Overview of the NTCIR-11 IMine Task

Yiqun Liu[1], Ruihua Song[2], Min Zhang[1], Zhicheng Dou[2], Takehiro Yamamoto[3], Makoto Kato[3], Hiroaki Ohshima[3], Ke Zhou[4]
[1]Tsinghua University, [2]Microsoft Research Asia, [3]Kyoto University, [4]University of Edinburgh
{yiqunliu, z-m}@tsinghua.edu.cn; {song.ruihua, zhichdou@microsoft.com}; {tyamamot, kato, ohshima@dl.kuis.kyoto-u.ac.jp}; zhouke.nlp@gmail.com

ABSTRACT
In this paper, we provide an overview of the NTCIR IMine task, which is a core task of NTCIR-11 and also a succeeding work of INTENT@NTCIR-9 and INTENT2@NTCIR-10 tasks. IMine is composed of a subtopic mining (SM) task, a document ranking (DR) task and a TaskMine (TM) pilot task. 21 groups from Canada, China, Germany, France, Japan, Korea, Spain, UK and United States registered to the task, which makes it one of the largest tasks in NTCIR-11. Finally, we receive 45 runs from 10 teams to the SM task, 25 runs from 6 groups to the DR task and 3 runs from 2 groups to the TM task. We describe the task details, annotation of results, evaluation strategies and then the official evaluation results for each subtask.

Keywords
Intent, ambiguity, diversity, evaluation, test collection.

1. INTRODUCTION
Many queries are short and vague in practical Web search environment. By submitting one query, users may have different intents. For an ambiguous query, users may seek for different interpretations. For a query on a broad topic, users may be interested in different subtopics. Today mining users' underlying intents of a query is an interesting topic for both IR communities and commercial search engines. IMine task from NTCIR-11 [24] aims to provide common data sets and evaluation methodology to researchers who want to investigate into the techniques for better understanding user intents behind ambiguous or broad queries. IMine is short for search Intent Mining and it also pronounces like "暧昧" which means "ambiguous" in Chinese and Japanese.

Through IMine task, we expect participants to advance the state-of-the-art techniques explored in INTENT [1] and INTENT2 [2] and to gain further insight into the right balance between relevance and diversity. We involve more user behavior data both for participants and in the annotation process to help assessors for subtopic clustering and importance estimation. We are also interested in comparing the differences between diversified search annotations from a small number of professional assessors and a relatively large number of untrained users as crowd sourcing efforts.

Similar with INTENT tasks, the IMine task consists of two subtasks: Subtopic Mining and Document Ranking. While the SM task may be regarded as a pre-DR task for identifying explicit intents, it can also be useful for other practical tasks such as query suggestion and auto-completion. We also setup a pilot subtask named TaskMine which focus on exploiting the techniques of understanding the relationship among tasks for supporting the Web searchers. We involve dealing with three different languages including English, Chinese and Japanese in IMine task. Query topics for all three languages were developed for SM task while only English and Chinese DR tasks are required since few participants show interests in Japanese DR. The major differences between IMine and previous INTENT2 tasks are shown in Table 1.

Table 1. Differences between IMine and INTENT2 tasks

	INTENT2	IMINE
Number of topics	Chinese: 100 Japanese: 100 English: 50	Chinese: 50 Japanese: 50 English: 50
DR task corpus	Chinese: SogouT Japanese: ClueWeb JA	Chinese: SogouT English: ClueWeb12-B13
Crowd sourcing	No	Crowd sourcing or Chinese DR
Subtopic organization	One level	Two level: no more than 5 first-level subtopics with at most 10 second-level subtopics each
Subtopic candidate	Query suggestions from Bing, Google, Sogou and Baidu	Query suggestions from Bing, Google, Sogou, Yahoo! and Baidu; Query facets generated by [3] from search engine results; Query facets generated by [4] from Sogou log data
Behavior data	SogouQ (data collected in 2008): appr. 2GB	SogouQ (data collected in 2008 and 2011): appr. 4GB

From Table 1 we can see that there are two major differences between IMine and previous INTENT tasks. The first difference lies that IMine requires participants to submit a two-level hierarchy of sub-intents for the query topics. In previous diversified search related studies, we notice the phenomena that some query subtopics belong to the concept of others (e.g. *IPhone* and *apple inc. products* are both regarded as subtopics for the query *apple*, while *IPhone* should be covered by *apple inc. products*). This may lead to difficulty in subtopic importance estimation and diversified ranking. Therefore, we introduce a two-level hierarchy of subtopics to better present the diversified intent structure of ambiguous/broad queries. This require extra efforts in assessment and a different design of evaluation metrics, which we will address in follow up sections.

The second major difference between IMine and previous tasks is that we try to incorporate more user behavior data and introduce the evaluation framework based on crowd sourcing. Recently, several metrics have been proposed to evaluate a diversified search result with different types of user behavior assumptions, considering relevance, diversity, novelty, user intent, and so on. To validate the credibility of these evaluation metrics, a number of methods that "evaluate evaluation metrics" are also adopted in diversified search evaluation studies, such as Kendall's tau [5], Discriminative Power [6], and the Intuitiveness Test [7]. These methods have been widely adopted and have aided us in gaining much insight into the effectiveness of evaluation metrics. However, they also follow certain types of user behaviors or statistical assumptions and do not take the information of users' actual search preferences into consideration. In IMine task, we want to take user preferences collected with crowd sourcing efforts as the ground truth to investigate into both the performance of participants' runs and diversified evaluation metrics.

21 groups from Canada, China, Germany, France, Japan, Korea, Spain, UK and United States registered to the IMine task, which makes it one of the largest tasks in NTCIR-11. Finally, we receive we receive 45 runs from 10 teams to the SM task, 25 runs from 6 groups to the DR task and 3 runs from 2 groups to the TM task. Names and organizations of the participants which submitted results are shown in Table 3 and Table 4.

Table 3. Organization of the participating groups in IMine

Group Name	Organization
UDEL	University of Delaware, United States
SEM13	Toyohashi University of Technology, Japan
HULTECH	University of Caen, France
THU-SAM	Joint team of Tsinghua University, China and Samsung Electronics, Korea
FRDC	Fujitsu Research & Development Center Co., LTD., China
TUTA1	The University of Tokushima, Japan
CNU	Capital Normal University, China
KUIDL	Kyoto University, Japan
UM13	University of Montreal, Canada
KLE	POSTECH, Korea
uhyg	University of Hyogo, Japan
Interactive MediaMINE	Kogakuin University, Japan

Table 4. Result submission from groups in IMine

Group	Chinese SM	Japanese SM	English SM	Chinese DR	English DR	TM
UDEL			1		5	
SEM13			5		5	
HULTECH			4			
THU-SAM	5		2	4		
FRDC	5			5		
TUTA1	1		1	1	2	
CNU	4					
KUIDL		1	1			
UM13			3		3	
KLE	4	4	4			
uhyg						2
Interactive MediaMINE						1
#Group	**5**	**2**	**8**	**3**	**4**	**2**
#Run	**19**	**5**	**21**	**10**	**15**	**3**

The remainder of the paper is organized as follows: Section 2 describes the details of the three subtasks, including the query set, supporting data resources and the test corpus adopted. The evaluation metrics and result assessment process are introduced in Section 3. Official evaluation results based on cranfield methodology are presented in Section 4. User preference test results are reported and compared with cranfield-like approaches in Section 5. Section 6 concludes this paper and the Appendix contains the details of each run as well as significance test results.

2. TASKS AND DATASETS

2.1 Query set

The same query topics are adopted in both Subtopic Mining and Document Ranking subtasks for all languages. These topics are sampled from the median-frequency queries collected from both Sogou and Bing search logs. We avoid top or tail queries because search performance of top queries are already quite high for most commercial search engines while many tail queries may contain typos, language mistakes or even illegal contents. Approximately equal amounts of ambiguous, broad and clear queries are included in the query topic set. Several topics are shared among different languages for possible future cross-language research purposes. Detailed information of the constructed query set is shown in Table 5. For SM task, queries with clear intents are not evaluated because they are not expected to contain subtopics.

Table 5. Statistics of the IMine query topic set

Language	#topic			#shared topics
	Ambiguous	Broad	Clear	
English	16	17	17	14 shared by all languages, 8 shared by English and Chinese
Chinese	16	17	17	
Japanese	17	17	16	

We follow the query intent classification framework proposed in [8] and group the queries into three groups: Ambiguous, Broad and Clear. Both ambiguous and broad queries are adopted in the SM task for query intent analysis while all queries are evaluated in the DR task (for clear queries, we just evaluate the ad-hoc retrieval performance instead of diversified search performance).

Table 6. IMine query topic set (for Intent, a: ambiguous, b: broad, c: clear)

ID	Topic	Intent	Shared
0001	先知	a	CEJ
0002	波斯猫	a	CE
0003	猫头鹰	a	CEJ
0004	Adobe	a	CEJ
0005	传奇	a	CEJ
0006	小米	a	
0007	中国水电	a	
0008	云轩	a	
0009	遮天	a	
0010	舍得	a	
0011	秋菊	a	
0012	三字经	a	

0013	三毛	a	
0014	阳光	a	
0015	嫦娥	a	
0016	程序员	a	
0017	泰国特产	b	CE
0018	科学美国人	b	CEJ
0019	黄金	b	CE
0020	浴缸	b	CEJ
0021	婚戒	b	CEJ
0022	三星	b	CEJ
0023	饥饿游戏	b	CEJ
0024	心理测试	b	
0025	椰岛造型	b	
0026	野葛根	b	
0027	秧歌	b	
0028	卫子夫	b	
0029	佛教音乐	b	
0030	浏览器下载	b	
0031	相亲节目有哪些	b	
0032	哈利波特	b	
0033	安卓 2.3 游戏下载	b	
0034	男鞋尺码对照表	c	CEJ
0035	奥巴马简历	c	CE
0036	肥胖的原因	c	CEJ
0037	什么是自然数	c	CEJ
0038	牙齿黄怎么办	c	CE
0039	治疗近视的方法	c	CE
0040	央金兰泽的歌曲	c	
0041	声卡是什么	c	
0042	乘法口诀	c	
0043	学雷锋作文	c	
0044	联通网上营业厅	c	
0045	怎么查 ip 地址	c	CEJ
0046	邮编号码查询	c	CEJ
0047	在线冲印照片	c	CE
0048	qq 加速器下载	c	
0049	冬季恋歌国语全集	c	
0050	初恋这件小事	c	
0051	apple	a	
0052	cathedral	a	
0053	eclipse	a	
0054	fas	a	
0055	flesh	a	
0056	ir	a	
0057	lost	a	
0058	shrew	a	
0059	symmetry	a	

