When time meets information retrieval: Past proposals, current plans and future trends

Bilel Moulahi
University of Toulouse Paul Sabatier, France
University of Tunis El-Manar, Tunisia

Lynda Tamine
University of Toulouse Paul Sabatier, France

Sadok Ben Yahia
University of Tunis El-Manar, Tunisia
TELECOM SudParis, France

Abstract
With the advent of Web search and the large amount of data published on the Web sphere, a tremendous amount of documents become strongly time-dependent. In this respect, the time dimension has been extensively exploited as a highly important relevance criterion to improve the retrieval effectiveness of document ranking models. Thus, a compelling research interest is going on the temporal information retrieval realm, which gives rise to several temporal search applications. In this article, we intend to provide a scrutinizing overview of time-aware information retrieval models. We specifically put the focus on the use of timeliness and its impact on the global value of relevance as well as on the retrieval effectiveness. First, we attempt to motivate the importance of temporal signals, whenever combined with other relevance features, in accounting for document relevance. Then, we review the relevant studies standing at the crossroads of both information retrieval and time according to three common information retrieval aspects: the query level, the document content level and the document ranking model level. We organize the related temporal-based approaches around specific information retrieval tasks and regarding the task at hand, we emphasize the importance of results presentation and particularly timelines to the end user. We also report a set of relevant research trends and avenues that can be explored in the future.

Keywords
Relevance; time; temporal queries; temporal ranking; timelines

1. Introduction and motivations

Searching valuable information in growing masses of data is still an open and thriving challenge in the community of information retrieval (IR). Alongside this progress, the Web users address the need of enhanced search facilities to effectively retrieve the resources that match their actual search intents. In order to provide the most relevant results to users, most state-of-the-art approaches rank documents by computing single scores separately with respect to one single objective criterion, expressing the documents topical overlap with the user's information need [1]. However, several studies stressed on the fact that relevance is a multidimensional concept that goes beyond this too simple topicality criterion towards a much richer set of relevance criteria [1–3]. This multidimensional property has been recently witnessed in many IR contexts to tailor the search results according to the given set of criteria as well as the IR application at hand

Corresponding author:
Lynda Tamine, IRIT Laboratory, University of Toulouse Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9, France.
Email: tamine@irit.fr

Figure 1. Google search results on the query 'Karim Benzema', accessed on June 2014.

[4, 5]. Specifically, the time criterion has already been the core concept of recent IR ranking models, given that most of documents include a high level of temporal information [6]. Besides, a large part of work, carried out into this scope argued that the search intent, behind queries and user search behaviour, change over time, so that the temporal information contained in the documents may be exploited to enhance numerous aspects of the Web. This issue has been recently tackled in real world search engines. For instance, the Google[1] search results on the query 'Karim Benzema' (*Cf.*, Figure 1) in the time period of the 2014 FIFA World Cup show a compact statistic table about the performance of 'Benzema' in the tournament on the top of the page result, and below appears some recent related news. More interestingly, unlike the usual search results given as an answer to this kind of queries, the Wikipedia page of the player figures at the bottom of the documents result list.

   To make matters more concrete, consider a user who enters the query 'Michael Jackson'. The latter may wish to find results about Michael Jackson's biography, his official website or some related news. However, as Michael Jackson's death was in 2009 (i.e. not recent) and the query is too short, it may be difficult to accurately find the users' intent behind the submitted couple of keywords or their sensitivity to time. Likewise, if we take another completely different query like 'Cricket World Cup', it is not clear if the latter is about the 2015 Cricket World Cup or another previous or even future one. Thus, understanding and exploiting temporal indicators contained on queries and documents may be of potential benefit for search engines. Indeed, temporal information can be used to favour documents whose contents refer to a time period in the past as per the query 'Michael Jackson', or more fresh documents such as that of '2015 Cricket World Cup'. If a time is explicitly stated in the query, this information could be taken into account to decide whether to return data matching the query on that particular point time in the past, or earlier related information. Taking in consideration the information about time is becoming increasingly prominent in the literature [7–9].

State-of-the-art approaches are often based on ranking models encompassing time in a relatively crude manner. Most current methods sort the results based on the publication time of document content, or ask the user to restrict the results to a particular period of time. Other works include temporal factors in the core of the ranking model. A large number of productive research has followed this direction so far. All these methods give result to various domain applications within the Web search community ranging from temporal ranking to temporal summaries, spatio-temporal information extraction, temporal indexing, temporal dynamics and temporal clustering, to cite but a few.

In this work, we provide a comprehensive and a comparative overview of most important work on both time and IR. We adopt a categorization inspired from three basic IR components namely query, documents and ranking models, and we additionally review the different techniques used to represent time-sensitive search results. There are some surveys studies concerning temporal IR on the past, where a set of representative methods for temporal document ranking were presented [6, 7, 10, 11]. Alonso et al. [7] and Nunes et al. [12] were among the first to emphasize the exploitation of temporal information. Nunes et al. [12] proposed a classification of the sources of temporal evidence based on two groups, namely document-based evidence (i.e. features extracted from individual documents) and Web-based evidence (i.e. features obtained from the whole Web). Mathews and Kanmani [6] proposed a brief overview on T-IR systems. They distinguished two different methods, namely time-aware retrieval methods and temporal ranking methods. Recently, Campos et al. [11] provided a general overview on temporal IR as well as some related applications such as temporal clustering, temporal text classification, temporal search engines and future IR. The survey presented in this article is fundamentally different from the previous ones according to several points of view. First, our work is more general than those already carried out and is not only restricted to approaches proposed on ad-hoc Web search. We cover many IR-related topics on diverse areas including traditional Web search, social search, mobile search and geo-temporal search that heavily involve document ranking and time-aware IR. We present these studies according to four IR aspects: queries, document's contents, ranking models and result presentation. Second, the aim of our article is to provide a broad picture on timeliness from the IR perspective across various related domains. The focus is put regarding the aforementioned research axis and mainly from the query and ranking levels perspective. This could inspire researchers from other fields on how timeliness is important, and show them insight to what extent time-aware applications could also meet their specific research area. We also present the evaluation frameworks and techniques used in order to assess the effectiveness of these temporal IR (T-IR) methods and show the major results that emerged from them. We do not extensively compare these techniques in terms of their performance, because they have not all been tested on the same corpus. Nevertheless, we show the outcome of the most important challenges and tracks that focused on T-IR access models and techniques. We finally report some open, challenging and promising research directions. We note that by no means is this work meant to be exhaustive and to cover all T-IR domains. We refrain from detailing studies on Web crawling and Web archiving as well as temporal indexing and future IR. We only report some studies proposed into this scope that could affect query-document matching models from a typical IR perspective. A good overview of these topics can be found in Campos et al. [11].

The remainder of this paper is organized as follows. In the next section, we remind some definitions that will be of interest throughout the rest of the paper. Section 3 describes our general categorization of temporal information access methods. In Section 4, we review the state-of-the-art work based on this categorization. Then, we discuss how the evaluation of T-IR approaches is carried out in the literature and we present some of the most prominent evaluation tracks that focused on T-IR models and techniques. Finally, we review the research trends for future work.

2. Definitions and terminology

Query. A query expresses a user's information need using generally a set of keywords [13, 14]. If a query contains explicit temporal expressions, it is classified into time-sensitive queries. However, some time-sensitive queries may not contain any temporal expression and may also refer to different points in time. In this case, a temporal query understanding is crucial in order identify the time period to which the query refers to.

Document. A document is frequently represented through a set of keywords [14] regardless of their origin of generation, e.g. a collection of Web snippets, (micro)-blogs, bookmarks or even images or videos. A document could incorporate temporal signals, including temporal expressions, document publication time or document last-modified date, etc. In this work, we consider a document as a Web page or news article regardless of the generating source.

Time. Time has been a prime motivation in many research areas. One of the most appropriate definitions for the IR realm is that defining it: 'as a measure in which events can be ordered from the past through the present into the future, and also the measure of durations of events and the intervals between them'[2]. Events may be defined as items, facts or changes arranged in a time order and presented in texts, tables, charts or timelines. In traditional IR, time is often represented by documents timestamps (document
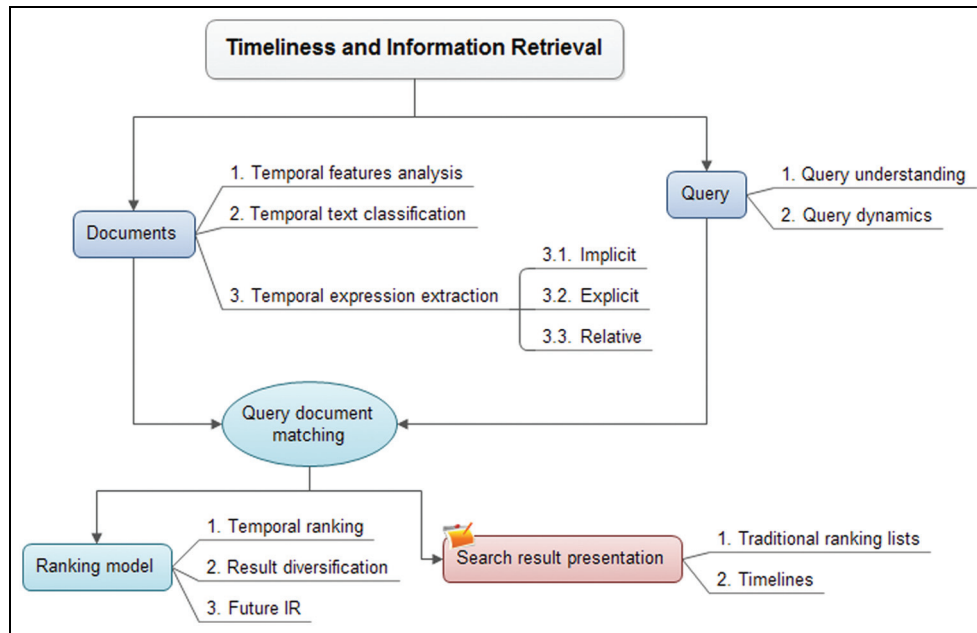
Figure 2. General framework of the T-IR area.

creation time), query submission time or temporal expressions contained in documents or queries. The time dimension may be exploited within two main relevance criteria:

1. *Recency*. Recency is a relevance criterion used in ranking models in order to boost newly published documents rather than older ones [15]. Recency could also refer to some types of queries where the user expects documents which are both topically relevant and fresh (timely). This criterion is usually computed as the difference between the document's publication time and that of the query timestamp.
2. *Freshness*. There is no clear distinction between freshness and recency in the literature. Freshness can be interpreted in different ways depending on the nature of queries. For instance, for queries related to news, freshness is more concerned about documents reflecting new information [16]. However, for non-temporal queries, freshness is interpreted as the recency of pages.

