NTCIR Temporalia: A Test Collection for Temporal Information Access Research

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! Research
Barcelona, Spain
roi@yahoo-inc.com

ABSTRACT

Time is one of the key constructs of information quality. Following an upsurge of research in temporal aspects of information search, it has become clear that the community needs standardized evaluation benchmark for fostering research in Temporal Information Access. This paper introduces Temporalia (Temporal Information Access), a new pilot task run at NTCIR-11 to create re-usable datasets for those who are interested in temporal aspects of search technologies, and discusses its task design in detail.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Measurement, Theory.

Keywords

Data challenge, NTCIR, Temporal IR

1. INTRODUCTION

Temporal Information Retrieval (e.g., [1,2,3,4,5,6,7,8] has been gaining a lot of interest in IR and related research communities. It can be defined as a subset of document retrieval in which time plays crucial role in estimating document relevance. Our recent analysis [5] suggests that although many users search for recent (fresh) information, a good proportion of them also seek for information about past incidents as well as future incidents. Given the fact that time plays crucial role in estimating information relevance and validity we believe that successful search engines must consider temporal aspects of information in greater detail. Otherwise a mismatch can occur resulting in poor user satisfaction. Consider a query with future-focused intent such as "Summer Olympics" for which search engine returns results about the past Olympics, or query about high-recency topics (e.g., "Kyoto weather", "Facebook stock price") that returns pages with obsolete information.

Although there are several evaluation tasks that involve search and filtering of temporal information (e.g., TDT, NTCIR GeoTime, TREC Temporal Summarization), nonce are designed to measure the performance of search applications across categories of temporal information needs such as Past, Recent, Seasonal, and Future (cf. [7]) in a systematic way. We thus propose a challenge called "Temporal Information Access" (Temporalia) that establishes common grounds for designing and analyzing time-aware information access systems.[1] The objective of this task is to systematize various requirements in Temporal IR and offer a standardized challenge based on which competing systems can be compared and analyzed.

Temporalia is hosted by the 11th NTCIR Workshop on Evaluation of Information Access Technologies (NTCIR-11).[2] It offers two subtasks to address temporal information access technologies as follows:

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

TQIC subtasks aims at classifying queries into four predetermined temporal classes based on their implicit or explicit 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 a study performed on AOL query dataset [9], 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 (see for example [3,4,8]).

On the other hand, TIR subtask focuses on ranking documents within the prepared collection for different temporal query classes. In such ranking both the topical and temporal relevance should be considered. There is a large body of research in ranking documents for temporal queries (e.g., [2,6]), hence, we believe

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579044

[1] At the time of writing this paper, 20 teams have registered.

[2] http://research.nii.ac.jp/ntcir/ntcir-11/

there should be significant interest in this particular task. Note that both TQIC and TIR are independent of each other and thus interested teams can choose one of the tasks or participate in both of them.

The remainder of this paper is structured as follows. In the next section we describe other related tasks focusing on their differences from Temporalia pilot task. Section 3 describes both TQIC and TIR subtasks, while Section 4 contains overview of the data collection. We outline the task schedule in Section 5 and then conclude the paper in Section 6.

2. RELATED TASKS

TREC Temporal Summarization and TREC Knowledge Base Acceleration are perhaps the closest tasks to Temporalia.

TREC Temporal Summarization task[3] (TempSum) is composed of two subtasks: Sequential Update Summarization and Value Tracking. The task of Sequential Update Summarization is to find timely, sentence-level, reliable, relevant and non-redundant updates about developing events. Value tracking aims at tracking values of event-related attributes that are of high importance to the event. Examples are the number of fatalities or financial impact. Both subtasks have clear temporal character since the updates have to be timely and relevant. Their outcome is then related to the subtask of ranking documents for the “recency” class of queries (as in Temporal Information Retrieval subtask of Temporalia). However, in the case of TempSum, the scope is limited to the information about a concrete past event or to a particular type of attribute-like information such as a numerical value, etc. Also, the returned information needs to be short and non-redundant. TIR subtask in Temporalia, on the other hand, can take any category of temporal query as input. The results are in the form of ranked list of documents for which neither redundancy nor text length plays any special role. Hence, the overlap of TemSum and Temporalia is rather minimal.