0060	the presidents of the united states of america	a	
0061	windows	a	
0062	prophet	a	CEJ
0063	gold	a	CE
0064	owl	a	CEJ
0065	adobe	a	CEJ
0066	legend	a	CEJ
0067	beijing subways	b	
0068	camera	b	
0069	free dvd burner	b	
0070	lost season 5	b	
0071	mobile phones	b	
0072	programming languages	b	
0073	tom cruise	b	
0074	top ipad games	b	
0075	watches	b	
0076	thai specialties	b	CE
0077	scientific american	b	CEJ
0078	persian cat	b	CE
0079	bathtub	b	CEJ
0080	wedding ring	b	CEJ
0081	samsung	b	CEJ
0082	the hunger games	b	CEJ
0083	harry potter	b	CE
0084	21 weeks pregnant	c	
0085	7zip	c	
0086	appendix pain symptoms	c	
0087	brad paisley lyrics	c	
0088	craig's list phoenix	c	
0089	mcdonalds nutrition guide	c	
0090	sausalito art festival	c	
0091	tennessee unemployment	c	
0092	men's shoe sizes conversion	c	CEJ
0093	obama biography	c	CE
0094	causes of obesity	c	CEJ
0095	what is a natural number	c	CEJ
0096	yellow teeth treatment	c	CE
0097	myopia treatment	c	CE
0098	how to find my ip address	c	CEJ
0099	postcode finder	c	CEJ
0100	online photo printing	c	CE
0101	シド	a	
0102	ダム	a	
0103	R	a	
0104	ハヤブサ	a	
0105	ナポレオン	a	
0106	アバター	a	
0107	ジップ	a	
0108	ウォッカ	a	
0109	横浜	a	

0110	伝奇	a	CEJ
0111	アドビ	a	CEJ
0112	予言者	a	CEJ
0113	オウル	a	CEJ
0114	赤とうがらし	a	
0115	銀シャリ	a	
0116	嵐	a	
0117	フランクフルト	a	
0118	東方神起	b	
0119	円形脱毛症	b	
0120	柿の葉すし	b	
0121	シャネル	b	
0122	女子バレー	b	
0123	TPP	b	
0124	ドラえもん	b	
0125	ビートルズ	b	
0126	ボーカロイド	b	
0127	年賀状	b	
0128	うつ病	b	
0129	サムスン	b	CEJ
0130	**タイ 特産**	b	CEJ
0131	浴槽	b	CEJ
0132	ハンガーゲーム	b	CEJ
0133	結婚指輪	b	CEJ
0134	サイエンティフィック・アメリカン	b	CEJ
0135	櫻井歯科診療所 ホームページ	c	
0136	京葉タクシー 電話番号	c	
0137	湘南新宿ライン 路線図	c	
0138	旭山動物園 アクセス	c	
0139	秋田中央交通 時刻表	c	
0140	羽田空港 リムジンバス 時刻表	c	
0141	水平投射運動 速度の求め方	c	
0142	のし袋 書き方	c	
0143	facebook 退会方法	c	
0144	タロットカード 吊るされた男 意味	c	
0145	少々お待ちください 英語	c	
0146	肥満の原因	c	CEJ
0147	自然数とは	c	CEJ
0148	IP アドレスを確認するには	c	CEJ
0149	メンズ靴サイズ対応表	c	CEJ
0150	郵便番号検索	c	CEJ

2.2 Subtopic Mining Subtask

In the Subtopic Mining task, a subtopic could be an interpretation of an ambiguous query or an aspect of a broad query. Participants are expected to generate a two-level hierarchy of underlying subtopics by analysis into the provided document collection, user behavior data set or other kinds of external data sources. A list of query suggestions/completions collected from popular commercial search engines as well as some queries mined from search logs/SERPs (see Table 1) are provided as possible subtopic candidates while participants can also use other information sources (e.g. Wikipedia, search behavior logs) to generate their own candidates. For both Subtopic Mining and Document Ranking subtasks, SogouQ search user behavior data collection is available for participants as additional resources. The collection contains queries and click-through data collected and sampled by China's second largest search engine Sogou.com in 2008 and 2012, separately.

As for the two-level hierarchy of subtopics, we can take the ambiguous query "windows" as an example. The first-level subtopic may be Microsoft Windows, software in windows platform or house windows. In the category of Microsoft Windows, users may be interested in different aspects (second-level subtopics), such as "Windows 8", "Windows update", etc.

From our experiences in past INTENT/INTENT2 tasks, we found that the relationship among subtopics for some queries are not trivial. For example, for topic #0205 in INTENT2 task (功夫/kung fu), the subtopics 功夫【电影《功夫》】 (kung fu the movie), 功夫【影片下载】(movie download), 功夫【在线观看】(movie online), 功夫【影视作品】(other movies related with kungfu) should be grouped into a same category "movies related with kungfu" instead of several different subtopics. We believe that organizing such a hierarchical structure of subtopics will help search engines to present a better ranking of results.

The hierarchical structure of subtopics is close related with knowledge graph which has been well studied in Web search researches recently. Some participants in INTENT/INTENT2 tasks also adopted existing knowledge graphs such as wikipedia, freebase (e.g. THCIB and THUIS in INTENT2) in developing subtopic candidate sets. However, we believe that the hierarchical subtopics for a certain query is used to describe users' possible information needs behind this query instead of the knowledge structure of the entity named this query. Therefore, even when a knowledge graph exists for a given query (which is not usually the case since Web queries are so complicated), we should not use the graph directly as the hierarchy of query intents.

In this year's IMine task, at most FIVE first-level subtopics with no more than TEN second-level subtopics each should be returned for each query topic. There is no need to return subtopics for clear queries but participants will not be penalized for doing this in the evaluation. The first-level subtopics for broad queries will not be taken into consideration in the evaluation process because there may be various standards for organizing high-level aspects for these queries. Besides the hierarchy of subtopics, a ranking list of all first-level subtopics and a separate ranking list of all second-level subtopics should also be returned for each ambiguous/broad query. It means that the submitted second-level subtopics should be globally ranked across different first-level subtopics within a same query topic. With these ranking lists, the importance estimation results could be evaluated and compared among different participant runs.

2.3 Document Ranking Subtask

In document ranking task, Participants are asked to return a diversified ranked list of no more than 100 results for each query. Participants are encouraged to selectively use diversification algorithms in ranking because diversification is not necessary for all queries (e.g. for clear queries). Based on the subtopic mining results, participants are supposed to select important first-level/second-level subtopics and mix

Ithem to form a diversified ranking list. The goals of diversification are (a) to retrieve documents that cover as many intents as possible; and (b) to rank documents that are highly relevant to more popular intents higher than those that are marginally relevant to less popular intents.

SogouT (http://www.sogou.com/labs/dl/t-e.html) is adopted as the document collection for Chinese topics in Document Ranking subtask. The collection contains about 130M Chinese pages together with the corresponding link graph. The size is roughly 5TB uncompressed. The data was crawled and released on Nov 2008. In order to help participants who are not able to construct their own retrieval platforms, the organizers provide a non-diversified baseline Chinese DR run based on THU-SAM's retrieval system.

As for English Document Ranking subtask, the ClueWeb12-B13 (http://lemurproject.org/clueweb12/) data set is adopted, which includes 52M English Web pages crawled in 2012. A search interface is provided by Lemur project so that the retrieval baseline could be obtained without having to construct one's own search index.

2.4 TaskMine Subtask

TaskMine subtask is a new subtask that starts from this NTCIR-11. The goal of TaskMine is to understand the relationship among tasks for supporting the Web searchers. Particularly, this year's TaskMine aims to explore the methods of automatically finding subtasks of a given task. In this subtask, participants are required to automatically find the subtasks of a given task. More specifically, given a task (i.e., query), participants are expected to return a ranked list of not more than 50 task strings ordered by their importance. For example, for a given task "lose weight," the possible outputs can be "do physical exercise," "take calories intake," "take diet pills" and so on.

We accept a tab-delaminated-values (TSV) file as an output of TaskMine subtask run, where each line must represents a single task, and be of the following format:

qid *tid* *task_string* *score* *source*

qid *tid* *task_string* *score* *source*

where *qid* and *tid* represent a query ID and a task ID, respectively, t*ask_string* is the text that represent the task, *score* is the importance of the task, and *source* is a URL in which a participant generated the task. Participants are allowed to use any Web resources such as Search engine result pages, Web pages, Wikipedia and Community Q&A corpus so that we can explore effective approaches to mining tasks from the Web.

As for queries, this year's TaskMine subtask focused on Japanese queries and textual outputs. Table 7 shows the query set of the TaskMine subtask. We distributed 50 queries, which are composed of four query categories: *Health*, *Education*, *Daily Life* and *Sequential*. First three categories are known as typical domains in which searchers' information needs are complex and cannot be satisfied with single query or search session [20][21]. Information needs behind queries in *Sequential* category require searchers to accomplish a sequence of its subtasks to achieve their goal. These 50 queries represent typical queries in relation to our subtask.

Table 7. TaskMine subtask queries (for Category, h: Health, e: Education, d: Daily Life and s: Sequential).

Query ID	Topic	Category
TM-001	視力矯正をする	h
TM-002	茶髪にする	h
TM-003	早起きする	h
TM-004	肩こりを解消する	h
TM-005	風邪を予防する	h
TM-006	しゃっくりを止める	h
TM-007	口内炎を治す	h
TM-008	痛風の痛みを和らげる	h
TM-009	ダイエットをする	h
TM-010	禁煙をする	h
TM-011	歯を白くする	h
TM-012	目の疲れをとる	h
TM-013	エコノミー症候群対策をする	h
TM-014	ストレスを解消する	h
TM-015	高血圧を予防する	h
TM-016	肌荒れを治す	h
TM-017	育毛する	h
TM-018	頭痛を治す	h
TM-019	歯周病を治療する	h
TM-020	二重にする	h
TM-021	Illustrator の使い方を習得する	e
TM-022	九九を覚える	e
TM-023	レーザーカッターを使う	e
TM-024	将棋が上手くなる	e
TM-025	Python を勉強する	e
TM-026	中国語を勉強する	e
TM-027	字が上手くなる	e
TM-028	小型船舶操縦免許を取る	e
TM-029	料理を上達させる	e
TM-030	日本史を勉強する	e
TM-031	人を笑わせる	d
TM-032	家を建てる	d
TM-033	映画を撮る	d
TM-034	子供をなだめる	d
TM-035	クレー射撃を体験する	d
TM-036	ペットを預ける	d
TM-037	ホエールウォッチングをする	d
TM-038	出産する	d
TM-039	マンションを購入する	d
TM-040	会社を辞める	d

TM-041	結婚する	s
TM-042	レストランでワインを飲む	s
TM-043	神社でお参りをする	s
TM-044	免許を取得する	s
TM-045	株式会社を設立する	s
TM-046	家具を廃棄する	s
TM-047	食パンを作る	s
TM-048	テントを組み立てる	s
TM-049	ロイヤルミルクティーを淹れる	s
TM-050	朝顔を育てる	s

3. EVALUATION METRICS

3.1 Subtopic Mining and Document Ranking Subtasks

Search result evaluations are based on the document relevance assessments with respect to certain queries. Supposing that these documents are assessed with level 0 to h where 0 means irrelevant and h means the highest relevant. Hence h=1 means a binary relevance assessment. Let N_x denote the number of relevant documents at level x ($0 < x < h$), then $N = \sum_x N_x$ means the total number of relevant documents. Let d_r denote the document at rank r in the result list and define $J(r)$=1 if d_r is relevant to a query at level x ($0 < x < h$), otherwise $J(r)$=0. We denote the cumulative number of relevant documents as $C(r) = \sum_{i=1}^{r} J(i)$.