Temporal Information Retrieval − *Temporal Ranking*. Temporal IR is a research area that aims to retrieve temporal relevant documents. Mathews and Kanmani [6] identified two interesting subareas in this field: Time-Aware Retrieval Models (T-R Model) and Temporal Ranking (T-Rank). In time-aware IR, there are two main temporal features: document timestamp and temporal relevance of the document content. Whereas, in temporal ranking, time is either taken into account directly in the ranking model or through models that aim to understand the query nature and address the ranking according to its type. In the remainder, we do not make distinction between temporal ranking and time-aware IR.

3. Time-aware information retrieval: A general classification

Regarding the recent research focus on time and relevance, we can distinguish three main categories of approaches: time at the query level, time within documents and time as a factor within a ranking model. Our classification comes from the general schema of typical IR models. In Figure 2, we show each of these components through their connection with time.

Besides the three major components emphasized in Figure 2, we also present the impact of time on search result presentation. Along with each component, we highlight the most commonly prominent time-dependent applications. For instance, one of the widely encountered problems when it comes to dealing with time at the document content level, is the extraction of the temporal expressions therein. Similarly, a close problem when responding to time-(in)sensitive queries is the query understanding [17], in order to identify relevant time periods corresponding to queries. To further

ease the presentation of our overview of studies on T-IR regardless of the IR application at hand, we briefly describe the levels at which 'time' may be incorporated or exploited as a relevance criterion or a feature to tailor the search results:

- *Time on the query level*: in this category, the focus is specifically put on the query. The primary aim is often to understand the temporal intent behind user queries and tailor the retrieval model based on the query dynamic or query type. The main challenge here is to identify the time-period the queries refer to and to deal with query temporal ambiguity, i.e. especially when considering queries without any temporal expression or those that could refer to many time points [18]. This category witnessed successful applications to areas as diverse as temporal query understanding, temporal dynamics and query processing, etc. Exploring these issues may enhance the retrieval effectiveness and help understand the information goals behind queries. For instance, modelling and predicting the change in queries that people issue (query dynamics) and the number of times that URLs are clicked for that query over time could help in estimating trends, periodicities and predicting the future.
- *Time on the document content level*: in this second category, research studies handle temporal information existing on the document contents level. The first challenge here is to find the best ways to represent the temporal expressions contained in documents in natural language. Then, the second challenge consists in identifying, extracting and normalizing these temporal expressions [19]. These temporal descriptors could be extracted and anchored in time for explicit document exploration purposes. We could distinguish between explicit, implicit (e.g. names of holidays) and relative temporal expressions [10]. It is worth noting that difficulties often arise when dealing with implicit (e.g. 'Independence day') or relative temporal expressions (e.g. 'today'), while the use of explicit temporal expressions is straightforward.
- *Time on the document matching level*: this category gathers works on temporal retrieval area or time-sensitive ranking, where temporal relevance attributes are encompassed into the core of the ranking model [20, 21]. Studies in this category are closely related to those laying within both the first and second categories, given that time-based ranking models could be induced thanks to the temporal nature of queries. Here, the challenge is to leverage the traditional IR models where documents are basically ranked according to their creation time to more enhanced time-aware ranking models. To tackle this challenge, models often integrate these features into query-likelihood language modelling approaches or linear combination mechanisms of textual and temporal parts of the documents and/or queries [21]. On the other side, future IR, which fall into ranking models as depicted in Figure 2, focuses on future related information such as news about a new product to be released, information about an upcoming event, etc.

4. Literature review: Analysis of time-aware information retrieval approaches

In this section, we present a comprehensive overview of almost existing time-aware IR access methods. More specifically, we are interested in temporal-aware information access models especially from the query and ranking level perspectives. To do so, we select the most relevant research papers in the literature according to their impact on timeliness across various related domains. Figure 3 illustrates the time relationship to different temporally-aware IR areas. The
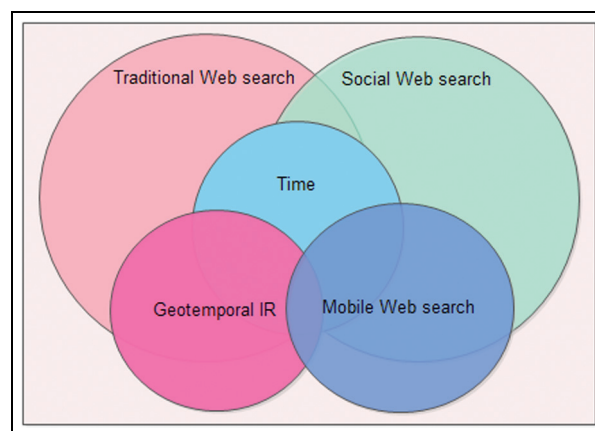

Figure 3. Time relationship to various IR areas.

figure depicts to what extent time is exploited in each IR area. Considering notions of clusters, each circle represents a range of work done in a particular IR area. The circle time represents any temporal evidence or feature that could be of benefit in IR systems. As can be see seen from Figure 3, many works fall under more than one IR task. For example, work on social Web search may be considered as part of social, traditional and geo-temporal search. Temporal evidence may be used at the intersection of all of these IR areas. The cluster size presents the amount of significant research papers carried out in each field. Although Figure 3 is not drawn to scale by any measure, it only provides a schematic view of how time could be encompassed in various IR tasks. In the following, we will discuss the implications of the integration of this concept on the performance of IR systems in each of the presented areas.

4.1. Traditional Web search

In traditional Web search, people usually aim to find relevant information satisfying informational, navigational or transactional goals [22]. Informational search is intended to find information about a specific topic, whereas in navigational search, users often look for a specific website in mind. Transactional search is intended to perform a Web-mediated activity. Whenever ranking documents in response to a user need, there are a variety of temporal query and document signals that may be used to tailor the user search results.

In the following, we describe the relevant work carried out in the scope of traditional Web search from the query, document and temporal ranking level perspective. Then, we report a summary of all of these works depending on the class of query, the ranking model and the used datasets.

4.1.1. Time at the query level.
Recent research on Web search showed that a lot of Web queries contain either implicit or explicit temporal expressions. An analysis of Web user query logs shows that 7% of Web queries have temporal intent implicitly provided [17]. Jones and Diaz [18] identified three temporal general types of query profiles: atemporal, temporally unambiguous and temporally ambiguous. Atemporal queries refer to non-time sensitive queries (e.g. 'Amazon'). While temporally unambiguous are those referring to a precise time period (e.g. '2015'), temporally ambiguous ones are those pointing to different points of time (e.g. 'US open tennis'). For example, a query containing an explicitly stated time period such as 'Australian Open Tennis Championships 2015' might only be matched with documents published at this particular period '2015'. However, for many queries, this time expression is not usually specified. Nunes et al. showed that only 1.5% of queries are explicitly provided with temporal criteria [19]. Consider the query 'Michael Jackson', the latter might be (implicitly) about articles published in (or before) the period of 25 June 2009, the date of death of Michael Jackson. However, it can also be about documents (articles, news, videos) related to his death anniversary or even his new song 'Never Felt So Good' first released on 2 May 2014. Thus, the relevant time interval for the query remains unclear. Instead of placing the burden on the users to explicitly deal with time when querying, a great deal of research has been done in order to understand the temporal nature of queries and systematically handle all of its families.

Roadmap.
In the following, we present the most representative work regarding two main research directions, namely query understanding and query dynamics, with respect to the schema presented in Figure 2 (*Cf.*, Section 3). The impact of these studies on the ranking process will be discussed at the document-query matching level (*Cf.*, Section 4.1.3).

The first line of research concerning query understanding aims to understand the temporal intent behind user queries. We present two main types of queries that have been extensively studied in the last decades.

Recency-Oriented Queries. Recency queries are those occurring right after breaking news or more recent events such as 'Charlie Hebdo newspaper shooting', 'Ebola virus disease', etc. The work of Li and Croft [20] was among the first researches in handling recency-oriented queries. Queries are classified based on the temporal distribution of documents over a collection of TREC queries (from volumes 4 and 5). Dakka et al. [23] defined a time sensitive query over an archive of time-stamped news documents as: 'queries for which relevant documents are not spread uniformly over time but rather tend to be concentrated in restricted time intervals'. If the relevant time period for a time sensitive query is unspecified, the authors suggest the computation of time based probability $p(q|t)$ for each time t and query q using the query-likelihood model.

Periodic Oriented Queries & Bursts: Periodic queries are those occurring based on a recurring basis, e.g. 'election', 'Cannes film festival', etc. Vlachos et al. [24] represented these classes in terms of coefficients in a Fourier transform based on time series. These time series are built for each query word or phrase from a large collection of MSN search engine query logs. Important periods are then identified by using outliers according to an exponential distribution. The significant periods are those having powers that deviate from the power content of the majority of the periods. After the extraction of important periods, the authors proposed an interactive burst discovery that is mainly based on the computation of the moving average (MA). Closest to periodic queries

are efforts on query bursts detection. Subasic et al. [25] defined a query burst as: 'a period of heightened interest of users on a topic which yields a higher frequency of the search queries related to it'. For instance, the query 'Harrison Ford plane crash' suddenly becomes among the most trending queries in the USA[3], on 5 March 2015 after a plane crash in which the actor Harrison Ford was injured. Subasic et al. [25] grouped query bursts into three classes: (1) bursts that fade-out completely after a certain period; (2) bursts that create new topics; and (3) bursts on existing topics. The authors suggest interesting findings for content providers as well as search engines in order to satisfy user's temporal needs and provide the most relevant results. For instance, to capture user's attention on emerging topics, content providers should either publish early or target query bursts on emerging topics that did not exist before.

In the second line of research that concerns query dynamics, the focus is mainly put on the usefulness of queries temporal patterns on research areas such as event tracking or temporal intent detection. In the sequel, we present the studies that have investigated the elicitation of users' search behaviours as well as the work on how temporal queries analysis may be of benefit in tracking events.

Event Tracking and prediction. Ginsberg et al. [26] showed that Google Web search queries may be a valuable source of information to track influenza-like illness in a population. The authors computed time series of about 50 million of the most common search queries in the United States and designed a method to select the influenza-like illness queries. This work is followed by Diaz [27] who studied the newsworthiness of queries. The author developed efficient methods for identifying queries that are related to breaking news. News intent is identified by means of either query dynamic understanding or using click-through prediction. Furthermore, the author investigated the impact of integrating search results from the news vertical search into Web core search results.

User behaviour prediction. Radinsky et al. [28] examined how to model and predict user behaviour over time using time series. They explored different dynamics of Web behaviours including trend, periodicity, noise, surprise and seasonality detection. The authors showed that the proposed time-aware modelling of user behaviour could be incorporated in many search-related applications such as query click prediction or query-URL prediction. Joho et al. [8] carried out a user survey in order to shed light on temporal characteristics of user Web search activities. This study suggests that an interplay of seasonal interests, technicality of information needs, target time of information, re-finding behaviour, and freshness of information can be important factors for the application of temporal search. Among the significant findings elicited from the conducted analysis is the importance of the recency of retrieved information.

