Overview of NTCIR-11 Temporal Information Access (Temporalia) 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
Kyoto University, Yoshida-Honmachi, Sakyo-ku, 606-8501 Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

Roi Blanco
Yahoo Labs, Barcelona, Spain
roi@yahoo-inc.com

Hajime Naka
Graduate School of Library, Information, and Media Studies, University of Tsukuba, Japan.
naka@slis.tsukuba.ac.jp

Shuhei Yamamoto
Graduate School of Library, Information, and Media Studies, University of Tsukuba, Japan.
yamahei@ce.slis.tsukuba.ac.jp

ABSTRACT

This paper describes the overview of NTCIR-11 Temporal Information Access (Temporalia) task. This pilot task aims to foster research in temporal aspects of information retrieval and search. Temporalia is composed of two subtasks: Temporal Query Intent Classification (TQIC) and Temporal Information Retrieval (TIR) subtask. TQIC attracted 6 teams which submitted a total of 17 runs, while 6 teams took part in TIR proposing 18 runs. In this paper we describe both subtasks, datasets, evaluation methods and results of meta analyses.

Subtasks

Temporal Query Intent Classification (TQIC)
Temporal Information Retrieval (TIR)

Keywords

temporal information retrieval, evaluation, temporal information access, test collections

1. INTRODUCTION

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

The importance of time in search scenarios has been known since long. Joho et al [3] conducted a behavioral study to shed light on some temporal aspects of web search experience. Importantly, the results indicated that there is a high demand for *recent* retrieved information among other temporal information needs, although a good number of users also demand for efficient retrieval of past and future information. Similarly, a good proportion of information needs can be related to seasonal interests and continuous interests. The study also provides evidence that context such as search devices and locations could be exploited to elicit temporal aspects of information needs such as seasonal interest and target time of information. Furthermore, seasonal interests, technicality of information needs, target time of information, re-finding behaviour, and freshness of information appear to all interplay to formulate temporal aspects of web searches.

As the pilot task we focused on two major search subproblems: query intent understanding and document ranking considering their temporal aspects. The first one called *Temporal Query Intent Classification* (TQIC) asks participants to classify provided queries into temporal classes following the intuitive time understanding: past-related, recency-related and future-related. For comparison we have also added atemporal class that characterizes queries without any underlying temporal intent.

This task should be useful challenge for any research that aims to recognize underlying temporal aspects of queries. With successful solutions, search engines could then treat temporal queries accordingly to their underlying temporal classes. According to the study performed on the AOL query dataset [7], about 1.5% of queries are explicit temporal queries, that is, they contain some temporal expressions. Examples of such queries are: "Germany 1920s", "Olympics 2012" or "top movies 1990s". 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., "Einstein childhood", "WWII major battles", "USA debt size", "Rio de Janeiro Olympics") whose rate has not been measured so far. The community has already embarked on the challenge of categorizing queries based on their temporalities.

The second problem relates to ranking search results for queries that contain a strong temporal requirement. The *Temporal Information Retrieval* (TIR) subtask in Tempo-

[1] https://sites.google.com/site/ntcirtemporalia/

ralia requires users to output documents within the prepared collection for different temporal query classes. Obviously, both the topical and temporal relevance need to be considered to satisfy user search needs. We note that both TQIC and TIR are independent of each other and teams could choose one of the tasks or participate in both.

The remainder of this overview paper is organized as follows. Section 2 introduces the data collection compiled for the task. Section 3 presents in more detail the tasks at hand, and Section 4 describes the data collecting for evaluation. Section 5 presents the main results for the subtasks, as well as a description of different methods employed by different participants. Section **??** describes some related existing competitions and datasets, and the paper is concluded in Section 6.

2. DATA COLLECTION

NTCIR-11 Temporalia used a document corpus called "LivingKnowledge news and blogs annotated sub-collection", constructed by the LivingKnowledge project and distributed by the Internet Memory Foundation [6]. The collection is approximately 20GB large when uncompressed and over 5GB large when zipped. It spans from May 2011 to March 2013 and contains around 3.8M documents collected from about 1500 different blogs and news sources. The data is split into 970 files, named after the date of that day and its sources (there might be more than one file per day).

```
<doc id=lk-20130223040102_592>
<meta-info>
<tag name="host">www.somesite.de</tag>
<tag name="date">2013-02-22</tag>
<tag
name="url">http://www.somesite.de/
    international/bus iness/eu-widens-libor
    -scandal-investigation-and- threatens-
    heavy-fines-a-884948.html#ref=rss</tag>
     <tag name="sourcerss">
http://www. .de/international/index.rss</
    tag >
<tag name="title">EU Widens LIBOR Scandal
    Investigation and Threatens Heavy Fines
    </tag> </meta-info>
<text>
\caption{Example document from the
    LivingKnowledge collection}
```

