Overview of NTCIR-12 Temporal Information Access (Temporalia-2) Task

Hideo Joho
Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba, Japan
hideo@slis.tsukuba.ac.jp

Adam Jatowt
Graduate School of Informatics, Kyoto University, Japan
adam@dl.kuis.kyoto-u.ac.jp

Roi Blanco
IR Lab, University of A Coruña, Spain
rblanco@udc.es

Haitao Yu
Faculty of Library, Information, and Media Science, University of Tsukuba, Japan.
yuhaitao@slis.tsukuba.ac.jp

Shuhei Yamamoto
Faculty of Library, Information, and Media Science, University of Tsukuba, Japan.
yamahei@ce.slis.tsukuba.ac.jp

ABSTRACT

This paper overviews NTCIR-12 Temporal Information Access (Temporalia-2) task. The task aims to foster research in temporal aspects of information retrieval and search, and is a continuation of Temporalia-1 task at NTCIR-11. Temporalia-2 is composed of two subtasks: Temporal Intent Disambiguation (TID) and Temporally Diversified Retrieval (TDR). Both the subtasks have English and Chinese language versions. A total of 47 runs were submitted by 15 teams across the world. This was 40% improvement in the number of participated teams when compared to the previous edition. TID in English attracted 12 teams which submitted a total of 30 runs, while its Chinese version attracted 3 teams submitting 7 runs. 4 teams including the organizer's team took part in TDR English language subtask with the total of 10 runs. In this paper we describe both the subtasks, datasets, evaluation methods and the results of meta analyses.

Subtasks

Temporal Intent Disambiguation (TID)
Temporally Diversified Retrieval (TDR)

Keywords

temporal information retrieval, evaluation

1. INTRODUCTION

Temporal Information Access (*Temporalia*)[1] task has been hosted at the 12th NTCIR Workshop on Evaluation of Information Access Technologies (NTCIR-12) [14] as one of core tasks. It is a continuation of Temporalia-1 task held at NTCIR-11 [11]. The task is an answer to the recent interest in temporal aspects of Information Retrieval and an attempt to establish common grounds for designing and analyzing time-aware information access systems. Temporal Information Retrieval [1, 2, 13] can be defined as a subset of document retrieval in which time plays crucial role in estimating document relevance.

As the core task we extended the two major search subproblems which were approached at Temporalia-1: query intent understanding and ranking documents considering their temporal aspects. The first one called *Temporal Intent Disambiguation* (TID) asks participants to estimate distribution of queries over four temporal classes following the intuitive time understanding of queries: past-related, recency-related and future-related. For comparison we have also added atemporal class that characterizes queries without any significant underlying temporal intent.

This task should be useful challenge for any researches that aim to recognize underlying temporal aspects of queries. With successful solutions, search engines could then treat temporal queries in a way that corresponds to their underlying temporal classes. According to the study performed on the AOL query dataset [22], about 1.5% of queries are explicit temporal queries, that is, they contain some temporal expressions. Examples of such queries are: "Tokyo Olympics 1964" or "most popular songs 2000s". Considering the popularity and importance of Web search in our lives, this rate amounts to quite a huge number of searches. In addition, there are also implicit temporal queries (e.g., "weather forecast Tokyo", "Berlin Wall collapse", "Tokyo Olympics") whose rate has not been measured so far. The community has then already embarked on the challenge of categorizing queries based on their temporal characteristics (see [1, 2, 13] for overview).

The second problem relates to ranking search results for queries that contain temporal requirement. *Temporally Diversified Retrieval* (TDR) subtask in Temporalia-2 requires participants to retrieve a set of documents relevant to each of four temporal intent classes for a given topic description. Obviously, both the topical and temporal relevance should be considered to satisfy user search needs.

We note that both TID and TDR subtasks are independent of each other and participating teams could choose one of them or participate in both. Table 1 overviews the schedule of Temporalia-2, while Table 5 shows the participating teams and the subtask they participated. As it can be seen, the participating teams were coming from geographically diverse regions.

[1] http://ntcirtemporalia.github.io/index.html

Table 1: Temporalia-2 important dates.

Date	Event
Feb 28, 2015	NTCIR-12 Kick-off Event
Jul 1, 2015	Final Task Guideline release
Jul 1, 2015	Document Collection release
Jul 7, 2015	Dry run testing data release
Jul 30, 2015	Task Registration Due
Sep 1, 2015	Deadline for dry run submissions
Oct 1, 2015	Return of dry run results
Nov 08, 2015	Release of formal run topics/queries
Dec 10, 2015	Deadline for formal run submissions
Feb 01, 2016	Evaluation results release
Mar 01, 2016	Participant papers due
May 01, 2016	All camera-ready copy due
Jun 07-10, 2016	NTCIR-12 Conference

The remainder of this overview paper is organized as follows. Section 2 introduces the document collection used in the task. Section 3 presents in more detail the tasks at hand, and Section 4 describes the data collecting for evaluation. Section 5 provides system descriptions of participated systems. Section 6 presents the meta analyses conducted for all submitted runs. Finally, the paper is concluded in Section 7.

