTIR-S-C-1	**0.6997**	0.6162	**0.6579**
THU-S-C-3	0.6311	0.6727	0.6519
THU-S-C-2	0.6227	**0.6739**	0.6483
ICTIR-S-C-2	0.6444	0.6460	0.6452
ISCAS-S-C-3	0.6478	0.6370	0.6424
THU-S-C-1	0.6311	0.6508	0.6409
ISCAS-S-C-1	0.6406	0.6387	0.6397
HITIR-S-C-5	0.6421	0.6180	0.6300
MSINT-S-C-2	0.6066	0.6462	0.6264
HITIR-S-C-1	0.6316	0.6213	0.6264
MSINT-S-C-4	0.6293	0.6008	0.6150
HITIR-S-C-4	0.6235	0.6027	0.6131
HITIR-S-C-3	0.6235	0.6027	0.6131
MSINT-S-C-5	0.6069	0.6122	0.6096
MSINT-S-C-3	0.5962	0.5852	0.5907
ICTIR-S-C-4	0.6206	0.5579	0.5893
ICTIR-S-C-3	0.5849	0.5913	0.5881
MSINT-S-C-1	0.6187	0.5506	0.5846
TUTA1-S-C-1	0.4794	0.6677	0.5735
HIT2jointNLPLab -S-C-2	0.5583	0.5736	0.5660
ICTIR-S-C-5	0.5803	0.5427	0.5615
HIT2jointNLPLab -S-C-1	0.5116	0.5498	0.5307
uogTr-S-C-5	0.4947	0.4278	0.4612
NTU-S-C-2	0.4683	0.4242	0.4463
NTU-S-C-3	0.4807	0.4090	0.4449
DBIIR-S-C-1	0.4926	0.3948	0.4437
UWat-S-C-2	0.4945	0.3893	0.4419
KLE-S-C-3	0.4482	0.4344	0.4413
KLE-S-C-2	0.4443	0.4329	0.4386
KLE-S-C-1	0.4443	0.4326	0.4385
ISCAS-S-C-4	0.4053	0.4626	0.4340
ISCAS-S-C-2	0.3922	0.4434	0.4178
uogTr-S-C-1	0.4187	0.3670	0.3929
THU-S-C-5	0.3567	0.4286	0.3926
III_CYUT_NTHU -S-C-1	0.3890	0.3946	0.3918
uogTr-S-C-4	0.4170	0.3662	0.3916
THU-S-C-4	0.3568	0.3967	0.3767
NTU-S-C-1	0.4335	0.3140	0.3738
UWat-S-C-1	0.3236	0.2459	0.2847
uogTr-S-C-2	0.2407	0.1691	0.2049
uogTr-S-C-3	0.2245	0.1796	0.2020
HITIR-S-C-2	0.0416	0.0226	0.0321

Table 7. Chinese Subtopic Mining runs ranked by mean D♯-nDCG@30. The highest value in each column is shown in bold.

run name	I-rec@30	D-nDCG@30	D♯-nDCG@30
ICTIR-S-C-2	0.7070	0.5895	**0.6482**
THU-S-C-3	0.6844	**0.6101**	0.6473
ISCAS-S-C-1	0.6861	0.5783	0.6322
ICTIR-S-C-1	**0.7224**	0.5299	0.6261
THU-S-C-2	0.6663	0.5750	0.6206
THU-S-C-1	0.6686	0.5667	0.6176
ISCAS-S-C-3	0.6884	0.5419	0.6152
HITIR-S-C-5	0.6730	0.5529	0.6130
HITIR-S-C-1	0.6634	0.5531	0.6083
MSINT-S-C-5	0.6500	0.5412	0.5956
MSINT-S-C-4	0.6638	0.5150	0.5894
MSINT-S-C-2	0.6275	0.5390	0.5832
HITIR-S-C-4	0.6479	0.5182	0.5830
HITIR-S-C-3	0.6479	0.5182	0.5830
MSINT-S-C-3	0.6218	0.5022	0.5620
MSINT-S-C-1	0.6432	0.4662	0.5547
ICTIR-S-C-3	0.6062	0.4867	0.5464
ICTIR-S-C-4	0.6340	0.4441	0.5390
TUTA1-S-C-1	0.4982	0.5602	0.5292
HIT2jointNLPLab -S-C-2	0.5775	0.4681	0.5228
ICTIR-S-C-5	0.5924	0.4407	0.5165
HIT2jointNLPLab -S-C-1	0.5297	0.4566	0.4931
KLE-S-C-2	0.4769	0.3712	0.4241
KLE-S-C-1	0.4769	0.3709	0.4239
ISCAS-S-C-4	0.4394	0.4066	0.4230
KLE-S-C-3	0.4776	0.3677	0.4226
ISCAS-S-C-2	0.4320	0.4059	0.4189
UWat-S-C-2	0.5110	0.3154	0.4132
uogTr-S-C-5	0.4947	0.3309	0.4128
DBIIR-S-C-1	0.4953	0.3068	0.4010
NTU-S-C-3	0.4807	0.3163	0.3985
NTU-S-C-2	0.4683	0.3278	0.3980
uogTr-S-C-4	0.4534	0.3386	0.3960
uogTr-S-C-1	0.4534	0.3386	0.3960
THU-S-C-4	0.3917	0.3507	0.3712
THU-S-C-5	0.3803	0.3598	0.3700
III_CYUT_NTHU -S-C-1	0.3890	0.3042	0.3466
NTU-S-C-1	0.4335	0.2432	0.3384
UWat-S-C-1	0.3270	0.1926	0.2598
uogTr-S-C-2	0.2640	0.1470	0.2055
uogTr-S-C-3	0.2422	0.1650	0.2036
HITIR-S-C-2	0.0432	0.0197	0.0315

Figure 6. I-rec (x-axis) / D-nDCG (y-axis) graphs for Chinese Subtopic Mining. Top five runs in terms of mean D♯-nDCG are indicated by a baloon.

Table 8. Japanese subtopic mining runs with their SYSDESC fields.

run name	SYSDESC field
ORG-S-J-1	MSINT: anchor
ORG-S-J-2	MSINT: clustering
RMIT-S-J-1	pattern match from Japanese Wikipedia
uogTr-S-J-1	ClueWeb09-JA anchor-text subtopic sample comprising anchors that refer to URLs referred to by each test query
uogTr-S-J-2	Query suggestions provided by a web search engine
WSE-S-J-1	Bing Related results
WSE-S-J-2	Bing Suggested results
WSE-S-J-3	Google Related results
WSE-S-J-4	Google Suggested results
WSE-S-J-5	Yahoo Japan related results
(NTU-S-J-1)	document clustering
(NTU-S-J-2)	using google's related search
(NTU-S-J-3)	using bing's related search
(NTU-S-J-4)	using yahoo's related search

Table 9. Japanese Subtopic Mining runs ranked by mean D♯-nDCG@10 (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-2	**0.4513**	0.4022	**0.4268**
WSE-S-J-3	0.4346	0.4081	0.4214
ORG-S-J-1	0.4355	0.4056	0.4205
uogTr-S-J-2	0.4321	0.4071	0.4196
WSE-S-J-1	0.4290	0.3982	0.4136
WSE-S-J-4	0.4084	**0.4128**	0.4106
WSE-S-J-5	0.3928	0.3968	0.3948
WSE-S-J-2	0.3140	0.3636	0.3388
RMIT-S-J-1	0.0876	0.0973	0.0925
uogTr-S-J-1	0.0113	0.0045	0.0079

Table 10. Japanese Subtopic Mining runs ranked by mean D♯-nDCG@20 (Official). The highest value in each column is shown in bold.

run name	I-rec@20	D-nDCG@20	D♯-nDCG@20
ORG-S-J-1	**0.6307**	0.4201	**0.5254**
ORG-S-J-2	0.6159	**0.4202**	0.5180
WSE-S-J-3	0.4346	0.2958	0.3652
uogTr-S-J-2	0.4321	0.2979	0.3650
WSE-S-J-1	0.4290	0.2916	0.3603
WSE-S-J-4	0.4084	0.3012	0.3548
WSE-S-J-5	0.3928	0.2883	0.3405
WSE-S-J-2	0.3140	0.2643	0.2892
RMIT-S-J-1	0.0876	0.0709	0.0793
uogTr-S-J-1	0.0270	0.0077	0.0173