Each file contains a number of text documents. For each document the following information is available (see document excerpt shown above). The <doc id> refers to a unique document identifier in the collection. The host contains the hostname the text was pulled from, date is the publishing date of the document, url is the URL the text was pulled from, sourcerss is the RSS address that was accessed to retrieve the page, and finally, title is the title of the page. Between the <text> tags, there is the content of the page. This collection has been automatically tagged with different semantic annotations (see [6] for a more detailed description of how the annotations were produced). In particular, we provide three kinds of annotations: sentence splitting, named entities, and time expressions. Sentences are surrounded by `<SE>` tags whereas named entities are surrounded by `<E>` tags. Entity types are included inside the tag, for instance `<E type="E:ORGANIZATION:CORPORATION">YouWalkAway .com</E>`. Time expressions are surrounded by `<T>` tags. For example, `<T val="2012">the end of 2012</T>` contains a val element referring to the estimated point in time the annotation is referring to. Time expressions in text are of course directly useful for any time-related search or mining tasks. Entities, on the other hand, can be used indirectly via entity linking procedure with external databases such as Wikipedia that are rich in metadata including time-related aspects or with timestamped external document collections. Both time expressions and named entities should also constitute good features for procedures manipulating the collection on event level (e.g., event detection or event linking) should they be required.

3. TASKS

3.1 Temporal Query Intent Classification

Teams participating in TQIC subtask were asked to classify the query into one of the following classes: past, recency, future and atemporal. Below we define conceptual definitions of individual query classes.

Past query: query about past events, whose search results are not expected to change much along with time passage.

Recency query: query about recent things, 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 query: query about predicted or scheduled events, the search results of which should contain future-related information.

Atemporal query: query without any clear temporal intent (i.e., its returned search results are not expected to be related to time neither 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 classify each of the query strings to one of the four above-mentioned temporal classes. As this problem rather requires different kinds of knowledge (e.g., historical information or information on planned events), the participants were allowed to use any external resources to complete the TQIC subtask as long as the details of external resource usage are described in their reports. Each participating team was asked to submit a temporal class (past, recency, future, or atemporal) for each one of the queries. The performance of submitted runs was measured by the number of queries with correct temporal classes divided by the total number of queries. Table 1 shows examples of queries in TQIC subtask.

3.2 Temporal Information Retrieval

TIR subtask asks participants to retrieve a set of documents in response to a search topic that incorporates time factor. In addition to a typical search topic description (i.e., title, description, and sub topics), TIR search topic description contains also a query submitting date. This subtask requires to index the document collection with any standard information retrieval toolkit. Participants were asked to submit the top 100 documents for each of temporal questions per topic (i.e., top 100 documents for past question,

Table 1: Example queries for the TQIC subtask (Dry Run)

Query class	Query example
Past	price hike in bangladesh 2008
Past	Who Was Martin Luther
Past	when did the titanic sink
Past	Yuri Gagarin Cause of Death
Past	History of Coca-Cola
Recency	apple stock price
Recency	Number of Millionaires in USA
Recency	time in london
Recency	Trendy Plus Size Clothing
Recency	Did the Pirates Win Today
Future	2013 MLB Playoff Schedule
Future	release date for ios7
Future	College Baseball Regional Projections
Future	disney prices 2014
Future	long term weather forecast
Atemporal	blood pressure monitor
Atemporal	distance from earth to sun
Atemporal	how to start a conversation
Atemporal	New York Times
Atemporal	lose weight quickly

Table 2: Example topics for the TIR subtask (Dry Run)

	Girl with the Dragon Tattoo
Description	I've recently watched a film called Girl with the Dragon Tattoo, and really liked it. Therefore, I would like to gather information about the movie.
Past question	How did the casting of the film develop?
Recency question	What did the recent reviews say about the film?
Future question	Is there any plan about its sequel?
Atemporal question	What are the names of main actors and actresses of the film?
Search date	28 Feb 2013 GMT+0:00

another 100 for recency question, etc.). The retrieval effectiveness was evaluated by the precision at 20 for each of the temporal questions. Similarly to the TQIC subtask, the results section provides the breakdown of the performance across temporal questions.

4. DATASETS

This section describes how we created queries, topics, and answer sets for TQIC and TIR subtasks at NTCIR-11 Temporalia.

4.1 TQIC

4.1.1 Query creation

A set of seed temporal expressions (approx. 300) were first collected by the organisers from literatures, 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 (approx. 1.5K queries) were independently annotated by three of the organisers for their temporal intent class (i.e., atemporal, past, recency, future).

In dry runs, we selected 100 queries where the agreement of annotations was high and thus ambiguity of their intension was low. In formal runs, we selected a total of 300 new queries where 75% had high agreement and low ambiguity and 25% had medium agreement and higher ambiguity.

4.2 TIR

4.2.1 Topic creation

A series of workshops was held to create candidate topics for both dry runs and formal runs. Participants of the workshops were international students from the University of Tsukuba with a high level of English fluency. At the workshops, participants were asked to use an Apache Solr-based search interface developed by the organisers to access the LivingKnowledge news and blogs annotated sub-collection. Participants were asked to explore the topics they were genuinely interested first, and asked to find any documents that can be seen as relevant to their topics. When they were successful in identifying several relevant documents, they were asked to expand their main topic from the perspective of four temporal subtopics (atemporal, past, recency, and future). Again, they were asked to identify several relevant documents for each of subtopics. They were allowed to refer to any external resources (e.g., search engines, Wikipedia) during topic creation.

In the formal run topic creation, over 80 candidate topics (each has four subtopics) were suggested with varying degrees of completeness. The organisers went through all candidate topics and grouped those similar topics and discarded those that were below expected quality. As a result, a total of 50 topics (200 subtopics) was selected for the formal run. The organisers further went throughout the descriptions of all topics and edited texts where appropriate. Table 2 shows examples of queries in TIR subtask.

