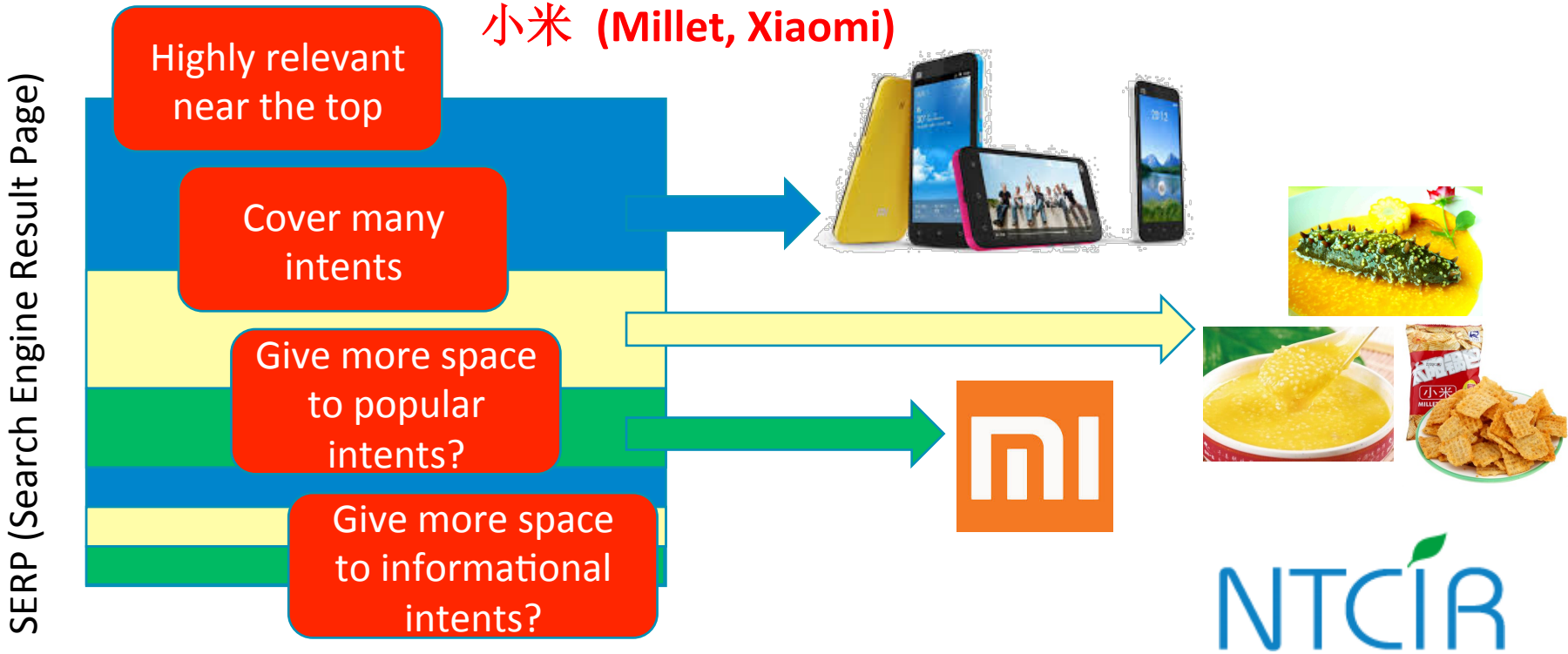
Overview of The NTCIR-11 IMine Task

Yiqun Liu, Ruihua Song, Min Zhang, Zhicheng Dou, *Takehiro Yamamoto*, Makoto P. Kato, Hiroaki Ohshima, Ke Zhou

