Building Test Collections for Evaluating Temporal IR

Hideo Joho
University of Tsukuba, Tsukuba, Japan
hideo@slis.tsukuba.ac.jp

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

Roi Blanco
University of A Coruña
rblanco@udc.es

Haitao Yu
University of Tsukuba, Tsukuba, Japan
yuhaitao@slis.tsukuba.ac.jp

Shuhei Yamamoto
University of Tsukuba, Tsukuba, Japan
yamahei@ce.slis.tsukuba.ac.jp

ABSTRACT

Research on temporal aspects of information retrieval has recently gained considerable interest within the Information Retrieval (IR) community. This paper describes our efforts for building test collections for the purpose of fostering temporal IR research. In particular, we overview the test collections created at the two recent editions of Temporal Information Access (Temporalia) task organized at NTCIR-11 and NTCIR-12, report on selected results and discuss several observations we made during the task design and implementation. Finally, we outline further directions for constructing test collections suitable for temporal IR.

CCS Concepts

•**Information systems** → **Test collections;**

Keywords

Temporal IR; test collections; temporal query intents; temporal search result diversification

1. INTRODUCTION

Topical document relevance has traditionally been the key concern in the development and evaluation of Information Retrieval (IR) systems. Recently, however, IR community has been interested in exploring other dimensions which underlie effective search processes such as the novelty, diversity, coverage, and/or readability. Timeliness is one such quality aspect of documents which, to a significant degree, determines user satisfaction, at least, for the subset of queries whose underlying search intents contain strong temporal component (e.g., `Pisa weather`, `Olympics 2020`). In consequence, Temporal IR [1, 7] has started to emerge as one challenge of information retrieval in which time and temporality play crucial roles.

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914673

Table 1: Number of formal runs (number of teams) in subtasks of Temporalia-1 and Temporalia-2. Number of teams increased by 40% in Temporalia-2.

Task	TQIC	TID	TIR	TDR
Temporalia-1 EN	17 (6)	-	19 (6)	-
Temporalia-2 EN	-	30 (12)	-	9 (3)
Temporalia-2 CH	-	7 (3)	-	0 (0)

Time-related aspects can be actually bound to many components in the architecture of a search engine system and to many steps of a retrieval process. For example, temporality may be already estimated on the level of input queries (query issue time and temporal intent class). Documents can be associated with different types of time (e.g., publication time, content focus time or content modification time) allowing construction of temporally-aware search indexes. Finally, retrieval mechanisms can output results according to the match between the temporal aspects of query intents and the estimated time scopes of documents [3].

While some prior attempts tried to survey the growing field of Temporal IR [1, 7] or to investigate the actual user needs and tactics while searching for information of temporal character [4], the community lacks common evaluation grounds and standardized test beds for effective comparison and evaluation of search technologies.

Temporal Information Access (*Temporalia*)[1] tasks hosted at NTCIR-11 [5] and NTCIR-12 [6] aimed to fill in this gap. Both *Temporalia-1* and *Temporalia-2* consisted of two subtasks focusing on two crucial components necessary for designing an effective time-aware IR process. *Temporal Query Intent Categorization* (*TQIC*) in Temporalia-1 and *Temporal Intent Disambiguation* (*TID*) in Temporalia-2 dealt with the problem of recognizing temporal class of search queries. On the other hand, *Temporal Information Retrieval* (*TIR*) and *Temporally Diversified Retrieval* (*TDR*) in Temporalia-1 and Temporalia-2, respectively, concentrated on evaluating document retrieval for information needs with underlying temporal aspect. While English was our primary focus, the second edition of Temporalia extended also the evaluation to Chinese. Tab. 1 shows the overall participation in both editions of the task, while Tab. 2 summarizes the resulting test collections.

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

Table 2: Collection statistics.

Collection	Language	Description	Size
Temporalia-1 TQIC	EN	Queries with temporal classes	300
Temporalia-1 TIR	EN	Topics with temporal subtopics	50
Temporalia-2 TID	EN	Queries with distribution over temporal classes	300
Temporalia-2 TID	CH	Queries with distribution over temporal classes	300
Temporalia-2 TDR	EN	Topics with temporal subtopics	50
Temporalia-2 TDR	CH	Topics with temporal subtopics	50

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
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 topic for TDR subtask (Dry Run).