Table 11. Japanese Subtopic Mining runs ranked by mean D♯-nDCG@30 (Official). The highest value in each column is shown in bold.

run name	I-rec@30	D-nDCG@30	D♯-nDCG@30
ORG-S-J-1	**0.6533**	**0.3867**	**0.5200**
ORG-S-J-2	0.6266	0.3823	0.5044
uogTr-S-J-2	0.4321	0.2629	0.3475
WSE-S-J-3	0.4346	0.2597	0.3472
WSE-S-J-1	0.4290	0.2574	0.3432
WSE-S-J-4	0.4084	0.2649	0.3366
WSE-S-J-5	0.3928	0.2517	0.3222
WSE-S-J-2	0.3140	0.2316	0.2728
RMIT-S-J-1	0.0876	0.0624	0.0750
uogTr-S-J-1	0.0270	0.0081	0.0175

Table 12. Japanese Subtopic Mining runs ranked by mean D♯-nDCG@10 (Revised). The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
WSE-S-J-3	0.4442	0.4244	**0.4343**
NTU-S-J-2	0.4442	0.4244	**0.4343**
ORG-S-J-2	**0.4523**	0.3991	0.4257
WSE-S-J-1	0.4373	0.4131	0.4252
uogTr-S-J-2	0.4353	0.4120	0.4236
WSE-S-J-4	0.4172	**0.4288**	0.4230
ORG-S-J-1	0.4355	0.4025	0.4190
WSE-S-J-5	0.4092	0.4187	0.4139
NTU-S-J-3	0.4205	0.3913	0.4059
NTU-S-J-4	0.3935	0.4060	0.3998
WSE-S-J-2	0.3238	0.3800	0.3519
NTU-S-J-1	0.3021	0.2409	0.2715
RMIT-S-J-1	0.0876	0.0961	0.0918
uogTr-S-J-1	0.0113	0.0045	0.0079

Table 13. Japanese Subtopic Mining runs ranked by mean D♯-nDCG@20 (Revised). The highest value in each column is shown in bold.

run name	I-rec@20	D-nDCG@20	D♯-nDCG@20
ORG-S-J-1	**0.6344**	**0.4142**	**0.5243**
ORG-S-J-2	0.6188	**0.4142**	0.5165
WSE-S-J-3	0.4442	0.3043	0.3742
NTU-S-J-2	0.4442	0.3043	0.3742
WSE-S-J-1	0.4373	0.2997	0.3685
uogTr-S-J-2	0.4353	0.2988	0.3670
WSE-S-J-4	0.4172	0.3092	0.3632
WSE-S-J-5	0.4092	0.3001	0.3546
NTU-S-J-3	0.4205	0.2831	0.3518
NTU-S-J-4	0.3935	0.2904	0.3420
WSE-S-J-2	0.3238	0.2733	0.2986
NTU-S-J-1	0.3021	0.1741	0.2381
RMIT-S-J-1	0.0876	0.0690	0.0783
uogTr-S-J-1	0.0270	0.0075	0.0172

Table 14. Japanese Subtopic Mining runs ranked by mean D♯-nDCG@30 (Revised). The highest value in each column is shown in bold.

run name	I-rec@30	D-nDCG@30	D♯-nDCG@30
ORG-S-J-1	**0.6583**	**0.3812**	**0.5197**
ORG-S-J-2	0.6318	0.3745	0.5031
WSE-S-J-3	0.4442	0.2647	0.3544
NTU-S-J-2	0.4442	0.2647	0.3544
WSE-S-J-1	0.4373	0.2622	0.3498
uogTr-S-J-2	0.4353	0.2614	0.3483
WSE-S-J-4	0.4172	0.2691	0.3432
WSE-S-J-5	0.4092	0.2595	0.3343
NTU-S-J-3	0.4205	0.2469	0.3337
NTU-S-J-4	0.3935	0.2509	0.3222
WSE-S-J-2	0.3238	0.2371	0.2805
NTU-S-J-1	0.3021	0.1522	0.2272
RMIT-S-J-1	0.0876	0.0601	0.0739
uogTr-S-J-1	0.0270	0.0078	0.0174

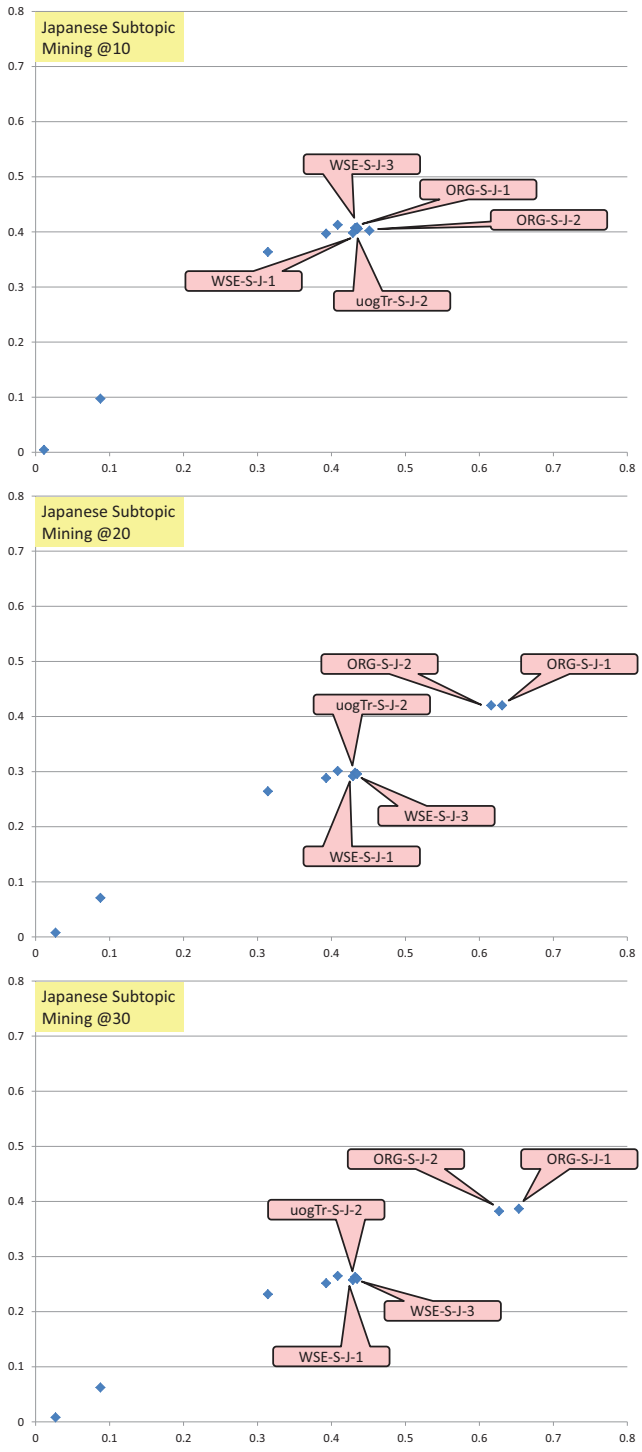

Figure 7. I-rec (x-axis) / D-nDCG (y-axis) graphs for Japanese Subtopic Mining (Official). Top five runs in terms of mean D♯-nDCG are indicated by a baloon.