Query classification for user's query temporal intent detection. Search intentions of two queries could be considered as similar as far as they present similar temporal profiles [29]. This issue has been studied by Radinsky et al. [30] through a Temporal Semantic Analysis (TSA) based approach. The authors used TSA to represent the semantics of natural language words, and to compute the semantic relatedness between words. Therein, a word is represented by a weighted vector of concept time series derived from several sources such as Wikipedia, Flickr and del.icio.us. Experiments on the New York Times archive showed that TSA leads to significant improvements in computing query-terms relatedness.

Earlier, Asur and Buehrer [31] studied whether the different frequency profiles could be of use to improve query classification. Three different families of queries have been investigated, i.e. Navigational, Adult and News queries. To perform query classification, authors have computed trends in query-clicks over time. In the same context, Ren et al. [32] proposed a time-based query classification approach to automatically detect user's query temporal intent. They identified several time-related latent semantics under queries, such as full-time intent, most recent time intent or burst time intent. Then, queries were grouped according to their temporal intents based on a query taxonomy and their search frequency distributions over time in Web query logs. The work presented in Costa et al. [33] is also worth noting since it designed ranking features that exploit correlations between archived data and relevance. The author's hypothesis was that closer periods are more likely to hold similar Web characteristics. For this purpose, a temporal-dependent ranking framework exploiting the variance of Web characteristics over time has been proposed. Their study has been performed on a Web archive containing over 14 years of Web snapshots and it has been proved to be very effective in ranking problems over Web archives.

4.1.2. Time at the document content level.
In the following, we overview the research studies that tackled the problem of time and relevance at the document content level.

The main challenge here consists in the identification and extraction of temporal expressions contained on the Web documents. As discussed in Section 3 (*Cf.* Figure 2), temporal expressions may be classified, respectively, into implicit, relative or explicit. The extraction of these expressions is a crucial step in identifying the focused time for Web pages (i.e. adequate time associated with Web pages) and thus in improving the retrieval effectiveness of T-IR models. While explicit temporal expressions are often easy to identify (e.g. from a calendar), implicit temporal expressions are difficult to deal with, as they require a reference time. For instance, a temporal expression such as 'today' needs to be matched to the right date to which it refers in the document (e.g. document publication or last modified time). However, when the document creation time is missing, implicit temporal expressions should be matched with other temporal expressions

from the document. This task is part of the tasks of so-called *temporal taggers* and is the most challenging one in the temporal document annotation process. To deal with this problem, there is some crucial pre-processing steps that should be performed on text documents: (1) first, tokenizing information mentioned in documents into sets of phrases; (2) identification of the sentences; (3) POS: assigning part of speech to the set of tokens; and (4) NER: Named-Entity Recognition. Once the temporal expressions extracted, they shall be normalized. One of the most used temporal taggers that have been proposed is HeidelTime[4]. HeidelTime is a multilingual temporal tagger that is able to correctly identify the reference time of temporal expressions (e.g. 'today', In 'November') [34]. The extraction is performed depending on the domain of the documents that are to be processed: news, narratives (e.g. Wikipedia articles), colloquial (e.g. SMS, tweets) and scientific (e.g. biomedical studies). When evaluated in the TempEval-2 challenge[5], this system was the best performing one among eight other competitors.

Lin et al. [9] employed the GUTime tool[6] to extract temporal expressions and determine the focused time of documents. The authors classify temporal expressions into two classes: (1) global time; and (2) local time. While the latter takes the publication time as the referent, the former makes reference to the narrative time in the text, which also depends on the document context. After extraction, these temporal expressions as well as the corresponding focused time are used within a temporal search engine.

Detecting temporal expressions was also the aim of some evaluation tasks such as SemEval [35]. For instance, in the SemEval-2007 Task 15: 'TempEval Temporal Relation Identification' [35], the aim was to propose new methods for the automatic identification of all temporal referring expressions, events and temporal relations within a text. Given that this task is judged to be difficult, the organizers of this evaluation challenge proposed three tasks to simplify it to the participants. The first one (Task A) addresses only the temporal relations holding between time and event expressions that occur within the same sentence. The second task (Task B) addresses only the temporal relations holding between the Document Creation Time (DCT) and event expressions. Assigning the temporal relation between the main events of adjacent sentences was the aim of Task C. To address these tasks, several NLP based approaches have been successfully applied [35–37]. These NLP methods include the part-of-speech (POS) tagging tools, NER and coreference resolution systems, to cite but a few.

The integration of the temporal information identified from documents into the Web search will be presented in the remainder.

4.1.3.   Time at the ranking level.   In this section, we show how time could be of use at the query document-matching level, as sketched in Figure 2 (Cf., Section 3). The most commonly used approach to rank temporally relevant documents is to combine the topical relevance with temporal relevance features. One of the most used techniques is to incorporate the temporal expressions into a language modelling framework [21]. The documents are then ranked according to their probability of generating the query. Another promising idea that is based on the assumption that queries (popularity) and documents (content) change over time [38] consists in analysing these changes to understand the search experience. This helps in giving insights into what the document is fundamentally about and how temporally sensitive it is, and this can be appropriately matched to the user's intent [28, 39]. This issue has been mainly tackled using time series analysis [28, 39–41]. Thus, the bursts or highest peaks in the time series of relevant documents are coupled to the events causing the interest for the topic of these documents.

Roadmap.   We split the dedicated related work into two significant lines of research, according to the way time is defined and in which context it is exploited.

- *Time as a Dimension of Relevance*: Here, we present studies in which time is defined as a dimension of relevance and combined with other relevance factors. Li and Croft [20] proposed a time-based language model framework that incorporates time into both query likelihood language model and relevance-based language model. The scores of the most recent documents are boosted to privilege recent articles over older ones. Campos et al. [42] defined a novel temporal ranking model, called GTE-Rank, which takes into account both content importance and temporal distance to re-rank Web snippets. The authors studied the impact of the incorporation of the temporal classification model on the retrieval effectiveness and proposed a set of measures that enable to push down non-relevant ones. Berberich et al. [21] studied the inherent uncertainty problem in temporal expressions of users' queries (e.g. 'in the 1990's'). Interestingly enough, the authors showed how temporal patterns can be integrated into a language modelling framework and how they could be leveraged to improve the retrieval effectiveness for temporal information needs. Then, documents are ranked according to their estimated probability of (temporally and textually) generating the query.

In the same line of research, Dakka et al. [23] defined the relevance of a document as a combination of topic similarity and time relevance. The authors proposed a general framework to incorporate time into the retrieval task in a principled manner. For a given time-sensitive query over a news archive, the approach automatically identifies significant time intervals for the query and uses them to adjust the document relevance scores by boosting the scores of documents published within the important intervals. The proposed model ranks documents in decreasing order of their probability of relevance based on their temporal $(P(t|q))$ and topical $(P(q|d))$ relevance: $P(dt |q) = P(d, t|q) \propto P(q|d)P(t|q)$, where $P(q|d)$ denotes the query likelihood on document d, and the second factor $P(t|q)$ conveys the relative importance of the time point t for the whole query q. While the first factor (i.e. topical relevance) can be estimated using a standard text-based query likelihood method, the second factor (i.e. temporal evidence) could be estimated using different methods, e.g. the maximum likelihood model, which is defined as the normalized sum of the relevance scores of documents published at time t for query q.

Recently, Lin et al. [43] studied the extraction of the focused time for Web pages (i.e. the most appropriate time associated with it). After extracting temporal expressions and determining the focused time among all the extracted information, the authors suggested a score model based on both the frequency of temporal information as well as the containment relationship among temporal information. As for most T-IR models, the authors combined the textual similarity and the temporal similarity between queries and documents within the ranking process. A Time-Aware Search Engine (TASE) has also been proposed for that purpose. Harper and Chen [44] assume that the World Wide Web is in constant evolutionary change and hypothesize that the text similarity of a page to a query does not change over time, unlike its importance which effectively changes. Pages are ranked based on their text content and the included temporal information as well as the page importance. As an application, the authors introduce a model for PageRank, which considers the relevance between the text of a page and a query, the temporal information, as well as the outstanding score of the page.

On the other hand, a wide range of work focused on the temporal properties of standard IR measures and investigated how the corpora changes over time may be exploited in document ranking. Perkiö et al. [45] assumed that the relevance of documents is not static over time and proposed a statistical topic model that takes into account the temporal behaviour of the underlying statistical properties of the documents. The basic assumption of this work was that: 'the ranking of the results for a query Q at time t should promote documents whose most prominent topics are the same that are the most active within the whole corpus at time t'. The authors modelled this hypothesis based on a temporal adaptation of the TF-IDF weighting formula. The experimental evaluation using a news corpus have shown encouraging results when relying on the temporally adaptive ranking model.

More recently, Efron [40] used temporal behaviour of terms in the entire collection, rather than in individual documents to derive the weight of each word to be used during the document ranking. Unlike traditional measures of term importance, in which rare terms receive higher weights than common terms receive, Efron defined a term's time-based weight as a function of its behaviour over the lifetime of the corpus. The author proposed three time-based measures relying on time series analysis. These methods are shown to yield statistically significant improvements over two baseline retrieval models, using data from several tasks undertaken at TREC. In the same line of research, Karkali et al. [46] extended the traditional *Inverse Document Frequency* (IDF) measure and show that it could be efficiently used as a measure for document novelty on text streams. Given that document streams may shift over time, the authors inject a time decay technique in the scoring function in order to give recent documents higher weight values than older ones when performing novelty detection. Experiments within a Google news dataset[7] as well as a Twitter corpus showed that using a temporal variant of IDF lead to significant improvement in a document novelty detection task. Aji et al. [47] built upon the work of Efron [40] and proposed a novel term weighting scheme that rely on the revision history of documents. The main intuition of this work is that the importance of a term in a document can be measured by analysing the revision history (e.g. the edit history of a page in Wikipedia) that incorporates a significant amount of human editors' knowledge about the world. That is, as some documents may present extensive changes due to revisions that reflect news events for a given topic, important terms may be introduced early in the life of a document. For example, in the earlier revisions (June 2006) of the Wikipedia page devoted to the movie 'Avatar', there was little editing activity and little content, the page simply mentioned that James Cameron would direct the film. However, in October 2006, there is a significant change to the content as new details about the plot, budget and development are added. The authors argue that term weights should be adjusted by incorporating burst history of the movie page. This observation gives rise to a ranking model that combines the term importance information distilled from the revision history with two conventional statistical approaches, namely, BM25 and the generative statistical language model (LM). Experiments using the INEX 2009 ad-hoc track evaluation and a set of TREC ad-hoc queries show significant improvements over the conventional BM25 and LM retrieval models. However, the proposed model was shown

Table 1. A synthetic overview of empirical studies on temporal query understanding.