4.2.2 Relevance assessments

Over 36K documents were identified in our pool of TIR formal runs (depth was 20) for relevance judgements. From our experience of relevance judgements of dry runs, we judged that it was infeasible to complete the relevance assessments for formal runs using our limited resources. Therefore, we had a combination of workshops and crowdsourcing in formal runs. In another series of workshops, participants (not necessarily the same people as topic creators) were asked to read the formal run topic descriptions carefully, and assess the relevance of retrieved documents. For each assigned subtopic, they were asked to identify at least one highly relevant and one irrelevant document. They were also asked to note the relevant text from original documents in the case of highly relevant documents. The relevance of these documents were verified by a third person during the workshop to ensure their reliability.

The documents initially identified by the workshop participants were then used as "test questions" of crowdsourcing jobs. Test questions are questions that crowdsourcing workers had to pass to take part in our relevance assessment jobs.

We used CrowdFlower[2] to run relevance judgements. Our configuration of crowdsorcing is based on common settings used by various IR evaluations (e.g., [4]).

- Each task had five documents to judge
- Ten cents were paid to one task
- Each task had 120 seconds minimum work time
- Each document had at least three judgements

We had several iterations of revising job instructions and relevance criteria before running all formal run subtopics. We tested both detailed instructions and simple instructions, but received mixed responses from workers. Also, detailed instructions were taking too long to complete relevance assessments. After several iterations, we decided to use the following three-levels relevance criteria (c.f., [5], [8]).

Not Relevant The web page does not contain any information to answer the search question.

Highly Relevant The web page discusses the answer of the search question exhaustively. In case of a multi-faceted search question, all or most sub-themes or viewpoints are covered. Typical extent: several text paragraphs, at least 4 sentences or facts.

Relevant The web page contains some information to answer the question, but the presentation is not exhaustive. In case of a multi-faceted search question, only some of the sub-themes or viewpoints are covered. Typical extent: one text paragraph, 1-3 sentences or facts.

During the dry run assessments where the organisers were able to give a hands-on tutorial to participants about how to judge relevance, we used two dimensional scoring system (i.e., one for topical relevance, and another for temporal relevance). However, it was difficult to convey the accurate criteria to crowdsourcing workers who tended to have a short-span of interest. Therefore, we decided to have a uniform scoring system given above. An advantage of this criteria was that we were able to use the identical criteria to temporal subtopics as well as atemporal subtopics, making their scores comparable. Further revision of the criteria is open for discussion.

The distribution of three level judgements is shown in Table 5, where Highly Relevant, Relevant, Not Relevant are denoted as $L2$, $L1$, $L0$, respectively. As can be seen, overall, nearly half of documents in the pool was judged as Not Relevant, 28% as Relevant, and 22% as Highly Relevant. The proportion of three levels slightly varies across temporal classes. Official evaluation metrics adapted by NTCIR-11 Temporalia TIR subtask were Precision at 20 (P@20) and nDCG at 20 (nDCG@20). We also report Q-measure [17] at 20 (Q@20) as reference.

5. META ANALYSES

The tasks attracted a total of 36 runs submitted by 8 teams from 7 countries (3 teams participated in both TQIC and TIR subtasks). Table 3 shows the participating teams and the subtask they participated. As can be seen, the participated teams were geographically diverse, ranging from Asia to Europe and North America. This section presents the results of meta-analysis of submitted runs.

[2] http://www.crowdflower.com/

5.1 TQIC

TQIC subtask had 17 runs from 6 teams. TQIC in formal run had 300 queries to classify. The results of classification accuracy across four temporal classes are shown in Table 6, where runs are ordered by the overall accuracy score. As can be seen from the mean value (2nd bottom row in the table), classification of recency queries were found to be most difficult with 56%, and past queries were the easiest with 73%. Another overall trend was that no single run was effective across the four temporal classes. For example, *TUTA1-TQIC-RUN-1* had the highest overall score of 74%, but accuracy of Recency class was low. On the other hand, *TQIC-HULTECH-Run1* and *TQIC-HULTECH-Run2* had the highest accuracy in Recency class with varied performance in other classes.

Table 7 shows a confusion matrix between answer classes and estimated classes. The table indicates that 1) atemporal queries are likely to be confused as either recency or past queries (16.7% and 9.6%, respectively), 2) past queries are likely to be confused as atemporal queries (13.1%), 3) recency queries are likely to be confused as future (28.2%) or atemporal (13.5%) queries, and finally, 4) future queries are tend to be confused as recency queries (25.9%).

Figure 1 shows Pearson's correlation of classification accuracy between four temporal classes. As can be seen, many temporal class pairs have a weak correlation between $r = -0.180$ and $r = 0.316$. The largest correlation was observed between Atemporal class and Recency class with $r = -0.687$. This suggests that those runs who performed well on atemporal queries tended to have a poor performance on recency queries. The second largest correlation was between Recency class and Future class with $r = -0.463$. Again, this suggests that those runs who performed well on future queries tended to have a poor performance on recency queries.

Participating systems.