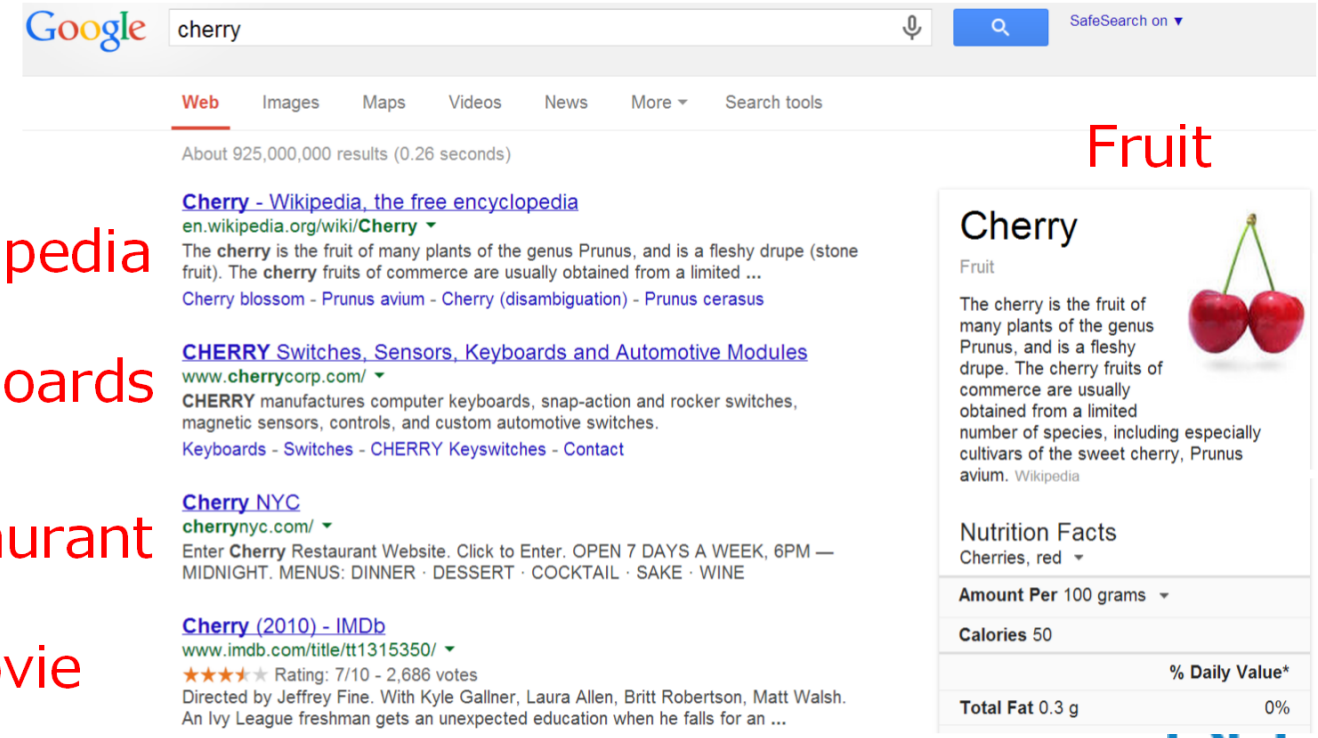
Background: Diversified Search

- Given an ambiguous/underspecified query, produce a single result page that satisfies different user intents!
- Challenge: balancing relevance and diversity with results from heterogeneous information sources



Background: Diversified Search

- Possible framework for diversified search
  - Identification of *ambiguous*/*broad*/*clear* queries
  - Generation of subtopics for *ambiguous*/*broad* queries
  - Search result diversification for better ranking



The IMine task

- IMine (曖昧, *ambiguous* in Japanese) Task Goal
  - To explore and evaluate the technologies of mining and satisfying different user intents behind a Web search query
- A core task in NTCIR-11 and succeeding work of INTENT@NTCIR-9 and INTENT2@NTCIR-10 tasks
- Three subtasks
  - TaskMine (TM) subtask: to find subtasks of a given task described by a query.
  - Subtopic Mining (SM) subtask: automatically estimating different intents of a given query.
  - Document Ranking (DR) subtask: Selectively diversifying search results by balancing between relevance and diversity

NTCIR

Differences from Previous Tasks

- Mining and evaluating hierarchical user intents
- More subtopic candidates provided (from commercial search engine, user behavior log mining and result page analysis)
- New corpus(ClueWeb12-B13), More public user behavior data (doubled size)
  - 1.85GB => 3.85GB, over 40M user clicks
- User preference test v.s. Cranfield-like evaluation with professional assessment in diversified search evaluation

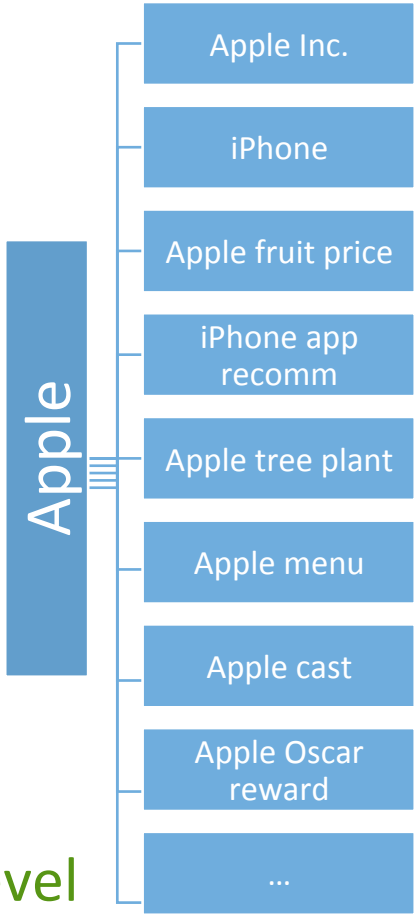
NTCIR

IMine Task Timetable

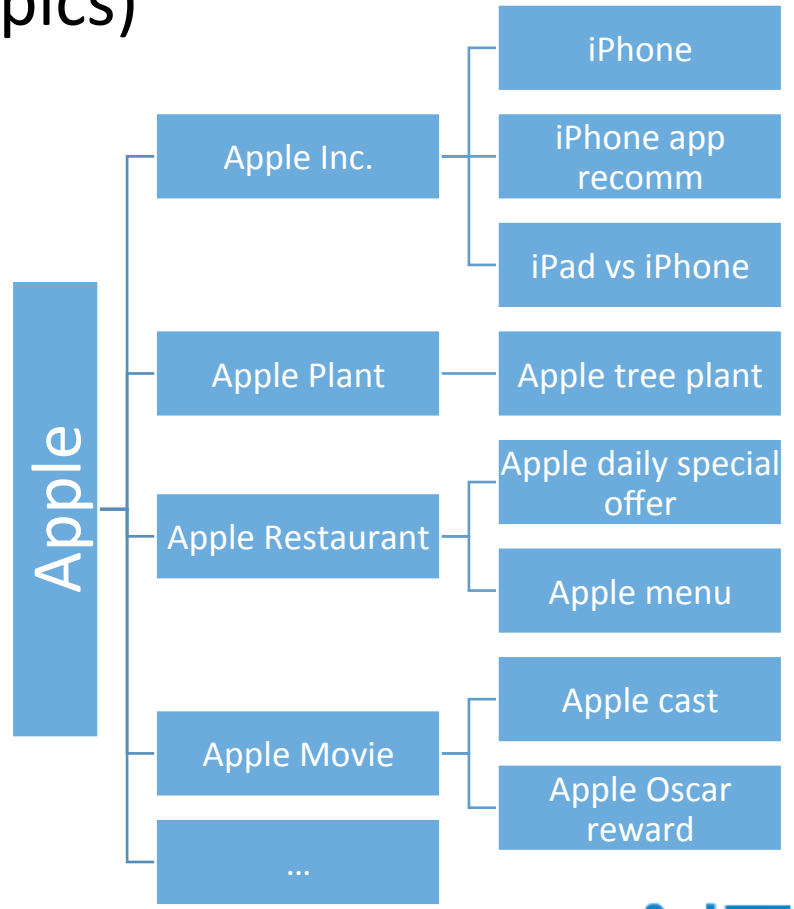
- Corpus available: Aug 31, 2013
- Call for participants: Aug 31, 2013
- Task participant registration Due: Jan 20, 2013
- Topics and non-diversified baseline DR runs released: Jan 21, 2014
- SM and DR submissions due: May 23, 2014
- Evaluation results available: Aug 15, 2014 (delayed 2 weeks)
- Early draft overview paper available: Aug 22, 2014 (delayed 3 weeks)
- Draft participant paper submission due: Sept 15, 2014
- Final Overview paper available: Oct 1, 2014
- Camera-ready participant paper submission due: Nov 1, 2014