Let $g(r)$ denote the document gain of d_r, then $cg(r) = \sum_{i=1}^{r} g(i)$ means the cumulative gain at rank r. Also, the gain and cumulative gain of the ideal ranked list are denoted as $g^*(r)$ and $cg^*(r)$ respectively. Then we can define *nDCG* at document cutoff l as:

$$nDCG@l = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \tag{1}$$

Diversified search evaluation requires document relevance assessments with respect to subtopics instead of queries, which is different from the traditional evaluations. Document gains are therefore evaluated in terms of subtopics underlying the query. Let $g_i(r)$ denote the gain of d_r with respect to subtopic i, N_i denote the total number of documents relevant to subtopic i, and $J_i(r)$ indicate whether d_r is relevant to subtopic i. Furthermore, we suppose that there are n subtopics underlying a query q and denote the probability distribution of subtopic i as $P(i|q)$, therefore $\sum_{i=1}^{n} P(i|q) = 1$.

In INTENT/INTENT2 tasks, the major evaluation metric is *D#-measures* which is proposed in [9] to more intuitively evaluate the diversity of a ranked list. The main idea is that the abandonment of the separate calculation of measures for each subtopic, which is leveraged in previous IA measures proposed in [10] and [11]. By introducing a new document gain (named *Global Gain*), the original document gains calculated in terms of each subtopic are linearly combined. The *Global Gain* is defined as follows:

$$GG(r) = \sum_{i=1}^{n} P(i|q) g_i(r) \tag{2}$$

Then document gains in the traditional measures are replaced by this *Global Gain* factor. After this replacement, these measures (referred to as *D-measures*) capture all the properties of the original measures. Furthermore, the Global Gain linearly combines the original document gain with the respective subtopic probability for each document in an overall perspective, which directly reflects the diversity. To evaluate the subtopic recall, [9] also defined the measure namely *I-rec*[1], which is the proportion of subtopics covered by documents:

$$I\text{-}rec@l = |\textstyle\bigcup_{r=1}^{l} I(r)| / n \tag{3}$$

where $I(r)$ stands for the set of subtopics which d_r is relevant to. Linearly combining the *D-measures* with *I-rec* for documents at cutoff l, [9] defined the *D#-measures* as follows:

$$D\#\text{-}measure@l = \lambda\, I\text{-}rec@l + (1\text{-}\lambda)\, D\text{-}measure@l \tag{4}$$

where λ is the tradeoff between the diversity and the subtopic recall and is set to 0.5 in [12]. The *D#-measures* are adopted in both subtopic mining and document ranking tasks in INTENT/INTENT2 tasks as the main evaluation metric.

Besides the *D#-measures*, *DIN-measures* were also adopted in INTENT2 task. According to [14], diversity evaluation should distinguish the navigational subtopic from the informational one. The reason lies that when a certain subtopic is a navigational one, the user wants to see only one particular web page; while the user is happy to see many relevant pages when the subtopic is informational. Therefore, the types of information needs behind subtopics should be taken into account and different measures should be leveraged for evaluating subtopics in different types. Based on this assumption, the reformulation of the *Global Gain* factor in *DIN-measures* is described as follows:

$$GG^{DIN}(r) = \sum_{i} P(i|q) g_i(r) + \sum_{j} isnew_j(r) P(j|q) g_j(r) \tag{5}$$

where $\{i\}$ and $\{j\}$ denote the sets of informational and navigational subtopics for query q. And $isnew_j(r)$ is an indicator that if there is no document relevant to the navigational subtopic j between ranks 1 and r-1, $isnew_j(r)$ is set to 1, otherwise $isnew_j(r)$ is set to 0. In this way, GG^{DIN} evaluates the informational and navigational subtopics in different ways. From this definition, we can find that GG^{DIN} evaluate the informational subtopic in the same way as *D#-measures*, but for the navigational subtopic j, it leverages the indicator $isnew_j(r)$ to guarantee that only the first relevant document is considered. The *DIN-measures* are then calculated by replacing the $GG(r)$ of *D#-measures* with GG^{DIN}.

In IMine task, we follow the settings in INTENT/INTENT2 and choose *D#-nDCG* as the main evaluation metric for Document Ranking subtask. However, since a hierarchy instead of a single list of subtopics are submitted for each query topic in the new Subtopic Mining task, new metrics should be designed to evaluate the performance of the submitted two-level hierarchy of subtopics.

For the IMine Subtopic Mining task, we propose to use the *H-measures* (evaluation measures of Hierachical subtopic structure) as the main evaluation metric. The definition of H-measure is as follows:

$$\begin{aligned} &H-measure \\ &= Hscore*(\alpha * Fscore + \beta * Sscore), \qquad (\alpha+\beta=1) \end{aligned} \tag{6}$$

The definitions of *Hscore*, *Fscore* and *Sscore* are as follows and they each describe one aspect of the submitted hierarchy.

Hscore measures the quality of the hierarchical structure by

[1] In [12] the authors renamed the *S-Recall* in [13] as *I-rec*.

whether the second-level subtopic is correctly assigned to the appropriate first-level subtopic.

$$Hscore = \frac{\sum_{i=1}^{N^{(1)}} accuracy(i)}{N^{(1)}} \quad (7)$$

Here $N^{(1)}$ is the number of first-level subtopics for a certain query topic in the submission (no more than 5). *Accuracy*(i) is the percentage of correctly-assigned second-level subtopics for first-level subtopic i. If first-level subtopic i is not relevant to the query topic, then *Accuracy*(i) should be 0. Irrelevant second-level subtopics should not be regarded as "correctly-assigned" ones.

Fscore measures the quality of the first-level subtopic by whether the submitted first-level subtopics are correctly ranked and whether all important first-level subtopics are found:

$$Fscore = D\#\text{−}measure(FS_1, FS_2, ..., FS_{N^{(1)}}) \quad (8)$$

Here $\{FS_i\}$ is the first-level subtopic list for a certain query topic ranked by the score contained in submission file.

Similar with *Fscore*, *Sscore* measures the quality of the second-level subtopic with the following equation:

$$Sscore = D\#\text{−}measure(SS_1, SS_2, ..., SS_{N^{(2)}}) \quad (9)$$

Here $\{SS_i\}$ is the second-level subtopic list for a certain query topic ranked by multiplying the scores of the second-level subtopic and its corresponding first-level subtopic. Notice that all second-level subtopics are globally ranked in the submitted results so that a single $\{SS_i\}$ list could be derived.

We can see that the parameters α and β are used to balance the scores of first-level and second-level subtopics. Note that the first level subtopics are not considered in the evaluation of broad queries because there may be different categories to group the second level subtopics (e.g. book/character/film or secret chamber/order of phoenix/death hollow for the query harry potter). Therefore, α is set to 0 for all broad queries. As for ambiguous queries, we choose equal values of α and β ($\alpha = \beta = 0.5$).

3.2 TaskMine Subtask

In the TaskMine subtask, runs submitted by participants include a ranked list of tasks, called *participant tasks,* for each query. Meanwhile, for each query, we have the tasks, called *gold-standard tasks*, which are extracted by the organizers. We first identify gold-standard tasks covered by each participant task from the manual assessment (See Section 4.3 for more details of the assessment process in the TaskMine subtask). From this assessment, we obtain a ranked list of gold-standard tasks for a given list of participant tasks. More precisely, let $P = (p_1, p_2, \ldots)$ be a ranked list of participant tasks, where p_i is the i-th participant task, we can generate the list of corresponding gold-standard tasks $G = (g_1, g_2, \ldots)$. Note that g_i can be empty (ε) since a participant task may not contain any gold standard tasks. We use the generated list G to compute the effectiveness of submitted runs.

To evaluate the effectiveness of runs submitted by participants, two principles are considered: (1) a run receives a higher score if it ranks more important task at higher ranks, (2) a redundant task in a list does not obtain any gain. In order to follow these principles, we adopt nDCG to a metric that penalizes redundancy. We define the global gain of gold-standard task g_i at rank i as follows:

$$\text{gain}(g_i) = \begin{cases} 0, & \exists j < i \quad g_i = g_j \\ w(g_i), & \text{otherwise} \end{cases}$$

where $w(g_i)$ is a weight of gold-standard task g_i. Note that the weight of empty task (ε) is set to 0. If the same gold-standard task g_i has appeared at the higher rank in a list, task g_i at rank i does not obtain any gain. By using this gain function, we compute nDCG@k, which is the official metric in the TaskMine subtask:

$$\text{nDCG@}k = \frac{\sum_i^k \text{gain}(g_i)/\log_2(1+i)}{\sum_j^k \text{gain}(g_j^*)/\log_2(1+j)}$$

where g_j^* is the j-th gold-standard task in an ideal ranked list, which can be constructed by sorting all the gold-standard tasks for a query by their weight.

4. RESULT ASSESSMENT

The result assessment process is completed by different groups of assessors. As for the Chinese and English SM/DR subtasks, a vendor company is hired by NII to finish the annotation. Meanwhile, assessment of the Japanese SM task is completed by volunteers recruited in Kyoto University. All the annotation tasks are completed by native speakers to guarantee quality.