2. DOCUMENT COLLECTIONS

According to the language difference, multiple document collections are designated for English Temporally Diversified Retrieval (TDR) subtask and Chinese TDR subtask, respectively. Please refer to the Temporalia website for detailed information.

2.1 English Collection

For English TDR subtask, NTCIR-12 Temporalia-2 used the same document corpus as the one used by the Temporal Information Retrieval (TIR) subtask in Temporalia-1. Please refer to Temporalia-1 Overview paper [11] for the detail descriptions of annotations available in this collection.

2.2 Chinese Document Collection

For Chinese TDR subtask, NTCIR-12 Temporalia-2 used SogouCA-2012[2] and SogouT-2012 [3] for dry-run and formal run, respectively. Table 2 shows the basic structure of a Chinese document, where the used tags are a little bit different from the ones in a English documents.

The "host" contains the hostname the document was pulled from, the "date" is the published date of the document, the "url" identifies where the document was pulled from, and finally, the "title" the title of the page. Between the <text> tags, there is the content of the page. This collection also provides three kinds of annotations: sentence splitting, named entities, and time annotations. Each sentence in the content of the page is surrounded by <SE> tags. Each identified named entity is surrounded by <E> tags. The type of the entity is included inside the tag. Specifically, for named entities, there are totally 5 types of tags. Namely, PERSON, LOCATION, ORGANIZATION, GPE (geo-political entity) and MISC (i.e., miscellaneous) [27]. For annotating temporal expressions, a variant of the standard format TIMEX3 used in TempEval task is applied[4]. Furthermore, the script for generating the untagged and tagged collections from SogouCA-2012 for dry-run is provided, as well as the script for extracting the subset from SogouT-2012 for the formal run of the Chinese TDR subtask[5].

Table 2: The structure of a Chinese document.

```
<doc id=***>
<meta-info>
<tag name="host">***</tag>
<tag name="date">***</tag>
<tag name="url">***</tag>
<tag name="title">***</tag>
<tag name="source-encoding">***</tag>
</meta-info>
<text>***</text>
</doc>
```

3. TASKS

3.1 Temporal Intent Disambiguation (TID)

Teams participating in TID subtask were asked to determine distribution of a query over four following classes denoting the types of temporal intent: past, recency, future and atemporal. Below, we give their conceptual definitions.

Past: class characterizing queries about past entities/events, whose search results are not expected to change much along with time passage.

Recency: class characterizing queries about recent entities/events, whose search results are expected to be timely and up-to-date. The information contained in the search results usually changes quickly along with the time passage. Note that this type of query usually refers to events that happened in near past or at the present time. On the contrary, the "past" query category tends to refer to events in relatively distant past.

Future: class characterizing queries about predicted or scheduled events, the search results of which should contain future-related information.

Atemporal: class characterizing queries without any clear temporal intent (i.e., their returned search results are not expected to be related to time neither should change much over time). Navigational queries are considered to be atemporal.

Participants were handed a set of query strings and query submitting dates, and were asked to develop a system to determine the membership degree of each of the query strings to every of the four above-mentioned temporal classes. As this problem may require different kinds of knowledge (e.g., historical information or information on planned events), participants were allowed to use any external resources to complete the subtask as long as the details of external resource usage are described in their reports. Each team was asked to submit a probability distribution over the four temporal classes (past, recency, future, or atemporal) for each query. The performance of submitted runs was then mea-

[2]http://www.sogou.com/labs/dl/ca.html
[3]http://www.sogou.com/labs/dl/t-e.html
[4]http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/timex3guidelines-072009.pdf
[5]http://ntcirtemporalia.github.io/NTCIR-12/collection.html

Table 3: Example queries for the TID subtask (Dry Run) with query submission date of May 1, 2013.

Query	Past	Recency	Future	Atem.
Australian Open	0.091	0.0	0.455	0.455
motorcycle accident june	0.7	0.0	0.3	0.0
NBA Finals	0.1	0.0	0.4	0.5
NBA playoff schedule	0.0	0.2	0.6	0.2
price of oil	0.0	0.9	0.0	0.1
how to lose weight	0.0	0.1	0.0	0.9
time in India	0.0	1.0	0.0	0.0
history of volleyball	1.0	0.0	0.0	0.0

Table 4: Example topics for TDR subtask.

	Junk food health effect
Description	I am concerned about the health effects of junk food in general. I need to know more about their ingredients, impact on health, history, current scientific discoveries and any prognoses.
Past question	When did junk foods become popular?
Recency question	What are the latest studies on the effect of junk foods on our health?
Future question	Will junk food continue to be popular in the future?
Atemporal question	How junk foods are defined?
Search date	29 May 2013 GMT+0:00

sured by (i) the averaged per-class absolute loss and by (ii) the cosine similarity between the ground truth temporal class distribution and the temporal class distribution estimated by the participating systems.