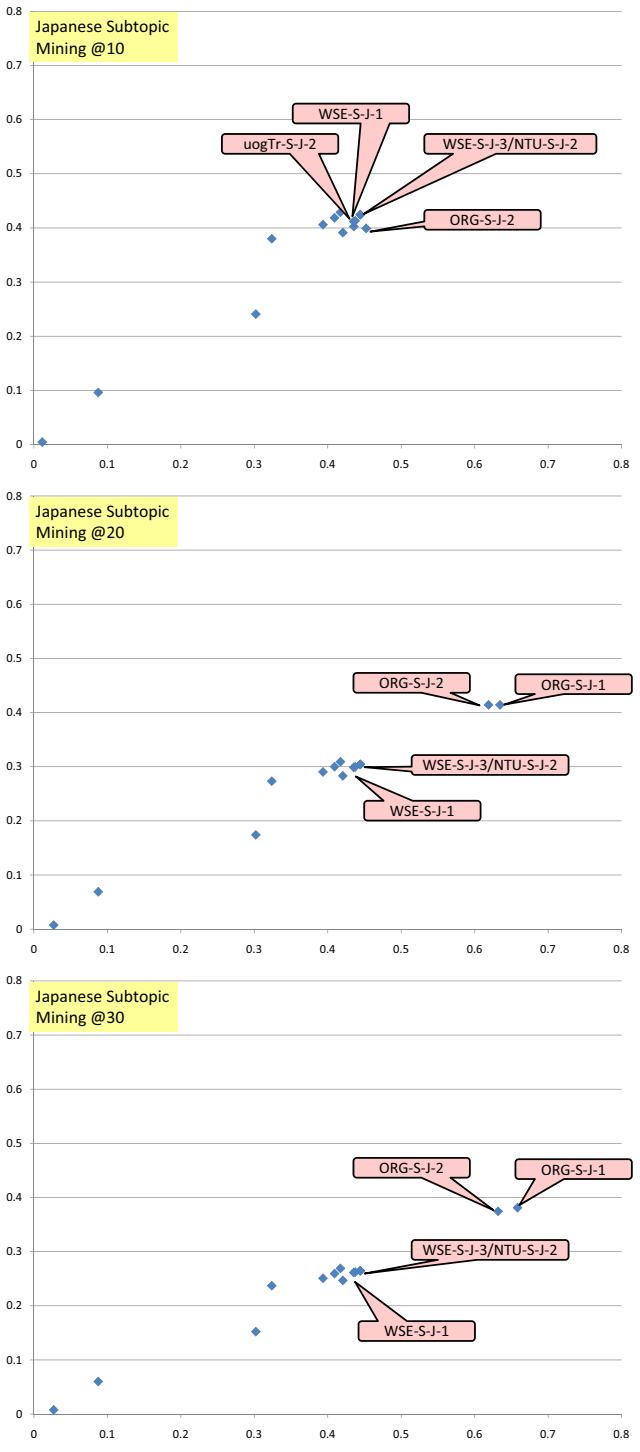

Figure 8. I-rec (x-axis) / D-nDCG (y-axis) graphs for Japanese Subtopic Mining (Revised). Top five runs in terms of mean D♯-nDCG are indicated by a baloon.

7. Document Ranking Results

7.1 Chinese Document Ranking Results

Table 15 shows the SYSDESC (system description) fields of all Chinese Document Ranking runs. Tables 16-18 show the mean I-rec, D-nDCG and D♯-nDCG values for $l = 10, 20, 30$ for all of these runs. The runs are sorted by mean D♯-nDCG.

Figure 9 shows the I-rec / D-nDCG graphs for the Chinese Document Ranking runs to visualise the trade offs between diversity and overall relevance.

Among the top performers, it can be observed that THUIR-D-C-1, THUIR-D-C-5 and uogTr-D-C-5 are ranked within top five in terms of D♯-nDCG for all three document cutoffs. On the other hand, Tables 16-18 also show that the runs from MSINT generally outperform other runs in terms of I-rec.

The set of 24 runs were tested for statistical significance with D♯-nDCG using randomised two-sided Tukey's HSD with 1000 trials. Run pairs significantly different at $\alpha = 0.05$ are listed up in Figures 17-19 (**Appendix**). Of the 276 run pairs, 140 pairs were significantly different with $l = 10$; 138 pairs were significantly different with $l = 20$; and 133 pairs were significantly different with $l = 30$. The aforementioned three top performers and the four MSINT runs (MSINT-D-C-{1,2,4,5}) are not significantly different from one another.

For these runs, the correlations between D-nDCG and I-rec are relatively weak: .442, .210, .188 in Kendall's τ; .611, .355, .398 in Pearson correlation, for $l = 10, 20, 30$.

Table 16. Chinese Document Ranking runs ranked by mean D♯-nDCG@10. The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
THUIR-D-C-5	0.6861	**0.4573**	**0.5717**
THUIR-D-C-1	0.6893	0.4542	0.5717
uogTr-D-C-5	0.6624	0.4374	0.5499
MSINT-D-C-1	0.7068	0.3854	0.5461
uogTr-D-C-2	0.6600	0.4316	0.5458
MSINT-D-C-4	**0.7091**	0.3822	0.5456
uogTr-D-C-4	0.6474	0.4423	0.5449
MSINT-D-C-2	0.7003	0.3783	0.5393
uogTr-D-C-3	0.6301	0.4480	0.5390
MSINT-D-C-5	0.6936	0.3783	0.5359
uogTr-D-C-1	0.6406	0.4252	0.5329
THUIR-D-C-2	0.6495	0.3853	0.5174
HIT2jointNLPLab -D-C-2	0.5794	0.3704	0.4749
NTU-D-C-1	0.6180	0.3314	0.4747
SJTUBCMI-D-C-2	0.6008	0.3317	0.4663
MSINT-D-C-3	0.5987	0.3222	0.4604
SJTUBCMI-D-C-3	0.5856	0.3288	0.4572
SJTUBCMI-D-C-5	0.6228	0.2816	0.4522
SJTUBCMI-D-C-4	0.6108	0.2756	0.4432
SJTUBCMI-D-C-1	0.6038	0.2654	0.4346
THUIR-D-C-3	0.5979	0.2598	0.4288
THUIR-D-C-4	0.6001	0.2569	0.4285
HIT2jointNLPLab -D-C-1	0.4716	0.3573	0.4144
III_CYUT_NTHU -D-C-1	0.4630	0.2040	0.3335

7.2 Japanese Document Ranking Results

Table 19 shows the SYSDESC (system description) fields of all Japanese Document Ranking runs. Tables 20-22 show the mean I-rec, D-nDCG and D♯-nDCG values for $l = 10, 20, 30$ for all of these runs. The runs are sorted by mean D♯-nDCG.

Figure 10 shows the I-rec / D-nDCG graphs for the Japanese Document Ranking runs to visualise the trade offs between diversity and overall relevance.

It can be observed that the runs from MSINT are the overall winners. Their performances are high both in terms of D-nDCG and I-rec. On the other hand, uogTr-D-J-1 and uogTr-D-J-2 seem to do very well in terms of D-nDCG.

The set of 18 runs were tested for statistical significance with D♯-nDCG using randomised two-sided Tukey's HSD with 1000 trials. Run pairs significantly different at $\alpha = 0.05$ are listed up in Figures 20-22 (**Appendix**). Of the 105 run pairs, 50 pairs were significantly different with $l = 10$; 45 pairs were significantly different with $l = 20$; and 41 pairs were significantly different with $l = 30$. The top performer MSINT-D-J-3 is significantly better than uogTr-D-J-{3,4,5} with $l = 10, 20, 30$. Other differences among these top two teams are not statistically significant.

The correlations between D-nDCG and I-rec are high: .657, .581, .562 in Kendall's τ; .767, .660, .701 in Pearson correlation, for $l = 10, 20, 30$.