Subtopic Mining Settings

- Goal: a two-level hierarchical list of subtopics for each query topic (5*10 subtopics)



Single-level



Two-level

NTCIR

Subtopic Mining Settings

- Query set

Language	#Topics			#Shared Topics
	Ambiguous	Broad	Clear	
English	16	17	17	14 shared topics for E/C/J (another 8 for E/C)
Chinese	16	17	17	
Japanese	17	17	16	

- Candidate subtopics provided
  - Query suggestions collected from Bing, Google, Sogou, Yahoo! and Baidu
  - Query dimensions generated by rule-based method (Dou et al., 2011) from search results
  - Query facets generated by keyword extraction from clicked snippets on SERPs (Liu et al., 2011)

NTCIR

Participants' Techniques (SM)

- Additional candidate sources
  - Disambiguation items from Wikipedia/Baidu Baike (e.g. FRDC, THUSAM)
  - Random walk on query-result bipartite graph with user behavior logs (e.g. THUSAM)
  - Title / keyword / anchor of landing pages (e.g. KUIDL)
- Generating two-level hierarchy
  - Clustering candidates to find similar second-level subtopics (e.g. FRDC, THUSAM)
  - Extracting first-level subtopics from clusters with word embedding, semantic expansion or rule-based methods (e.g. KLE, KUIDL, hultech)
  - Web page structures are used to identify the matching of first-level and second-level subtopics (e.g. KUIDL)

Subtopic Mining Evaluation

- A new metric considering both the importance of subtopics and the quality of the subtopic hierarchy.
- A mixture of three factors:
  - ***H-score***: evaluate the matching of first-level and second-level subtopics (accuracy-based)
  - ***F-score***: evaluate ranking of first-level subtopics (D#-nDCG based)
  - ***S-score***: evaluate ranking of second-level subtopics (D#-nDCG based)

$$H-measure$$
$$= Hscore * (\alpha * Fscore + \beta * Sscore), \qquad (\alpha + \beta = 1)$$

  - For ambiguous queries, $\alpha = \beta = 0.5$
  - For broad queries, $\alpha = 0$, $\beta = 1.0$

NTCIR

Document Ranking Settings

- Goal: a diversified ranked list of no more than 100 results for each query topic
- Chinese corpus: SogouT (ver. 2008)
  - 130M Chinese pages
  - Organizer provided a non-diversified baseline (adopted by TUTA and THUSAM)
- English corpus: ClueWeb12-B13
  - 52M English Web pages
  - A search interface is provided by Lemur project
  - Many thanks to Prof. Jamie Callan and his team
- Evaluation: D#nDCG (weight=0.5)

$$D\#nDCG = \lambda \cdot I - recall + (1-\lambda) \cdot D - nDCG$$

NTCIR

Participants' Techniques (DR)

- External sources adopted
  - Query logs, Wikipedia, ConceptNet and query suggestions
- Result diversification based on subtopics
  - Result combination via filling up multiple knapsacks (TUTA)
  - Result selection based on pruned exhaustive search (THUSAM)
  - Result selection based on greedy search (UM13, SEM13)
  - Result aggregation based on original ranking (udel)
- Result diversification based on novelty detection or redundancy detection
  - Result re-ranking with HITS (THUSAM)

NTCIR

Document Ranking Evaluation

- User preference test v.s. Cranfield-like approach
  - 30 students were recruited to finish the preference test
  - Each pair of results are annotated by 3 students