3.2 Temporally Diversified Retrieval (TDR)

TDR subtask required participants to retrieve a set of documents relevant to each of four temporal intent classes for a given topic description. Participants were also asked to return a set of documents that is temporally diversified for the same topic. They received a set of topic descriptions, query issuing time, and indicative search questions for each of temporal classes (Past, Recency, Future, and Atemporal). The objective of the indicative search questions is to show one possible subtopic under a particular temporal class. Participants were asked to develop systems that can produce a total of five search results per topic (Past, Recency, Future, Atemporal, and Diversified).

For the evaluation, the standard Cranfield methodology was used. In particular, a pool of possibly relevant documents was created based on the top-ranked documents from participants' submitted runs. Then, each document in the pool was assessed through online crowdsourcing, and its relevance grade was judged. Given a ranked list generated for a specific temporal subtopic, the system performance was evaluated using nDCG metric [10].

For a diversified ranked list generated to satisfy all possible temporal classes, the performance was evaluated using α-nDCG [5] and D#-nDCG [25].

4. DATASETS

This section describes how we created queries, topics, and answer sets for TID and TDR subtasks.

4.1 TID

4.1.1 English TID

A set of seed temporal expressions (approx. 300) were first collected by the organisers from literature, dictionaries, and query logs. These seed expressions were then submitted to three major commercial search engines, and the alternate queries (typically 10-20 queries) suggested by the search engines were recorded, and finally duplicates were removed. Resulted queries were independently annotated using CrowdFlower[6] for their temporal intent class (i.e., atemporal, past, recency, future). As test queries we have used the formal run data from Temporalia-1 after mapping it to distributions. The organizers picked up 300 queries from results in a way to balance the numbers of queries with given distributions over classes as much as possible.

4.1.2 Chinese TID

The Chinese TID topics were generated through the following steps. First, we randomly extracted 10,000 Chinese search queries from the Chinese query log SogouQ2012. In order to prevent the obtained data from being dominated by the extremely frequent or sparse queries, a query whose collection frequency (w.r.t. SogouQ2012) is not in the interval [5, 50] was not considered. Moreover, noisy queries (e.g., including ones without any Chinese characters or having spelling errors) were filtered out. Second, by sequentially scanning the query list obtained via Step-1, 1,052 Chinese queries were selected by one organizer (a native Chinese speaker). In this process, we mainly focused on queries with temporal intents (without further differentiation of each intent class). The last steps essentially are the same as the ones in selecting English formal run TID topics, e.g., by using CrowdFlower, etc.

In both English and Chinese queries, the answer distribution of temporal classes was estimated by the number of votes given by 10 crowd workers. Crowd workers were asked to vote for the temporal class that they would be most likely to be searching for, given a query.

4.2 TDR

4.2.1 English TDR

The organizers have manually created initial topics and example questions in subtopics (atemporal, past, recency, and future), based on their experience of supervising topic creation workshops at Temporalia-1. The organizers then went through all the initial candidate topics grouping those similar topics, discarding those below expected quality and editing text when appropriate. As a result, a total of 50 topics was selected for the formal run.

All the documents returned by participating systems were combined into pool of TIR formal runs (depth was 20) for relevance judgments. A total of 11,154 documents in the pool were evaluated against each of the four classes. The way the relevance assessments were conducted in Temporalia-2 was similar to those in Temporalia-1. Therefore, please refer to our Temporalia-1 overview paper for the detail [11].

[6] `http://www.crowdflower.com/`

4.2.2 Chinese TDR

For generating the Chinese TDR topics, we asked several Ph.D. candidates and master students from different universities whose research fields are related to this task (all of them are native Chinese speakers) to create candidate topics. We provided them with detailed instructions in order to obtain high quality candidate topics. The instructions include the necessary background information, definitions of each temporal class, the structure of an expected topic, the number of example topics, etc. Given the candidate topics, four native Chinese speakers (including one task-organizer and 3 Ph.D. candidates) checked each topic and made changes if necessary. Finally, a total number of 50 topics were adopted.

5. PARTICIPATED SYSTEMS

This section provides the system descriptions of submitted runs. See Table 5 for the summary of participated teams.

5.1 TID

KDETM team [4] combined a rule-based classifier with weakly supervised classifiers. They defined a set of rules for the rule-based classifier based on the temporal distance, temporal reference, and POS-tag detection, whereas a small set of query with their temporal polarity knowledge were applied to train the weakly supervised classifier. For weakly supervised classifier, the team used the bag-of-words feature and TF-IDF score as a feature weight.

IRISM team's [17] approach to TID subtask relied on supervised machine learning classification technique that involves building models on different standard classifiers based on probabilistic and entropy models from MALLET, a Natural Language Processing tool. They focused on feature engineering that includes temporal & linguistic feature to predict the probability distribution of given temporal classes for search queries. Three runs were submitted based on MaxEnt, Naive Bayes and C4.5 Decision Tree classifiers.