Table 17. Chinese Document Ranking runs ranked by mean D♯-nDCG@20. The highest value in each column is shown in bold.

run name	I-rec@20	D-nDCG@20	D♯-nDCG@20
THUIR-D-C-5	0.7937	0.4764	**0.6351**
THUIR-D-C-1	0.7937	0.4755	0.6346
uogTr-D-C-5	0.7603	0.4661	0.6132
uogTr-D-C-4	0.7500	0.4748	0.6124
uogTr-D-C-3	0.7430	**0.4782**	0.6106
uogTr-D-C-2	0.7550	0.4658	0.6104
uogTr-D-C-1	0.7458	0.4517	0.5987
MSINT-D-C-5	**0.8095**	0.3801	0.5948
MSINT-D-C-1	0.8055	0.3836	0.5946
MSINT-D-C-2	0.8010	0.3828	0.5919
MSINT-D-C-4	0.8013	0.3806	0.5909
THUIR-D-C-2	0.7633	0.4015	0.5824
SJTUBCMI-D-C-2	0.7213	0.3536	0.5375
HIT2jointNLPLab -D-C-2	0.6815	0.3928	0.5371
NTU-D-C-1	0.6952	0.3706	0.5329
MSINT-D-C-3	0.7245	0.3304	0.5274
SJTUBCMI-D-C-3	0.6918	0.3476	0.5197
THUIR-D-C-3	0.7553	0.2719	0.5136
SJTUBCMI-D-C-5	0.7424	0.2845	0.5134
THUIR-D-C-4	0.7484	0.2714	0.5099
SJTUBCMI-D-C-4	0.7388	0.2772	0.5080
SJTUBCMI-D-C-1	0.7464	0.2615	0.5039
HIT2jointNLPLab -D-C-1	0.5499	0.3819	0.4659
III_CYUT_NTHU -D-C-1	0.5658	0.2179	0.3919

Table 15. Chinese document ranking runs with their SYSDESC fields.

run name	SYSDESC field
HIT2jointNLPLab-D-C-1	we rerank the documents by using the Maximal Marginal Relevance method.
HIT2jointNLPLab-D-C-2	we rerank the documents by integrating all the results of sub-topics according to the score.
III_CYUT_NTHU-D-C-1	clustering retrieve and clustering rank
MSINT-D-C-1	Diversification by sites, anchor, log, top results v1, and v2
MSINT-D-C-2	Diversification by sites, anchor, and merged log, v1 and v2
MSINT-D-C-3	A baseline ranking without diversificaton
MSINT-D-C-4	Diversification by sites, anchor, log, and top results v1
MSINT-D-C-5	Diversification by sites, anchor, log, and top results v2
NTU-D-C-1	document clustering
SJTUBCMI-D-C-1	BM25F, PageRank, Similar Pages Removal
SJTUBCMI-D-C-2	BM25F, PageRank(lamdba=0.4)
SJTUBCMI-D-C-3	BM25F
SJTUBCMI-D-C-4	BM25, PR, similar pages removal(0.1, -0.9)
SJTUBCMI-D-C-5	BM25, similar pages removal(0,-0.9)
THUIR-D-C-1	base+HITS+RemoveDuplicate
THUIR-D-C-2	base+HITS+Novelty
THUIR-D-C-3	DMeasure+selected
THUIR-D-C-4	DMeasure+selected+browse+graph
THUIR-D-C-5	base+HITS
uogTr-D-C-1	xQuAD with learned models uniformly applied with query suggestions provided by a web search engine
uogTr-D-C-2	xQuAD with learned models selectively applied with query suggestions provided by a web search engine
uogTr-D-C-3	xQuAD with learned models uniformly applied with SogouQ subtopics ranked by a data-driven approach
uogTr-D-C-4	xQuAD with learned models selectively applied with SogouQ subtopics ranked by a data-driven approach
uogTr-D-C-5	xQuAD with learned models uniformly applied with SogouQ subtopics ranked by reformulation likelihood

Table 19. Japanese document ranking runs with their SYSDESC fields.

run name	SYSDESC field
MSINT-D-J-1	MSINT: WSE sugggestion, WSE related queries, and site domain
MSINT-D-J-2	MSINT: WSE sugggestion and site domain
MSINT-D-J-3	MSINT: WSE sugggestion
MSINT-D-J-4	MSINT: WSE related queries and site domain
MSINT-D-J-5	MSINT: baseline
RMIT-D-J-1	BM25 Rank with neWT on Merged Diversity Query Set 1
RMIT-D-J-2	BM25 Rank with neWT on Vanilla Query Set 2
RMIT-D-J-3	Raw Rank with neWT on Merged Diversity Query Set 1
RMIT-D-J-4	BM25 Rank with neWT on Vanilla Query Set 1
RMIT-D-J-5	Raw Rank with neWT on Vanilla Query Set 2
uogTr-D-J-1	xQuAD with learned models uniformly applied with query suggestions provided by a web search engine
uogTr-D-J-2	xQuAD with learned models selectively applied with query suggestions provided by a web search engine
uogTr-D-J-3	xQuAD with learned models uniformly applied with ClueWeb09-JA anchor-text subtopics
uogTr-D-J-4	xQuAD with learned models selectively applied with ClueWeb09-JA anchor-text subtopics
uogTr-D-J-5	Learned models leveraging several features
(NTU-D-J-1)	document clustering
(NTU-D-J-2)	use google related query to rerank top 1000 documents
(NTU-D-J-3)	use google related query to rerank top 1000 documents

Table 18. Chinese Document Ranking runs ranked by mean D♯-nDCG@30. The highest value in each column is shown in bold.

run name	I-rec@30	D-nDCG@30	D♯-nDCG@30
uogTr-D-C-4	0.7911	**0.4758**	**0.6334**
uogTr-D-C-5	0.8028	0.4632	0.6330
THUIR-D-C-5	0.8184	0.4473	0.6329
THUIR-D-C-1	0.8192	0.4456	0.6324
uogTr-D-C-3	0.7851	0.4698	0.6274
uogTr-D-C-2	0.7931	0.4600	0.6265
uogTr-D-C-1	0.7781	0.4331	0.6056
MSINT-D-C-1	0.8343	0.3645	0.5994
MSINT-D-C-4	**0.8349**	0.3609	0.5979
MSINT-D-C-2	0.8327	0.3625	0.5976
MSINT-D-C-5	0.8330	0.3619	0.5974
THUIR-D-C-2	0.7984	0.3905	0.5945
SJTUBCMI-D-C-2	0.7620	0.3351	0.5485
MSINT-D-C-3	0.7719	0.3156	0.5437
HIT2jointNLPLab -D-C-2	0.7057	0.3656	0.5357
NTU-D-C-1	0.7169	0.3473	0.5321
SJTUBCMI-D-C-5	0.7819	0.2673	0.5246
SJTUBCMI-D-C-3	0.7315	0.3157	0.5236
SJTUBCMI-D-C-4	0.7803	0.2596	0.5200
THUIR-D-C-4	0.7833	0.2550	0.5192
THUIR-D-C-3	0.7832	0.2527	0.5180
SJTUBCMI-D-C-1	0.7738	0.2409	0.5073
HIT2jointNLPLab -D-C-1	0.5752	0.3499	0.4625
III_CYUT_NTHU -D-C-1	0.5821	0.1869	0.3845