HITSZ team [14] approach to TQIC subtask relied on extracting various linguistic level features from queries and deriving features from search results received after issuing the queries to Bing[3] search engine. They then applied rule based method and combined the classification results produced by rule based method and multi-classifier voting. The rule based method used distance between the date in user query and query issue time, time-sensitive word disctionary and the combination of date distance and verb tense. Classifier features contained the following groups: N-gram words of query, POS n-grams, named entities, normalized date, date distance and special words from the time-sensitive word dictionary.

Team HULTECH [16] experimented with an ensemble learning paradigm to reduce bias by combining multiple classifiers instead of a single one. They considered eleven types of features from three different information sources: TempoWordNet[4], Web snippets results and the query itself seen as a sentence. Different feature subsets were used for submitted runs. Their system reached average results but managed to outperform other participants for the temporal class Recency in terms of F-measure.

TUTA1 team [13] used semi-supervised and supervised

[3] http://www.bing.com/

[4] https://tempowordnet.greyc.fr/

Table 3: Participating Teams.

Team ID	Team Name	Country	TQIC	TIR
HITSZ	Graduate School of Harbin, Institute of Technology at Shenzhen	P.R.C.	yes	yes
HULTECH	University of Caen	France	yes	no
TUTA1	The University of Tokushima	Japan	yes	yes
Andd7	Dhirubhai Ambani Institute of Information and Communication Technology	India	yes	yes
MPII	Max Planck Institute for Informatics	Germany	yes	no
UniMAN	The University of Manchester	UK	yes	no
BRKLY	U.C. Berkeley	USA	no	yes
OSKAT	Sato Laboratory, Osaka Kyoiku University	Japan	no	yes
ORG	Temporalia Organiser	Japan, Spain	no	yes

linear classifiers to predict temporal classes. The features they utilized are time gap features, verb tense features and lemmas, and named entities. The team's approach made use also of AOL 500K query session dataset to provide more training data. Four runs have been proposed: two using Logistic Regression Classifiers with different parameter settings, one with SVMlin classifier using additional data from AOL dataset and one with Logistic Regression Classifier with only lemma and named entity features.

Andd7 team [12] has used classifiers that employ following feature groups: bag of words, query length, difference of query issue time and temporal expression in query and verbs in query. Three runs have been applied, using either SVM, Naive Bayes Classifier or agreement decision between the two classifiers. The results reveal terms specific for particular temporal classes like will, forecast for future class or today, current for recency class.

MPII's team [10] method is based on classification with range of features derived by applying POS tagger, DMOZ directory[5] information, publication dates and the content of document collections used for finding query-time associations. Experiments were done using Naive Bayes with extensive evaluation of contributions coming from diverse feature groups. Three runs are submitted with feature sets picked by simulated annealing.

Team UniMan [9] proposed classification approach for TQIC subtask that specially engineered features in a way to minimize feature scarcity. Nineteen feature groups were used where some were computed from query, some derived from submission date and some from both. For feature computation the team used POS tagger, TempoWordnet and comparison with Wikipedia titles. The authors constructed three runs with different feature sets, and the minimal feature set was found to work best.

5.2 TIR

TIR subtask had 15 runs from 5 participant teams. 3 runs were added to the pool by the organizer team. Therefore, we had a total of 18 runs. TIR in formal run had a total 50 main topics and each had 4 subtopics, making it a total of 200 subtopics. The average retrieval performance over the 200 topics is shown in Table 8. As can be seen, one of the baseline runs submitted by the organizer team had the highest score in both nDCG@20 and P@20. However, the performance varied across four temporal classes. For example, *tir-OKSAT-TF01* obtained the highest score for atemporal subtopics (see Table 9), while *tir-HITSZ-LTRNC2* achieved the highest score for past subtopics (See Table 10). *tir-org-sqd* achieved the highest score for recency and future subtopics (See Table 11 and 12). Like TQIC results, this suggests that it is difficult to generate an effective ranking for all temporal classes based on a single IR strategy.

Finally, Figure 2 shows a topic breakdown of nDCG@20 scores across four temporal classes. The bottom figure in Figure 2 shows a stacked performance of all classes which are helpful to identify easy or difficult topics in the formal run dataset. The stacked bar chart suggests that Topic 8 was particularly difficult, followed by a secondary group of Topic 4, 9, 35, 40, 42, and 44. On the other hand, a group of Topic 16, 19, 24, 28 and 32 was found to be easy. Topic 8 was about *English as a second language*, and subtopics asked for the definition of second languages, past learning methods in a particular time period (before 2000), latest technologies available, and future learning methods. The number of relevant documents on this topic was lower than other topics, and this could be one factor for poor performance.

Another insight we can gain from the stacked bar chart is that the performance can vary across temporal subtopics within main topics. For example, Topic 19 is one of the best performing topics among 50 topics. However, the average performance of past subtopics is very low. Similarly, some main topics have difficult recency subtopics, while others have difficult future subtopics.

Participating systems.

The organiser team provided three runs as a baseline. Our systems are out of box of Apache Solr with BM25 weighting scheme. We did not use any temporal annotations available in the document collection. Only difference among runs was the use of fields in topic descriptions for query generation. The *tir-org-t* used only the title field of the topic description. This means that this run returned exactly the same results to all four subtopics. The *tir-org-sq* used the title and search question fields to generate queries, while *tir-org-sqd* used the title, search question, and description fields.