Title	Father's day
Description	I am from a country where Father's Day is not a common custom, and I would like to learn more about Father's day and its relation to the role of father in society.
Past question	What is the history and origin of the Father's day?
Recency question	What famous persons has lately done during Father's day?
Future question	What is the future outlook for the problem of fatherlessness and how the father's role is supposed to change?
Atemporal question	What is Father's day?
Search date	1 Mar 2013 GMT+0:00

2. RELATED TEST COLLECTIONS

Although many tasks at TREC, CLEF or NTCIR have been devoted to query analysis and document retrieval, only a few of them particularly focused on temporal aspects in search. Furthermore, temporal notions in other tasks were typically limited to the timeliness and recency requirements put on extracted information. *TempSum* and *KBA* are the most relevant ones.

TREC's *Temporal Summarization task*[2] (*TempSum*) is composed of two subtasks: *Sequential Update Summarization* and *Value Tracking*. Sequential Update Summarization asks to find timely, sentence-level, reliable, relevant and non-redundant updates about developing events, while Value Tracking requires tracking values of event-related attributes (e.g., fatalities count) being of high importance to the event. Both subtasks have clear temporal character since the updates have to be not only relevant but also timely. Although, we also test ranking methods for the "recency" class of search needs in our tasks, we do not limit the scope to the information needs about recent or ongoing events, or to a particular type of attribute-like information. Hence, any temporal query category may become an input, and the results are output as lists of ranked documents rather than information pieces.

One of its subtasks, *Streaming Slot Filling*, tracks the attributes and relations of a selected entity over time. Similarly to TempSum, the temporal information need is in the form of a recency requirement. In addition, only documents related to a particular entity such as a person or organization are deemed relevant.

3. TEMPORAL INFORMATION ACCESS

First we discuss time categorization used in Temporalia. Several categorizations of time are possible such as ones according to calendar, frequency, periodicity, distance, etc. However, the most common and the most natural distinction is the division of timeline into the past, present and future from the viewpoint of a speaker or information searcher. Therefore, we have constructed our task around this natural time distinction assuming the query issuing time as a center (i.e., point "now"). In addition, we provided an atemporal class for queries without any obvious time focus. The four classes are defined below. We also show example queries with their per-class probability distributions in Tab. 3.

Past: class of queries about past entities or events. Content relevant to past-related queries is not expected to change significantly over time.

Recency: class of queries about present or recent entities or events. Relevant content for queries belonging to the Recency class generally changes frequently and quickly hence the search results are expected to contain timely and up-to-date documents.

Future: class of queries about predicted, scheduled or expected events. Search results should contain documents with future-related information when compared to the query issuing time.

Atemporal: class characterizing queries without any clear temporal intent. Search results returned for atemporal queries are not expected to be related to time, neither, their content should change much over time. Navigational queries are considered to be atemporal.

3.1 Temporal Intent Detection

In both the task editions we have approached the problem of detecting temporal classes of intent underlying search queries. According to the well-known study conducted in 2008 on the AOL query dataset [10], about 1.5% of queries have been found to contain explicit temporal expressions (e.g., `Tokyo Olympics 1964`, `popular songs 2000s`). Although this rate may seem small at first, it amounts to quite a large number of searches, and it does not cover queries which lack clear temporal expressions which contain strong implicit temporal component (e.g., `Berlin Wall`, `economic forecast` or `Tokyo Olympics`).

Given the set of query strings associated with their submitting times, participants had to develop systems to de-

[2]http://www.trec-ts.org/

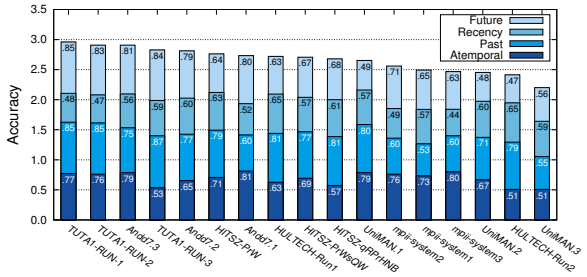

Figure 1: Performance in TQIC subtask.

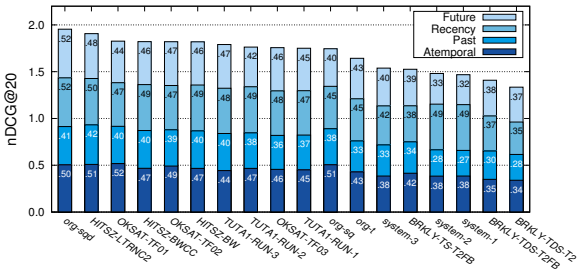

Figure 2: Performance in TIR subtask.