Main references	Query class	Ranking model	Dataset
Li and Croft [20]	Recency	Temporal language model	TREC queries 301–400 over collections from TREC volumes 4 and 5
Vlachos et al. [24]	Bursts, periodic	Time-Series Analysis	MSN search engine query logs
Jones and Diaz [18]	Atemporal, temporally unambiguous, temporally ambiguous	–	News articles and Web search query logs
Diaz [27]	–	Language model	News articles
Metzler et al. [17]	Implicit year qualified queries	–	Query logs
Asur and Buehrer [31]	Navigational, Adult and News queries	–	–
Berberich et al. [21]	Temporal expressions	Temporal language model	New York Times Annotated Corpus, English Wikipedia
Subasic and Castillo [25]	Bursts	Temporal language model	Query logs from Yahoo
Shokouhi [51]	Seasonal	–	Search logs
Kulkarni et al. [39]	Spikiness, spike shape, periodicity, overall trend	–	Query logs from the Bing search engine, daily Web crawl, periodic human relevance judgements
Radinsky [28]	Trend, periodicity, noise, surprise, seasonality	Time-Series Analysis	Query and URL logs from Bing
Dakka et al. [23]	Time-sensitive	Temporal language model	News article datasets
Jabeur et al. [52]	–	Bayesian Model	TREC 2011 Microblog track dataset
Efron [40]	–	Time-Series Analysis	TREC Tipster collection, TREC robust track

to be outperformed by the baselines for ambiguous queries which spans multiple topics such as the query 'earthquake prediction' that include geology, natural disasters, etc. Following the same direction, Nunes et al. [48] used the same document's revision history as a source of temporal evidence and propose several different term-weighting measures. The basic idea of the approach is to give higher weight values to terms that have existed for a longer time in a document, since its first version should be valued higher than a term that was introduced only in the latest revision made. The proposed method has been evaluated using a collection of Wikipedia articles. The results showed that the proposed revision term frequency and revision term frequency spans perform the traditional TF measure.

- *Time as a Query Topic*: A novel interesting IR problem that has been recently studied consists in considering the time as a query topic instead of using a text query as in traditional search. The work of Kim et al. [49] was among the first ones that tackled such issue. The authors have considered the query as a time series for a time period. The retrieval task is to find relevant documents in a text collection of the same period, which contain topics that are correlated with the query time series [49]. After testing several ranking methods that rely on ranking text documents based on the assessment to what extent their terms are correlated with the query time series, they showed that their method supports users in finding the documents that are relevant to the time series queries, which could help them in analysing the variation patterns of the time series.

4.1.4. Synthesis. Prior research on the temporal dimension of search showed the importance of time in document ranking [12, 50]. Many sources of temporal evidence were presented ranging from document-based evidence to Web-based features. These research could be structured according to three types, depending on the main source of temporal evidence being explored [12]: (1) Link-based research uses links, both in-links and out-links, in a temporal context to refine information retrieval; (2) content-based research examines document's contents from a temporal point of view; and (3) metadata-based research uses evidence gathered from HTTP headers and external services (e.g. PageRank values).

This temporal evidence is exploited to improve IR on the Web. As previously mentioned, we have organized these work on three levels. Table 1 reports a synthesis of the discussed work on the query level, with the type of the

investigated query, the used model and the datasets used for evaluating the underlying research contributions. Most of the studied queries could be classified into two families: temporal and atemporal. The former are those for which the user's interest is about documents published in particular points of time in the past or recently published ones, which are about fresh events. To understand the temporal nature of queries, these studies exploit the distribution of the documents respective dates to identify time intervals that are likely to be of interest to the latter. For instance, when searching for 'earthquakes', the peaks in the distribution may indicate when 'earthquakes' occurred. Time series analysis is shown to be a good candidate to tackle this kind of issue [24, 41, 49, 51]. This solution is of particular relevance especially in the context of news collections and query logs. For instance, as reported in Kulkarni et al. [39], analysing the changes in query popularity, Web content and query intent could help search engines to decide when it is appropriate to interject news into the result page. These approaches use time series with a special focus towards identifying features by which changes in query-term frequency or popularity can be classified including spikiness, periodicity, trend, etc. [24, 28, 38, 51].

From the ranking perspective, an ongoing body of work exploit the temporal change of user's search behaviour and documents collections in order to enhance the retrieval task [28]. While some work investigated specific temporal year queries to modify the language model [17], others added temporal factors into the same model and relevance to re-rank documents results [20]. The term frequency changes over time has also been investigated to weight terms in dynamic collections of documents [40]. Another line of research has focused on improving the ranking of recent information within several IR tasks [27, 52]. The exploitation of the language modelling framework is straightforward and it is shown to be effective in solving such IR problems [27]. The temporal variant of this model has been successfully applied in a wide scope of applications ranging from real-time search, temporal summarization, topic detection and within different evaluation frameworks. It is worth mentioning the common interest on news archive collections and query logs. This explains their interest in many seminars dedicated to T-IR applications, such as the Time-aware Information Access temporal (TAIA) series [53, 54]. Most of the research done into this scope are based on Twitter and Wikipedia.

4.2. Social Web search

Our motivation behind a separate section for social Web search comes from the fact that social Web content differs from that of the classical Web content in several points [55]. First, the users' search intentions are slightly different given that in social search, users usually seek for user-generated social content. Second, contents are different, for instance, in social systems such as Twitter, tweets are short and do not change after being posted [55] while Web pages evolve and are richer. Moreover, time is gaining a surging interest in this area, given that this task is temporal in nature as the relevance of the search is time-dependent.

Roadmap. To the best of our knowledge, most of the work done in the scope of the temporal-dependent social Web search primarily focus on the ranking process, unlike traditional Web search where a spate of research is given to query understanding and processing (*Cf.*, Figure 2, Section 3). Thus, in the following, we report the studies that typically aim to exploit time in social IR settings. We split these studies according to the type of documents that users are seeking for.

Tweet search. Many temporal characteristics have been incorporated to account for tweet relevance [56–58]. Elsweiler et al. [57] tried to understand what people are really trying to achieve in Twitter. Their conclusion was that the search is motivated by: (1) temporal patterns (e.g. trending event information); (2) social content (e.g. searching for a particular person); and (3) topical content. The tweet search task is driven by a variety of criteria such as authority, topicality and recency of tweets. Time is represented by the recency criterion which is often defined as the difference between the time a document was published and the query submission timestamp. Jabeur et al. [52] proposed a Bayesian model that aggregated relevance factors such as the social importance of microbloggers and the temporal magnitude of tweets. The temporal magnitude of microblogs is estimated based on temporal neighbours that present similar query terms.

In Wang et al. [58], the authors made use of learning to rank methods in order to aggregate the relevance features in a microblog search setting. The authors divide features into two categories: entity-related and temporal-related features. Temporal features include some clues about the temporal distribution of documents, the average time of the whole collection, time interval between queries and a given document, etc. Then, the Listnet learning to rank algorithm [59] is used to combine all these features. The authors showed that learning to rank is efficient in combining temporal evidence into ranking models. Miyanishi et al. [60] proposed a concept-based query expansion method based on a temporal relevance model that uses the temporal variation of concepts (e.g. terms and phrases) on microblogs. Their model is based on important concepts that are frequently used within a particular time period associated with a given topic, which better discriminates, rather than words, between relevant and non-relevant microblog documents.

Event-episode discovery. This task focuses on the retrieval of new events from news documents. Wei et al. [61] have investigated the problem of integrating temporal information for event episode discovery. The authors developed a temporal based method to effectively discover event episodes from a sequence of news documents pertaining to a specific event. Two temporally enhanced metrics inspired from the classical $TF - IDF$ metric have been then proposed: TF-IDF Tempo and TF-Enhanced-IDF Tempo. The authors show that incorporating a time-decaying function can significantly improve the effectiveness of their event episode discovery technique.

Beyond the task of real-time search, some work have paid attention to the usability of some external sources such as Wikipedia in reflecting the latency of events in the real world. Osborne et al. [62] examined Wikipedia pages that exhibit unusual large spikes in page views. Then, they compared the obtained tweets and Wikipedia pages over textual and time dimensions to identify common aspects of both systems. The authors showed that there is a delay of 2 h between events breaking on Twitter and the time when users start to search Wikipedia for information about it. Close to this work, Whiting et al. [63] discussed the main content and meta-data temporal signals available within the website. The source and nature of each signal as well as its extraction have also been studied. Some datasets have been released in this work in order to support temporal research based on Wikipedia. Some approaches that work on event detection based on Wikipedia edits and linkage highlighted that the burst changes in both the linkage and content of a Wikipedia page in a particular time could be a result of a trending event related to the topic of the page. Based on this, the changed page is considered relevant if relevant changes can be observed in the interlinked pages as well [64]. Julianna et al. [64] have combined a textual relevance score with a temporal relevance score that measures the changes in time of both linkage and content. The authors claimed that their methods can directly be applied for any hyperlinked collection.

Tag recommendation. This task focuses on helping users in annotating documents on the Web. Zhang et al. [65] proposed a time-dependent tag-recommender algorithm that is based on the frequency and the temporal usage of a user's tag assignments. The algorithm models the temporal tag usage with an exponential distribution based on the occurrence patterns of tags over time. The intuition behind this was that the users' preferences towards tags could change over time. To assess a tag for a user-document pair, the authors first considered the user's general interest in the tag, then computed its recurrence probability based on the first- and last-time usage of the tag. The time-dependent algorithm has been tested on two real-life datasets extracted from Bibsonomy and Delicious. Experimental results have shown that the proposed temporal model outperforms state-of-the-art tag prediction methods.

Time-aware social search. Given the tremendous amount of users' generated contents in social networks, several works harnessed the available temporal features to improve the ranking effectiveness. Khodaei and Alonso [66] used a set of temporally-aware social signals and show how they can add another aspect to the existing social signals. The authors assumed that documents can be socially relevant to certain people (or networks) for specific time periods or with some temporal patterns. The time of (social) actions is used to categorize a user's social interests into five categories: (1) recent interests; (2) ongoing; (3) past; (4) seasonal; and (5) random interests. Then, social results could be filtered based on one of these categories. The authors show through experiments on Facebook and Twitter sample data that these temporal signals could be used in existing search engines. Inagaki et al. [67] proposed a set of click features to improve machine learned recency ranking. Among these features a temporal factor called *click buzz* that captures the spiking interest of a *URL* for a query. The latter favours *URL* that have recent interests to the user. In a closest social IR task, Badache et al. [68] proposed an approach based on social signals to estimate document's relevance. The authors made use of a recency criterion assuming that a resource is fresh whenever there are recent social signals associated with it. Freshness is defined as: 'a date of each social action (e.g. date of comment, date of share) performed on a resource on social'. The authors show that temporal features improve relevance ranking of conventional text search.