4.1 Subtopic Mining Subtask

For the subtopic mining subtask, each ambiguous and broad query should be annotated by assessors to get a two-level hierarchy of subtopics. Clear queries are not considered in this subtask. The annotation process is completed in the following steps:

- Result pool construction: Result pool of the Chinese SM task contains 1,630 first-level subtopics, 6,594 second-level subtopics and 13,251 subtopic pairs (each pair is composed of a first-level subtopic and a corresponding second-level one as submitted by participating groups). Result pool of the Japanese SM task contains 539 first-level subtopics, 3,500 second-level subtopics and 5,467 subtopic pairs. Result pool of the English SM task contains 2,537 first-level subtopics, 13,993 second-level subtopics and 23,981 subtopic pairs.
- Annotation task 1 (relevance judgment): for each submitted first-level and second-level subtopic, the assessors are required to decide whether it is relevant to the query topic or not. Any irrelevant ones will be removed from the result pool and not dealt with in the following annotation tasks.
- Annotation task 2 (Subtopic relationship verification): For each second-level subtopic in a submission, the assessors are required to decide whether the submission correctly assigns its first-level subtopic.
- Annotation task 3 (first-level clustering): For all submitted hierarchy of subtopics, the assessors are required to cluster all the first-level subtopics into several clusters.
- Annotation task 4 (importance voting for first-level): For all first-level clusters, the assessors are required to vote for its importance and select the FIVE most important ones.
- Annotation task 5 (post-clustering classification): For all second-level subtopics, the assessors are required to decide which of the five most important first-level subtopic cluster it should belong to or it doesn't fit for any. The second-level subtopics that are not relevant to any first-level subtopic should be regarded as irrelevant.
- Annotation task 6 (second-level clustering): For each of the five most important first-level subtopics, the assessors are required to cluster all the second-level subtopics which belong to it into several clusters.
- Annotation task 7 (importance voting for second-level): For all second-level clusters, the assessors are required to vote for its importance and retain at most TEN ones for each

first-level cluster. The importance voting is for all second-level subtopics of the corresponding query instead of particular first-level subtopics.

With the above procedure, the two-level hierarchy of subtopics could be generated for each ambiguous/broad query topics. *Hscore* could be estimated with the results from Annotation task 2. Meanwhile *Fscore* and *Sscore* are estimated with the results generated in Annotation task 4 and Annotation task 7, separately. Note that in the calculation of *Hscore*, we do not consider whether the first-level or second-level subtopics are finally chosen as qrels or not. Instead, we want to evaluate whether the submitted hierarchy is self-consistent.

According to the assessment results for SM task, we have 116 first-level subtopics and 501 second-level subtopics for the 33 unclear queries in Chinese SM (3.51 first-level subtopics per query and 4.32 second-level subtopics per first-level subtopic on average). In English SM, we have 125 first-level subtopics and 373 second-level subtopics for the 33 unclear queries (3.79 first-level subtopics per query and 2.98 second-level subtopics per first-level subtopic on average). In Japanese SM, we have 145 first-level subtopics and 477 second-level subtopics for the 34 unclear queries (4.26 first-level subtopics per query and 3.29 second-level subtopics per first-level subtopic on average).

Although the participants are required to submit up to 10 second-level subtopics for each first-level subtopic, the assessment shows a much smaller number of second-level subtopics. We believe that the assessment is more proper because a hierarchical structure with too fine-grain subtopics will not help improve search ranking given the fact that there are only 10 ranking positions available on the first SERP.

4.2 Document Ranking Subtask

For the Document Ranking subtask, relevance judgment should be performed to result documents for all queries including clear, ambiguous and broad ones. To help assessors to finish the relevance judgment task, we extract all result documents in the pool from SogouT and ClueWeb. HTML documents are transformed into JPG version so that the appearance of documents to each assessor is the same. It can also reduce the efforts of assessors to load a Web page from its HTML version. The annotation process is completed in the following steps:

- Result pool construction: Due to limited annotation resources, we only cover a number of top results from a selection of submitted runs from participating groups. For the Chinese DR task, we choose top 20 results from runs with top priority from each group. While for English DR task, we choose The result top 10 results from runs with top priority from each group. The result pool for Chinese and English DR tasks contain 2,525 and 1,930 result documents, separately.
- Annotation task 1 (relevance judgment): for each document-query pair, the assessors are required to decide whether the document is relevant to the query with a 4-grade score (3: highly-relevant, 2: relevant, 1: irrelevant, 0: spam).
- Annotation task 2 (subtopic judgment) For a result document annotated as 2 or 3 in the first step for a broad or ambiguous query, the assessors should point out which first-level and second-level subtopic this document is relevant to. If one document isn't relevant to any of the subtopics, it shouldn't be regarded as a relevant one. For clear queries, there is no need to finish this step.

With the above procedure, we obtain the document relevance assessment result both to queries and to corresponding subtopics. For clear queries, the original NDCG score is calculated as the evaluation result. For ambiguous and broad queries, we choose corresponding first-level subtopics in the calculation of *D#-measures*. Second-level subtopics are not involved in the evaluation of DR tasks because the number of subtopics are too many (about 50) for a practical Web search scenario.

4.3 TaskMine Subtask

For the TaskMine subtask, we first prepare the gold standard tasks for each query and then match the tasks returned by the participants' runs with the gold standard tasks. The actual annotation process is completed in the following steps:

- Preparing gold standard tasks: For each query, we asked an assessor to list up all the tasks that help to achieve the given topic. The assessors are allowed to use Web search engines to learn about the topic of the query.
- Matching participants' task with gold standard task: For each task returned by the participants' run, one assessor is asked to match it with *one* of the gold standard tasks. If the assessor judged that the participants' task is not effective to achieve the given topic, he/she annotates it as an irrelevant task. If the assessor find the participants' task is effective but there is no suitable task in the gold standard tasks, he/she add the participant task to the gold standard tasks. Finally, 2,716 gold-standard tasks are extracted for 50 queries.
- Importance voting for each gold standard task: For each gold standard task of a query, two assessors are asked to vote for its importance. The assessors annotate the importance of gold standard tasks with a two grade score: (1: the task is moderately effective to achieve the query, 2: the task is highly effective to achieve the query). We take the average of results from two assessors and treat it as a weight of the gold standard task.

With the above procedure, given a ranked list of participant tasks, we can generate the corresponding list of gold-standard tasks with their importance. We compute nDCG@k for the generated list.

5. OFFICIAL EVALUATION RESULTS

We will present the evaluation results in the following two sections. At first, Cranfield-like approach is adopted based on the result assessment described in Section 4. These results should be regarded as official results because they could be compared with existing testing results such as those in INTENT/INTENT2. The test collection could also be reused by researchers who do not participate in the IMine task. After that, we will show the user preference test results for Chinese DR task. Although those results could not be reused or compared with previous Cranfiled-like evaluation results, we believe that comparison of these two results should help further our understanding in the research of diversified search evaluations.

5.1 Subtopic Mining Subtask

While reporting the evaluation results for the Subtopic Mining subtask, we will at first compare the performance of different participating groups in terms of *Hscore*, *Fscore* and *Sscore*, separately. We will also test different parameters of H-measures. After that, we will show the evaluation results with H-measures for both ambiguous and broad queries.

5.1.1 *Hscore comparison*

Comparison of *Hscore* of different participating runs are shown in Figure 1. According to the figure we can see that for the Chinese SM task, CNU performs best with *Hscore* of 0.5789. Meanwhile, best runs from THUSAM, FRDC and KLE also gain promising results. Significance test results (two-tailed t-Test with *p*-value<0.01) show that the best results of CNU, KLE, THUSAM and FRDC cannot be separated from each other. According to these participants' descriptions, clustering technique was adopted by most of these runs to group the provided candidates and word embedding as well as semantic expansion were also employed to extract subtopics.

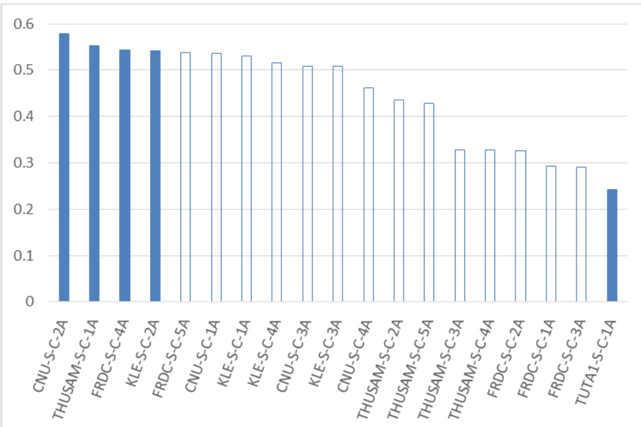

Figure 1. *Hscores* of submitted runs for unclear queries in Chinese Subtopic Mining (run with the highest performance for each participant is shown as a colored block while other runs are shown as non-colored blocks)

Figure 2 shows the *Hscore* distribution of proposed runs in English Subtopic Mining task. We can see that KUIDL and THUSAM gain best performances and their *Hscores* are much higher than those of other runs and their performance differences is not significant (two-tailed t-Test with *p*-value<0.01). According to their descriptions for submitted runs, KUIDL adopted the content from search engine result pages and THUSAM rely on Wikipedia page structures. One common feature from both runs is that first-level subtopics are always a sub-string for their corresponding second-level subtopics. The assessors tend to believe that this kind of second-level subtopics belong to the scope of first-level ones and annotate them as correct ones.

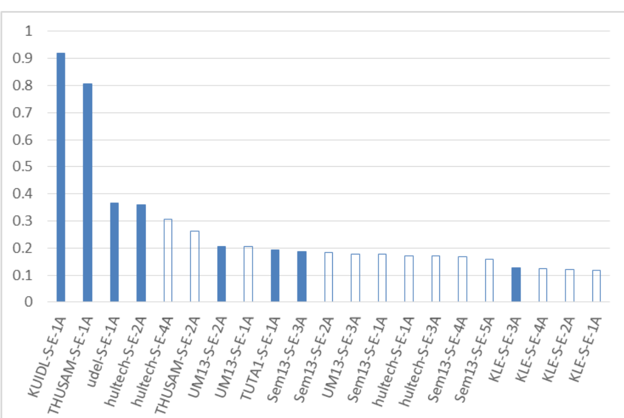

Figure 2. *Hscores* of submitted runs for unclear queries in English Subtopic Mining (run with the highest performance for each participant is shown as a colored block while other runs are shown as non-colored blocks)

This is the first year that we introduce a hierarchical structure in subtopic extraction tasks. The relationship between first-level and second-level subtopics shares similar characteristics with the relationship between entities in knowledge graphs. Meanwhile, diversified search mainly focuses on covering more popular user interests behind these topics. From the above results, we can see that the best runs from Chinese SM task focus on clustering technique while those in English prefer candidate pairs in which first-level subtopics are substrings for corresponding second-level ones. We hope to see how the introduction of user behavior data (the organizers shared some user behavior data for Chinese SM task while participants can also acquire English/Japanese query frequency data from services such as google trends) could improve these methods in the future tasks or discussions.