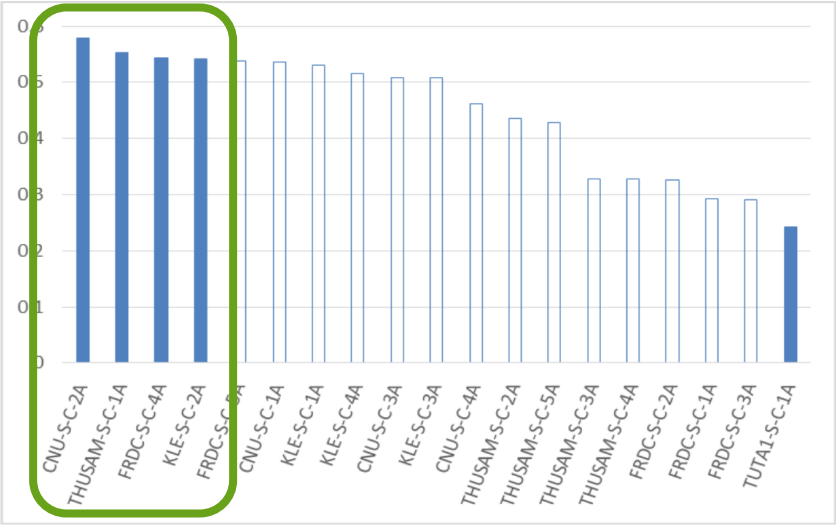
Result Submissions

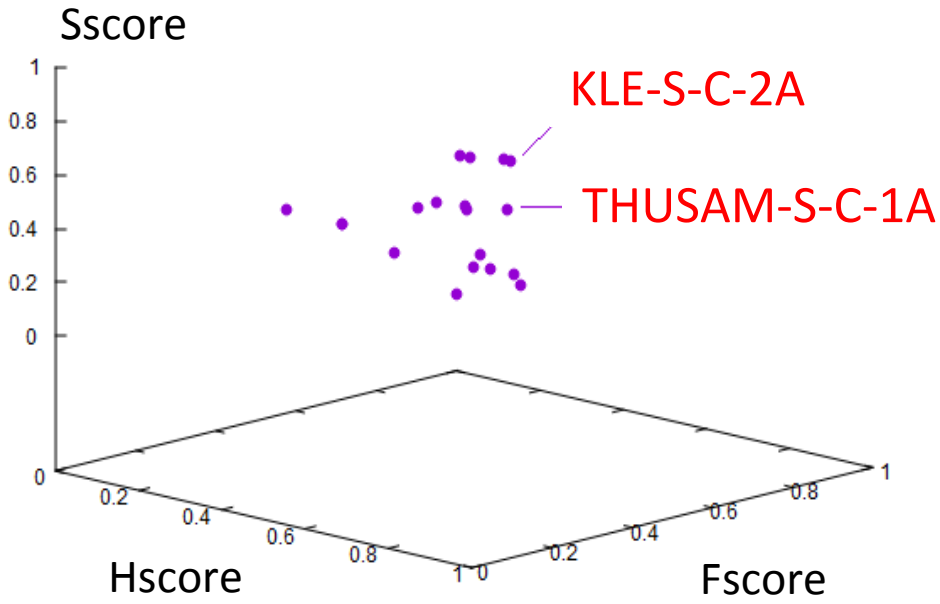
- 10 teams submitted results
  - Universities and research institutes from Canada, China, France, Japan, Korea and U.S.

Group	SM-C	SM-J	SM-E	DR-C	DR-E
UDEL			1		5
SEM13			5		5
HULTECH			4		
THU-SAM	5		2	4	
FRDC	5			5	
TUTA1	1		1		2
CNU	4				
KUIDL		1	1		
UM13			3		3
KLE	4	4	4		
#Group	5	2	8	2	4
#Run	19	5	29	9	15

Evaluation Results (SM)

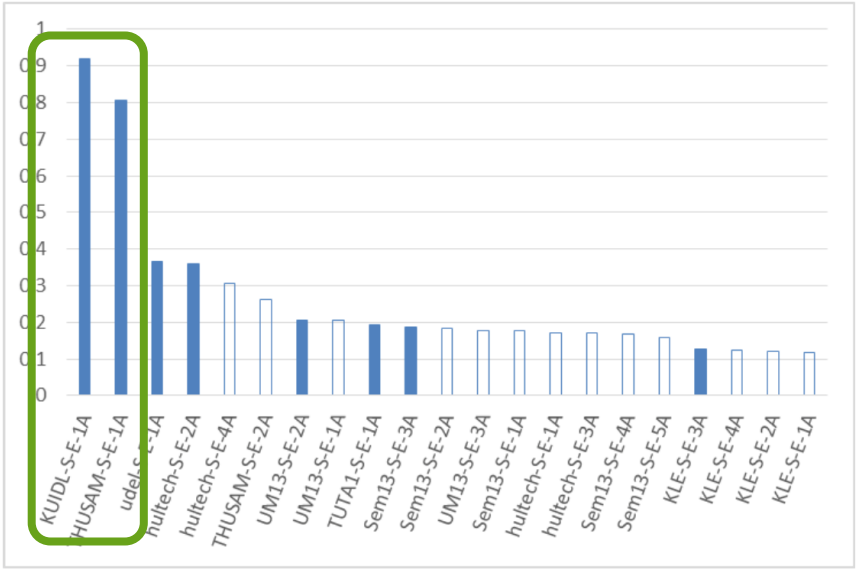
- KLE performs best in Chinese and Japanese SM task
  - *S-score* oriented ranking (best *S-score* for SME/SMC/SMJ)
  - *H-scores* of CNU, KLE, THUSAM and FRDC are not significantly different from each other for SMC
  - KLE approach in SLS extraction: semantic pattern matching in top-ranked search results (KLE oral report: 15:00@Day3, session A-2)

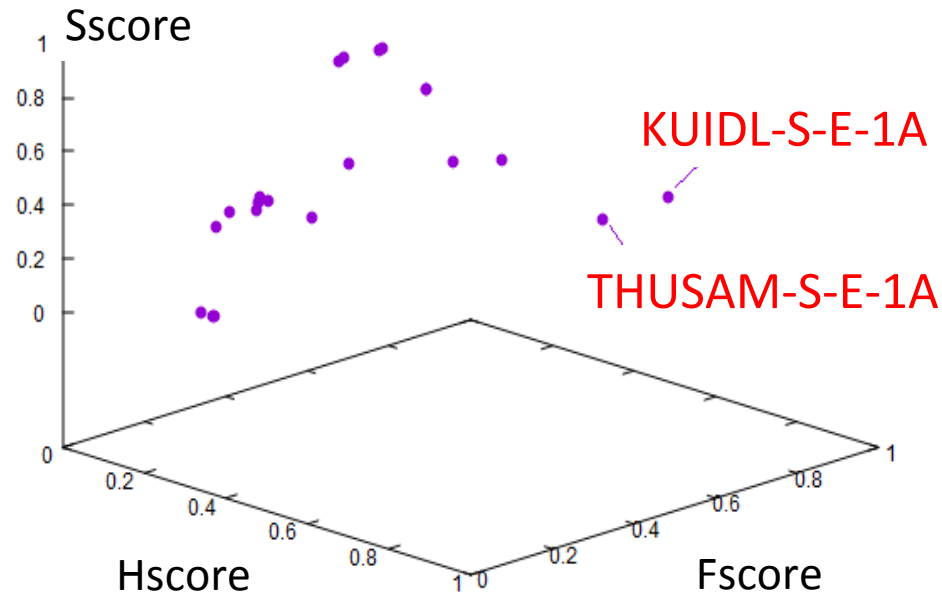




Evaluation Results (SM)

