

Survey of Temporal Information Retrieval and Related Applications

RICARDO CAMPOS, Polytechnic Institute of Tomar, Tomar, Portugal, LIAAD – INESC TEC, Center of Mathematics, University of Beira Interior, Covilhã, Portugal

GAËL DIAS, HULTECH/GREYC, University of Caen Basse-Normandie, France

ALÍPIO M. JORGE, LIAAD – INESC TEC, DCC-FCUP, University of Porto, Porto, Portugal

ADAM JATOWT, Graduate School of Informatics, Kyoto University, Japan, Japan Science and Technology Agency, Japan

Temporal information retrieval has been a topic of great interest in recent years. Its purpose is to improve the effectiveness of information retrieval methods by exploiting temporal information in documents and queries. In this article, we present a survey of the existing literature on temporal information retrieval. In addition to giving an overview of the field, we categorize the relevant research, describe the main contributions, and compare different approaches. We organize existing research to provide a coherent view, discuss several open issues, and point out some possible future research directions in this area. Despite significant advances, the area lacks a systematic arrangement of prior efforts and an overview of state-of-the-art approaches. Moreover, an effective end-to-end temporal retrieval system that exploits temporal information to improve the quality of the presented results remains undeveloped.

Categories and Subject Descriptors: H 3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Clustering, query formulation, retrieval models, search process*

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Temporal information retrieval, temporal queries, query classification, time-based clustering, temporal search engines

ACM Reference Format:

Ricardo Campos, Gaël Dias, Alípio M. Jorge, and Adam Jatowt. 2014. Survey of temporal information retrieval and related applications. *ACM Comput. Surv.* 47, 2, Article 15 (July 2014), 41 pages.

DOI: <http://dx.doi.org/10.1145/2619088>

This research was funded by Project NORTE-07-0124-FEDER-000059 which is financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency (FCT). It was also financed by the ERDF through the COMPETE Programme (operational programme for competitiveness), by National Funds through the FCT within project «FCOMP-01-0124-FEDER-037281» and by the Center of Mathematics, University of Beira Interior, within project «PEst-OE/MAT/UI0122/2014». It was also supported by MEXT Grant-in-Aid for Young Scientists B (#22700096) and by JST research promotion program Presto.

Authors' addresses: R. Campos, Polytechnic Institute of Tomar - ICT Department, Av. Dr. Aurélio Ribeiro, 2300 - 313 Tomar, Portugal; email: ricardo.campos@ipt.pt; LIAAD – INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 4200 - 465 Porto, Portugal; email: ricardo.campos@inesctec.pt; Center of Mathematics, Rua Marquês d'Ávila e Bolama, 6200-001 Covilhã, Portugal; G. Dias, University of Caen Basse-Normandie - HULTECH/GREYC, Campus Côte de Nacre, F-14032 Caen Cedex France; gaël.dias@unicaen.fr; A. M. Jorge, LIAAD – INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 4200 - 465 Porto, Portugal; email: alipio.jorge@inescporto.pt; DCC – Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169 - 007 Porto, Portugal; email: amjorge@fc.up.pt; A. Jatowt, Kyoto University - Graduate School of Informatics, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan; email: adam@dl.kuis.kyoto-u.ac.jp.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2014 ACM 0360-0300/2014/07-ART15 \$15.00

DOI: <http://dx.doi.org/10.1145/2619088>

1. INTRODUCTION

Information retrieval is the process that provides users with the most relevant documents from an existing collection. The user information needs are expressed by a query, typically in a short textual form. In recent years, time has been gaining increasing importance within search contexts, leading to a new research area known as temporal information retrieval (T-IR) that contains a number of different challenges. In general, T-IR aims to satisfy search needs by combining the traditional notion of *document relevance* with *temporal relevance*. For example, users may require documents that describe the past (e.g., queries about historical figures); documents containing the most recent, up-to-date information (e.g., queries about weather forecasts or currency rates); or even future-related information (e.g., queries about planned events to be held in a certain area). The value of information and its quality are intrinsically time dependent. Information science researchers [Metzger 2007] tend to consider timeliness or currency as one of the five key aspects that determine a document's quality; the others are relevance, accuracy, objectivity, and coverage.

The huge volume of the web, however, makes T-IR a difficult task. First, since the web is constantly changing, maintaining up-to-date indexes is becoming more and more difficult. Second, a clear understanding of the temporal nature of queries is difficult due to query ambiguity, different possible temporal characteristics of queries and unknown users' expectations toward the required temporality of search results. Third, it is not easy to retrieve web documents so that their temporal dimension will meet the user temporal intent underlying the query. Nevertheless, the community started to address the problem of retrieving web pages that are not only topically relevant but also created during (or that refer to) the most relevant time periods. They also approached the problem of determining various temporal dimensions of documents and queries. These contributions can greatly benefit the process of indexing documents, as well as the ranking and clustering of web search results.

The importance of considering temporal aspects in IR and the need for a continuous search for effective T-IR solutions become clear in light of the recent emergence of numerous *time-related initiatives* and *applications*. One of the first is the Internet Archive project [Kahle 1997], which is compiling a digital library of websites. Its objective is to store past versions of websites based on their periodical crawls. The archive has been used by computer scientists, information scientists and historians as a way to preserve, provide access, search, extract and visualize the different past versions of a web page. The information collected reportedly grows at a rate of 100 terabytes each month reaching an impressive number of over 350 billion archived web pages. Further evidence is the recent development of research projects that address the archiving, the analysis, and the access to the temporal web. Examples include ARCOMEM,¹ LAWA,² LiWA,³ and LivingKnowledge.⁴ There is also much research on using temporal information for exploration and search purposes. For instance MIT has developed SIMILE Timeline Visualization,⁵ a web widget prototype for visualizing temporal data.

Within the context of knowledge bases, YAGO2 [Hoffart et al. 2011] provides a search interface to seek temporal and spatial knowledge facts. Figure 1 shows an example of the SVG-based browser interface for the query “*David Beckham*”. We can observe that this famous athlete was born on May 2, 1975, by looking at the relation `<wasBornOnDate>`.

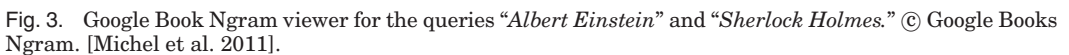
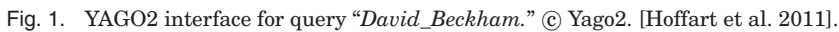
¹<http://www.arcomem.eu/> [May 26, 2014].

²<http://www.lawa-project.eu/> [May 26, 2014].

³<http://www.liwa-project.eu/> [May 26, 2014].

⁴<http://livingknowledge.europarchive.org/> [May 26, 2014].

⁵<http://www.simile-widgets.org/timeline/> [May 26, 2014].



More recently, Google NGram Viewer [Michel et al. 2011] (Figure 3) was released as a visualization tool that shows the rise and fall of particular keywords across a temporally arranged collection constructed from over 5 million books. All of these large-scale projects suggest high interest of the public in temporal aspects of information and support the claim that T-IR should be regarded as a promising research area.

⁶<https://www.recordedfuture.com/> [May 26, 2014].

In addition, the creation of annotation standard corpora such as the TimeBank [Pustejovsky et al. 2006] and annotation schemas such as TimeML⁷ [Pustejovsky et al. 2003], and the development of temporal taggers (see Section 2.5) and specialized test collections show the interest of using time as a key feature in document analysis and in IR processes. Another indication of T-IR's importance is the realization of an increasing number of contests and workshops that focus on the temporal aspects of information. For the former, different competitions have been proposed, such as the Message Understanding Conference (MUC) with specific tracks on the identification of temporal expressions (MUC6 and MUC7); the Automated Content Extraction (ACE) evaluation program, organized by the National Institute of Standards and Technology (NIST), the Time Expression Recognition and Normalization (TERN), which has been recently associated with the Text Analysis Conferences (TACs); TempEval within the SemEval competition and the NTCIR Temporal Information Access (Temporalia). WWW Temporal Web Analytics workshop (TAWW 2011; TempWeb 2012, 2013, and 2014) and the SIGIR Time-Aware Information Access workshops (TAIA 2012, 2013, and 2014) are examples of seminars dedicated to temporal information search and processing. A detailed description of existing evaluation challenges, annotation schemes, and datasets can be found in Mazur [2012] and Costa [2013].