5.1.2 *Fscore Comparison*

Fscore evaluates whether the submitted ranking lists of first-level subtopics meet users' diversified search intents. Comparison results for the participating runs are shown in Figures 3 and 4 for Chinese and English SM tasks. Note that only ambiguous queries are evaluated in this part because there may be several different groups of first-level subtopics that are all reasonable for broad queries.

We can see that for Chinese SM task, FRDC gain highest *Fscores* with the runs FRDC-S-C-1A and FRDC-S-C-3A. Detailed analysis show that their runs gain both good I-recall (0.76 on average) and D-nDCG (0.67 on average) values. One interesting finding lies that their best performing run according to *Hscore* (FRDC-S-C-4A) fails to get high *Fscore* value while the two runs that gain best result in *Fscore* (FRDC-S-C-1A and FRDC-S-C-3A) don't got promising results in *Hscores*, either. According to the system description provided by participant, we find that FRDC adopts the same strategy in developing first-level subtopics in FRDC-S-C-1A and FRDC-S-C-3A. Query subtopics provided by organizers as well as knowledge graph entries (from Baidu Baike) are adopted as candidates, which are clustered based on corresponding SERPs collected from Google. After that, new word detection techniques are employed to generate the first level subtopics.

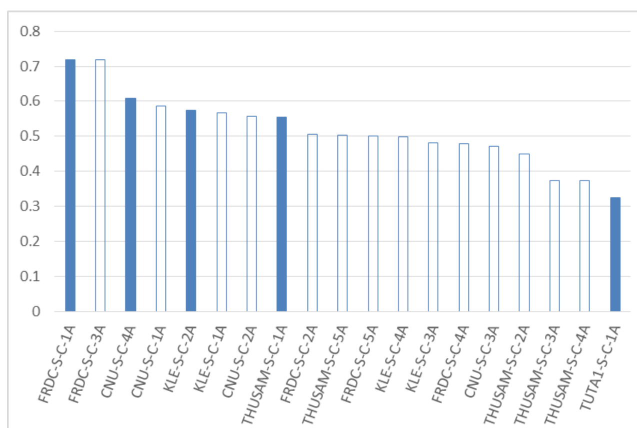

Figure 3. *Fscores* of submitted runs for ambiguous queries in Chinese Subtopic Mining (run with the highest performance for each participant is shown as a colored block while other runs are shown as non-colored blocks)

For English SM task, we can see that hultech gains best performance in *Fscores*. While the difference between the best results from hultech, KLE and KUIDL are not significant (two-tailed t-Test with p-value<0.01). According to participants' descriptions, KLE and hultech both adopt pattern matching on the provided subtopic candidates to generate first-level subtopics. Meanwhile, KUIDL employs a different strategy by extracting first-level subtopics from top SERPs for given queries.

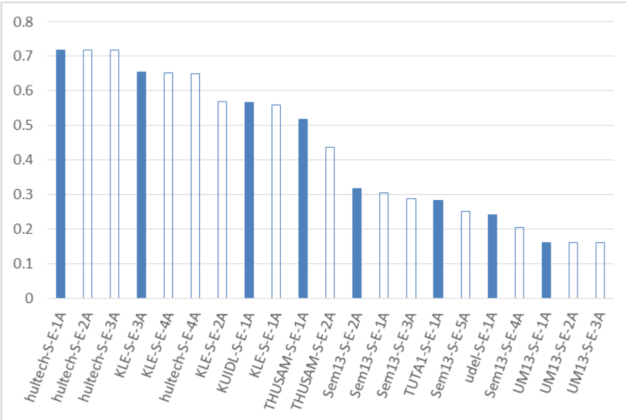

Figure 4. *Fscores* of submitted runs for ambiguous queries in English Subtopic Mining (run with the highest performance for each participant is shown as a colored block while other runs are shown as non-colored blocks)

5.1.3 Sscore Comparison

Sscore shows the fine-grained subtopic mining performance of participating runs. As stated in previous sections, at most 50 second-level subtopics are submitted in each run and they should be ranked within the whole query instead of within corresponding first-level subtopics. By this means, we could evaluate the system's performance in meeting fine-grained search intents.

According to the results shown in Figure 5, we can see that KLE obtains best *Sscore* performance in Chinese SM task. The difference between their best performing run and that from the second best group (THUSAM) is significant (two-tailed t-Test with p-value<0.01). From the descriptions provided by participants, we can see that the four runs submitted by KLE all adopt similar strategy (with different parameters). They are based on the provided subtopic candidates (query suggestion, query dimension, related queries and baseline documents) and combined with certain re-ranking techniques.

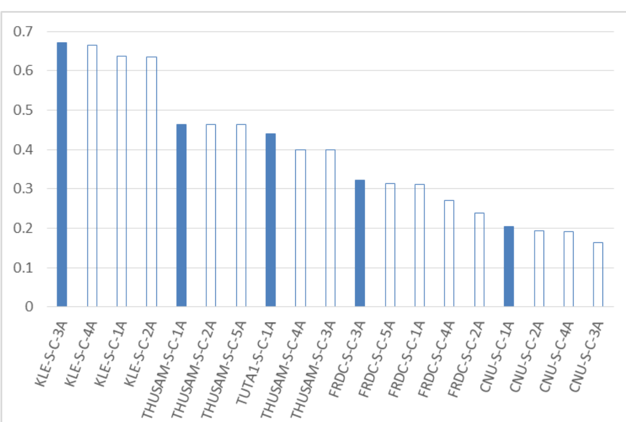

Figure 5. *Sscores* of submitted runs for unclear queries in Chinese Subtopic Mining (run with the highest performance for each participant is shown as a colored block)

KLE also gains best performance in *Sscore* according to the English SM results shown in Figure 6. The difference between their best performing result and the second best one (from KUIDL) is also significant. We can see that a similar strategy (combination of candidates from different sources) is adopted in both Chinese and English mining tasks and their submitted runs are all based on this strategy with different parameters. According to the participant's technical paper, we found that their candidate subtopics are mainly generated with a number of effective patterns that contain both the query terms and words providing specified information. We believe that this candidate selection process is effective and may be the reason of their success.

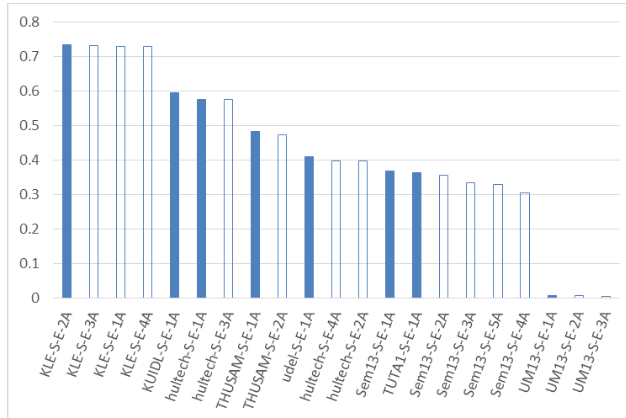

Figure 6. *Sscores* of submitted runs for unclear queries in English Subtopic Mining (run with the highest performance for each participant is shown as a colored block while other runs are shown as non-colored blocks)

5.1.4 H-Measure Comparison

With the *Hscore*, *Fscore* and *Sscore* result comparisons in previous sections, we generate the *H-measure* results according to Equation (6) in Tables 8, 9 and 10. The best performing results are also shown in Figures 7, 8 and 9 so that we can see in what way these results outperform other runs. Note that the first-level subtopics for broad queries will not be taken into consideration in the evaluation because there may be various standards for organizing high-level aspects for these queries. For example, "harry potter movies/harry potter books/harry potter games" and "harry potter and the prisoner of Azkaban/harry harry potter and the goblet of fire/harry potter and the half blood prince" may both be good categories of subtopics for the query "harry potter" (IMINE 0083), but they lead to quite different first-level subtopic evaluation results. Therefore, for the "broad" queries in Table 6, the parameter α in *H-measure* calculation is set to 0 while β is set to 1.0 in Equation (6). As stated in Section 3.1, we choose equal values of α and β ($\alpha = \beta = 0.5$) for ambiguous queries.

For Chinese SM task, KLE gain best performance with all four submitted runs. We can see that *Sscore* contributes most to their performance and they also gain nice results in *Hscores* and *Fscores*. As stated in Section 5.1.3, the four runs submitted by them all adopt similar strategy (with different parameters).

Table 8. Chinese Subtopic Mining runs ranked by *H-measure* (official result) over 33 unclear topics. The highest value in each column is shown in bold.

	Hscore	*Fscore*	*Sscore*	*H-measure*
KLE-S-C-2A	0.5413	0.5736	0.6339	**0.3360**
KLE-S-C-1A	0.5306	0.5666	0.6360	0.3303
KLE-S-C-4A	0.5148	0.4986	0.6640	0.3279
KLE-S-C-3A	0.5072	0.4817	**0.6718**	0.3255
THUSAM-S-C-1A	0.5527	0.5537	0.4634	0.2773
THUSAM-S-C-5A	0.4287	0.5040	0.4626	0.2224

THUSAM-S-C-2A	0.4347	0.4498	0.4633	0.2204
FRDC-S-C-5A	0.5377	0.5004	0.3139	0.1757
CNU-S-C-2A	**0.5789**	0.5569	0.1932	0.1748
CNU-S-C-1A	0.5353	0.5867	0.2045	0.1739
FRDC-S-C-4A	0.5436	0.4782	0.2715	0.1724
CNU-S-C-4A	0.4611	0.6073	0.1910	0.1407
THUSAM-S-C-4A	0.3284	0.3744	0.3993	0.1404
THUSAM-S-C-3A	0.3284	0.3744	0.3981	0.1400
FRDC-S-C-1A	0.2931	**0.7191**	0.3110	0.1327
FRDC-S-C-3A	0.2897	**0.7191**	0.3214	0.1326
CNU-S-C-3A	0.5086	0.4708	0.1626	0.1189
TUTA1-S-C-1A	0.2419	0.3242	0.4391	0.1126
FRDC-S-C-2A	0.3257	0.5045	0.2381	0.1032

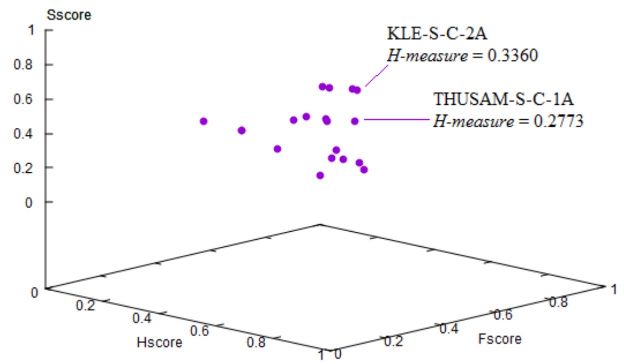

Figure 7. Best performing runs in Chinese SM task and their performance comparison

Different from the Chinese SM task, *Hscore* plays a central part in English SM result comparisons. It is possibly due to the fact that there exist large differences between runs in *Hscore* for English SM task (see Figure 2). KUIDL-S-E-1A achieves both the best *Hscore* and best *H-measure* in Table 9. From Figures 4 and 6 we can also see that KUIDL's *Fscore* and *Sscore* results are also quite nice compared with other runs. According to participant's result descriptions, they try to extract hierarchical intents from search result landing pages' structures. First-level subtopics are at first extracted from Web search results and then second-level ones are extracted by counting the co-occurrence of words in different page portions. According to the participant's paper, this method is proposed by [22] and assumes that terms appear in the title of documents are likely to represent the overall subject while terms appear in the body of documents are likely to represent the detailed topic of the subject. Therefore, they use the occurrence of a certain subtopic candidate in the title part as a sign for its being first-level subtopic and other occurrences as signs for being second-level ones. This seems an effective strategy.