4.3. Mobile Web search

With the proliferation of mobile devices, recent trends in information technology show an emerging interest on mobile Web search. In a mobile IR setting, users are interested in relevant information while they are moving. Relevance often depends on the user's situation (location/time) and some social (surrounding persons) dimensions [69, 70]. As time became a paramount relevance factor in mobile search settings, we report in the following some research on time and mobile Web search.

Church et al. [71] examined users' goals behind mobile search and argued that taking advantage of temporal, location and preference-based contexts may be of potential benefit in mobile information scenarios. The authors carried out a 4-week diary study in order to understand the intent behind information needs and how those information change based on context. They found 8.4% of diary entries expressing information need including explicit temporal cues such as 'tonight, tomorrow, next week, etc.'. Interestingly, Church et al. found that most of the geographical information needs (30%) were temporally dependent, even though most of them did not include explicit temporal cues (e.g. 'Where is nice for lunch near Jervis St.?'). In this case, it is obvious that the user is looking for fresh information that is only relevant for this particular point of time.

In the same line of research, Teevan et al. [72] investigated the importance of location, time, and people in mobile local search behaviour. To do so, the authors analysed the results of a survey of 929 employees at Microsoft whom have been asked to report their most recent mobile local search. The authors highlighted three important aspects that emerged

from the data: temporal features, location and social context. Time was the most important aspect that influenced the topics people search for. For instance, people wanted search result to be near them in term of time (e.g. 9% of respondents did not care how long it would take to reach the destination, while 26% were interested in having it to be as close as possible) [72].

Yau et al. [69] combine situation-based adaptation and profile-based personalization into an IR model. A situation is represented as a set of past context attributes and/or actions such as: location, time, light, device, etc. A user profile includes a usage history and general interests that have been automatically learned using a modified Naive Bayesian classifier. Here, the time is simply represented by the date when the search occurs.

In summary, in most of research on mobile IR, time is used as a simple but important relevance feature that yields information about when a query is submitted and allows to tailor the search results to users accordingly.

4.4. Geo-temporal information retrieval

In this section, we present studies on geo-temporal IR that are particularly done at the document ranking level.

The spatio-temporal information exploration is also an emerging field which heavily relies on temporal information. Works on this area often combine spatial and temporal information [73–76]. Spatial information is normalized according to the given latitude/longitude and is extracted through the so-called *geo-taggers*.

Strötgen et al. [73] have proposed a new model to rank documents according to combined textual, temporal and geographic queries. Unlike previous works, the authors abandoned the independence assumption between the query parts by computing proximity scores. Therefore, the higher number terms and expressions fulfilling the different query parts occur close to each other in a document, the more relevant the document is. An extended version of the Okapi BM25 ranking model has been proved to be efficient for such a purpose. Mishra et al. [74] proposed a geo-temporal re-ranking approach based on counting the number of geographic and temporal expressions from an initial search. The documents final ranking relies on the initial probability coupled with weighting of the counts.

Later, Daoud and Huang [75] proposed a geo-temporal retrieval strategy that models and exploits geo-temporal context-dependent evidence extracted from pseudo-relevant feedback documents. The final score of the document is based on combining the content-based score, the temporal score, the geographic score, and the proximity score using a linear combination mechanism. Experiments on the New York Times news collection and the TREC 2004 robust retrieval track collection showed a positive correlation between the geographic and temporal query sensitivity and the retrieval performance.

4.5. Result presentation

With the rapid growth of the World Wide Web, data are generated at a furious pace. It becomes difficult to users to choose the documents to access, especially when it comes to temporal-dependent information needs. Consider, for instance, a user looking for information about the 'Charlie Hebdo newspaper shooting'. It might be possible that the user's information need is about the number or names of the victims and survivors after the attack. However, the user might also have a myriad of general interest about subsequent relevant information regarding the suspects, involved entities and political aftermath. General search engines and content providers simply return a set of documents ranked by their relevance to the query. However, presenting results in the form of a complete and updatable timeline would be a promising issue. More recently, Google search engine proposed a kind of timeline for some types of queries (Cf., Figure 4). In Figure 4, the Google search result on the query 'Oscars academy awards' shows some recent related news about the event, in addition to a table showing information about the winners of the current and previous editions.

Some news websites presented similar approaches based on manually created timelines to show news about trending events. For instance, in Figure 5, we show the timeline manually created by the CBC news Website[8] during the 'Charlie Hebdo newspaper shooting' event. We can see that events are chronologically ordered in time (hour and date), with an illustrative map showing the geographical locations of the most important incidents, as well as some additional audio recording about the events.

Timeline construction has recently been tackled by some challenges and evaluation tracks such as the Tweet timeline generation (TTG) task of the TREC Microblog track and the Cross-document Event Ordering task of SemEval[9] 2015. The ultimate goal of these tracks is to unveil a set of significant sentences in response to a query or a given entity that could be a person, an organization or even a product, and put them within a chronological time span. This visualization approach allows end users to have an overall picture about the information (events or temporal query) they are looking for. In the context of a social IR, Li and Cardie [77] have tackled the problem of timeline generation for Twitter users. They proposed an unsupervised approach based on a Dirichlet Process Mixture Model to extract two significant types of

Figure 4. Google search results on the query 'Oscars academy awards', accessed in February 2015.

Figure 5. Timeline of the 'Charlie Hebdo newspaper shooting' event on the CBC News Website, accessed in January 2015.

information from the joint Twitter feed: personal tweets and time-specific information (to exclude time-general events). The authors found that topic models are not adequate for this task, as they harness word frequency as a main criterion for topic modelling, except for topics about celebrities that are widely discussed. Moreover, the time-specific feature is not found to be yet effective, as it fails to distinguish between short-term interests and personally important event topics.

So far, Tran et al. [78] proposed a timeline summarization approach for news articles based on a learning to rank algorithm. Due to the absence of public available datasets for this task, the authors manually constructed a collection of timelines[10] that are published by popular news agencies such as BBC, CNN, etc. A timeline summary is defined as a temporally ordered list of day summaries, where each day summary consists of a couple: a day's identifier and a set of sentences describing the main news event(s) that occurred on that day [78]. To measure the quality of summaries, the authors made use of three criteria: sentence relevance, sentence novelty and content continuity.

5.   Main tracks for the evaluation of time-aware information access models

A key innovation in the IR academic community was the early recognition of the importance of building and crucially sharing test collections [13]. This offer research groups the opportunity to evaluate their own IR systems and compare them with other published results obtained on the same shared collection. Motivated by the theoretical interest in the problems of temporal IR, there has been a spate of recent evaluation tracks and challenges that focused on T-IR access. These tracks present a focus for many time-aware international collaborative research exercises including temporal summarization, timeline generation, semantic evaluation and information filtering. Some of these activities include, but are not limited to:

SemEval 2015 - Time and Space track[11]: this evaluation track includes five tasks: (1) TimeLine: Cross-Document Event Ordering; (2) QA TempEval; (3) Clinical TempEval; (4) Diachronic Text Evaluation; and (5) SpaceEval. The Cross-Document Event Ordering is a new task on SemEval that focuses on the timeline generation problem. QA TempEval revolves around the temporal question-answering (QA) accuracy and the evaluation of temporal information understanding. The task mainly consists in building a knowledge base for obtaining answers for temporal questions about the documents and then compare them to a human answer key. Clinical TempEval focuses on identifying and describing events, times and the relations between them in clinical text. The Diachronic Text Evaluation task proposes to tackle the problem of automatically identifying the time period whenever a piece of news was written. The SpaceEval task consists in identifying and classifying items from an inventory of spatial concepts based on a set of features including time and space measurements, locations, motions, etc. Most of these tasks require the identification and extraction of temporal expressions included in the given document collections.

TREC Temporal Summarization Track[12] *(TS)*: aims to return relevant documents about an event from a time-stamped document collection. TREC 2013 was the first iteration of this track. It consists of two tasks: (1) Sequential Update Summarization where participants should return relevant and novel updates (sentences) for an event; and (2) Value Tracking that intend to estimate values for a particular attribute for an event. Participants of this task made use of the TREC KBA Stream Corpus, which consists of a set of time-stamped documents from a variety of news and social media sources. Many temporal metrics have also been proposed to evaluate the proposed systems. Eight groups have participated to the TREC 2013 TS track, the best run is that of PRIS [79]. For the first task, the system used the hierarchical Latent Dirichlet Allocation (LDA) for sentence scoring, while it exploits a Conditional Random Field (CRF) model for the second task. For the TREC TS 2014 task, participants were also given another pre-filtered corpus which only contains documents from the 2014 event topic periods and they are asked to produce temporal summaries for 15 different events, spanning accidents, natural disasters, storms, shootings and protests.

TREC Knowledge Base Acceleration Track (KBA) [80]: the KBA track intends to help users expand and reduce efforts with the maintenance of knowledge bases like Wikipedia, by automatically filtering the time-ordered corpus for the documents that are highly relevant to a predefined list of entities [80]. For TREC 2012, the organizers selected a set of filter topics based on 29 Wikipedia target entities: 27 people and two organizations. Participants are asked to apply their systems to a set of hourly directories of corpus data in a chronological order. The goal is to find documents having a significant likelihood indicating whether the latter is citation worthy for a given entity. F-score and a utility evaluation measure are used to assess the submitted runs. The highest performing systems in KBA 2012 (11 participating teams, 43 runs) were based either on rich feature engineering from the KBA or on learning to rank algorithms (e.g. SVMs). TREC 2013 proposed two tasks: (1) Cumulative Citation Recommendation (CCR), the same as KBA 2012; and (2) Streaming Slot Filling (SSF), a new task that builds on CCR by specifying a slot name for each entity (i.e. detects changes to the slot values). For KBA 2013, 141 entities have been released and 15 teams (140 runs) took part of the track. The best performing run in the evaluation is based on machine learning methods. As the track focuses on identifying vital documents that contain timely new information, most of participant systems are based on temporal features to retrieve vital documents. For instance, Abbes et al. [81] used a burst-based factor relying on the assumption that the higher the number of matching documents in a short period is, the higher the probability of having vital documents would be. This factor is linearly combined with a Language Model-based factor.

NTCIR Temporal Information Access (Temporalia) Task[13]: offers two subtasks to address temporal information access: (1) Temporal Query Intent Classification (TQIC) subtask; and (2) Temporal Information Retrieval (TIR) subtask. While the former

Table 2. Evaluation of T-IR models.