MPII team's system [8] — *TimeSearch* — is a probabilistic framework, that utilizes an unique time model to understand temporal expressions. Building on this model it identifies interesting time intervals for a given keyword query. These time intervals are then also used to rank and diversify documents in a time-sensitive manner.

L3S team [6] used a rule based voting method that comprised of classification rules designed from the dry run queries, which if satisfied contributed a number of votes to a particular class. The rules used various features extracted from the query such as tense of the root verb, distance between the date identified in the query and query issue time and the multinomial distribution of the n-grams extracted from the queries. To understand the implicit temporal nature of a query, candidate years returned from the GTE web service[7] was also used.

Kyoto team [24] took a fully supervised machine learning approach, using features of POS, verb tense and word vectors. They also incorporated knowledge about temporal and holiday expressions. The first submitted run is based on Neural Network (NN), the second run uses the average scores of NN, SVM and SVR, and the third run uses the average scores of NN, SVM, SVR and CNN.

[7] `http://www.ccc.ipt.pt/~ricardo/software.html`

WHUIR team's work [7] for NTCIR-12 TID subtask was mainly based on the research work or research findings of NTCIR-11 TID subtasks with the basic idea of classification. They argued that using features in external retrieved documents from other search engines are not an accurate way for this task. As a result, their system considered nineteen features in total from query itself only and used all the features for SVR in different parameter sets, and chose the best three sets on the dry run data for the formal runs.

DUT-NLP-EN team [18] submitted 3 runs. In RUN1, four groups of features were used including trigger word, word POS, explicit time gap, temporal probability of words. Implicit time gap was added in the form of rule-based time gap in RUN2 and in the form of time-series statistics in RUN3. For all the three runs, logistic regression was used to predict the probability distribution of the four categories.

DUT-NLP-CH team [23] proposed classification method for Chinese TID subtask with explicit features derived from query text and implicit features extracted by analyzing Google Trends data. In formal run, three classifiers have been applied, that is, linear C-Support Vector Classifier, Logistic Regression Classifier and Random Forest Classifier. In posteriori research, the team have compared different models and feature composition and got a better performance.

KGO team [12] developed a deep neural network for the English TID regression problem. This work focused on exploring the temporal information in query entities and constructing the deep neural networks for temporal intent disambiguation. To infer the temporal intents in search queries, explicit temporal features such as the Uppermost Verb Tense and Time Gap features, inexplicit temporal features with respect to the Temporal Named Entities and Holidays, as well as other temporal information in the People names, Time expressions, and word Lemmas have been extracted from Wikipedia and many other knowledge bases on the Internet. The structure and training procedure of the deep neural networks have been carefully selected to avoid overfitting. They thoroughly analyzed the importance in different temporal features and discussed the impact of neural network structures to the TID results.

In the system by *WIS* team [26], features were not only derived from the content of queries considering the issue time, but also extracted from related Wikipedia concepts and their corresponding page views (time-series data). The machine learning approaches leveraged in WIS system are the regression with multiple dependent variables and the probabilistic classification with data extension. In the regression with multiple dependent variables, queries (i.e. features and tags) can be fed to the model directly. In the probabilistic classification with data extension, each query in training data with multi-value tag is extended to several same queries (i.e. same features) with different single-value tags.

GIR team [3] explored the rich temporal information in the labeled and unlabeled search queries. For those search queries with explicit temporal expression or predicate verbs, they extracted the "time gap" and "verb tense" features, separately. For those with no temporal information, the team submitted them to Google search engine to collect temporal indicators for inferring their temporal state. Given the temporal features identified from both search queries and Google results, a semi-supervised linear classifier was then built up to predict the temporal classes for each search query.

HITSZ team [9] merged results of rule based method and

Table 5: Participating Teams.

Team ID	Team Name	Country	TID	TDR
DUTEN	Dalian University of Technology	P.R.C.	En	
DUTCH	Dalian University of Technology	P.R.C.	Ch	
GIR	University of Glasgow	UK	En	
HITSZ	Harbin Institute of Technology, Shenzhen Graduate School	P.R.C.	En, Ch	En
Ho-tm	Japan Advanced Institute of Science and Technology	Japan	En	
IRISM	Indian School of Mines, Dhanbad	India	En	
KDETM	Toyohashi University of Technology	Japan	En	
KGO	University of Tokushima	Japan	En	
kyoto	Kyoto University	Japan	En	
L3S	Leibniz University Hannover	Germany	En	En
MPII	Max Planck Institute for Informatics	Germany	En	En
TUTA1	University of Tokushima	Japan	Ch	
WHUIR	Wuhan University	P.R.C.	En	
WIS	TU Delft	The Netherlands	En	
ORG	Temporalia Organiser	Japan, Spain		En

word intent classes vector based method to estimate temporal intent classes distribution on Chinese queries and English queries. The rule based method was improved from the method the team used in Temporalia-1, which was based on time-sensitive word dictionary, date interval between date in query and query issue time, and the verb tense. The word intent classes vector based method estimated temporal intent classes distribution by normalizing the sum of temporal intent classes vectors of all words in the query.