Despite the recent improvement of search and retrieval applications in the temporal context, one can still find examples when the returned search results do not satisfy the user information needs due to problems of a temporal nature. For example, search results might contain obsolete data, even though a user was actually searching for fresh information (e.g., queries about weather forecasts, stock prices, traffic conditions), or past-related information despite the user's search intent being focused on the future (e.g., queries about a company's future plans).

Based on all these factors, an upsurge of applications is expected in the near future, mostly concerning temporal information exploration, new forms of presenting the search results, and applications concerning temporally focused retrieval in microblogs (e.g., Blog, Twitter, and Facebook posts).

In particular, various research studies have already been proposed in different sub-areas of T-IR, such as query understanding [Jones and Diaz 2007; Metzler et al. 2009], temporal snippet generation [Alonso et al. 2009b; Alonso et al. 2011], temporal ranking of documents [Li and Croft 2003; Berberich et al. 2005; Dong et al. 2010a, 2010b; Elsas and Dumais 2010; Dai et al. 2011; Kanhabua and Nørvåg 2012], temporal clustering [Alonso and Gertz 2006; Campos et al. 2012b], future-related information retrieval [Baeza-Yates 2005; Jatowt et al. 2009; Radinski and Horvitz 2013], and temporal image retrieval [Dias et al. 2012].

To the best of our knowledge, this is the first comprehensive review of the state of the art in T-IR. Our main objective is to provide an overview of the research carried out in T-IR so far and to present a number of open issues as well as to propose promising directions that in our opinion offer much research potential yet remain generally unexplored by the scientific community. Although some prior work already described the field and indicated future promising directions [Alonso et al. 2011], we provide a more extensive overview of related works and comparatively present them in various contexts. A recent tutorial⁸ [Radinski et al. 2013] at the WSDM 2013 conference as well as the Wikipedia article entitled "Temporal Information Retrieval"⁹ can be considered other attempts to overview and systematize the field.

⁷<http://www.timeml.org/site/index.html> [May 26, 2014].

⁸http://research.microsoft.com/en-us/um/people/sdumais/WSDM2013-Tutorial_Final.pdf [May 26, 2014].

⁹http://en.wikipedia.org/wiki/Temporal_information_retrieval [May 26, 2014].

Like most recent efforts, we focus on T-IR within the context of the web. Since this is a relatively new area, no comprehensive overview exists that positions the existing research in the field. Given this, we introduce a set of models to serve as a framework to enable comparisons between different approaches in the web context.

This article is intended for academics and practitioners interested in information retrieval (IR) who do not have a detailed knowledge of T-IR and lack an organized overview of this research area. It also may provide value to scientific professionals interested in a broad snapshot of this field. Since we also indicate several directions and open challenges for future research, we hope it could become a source of new ideas for researchers who are already working on related topics or for those who are considering their own research studies.

The remainder of this survey is organized as follows. Section 2 gives a high-level overview of T-IR and formalizes the definitions of time, events, and timelines. In particular, we introduce the notion of temporal expressions and discuss the extraction of temporal information from texts, further listing available temporal taggers. Section 3 presents different approaches that can be used to extract time-related features within web collections. More specifically, we distinguish among metadata, content, and usage approaches and emphasize their main characteristics, challenges, and available data sources. Section 4 explores the key advances in temporal web information retrieval. We extensively detail in a classical IR fashion the different approaches used in the execution of any process of a web T-IR (i.e., crawling, indexing, query processing, and ranking). In addition, we present works in the field of temporal clustering, temporal text classification, temporal search engines, and recent research conducted in the scope of future-related information retrieval. Section 5 highlights possible future trends and unexplored areas of temporal IR, which although already proposed, still lack further developments. Finally, Section 6 concludes this survey with some final remarks.

2. MODELS OF TEMPORAL ANNOTATIONS OF DOCUMENTS

In the following sections, we introduce different temporal dimensions. Section 2.1 provides definitions of the concepts related to the notion of time. Section 2.2 describes the underlying relation between time and events. Next, Section 2.3 introduces timelines as a means of graphically representing the effects of time's passage. Section 2.4 describes different types of temporal expressions occurring in texts. Finally, Section 2.5 outlines the methodologies behind the extraction of temporal information that are usually used as preprocessing stages in T-IR systems.

2.1. Notion of Time

One of the first works to present a formal model for temporal references was presented by Bruce [1972]. He defined time as an ordered pair, (time, \leq) , where time is a set whose elements are called time points and \leq is a relation that partially orders time. Another formal approach is the work of Allen [1983], which describes a set of 13 possible temporal relationships between any two time intervals.

In a less formal way, time can be seen as an inherent construct in human life since our thinking is often in the form of chronologically arranged events stretching from past, to the present, and to the future. Each instance of time is a point-in-time value. Point-in-time values can be represented by units ranging from the finest granularity to the coarsest significant granularity: day (D), week (w), month (M), semester (s), quarter (q), year (Y), decade (de), and century (c). Note that a day can also include finer time points, such as hours, minutes, seconds, fractions of a second, and so forth.

Time values can be physically represented in a calendar, which is a timekeeping system that organizes time into several different granularities. The most widely used calendar in the world today is the Gregorian calendar (also called the Christian

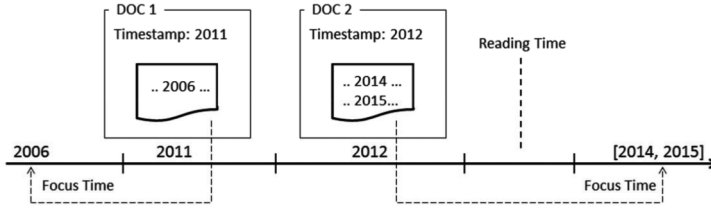


Fig. 4. Different types of time in documents.

calendar). In some countries, this calendar is substituted by or complemented with local ones (e.g., Jewish, Hindu, Chinese, and Islamic calendars, to name a few). Following the ISO-8601:2004¹⁰ standard, a date in the Gregorian calendar is usually represented in the form of *YYYY-MM-DD*, where *[YYYY]* indicates a year, *[MM]* indicates a month, and *[DD]* indicates a day of that month. Although less common, the date representation can also include the number of the week. In this case, the month is replaced by the corresponding week, which results in the format *YYYY-Www-DD*, where *ww* represents the week's number from W_{01} to W_{52} . Note that specialized calendars also exist, such as fiscal, sports, business, or academic ones.

When addressing the time issue within the realm of database research areas, two types of time are usually distinguished: *valid* and *transaction* [Snodgrass and Ahn 1985]. Valid time is related to the period of time during which events occur in real life (i.e., the time of the fact itself), while the transaction time refers to the specific time when the fact was stored in a database. In the web context, *focus time* is the time mentioned or implicitly referred to in the content of web pages. We regard it as a counterpart of valid time. Naturally, since a web page can refer to different points in time, its focus time should be represented by a set of time intervals rather than as a single point in time. Transaction time, on the other hand, is treated as a parallel of a *document timestamp*, that is, the point in time when the web page was either created (creation time—*ct*), modified (last-modified date—*lmd*), or published (publication time—*pt*). For instance, we may have an interval bounded by the *[start focus time point, end focus time point]* of the document, but also by the *[ct, lmd]* and *[pt, lmd]* time references. However, in cases where neither the focus time nor the timestamp can be determined, we can consider the first crawling date as the birth time and the most recent crawling time as the end time.

We also consider *reading time* and *document age* as additional types of time. In web search scenarios, a document's reading time is assumed to be the same as the time a search query was issued since users often access search results immediately after performing a search. On the other hand, a document's age is the difference between the reading time and the timestamp.

Figure 4 shows a visual example of different types of time. We start by analyzing document *Doc1*, which was created in 2011. Supposing that the current reading time is 2013, this document is considered 3 years old. Temporal references in its content define its focus time as equal to 2006. Naturally, since the document content can be re-edited, the document's focus time can change with time. In this sense, a document that previously included references to past time can later refer to the present, to future time, or to other past dates.