HITSZ team [14] approached TIR task by using learning to rank method and BM25 search model. They first returned the set of top-ranked documents using BM25 ranking for each subtopic and then classified any occurring time expressions as future, past or recency with respect to the query issuing time so as to compute temporal relevance scores of documents. Since the class information of subtopics was not

[5] `http://www.dmoz.org/`

supposed to be used, the team employed their TQIC classification method for estimating subtopic class. The first and the second run were returned by methods that used the linear combinations of the content relevancy and temporal relevancy. The third run used learning to rank algorithm on a range of features related to topical and temporal relevancy.

TUTA1 team [13] also used learning to rank technique. They used features derived from temporal expressions and verb tenses in documents which were also differentiated as whether they occur in sentences related to search queries or not. Sentence relevance estimation was bound to the occurrence of same nouns as ones in search topics. The first and second run were based on learning to rank model trained over entire dataset or class-specific datasets, respectively, while the third run used learning model trained on entire dataset with TFIDF retrieval model used.

Andd7 team [12] submitted results of three runs. The first submitted system used nouns, verbs and adjectives derived from subtopic content as queries against the title fields of documents. The second was different as it used both the words from the content of topic to find documents by matching their title fields and words from content of subtopics to return results by matching document content. In order to construct the third system the team classified subtopics into one of temporal classes based on words selected from queries used in TQIC subtask. The last run used the combination of the topic relevance and temporal relevance that was computed for different temporal classes of subtopics by comparing mean dates in documents with their timestamps.

BRKLY team [11] offered baseline text retrieval system that does not explicitly use temporal features. The team used bag-of-words representation and retrieval model based on logistic regression model of probabilistic IR. Logistic regression was trained on TREC2[6] data. As a suplement to the base retrieval model the team implemented also the blind relevance feedback that uses probabilistic term relevance weighting formula. For the blind relevance feedback they used top ten terms from ten top-ranked documents.

OKSAT team [15] submitted three runs, where the first one returned search results using words from topics and subtopics after removing stopwords. The second run was created by issuing queries extended with the most common words derived from Wikipedia[7] and Web search results. The search results used for query extension were obtained by issuing queries used in the first run. The third run was based on issuing plural sets of terms.

6. CONCLUSIONS AND FUTURE DIRECTIONS

This paper presented the first test collection of NTCIR Temporal Information Access (Temporalia) Task. Our test collection was 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 classification technologies and temporal document ranking technologies. Both subtasks had a respectable number of queries and topics for system evaluation and user studies. With the participation of 8 teams, NTCIR-11 Temporalia was able to set the foundation of temporal information access technology evaluation.

There are several directions for our test collections. First, we can extend TQIC subtask by including the detection of temporal ambiguity. Some queries are temporally ambiguous while others are reasonably fixed. Therefore, one can ask systems to detect whether a given query should diversify search results with relevant temporal classes. Although NTCIR-11 collection has a disjoint relationship between TQIC and TIR subtasks, a closer connection between the two subtasks is also of our interest in the future.

Another direction is to extend TIR subtask with multi-document summarization. For example, a typical Wikipedia page has a historical order of people, organisations, or events. One can ask systems to create such a chronological summary of a given query or topic, but including future information too. Visualisation of temporal search results is also of potential interest.

[6] http://trec.nist.gov/pubs/trec2/t2_proceedings.html

[7] https://www.wikipedia.org/

Table 4: Temporalia important dates.

Date	Event
Sep 02, 2013	NTCIR-11 Kick-off Event
Jan 05, 2014	Document collection release
Jan 20, 2014	Task Registration Due
Jan 23, 2014	Release of dry run topics/queries
Mar 31, 2014	Deadline for dry run submissions
Apr 15, 2014	Return of dry run results
May 08, 2014	Release of formal run topics/queries
Jun 30, 2014	Deadline for formal run submissions
Aug 01, 2014	Evaluation results release
Aug 21, 2014	Early overview draft release
Sep 15, 2014	Participant papers due
Nov 01, 2014	All camera-ready copy due
Dec 09-12, 2014	NTCIR-11 Conference

Table 8: Mean retrieval performance over 200 topics. The highest value in each column is shown in bold.

run name	Q@20	P@20	nDCG@20
tir-org-sqd	**0.416**	**0.618**	**0.488**
tir-HITSZ-LTRNC2	0.410	0.602	0.477
tir-OKSAT-TF01	0.383	0.583	0.457
tir-HITSZ-BWCC	0.385	0.590	0.455
tir-OKSAT-TF02	0.383	0.584	0.455
tir-HITSZ-BW	0.384	0.590	0.454
TUTA1-TIR-RUN-3	0.370	0.583	0.447
TUTA1-TIR-RUN-2	0.369	0.579	0.441
tir-OKSAT-TF03	0.372	0.572	0.439
TUTA1-TIR-RUN-1	0.362	0.568	0.438
tir-org-sq	0.364	0.559	0.436
tir-org-t	0.348	0.531	0.411
system-3	0.298	0.500	0.385
TIR-BRKLY-TS-T2FB	0.304	0.501	0.382
system-2	0.290	0.478	0.370
system-1	0.286	0.475	0.367
TIR-BRKLY-TDS-T2FB	0.274	0.468	0.352
TIR-BRKLY-TDS-T2	0.251	0.445	0.334

7. REFERENCES

[1] Joho, H. and Kishida, K. Overview of NTCIR-11. In:

Table 5: Temporalia relevance assessment statistics.