Table 9. English Subtopic Mining runs ranked by *H-measure* (official result) over 33 unclear topics. The highest value in each column is shown in bold.

	Hscore	*Fscore*	*Sscore*	*H-measure*
KUIDL-S-E-1A	**0.9190**	0.5670	0.5964	**0.5509**
THUSAM-S-E-1A	0.8065	0.5179	0.4835	0.4257
hultech-S-E-2A	0.3596	**0.7184**	0.3977	0.1562
hultech-S-E-4A	0.3055	0.6496	0.3981	0.1384
udel-S-E-1A	0.3658	0.2420	0.4103	0.1180
THUSAM-S-E-2A	0.2634	0.4361	0.4732	0.1179
KLE-S-E-3A	0.1291	0.6539	0.7317	0.0980
KLE-S-E-4A	0.1260	0.6511	0.7294	0.0938
KLE-S-E-2A	0.1200	0.5698	**0.7342**	0.0893
hultech-S-E-1A	0.1703	**0.7184**	0.5754	0.0888
hultech-S-E-3A	0.1703	**0.7184**	0.5754	0.0888
KLE-S-E-1A	0.1185	0.5591	0.7298	0.0873
TUTA1-S-E-1A	0.1933	0.2833	0.3647	0.0694
Sem13-S-E-1A	0.1762	0.3043	0.3689	0.0581
Sem13-S-E-3A	0.1869	0.2882	0.3333	0.0569
Sem13-S-E-2A	0.1844	0.3174	0.3566	0.0565
Sem13-S-E-4A	0.1672	0.2056	0.3039	0.0460
Sem13-S-E-5A	0.1580	0.2511	0.3285	0.0437
UM13-S-E-2A	0.2064	0.1624	0.0059	0.0049
UM13-S-E-1A	0.2056	0.1624	0.0059	0.0047
UM13-S-E-3A	0.1766	0.1624	0.0049	0.0037

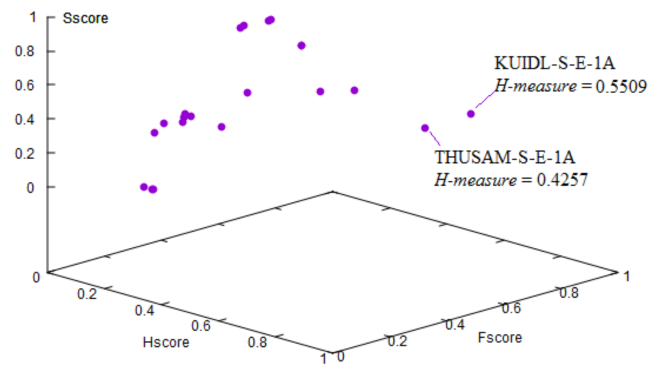

Figure 8. Best performing runs in English SM task and their performance comparison

As for Japanese SM task, since there are only two participating groups, we just compare their *H-measure* performances in this section and don't present the *Hscore*, *Fscore* and *Sscore* comparisons, separately. From the results shown in Table 10 and Figure 9 we can see that KLE gain better performance than KUIDL in *H-measure* but the difference between their best performing runs is not significant (two-tailed t-Test with *p*-value<0.01). From the descriptions in submitted runs, we find that both KLE and KUIDL adopt similar strategies in different languages. They gain best performance in Chinese SM and English SM, separately. We expect the two participating groups to compare their runs in different languages in the future (they didn't provide such analysis in their participants' papers).

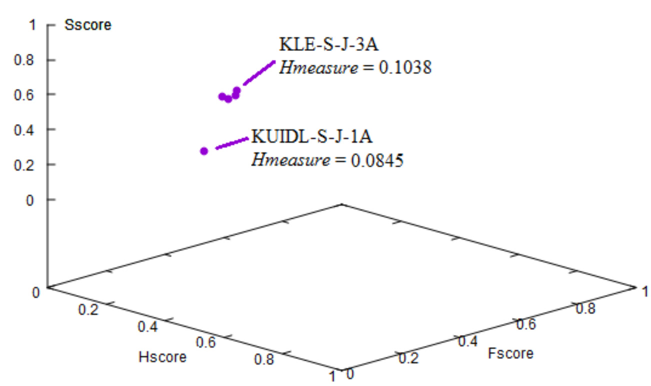

Figure 9. Submitted runs in Japanese SM task and their performance comparison

Table 10. Japanese Subtopic Mining runs ranked by *H-measure* (official result) over 34 unclear topics. The highest value in each column is shown in bold.

	Hscore	*Fscore*	*Sscore*	*H-measure*
KLE-S-J-3A	0.2030	0.4416	**0.5086**	**0.1038**
KLE-S-J-4A	0.2025	0.3920	0.4997	0.1008
KLE-S-J-2A	0.1867	**0.4502**	0.4697	0.0908
KLE-S-J-1A	0.1759	0.4372	0.4509	0.0853
KUIDL-S-J-1A	**0.2702**	0.2629	0.2848	0.0845

5.2 Document Ranking Subtask

As stated in Section 3.1, we follow the settings in INTENT/INTENT2 and choose *D#-nDCG* as the main evaluation metric for Document Ranking subtask. Since a hierarchy of subtopics is provided for each unclear query topic, we actually have two lists of subtopics for each of these queries: a first-level subtopic list and a second-level one. Therefore, we could evaluate the submitted runs with either fine-grained or coarse-grained search intents. The evaluation results are shown in Tables 11 and 12 for Chinese DR and English DR tasks, separately. In the results shown in these tables, the performance of clear queries are evaluated with *nDCG*, which could be regarded as a special case for *D#-nDCG* with no diversified subtopic lists.

Table 11. Chinese Document Ranking runs ranked by coarse-grain *D#-nDCG* (official result) over all query topics. The highest value in each column is shown in bold.

	Coarse-grain results (evaluated with first-level subtopics)	Fine-grain results (evaluated with second-level subtopics)
TUTA1-D-C-1B	**0.7334**	**0.6538**
THUSAM-D-C-1A	0.6965	0.6127
THUSAM-D-C-1B	0.6943	0.6106
FRDC-D-C-1A	0.4619	0.4118
FRDC-D-C-3A	0.4440	0.3950
FRDC-D-C-2A	0.3899	0.3402
FRDC-D-C-5A	0.3841	0.3338
FRDC-D-C-4A	0.3746	0.3240
THUSAM-D-C-2B	0.3697	0.2711
THUSAM-D-C-2A	0.3502	0.2623

Evaluation results in Tables 11 and 12 show that the coarse-grain results and fine-grain results are highly correlated (correlation values are both over 0.99). In Chinese DR task, TUTA gains best performance for both coarse-grain and fine-grain subtopic lists and the difference between their best run (TUTA1-D-C-1B) and the second best run (THUSAM-D-C-1A) is significant (two-tailed t-Test with p-value<0.01). According to descriptions given by TUTA, they adopt the subtopic list submitted to Chinese SM task and use different ranking strategies for different kinds of topics. This run is based on the non-diversified baseline provided by organizers. Considering the fact that TUTA doesn't gain very promising results in SM task (no better than FRDC and THUSAM), we believe that the ranking strategy they adopt must be effective and we would like to read more details in the technical paper.

From the results in Figure 10, we can see that the coarse-grain *D-nDCG* value of THUSAM-D-C-1A is higher than that of TUTA-D-C-1B. It probably show that the THUSAM run tends to adopt a relevance-oriented strategy while the TUTA one focuses more on intent recall. We can also see that these two runs gain much better performance than the other runs according to both coarse-grain and fine-grain results.

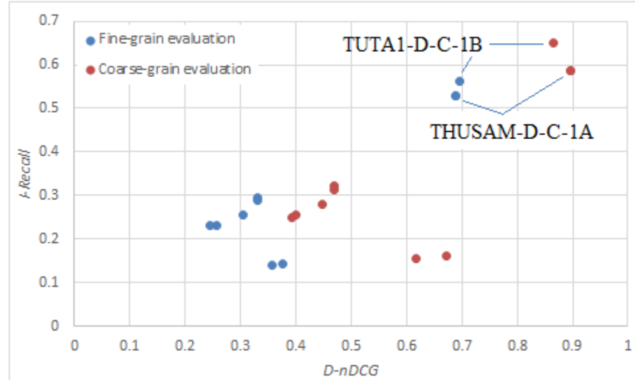

Figure 10. Best performing runs in Chinese DR task and their relationship with other submitted runs

According to evaluation results in Table 12, udel gains best performance with coarse-grain subtopic lists but the differences among best runs of udel, UM13, TUTA1 and Sem13 are not significant (two-tailed t-Test with p-value<0.01). Similarly, *D#-nDCG*s of the best performing runs of udel, UM13, TUTA1 and Sem13 with fine-grain subtopic lists are not significantly different, either. It is probably due to the fact that the pool depth for English DR runs are a bit shallow (covers top 10 results of the top priority runs). For the top performing runs, udel adopts query suggestions as inputs and use data fusion techniques to combine different ranking lists. TUTA adopts the same strategy in Chinese DR and UM13 employs a number of external resources including query logs, Wikipedia, ConceptNet and query suggestions from commercial search engines.

Table 12. English Document Ranking runs ranked by coarse-grain *D#-nDCG* (official result) over all query topics. The highest value in each column is shown in bold.