TREC Knowledge Base Acceleration task[4] (KBA) is a task for filtering a large stream of text to find documents that can help update knowledge bases like Wikipedia or Freebase. It is composed of two subtasks: Cumulative Citation Recommendation and Streaming Slot Filling. The former is about filtering documents that are worth citing in a profile of selected entities (profile could be a Wikipedia page of the entity). It does not add any requirement on temporality neither novelty of the content. On the other hand, the Streaming Slot Filling tracks the attributes and relations of a selected entity over time. As the organizers state, it virtually “allows the target entity to "move" as accumulated content implies modifications to the attributes, relations, and free text associated with” an entity. Thus, again, like in the case of TempSum, the temporal information need is in the form of the recency requirement put on documents related to a particular entity such as a person or organization.

Space and time are commonly treated as orthogonal or parallel dimensions. GeoCLEF 2008 task[5] required to categorize queries into geographic queries and non-geographic queries. In a similar fashion, in Temporalia the query categorization task consists of deciding whether query is temporal or non-temporal. GeoCLEF also looked deeper into components of geographic queries differentiating typical three subparts: “where”, “geo-relation” and “what” components. The “what” component is of a map type (e.g., river, beach, mountain, and monuments), yellow page type (e.g., businesses or organizations, like hotels, restaurants, hospitals) or information type. In contrast to GeoCLEF, in Temporalia we provide deeper, hierarchy of query categorization proposing diverse query types such as recency, past-related or future-related queries. We then set up the second subtask consisting of ranking documents for diverse temporal query types.

Finally, the GeoTime task[6] aims at answering mixed geo-temporal information needs represented by questions such as “When and where did George Kennan die?” or “When and where were the last three Winter Olympics held?” The temporal answer is in the form of date or period/interval type variable. In Temporalia we focus on more diverse types of temporal search needs (e.g., ones for which freshness and timeliness of answers play key role and so on).

3. SUBTASKS

3.1 Temporal Query Intent Classification

TQIC subtask asks participants to classify a given query string to one of temporal classes. Figure 1 shows the hierarchy of query classes on which TQIC is based.

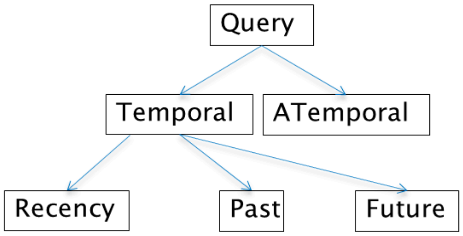


Figure 1. Hierarchy of query classes.

In particular, participating teams will be 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). Navigational queries are considered to be atemporal.

Table 1 shows examples of queries from each category.

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

[4] http://trec-kba.org/trec-kba-2013.shtml

[5] http://www.uni-hildesheim.de/geoclef/

[6] http://metadata.berkeley.edu/NTCIR-GeoTime/

Table 1. Example queries for TQIC subtask[7].

Query category	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

Participants will receive a set of query strings and query submitting date, and will be asked to develop a system to classify each of the query strings to one of the four temporal classes. As this problem rather requires different kinds of knowledge (e.g., historical information or information on planned events), the participants will be allowed to use any external resources to complete the TQIC subtask as long as the details of external resource usage are described in their report.

Each participating team is asked to submit a temporal class (past, recency, future, or atemporal) for each of the queries. The performance of submitted runs will be measured by the number of queries with correct temporal classes divided by the total number of queries. We will provide the breakdown of the performance based on temporal classes.

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 will contain a query submitting date information to indicate the relationship between the query and searcher. This subtask requires indexing our document collection (see Section 4).

We show in Table 2 example search topic with its corresponding temporal questions (past, recency, future and atemporal).

Table 2. Example topics for TIR subtask.