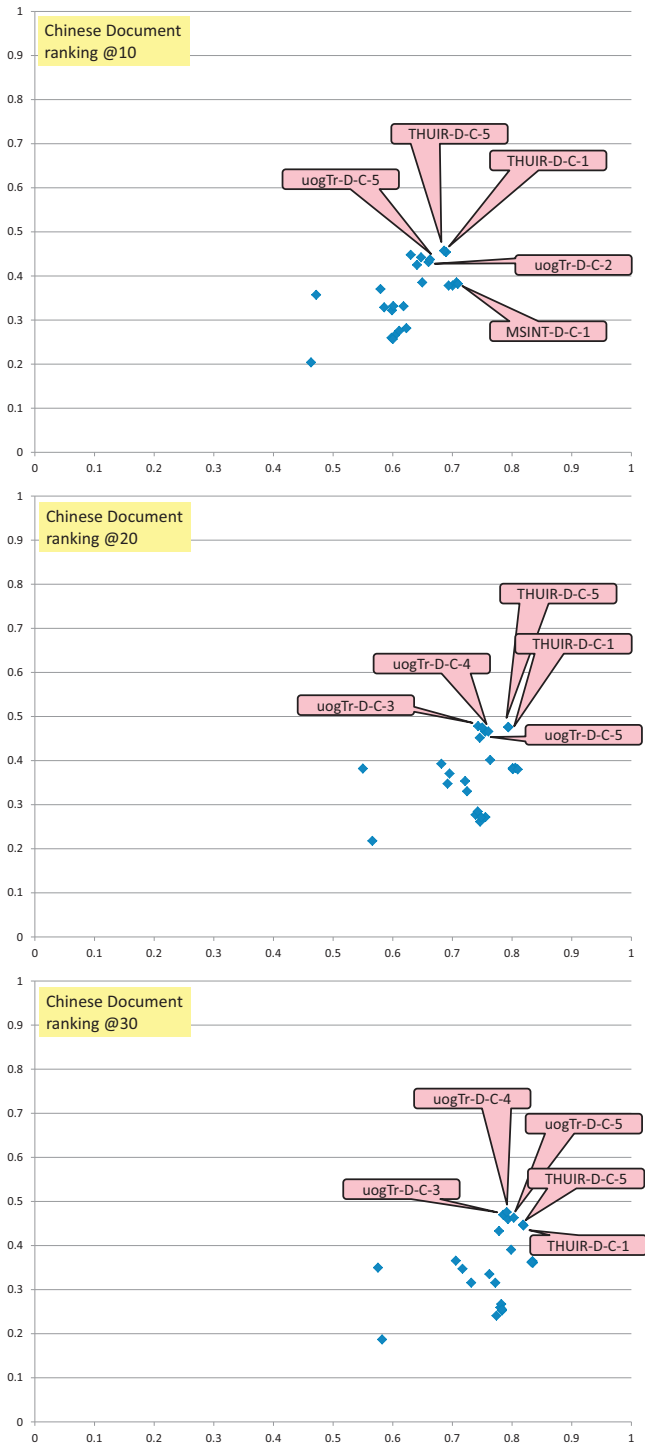

Figure 9. I-rec (x-axis) / D-nDCG (y-axis) graphs for Chinese Document Ranking. Top five runs in terms of mean D♯-nDCG are indicated by a baloon.

Table 20. Japanese Document Ranking runs ranked by mean D♯-nDCG@10. The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
MSINT-D-J-3	**0.7554**	0.4444	**0.5999**
MSINT-D-J-2	0.7548	0.4329	0.5938
MSINT-D-J-1	0.7488	0.4377	0.5933
MSINT-D-J-4	0.7526	0.4210	0.5868
uogTr-D-J-1	0.6845	**0.4715**	0.5780
uogTr-D-J-2	0.6840	0.4673	0.5756
MSINT-D-J-5	0.7063	0.4257	0.5660
uogTr-D-J-3	0.6702	0.4091	0.5397
uogTr-D-J-4	0.6660	0.4088	0.5374
uogTr-D-J-5	0.6697	0.4034	0.5365
RMIT-D-J-3	0.6723	0.3664	0.5193
RMIT-D-J-1	0.6356	0.3540	0.4948
RMIT-D-J-4	0.6489	0.3301	0.4895
RMIT-D-J-2	0.6306	0.3283	0.4795
RMIT-D-J-5	0.5639	0.2989	0.4314

Table 21. Japanese Document Ranking runs ranked by mean D♯-nDCG@20. The highest value in each column is shown in bold.

run name	I-rec@20	D-nDCG@20	D♯-nDCG@20
MSINT-D-J-3	0.8697	0.4821	**0.6759**
MSINT-D-J-1	0.8669	0.4644	0.6656
MSINT-D-J-2	**0.8720**	0.4577	0.6648
MSINT-D-J-4	0.8563	0.4519	0.6541
MSINT-D-J-5	0.8372	0.4692	0.6532
uogTr-D-J-1	0.7876	0.5138	0.6507
uogTr-D-J-2	0.7835	**0.5143**	0.6489
uogTr-D-J-4	0.7752	0.4522	0.6137
uogTr-D-J-5	0.7701	0.4531	0.6116
uogTr-D-J-3	0.7692	0.4493	0.6093
RMIT-D-J-3	0.7649	0.4037	0.5843
RMIT-D-J-4	0.7695	0.3803	0.5749
RMIT-D-J-2	0.7567	0.3725	0.5646
RMIT-D-J-1	0.7465	0.3796	0.5630
RMIT-D-J-5	0.6580	0.3324	0.4952

Table 22. Japanese Document Ranking runs ranked by mean D♯-nDCG@30. The highest value in each column is shown in bold.

run name	I-rec@30	D-nDCG@30	D♯-nDCG@30
MSINT-D-J-3	**0.9058**	0.4813	**0.6936**
MSINT-D-J-1	0.9025	0.4678	0.6852
MSINT-D-J-2	0.8968	0.4464	0.6716
uogTr-D-J-2	0.8231	**0.5014**	0.6622
MSINT-D-J-4	0.8850	0.4383	0.6617
uogTr-D-J-1	0.8224	0.5002	0.6613
MSINT-D-J-5	0.8679	0.4486	0.6583
uogTr-D-J-4	0.8046	0.4502	0.6274
uogTr-D-J-5	0.8028	0.4477	0.6252
uogTr-D-J-3	0.7972	0.4425	0.6198
RMIT-D-J-3	0.7836	0.3800	0.5818
RMIT-D-J-4	0.8012	0.3617	0.5814
RMIT-D-J-2	0.7977	0.3575	0.5776
RMIT-D-J-1	0.7752	0.3617	0.5684
RMIT-D-J-5	0.6759	0.3118	0.4938

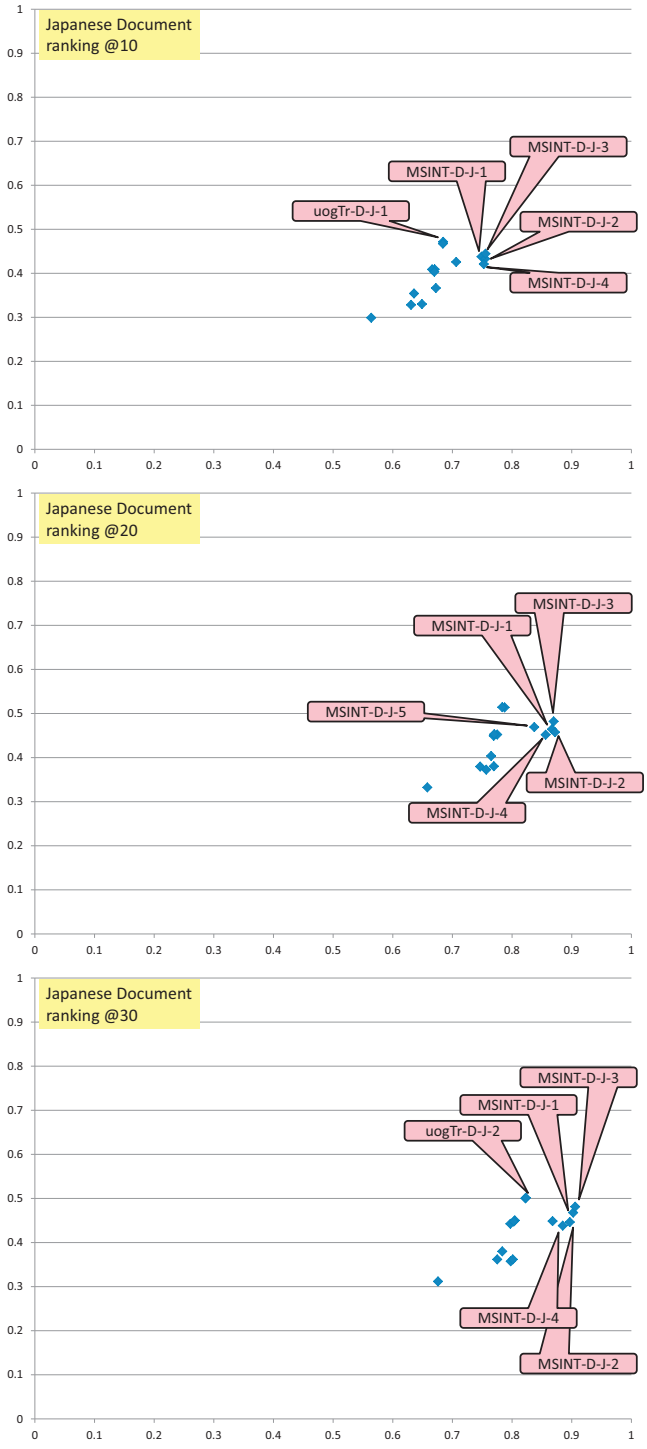

Figure 10. I-rec (x-axis) / D-nDCG (y-axis) graphs for Japanese Document Ranking. Top five runs in terms of mean D♯-nDCG are indicated by a baloon.