	Atemporal	Past	Recency	Future	All
L2	1,985 (23.3%)	1,719 (18.5%)	2,271 (24.6%)	2,102 (22.6%)	8,077 (22.2%)
L1	2,547 (29.9%)	2,361 (25.5%)	2,555 (27.6%)	2,747 (29.5%)	10,210 (28.1%)
L0	3,976 (46.7%)	5,196 (56.0%)	4,417 (47.8%)	4,449 (47.8%)	18,038 (49.7%)
Total	8,508	9,276	9,243	9,298	36,325

Table 6: Classification accuracy of TQIC. The highest value in each column is shown in bold.

Run name	Atemporal	Past	Recent	Future	All
TUTA1-TQIC-RUN-1	0.773	0.853	0.480	**0.853**	**0.740**
TUTA1-TQIC-RUN-2	0.760	0.853	0.467	0.827	0.727
tqic-Andd7.3	0.787	0.747	0.560	0.813	0.727
TUTA1-TQIC-RUN-3	0.533	**0.867**	0.587	0.840	0.707
tqic-Andd7.2	0.653	0.773	0.600	0.787	0.703
tiqc-HITSZ-PrW	0.707	0.787	0.627	0.640	0.690
tqic-Andd7.1	**0.813**	0.600	0.520	0.800	0.683
TQIC-HULTECH-Run1	0.627	0.813	**0.653**	0.627	0.680
tiqc-HITSZ-PrWsQW	0.693	0.773	0.573	0.667	0.677
tiqc-HITSZ-qRPrHNB	0.573	0.813	0.613	0.680	0.670
tqic-UniMAN.1	0.787	0.800	0.573	0.493	0.663
tqic-mpii-system2	0.760	0.600	0.493	0.707	0.640
tqic-mpii-system1	0.733	0.533	0.573	0.653	0.623
tqic-mpii-system3	0.800	0.600	0.440	0.627	0.617
tqic-UniMAN.2	0.667	0.707	0.600	0.480	0.613
TQIC-HULTECH-Run2	0.507	0.787	**0.653**	0.467	0.603
tqic-UniMAN.3	0.507	0.547	0.587	0.560	0.550
Mean	0.687	0.733	0.565	0.678	0.665
SD	0.105	0.112	0.064	0.128	0.051

Proceedings of the 11th NTCIR Conference, Tokyo, Japan, 2014.

[2] O. Alonso, R. Baeza-yates, J. Strotgen, and M. Gertz. M.: Temporal information retrieval: Challenges and opportunities. In: *TempWeb 2011*, pages 1–8, 2011.

[3] H. Joho, A. Jatowt, and B. Roi. A survey of temporal web search experience. In: *TempWeb 2013*, pages 1101–1108, 2013.

[4] G. Kazai, J. Kamps, and N. Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178, 2013.

[5] G. Kazai, S. Masood, and M. Lalmas. A study of the assessment of relevance for the inex'02 test collection. In: *ECIR 2004*, pages 296–310, 2004.

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

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

[8] E. Sormunen. Liberal relevance criteria of trec -: Counting on negligible documents? In: *SIGIR 2002*, pages 324–330, 2002.

[9] Filannino, M. and Nenadic, G. Using machine learning to predict temporal orientation of search engines' queries in the Temporalia challange. In: *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[10] Burghartz, R. and Berberich, K. PI-INF at the NTCIR-11 Temporal Query Classification Task. In: *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[11] Larson, R. A Logistic Regression Approach for NTCIR-11 Temporalia. In: *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[12] Shah, A., Shah, D., and Majumder, P. Andd7 @ NTCIR-11 Temporal Information Access Task. In: *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[13] Yu, H., Kang, X., and Ren, F. TUTA1 at the NTCIR-11 Temporalia Task. In: *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[14] Hou, Y., Tan, C., Xu, J., Pan, Y., Chen, Q., and Wang, X. HITSZ-ICRC at NTCIR-11 Temporalia Task. In: *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[15] Sato, T., and Aoki, S. OKSAT at NTCIR-11 Temporalia - Plural Sets of Search Terms for a Topic -. In: *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[16] Hasanuzzaman, M., Dias, G., and Ferrari, S. HULTECH at the NTCIR-11 Temporalia Task: Ensemble Learning for Temporal Query Intent Classification. In: *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[17] Sakai, T. Ranking the NTCIR Systems Based on Multigrade Relevance. In: *AIRS 2004*, pages 251–262, 2005.

Table 7: Confusion matrix between answer and estimated classes.