Title	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
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 the 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?.
Search date	1 Mar 2013 GMT+0:00

Participants are asked to submit the top 100 documents for each of temporal questions per topic (i.e., top 100 documents for past question, another 100 for recency question, etc.). The retrieval effectiveness will be evaluated by the precision at 20 for each of the temporal questions. Same as in TQIC subtask, we will provide the breakdown of the performance across temporal questions.

4. DOCUMENT COLLECTION

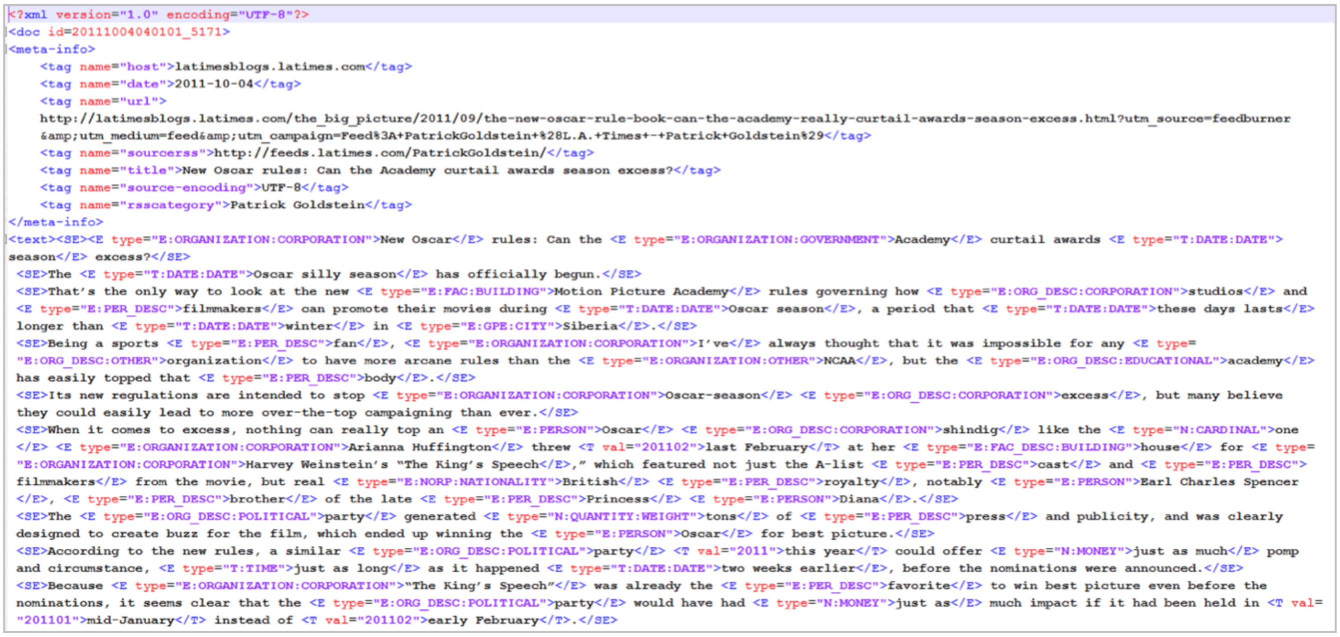
NTCIR-11 Temporalia uses a document corpus, called “LivingKnowledge news and blogs annotated subcollection”, constructed by the LivingKnowledge project[8] and distributed by the Internet Memory Foundation[9]. The collection is approximately 20GB uncompressed and over 5GB zipped in size. It spans from 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 some information about its sources (there might be more than one file per day).

Each file contains a number of text documents. For each document the following information is available (See Figure 2). 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 data 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

[7] Query submitting date is assumed to be on Feb 28, 2013 GMT+0:00 for all queries shown in Table 1.