8. Conclusions and Future Work

The NTCIR-9 INTENT task attracted participating teams from seven different countries/regions – 16 teams for Subtopic Mining and 8 teams for Document Ranking, or 17 unique teams for the entire INTENT task. The Subtopic Mining subtask received 42 Chinese runs and 14 Japanese runs; the Document Ranking subtask received 24 Chinese runs and 18 Japanese runs. Both subtasks used D♯-nDCG as the primary evaluation metric, which explicitly combines the *overall relevance* across intents as measured by D-nDCG and the *diversity* as measured by intent recall. Our macro-level conclusions at this point are:

1. High-relevance runs (i.e. high D-nDCG runs) are generally also high-diversity runs (i.e. high I-rec runs), the exception being the Chinese Document Ranking runs which showed relatively low correlations between these two axes (Figure 9);
2. Some top performers achieve a high D♯-nDCG by achieving high relevance (e.g. the Chinese Subtopic Mining runs from THU, the Chinese Document Ranking runs from THUIR and uog, and the Japanese Document Ranking runs from uog); while other top performers achieve a high

   D♯-nDCG by achieving high diversity (e.g. the Chinese Subtopic Mining runs from ICTIR, the Chinese Document Ranking runs from MSINT, and the Japanese Document Ranking runs from MSINT).

Item 1 is not surprising, as a high D-nDCG implies that the system managed to cover many relevant documents for different intents, in which case intent recall will inevitably be high as well. The reason why we combine D-nDCG and intent recall explicitly is to discourage runs that retrieve many documents that are highly relevant to only a few popular intents [8].

The interested reader is encouraged to read the INTENT participants's papers.

9. Acknowledgments

We would like to thank the following people/institutions:

- INTENT participants for their research efforts and cooperation;
- Noriko Kando and the other NTCIR chairs for their support;
- Yijiang Jin for helping us set up and maintain the INTENT websites;
- Jaime Callan for providing the ClueWeb09-JA document collection;
- Sogou.com for providing SogouT, SogouQ and other related resources;
- Nick Craswell for his valuable advice.

10. References

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

[2] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, to appear.

[3] C. L. A. Clarke, N. Craswell and I. Soboroff. Overview of the TREC 2009 web track. In *Proceedings of TREC 2009*, 2010.

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

[5] T. Sakai. NTCIREVAL: A Generic Toolkit for Information Access Evaluation. In *Proceedings of the Forum on Information Technology 2011*, Vol. 2, pages 23–30, 2011.

[6] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama and C.-Y. Lin. Using graded-relevance metrics for evaluating community QA answer selection. In *Proceedings of ACM WSDM 2011*, pages 187–196, 2011.

[7] T. Sakai and N. Kando. Are Popular Documents More Likely To Be Relevant? A Dive into the ACLIA IR4QA Pools. In *Proceedings of EVIA 2008*, pages 8–9, 2008.

[8] T. Sakai and R. Song. Evaluating Diversified Search Results Using Per-Intent Graded Relevance. In *Proceedings of ACM SIGIR 2011*, pages 1043–1052, 2011.

[9] T. Sakai, M. P. Kato and Y.-I. Song. Overview of NTCIR-9 1CLICK. In *Proceedings of NTCIR-9*, to appear, 2011.

[10] R. L. T. Santos, C. Macdonald and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW 2010*, pages 881–890, 2010.

[11] R. Song, Z. Luo, J.-Y. Nie, Y. Yu, and H.-W. Hon. Identification of ambiguous queries in web search. *Information Processing and Management*, 45(2), 2008.

[12] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to Cluster Web Search Results. In *Proceedings of ACM SIGIR 2004*, pages 210–217, 2004.

Appendix

DBIIR-S-C-1 with HITIR-S-C-2,III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
HIT2jointNLPLab-S-C-1 with HITIR-S-C-2,III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,
uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
HIT2jointNLPLab-S-C-2 with HITIR-S-C-2,III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,
uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-1 with HITIR-S-C-2,III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-2 with HITIR-S-C-3,HITIR-S-C-4,HITIR-S-C-5,ICTIR-S-C-1,ICTIR-S-C-2,ICTIR-S-C-3,ICTIR-S-C-4,ICTIR-S-C-5,III_CYUT_NTHU-S-C-1,ISCAS-S-C-1,ISCAS-S-C-2,
ISCAS-S-C-3,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,
THU-S-C-2,THU-S-C-3,THU-S-C-4,THU-S-C-5,TUTA1-S-C-1,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-3 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,
UWat-S-C-1,UWat-S-C-2
HITIR-S-C-4 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,
UWat-S-C-1,UWat-S-C-2
HITIR-S-C-5 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,
UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-1 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-2 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-3 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,
UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-4 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-5 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,
UWat-S-C-1,UWat-S-C-2
III_CYUT_NTHU-S-C-1 with ISCAS-S-C-1,ISCAS-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,THU-S-C-2,THU-S-C-3,
TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-5,UWat-S-C-1
ISCAS-S-C-1 with ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
ISCAS-S-C-2 with ISCAS-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-5,UWat-S-C-1
ISCAS-S-C-3 with ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
ISCAS-S-C-4 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,
uogTr-S-C-3,uogTr-S-C-5,UWat-S-C-1
KLE-S-C-1 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-5,UWat-S-C-1
KLE-S-C-2 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-5,UWat-S-C-1
KLE-S-C-3 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-5,UWat-S-C-1
MSINT-S-C-1 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-2 with NTU-S-C-1,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-3 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-4 with NTU-S-C-1,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-5 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
NTU-S-C-1 with THU-S-C-1,THU-S-C-2,THU-S-C-3,THU-S-C-4,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
NTU-S-C-2 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
NTU-S-C-3 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
THU-S-C-1 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
THU-S-C-2 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
THU-S-C-3 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
THU-S-C-4 with TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-5
THU-S-C-5 with TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-5,UWat-S-C-1
TUTA1-S-C-1 with uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
uogTr-S-C-1 with uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-5
uogTr-S-C-2 with uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-2
uogTr-S-C-3 with uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-2
uogTr-S-C-4 with uogTr-S-C-5
uogTr-S-C-5 with UWat-S-C-1,UWat-S-C-2

Figure 11. Significantly different Chinese Subtopic Mining run pairs in terms of D♯-nDCG@10 (randomised Tukey's HSD at $\alpha = 0.05$).

DBIIR-S-C-1 with HIT2jointNLPLab-S-C-2,HITIR-S-C-1,HITIR-S-C-2,HITIR-S-C-3,HITIR-S-C-4,HITIR-S-C-5,ICTIR-S-C-1,ICTIR-S-C-2,ICTIR-S-C-3,ICTIR-S-C-4,ICTIR-S-C-5,
ISCAS-S-C-1,ISCAS-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
HIT2jointNLPLab-S-C-1 with HITIR-S-C-2,ICTIR-S-C-1,III_CYUT_NTHU-S-C-1,NTU-S-C-1,THU-S-C-2,THU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-4,UWat-S-C-1
HIT2jointNLPLab-S-C-2 with HITIR-S-C-2,III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,
THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-1 with HITIR-S-C-2,III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,
uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-2 with HITIR-S-C-3,HITIR-S-C-4,HITIR-S-C-5,ICTIR-S-C-1,ICTIR-S-C-2,ICTIR-S-C-3,ICTIR-S-C-4,ICTIR-S-C-5,III_CYUT_NTHU-S-C-1,ISCAS-S-C-1,ISCAS-S-C-2,
ISCAS-S-C-3,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,
THU-S-C-2,THU-S-C-3,THU-S-C-4,THU-S-C-5,TUTA1-S-C-1,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-3 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-4 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-5 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-1 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-2 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-3 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-4 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-5 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,
uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
III_CYUT_NTHU-S-C-1 with ISCAS-S-C-1,ISCAS-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,
uogTr-S-C-2,uogTr-S-C-3
ISCAS-S-C-1 with ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ISCAS-S-C-2 with ISCAS-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
UWat-S-C-1
ISCAS-S-C-3 with ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,
uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ISCAS-S-C-4 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
KLE-S-C-1 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
KLE-S-C-2 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
KLE-S-C-3 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
MSINT-S-C-1 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-2 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-3 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-4 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-5 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
NTU-S-C-1 with THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3
NTU-S-C-2 with THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
NTU-S-C-3 with THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
THU-S-C-1 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
THU-S-C-2 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
THU-S-C-3 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
THU-S-C-4 with TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3
THU-S-C-5 with TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3
TUTA1-S-C-1 with uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1,UWat-S-C-2
uogTr-S-C-1 with uogTr-S-C-2,uogTr-S-C-3
uogTr-S-C-2 with uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-2
uogTr-S-C-3 with uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-2
uogTr-S-C-5 with UWat-S-C-1
UWat-S-C-1 with UWat-S-C-2