5.2 TDR

MPII team's probabilistic framework [8] called TimeSearch uses an unique time model to understand temporal expressions and by this to identify interesting time intervals for a given query. These time intervals are then used to rank and diversify documents in a time-sensitive manner.

L3S team [6] used a learning to rank approach that utilized features extracted from verb tenses of sentences related to search queries or not in documents, similarity of the topic, subtopic against the title and content of the document, textual relevance score returned by statistical language model and temporal relevance based on the distribution of time references in the document with temporal intent specific filters. They trained separate learning to rank models for each temporal intent. In order to classify the subtopics into one of the temporal classes, a joint classifier based on verb tense and dictionary features was used. For the diversified results they combined the top 100 documents retrieved for each intent by computing the earth mover's distance between the distribution of time references in the diversified set and the distribution of time references in each of the remaining documents. They then considered the document with the maximum distance to be added to the diversified set.

HITSZ team [9] used TIR system in Temporalia-1 to get ranked documents list for each subtopic. For the temporal diversified ranking, they used all documents in each subtopic result list as candidate documents set for a query topic, and ranked each document in the candidate set based on: the document relevant score in each subtopic list, the temporal intent class of each temporal expression in the document and the subtopic list of previous ranked documents belong to.

Finally, the Temporalia Organizer team (*ORG*) submitted three runs based on the same configuration used in Temporalia-1. We took a round-robin approach to select top ranked documents from each of four class retrieval results, to create diversified runs.

6. META ANALYSES

This section presents the results of meta analyses conducted for all submitted runs in TID and TDR.

6.1 TID

The results based on the mean absolute loss in English TID subtask are given in Figure 1 and the ones by the average cosine similarity are in Figure 2 where runs are ordered by their total scores. The lowest mean absolute loss and maximum average cosine similarity are achieved by *HITSZ-TID-E-1* and *kyoto-TID-E-2*, respectively. The same evaluation results in Chinese TID subtask are shown in Figure 3 and 4. *HITSZ-TID-C-2* had the minimum mean absolute loss and the highest average cosine similarity.

Figures 5 and 6 show breakdown results of average absolute loss and mean cosine similarity across four temporal classes in English TID subtask. The results are in the form of stacked bars making it possible to identify strong or weak temporal classes of each run. For example, in Figure 5, *HITSZ-TID-E-1* achieved the minimum average absolute loss value in Past class compared to other runs. On the other hand, the lowest score of Future class is achieved by both *KGO-TID-E-1* and *KGO-TID-E-2*. The breakdown results in the case of Chinese TID subtask are shown in Figures 7 and 8. Looking at Figure 8, we see that *HITSZ-TID-C-2* had the maximum average cosine similarities of Past and Atemporal classes. *DUT-NLP-CH-TID-C-2* achieved the highest score of Recency and Future classes in all runs.

To determine easy and difficult temporal classes, we calculated the average absolute values per temporal class over all runs, showed in Table 6. In both subtasks of English and Chinese, Atemporal class was found to be most difficult.

Next, we analyzed the relationships between query complexity of temporal distribution and performance. Each query was classified into four categories based on its number of classes with zero probability value in ground truth distri-

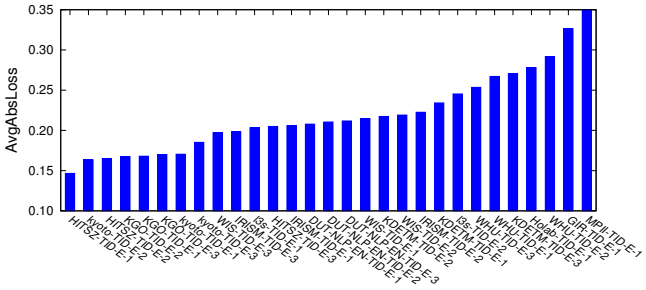

Figure 1: System ranking based on average loss for English TID.

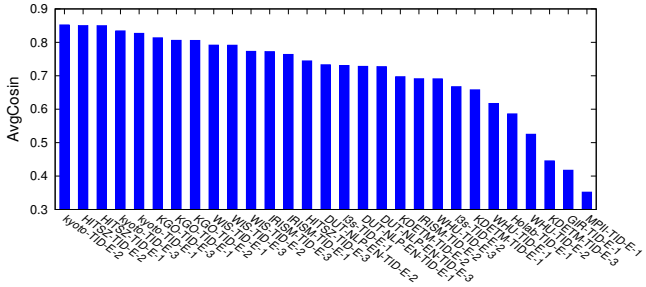

Figure 2: System ranking based on cosine similarity for English TID.