[8] http://livingknowledge.europarchive.org/

[9] http://internetmemory.org/



```
<?xml version="1.0" encoding="UTF-8"?>
<doc id=20111004040101_5171>
<meta-info>
    <tag name="host">latimesblogs.latimes.com</tag>
    <tag name="date">2011-10-04</tag>
    <tag name="url">
    http://latimesblogs.latimes.com/the_big_picture/2011/09/the-new-oscar-rule-book-can-the-academy-really-curtail-awards-season-excess.html?utm_source=feedburner
    &amp;utm_medium=feed&amp;utm_campaign=Feed%3A+PatrickGoldstein+%28L.A.+Times+-+Patrick+Goldstein%29</tag>
    <tag name="sourcerss">http://feeds.latimes.com/PatrickGoldstein/</tag>
    <tag name="title">New Oscar rules: Can the Academy curtail awards season excess?</tag>
    <tag name="source-encoding">UTF-8</tag>
    <tag name="rsscategory">Patrick Goldstein</tag>
</meta-info>
<text><SE><E type="E:ORGANIZATION:CORPORATION">New Oscar</E> rules: Can the <E type="E:ORGANIZATION:GOVERNMENT">Academy</E> curtail awards <E type="T:DATE:DATE">
season</E> excess?</SE>
 <SE>The <E type="T:DATE:DATE">Oscar silly season</E> has officially begun.</SE>
 <SE>That's the only way to look at the new <E type="E:FAC:BUILDING">Motion Picture Academy</E> rules governing how <E type="E:ORG_DESC:CORPORATION">studios</E> and
 <E type="E:PER_DESC">filmmakers</E> can promote their movies during <E type="T:DATE:DATE">Oscar season</E>, a period that <E type="T:DATE:DATE">these days lasts</E>
 longer than <E type="T:DATE:DATE">winter</E> in <E type="E:GPE:CITY">Siberia</E>.</SE>
 <SE>Being a sports <E type="E:PER_DESC">fan</E>, <E type="E:ORGANIZATION:CORPORATION">I've</E> always thought that it was impossible for any <E type=
 "E:ORG_DESC:OTHER">organization</E> to have more arcane rules than the <E type="E:ORGANIZATION:OTHER">NCAA</E>, but the <E type="E:ORG_DESC:EDUCATIONAL">academy</E>
 has easily topped that <E type="E:PER_DESC">body</E>.</SE>
 <SE>Its new regulations are intended to stop <E type="E:ORGANIZATION:CORPORATION">Oscar-season</E> <E type="E:ORG_DESC:CORPORATION">excess</E>, but many believe
 they could easily lead to more over-the-top campaigning than ever.</SE>
 <SE>When it comes to excess, nothing can really top an <E type="E:PERSON">Oscar</E> <E type="E:ORG_DESC:CORPORATION">shindig</E> like the <E type="N:CARDINAL">one
 </E> <E type="E:ORGANIZATION:CORPORATION">Arianna Huffington</E> threw <T val="201102">last February</T> at her <E type="E:FAC_DESC:BUILDING">house</E> for <E type=
 "E:ORGANIZATION:CORPORATION">Harvey Weinstein's "The King's Speech</E>," which featured not just the A-list <E type="E:PER_DESC">cast</E> and <E type="E:PER_DESC">
 filmmakers</E> from the movie, but real <E type="E:NORP:NATIONALITY">British</E> <E type="E:PER_DESC">royalty</E>, notably <E type="E:PERSON">Earl Charles Spencer
 </E>, <E type="E:PER_DESC">brother</E> of the late <E type="E:PER_DESC">Princess</E> <E type="E:PERSON">Diana</E>.</SE>
 <SE>The <E type="E:ORG_DESC:POLITICAL">party</E> generated <E type="N:QUANTITY:WEIGHT">tons</E> of <E type="E:PER_DESC">press</E> and publicity, and was clearly
 designed to create buzz for the film, which ended up winning the <E type="E:PERSON">Oscar</E> for best picture.</SE>
 <SE>According to the new rules, a similar <E type="E:ORG_DESC:POLITICAL">party</E> <T val="2011">this year</T> could offer <E type="N:MONEY">just as much</E> pomp
 and circumstance, <E type="T:TIME">just as long</E> as it happened <E type="T:DATE:DATE">two weeks earlier</E>, before the nominations were announced.</SE>
 <SE>Because <E type="E:ORGANIZATION:CORPORATION">"The King's Speech"</E> was already the <E type="E:PER_DESC">favorite</E> to win best picture even before the
 nominations, it seems clear that the <E type="E:ORG_DESC:POLITICAL">party</E> would have had <E type="N:MONEY">just as</E> much impact if it had been held in <T val=
 "201101">mid-January</T> instead of <T val="201102">early February</T>.</SE>
```