	Estimated class				
Answer class	Atemporal	Past	Recency	Future	total
Atemporal	783 (67.7%)	111 (9.6%)	193 (16.7%)	69 (6.0%)	1,156
Past	147 (13.1%)	836 (74.2%)	61 (5.4%)	82 (7.3%)	1,126
Recency	154 (13.5%)	28 (2.5%)	638 (55.9%)	322 (28.2%)	1,142
Future	51 (4.4%)	18 (1.6%)	299 (25.9%)	786 (68.1%)	1,154
total	1,135 (24.8%)	993 (21.7%)	1,191 (26.0%)	1,259 (27.5%)	4,578

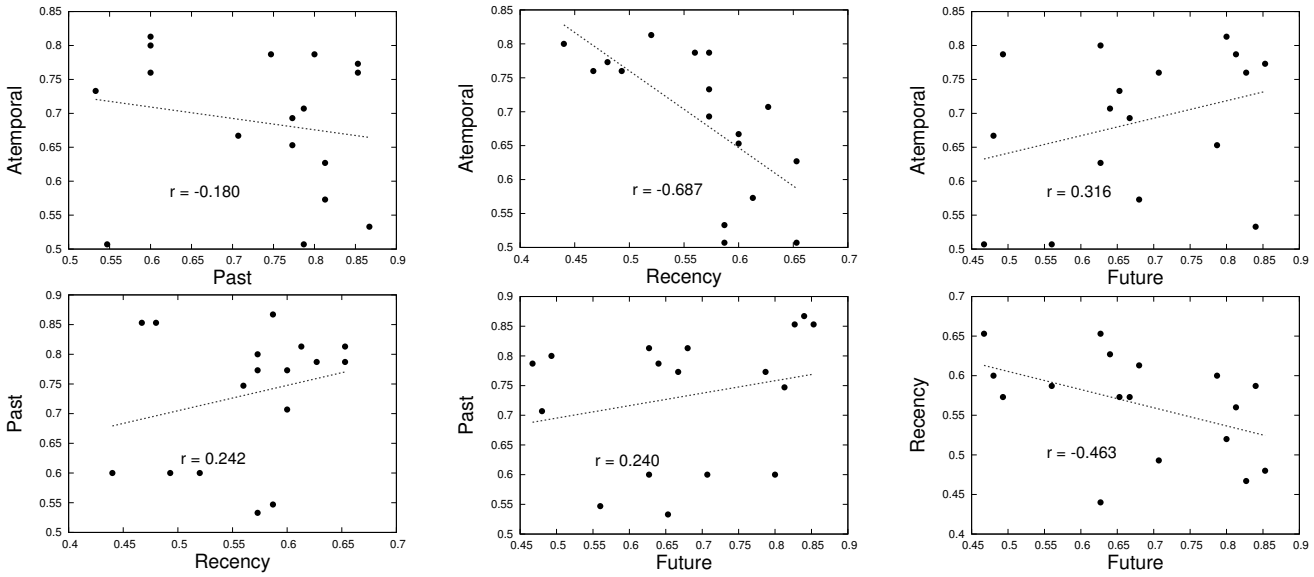

Figure 1: Pearson's correlation coefficient among temporal classes.

Table 9: Mean retrieval performance over atemporal 50 topics. The highest value in each column is shown in bold.

run name	Q@20	P@20	nDCG@20
tir-OKSAT-TF01	0.445	**0.646**	**0.518**
tir-HITSZ-LTRNC2	**0.452**	0.640	0.509
tir-org-sq	0.448	0.630	0.506
tir-org-sqd	0.436	0.641	0.505
tir-OKSAT-TF02	0.424	0.626	0.492
tir-HITSZ-BWCC	0.401	0.597	0.468
TUTA1-TIR-RUN-2	0.406	0.602	0.467
tir-HITSZ-BW	0.402	0.597	0.467
tir-OKSAT-TF03	0.380	0.583	0.456
TUTA1-TIR-RUN-1	0.381	0.584	0.451
TUTA1-TIR-RUN-3	0.381	0.594	0.443
tir-org-t	0.359	0.550	0.429
TIR-BRKLY-TS-T2FB	0.340	0.528	0.415
system-1	0.296	0.498	0.385
system-3	0.296	0.498	0.385
system-2	0.294	0.496	0.384
TIR-BRKLY-TDS-T2FB	0.274	0.456	0.350
TIR-BRKLY-TDS-T2	0.263	0.456	0.340

Table 10: Mean retrieval performance over past 50 topics. The highest value in each column is shown in bold.

run name	Q@20	P@20	nDCG@20
tir-HITSZ-LTRNC2	**0.342**	**0.531**	**0.423**
tir-org-sqd	0.317	0.524	0.409
tir-HITSZ-BWCC	0.328	0.522	0.403
tir-HITSZ-BW	0.323	0.520	0.401
tir-OKSAT-TF01	0.323	0.505	0.398
TUTA1-TIR-RUN-3	0.313	0.512	0.397
tir-OKSAT-TF02	0.315	0.507	0.387
tir-org-sq	0.302	0.501	0.385
TUTA1-TIR-RUN-2	0.310	0.513	0.380
TUTA1-TIR-RUN-1	0.282	0.489	0.372
tir-OKSAT-TF03	0.300	0.487	0.363
TIR-BRKLY-TS-T2FB	0.266	0.447	0.336
system-3	0.245	0.428	0.333
tir-org-t	0.266	0.434	0.331
TIR-BRKLY-TDS-T2FB	0.227	0.405	0.303
system-2	0.201	0.365	0.281
TIR-BRKLY-TDS-T2	0.200	0.363	0.276
system-1	0.192	0.362	0.274

Table 11: Mean retrieval performance over recency 50 topics. The highest value in each column is shown in bold.