In the same figure, document *Doc2* represents a document whose focus time is defined by a time interval. Created in 2012, it currently contains future-related information, since its focus time is defined by a time interval [2014, 2015].

In the next section, we describe the underlying association between time and events.

¹⁰http://www.iso.org/iso/date_and_time_format [May 26, 2014].

2.2. Time and Events

Time is inherently associated with events [Setzer and Gaizauskas 2000]. A simple example is the phrase “**An airplane** coming from Brazil **fell** into the Atlantic Ocean **on Monday**,” which uses the temporal expression *on Monday* as a point in time that defines the fall of an airplane. We define an event simply as a change (“occurrence”) that happens at a given time in a given place and that can be thus mapped into a bidimensional spatiotemporal view. Events are usually considered to be a change or a disruption of a normal course that is important to society or to a large group of people; thus, they are worth reporting and publicizing.

One initial effort to automatically determine and track events was introduced by Allan et al. [1998] through the Topic Detection and Tracking (*TDT*) initiative. The goal of the TDT project, which was one of the first significant research initiatives related to automatic news management, is the exploration of techniques that identify the occurrence of new events in order to follow their track over time. Swan and Allan [1999] and Swan and Jensen [2000] proposed using a classical hypothesis test to discover time-dependent features that identify important topics in text documents. Makkonen and Ahonen-Myka [2003] suggested an alternative solution by comparing one document with another through a temporal similarity measure. Kumaran and Allan [2004], on the other hand, detected new stories by measuring the degree overlap of one story with those that occurred in the past. Shaparenko et al. [2005] correlated topic events with texts used in a document collection to provide an overview of how topics evolve over time. The underlying assumption is that as events change, the text used in documents will change as well. Changes in a text are detected by a K-Means clustering algorithm, where each cluster represents an important topic. The popularity of the topic over time is given by the number of documents that fall into each cluster. More recently, Vandembussche and Teissèdre [2011] introduced an experimental end-user prototype as a first step for query-event retrieval by offering users the possibility of querying a specific music dataset (enriched with web semantic data) for events occurring in a given time period at a specific location.

2.3. Timelines

A sequence of events is usually represented in a timeline. A timeline, also known as a chronology, is a graphic representation listing important events within a particular time span. Timelines are particularly useful to give a topic a historical context and to provide a comprehensive temporal understanding of it. An example of a timeline is what one could construct to represent the history of Haitian earthquakes (Figure 5).

Depending on their purpose, timelines of different granularities can be constructed, either more finely grained (e.g., quarters, semesters, months, weeks, and days) or more coarsely grained (decades and centuries). In the example in Figure 5, we use year, month, and day granularities to represent *Haiti earthquakes*. In what follows we describe different types of temporal expressions.

2.4. Temporal Expressions

Temporal expressions are a rich form of natural language that can be defined as a sequence of tokens with temporal meaning. The greatest difficulty in developing an automatic system for detecting temporal expressions is the large diversity of ways in which time can be expressed [Mazur 2012]. Following the work of Schilder and Habel [2005], we distinguish the following:

- Explicit temporal expressions
- Implicit temporal expressions
- Relative temporal expressions

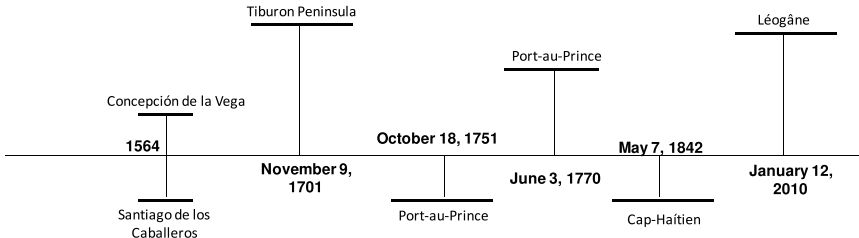


Fig. 5. Timeline for *Haiti earthquakes*.

Explicit temporal expressions, which were first referenced in 1993 [Setzer and Gaizauskas 2000] during MUC-5 [Advanced Research Projects Agency 1993], denote a precise moment in time and can be anchored on timelines without further knowledge. Depending on the granularity level, we may for example have “2009” for the year’s granularity, “December 2009” for the month’s granularity, and “25.12.2009” for the day’s granularity.

Implicit expressions are often associated with events carrying an implicit temporal nature. They are often difficult to be positioned in time due to the lack of a clear temporal target or an unambiguously associated time point. For example, expressions such as “Christmas Day” embody a temporal nature that is not explicitly specified. Therefore, as observed by Alonso et al. [2009b], these expressions require that at least an absolute time expression appears somewhere close in the text to establish accurate temporal values.

Relative temporal expressions, which were referenced for the first time in 1998 during MUC-7 [Setzer and Gaizauskas 2000], depend on the document publication date or another date nearby in the text. For instance, the expressions *today*, *last Thursday*, or *45 minutes after* are all relative to the document timestamp or to the absolute dates occurring nearby in the text. As such, collecting the document timestamp or related explicit temporal expressions is important, so that the expression can be mapped directly on the timeline as an explicit expression. Even though such information is usually available in news documents, it is difficult to be detected within web documents, as we discuss in Section 3.1. Besides, the document timestamp or any contextual clues, although important, might be insufficient in the case of more ambiguous phrases. An example is the expression “on Thursday,” which, as observed by Alonso et al. [2011], can either refer to the previous or to the next Thursday.

2.5. Temporal Information Extraction

The identification of temporal information is a nontrivial task that requires the pre-processing stage of a document and usually involves four steps. The first one is *Tokenization*, which divides the text into words or phrases. The second is the *Sentence Extraction* process, which identifies the set of all sentences in texts. The third is the part-of-speech tagging (POS) process where tokens are assigned a part of speech. Finally, the fourth step, Named-Entity Recognition (NER), identifies the proper nouns in documents. Interestingly, detecting temporal expressions was a part of the NER process before. However, since 2004, after the introduction of the TERN task as part of the ACE program, Temporal Information Extraction (T-IE) has become an independent task. As such, once text processing is underway, the T-IE process can start. It consists of three main tasks. The first is the *Extraction* or *Recognition* of temporal expressions. The second is *Normalization* to unify the different ways in which temporal expressions can be expressed. Finally, the last task, called *Temporal Annotation*, expresses temporal expressions in a standard format. The result is a set of texts where temporal

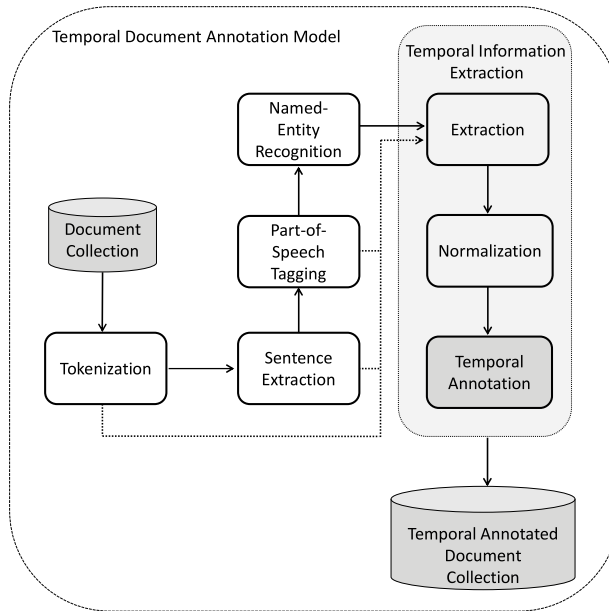


Fig. 6. Temporal document annotation model.

expressions are usually annotated with TimeML [Pustejovsky et al. 2003], which is a temporal formal specification XML language. Figure 6 shows the entire process. Note that not all the preprocessing steps are always necessary to perform temporal information extraction.