		Estimated class				
		Atemporal	Past	Recency	Future	total
Answer class	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 3: Confusion matrix in TQIC subtask.

termine either the dominant temporal class (past, recency, future and atemporal) of each of the queries (TQIC subtask in Temporalia-1) or their per-class distribution (TID subtask in Temporalia-2). Queries for both the dry and formal runs were created based on seed set of temporal expressions and were then collected from search engine query logs or search suggestions given by major search engines. We have then administered crowdsourcing tasks for determining the distribution of queries across the temporal classes. Each query received 20 judgments.

The performance of submitted runs in TQIC was measured by the average rate of queries with correct temporal classes and, in TID, by (i) the averaged per-class absolute loss as well as (ii) by the cosine similarity between the ground truth temporal class distribution and the distribution estimated by the participating systems.

In Fig.1 we show the system performance in Temporalia-1 TQIC subtask. Temporalia-2 TID results will be reported at NTCiR-12 Conference, June 2016, Tokyo, Japan.

3.2 Temporal Ranking

The second problem relates to ranking results for queries containing temporal requirement. First, *Temporal Information Retrieval* (TIR) at Temporalia-1 asked participants to retrieve a set of documents for a search topic that incorporates time factor. In addition to a standard search topic data (i.e., title, description, and sub topics), TIR search topic contained the query submitting date information to indicate the relationship between the query and searcher.

Temporally Diversified Retrieval (TDR) extended TIR by requiring participants to separately retrieve ranked lists of documents relevant to each of four temporal intent classes for a given topic. Participants were also asked to return a set of documents that is temporally diversified for the same topic. TDR participants 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 was to demonstrate one possible subtopic under a particular temporal class. In Tab. 4 we show an example topic with its subtopics.

For the evaluation, the standard Cranfield methodology was used. In particular, a pool of potentially 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 (3 judgments per document for each class), and its relevance grade was judged. Given a ranked list generated for a specific temporal subtopic, the system performance was evaluated using Precision, Q-measure and nDCG metric. For a diversified ranked list generated to satisfy all possible temporal classes, the performance was evaluated using α-nDCG and D#-nDCG.

In Fig.2 we show the system performance in Temporalia-1's TIR subtask. Temporalia-2's TDR results will be reported at NTCIR-12 Conference.

3.2.1 Document Collections

Both, Temporalia-1 and Temporalia-2 used the same document corpus for English called *LivingKnowledge News and Blogs Annotated Sub-collection* constructed by the LivingKnowledge project and distributed by the Internet Memory Foundation [9]. This collection of approximately 20GB size spans the time period from May 2011 to March 2013 and contains around 3.8M documents from about 1,500 different blogs and news sources. We chose this collection since it has a relatively long temporal span of documents. To facilitate information extraction, we have indicated sentence boundaries as well as mentions of named entities, and we have also disambiguated detected temporal expressions.

Temporalia-2 has also introduced Chinese language. SogouCA [3] and SogouT [4] document collections were used for the dry run and formal run, respectively. Three kinds of annotations were provided: sentence splitting, named entities, and time annotations.

4. DISCUSSIONS

In this section we discuss several observations obtained during the task design and we outline some future directions:

Challenges with Recency and Future classes: Recency and Future classes seem to be the most challenging when it comes to recognizing temporal intent. The mean classification accuracy for these classes were 0.565 and 0.678, respectively, compared to 0.687 and 0.733 for the Atemporal and Past classes, respectively. This can be also seen in Fig. 3, which shows the confusion matrix for TQIC subtask.

Balancing queries over temporal classes: When creating test sets for TID subtask we faced the problem with

[3]http://www.sogou.com/labs/dl/ca.html
[4]http://www.sogou.com/labs/dl/t-e.html

Team	Runs	Num. of features	External data sources	External tools	Classifiers
HITSZ	3	3 features + ngrams and POS ngrams	Web search results (titles, snippets)	POS tagger, NER	Classifier voting & rule-based method
HULTECH	2	11 features	Web search results (snippets)	TempoWordnet, GTE	Ensemble learning (8 classifiers)
TUTA1 (top)	3	3 feature groups	AOL 500K query session	POS tagger, SUTime, NER	Semi-supervised & supervised classifiers (Log. Reg., SVMLin)
Andd7	3	3 features + bag of words	-	Porter Stemmer	SVM, NB, DT & their agreement
MPII	3	6 feature groups + ngrams	DMOZ directory	2 thesauri, POS tagger, NER	NB, DT & simulated annealing
UniMan	3	19 features	Wikipedia	POS tagger, TempoWordnet, ManTIME	SVM & Random Forrest

Figure 4: Overview of techniques used in TQIC.

constructing a query set that would be appropriately balanced over query classes. For example, Atemporal class was ubiquitous for each query making it difficult to find many queries without substantial probability score in the Atemporal class. Relatedly, the combinations in which relatively high rate was observed in the Future as well as in the Past class were quite rare. All these issues made the construction of symmetrical test set collections a difficult challenge.