Table 6: Average absolute loss of each temporal class over all systems.

	English		Chinese	
Class	Mean	SD	Mean	SD
Past	0.180	0.071	0.159	0.043
Recency	0.170	0.031	0.152	0.014
Future	0.190	0.052	0.126	0.021
Atemporal	0.293	0.064	0.252	0.041

butions: one, two, three, and four non-zero(s) classes. For example, a query *Australian Open* in Table 3 was classified to belong to the category of three non-zeros because it had three non-zero probabilities, except the Recency class. Figures 9 and 10 show breakdown results of average absolute loss across four categories based on the number of non-zero probabilities in English and Chinese, respectively. In English TID subtask, *HITSZ-E-1* achieved the minimum average absolute loss in the categories of one and two non-zero(s) probabilities. On the other hand, *WIS-E-2* and *WHU-E-3* had the lowest score in the four non-zeros category.

We next calculated the Pearson's correlation coefficients between these categories over all runs in English TID subtask (see Table 7). The highest correlation coefficients was between the categories of three and four non-zero probabilities at 0.903. The second highest one was between one and two non-zero(s) probabilities' categories at 0.838. The lowest one was between the one non-zero's category and the four non-zeros' category at 0.110. This suggests that runs had high correlation of performance between simple queries, i.e., one and two non-zero(s) categories, or between complex queries, i.e., three and four non-zeros categories. On the other hand, there is basically no correlation between the simplest and the most complex categories.

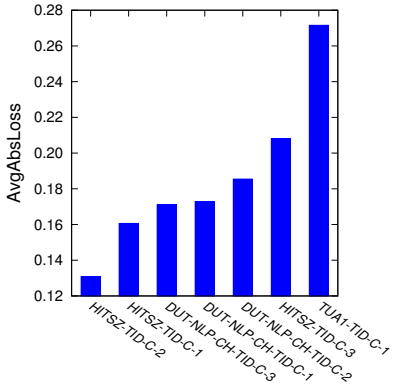

Figure 3: System ranking based on average loss for Chinese TID.

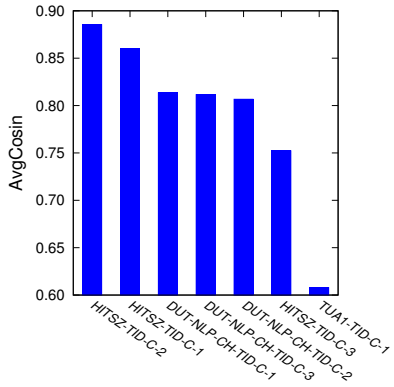

Figure 4: System ranking based on cosine similarity for Chinese TID.

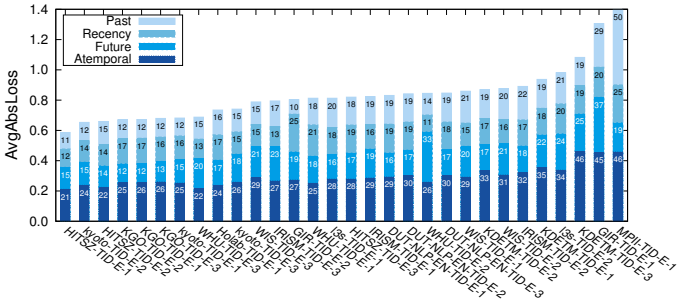

Figure 5: Per class performance of systems in English TID subtask using absolute loss.

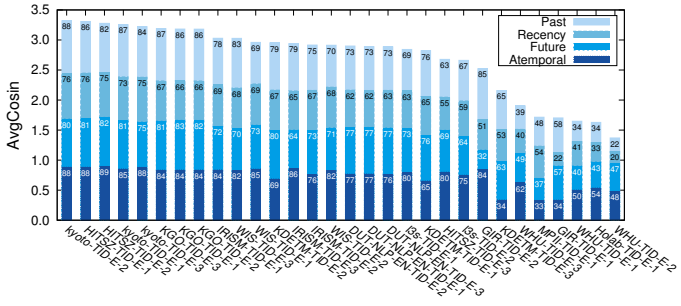

Figure 6: Per class performance of systems in English TID subtask using cosine similarity measure.

6.2 TDR

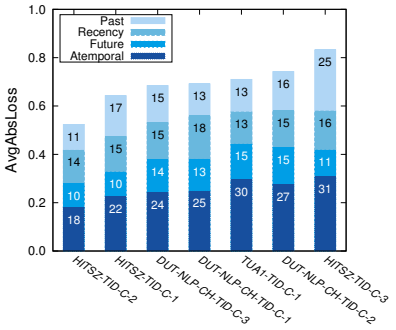

Figure 7: Per class performance of systems in Chinese TID subtask using absolute loss.