run name	Q@20	P@20	nDCG@20
tir-org-sqd	**0.448**	**0.641**	**0.520**
tir-HITSZ-LTRNC2	0.432	0.617	0.495
TUTA1-TIR-RUN-2	0.416	0.630	0.492
tir-HITSZ-BWCC	0.427	0.632	0.491
tir-HITSZ-BW	0.425	0.629	0.490
system-1	0.417	0.614	0.489
system-2	0.417	0.614	0.488
TUTA1-TIR-RUN-3	0.393	0.609	0.483
tir-OKSAT-TF03	0.410	0.600	0.476
TUTA1-TIR-RUN-1	0.409	0.608	0.474
tir-OKSAT-TF02	0.397	0.601	0.473
tir-OKSAT-TF01	0.397	0.606	0.466
tir-org-sq	0.387	0.581	0.453
tir-org-t	0.386	0.571	0.451
system-3	0.341	0.542	0.417
TIR-BRKLY-TS-T2FB	0.298	0.498	0.385
TIR-BRKLY-TDS-T2FB	0.293	0.496	0.374
TIR-BRKLY-TDS-T2	0.257	0.456	0.346

Table 12: Subtopic mining runs ranked by mean nDCG@20 over future 50 topics. The highest value in each column is shown in bold.

Future			
run name	Q@20	P@20	nDCG@20
tir-org-sqd	**0.461**	**0.667**	**0.520**
tir-HITSZ-LTRNC2	0.413	0.619	0.480
tir-OKSAT-TF02	0.395	0.603	0.468
TUTA1-TIR-RUN-3	0.393	0.618	0.467
tir-OKSAT-TF03	0.397	0.619	0.462
tir-HITSZ-BW	0.384	0.612	0.461
tir-HITSZ-BWCC	0.382	0.610	0.459
TUTA1-TIR-RUN-1	0.376	0.592	0.454
tir-OKSAT-TF01	0.368	0.574	0.444
tir-org-t	0.379	0.568	0.432
TUTA1-TIR-RUN-2	0.346	0.572	0.424
system-3	0.309	0.534	0.403
tir-org-sq	0.320	0.525	0.401
TIR-BRKLY-TS-T2FB	0.311	0.530	0.390
TIR-BRKLY-TDS-T2FB	0.302	0.514	0.382
TIR-BRKLY-TDS-T2	0.285	0.505	0.373
system-2	0.246	0.439	0.328
system-1	0.240	0.428	0.320

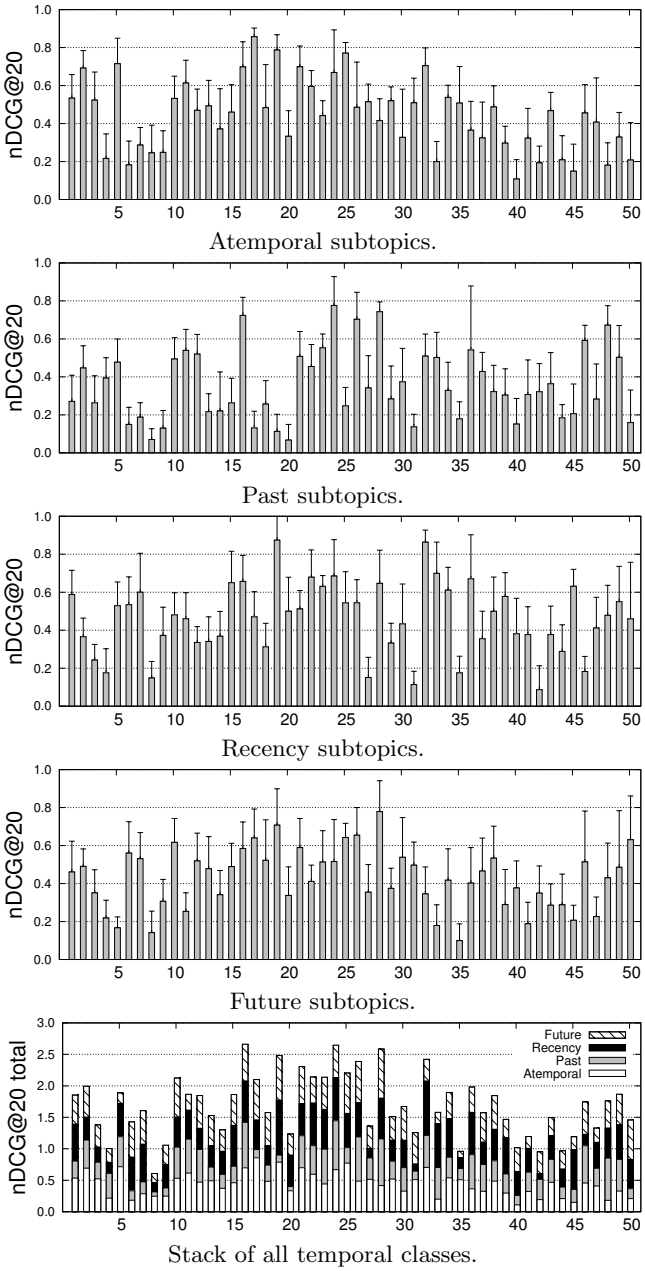

Figure 2: nDCG values across four temporal classes.