Figure 12. Significantly different Chinese Subtopic Mining run pairs in terms of D♯-nDCG@20 (randomised Tukey's HSD at $\alpha = 0.05$).

DBIIR-S-C-1 with HIT2jointNLPLab-S-C-2,HITIR-S-C-1,HITIR-S-C-2,HITIR-S-C-3,HITIR-S-C-4,HITIR-S-C-5,ICTIR-S-C-1,ICTIR-S-C-2,ICTIR-S-C-3,ICTIR-S-C-4,ICTIR-S-C-5,
ISCAS-S-C-1,ISCAS-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
HIT2jointNLPLab-S-C-1 with HITIR-S-C-1,HITIR-S-C-2,HITIR-S-C-5,ICTIR-S-C-1,ICTIR-S-C-2,III_CYUT_NTHU-S-C-1,ISCAS-S-C-1,ISCAS-S-C-3,NTU-S-C-1,THU-S-C-1,THU-S-C-2,
THU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
HIT2jointNLPLab-S-C-2 with HITIR-S-C-2,ICTIR-S-C-2,III_CYUT_NTHU-S-C-1,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,
uogTr-S-C-3,uogTr-S-C-4,UWat-S-C-1
HITIR-S-C-1 with HITIR-S-C-2,III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,
uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-2 with HITIR-S-C-3,HITIR-S-C-4,HITIR-S-C-5,ICTIR-S-C-1,ICTIR-S-C-2,ICTIR-S-C-3,ICTIR-S-C-4,ICTIR-S-C-5,III_CYUT_NTHU-S-C-1,ISCAS-S-C-1,ISCAS-S-C-2,
ISCAS-S-C-3,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-1,
THU-S-C-2,THU-S-C-3,THU-S-C-4,THU-S-C-5,TUTA1-S-C-1,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-3 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-4 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
HITIR-S-C-5 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-1 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-2 with ICTIR-S-C-5,III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,
TUTA1-S-C-1,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-3 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-4 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,
uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ICTIR-S-C-5 with III_CYUT_NTHU-S-C-1,ISCAS-S-C-1,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,
UWat-S-C-1
III_CYUT_NTHU-S-C-1 with ISCAS-S-C-1,ISCAS-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,
uogTr-S-C-2,uogTr-S-C-3
ISCAS-S-C-1 with ISCAS-S-C-2,ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,
uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ISCAS-S-C-2 with ISCAS-S-C-3,MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
ISCAS-S-C-3 with ISCAS-S-C-4,KLE-S-C-1,KLE-S-C-2,KLE-S-C-3,NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,
uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
ISCAS-S-C-4 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
KLE-S-C-1 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
KLE-S-C-2 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
KLE-S-C-3 with MSINT-S-C-1,MSINT-S-C-2,MSINT-S-C-3,MSINT-S-C-4,MSINT-S-C-5,THU-S-C-1,THU-S-C-2,THU-S-C-3,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
MSINT-S-C-1 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-2 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-3 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-4 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
MSINT-S-C-5 with NTU-S-C-1,NTU-S-C-2,NTU-S-C-3,THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
NTU-S-C-1 with THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3
NTU-S-C-2 with THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
NTU-S-C-3 with THU-S-C-1,THU-S-C-2,THU-S-C-3,TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
THU-S-C-1 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
THU-S-C-2 with THU-S-C-4,THU-S-C-5,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
THU-S-C-3 with THU-S-C-4,THU-S-C-5,TUTA1-S-C-1,uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
THU-S-C-4 with TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
THU-S-C-5 with TUTA1-S-C-1,uogTr-S-C-2,uogTr-S-C-3
TUTA1-S-C-1 with uogTr-S-C-1,uogTr-S-C-2,uogTr-S-C-3,uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-1,UWat-S-C-2
uogTr-S-C-1 with uogTr-S-C-2,uogTr-S-C-3,UWat-S-C-1
uogTr-S-C-2 with uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-2
uogTr-S-C-3 with uogTr-S-C-4,uogTr-S-C-5,UWat-S-C-2
uogTr-S-C-4 with UWat-S-C-1
uogTr-S-C-5 with UWat-S-C-1
UWat-S-C-1 with UWat-S-C-2

Figure 13. Significantly different Chinese Subtopic Mining run pairs in terms of D♯-nDCG@30 (randomised Tukey's HSD at $\alpha = 0.05$).

NTU-S-J-1 with NTU-S-J-2,NTU-S-J-3,NTU-S-J-4,ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1,uogTr-S-J-2,WSE-S-J-1,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
NTU-S-J-2 with RMIT-S-J-1,uogTr-S-J-1
NTU-S-J-3 with RMIT-S-J-1,uogTr-S-J-1
NTU-S-J-4 with RMIT-S-J-1,uogTr-S-J-1
ORG-S-J-1 with RMIT-S-J-1,uogTr-S-J-1
ORG-S-J-2 with RMIT-S-J-1,uogTr-S-J-1
RMIT-S-J-1 with uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
uogTr-S-J-1 with uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5

Figure 14. Significantly different Japanese Subtopic Mining run pairs in terms of D♯-nDCG@10, Revised (randomised Tukey's HSD at $\alpha = 0.05$).

NTU-S-J-1 with NTU-S-J-2,NTU-S-J-3,NTU-S-J-4,ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1,uogTr-S-J-2,WSE-S-J-1,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
NTU-S-J-2 with ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1
NTU-S-J-3 with ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1
NTU-S-J-4 with ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1
ORG-S-J-1 with RMIT-S-J-1,uogTr-S-J-1,uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
ORG-S-J-2 with RMIT-S-J-1,uogTr-S-J-1,uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
RMIT-S-J-1 with uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
uogTr-S-J-1 with uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5

Figure 15. Significantly different Japanese Subtopic Mining run pairs in terms of D♯-nDCG@20, Revised (randomised Tukey's HSD at $\alpha = 0.05$).

```
NTU-S-J-1 with NTU-S-J-2,NTU-S-J-3,NTU-S-J-4,ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1,uogTr-S-J-2,WSE-S-J-1,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
NTU-S-J-2 with ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1
NTU-S-J-3 with ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1
NTU-S-J-4 with ORG-S-J-1,ORG-S-J-2,RMIT-S-J-1,uogTr-S-J-1
ORG-S-J-1 with RMIT-S-J-1,uogTr-S-J-1,uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
ORG-S-J-2 with RMIT-S-J-1,uogTr-S-J-1,uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
RMIT-S-J-1 with uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
uogTr-S-J-1 with uogTr-S-J-2,WSE-S-J-1,WSE-S-J-2,WSE-S-J-3,WSE-S-J-4,WSE-S-J-5
```