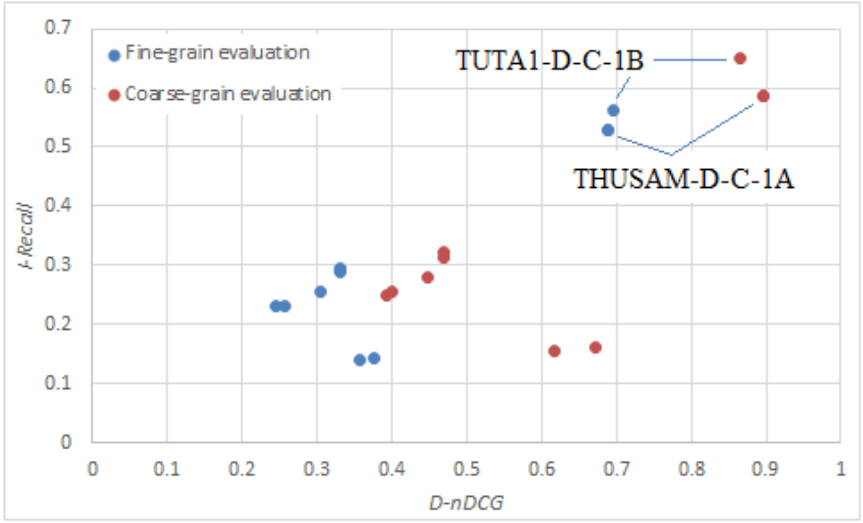
- KUIDL performs best in English SM task
  - *H-score* plays the most important part (KUIDL and THUSAM gain best performance with no significant difference)
  - KUIDL approach: document structure of result pages (KUIDL oral report: 14:30@Day3, session A-2)
  - Similar strategy in KUIDL and THUSAM: FLS is a substring of corresponding SLS

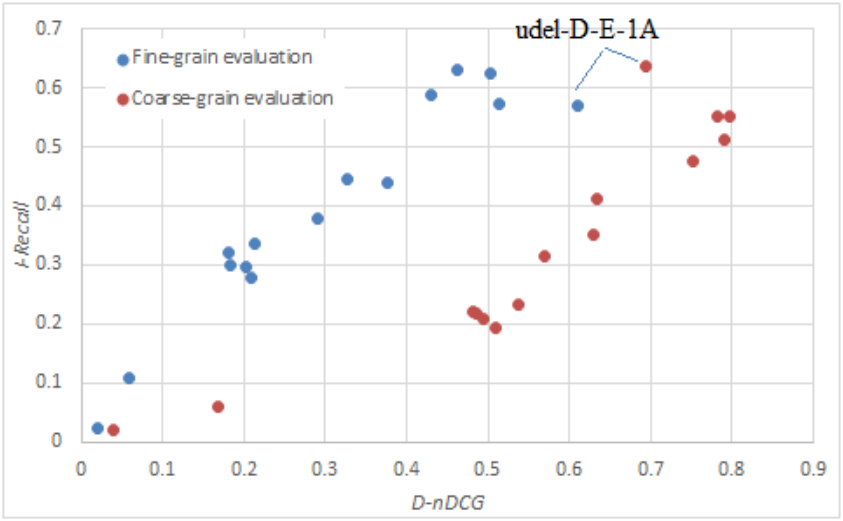




Evaluation Results (DR)

- Fine-grain v.s. Coarse-grain evaluation
  - Evaluation based on second-level or first-level subtopic list
- TUTA performs best for DRC and fine-grain DRE; Udel performs best for coarse-grain DRE
  - No significant differences with other top runs
  - Top performers gain more balanced results compared with previous INTENT tasks (both I-recall and D-nDCG are high)





Evaluation Results (DR)

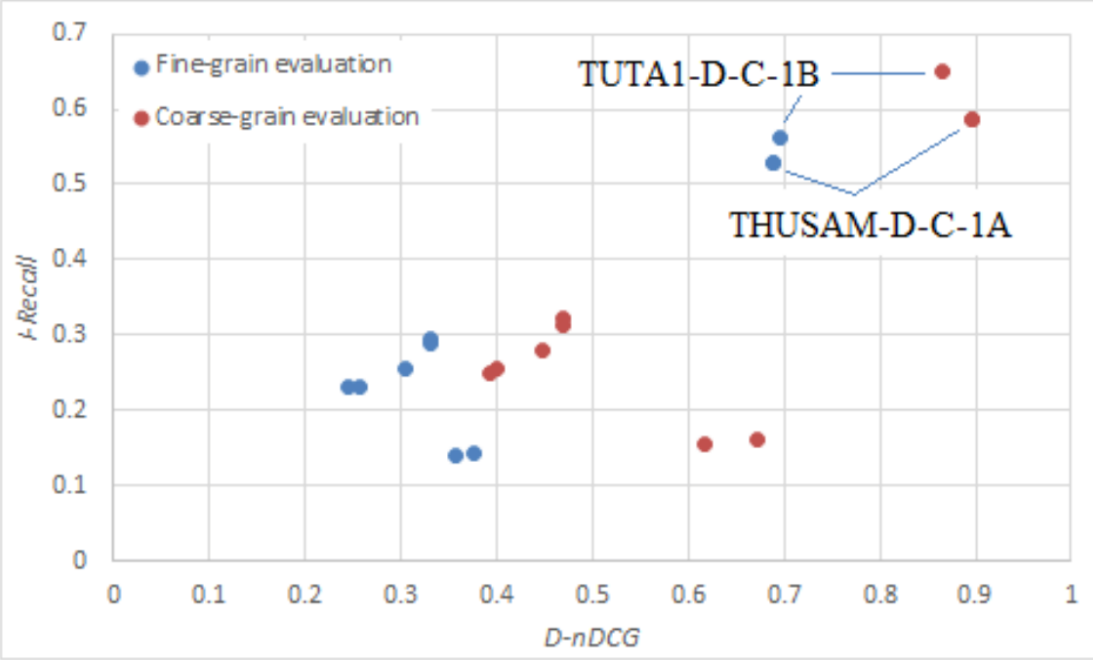
- User preference test v.s. Cranfield-like approach