Track (or Task)	Dataset	Temporal span	Availability	Evaluation metrics
SemEval 2015	Collection of news articles, wiki, blogs, Clinical data	Different times spans (from 1960 to 2014)	☐	MAP, Recall, Precision
TREC TS	TREC KBA Stream Corpus	From October 2011 through mid-February 2013	☐	EG metric, coverage, EL metric, Harmonic Mean of normalized EL, Latency Comprehensiveness
TREC KBA	TREC KBA Stream Corpus	From October 2011 through mid-February 2013	☐	F_1 accuracy, Scaled Utility
NTCIR Temporalia	LivingKnowledge news and blogs	2014	☐	Precision, nDCG, Q-measure
TREC Microblog TTG	TREC Microblog dataset	2014	☐	Cluster precision, Cluster recall

aims at classifying queries into four predetermined temporal classes (past, recency, future and atemporal) based on their implicit or explicit temporal intent, the latter focuses on retrieving a set of documents in response to a search topic that incorporates a time factor [82]. Compared to the TREC temporal summarization track where the scope is related to the ranking documents for the 'recency' class of queries, the Temporalia task is limited to information about concrete past events or to a particular type of attribute-like information (e.g. numerical value). The task participants are given a document corpus, called 'LivingKnowledge news and blogs annotated sub-collection' collected from blog and news sources. For the Temporalia 2014 task, eight teams have participated with 36 runs. The retrieval effectiveness is evaluated by the precision and nDCG at 20. The main outcome from the TQIC subtask was that past queries were found to be quite easy to classify (73%) while recency queries were the most difficult (56%). The best performing system used logistic regression classifiers and SVMlin classifier to predict temporal query classes [83]. The team's methods were also based on additional data from the AOL 500K query session dataset. For the TIR subtask, the best performing system was a baseline submitted by the organizer team relying on a BM25 weighting scheme and using some fields that are given in the topic descriptions. One of the outcomes from this subtask, as well as for the TQIC subtask, suggests that it is difficult to generate effective models for all temporal classes based on a single IR method. For future research, the task organisers suggest the extension of the TIR subtask with multi-document summarization, which consists in creating temporal summaries from many documents including historical order of peoples, organizations or events [84].

Tweet timeline generation (TTG) task of the TREC Microblog track [85]: is a new task firstly proposed on the 2014 TREC Microblog track and tries to respond to the following: 'I have an information need expressed by a query Q at time t and I would like a summary that captures relevant information'. Two challenges are to be addressed for this task: (1) detect and eliminate redundant tweets (novelty); and (2) automatically identify the number of relevant tweets to return. In the 2014 TREC TTG, 15 teams have participated with 50 runs. Participating systems are evaluated based on, respectively, recall, precision and F-measure. The best performing system proposed a graph-based ranking method representing tweets as nodes and relying on similarity measures to compute the relationship between the tweets in one hand, then the tweets and the query on the other hand [86].

Several T-IR models have been proposed and evaluated in the context of the above challenges. We present in Table 2 some of the proposed approaches with the document collection as well as the associated metrics used for evaluation. As shown in Table 2, news archives are the most used data collections to evaluate the T-IR methods.

6.   Research trends

Despite the great advance in the T-IR domain, some potential areas remain unexplored. We present in the following, some interesting avenues that could be investigated in the future.

Data sources and evaluation. With the pace at which the social Web grows, we argue that recency queries have almost the same overall trend over diverse news collections. As previously discussed, many sources have been used in the T-IR realm, the primary ones are Wikipedia and news sources. However, users tend to react to trending events across various social networks (e.g. Facebook, Twitter, Youtube, Reddit, 4Chan, Imgur, etc.). Thus, we believe that the relevance of one document from one document collection may provide useful information about the relevance of other documents from other collections with similar content that were published around the same time. This requires novel methods just like those proposed in the data aggregation field, but that leverage the traditional information presentation paradigm toward more time-sensitive search results. Some news websites are dealing with this issue, but the content remains manually created.

Another promising direction that recently emerged consists in proposing relevant search results published in diverse times of interest to the query. This issue was the main topic of temporal diversification approaches [87, 88]. In addition, a challenging constraint that should be taken into account is the comprehensiveness of the provided content regardless of the used data sources. A good search result should be easy, succinct and refer to summarized content that does not require cognitive effort for the user.

Considering the evaluation issue, despite the ever-growing maturity of T-IR approaches, most of the proposals are individually submitted and evaluated with specific datasets and evaluation methods to prove its superiority to previous approaches. We advocate that the superiority of a proposed model should be objectively tested with a whole collection of datasets that are issued by various sources in order to avoid the data bias problem. Moreover, another fundamental limitation that may be faced in this context is the evaluation scenario which is unrealistic to some extent. Consider for example the TREC KBA evaluation track, the latter asks the participants to automatically filter the whole time-ordered dataset for the documents that are highly relevant given a list of entities (topics). However, in real-time scenarii, a system is not given any future evidence or information about the entities or the events that may occur in future. This is not the case here as far as the system participants may have access to future information such as term statistics over the entire corpus, which would not have been available after the timestamp of a given topic. Needless to say, this issue exists in many real-time documents collections that are based on the Cranfield-style evaluation paradigm [89]. Another concern that is worth emphasizing is the system comparison procedure. Due to the complexity and large volumes of most document collections, the IR systems are facing many other systems engineering issues beyond the proposal of efficient retrieval models, such as choosing the tokenization method and the stemming algorithm, among others. This leads to different systems implementations which make it difficult to compare them only in term of effectiveness. Some of these issues have been addressed by the evaluation-as-a-service model [90], which was firstly implemented at TREC 2013 to address restrictions on data redistribution. The premise of this model is that instead of distributing the document collection, the track organizers provided a service API with which all participants could accomplish the evaluation task. This provides a solution to all the systems engineering issues, and consequently, the participants are only concerned about the retrieval models. However, this solution is not generalized yet and it is only used in the TREC Microblog track. In the same line of research, some interesting research proposed the so-called Living Labs as an attempt to further away from the Cranfield-style evaluation paradigm, and make evaluations more 'realistic'. The basic idea of living labs for IR is to provide a central and shared data repository and evaluation environment giving researchers the data required to test and evaluate IR models as well as methods and systems [91, 92]. Even though this idea is novel and worth interesting, it has not been operationalized yet for T-IR.

Time and Stream analysis. As previously mentioned, most of T-IR systems are evaluated on news articles and real-time data collections. Most of these datasets are considered as streams, as it is the case for the TREC KBA corpus or Twitter data. Analysing and proposing efficient methods for such data flows is a difficult task. Although many research papers have been proposed for tackling these problems, most of the techniques are either complicated to be implemented or oversimplified that cannot yield reasonable performances in handling such continuous volumes of data [93]. Moreover, as time is the essence in these datasets, it should be thoroughly considered at the three levels discussed in the paper (query, document content and the model). Novel hybrid techniques that are derived from the already existing approaches will be required and they should be designed specifically to gauge with the overflowing news streams.

Presentation of time sensitive results. Displaying relevant and visually compelling temporal results is an interesting means to further enhance the functionality of current T-IR systems. Recently, there has been several attempts into this insight. Google's news timeline[14] is an example of such applications. The presented news could be explored by time and returned from different news sources. Another similar interesting application is that of Newsmap[15], which visually reflects the constantly changing landscape of the Google News aggregator. The goal is to provide an accurate visualization tool that divides information into recognizable bands which reveal the underlying patterns in news reporting. The systems show news segments in constant changes around the globe. Nevertheless, these attempts need to be enhanced and studied to assist users seeking for urgent needs or news summaries. This avoids users from exploring pages with complex and lengthy textual descriptions and obfuscating advertisements and graphics [94]. The TREC temporal summarization track as well as the TREC Microblog timeline generation task are suitable tracks that would hopefully allow tackling these issues. However, with the proliferation of mobile devices, new requirements for results presentation are imposed due to the resource-constrained environments of these devices (size of the screen, energy consumption). Unfortunately, progress towards realizing dynamic user exploration techniques has been very limited.

Ranking models. T-IR ranking methods are mostly based on linear combination models. However, most of these methods: (1) assume the independency of the query terms; and (2) do not take into account the possible dependencies between the considered relevance factors. Though some probabilistic models based on Markov random field [95] account for term proximity, they are not exploiting the temporal information contained in query terms and documents. For instance, people have tended to talk about 'Jennifer Lawrence iCloud leak' only after the massive hack of the iCloud celebrity accounts, which happened in August 2014. That is, there are terms that are highly correlated only on a given period. Computing their scores regarding this information could boost the retrieval effectiveness. For instance, Figure 6 illustrates the temporal distribution of the terms 'Jennifer Lawrence', '4chan', 'celebrity', 'leak' and 'iCloud' worldwide in the past 90 days using Google Trends. Figure 6 shows a high correlation on the date '1, September 2014', i.e. after the recent iCloud security breach leading to hundreds of intimate celebrity pictures leaked and posted on the *4Chan* website. Such term dependencies have not till yet grasped the interest of the community. This dependency property has been investigated between the relevance dimensions, including the 'time' as a factor [4, 5], but the temporal factor was represented by the basic recency relevance criterion. This leads to the assumption that documents, that are topically
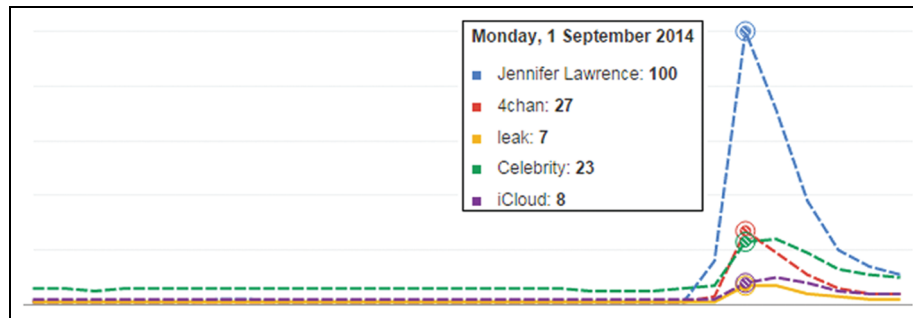
Figure 6. Google trend results on the terms 'Jennifer Lawrence', '4chan', 'celebrity' and 'leak', accessed in September 2014.

relevant for all the query terms and published in these bursty time periods, are more likely to be relevant in response to the whole query. Estimating the distribution of the query terms in the document collection considering the terms-temporal dependencies may improve the retrieval effectiveness.

Another interesting line of research that could be tackled is to supplement the standard challenges of ad-hoc retrieval with issues from topic detection and tracking, temporal summarization, novelty detection and other related areas given the resemblance of the proposed models within these tasks. For instance, many models proposed on topic detection and tracking have been successfully applied on temporal summarization. As a matter of fact, unifying those techniques could potentially lead to powerful systems.