Figure 16. Significantly different Japanese Subtopic Mining run pairs in terms of D♯-nDCG@30, Revised (randomised Tukey's HSD at $\alpha = 0.05$).

```
HIT2jointNLPLab-D-C-1 with III_CYUT_NTHU-D-C-1,MSINT-D-C-1,MSINT-D-C-2,MSINT-D-C-4,MSINT-D-C-5,THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,
 uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
HIT2jointNLPLab-D-C-2 with III_CYUT_NTHU-D-C-1,MSINT-D-C-1,MSINT-D-C-4,THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-2,uogTr-D-C-4,uogTr-D-C-5
III_CYUT_NTHU-D-C-1 with MSINT-D-C-1,MSINT-D-C-2,MSINT-D-C-3,MSINT-D-C-4,MSINT-D-C-5,NTU-D-C-1,SJTUBCMI-D-C-1,SJTUBCMI-D-C-2,
 SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-3,THUIR-D-C-4,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,
 uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
MSINT-D-C-1 with MSINT-D-C-3,NTU-D-C-1,SJTUBCMI-D-C-1,SJTUBCMI-D-C-2,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,
 THUIR-D-C-4
MSINT-D-C-2 with MSINT-D-C-3,SJTUBCMI-D-C-1,SJTUBCMI-D-C-2,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
MSINT-D-C-3 with MSINT-D-C-4,MSINT-D-C-5,THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
MSINT-D-C-4 with NTU-D-C-1,SJTUBCMI-D-C-1,SJTUBCMI-D-C-2,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
MSINT-D-C-5 with SJTUBCMI-D-C-1,SJTUBCMI-D-C-2,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
NTU-D-C-1 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-2,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-1 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-2 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-3 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-4 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-5 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
THUIR-D-C-1 with THUIR-D-C-3,THUIR-D-C-4
THUIR-D-C-2 with THUIR-D-C-3,THUIR-D-C-4
THUIR-D-C-3 with THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
THUIR-D-C-4 with THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
```

Figure 17. Significantly different Chinese Document Ranking run pairs in terms of D♯-nDCG@10 (randomised Tukey's HSD at $\alpha = 0.05$).

```
HIT2jointNLPLab-D-C-1 with HIT2jointNLPLab-D-C-2,III_CYUT_NTHU-D-C-1,MSINT-D-C-1,MSINT-D-C-2,MSINT-D-C-4,MSINT-D-C-5,SJTUBCMI-D-C-2,
 THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
HIT2jointNLPLab-D-C-2 with III_CYUT_NTHU-D-C-1,THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
III_CYUT_NTHU-D-C-1 with MSINT-D-C-1,MSINT-D-C-2,MSINT-D-C-3,MSINT-D-C-4,MSINT-D-C-5,NTU-D-C-1,SJTUBCMI-D-C-1,SJTUBCMI-D-C-2,
 SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-3,THUIR-D-C-4,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,
 uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
MSINT-D-C-1 with SJTUBCMI-D-C-1,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
MSINT-D-C-2 with SJTUBCMI-D-C-1,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
MSINT-D-C-3 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
MSINT-D-C-4 with SJTUBCMI-D-C-1,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
MSINT-D-C-5 with SJTUBCMI-D-C-1,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
NTU-D-C-1 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-1 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-2 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-3 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-4 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-5 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
THUIR-D-C-1 with THUIR-D-C-3,THUIR-D-C-4
THUIR-D-C-2 with THUIR-D-C-3,THUIR-D-C-4
THUIR-D-C-3 with THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
THUIR-D-C-4 with THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
```

Figure 18. Significantly different Chinese Document Ranking run pairs in terms of D♯-nDCG@20 (randomised Tukey's HSD at $\alpha = 0.05$).

```
HIT2jointNLPLab-D-C-1 with HIT2jointNLPLab-D-C-2,III_CYUT_NTHU-D-C-1,MSINT-D-C-1,MSINT-D-C-2,MSINT-D-C-3,MSINT-D-C-4,MSINT-D-C-5,
 NTU-D-C-1,SJTUBCMI-D-C-2,THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
HIT2jointNLPLab-D-C-2 with III_CYUT_NTHU-D-C-1,THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
III_CYUT_NTHU-D-C-1 with MSINT-D-C-1,MSINT-D-C-2,MSINT-D-C-3,MSINT-D-C-4,MSINT-D-C-5,NTU-D-C-1,SJTUBCMI-D-C-1,SJTUBCMI-D-C-2,
 SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-3,THUIR-D-C-4,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,
 uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
MSINT-D-C-1 with NTU-D-C-1,SJTUBCMI-D-C-1,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
MSINT-D-C-2 with SJTUBCMI-D-C-1,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
MSINT-D-C-3 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
MSINT-D-C-4 with SJTUBCMI-D-C-1,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
MSINT-D-C-5 with SJTUBCMI-D-C-1,SJTUBCMI-D-C-3,SJTUBCMI-D-C-4,SJTUBCMI-D-C-5,THUIR-D-C-3,THUIR-D-C-4
NTU-D-C-1 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-1 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-2 with THUIR-D-C-1,THUIR-D-C-5,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-3 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-4 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
SJTUBCMI-D-C-5 with THUIR-D-C-1,THUIR-D-C-2,THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
THUIR-D-C-1 with THUIR-D-C-3,THUIR-D-C-4
THUIR-D-C-2 with THUIR-D-C-3,THUIR-D-C-4
THUIR-D-C-3 with THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
THUIR-D-C-4 with THUIR-D-C-5,uogTr-D-C-1,uogTr-D-C-2,uogTr-D-C-3,uogTr-D-C-4,uogTr-D-C-5
```

Figure 19. Significantly different Chinese Document Ranking run pairs in terms of D♯-nDCG@30 (randomised Tukey's HSD at $\alpha = 0.05$).

```
MSINT-D-J-1 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
MSINT-D-J-2 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
MSINT-D-J-3 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5,uogTr-D-J-3,uogTr-D-J-4,uogTr-D-J-5
MSINT-D-J-4 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
MSINT-D-J-5 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-4,RMIT-D-J-5
RMIT-D-J-1 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-2 with uogTr-D-J-1,uogTr-D-J-2,uogTr-D-J-3
RMIT-D-J-3 with RMIT-D-J-5
RMIT-D-J-4 with uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-5 with uogTr-D-J-1,uogTr-D-J-2,uogTr-D-J-3,uogTr-D-J-4,uogTr-D-J-5
```

Figure 20. Significantly different Japanese Document Ranking run pairs in terms of D♯-nDCG@10 (randomised Tukey's HSD at $\alpha = 0.05$).

```
MSINT-D-J-1 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
MSINT-D-J-2 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
MSINT-D-J-3 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5,uogTr-D-J-3,uogTr-D-J-4,uogTr-D-J-5
MSINT-D-J-4 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
MSINT-D-J-5 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
RMIT-D-J-1 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-2 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-3 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-4 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-5 with uogTr-D-J-1,uogTr-D-J-2,uogTr-D-J-3,uogTr-D-J-4,uogTr-D-J-5
```

Figure 21. Significantly different Japanese Document Ranking run pairs in terms of D♯-nDCG@20 (randomised Tukey's HSD at $\alpha = 0.05$).

```
MSINT-D-J-1 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5,uogTr-D-J-3,uogTr-D-J-4,uogTr-D-J-5
MSINT-D-J-2 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
MSINT-D-J-3 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5,uogTr-D-J-3,uogTr-D-J-4,uogTr-D-J-5
MSINT-D-J-4 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
MSINT-D-J-5 with RMIT-D-J-1,RMIT-D-J-2,RMIT-D-J-3,RMIT-D-J-4,RMIT-D-J-5
RMIT-D-J-1 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2,uogTr-D-J-4,uogTr-D-J-5
RMIT-D-J-2 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-3 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-4 with RMIT-D-J-5,uogTr-D-J-1,uogTr-D-J-2
RMIT-D-J-5 with uogTr-D-J-1,uogTr-D-J-2,uogTr-D-J-3,uogTr-D-J-4,uogTr-D-J-5
```

Figure 22. Significantly different Japanese Document Ranking run pairs in terms of D♯-nDCG@30 (randomised Tukey's HSD at $\alpha = 0.05$).