The overall T-IE process is usually conducted by temporal taggers, which follow rule-based approaches that are based on regular expressions or local grammar-based techniques and usually involve hard work by experts. In the last few years, temporal taggers have become an important research area. However, the fact that they rely on language-specific solutions makes them difficult to build. Hence, most available temporal taggers are useful for only one language (typically English) and for one domain (usually, the news domain). Other challenges involve determining the document creation time; delimiting, classifying, and normalizing temporal expressions; recognizing events; and determining their temporal order [Costa 2013]. Moreover, the lack of an extensive collection of texts annotated with temporal information covering different languages forms an additional problem. Additionally, the multitude of different forms in which human language allows temporal information to be conveyed [Mazur 2012] and the language intricacy and ambiguity complicate the task of tagging temporal information in texts more than simply finding the part-of-speech functions of words.

The following temporal taggers have been proposed: *TempEx* [Mani and Wilson 2000], *GUTime*,¹¹ *Annie*,¹² *HeidelTime*¹³ [Strötgen and Gertz 2010a], and *SuTime*¹⁴ [Chang and Manning 2012]. When evaluating temporal taggers, the task of extracting and normalizing temporal expressions can be measured individually. The former aims to correctly identify temporal expressions. The latter aims to normalize the temporal

¹¹<http://www.timeml.org/site/tarsqi/modules/gutime/download.html> [May 26, 2014].

¹²<http://www.actors.org/technologies/annie/> [May 26, 2014].

¹³<http://dbs.ifi.uni-heidelberg.de/index.php?id=form-downloads> [May 26, 2014].

¹⁴<http://nlp.stanford.edu:8080/sutime/process> [May 26, 2014].

expression to a standard format, and its complexity depends on the type of temporal expression extracted (see Section 2.4 for the different types of temporal expressions). For both tasks, precision (P), recall (R) and F_β -measure (F_β) are computed according to the formulas that follow. TP (true positives) is the number of expressions correctly identified by the system as temporal expressions. FP (false positives) is the number of expressions incorrectly identified by the system as temporal expressions, and FN (false negatives) is the number of expressions incorrectly identified as non-temporal expressions.

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F_\beta = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 P + R}$$

TempEx was the first temporal expression tagger to be developed. It is a rule-based model that extracts temporal information, particularly explicit (e.g., *December 24, 2009*) and relative temporal expressions (e.g., *Monday*), and labels them with TIMEX2 tags. First, the document is tokenized into words and sentences and the part of speech tagged. Each sentence is then passed on to a module that identifies time expressions and thereafter to a discourse processing module which solves context-dependent time expressions such as indexicals. Another temporal tagger is GUTime, which extends the capabilities of TempEx [Mani and Wilson 2000] by adding TIMEX3 tags. GUTime has been evaluated on the TERN 2004 training corpus and achieved F_1 -measure scores of 0.85 and 0.82 for temporal expression recognition and normalization, respectively. *Annie* was also developed in 2002 as part of the GATE¹⁵ distribution [Cunningham et al. 2002]. More recently, SuTime and HeidelTime have been developed based on a rule-based system to extract and normalize temporal expressions. SuTime is optimized for English texts and HeidelTime is a multilingual temporal tagger (English, German, Dutch, Vietnamese, Arabic, Spanish, Italian, French, and Chinese) that is adapted not only to the news domain but also to narrative documents. Both have been evaluated in the TempEval-3 challenge [UzZaman et al. 2013] and achieved competitive results. Results were very similar for the extraction process, with SuTime achieving a score of 0.932 in terms of F_1 -measure and HeidelTime obtaining a score of 0.930. In contrast, HeidelTime achieved the best performance for the normalization process with an F_1 -measure of 0.776 as opposed to SuTime, which achieved a score of 0.674. A detailed description of the existing approaches can be found in Strötgen and Gertz [2012] and Mazur [2012].

While temporal taggers are quite effective for temporal information processing, some works simply use straightforward regular expressions to look for temporal expressions. Indeed, for certain applications there may be no need to use temporal taggers since they may require very specific information, such as year mentions in texts.

There has also been research on extracting relational facts (e.g., *Barack Obama is the president of the USA*) from text corpora, which has led to the emergence of large knowledge bases, such as DBpedia¹⁶ [Auer et al. 2007] and YAGO¹⁷ [Suchanek et al. 2007] as well as commercial ones, such as Freebase¹⁸ [Bollacker et al. 2007] and Wolfram Alpha.¹⁹ These knowledge bases provide countless factual relations among entities, such as people or locations. However, they often ignore the temporal dimension (e.g., *Barack Obama has been the president of the USA since 2008*) and mostly focus on identifying the most salient facts. One consequence is the inconsistency of information,

¹⁵<http://gate.ac.uk/download/index.html> [May 26, 2014].

¹⁶<http://dbpedia.org> [May 26, 2014].

¹⁷<http://www.mpi-inf.mpg.de/yago-naga/yago/> [May 26, 2014].

¹⁸<http://www.freebase.com> [May 26, 2014].

¹⁹<http://www.wolframalpha.com> [May 26, 2014].

since some facts might only be true for a particular time (e.g., the relation [Bill Clinton, president of the USA] is only valid from 1993 to 2001).

Leveraging such temporal information enables researchers to create knowledgeable retrieval mechanisms that support entity-related temporal queries instead of keyword-based ones. Such a new paradigm will likely improve the effectiveness of the results and the user experience by answering such queries as “*Who got 2nd place at Ballon d’Or in 2010?*” or “*Which player made the most assists to Cristiano Ronaldo at the Real Madrid FC during the 2012/2013 season?*”

Despite the importance of time, research on time-sensitive fact extraction has only recently been addressed. A few recent works [Wang et al. 2010, 2011, 2012; Hoffart et al. 2011; Talukdar et al. 2012; Kuzey and Weikum 2012] have explored temporal information for the automatic development of temporal knowledge bases. One of the first works addressing this problem was developed by Wang et al. [2010] with the T-YAGO knowledge base, which is an extension of YAGO [Suchanek et al. 2007]. T-YAGO uses regular expressions to extract temporal facts from semistructured data attached to Wikipedia articles, such as infoboxes. Although T-YAGO was an interesting first approach, its restriction to the football domain and the fact that it does not support the extraction of information from free text corpora limit its scope. Another extension of the YAGO knowledge base is the YAGO2 system [Hoffart et al. 2011], which focuses on temporal and spatial knowledge by gathering information from Wikipedia²⁰ infobox attributes, WordNet,²¹ and GeoNames.²² However, like its predecessor it fails to extract information from free text corpora, which is a limitation.

Wang et al. [2011] proposed PRAVDA to automatically harvest temporal facts from textual web sources, especially from news articles and biography texts. It uses a pattern-based approach to extract the temporal candidates of facts. Then label propagation, a semi-supervised learning algorithm, computes the confidence scores of the candidate facts. As noted earlier, Wang et al. [2012] employ a methodology that combines label propagation and an integer linear program that incorporates temporal constraints among correlated events to determine noisy facts (e.g., Cristiano Ronaldo cannot play football for Manchester United and Real Madrid at the same time). In contrast, CoTS [Talukdar et al. 2012] temporally scopes relational facts based on change detection in a time series of the number of facts from the Google Books Ngram [Michel et al. 2011] and Gigaword [Graff et al. 2003] datasets. As done previously, an integer linear program incorporating temporal constraints temporally scopes the correlated facts and guarantees their temporal consistency.

Another recent approach is the work of Kuzey and Weikum [2012], which is an extension of T-YAGO and YAGO2. As well as harvesting temporal facts and events from Wikipedia, especially from the infoboxes of articles devoted to named events (e.g., historical events, conferences, etc), it also extracts temporal facts from free text, namely, from the full content of Wikipedia articles. This allows the construction of an enhanced knowledge base when compared to T-YAGO and YAGO2. In contrast to the development of knowledge bases, Döhling and Leser [2014] propose finding, extracting, and aggregating numerical facts from the web, paying particular attention to reliable and timely data related to natural disasters.

3. EXTRACTING TEMPORAL INFORMATION FROM WEB RESOURCES

With the advent of the web, the world’s largest collection of data, a huge amount of temporal information has become available. This information can be found within a

²⁰<http://www.wikipedia.org/> [May 26, 2014].

²¹<http://wordnet.princeton.edu/> [May 26, 2014].