On the other hand, one of the major problems that affect the reproducibility of results and its application on other tasks is the large number of parameters to be tuned in the T-IR models. For instance, computing the time series on a given temporal window may yield completely different results than another timespan. Parameter-free datasets and models may be of benefit in these cases. However, there are many challenges to be overcame before this can be applied on real settings, including diverse data sources and appropriate evaluation protocols, as previously mentioned.

7. Conclusion

Temporal IR is a new emerging research field that aims to enhance the retrieval effectiveness by incorporating temporal characteristics of queries and documents in the core of the IR ranking models. Research directions that have been recently addressed in this concern showed the potential of embedding temporal signals into the ranking process and its effects on the retrieval effectiveness. In this paper, after motivating and presenting a number of examples and scenarios where temporal information could be very useful, we proposed a general categorization of the state-of-the-art research in the community. We have put the focus on three aspects in which time meets directly the IR realm. More particularly, we paid a special attention to work in which time is encompassed on the query level, the document content level and the document ranking level. We have also highlighted the importance of result presentation, in particular timeline visualisation. Recent advances in the areas of social Web search, mobile IR and geo-temporal ranking have also been provided in order to show how temporal and spatial or other relevance factors are jointly considered to account for documents relevance. The choice of these IR fields is not intended to be exhaustive, it is rather an attempt to provide a view of various fields as they thoroughly connect with time. Then, we have reported the leading tracks for the evaluation of time-aware information access models.

The scrutiny of the experience arisen from the state-of-the-art approaches reveals that there is a compelling need to shift these studies towards more targeted generalized frameworks that involve all the IR components. Interesting future work to fill this gap may consist in providing real world documents collections such as query logs or real users' search behaviours to empirically evaluate T-IR models. Another compelling experimental issue is that concerning the evaluation metrics used to assess the quality of the extracted temporal expressions as well as the document results. Most of the common methods rely on classical (adaptation) evaluation measures such as precision, recall and F-measure and to some extent ROUGE scores, among others, depending on the domain of the application. For instance, the ROUGE measure which is used for the automatic evaluation of summaries is found to be unable to distinguish between automatic and human summaries. This metric may be inappropriate when a large number of test points is available [96]. Further investigations are needed to study more temporally dependent metrics that consider the time as an important evaluation measure through the whole evaluation process.

Acknowledgements

Funding

Notes

1. http://www.google.com/
2. http://en.wikipedia.org/wiki/Time/
3. http://www.google.com/trends/hottrends/ (accessed 6 March 2015)
4. http://heideltime.ifi.uni-heidelberg.de/heideltime/
5. http://www.timeml.org/tempeval2/
6. http://timeml.org/site/tarsqi/modules/gutime/
7. The dataset is publicly available at: http://www.db-net.aueb.gr/GoogleNewsDataset/
8. http://www.cbc.ca/news/world/charlie-hebdo-newspaper-shooting-timeline-1.2892399
9. http://alt.qcri.org/semeval2015/task4/
10. http://www.l3s.de/~gtran/timeline/
11. http://alt.qcri.org/semeval2015/
12. http://www.trec-ts.org/
13. https://sites.google.com/site/ntcirtemporalia/
14. https://news.google.com/
15. http://newsmap.jp/

References

[1] Borlund P. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology* 2003; 54(10): 913–925.

[2] Saracevic T. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science* 2007; 58(13): 2126–2144.

[3] Taylor AR. User relevance criteria choices and the information search process. *Information Processing and Management* 2012; 48(1): 136–153.

[4] Eickhoff C, de Vries AP and Collins-Thompson K. Copulas for information retrieval. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '13; New York, NY: ACM; 2013, pp. 663–672.

[5] Moulahi B, Tamine L and Ben Yahia S. iAggregator: Multidimensional relevance aggregation based on a fuzzy operator. *Journal of the Association for Information Science and Technology* 2014; 65(10): 2062–2083.

[6] Mathews LK and Kanmani SD. A survey on temporal information retrieval systems. *International Journal of Computer Applications* 2012; 58(4): 24–28.

[7] Alonso O, Gertz M and Baeza-Yates R. On the value of temporal information in information retrieval. *SIGIR Forum* 2007; 41(2): 35–41.

[8] Joho H, Jatowt A and Roi B. A survey of temporal Web search experience. In: *Proceedings of the 22nd International Conference on World Wide Web Companion*. WWW '13 Companion; Geneva: International World Wide Web Conferences Steering Committee; 2013, pp. 1101–1108.

[9] Lin S, Jin P, Zhao X and Yue L. Exploiting temporal information in Web search. *Expert Systems with Applications* 2014; 41(2): 331–341.

[10] Alonso O, Baeza-yates R, Strötgen J and Gertz M. Temporal information retrieval: Challenges and opportunities. In: *1st Temporal Web Analytics Workshop at WWW. CEUR Workshop Proceedings*, 2011, pp. 1–8.

[11] Campos R, Dias G, Jorge AM and Jatowt A. Survey of temporal information retrieval and related applications. *ACM Computing Surveys* 2014; 47(2): 15:1–15:41.

[12] Yu PS, Li X and Liu B. On the temporal dimension of search. In: *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*. WWW Alt. '04; New York, NY: ACM, 2004, pp.448–449.

[13] Sanderson M. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval* 2010; 4(4): 247–375.

[14] Baeza-Yates RA and Ribeiro-Neto B. *Modern Information Retrieval*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[15] Dong A, Zhang R, Kolari P et al. Time is of the essence: Improving recency ranking using twitter data. In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10; New York, NY: ACM, 2010, pp. 331–340.

[16]   Amati G, Amodeo G and Gaibisso C. Survival analysis for freshness in microblogging search. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12; New York, NY: ACM, 2012, pp. 2483–2486.

[17]   Metzler D, Jones R, Peng F and Zhang R. Improving search relevance for implicitly temporal queries. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '09; New York, NY: ACM, 2009, pp. 700–701.

[18]   Jones R and Diaz F. Temporal profiles of queries. *ACM Transactions on Information Systems* 2007; 25(3): 14.

[19]   Nunes S, Ribeiro C and David G. Use of temporal expressions in Web search. In: *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*. ECIR '08; Berlin, Heidelberg: Springer-Verlag, 2008, pp. 580–584.

[20]   Li X and Croft WB. Time-based language models. In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. CIKM '03; New York, NY: ACM, 2003, p. 469–475.

[21]   Berberich K, Bedathur S, Alonso O and Weikum G. A language modeling approach for temporal information needs. In: *Proceedings of the 32Nd European Conference on Advances in Information Retrieval*. ECIR '10; Berlin, Heidelberg: Springer-Verlag, 2010, pp. 13–25.

[22]   Rose DE and Levinson D. Understanding user goals in Web search. In: *Proceedings of the 13th International Conference on World Wide Web*. WWW '04; New York, NY: ACM, 2004, pp. 13–19.

[23]   Dakka W, Gravano L and Ipeirotis PG. Answering general time-sensitive queries. *IEEE Transactions on Knowledge and Data Engineering* 2012; 24(2): 220–235.

[24]   Vlachos M, Meek C, Vagena Z and Gunopulos D. Identifying similarities, periodicities and bursts for online search queries. In: *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*. SIGMOD '04; New York, NY: ACM, 2004, pp. 131–142.

[25]   Subasic I and Castillo C. The effects of query bursts on Web search. In: *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology; vol. 1*. WI-IAT '10. Washington, DC: IEEE Computer Society, 2010, pp. 374–381.

[26]   Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M and Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* 2009; 457: 1012–1014.

[27]   Diaz F. Integration of news content into Web results. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. WSDM '09; New York, NY: ACM, 2009, pp. 182–191.

[28]   Radinsky K, Svore K, Dumais S, Teevan J, Bocharov A and Horvitz E. Modeling and predicting behavioral dynamics on the Web. In: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12; New York, NY: ACM, 2012, pp. 599–608.

[29]   Chien S and Immorlica N. Semantic similarity between search engine queries using temporal correlation. In: *Proceedings of the 14th International Conference on World Wide Web*. WWW '05; New York, NY: ACM, 2005, pp. 2–11.

[30]   Radinsky K, Agichtein E, Gabrilovich E and Markovitch S. A word at a time: Computing word relatedness using temporal semantic analysis. In: *Proceedings of the 20th International Conference on World Wide Web*. WWW '11; New York, NY: ACM, 2011, pp. 337–346.

[31]   Asur S and Buehrer G. Temporal analysis of Web search query-click data. In: *SNA-KDD*. Paris: ACM Press, 2009, pp. 1–8.

[32]   Ren P, Chen Z, Song X, Li B, Yang H and Ma J. Understanding temporal intent of user query based on time-based query classi-fication. In: *Natural Language Processing and Chinese Computing; vol. 400*. Berlin: Springer, 2013, pp. 334–345.

[33]   Costa M, Couto F and Silva M. Learning temporal-dependent ranking models. In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '14; New York, NY: ACM, 2014, pp. 757–766.

[34]   Strötgen J and Gertz M. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation* 2013; 47(2): 269–298.

[35]   Verhagen M, Saurí R, Caselli T and Pustejovsky J. Semeval-2010 task 13: Tempeval-2. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval '10; Stroudsburg, PA: Association for Computational Linguistics, 2010, pp. 57–62.

[36]   Pustejovsky J and Verhagen M. Semeval-2010 task 13: Evaluating events, time expressions, and temporal relations (tempeval-2). In: *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. DEW '09; Stroudsburg, PA: Association for Computational Linguistics, 2009, pp. 112–116.

[37]   UzZaman N, Llorens H, Allen JF, Derczynski L, Verhagen M and Pustejovsky J. Tempeval-3: Evaluating events, time expres-sions, and temporal relations. *CoRR* 2012; abs/1206.5333.

[38]   Wang P, Berry MW and Yang Y. Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology* 2003; 54(8): 743–758.

[39]   Kulkarni A, Teevan J, Svore KM and Dumais ST. Understanding temporal query dynamics. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11; New York, NY: ACM, 2011, pp. 167–176.

[40]   Efron M. Linear time series models for term weighting in information retrieval. *Journal of the Association for Information Science and Technology* 2010; 61(7): 1299–1312.

[41] Radinsky K, Svore KM, Dumais ST et al. Behavioral dynamics on the Web: Learning, modeling, and prediction. *ACM Transactions on Information Systems* 2013; 31(3): 16:1–16:37.

[42] Campos R, Dias G, Jorge AM and Nunes C. GTE-Rank: Searching for implicit temporal query results. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. CIKM '14; New York, NY: ACM, 2014, pp. 2081–2083.

[43] Lin S, Jin P, Zhao X and Yue L. Tase: A time-aware search engine. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12; New York, NY: ACM, 2012, pp. 2713–2715.

[44] Harper S and Chen AQ. Web accessibility guidelines: A lesson from the evolving Web. *World Wide Web* 2012; 15: 1–28.

[45] Perkiö J, Buntine W and Tirri H. A temporally adaptive content-based relevance ranking algorithm. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '05; New York, NY: ACM, 2005, pp. 647–648.