Figure 2. Example of document from the provided collection.

the page. Between the <text> tags, there's the content of the page. This collection has been automatically tagged with different semantic annotations (see [8] for a more detailed description of how the annotations are 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. An example of document in the provided collection can be found in Figure 2.

Finally, it should be emphasized that the dataset (both document collection and query/topic data) generated by Temporalia will be released to interested research groups after the NTCIR-11 workshop.

5. SCHEDULE

Temporalia at NTCIR-11 will run in the following schedule (see Table 3). As can be seen, we have a two-stage schedule for task participants to test and develop their systems in the dry run period, and to formally evaluate the performance in the formal run period.

Table 3. Schedule of NTCIR-11 Temporalia.

Jan 05, 2014	Document collection release
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 1, 2014	Release of formal run topics/queries
Jun 30, 2014	Deadline for formal run submissions
Aug 1, 2014	Evaluation results release
Aug 1, 2014	Early overview draft release
Sep 1, 2014	Participant papers due
Nov 1, 2014	All camera-ready copy due
Dec 09-12, 2014	NTCIR-11 Conference at Tokyo

6. CONCLUSIONS

This paper presented our ongoing project to create a test collection for Temporal Information Access Research in the context of NTCIR-11. This is a pilot task and we welcome any feedback on the design and methodology of test collections that can facilitate the research on temporal information access. For the details of the task and latest information (that's temporal), please visit our website.[10]

7. ACKNOWLEDGMENTS

The first author was supported in part by MEXT Grant-in-Aid for Young Scientists B (#24700239). The second author was supported in part by MEXT Grant-in-Aid for Young Scientists B (#22700096) and by the JST research promotion program Sakigake: "Analyzing Collective Memory and Developing

[10] https://sites.google.com/site/ntcirtemporalia/ (Main)
http://ntcir.nii.ac.jp/Temporalia/ (Mirror)

Methods for Knowledge Extraction from Historical Documents”. The authors also thank to the NTCIR project at NII for supporting our activity. Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] O. Alonso, R. Baeza-Yates, J. Strötgen, M. and Gertz. Temporal information retrieval: Challanges and opportunities. In TWAW 2011 Workshop, 2011.

[2] K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. A Language Modeling Approach for Temporal Information Needs. In ECIR 2010, pages 13–25, 2012.

[3] R. Campos, A. Mário Jorge, G. Dias, C. Nunes. Disambiguating Implicit Temporal Queries by Clustering Top Relevant Dates in Web Snippets. Web Intelligence 2012.

[4] R. Jones and F. Diaz. Temporal profiles of queries. ACM Transactions on Information Systems, 25(3):656–678, 2007.

[5] H. Joho, A., Jatowt, and R., Blanco. (2013). A Survey of Temporal Web Search Experience”. In: Proceedings of the TempWeb 2013 Workshop WWW 2013.

[6] N. Kanhabua, and K. Nørvåg. Determining Time of Queries for Re-Ranking Search Results. ECDL 2010.

[7] A. Khodaei and O. Alonso. Temporally-Aware Signals for Social Search, SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012).

[8] M.Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika and H. Zaragoza. Searching through time in the New York Times. HCIR, 2010.

[9] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. SIGIR 2009.

[10] S. Nunes, C. Ribeiro, and G. David. Use of Temporal Expressions in Web Search. ECIR2008.