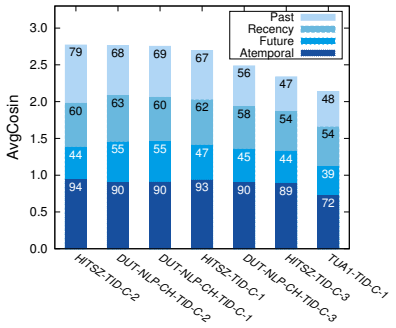

Figure 8: Per class performance of systems in Chinese TID subtask using cosine similarity measure.

Table 7: Pearson's correlation coefficients between query categories with given numbers of non-zero probabilities over all runs in English TID subtask.

	Two	Three	Four
One	0.838	0.383	0.110
Two		0.755	0.518
Three			0.903

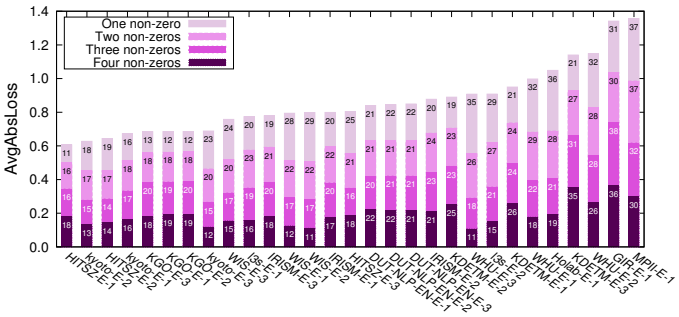

Figure 9: Performance of systems in English TID based on query complexity using absolute loss.

The results based on the nDCG@20 in TDR subtask are given in Figure 11 which is shown as stacked performance across four temporal classes. Runs are ordered by their total scores. In Past, Recency, and Atemporal classes, *HITSZ-TDR-E-3* achieved the highest nDCG@20 score in all runs. *I3s-TDR-E-1* had the maximum one in Future class. The TDR subtask results for diversified ranking by D#-nDCG@20 are shown in Figure 12. Same as in the case of nDCG@20 results, runs are odered by their scores. The maximum D#-nDCG@20 was achieved by *org-TDR-E-3*.

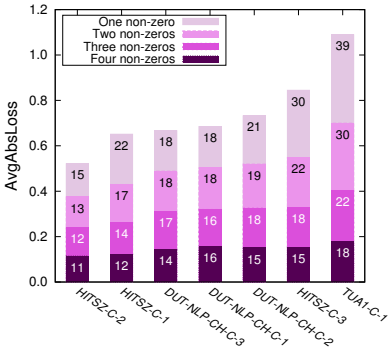

Figure 10: Performance of systems in Chinese TID based on query complexity using absolute loss.

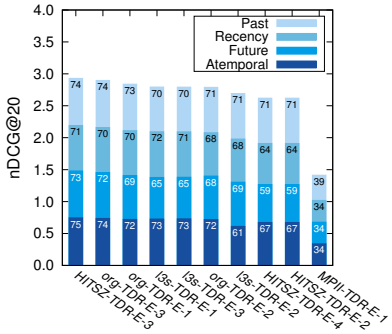

Figure 11: TDR subtask results for each temporal class by nDCG@20.

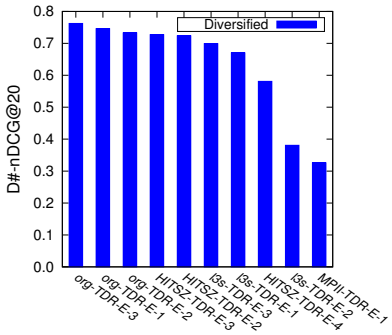

Figure 12: TDR subtask results for diversified ranking by D#-nDCG@20.

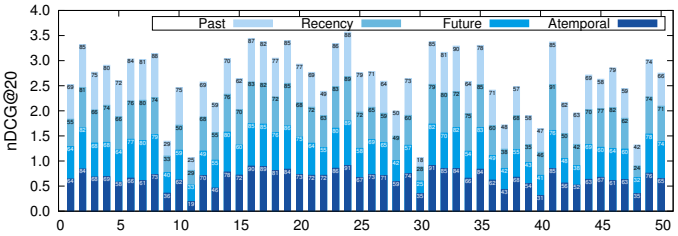

Figure 13: TDR topics by nDCG@20.

Figure 13 shows a topic breakdown of nDCG@20 scores across four classes which are helpful to identify easy or difficult topics in the formal run dataset. The stacked bar chart suggests that Topic 11 was particularly difficult, followed by a secondary group of Topic 9, 28, 30, 37, and 48.

7. CONCLUSIONS

This paper presented the 2nd NTCIR Temporal Information Access (Temporalia-2) Task. Our test collections were designed to offer an opportunity to evaluate temporal-aware search technologies across four temporal classes (atemporal, past, recency, and future) in a structured way. Two subtasks were devised to advance temporal query intent disambiguation and temporal document ranking technologies. Both subtasks had a respectable number of queries and topics for system evaluation and user studies. With 15 participating teams, Temporalia-2 was able to set the foundation of temporal information access technology evaluation.