²²<http://www.geonames.org> [May 26, 2014].

number of different web sources, from web query logs to collections of web pages to social networks such as Twitter. In this section, we describe approaches for extracting temporal features from web resources and we consider three different approaches related to the type of underlying data collection: web documents for *metadata* and *content* techniques, and web query logs for *usage* methodologies.

3.1. Metadata

The metadata approach extracts time information from the document's metadata. This includes the document's creation time, its publication time, and the last-modified date. But it may also embody the extraction of additional temporal information from the document structure, especially information extracted from the URL of the document or from the anchor texts associated to the links pointing to the document.²³ This information is usually available from news collections. One of the best-known longitudinal news article sources is the *New York Times Annotated Corpus* [Sandhaus 2008], which spans 20 years of newspapers between 1987 and 2007, including 1.8 million articles (more than 1.5 million manually annotated) and 650,000 article summaries.

While metadata information may be quite useful to solve relative temporal expressions found in the document's content (e.g., “today”) and to normalize them to some concrete date (e.g., “2012/12/31”), it may sometimes be inadequate since the timestamp of a document (creation, publication, or modification time) may differ significantly from its focus time (i.e., its content). A simple example is a document published in “2009” whose content concerns the year “2011.”

In addition, metadata information is particularly difficult to obtain from less structured collections, such as web pages, as opposed to news articles. One reason for this, as observed by Nunes et al. [2007], is due to the fact that web servers typically do not provide other temporal information than the crawling date. An alternative solution is to extract metadata information from the document content, for instance, searching for temporal expressions preceded by the phrase “last-modified.” This procedure demands a rule definition for each different case or language, which may be quite unfeasible for real-world applications.

3.2. Content

The content approach focuses on the analysis and extraction of temporal features within web contents, usually to determine a document's focus time. This includes looking for information within web pages, within posts in web micro-blog collections or different past page versions stored in web archives. When seeking a content-based collection, a good starting point is to consider the Clueweb09²⁴ dataset, which consists of 1 billion web pages in 10 different languages or its newer version Clueweb12.²⁵ Other available data sources can be found on Datamob²⁶ and Kevin Chai's²⁷ websites. Many collections are listed as open room for research on a number of different dimensions, including the temporal one.

Unlike metadata approaches, the content approach implies an increased level of difficulty since it usually involves linguistic analysis of texts, as discussed in Section 2.5. Since the web is heterogeneous, multilingual, multicultural, and highly multidomain,

²³Note that Metadata simply refers to the structured information embedded in the web source, excluding any reference to the document's content. This is the typical definition used in the T-IR field and should not be compared to the terminology used in digital libraries (e.g., Dublin Core).

²⁴<http://lemurproject.org/clueweb09/> [May 26, 2014].

²⁵<http://lemurproject.org/clueweb12/> [May 26, 2014].

²⁶<http://datamob.org/datasets> [May 26, 2014].

²⁷<http://kevinchai.net/datasets> [May 26, 2014].

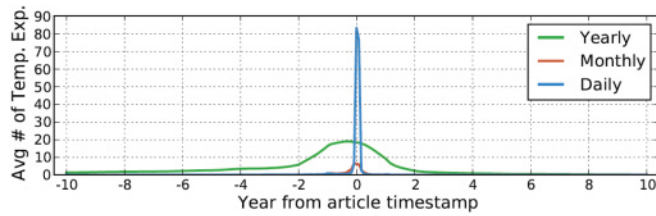


Fig. 7. Reference time of temporal expressions in news articles based on their granularity in relation to article timestamp [Jatowt and Yeung 2011].

ambiguity is common. An illustrative example is the expression “*New Year*” which refers to a different point in time in the United States or in China. Other problems are related to multilingual time formats (e.g., “*December 31, 2012*” is translated to “*31 de Dezembro de 2012*” in Portuguese). In this case, time taggers need to be built for each different language. Moreover, similar to the application of part-of-speech taggers, problems might surface when applying temporal taggers to micro-blog collections, such as web snippets or tweets. Indeed, their application may eventually result in poor outcomes, mostly due to a lack of background, which is caused by the small number of characters allowed for such sources (e.g., 140-tweet posts) and the specific language used to write these texts (e.g., “*tomorrow*” may be transcribed by “*Tomoz*”²⁸).

Jatowt and Yeung [2011], Campos et al. [2011b], Dias et al. [2011], and Jatowt et al. [2013b] recently conducted studies to understand the amount of temporal evidence embedded in web documents. Jatowt and Yeung [2011] studied the typical granularity and the range of temporal expressions in a collection of online news articles and concluded that news articles are more likely to contain daily granularity temporal expressions referring to the present, the immediate past, or the immediate future. This is shown in Figure 7, where most of the detected temporal expressions point to a 2-year-long time window with respect to the document timestamp. The near past and the near future tend to be referred to by fine granularity expressions, such as days.

Campos et al. [2011b] studied temporal evidence within a collection of 62,842 web snippets. They observed that roughly 10% contain explicit temporal information and that, similar to the work of Jatowt and Yeung [2011] and Jatowt et al. [2013b], most of the temporal expressions found within such collections are from the near past or the near future. Dias et al. [2011] analyzed a set of 508 web snippets looking for future dates. Their research results show that 82.48% of the future dates are related to the near future (i.e., a few months after the query time) and only 17.52% are related to a further future (i.e., at least 1 year after the query time).

3.3. Usage

Finally, the usage approach considers the extraction of temporal information from queries in a twofold perspective: *query timestamp* and *query focus time*.

The query timestamp, which is the date when the query was issued, is mostly used to understand changes in query popularity and search intent. This information is usually obtained from web query logs, which are flat sets of files that record server activity over time.

Query focus time is the content time of the query (i.e., the time to which the user’s query refers). Two types of queries are considered: (1) *explicit* temporal queries and (2) *implicit* temporal queries. Explicit temporal queries explicitly indicate a certain time period and contain a concrete date (e.g., “*Sapporo Olympics 1972*”) or an easily

²⁸http://en.wikipedia.org/wiki/SMS_language [May 26, 2014].

resolved temporal expression. Such queries represent about 1.5% of all queries [Nunes et al. 2008]. Another investigation [Campos et al. 2011b] reduced this value to 1.21% by eliminating queries that actually contain false-positive temporal expressions (e.g., “*form 1412*” or “*office 1997*”). Even though retrieving relevant documents related to explicit temporal queries appears to be a straightforward process, it can be problematic when the query contains false-positive temporal expressions. A simple solution is to make a temporal index that contains time intervals associated with each crawled document. These time intervals can be used to adjust the score of the document with regard to the query’s explicit temporal intent. As such, the documents about Blaise Pascal delimited by [1623, 1662] would not be retrieved for query “*Blaise Pascal 1450*” since the query time span falls outside the document boundaries. Although such a solution may be a step toward achieving a fully integrated temporal information retrieval solution, it still does not solve the problem of false-positive temporal query expressions whose time period fits within the document’s time span. In that case, a more elaborate approach that determines the time interval or a set of time intervals of a document is needed. We discuss this further in Section 4.6.

Implicit temporal queries point to a time period but do not contain any explicit date (e.g., “*Sapporo Olympics*” or “*Battle of Stalingrad*”). Following the work of Jones and Diaz [2007], such queries may be divided into three categories: (1) atemporal queries, which are queries not sensitive to time (e.g., “*rabbit*”); (2) temporal unambiguous queries, which are characterized by pointing to a concrete time period (e.g., “*first moon landing*”), and (3) temporal ambiguous queries, which refer to instances of either periodical events (occurring on a recurring basis, e.g., “*SIGIR*”) or aperiodic events (occasional peaks of popularity lacking periodicity, e.g., “*Haiti earthquakes*”).