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%**

NTCIR

Evaluation Results (DR)

- User preference test v.s. Cranfield-like approach (cont.)
  - TUTA1-D-C-1B v.s. THUSAM-D-C-1A: difference is not significant (two-tailed paired t-test p-value=0.13)



  - FRDC-D-C-1A/2A v.s. THUSAM-D-C-2A: FRDC systems sometimes return less than 10 results

Lessons Learned from This Round

- Less second level subtopic should be required
  - In this year, 5 first-level subtopics per query and 10 second-level subtopics per first-level subtopic is required
  - Too much assessment cost => reduced to 10-20 in practice
  - Not so reasonable in Web search scenario
- A more recent corpus should be adopted
  - One query from DRC fails to return any valid results: Android 2.3 game download
  - SogouT is crawled in 2008 (Android 2.3 didn't exist)
- Heterogeneous information sources (e.g. verticals) should be involved
  - Vertical results are necessary; users are not familiar with SERP without verticals

NTCIR

Take-home messages from SM and DR

- Subtopic structure is studied in SM task
  - Two-level hierarchy of subtopics requires much annotation efforts and sometimes cause confusions
  - KLE performs the best in SMJ and SMC with highest SLS mining performance (*Sscore*)
  - KUIDL performs the best in SME with highest FLS-SLS matching performance (*Hscore*)
- User preference test results are compared with Cranfield-like approach in DR task
  - Most results are similar with each other
  - *D#-nDCG* may not produce credible ranking when performances are close or lengths of result lists are different
- Top results are more balanced than previous tasks

Task Mining Subtask (TaskMine)

- As a subtask of IMine -

Takehiro Yamamoto

Makoto P. Kato

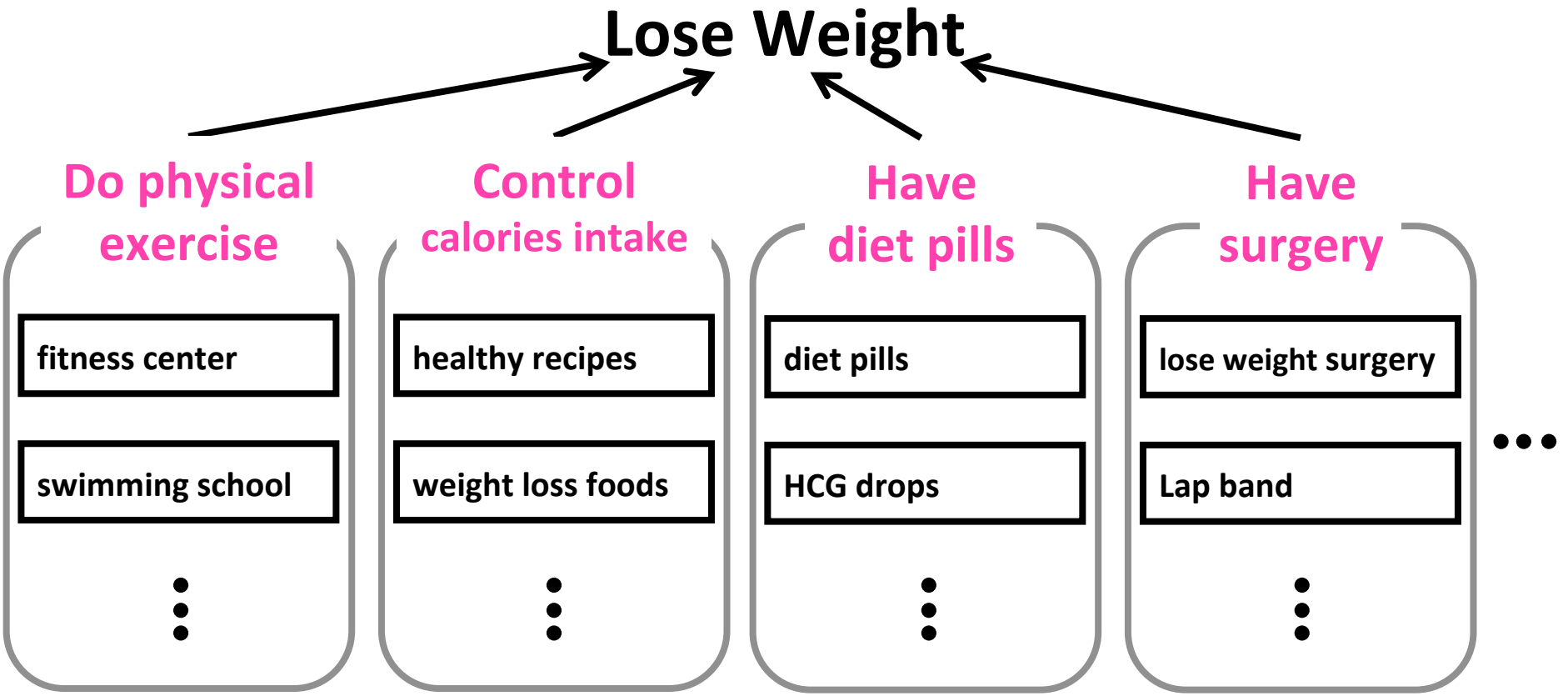
Hiroaki Ohshima

(Kyoto University)

December, 2014@NII, Tokyo

Background

Information needs of searchers are sometimes Complex



Goal of TaskMine subtask

- Understanding **the relationship among tasks** for supporting the Web searchers.
- Particularly, aims to explore the methods of automatically **finding subtasks of a given task**.