8. ACKNOWLEDGMENTS

This work was supported in part by MEXT Grant-in-Aid for Young Scientists B (#24700239 and #22700096), and by the JST research promotion program Sakigake. The authors also thank the NTCIR project at NII.

9. REFERENCES

[1] O. Alonso, R. Baeza-yates, J. Strotgen, and M. Gertz. Temporal Information Retrieval: Challenges and Opportunities. In: *TempWeb 2011*, 1–8, 2011.

[2] R. Campos, G. Dias, A.M. Jorge, and A. Jatowt. Survey of Temporal Information Retrieval and Related Applications. In: *ACM Comput. Surv.*, 15:1–15:41, 2014.

[3] L. Chen, H. Yu, F. Yuan and J. M Jose. GIR at the NTCIR-12 TID Task. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[4] A. N. Chy, M. Z. Ullah, M. Shajalal and M. Aono. KDETM at NTCIR-12 Temporalia Task: Combining a Rule-based Classifier with Weakly Supervised Learning for Temporal Query Intent Disambiguation. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[5] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In: *Proceedings of the 31st SIGIR*, 659–666, 2008.

[6] Z. T. Fernando, J. Singh and A. Anand. L3S at the NTCIR-12 Temporal Information Access (Temporalia-2) Task. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[7] S. Gui and W. Lu. WHUIR at the NTCIR-12 Temporal Intent Disambiguation Task. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[8] D. Gupta and K. Berberich. A Probabilistic Framework for Time-Sensitive Search: MPII at the NTCIR-12 Temporalia-2 Task. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[9] Y. Hou, J. Xu, X. Wang, C. Tan, and Q. Chen. HITSZ-ICRC at NTCIR-12 Temporal Information Access Task. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[10] K. Järvelin, and J. Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. In: *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[11] H. Joho, A. Jatowt, R. Blanco, H. Naka and S. Yamamoto. Overview of NTCIR-11 Temporal Information Access (Temporalia) Task. In: *NTCIR-11*, 2014.

[12] X. Kang, Y. Wu and F. Ren. KGO at the NTCIR-12 Temporalia Task: Exploring Temporal Information in Search Queries. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[13] N. Kanhabua, R. Blanco and K. Nørväg. Temporal Information Retrieval. In: *Foundations and Trends in Information Retrieval*, 9(2):91–208, 2014.

[14] M. P. Kato, and Kishida, K. Overview of NTCIR-12. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[15] G. Kazai, J. Kamps, and N. Milic-Frayling. An Analysis of Human Factors and Label Accuracy in Crowdsourcing Relevance Judgments. *Inf. Retr.*, 16(2):138–178, 2013.

[16] G. Kazai, S. Masood, and M. Lalmas. A Study of the Assessment of Relevance for the Inex'02 Test Collection. In: *ECIR 2004*, 296–310, 2004.

[17] J. Kumar, S. S. Prasad and S. Pal. IRISM @ NTCIR-12 Temporalia Task: Experiments with MaxEnt, Naive Bayes and Decision Tree classifiers. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[18] D. Li, X. Liu, Y. Zhang, D. Huang and J. Cao. Using Time-Series for Temporal Intent Disambiguation in NTCIR-12 Temporalia. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[19] N. Liu, M. He, C. Li, X. Kang and F. Ren. TUTA1 at the NTCIR-12 Temporalia Task. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[20] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, and H. Zaragoza. Searching through Time in the New York Times. In: *HCIR 2010*, 41–44, 2010.

[21] G. Moharasan and T. B. Ho. NTCIR-12: Temporal Intent Disambiguation Subtask: Using Machine Learning Approach to Predict Temporal Classes. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[22] S. Nunes, C. Ribeiro, and G. David. Use of Temporal Expressions in Web Search. In: *ECIR 2008*, 580–584, 2008.

[23] J. Pei, D. Huang, J. Ma, D. Song and L. Sang. DUT-NLP-CH @ Temporal Intent Disambiguation Subtask. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[24] T. Sakaguchi and S. Kurohashi. KYOTO at the NTCIR-12 Temporalia Task: Machine Learning Approach for Temporal Intent Disambiguation Subtask. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[25] T. Sakai and R. Song Evaluating Diversified Search Results Using Per-intent Graded Relevance. In: *Proceedings of the 34th SIGIR*, 1043–1052, 2011.

[26] Y. Zhao and C. Hauff. WIS @ the NTCIR-12 Temporalia-2 Task. In: *Proceedings of the 12th NTCIR Conference*, 2016.

[27] Finkel, Jenny Rose and Manning, Christopher D. Joint Parsing and Named Entity Recognition. In: *NAACL 2009*, 326-334, 2009.