[46] Karkali M, Rousseau F, Ntoulas A and Vazirgiannis M. Using temporal IDF for efficient novelty detection in text streams. *CoRR* 2014; abs/1401.1456.

[47] Aji A, Wang Y, Agichtein E and Gabrilovich E. Using the past to score the present: Extending term weighting models through revision history analysis. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM '10; New York, NY: ACM, 2010, pp. 629–638.

[48] Nunes S, Ribeiro C and David G. Term weighting based on document revision history. *Journal of the Association for Information Science and Technology* 2011; 62(12): 2471–2478.

[49] Kim HD, Nikitin D, Zhai C, Castellanos M and Hsu M. Information retrieval with time series query. In: *Proceedings of the 2013 Conference on the Theory of Information Retrieval*. ICTIR '13; New York, NY: ACM, 2013, pp. 14–56.

[50] Nunes S. Exploring temporal evidence in Web information fetrieval. In: MacFarlane A, Azzopardi L and Ounis I (eds) *BCS IRSG Symposium Future Directions in Information Access (FDIA 2007)*. London: BCS IRSG, 2007, pp. 44–50.

[51] Shokouhi M. Detecting seasonal queries by time-series analysis. In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '11; New York, NY: ACM, 2011, pp. 1171–1172.

[52] Jabeur LB, Tamine L and Boughanem M. Featured tweet search: Modeling time and social influence for microblog retrieval. In: *2012 IEEE/WIC/ACM International Conferences on Web Intelligence*. IEEE Computer Society, 2012, pp. 166–173.

[53] Diaz F, Dumais ST, Efron M, Radinsky K, de Rijke M and Shokouhi M. In: SIGIR 2013 workshop on time aware information access (#TAIA2013). In: SIGIR. 2013.

[54] Diaz F, Hauff C, Murdock V, de Rijke M and Shokouhi M. Sigir 2014 workshop on temporal, social and spatially-aware information access (#TAIA2014). In: SIGIR. New York, NY: ACM, 2014.

[55] Teevan J, Ramage D and Morris MR. #twittersearch: A comparison of microblog search and Web search. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11; New York, NY: ACM, 2011, pp. 35–44.

[56] Efron M, Lin J, He J and de Vries AP. Temporal feedback for tweet search with non-parametric density estimation. In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '14; New York, NY: ACM, 2014, pp. 33–42.

[57] Elsweiler D and Harvey M. Engaging and maintaining a sense of being informed: Understanding the tasks motivating twitter search. *Journal of the Association for Information Science and Technology* 2015; 66(2): 264–281.

[58] Wang S, Lu K, Lu X and Wang B. Query dependent time-sensitive ranking model for microblog search. In: *Web Technologies and Applications; vol. 8709 of Lecture Notes in Computer Science*. New York: Springer International Publishing, 2014, pp. 644–651.

[59] Cao Z, Qin T, Liu TY, Tsai MF and Li H. Learning to rank: From pairwise approach to listwise approach. In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07; New York, NY: ACM, 2007, pp. 129–136.

[60] Miyanishi T, Seki K and Uehara K. Time-aware latent concept expansion for microblog search. In: *Proceedings of the Eighth International Conference on Weblogs and Social Media*. ICWSM '14; Palo Alto, CA: AAAI Press, 2014.

[61] Wei CP, Lee YH, Chiang YS, Chen CT and Yang CC. Exploiting temporal characteristics of features for effectively discovering event episodes from news corpora. *Journal of the Association for Information Science and Technology* 2014; 65(3): 621–634.

[62] Osborne M, Petrovic S, McCreadie R, Macdonald C and Ounis I. Bieber no more: First story detection using Twitter and Wikipedia. In: *SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012)*. New York, NY: ACM, 2012.

[63] Whiting S, Jose J and Alonso O. Wikipedia as a time machine. In: *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. WWW Companion '14; Geneva: International World Wide Web Conferences Steering Committee, 2014, pp. 857–862.

[64] Göbölös-Szabó J and Benczur AA. Temporal wikipedia search by edits and linkage. In: *SIGIR 2013 workshop on temporal, social and spatially-aware information access (#TAIA2013)*. New York, NY: ACM, 2013.

[65] Zhang L, Tang J and Zhang M. Integrating temporal usage pattern into personalized tag prediction. In: *Web Technologies and Applications; vol. 7235 of Lecture Notes in Computer Science*. Berlin: Springer, 2012, pp. 354–365.

[66] Khodaei A and Alonso O. Temporally-aware signals for social search. In: *SIGIR 2012 workshop on temporal, social and spatially-aware information access (#TAIA2012)*. New York, NY: ACM, 2012.

[67] Inagaki Y, Sadagopan N, Dupret G et al. Session based click features for recency ranking. In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI 2010; Palo Alto, CA: AAAI Press, 2010.

[68] Badache I and Boughanem M. Harnessing social signals to enhance a search. In: *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Washington, DC, USA: IEEE; 2014, pp. 303–309.

[69] Yau SS, Liu H, Huang D and Yao Y. Situation-aware personalized information retrieval for mobile internet. In: *Proceedings of the 27th Annual International Conference on Computer Software and Applications*. Washington, DC: IEEE Computer Society, 2003.

[70] Göker A and Myrhaug H. Evaluation of a mobile information system in context. *Information Processing & Management* 2008; 44(1): 39–65.

[71] Church K and Smyth B. Understanding the intent behind mobile information needs. In: *Proceedings of the 14th International Conference on Intelligent User Interfaces*. IUI '09; New York, NY: ACM, 2009, pp. 247–256.

[72] Teevan J, Karlson A, Amini S, Brush AJB and Krumm J. Understanding the importance of location, time, and people in mobile local search behavior. In: *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. MobileHCI '11; New York, NY: ACM, 2011, pp. 77–80.

[73] Strötgen J and Gertz M. Proximity2-aware ranking for textual, temporal, and geographic queries. In: *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*. CIKM '13; New York, NY: ACM, 2013, pp. 739–744.

[74] Mishra A, Milchevski D and Berberich K. Vocabulary-based re-ranking for geographic and temporal searching at NTCIR geo-time task. In: *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-8)*. Tokyo: National Institute of Informatics, 2010, pp. 181–184.

[75] Daoud M and Huang JX. Modeling geographic, temporal, and proximity contexts for improving geotemporal search. *Journal of the American Society for Information Science and Technology* 2013; 64(1): 190–212.

[76] Strötgen J and Gertz M. Timetrails: A system for exploring spatio-temporal information in documents. *Proceedings of the VLDB Endowment* 2010; 3(1–2): 1569–1572.

[77] Li J and Cardie C. Timeline generation: Tracking individuals on twitter. In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14; New York, NY: ACM, 2014, pp. 643–652.

[78] Tran GB, Tran TA, Tran NK, Alrifai M and Kanhabua N. Leveraging learning to rank in an optimization framework for time-line summarization. In: *SIGIR 2013 workshop on temporal, social and spatially-aware information access (#TAIA2013)*. New York, NY: ACM, 2013.

[79] Zhang C, Xu W, Meng F, Li H, Wu T and Xu L. The information extraction systems of PRIS at temporal summarization track. In: *Text Retrieval Conference (TREC)*. Gaithersburg, MD: NIST, 2013.

[80] Frank JR, Kleiman-Weiner M, Roberts DA et al. Building an Entity-Centric Stream Filtering Test Collection for TREC 2012. In: *Proceedings of the Text REtrieval Conference (TREC)*. Gaithersburg, MD: NIST, 2012.

[81] Abbes R, Pinel-Sauvagnat K, Hernandez N and Boughanem M. IRIT at TREC KBA 2014. In: *Text REtrieval Conference (TREC)*. Gaithersburg, MD: NIST, 2014.

[82] Joho H, Jatowt A and Blanco R. NTCIR Temporalia: A test collection for temporal information access research. In: *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. WWW Companion '14; Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee; 2014, pp. 845–850.

[83] Yu H, Kang X and Ren F. TUTA1 at the NTCIR-11 Temporalia task. In: *Proceedings of the NTCIR-11 Conference*. Tokyo, Japan: National Institute of Informatics (NII); 2014.

[84] Joho Y, Jatowt A, Blanco HN and Yamamoto S. Overview of NTCIR-11 temporal information access (Temporalia) task. In: *Proceedings of the NTCIR-11 Conference*. Tokyo, Japan: National Institute of Informatics (NII); 2014.

[85] Lin J, Efron M, Wang Y and Sherman G. Overview of the TREC-2014 Microblog track. In: *Proceedings of the Text REtrieval Conference (TREC)*. Gaithersburg, MD: NIST, 2014.

[86] Zhang G, Yang Z and Si S. EBJUT at TREC 2014 microblog track. In: *Proceedings of the Text REtrieval Conference (TREC)*. Gaithersburg, MD: NIST, 2014.

[87] Nguyen TN and Kanhabua N. Leveraging dynamic query subtopics for time-aware search result diversification. In: *Advances in Information Retrieval - 36th European Conference on IR research, ECIR 2014; vol. 8416*. Berlin: Springer, 2014, pp. 222–234.

[88] Berberich K and Bedathur S. Temporal diversification of search results. In: *SIGIR 2013 workshop on temporal, social and spatially-aware information access (#TAIA2013)*. New York, NY: ACM, 2013.

[89] Harman D. *Information Retrieval Evaluation*. 1st ed. San Rafael, CA: Morgan & Claypool Publishers, 2011.

[90] Lin J and Efron M. Evaluation as a service for information retrieval. *SIGIR Forum* 2013; 47(2): 8–14.

[91] Kelly D, Dumais S and Pedersen JO. Evaluation challenges and directions for information-seeking support systems. *Computer* 2009; 42(3): 60–66.

[92] Azzopardi L and Balog K. Towards a living lab for information retrieval research and development: A proposal for a living lab for product search tasks. In: *Proceedings of the Second International Conference on Multilingual and Multimodal Information Access Evaluation*. CLEF '11; Berlin, Heidelberg: Springer-Verlag, 2011, pp. 26–37.

[93] Hong L, Ahmed A, Gurumurthy S, Smola AJ and Tsioutsiouliklis K. Discovering geographical topics in the twitter stream. In: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12; New York, NY: ACM, 2012, pp. 769–778.

[94] Mishra A, Milchevski D and Berberich K. Linking wikipedia events to past news. In: *SIGIR 2014 workshop on temporal, social and spatially-aware information access (#TAIA2014)*. New York, NY: ACM, 2014.

[95] Metzler D and Croft WB. A markov random field model for term dependencies. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '05; New York, NY: ACM, 2005, pp. 472–479.

[96] Nenkova A and McKeown K. Automatic summarization. *Foundations and Trends in Information Retrieval* 2011; 5(2–3): 103–233.