Several research efforts have studied the profile of implicit temporal queries. Jones and Diaz [2007] asked a few annotators to classify 51 TREC *ad hoc* queries using TREC topic descriptions and concluded that 54% of the queries belong to the atemporal class and 46% belong to the temporally ambiguous one. They also developed an automatic classification of queries on the basis of the time profiles of news document collections and obtained approximately the same values. Metzler et al. [2009], in turn, relied on web query logs to infer the implicit temporal values of queries based on similar explicit temporal queries and concluded that only 7% have an implicit temporal nature. A more recent study [Kulkarni et al. 2011] explored query intent changes over 10-week time spans and found that 10% of the queries never spiked (atemporal queries), 47% spiked once (temporal unambiguous queries), and 43% spiked multiple times (temporal ambiguous queries). In contrast to these studies, Campos et al. [2011c] addressed the profile of implicit temporal queries based on a collection of web snippets and found that 75% are atemporal and 25% have an implicit temporal nature.

Although relying on web query logs may be a valid solution to infer the temporal value of implicit temporal queries, access to real-world query logs outside big industrial labs is difficult and a huge impediment to information retrieval research. As pointed out by Callan and Moffat [2012], this is mostly due to legal concerns about privacy issues. In their report on the use of proprietary data, they point to the AOL incident²⁹ as one of the reasons for companies to embrace caution about providing query log data. In 2006, AOL Research released a file containing 21,011,240 queries for over 650,000 users collected over a 3-month period, including anonymized user IDs, the time at which the query was submitted for search, the rank of the search result which the user clicked, and its corresponding URL. Even though the file was only intended for research purposes, it had to be removed from the Internet after a journalist from the *New York Times Journal* identified an individual solely based on the available information.

²⁹http://en.wikipedia.org/wiki/AOL_search_data_leak [May 26, 2014].

Despite this incident, some other search query collections have been publicly provided over the last few years. Microsoft, for example, released three large-scale datasets. The first two,³⁰ MSLR-WEB30k and MSLR-WEB10K, were intended for ranking purposes and consist of 30,000 and 10,000 queries respectively. QRU-1³¹ [Li et al. 2012], on the other hand, promotes query representation and understanding and can be used in a variety of tasks, such as query rewriting, query suggestion, query segmentation, and query expansion. A further possibility is to access Google Trends³² or the *New York Times*' most popular search queries.³³ Since none of these collections includes a wide range of explicit temporal queries, the process of inferring the temporal nature of queries implicitly formulated is hampered. It also opens a wide field for debate about user search intents. Indeed, as stated by Campos et al. [2011b], the simple fact that a query is year-qualified does not necessarily mean that it has a temporal intent (e.g., “*Microsoft office 2007*” or “*HP 1430*”) or that the associated year is actually correlated with the query (e.g., “*football World Cup 2012*”—there was no World Cup in 2012). A further challenge is that, contrary to the extraction of information within metadata or web contents, which simply requires a set of web search results, extracting temporal information from web query logs implies that some versions of the query have already been issued, thus contributing to query dependency. To tackle all of these problems, several alternative directions have been proposed. A more detailed discussion of this topic is provided in Section 4.3.

Next we introduce research studies in a number of different T-IR areas and pinpoint some of the crucial shortcomings of each.

4. TEMPORAL WEB INFORMATION RETRIEVAL

Information retrieval studies the process of searching for relevant information. Typically, this involves looking for information within texts, structured databases, or the web. Documents are then processed with respect to their similarity with the query and displayed as ranked documents, clusters, or similar structures. The general IR framework consists of the following four main steps: (1) document processing, (2) indexing, (3) query processing, and (4) ranking documents.

In this section, we cover the basics of building a web IR system geared toward the temporal dimension and we refer to some relevant applications of T-IR. Since we already described the document temporal processing stage in Sections 2 and 3, we will omit its description here. The remainder of this section is structured as follows. Section 4.1 introduces related studies on web crawling and web archiving. Section 4.2 offers an overview of temporal indexing. Section 4.3 presents research devoted to query processing. Section 4.4 shows the recent improvements achieved in the field of temporal ranking. Recent advances in temporal clustering, temporal text classification, temporal search engines, and future-information retrieval are described in Sections 4.5, 4.6, 4.7, and 4.8, respectively.

4.1. Web Crawling and Web Archiving

The first step of any IR process is to crawl the web by fetching the content of pages. This is done by a software component called a web crawler or a web spider. For each web page, the web crawler captures a snapshot of it at a specific time. Although performing this task is relatively easy, the huge number of web documents and their highly dynamic

³⁰<http://research.microsoft.com/en-us/projects/mslr/> [May 26, 2014].

³¹<http://research.microsoft.com/en-us/downloads/d6e8c8f2-721f-4222-81fa-4251b6c33752/default.aspx> [May 26, 2014].

³²<http://www.google.com/trends/hottrends> [May 26, 2014].

³³<http://www.nytimes.com/most-popular-searched> [May 26, 2014].

character raises challenges. The evolution of web content has been widely studied over the years [Baeza-Yates et al. 2002; Cho and Garcia-Molina 2003; Fetterly et al. 2003; Ntoulas et al. 2004; Bordino et al. 2008; Adar et al. 2009; Elsas and Dumais 2010; Kulkarni et al. 2011]. Overall, the results show that the web is constantly changing mainly due to the creation of more and more new pages and, to a lesser extent, due to the modification of the content of existing ones.

In such a dynamic environment, web archives are of increased importance as serving to preserve documents and prevent information loss. They contain information about how the web has evolved over time and can greatly benefit researchers who wish to study a particular historical period of the web. One of the most famous archives, the Internet Archive, which started in 1996 [Kahle 1997], has saved more than 357 billion web pages. Access to archival content is enabled by the WayBack Machine, which allows particular versions of a given web page to be found. Unfortunately, the system still does not allow a free text query search. Another interesting project, started in 2004, is the Internet Memory Foundation³⁴ (formerly European Archive) which provides a large-scale, open memory of the Internet. Several other research projects, like ARCOMEM, LAWA, LiWA, and LivingKnowledge (introduced in Section 1), have been conducted in Europe. Many countries have also launched a number of national web archiving initiatives³⁵ [Gomes et al. 2011]. Masanès [2006] offered a comprehensive overview of the methods, tools, standards, and difficulties inherent in the development of a web archive system. Another useful summary is the work of Gomes et al. [2013] who described lessons learned while developing the Portuguese Web archive. They focused on web data acquisition, ranking search results, and user interface design.

A wealth of research also exists on usage scenarios and applications of content stored in web archives. For example, McCown and Nelson [2008] and Klein and Nelson [2010] proposed using them to recover lost information on the web, and Van de Sompel et al. [2009] introduced a framework called *Memento*³⁶ in which archived resources can seamlessly be reached by their original URI.

Since the web is constantly changing, it is also becoming more difficult for search engines to maintain up-to-date indexes, which threatens the effectiveness of the search process and the usefulness of search results. This is particularly evident for recency-sensitive queries, for which the relevant set of documents changes frequently. This problem, known as the freshness of search results, is related to the notion of the credibility of presented information. For example, Yamamoto et al. [2007] demonstrated that issuing a query following a real-world change (e.g., nomination of a new president or a change in numerical values such as a country's population or the number of EU states) may still result in outdated information being retrieved due to the self-correcting latency of the web.

One possibility is to let crawlers detect document age, so that their schedule becomes more precise. However, detecting the freshness of a web page can be quite difficult since temporal metadata may neither be provided nor be trustworthy. Indeed, as discussed in Section 3.1, it is generally quite difficult to determine (with a high degree of certainty) trustworthy metadata (i.e., document creation time, document publication time, or last-modified date) based on information extracted during the crawling process. This gives rise to a new challenge called temporal text classification, whose main goal is to determine the time of undated documents. A more detailed analysis of this task can be found in Section 4.6.

³⁴<http://internetmemory.org/en/> [May 26, 2014].

³⁵http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives [May 26, 2014].

³⁶<http://mementoweb.org/> [May 26, 2014].

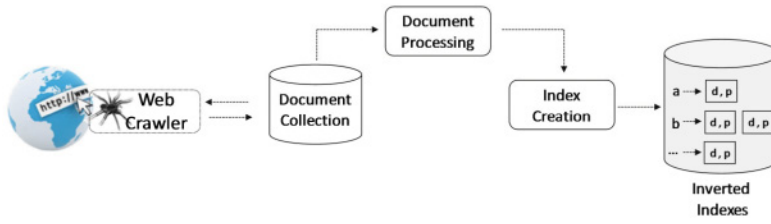


Fig. 8. Indexing process.