Sources of temporal signals: Participants used variety of external sources and tools for capturing temporal signals. For a rough overview, we show in Fig. 4 the types of temporal signals and techniques used by the participating teams in the TQIC subtask.

Relation to search result diversification tasks: *IMine* [8] and *Intent* [12] tasks focus on various aspects of search result diversification in IR. For example, *Query Understanding* (*QU*) and *Vertical Incorporating* (*VI*) subtasks in *IMine-2* require to identify the subtopics of a query or diversify organic results respectively. Temporal intent detection and temporal diversification of results could be then regarded as attempts at enforcing higher level diversification, i.e., diversification not on the level of subtopics but, on the level of temporal classes, each of which typically covers several possible subtopics that share the same temporal orientation.

Task participation: As indicated in Tab. 1, the number of participating teams in Temporalia-2 increased by 40% from Temporalia-1. This demonstrates the need for and benefits of Temporalia's test collections. As for the subtask breakdown, tasks related to estimating temporal query classes were more popular than ones for retrieving documents. Besides the presumed difficulty of indexing the document collection, the need to cope with large data sizes as well as with the content of varying subtopics, the other reasons for this preference could come from the problems with estimating temporality of large text chunks in contrast to short query strings that can be rather easily manipulated. We also note that Chinese language TDR failed to attract any participation in contrast to Chinese language TID. This may suggest that the additional difficulty of processing large amounts of ideograms in documents might have outweighed the benefits of solving the task in a language less commonly researched within the IR community.

Future of Temporal IR evaluation: Estimating focus time of documents [1, 3], which is the time to which documents refer to, could be one possible subtask in the future. Another challenge is estimating document creation time. In this context, we note that *SemEval2015* introduced *Diachronic Text Evaluation* task [11] which required to identify time interval when a given piece of news text (typically spanning between tens to a couple of hundred of words) was published on the time scale of 1700 and 2010.

We think that the combination of the document creation and focus time coupled with the estimated query's temporality and its issuing date should be sufficient for judging temporal correspondence of documents to the query. Finally, future challenges could be designed to testing temporal query suggestion or retrieval mechanisms over long term document archives.

5. CONCLUSIONS

In this paper we have reported the results of our recent efforts towards building test collections for fostering Temporal IR and related applications in the context of Temporalia-1 and Temporalia-2 tasks at NTCIR-11/12. Both tasks focused on two important objectives. The first one is to test methods for recognizing temporal aspects behind query intent. The second objective is to facilitate comparison of document ranking methods according to temporal classes of queries. Unlike other test collections, Temporalia's test collections are designed to let researchers compare performance of their systems across major temporal classes.

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.

6. REFERENCES

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

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

[3] A. Jatowt, C.M. Au Yeung and K. Tanaka. Estimating Document Focus Time. In: *CIKM'13*, 2273–2278.

[4] H. Joho, A. Jatowt, and R. Blanco. Temporal Information Searching Behaviour and Strategies. In: *Inf. Proc. & Manag.*, 51(6):834-850 (2015).

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

[6] H. Joho, A. Jatowt, R. Blanco, H. Yu and S. Yamamoto. Overview of NTCIR-12 Temporal Information Access (Temporalia-2) Task. In: *Proc. of NTCIR-12*, to appear, 2016.

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

[8] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. P. Kato, H. Ohshima, K. Zhou. Overview of the NTCIR-11 IMine Task. In: *Proc. of NTCIR-11*, 2014.

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

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

[11] O. Popescu, and C. Strapparava. SemEval2015, Task 7: Diachronic Text Evaluation. In: *Proc. of the International Workshop on Semantic Evaluation.*, 2015.

[12] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, M. P. Kato, R. Song, M. Iwata Summary of the NTCIR-10 INTENT-2 Task: Subtopic Mining and Search Result Diversification. In: *Proc. of IGIR 2013*, 761–764, 2013.