- Subtopic Mining subtask
  - Focus on the topical intent of a query
    - "I want to *find* this information!"
- TaskMine subtask
  - Focus on the task-oriented intent of a query
    - "I want to *accomplish* this task!"

NTCIR

Task

- Given a query (task), participants are required to return a ranked list of subtasks that help to achieve the query.

Input: **Query**

Output: **List of Subtasks**

Rank	Subtask
1	physical exercise
2	healthy food
3	diet pills
…	…

- Documents
  - Participants are allowed to use **any resources** on the Web

Queries

- Queries
  - 50 Japanese queries

Category	Examples
Health	禁煙する (quit smoking) ストレスを解消する(relief stress)
Education	中国語を勉強する(learn Chinese) 九九を覚える(master 9x9 table)
Daily Life	ペットを預ける(leave pet) ホエールウォッチングをする(whale watching)
Sequential	神社でお参りをする(pray in shrine) 食パンを作る (make bread)

Evaluation Methodology

1. Preparing gold-standard task
   - Ask assessors to create gold standard task for each query
   - Also ask assessors to vote the importance of each task
     - How the task effectively helps to achieve the given goal?

2. Matching gold-standard task and participant task
   - For each participant task, assessors were asked to select at most one corresponding gold-standard task

3. Evaluation Metric
   - nDCG (adopted to penalize the redundant output)

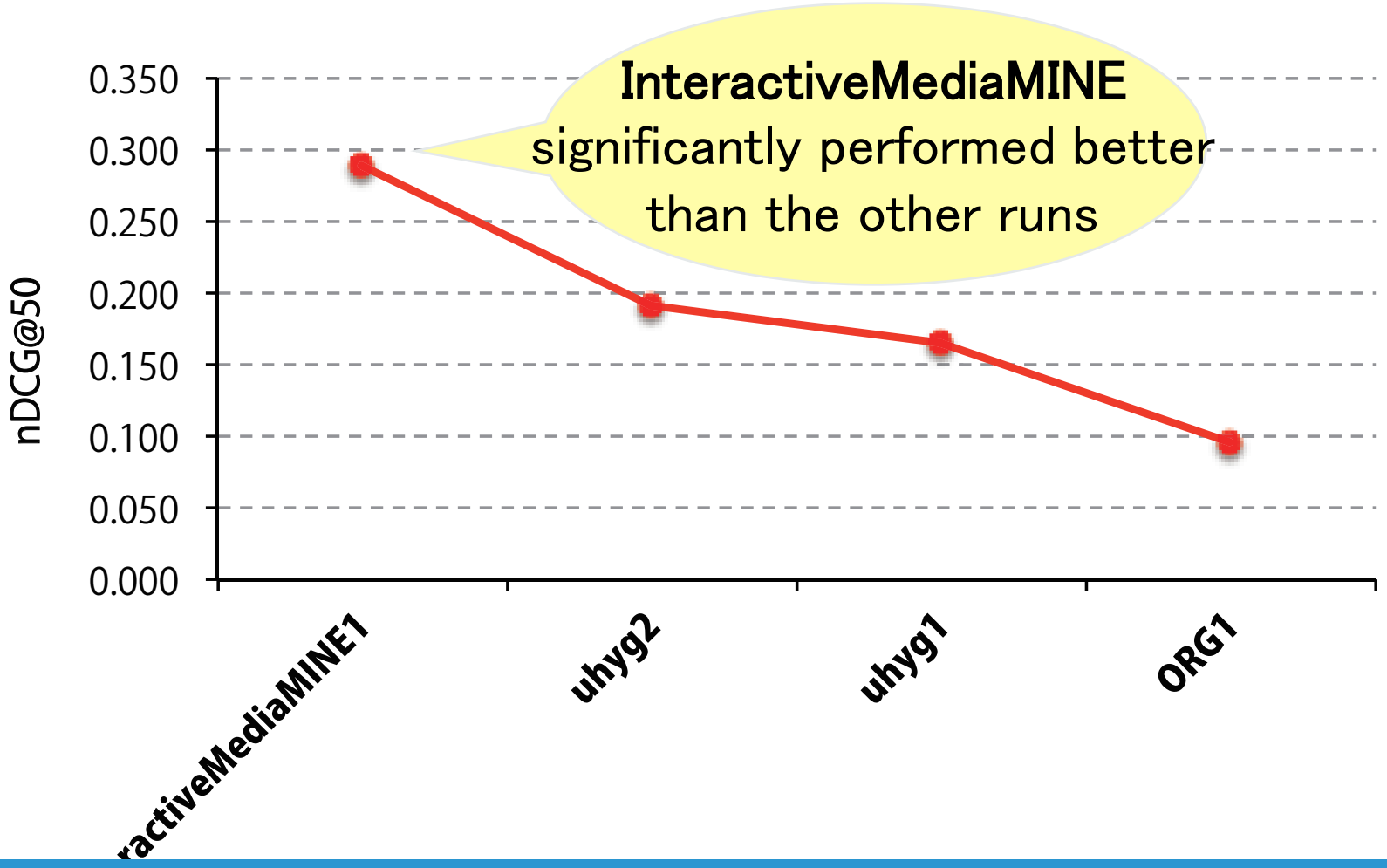
NTCIR

Participating Teams

- **uhyg** (University of Hyogo)
  - Use Web search engines
  - **Query expansion**
  - **Dependency parsing**
- **InteractiveMediaMINE** (Kogakuin University)
  - Use **Community Q&A corpus** (Yahoo! Chiebukuro)
  - **Dependency parsing**
- **Organizer's baseline**
  - Use Web search engines
  - Simple syntactic pattern with tf-idf weighting