4.2. Indexing

Before indexing, documents must first be converted into a standard format. The component responsible for this task is the document processing module that requires the execution of four steps: tokenization, stopping, stemming, and information extraction, which have been already described in Section 2.5.

The core of the indexing process is the inversion module which transforms document-term pairs into term-document ones. As described in the survey conducted by Zobel and Moffat [2006], this is usually done using an inverted index structure. Several models can be adopted for this purpose. In Figure 8, we depict an example of an inverted index structure, where the dictionary terms are associated with a posting that contains two numbers: document d and term position p .

Although common in text search systems, a completely comprehensive search feature is often missing in most time search mechanisms. One reason is that this process is treated as an inherent part of document processing, when in fact it should be an independent task, supported by a two-layered index framework, where both document and temporal features are considered. This problem was first addressed by Berberich et al. [2007a, 2007b], who proposed a solution for a time-travel text search by extending the inverted file index to make it operational for temporal searches. Temporal information is explicitly incorporated in the posting list as part of the position item. The postings are thus of the form $(d, p, [t_b, t_e])$, where d is the document, p is the positional information indicating where a term appears in the document, and $[t_b, t_e]$ is a time interval, where t_b is the birth time of the document and t_e is its end time. This research was later extended by Anand et al. [2012] to allow the incremental addition of new document versions without rebuilding the index structure.

In addition to the aforementioned studies, other proposals have been implemented. Song and JaJa [2008] proposed a novel indexing structure based on the concept of multiversion B-trees and a duplicate detection algorithm to avoid storing duplicate web content between two consecutive crawls. They proposed a key consisting of a URL and a time interval $[t_b, t_e]$ during which the corresponding web content has not changed. Jin et al. [2008] also proposed a temporal search engine to answer temporal ranged queries. Their solution is supported by a hybrid temporal text index for web pages, where time and textual keywords are grouped into one uniform index structure based on a MAP21 index [Nascimento and Dunham 1999]. Arikan et al. [2009] suggested the creation of two types of indexes: one to store text documents as direct and inverted indexes (discarding all documents that do not contain any temporal expressions related to a temporal query) and a second one to store the temporal data extracted from the content of the documents by regular expressions (Section 2.5). Pasca [2008] proposed a temporal question answering system where both the dates as well as the text are stored in a fact repository. Lastly, Matthews et al. [2010] suggested the creation of two indexes, one for each document in the collection and one for each sentence. For the sentence level index, a content date is computed based on the document's focus time

(i.e., time to which document content refers to). If this information is not available, then the document publication time is considered.

4.3. Query Processing

The process of searching for information is inherently temporal. Even though some user information needs may be explicitly expressed, most are implicit by nature. In this section, we provide an overview of the relevant literature regarding the estimation of temporal intent behind user search queries. Following the work of Cheng et al. [2013], we categorize past research into two classes: (1) works that target recency-sensitive queries for which users expect documents to be both topically relevant and up-to-date (fresh), and (2) works that target time-sensitive queries, where the results are from a specific time period.

4.3.1. Recency-Sensitive Queries. The importance of information timeliness was recently studied by Joho et al. [2013]. In user survey conducted, nearly half of the 110 respondents stated that the information for which they search is related to the present (information on the same day); yet a significant fraction also searches for future-related or past-related information. Sixty percent of the subjects confirmed that the freshness of search results was important in their recent search activities on the web, usually in response to the formulation of spiky queries (e.g., “Halloween” on October 31), fast-changing phenomena queries (e.g., “weather Miami” or “dollar/yen rate”) or news-related queries (e.g., “Nelson Mandela” at his death).

With respect to a document, the property of recency is inversely proportional to the time elapsed since its creation. Detecting whether a query requires special treatment in terms of recency analysis is not easy, given that user search intents are usually underspecified. Poor effectiveness of detecting temporal query intent may actually constitute a problem because, as reported by Efron and Golovchinsky [2011], applying temporal approaches to non-recency-sensitive queries might harm the quality of search results. Most pioneering approaches have tackled this problem based on newswire collections using query volume and the number of published documents as an indicator of the query’s recency sensitiveness. Diaz [2009], for example, proposed a solution that detects the news intent of a query by studying its dynamics and its click-through rates and modeling the click probability by logistic regression. They tackled a specific problem, called a news aggregated search, which refers to the integration of fresh content extracted from news article collections into “regular” web search results. König et al. [2009], on the other hand, proposed a supervised learning method to estimate the click-through rate for news search results. Dong et al. [2010a] used an automatic classifier to detect whether an incoming query is a breaking-news one. Even though all of these research studies perform well in the specific context of news, a more general solution that addresses this problem by resorting to any type of documents, such as “regular” web pages or microblog collections remain undeveloped. Zhang et al. [2010], for example, developed a machine-learning method that combines multiple features into a classifier to determine queries occurring at regular time intervals, such as public events, lottery drawings, public holidays, TV programs, and so on. Features are derived from web query logs based on time series analysis, where the query frequency is measured at constant intervals. Similarly, Styskin et al. [2011] trained a regression model classifier based on 30 features to predict the need for retrieving the recent contents for a given query. Shokouhi [2011] detected seasonal queries using time series analysis.

A further problem is related to queries that despite nonspiky behavior may still benefit from retrieving more recent documents (e.g., “fashionable haircuts” and “phone call prices”). However, in this case, volume-based techniques cannot be applied, since the number of published documents or issued queries does not clearly reflect their

temporality. To tackle this problem, Cheng et al. [2013] estimated query timeliness using the term distribution change of a query's relevant documents over time. They focused on recency-sensitive queries that are not driven by news events and for which there are no major spikes in query or document volumes over time.

4.3.2. Time-Sensitive Queries. Other works focus more on *time-sensitive* queries, where the results are preferably from a specific time period. Time-sensitive queries include those whose set of search intents (or the main intent) changes over time, such that the relevant answers may also vary over time. They can also be interpreted as seasonal queries [Shokouhi 2011] (i.e., cyclic queries related to seasonal events). For example, the query "*Halloween*" issued on October 29 most likely indicates that the user is looking for a costume or a party, but the same query in January probably indicates that the user simply wants to learn about Halloween. Thus, the Wikipedia page is relevant in this case.

A clear understanding of the query temporal nature offers search engines the chance to decide whether to return more historical or more recent information to provide relevant results. Several alternative directions for identifying relevant query time periods have been previously explored in the literature. The methods proposed to solve this problem can be broadly classified into three different classes: (1) metadata, (2) usage, and (3) content approaches.

Following a metadata approach, Jones and Diaz [2007] used a language model solution and a collection of web news documents to model the period of time that is relevant to a query. More specifically, they estimate distribution $P(t|q)$, where t is the day relevant to query q . They adopt a relevance modeling solution that considers not only the probability of the document's relevance, given by $P(q|d)$ but also the temporal information about the document, given by $P(t|d)$, where t is the day relevant to that document (0 if t day is not equal to the document timestamp and 1 otherwise). Kanhabua and Nørvåg [2010] proposed three different methods to determine the time of implicit temporal queries: (1) dating queries using only query keywords, (2) dating queries using the retrieved top- k documents, and (3) dating queries using the timestamp of the retrieved top- k documents. Dakka et al. [2012] proposed a solution that takes into account the publication times of documents to identify the important time intervals that are likely to be of interest to an implicit temporal query. Time was incorporated into language models to assign an estimated relevance value to each time period. They also built a framework that represents each document d by two components: content component c_d and temporal component t_d , where $P(c_d, t_d|q)$ represents the probability that c_d is topically relevant to query q in time period t_d .

Unfortunately, all of these approaches rely on the creation date of documents as correct temporal signals, which are far from being credible in most cases. Moreover, such information is not available in many documents. As observed by Kanhabua and Nørvåg [2010], the fact that all these studies are built on top of temporal language models involves drawbacks concerning document collection. In particular, the used documents must be timestamped and cover the time period of the queries.