	Coarse-grain results (evaluated with first-level subtopics)	Fine-grain results (evaluated with second-level subtopics)
udel-D-E-1A	**0.6297**	0.5469
UM13-D-E-1A	0.6254	0.5566
TUTA1-D-E-1B	0.6170	**0.5668**
Sem13-D-E-1A	0.6022	0.5291
UM13-D-E-2A	0.6001	0.5309
Sem13-D-E-3A	0.4735	0.3985
Sem13-D-E-2A	0.4495	0.3806
UM13-D-E-3A	0.4474	0.3770
udel-D-E-2A	0.3900	0.3181
udel-D-E-4A	0.3472	0.2808
TUTA1-D-E-2B	0.3314	0.2601
Sem13-D-E-4A	0.3227	0.2505
Sem13-D-E-5A	0.3081	0.2414
udel-D-E-3A	0.0985	0.0784
udel-D-E-5A	0.0932	0.0877

Table 13. Runs submitted to TaskMine subtask and their descriptions.

Run	Description
TM-TS-uhyg-1	Firstly, we search for seed Web pages by using the query string with the word "houhou", which means "method". We collect more pages in consideration of the anchor texts in the seed pages. Then, we find pairs of chunks satisfying predefined patterns by dependency parsing on the sentences. We extract a target and a postponed particle from a depending chunk, and extract an operation from a depended chunk. We regard ternaries of them as subtasks. The extracted subtasks are ranked by their frequency-based score.
TM-TS-uhyg-2	
TM-TS-InteractiveMediaMINE-1	Our system consists of four steps. First, obtaining ten Web pages from Yahoo! Chiebukuro using sentences user inputs as a search query. Second, Extracting the answers of the answerers from the pages. Third, extracting appropriate sentences to achieve the task using Morphological Analysis and Syntactic Analysis. Fourth, scoring sentences by term frequency in the set of extracted sentences.
TM-TS-ORG-1	Organizers baseline: This method first obtains a set of Web documents by issuing a query to a Web search engine. It then extracts phrases that match with the pattern <Noun><Preposition><Verb> and ranks the phrases by the TF-IDF weighting.

Table 14. nDCG@*k* for TaskMine subtask runs. Runs sorted by nDCG@50.

	nDCG@1	nDCG@5	nDCG@10	nDCG@50
InteractiveMediaMINE1	**0.323**	**0.330**	**0.320**	**0.289**
uhyg2	0.109	0.150	0.171	0.191
uhyg1	0.098	0.119	0.132	0.166
ORG	0.013	0.040	0.053	0.096

From the results shown in Figure 11, we can see that udel-D-E-1A gains much higher I-recall value compared with other runs in coarse-grain evaluation. It also gains best D-nDCG value in fine-grain evaluation. This difference shows that although the results given by coarse-grain and fine-grain evaluations are highly correlated, same strategy results in different performance evaluation results with subtopic lists in different grains.

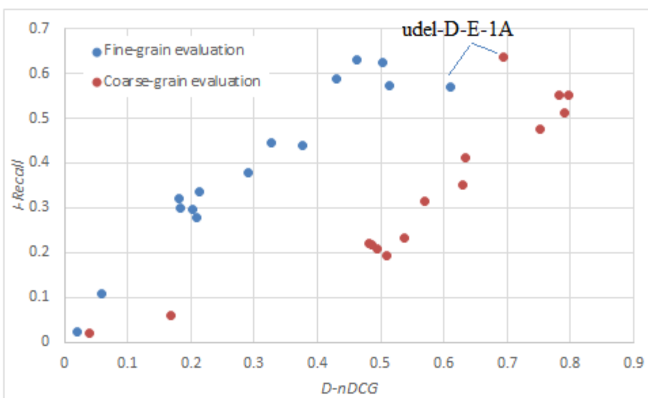

Figure 11. Best performing runs in English DR task and their relationship with other submitted runs

5.3 TaskMine Subtask

This section presents a list of submitted runs and the official results of the TaskMine subtask.

5.3.1 Submitted Runs

Table 13 shows runs submitted to the TaskMine subtask. We receive three runs from two groups to the TaskMine subtask. We can see that uhyg (TM-TS-uhyg-1 and TM-TS-uhyg-2) uses query modifications and anchor texts to obtain relevant documents, while InteractiveMediaMINE (TM-TS-InteractiveMediaMINE-1) uses the Community Q&A corpus (Yahoo! Chiebukuro) as the information resource.

The organizers also prepared a simple baseline method (TM-TS-ORG-1) to compare the effectiveness of submitted runs. It extracts phrases from the Web documents and ranks them by the TF-IDF weighting.

5.3.2 Results

Retrieval evaluation is performed for all queries and runs and the nDCG is computed for a variety of cutoff thresholds k. We first report the results of nDCG at k = 50 (nDCG@50), the primary measure in the TaskMine subtask. Figure 12 plots the results of nDCG@50 for all the runs. From the figure, we can see that InteractiveMediaMINE clearly performs the best among all the runs. As shown in Table 13, IntereactiveMediaMINE, which achieves the best performance, relies on the Community Q&A corpus to mine tasks for a query. This indicates that the Community Q&A corpus would be a useful resource to mine tasks. Also, we can find that all the submitted runs (TM-TS-uhyg-1, TM-TS-uhyg-2 and TM-TS-InteractiveMediaMINE-1) perform better than the organizer's baseline method.

To examine the significant differences among the runs in terms of nDCG@50, we performed a one-way ANOVA and the result show that there is a significant difference among runs (p-value < 0.01), and a Tukey's HSD post-hoc test shows that there are significant differences between all the pairs (p-value < 0.01), except one between uhyg-1 and uhyg-2.

Table 14 shows the results of nDCG@k for various k (k = 1, 5, 10, 50). We here see that InteractiveMediaMIME performs best in all the cutoff k. We can also see that InteractiveMediaMIME succeeds to rank important tasks at higher ranks as it achieves better nDCG@1 than nDCG@50.

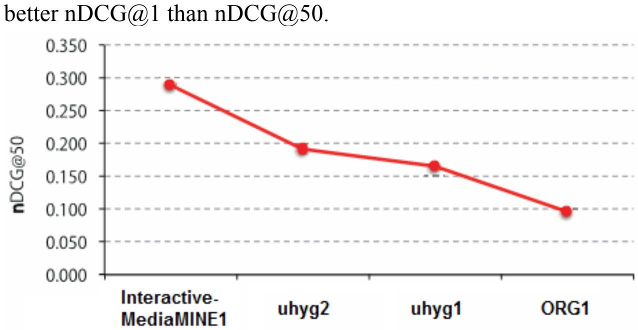

Figure 12. nDCG@50 averaged over all queries in TaskMine subtask.

As described in Section 2.4, queries can be broken down four categories. Figure 13 shows the results of nDCG@50 for different query categories. From the figure, we can see that the runs perform similarly on the various query categories, but uhyg performs well on the *Sequential* category. It is probably due to the uhyg's query modification approach is suitable to obtain tasks for Sequential queries.

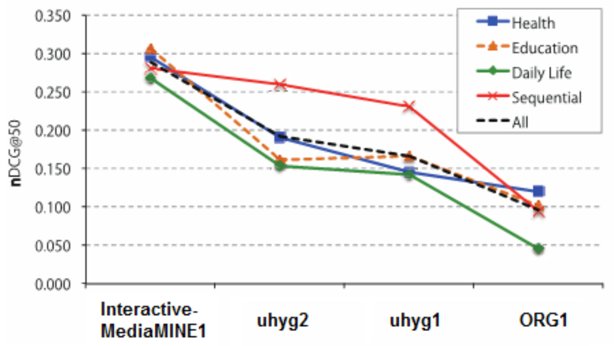

Figure 13. nDCG@50 averaged across each query category in TaskMine subtask.

Overall, runs submitted by the participants perform well compared to the baseline method prepared by the organizers. InteractiveMediaMINE uses a Community Q&A corpus and achieves the best among submitted runs. The approach taken by uhyg also works well, especially for the *Sequential* queries. On the other hand, there is lots of room for improvement in the TaskMine subtask since nDCG@50 of the top performer was less than 0.300. We need to continue to explore the effective methods for this task.

As for the evaluation process in this TaskMine subtask, we asked the assessor to match a participant task to one gold-standard task. One problem of this process would be a participant task with longer text is likely to be matched with a gold standard task, so a run that outputs longer task strings are likely to achieve higher nDCG. We need to treat this problem and devise more suitable metrics to our task.

6. USER PREFERENCE TEST RESULTS

As in most information retrieval evaluation researches, the evaluation metrics adopted in diversified search are based on a number of user behavior assumptions. For example, with the assumption that users always examine search results from top to bottom, most metrics leverage a ranking-based discount. These assumptions may not always hold in practical Web environment (e.g. user revisits results in a large proportion of Web search sessions according to [15]). To validate the credibility of diversity evaluation metrics, a number of methods that "evaluate evaluation metrics" are adopted in diversified search evaluation studies, such as Kendall's tau, Discriminative Power [16], and the Intuitiveness Test [17]. These methods have been widely adopted and have aided us in gaining much insight into the effectiveness of evaluation metrics. However, they also follow certain types of user behaviors or statistical assumptions and do not take the information of users' actual search preferences into consideration.

To look into the reliability of evaluation metrics and make sure it accords with practical users' preferences, we follow the works of [18] and [19] to compare evaluation results based on existing evaluation metrics and user preference tests. Different from these existing works, we are among the first to compare both fine-grained and coarse-grained diversified evaluation results with user preference tests thanks to the construction of two-level hierarchy of subtopics in IMine.

In user preference test, we select 5 of the 10 runs from the Chinese DR task according to the priorities noted by participants (TUTA1-D-C-1B, THUSAM-D-C-1A, THUSAM-D-C-2A, FRDC-D-C-1A and FRDC-D-C-2A). We don't involve all runs in the preference test because of limited resources and the fact that adding extra run will significantly increase the efforts. Each 2 of the 5 runs are then presented to users in a paralleled way to collect 7-graded preference as shown in Figure 14.

From the figure we can see that the users are required to give a confidence score according to their satisfaction with two paralleled ranking lists. The score ranged from -4 to 4 with integer numbers. A minus score means the left-side list performs better while a positive score represents a better right-side ranking list. The absolute value of score (user confidence) is used to show the degree of performance differences between two lists and the value of zero means it is difficult to tell which one is better.

Altogether 30 students from Tsinghua University were recruited to finish the user preference test. All of the participants are at the second year of undergraduate study and their majors include medicine, automobile engineering and computer science. For each pair of runs to be annotated, we have 3 assessors to review the paralleled ranking lists of all 49 query topics (Topic NO. 0033 is not included because no results are returned for that topic in all runs). Results from different runs are randomized so that they have equal opportunities to be presented at the left/right side of the annotation page. With these experimental settings, each participant is required to make 49 preference judgments. We believe that this workload is reasonable for most of our participants to guarantee quality and it usually takes them 30-50 minutes to finish the jobs. We collected 1470 preference judgments for the runs to be annotated with this procedure.