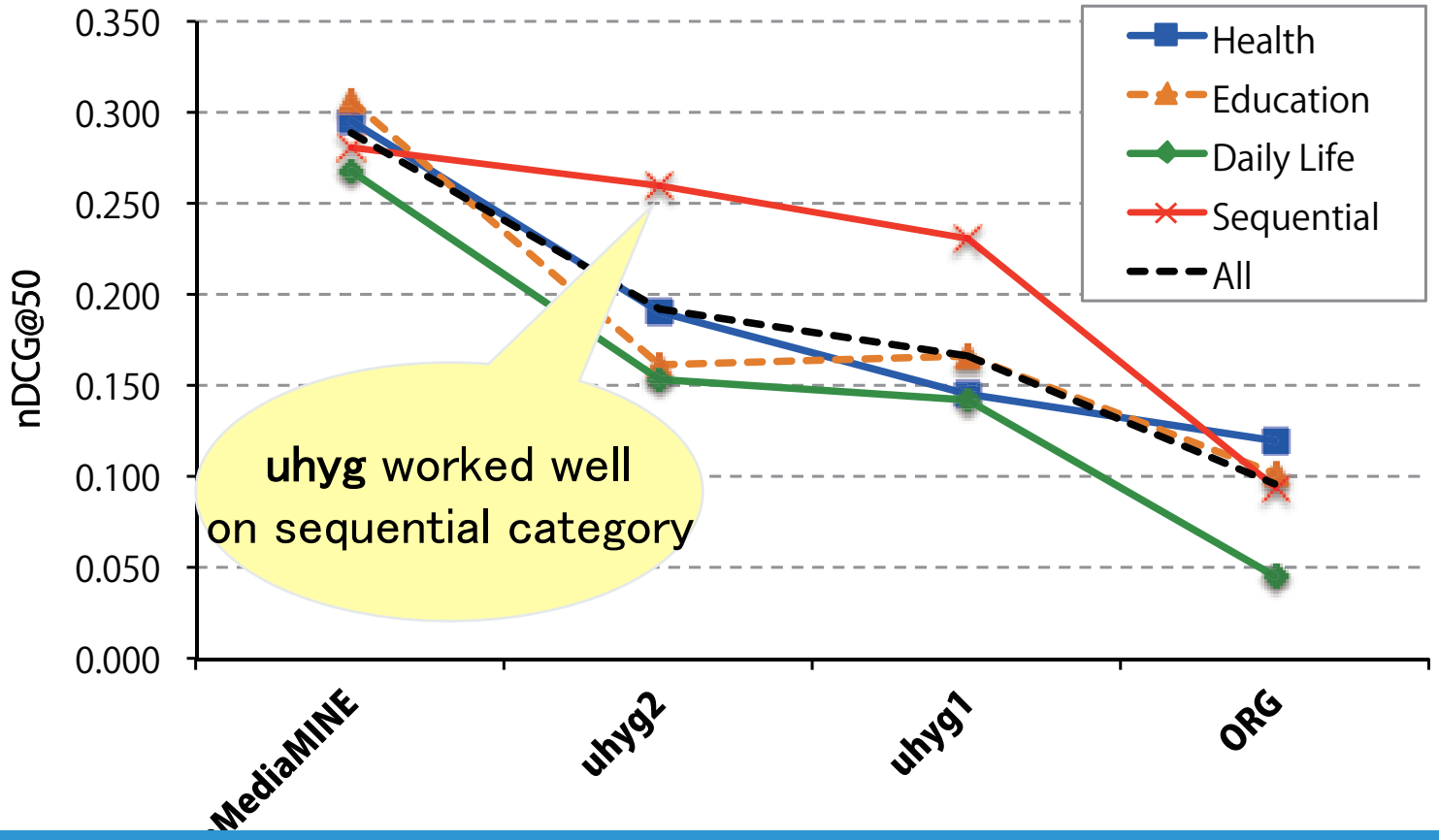
NTCIR

Overall Results



Community Q&A corpus
is useful resource for task mining

By Query Category



Query modification and dependency parsing are effective to mine sequential tasks

Summary of TaskMine

- Lessons Learned
  - Community Q&A corpus is a strong resource for the task mining
  - Query expansion and dependency parsing were effective on sequential types of query
  - InteractiveMediaMINE and uhyg have oral presentations @Day3 14:05～16:05
- Open Questions
  - How do the existing subtopic mining technique work well on TaskMine?
  - Can we aggregate heterogeneous information to find more effective tasks?

NTCIR

Future Plans of IMine-2

- In IMine-2, we will keep the basic task design in IMine-1, but more focus on vertical intents behind a query
  - More realistic to actual Web searches

- **Query Understanding subtask**
  - ≒ Subtopic Mining Subtask
  - Given a query, participants are required to identify its subtopics and their relevant verticals (web, news, image, movie, etc)
- **Vertical Incorporating subtask**
  - ≒ Document Ranking Subtask
  - Given a query, participants are required to return a diversified result and decide which result should be displayed with vertical results.

NTCIR

Thank you

http://www.thuir.org/IMine/
http://www.dl.kuis.kyoto-u.ac.jp/ntcir-11/taskmine/

NTCIR