To obtain user preference test results from the user judgments, we at first remove the preference judgments with confidence score equals to -1, 0 or 1. We believe that such judgments largely indicate the two runs provide comparable result rankings. After that, the scores of remaining judgments are averaged for each topic and the signal of these average scores are used as a preference judgment for the corresponding result pairs. With this method, the preference test result is shown in Table 15.

Table 15. User preference test results for Chinese DR task (A>B/A<B: participants prefer A(B) over B(A) with a confidence score of 2, 3 or 4; A=B: participants select "hard to tell the difference" or show preferences with score of 1)

Run A	Run B	A>B	A=B	A<B
TUTA1-D-C-1B	FRDC-D-C-1A	**53.7%**	19.5%	26.8%
TUTA1-D-C-1B	FRDC-D-C-2A	**48.7%**	28.2%	23.1%
TUTA1-D-C-1B	THUSAM-D-C-1A	29.2%	22.9%	**47.9%**
TUTA1-D-C-1B	THUSAM-D-C-2A	**45.8%**	14.6%	39.6%
THUSAM-D-C-1A	FRDC-D-C-1A	**56.1%**	31.7%	12.2%
THUSAM-D-C-1A	FRDC-D-C-2A	**51.3%**	20.5%	28.2%
THUSAM-D-C-1A	THUSAM-D-C-2A	**54.2%**	39.6%	6.3%
FRDC-D-C-1A	FRDC-D-C-2A	32.4%	**43.2%**	24.3%
FRDC-D-C-1A	THUSAM-D-C-2A	31.7%	12.2%	**56.1%**
FRDC-D-C-2A	THUSAM-D-C-2A	28.2%	15.4%	**56.4%**

From Table 15 we can see that 7 out of 10 run pairs gain same user preferences as the order of D#nDCG values shown in Table 11. Meanwhile, users prefer a different result run compared with the order of D#nDCG values in 3 result pairs (TUTA1-D-C-1B v.s. THUSAM-D-C-1A, FRDC-D-C-1A v.s. THUSAM-D-C-2A and FRDC-D-C-2A v.s. THUSAM-D-C-2A).

Figure 14. User preference labeling system interface (Up: original query, description of subtopics (first-level subtopics for the query topic); Middle: document ranking results from two competing runs; Bottom: 7-point preference score board). Note that a number of results from each ranking list are removed to save space in the paper so that only 4 results from each list are shown.

When we look into these 3 run pairs shown in Table 15, we find that for the run pair TUTA1-D-C-1B v.s. THUSAM-D-C-1A, both the user preference test results (39.6% v.s. 45.8%) and the D#nDCG results (0.7334 v.s. 0.6965, with two-tailed paired t-test p-value=0.13) are not so different from each other. It probably mean that user preference test and cranfiled-like approaches may not accord with each other for closely comparable runs in diversified search evaluation. Considering the test collection strategy proposed in [23], if we want to a minimum detectable range of 0.10 in terms of D#nDCG for 10 systems under Cohen's convention, we should construct a topic set with at least 170 topics, which is much larger than the topic set adopted in IMine. This may explain the fact that the runs with small differences in D#nDCGs are not so different from each other for users in the preference test.

As for the other two run pairs (FRDC-D-C-1A v.s. THUSAM-D-C-2A and FRDC-D-C-2A v.s. THUSAM-D-C-2A), although the differences in D#nDCG values are more significant (both p-values equal to around 0.02), a lot more users prefer THUSAM results than FRDC results. We find that for a large proportion of query topics, FRDC return much less results than THUSAM (28.7 per topic for FRDC-D-C-1A, 30.9 per topic for FRDC-D-C-2A compared with 200 per topic for THUSAM-D-C-2A). For many topics in FRDC results, less than 10 results or only 1 or 2 results are returned. Although the returned results may be relevant to an important aspect in user need, too few results will hurt user satisfaction and lead to not so promising results in preference test.

7. CONCLUSIONS AND FUTURE WORK

IMine task aims to mine users' diversified intents behind their simple, unspecified and sometimes ambiguous queries submitted to search engines. It follows the research framework of previous INTENT/INTENT2 tasks in NTCIR9/10 but features the mining of hierarchical subtopic structures. In this year's task, the organizers work with participants to develop new evaluation metrics for SM task (*Hscore*, *Fscore*, *Sscore* and *H-measure*) and employ them to evaluate the performance of this Web search intent mining task. The evaluation of DR task is also different from previous tasks in that both a fine-grain and a coarse-grain comparison can be obtained with the generated second-level and first-level subtopic lists, separately.

Through the evaluation results, we found that best performing runs for SM task in different languages are usually based on a combination of different information resources (query suggestions, query dimensions and related queries). *Hscore* plays a central role in English SM task but the best runs in Chinese and Japanese runs seem to be *Sscore*-oriented. It seems that rule-based methods generate high quality subtopic candidates and the structure of Web page helps improve the quality of subtopic hierarchies.

Although the pool depth for DR task is not so deep and the evaluation results are not significantly different from each other for English task, we still find several interesting findings through the evaluation process. We find that coarse-grain results and fine-grain results are highly correlated, which means that it may not be necessary to use fine-grain subtopic lists and we can therefore avoid extra annotation efforts. We also find that user preference test results and the evaluation results measured by D#-nDCG are quite similar with each other, especially for the runs with large differences in D#nDCG values. We plan to find out possible limitations with current evaluation methodology of IMine and improve the technique in the future round.

As for the TaskMine subtask, it attracts two groups with different approaches. Through the evaluation results, we found that the best performing run for this task relies on the Community Q&A corpus. Since the TaskMine task is a new task starts from this NTICR-11, we need to continue to explore more suitable evaluation processes and metrics, as well as methods to mine effective tasks.

8. ACKNOWLEDGMENTS

We appreciate Prof. Jamie Callan and his team for providing the ClueWeb collection, which dramatically reduces the working efforts of participants in Document Ranking subtask. We also benefit a lot from discussions with Dr. Tetsuya Sakai and the PC chairs of NTCIR-11. Part of this work was supported by a Research Fund FY14-RES-SPONSOR-111 from Microsoft Research Asia.

9. REFERENCES

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

[2] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the NTCIR-10 INTENT-2 Task. In Proceedings of NTCIR-10 Workshop Meeting. 2013.

[3] Zhicheng Dou, Sha Hu, Yulong Luo, Ruihua Song and Ji-Rong Wen.: Finding Dimensions for Queries, ACM CIKM 2011, October 2011.

[4] Yiqun Liu, Junwei Miao, Min Zhang, Shaoping Ma, Liyun Ru.: How Do Users Describe Their Information Need: Query Recommendation based on Snippet Click Model. Expert Systems With Applications. 38(11): 13847-13856, 2011.

[5] M. Kendall. A new measure of rank correlation. Biometrica, 30:81-89, 1938.

[6] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In In Proceedings of ACM SIGIR 2006. ACM, Seattle, Washington, USA, pages 525-532, August 2006.

[7] T. Sakai. Evaluation with informational and navigational intents. In In Proceedings of ACM WWW 2012. ACM, Lyon, France, pages 499{508, April 2012.

[8] Ruihua Song, Zhenxiao Luo, Jian-Yun Nie, Yong Yu, and Hsiao-Wuen Hon. Identification of ambiguous queries in web search. Information Processing & Management, 45(2):216–229, 2009.

[9] Sakai, T. and Song, R.: Evaluating Diversified Search Results Using Per-Intent Graded Relevance, ACM SIGIR 2011, pp.1043-1052, July 2011.

[10] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. (2009). Diversifying Search Results. In Proceedings of the Second ACM International Conference on Web Search and Data Mining. ACM, Barcelona, Spain (pp.5–14).

[11] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. (2009). Expected Reciprocal Rank for Graded Relevance. In Proceedings of ACM CIKM 2009. ACM, Hong Kong, China. 621–630.

[12] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C.-Y. Lin. (2010). Simple Evaluation Metrics for Diversified Search Results. In 3rd International Workshop on Evaluating Information Access. Tokyo, Japan (pp.42–50).

[13] Cheng Zhai, William W. Cohen, John Lafferty. (2003). Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In Proceedings of ACM SIGIR 2003. ACM, Toronto, Canada (pp.10–17).

[14] T. Sakai. (2012). Evaluation with Informational and Navigational Intents. In Proceedings of ACM WWW 2012. ACM, Lyon, France (pp.499-508).

[15] Danqing Xu, Yiqun Liu, Min Zhang and Shaoping Ma. Incorporating Revisiting Behaviors into Click Models. Proceedings of the 5th ACM International Conference on Web Search and Data Mining. WSDM 2012 (Seattle, U.S.). pp. 303-312.

[16] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In In Proceedings of ACM SIGIR 2006. ACM, Seattle, Washington, USA, pages 525{532, August 2006.

[17] T. Sakai. Evaluation with informational and navigational intents. In In Proceedings of ACM WWW 2012. ACM, Lyon, France, pages 499-508, April 2012.

[18] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In In Proceedings of ACM SIGIR 2010. ACM, Geneva, Switzerland, pages 555-562, July 2010.

[19] Fei Chen, Yiqun Liu, Zhicheng Dou, Keyang Xu, Yujie Cao, Min Zhang, and Shaoping Ma. Revisiting the Evaluation of Diversified Search Evaluation Metrics with User Preferences. Proceedings of the 10th Asia Information Retrieval Society Conference (AIRS 2014).

[20] T. Yamamoto, T. Sakai, M. Iwata, Y. Chen, J.-R. Wen and K. Tanaka. (2012). The Wisdom of Advertisers: Mining Subgoals via Query Clustering. In Proceedings of ACM CIKM 2012, pp.505-514, 2012.

[21] D. Donato, F. Bonchi, T. Chi and Y. Maarek. (2010). Do You Want to Take Notes?: Identifying Research Missions in Yahoo! Search Pad, In Proceedings of WWW 2010, pp. 321–330, 2010.

[22] S. Oyama and K. Tanaka. Query modification by discovering topics from web page structures. In Proceedings of the 6th Asia Pacific Web Conference, pp.553-564, 2004.

[23] Sakai, T., Designing Test Collections for Comparing Many Systems, Proceedings of ACM CIKM 2014.

[24] Joho, H. and Kishida, K. Overview of NTCIR-11. In: Proceedings of the 11th NTCIR Conference, Tokyo, Japan, 2014.