An alternative solution to using metadata was proposed by Vlachos et al. [2004], who developed a method to discover important time periods using the query logs of a commercial search engine. Likewise, Metzler et al. [2009] suggested mining query logs to identify implicit temporal information needs. They proposed a weighted measure that considers the number of times query q is pre- and postqualified with given year y . A query is then implicitly year-qualified if it is qualified by at least two different years (e.g., "*Miss Universe 1990*" and "*Miss Universe 1991*"). A relevance value is then given for each year found in a document. Based on this, they proposed a time-dependent ranking model that explicitly adjusts the score of a document in favor of

those matching the users' implicit temporal intents. Though an interesting first solution this approach lacks query coverage since it depends on the analysis of query logs. Shokouhi and Radinsky [2012] proposed a time-sensitive approach for query auto-completion by applying time series analysis. Their results show that predicting the popularity of queries by more elaborate time series analysis and periodicity estimation is more reliable than straightforwardly using information on past query popularity derived from web query logs.

While the aforementioned models rely on spikes in the distribution of relevant documents or queries, none extracted temporal information from web contents. To the best of our knowledge only one study [Campos et al. 2012a] has taken up this research so far. They proposed a language-independent temporal similarity measure called *GTE*, which, based on corpus statistics, associates relevant date(s) to a query while filtering out nonrelevant ones.

4.4. Temporal Ranking

Document relevance estimation can greatly benefit from the introduction of temporal aspects into ranking models. Based on such observation, researchers have started to address the problem of returning documents that are not only topically relevant but also that are also from the most important time periods. Under this assumption several works have been proposed that can be broadly divided into two classes: (1) those favoring more recent documents (recency-sensitive ranking) and (2) those that target documents from different time periods (time-sensitive ranking). Both classes are closely related to the categories defined in the query processing section since the method used to identify queries with a need of recency or time-sensitive treatment is intrinsically related to the methods used in the ranking stage.

4.4.1. Recency-Sensitive Ranking. One of the first works that implemented recency-sensitive ranking was proposed by Li and Croft [2003]. In their study, the authors introduced the notion of time-based language models as an extension of work proposed by Ponte and Croft [1998] to favor documents created in recent time. Instead of assuming uniform prior probabilities in the retrieval model, they assign document priors using an exponential decay distribution over the creation dates of documents. Thus, documents with a more recent creation date are assigned a higher probability. Berberich et al. [2005] introduced two approaches based on link analysis, T-Rank Light and T-Rank, taking into account both the freshness (i.e., the timestamps of the most recent updates) and the activity (i.e., update rates) of pages as well as links to retrieve recent documents. Similarly, Cho et al. [2005] relied on web link structure, especially its evolution, to propose a new ranking metric to solve the problem pointed out by Baeza-Yates et al. [2002], who demonstrated the temporal bias of PageRank. In its standard form, the PageRank algorithm fails to promote newly created relevant web pages because acquiring links usually requires considerable time. Subsequently, Li et al. [2008] tried to improve PageRank by assigning a nonfixed dumping factor governed by a function that depends on the time elapsed since the last update of the pages.

Next, Zhang et al. [2009] described a reranking score adjustment to give a ranking boost to fresh documents. The overall process assumes implicit temporal queries as input and relies on the extraction of temporal features from documents, especially from their titles, URLs, and anchor texts. Documents with more recent dates occurring in these fields are thus ranked higher. Dai and Davison [2010] estimated the web page authority by determining the variation of page and in-link freshness over time and incorporated this information into a temporal ranking probabilistic model called T-Fresh.

Within the context of learning to rank, Dong et al. [2010a] proposed a retrieval system to answer recency-sensitive breaking news queries. Such queries are first classified with regard to their recency sensitivity before being sent to the ranker. Document freshness is taken into account by combining multiple temporal features that represent document recency and the recency demotion of $\langle \text{query}, \text{urls} \rangle$ pairs in the training stage. The training data are then used to learn a ranking function. The models proposed to incorporate regular ranking trained data to solve the problem of insufficient recency information.

The methods put forward by Dong et al. [2010b] and Inagaki et al. [2010] use user click feedback features to identify how document relevance varies over time. More concretely, Dong et al. [2010b] incorporated fresh URLs extracted from Twitter into a general web search system. Using the labeled $\langle \text{query}, \text{url} \rangle$ training data pairs, a machine-learning ranking algorithm can predict the appropriate ranking of the search results for unseen queries. Inagaki et al. [2010] also proposed a set of novel temporal click features and query reformulation chains to improve the machine-learning recency-sensitive ranking by favoring URLs that have been of recent interest for the user's recency-sensitive query.

In contrast to Dong et al. [2010a], who selected one particular ranker per query type, Dai et al. [2011] proposed a framework where each query is run against a set of rankers. Consequently, weights vary based on the temporal profile of a query, thus minimizing the risk of poor performance when queries are misclassified in terms of recency intent. Another recent research is the work of Styskin et al. [2011], who relied on a recency-sensitive query classifier to apply result diversification by combining ordinary search results with fresh documents. Efron and Golovchinsky [2011], in turn, proposed an extension of the query likelihood model that considered not only when a document was published but also the relationship between the publication time and the query. They also proposed temporally informed smoothing, so that older documents that are further from the target time associated with the query are smoothed more aggressively.

Within the context of microblogging, Efron [2012] proposed survival analysis³⁷ to incorporate recency information into document ranking by following a query-dependent approach. While the results remain preliminary, their research opens up debate for future research directions.

More recently, Cheng et al. [2013] presented a language model that incorporates the timeliness factor to retrieve fresh recent results for nonspike timely queries. The proposed model can be used for different query types and does not depend on the distribution of documents over time since the timeliness factor of a query is determined using the term's distribution change of a query's relevant documents over time.

4.4.2. Time-Sensitive Ranking. Rather than concentrating on the retrieval of fresh documents, other studies propose more general time-sensitive frameworks where the results are adjusted upon longer time periods. Perkiö et al. [2005] automatically detected topical trends and their importance over time within a news corpus using a simple variant of TF.IDF. These trends were then used as the basis for temporally adaptive rankings; the ranking of the results for query q at time t should promote documents whose most prominent topics are the same as the most active topics within the whole corpus at time t . Jin et al. [2008] proposed a new ranking algorithm to sort results by applying a linear interpolation of three factors: text similarity, temporal information, and page importance. Text similarity represents the ranking scores of text relevance and depends on the frequency of query keywords and their corresponding locations in web

³⁷A branch of statistics applied in diverse fields, also called reliability analysis, which studies the amount of time until one or more events happen.

pages. Temporal similarity is the ranking score of temporal relevance based on the set of intersection conditions between the temporal query and the temporal expressions found in the web page. Page importance represents the ranking score of the importance of the web page based on the PageRank algorithm [Brin and Page 1998].

Metzler et al. [2009] considered a web query log dataset and a set of document fields (e.g., title and anchor text) to estimate both the times of the query and the document. Based on this, they proposed a time-dependent ranking model to explicitly adjust the score of a document in favor of documents matching the user's intent expressed by an implicit temporal query.

Arikan et al. [2009] were the first to propose an approach that integrates temporal expressions extracted from the document content in a language modeling framework. Similarly, Berberich et al. [2010] proposed a temporal retrieval model that integrates temporal expressions into query-likelihood language modeling. However, unlike in Arikan et al. [2009], uncertainty in temporal expressions is explicitly considered, both in the query and in the document, so that temporal expressions can be linked to the same time point even if they are not exactly equal (e.g., "1998" and "XX"). Elsas and Dumais [2010] developed a language-model-based ranking algorithm that incorporates the dynamics of document content changes using term frequency distribution over time. For example, although it may be advantageous for a recency-sensitive query to have a high weight set on recent terms, for navigational queries, it may be better to focus on content that is stable and present within many past versions. Kanhabua and Nørvåg [2010] used a query's determined time to improve the reranking of the web page results. The idea behind this research is that documents with creation dates that closely match the query's time are more relevant in the temporal dimension and thus should be ranked higher. To achieve this goal, they proposed a mixture model that linearly combined both textual and temporal similarity.

Aji et al. [2010] introduced a new term weighting model that uses the revision history analysis (RHA) of a document to redefine a term's importance, assuming that a term should be as relevant as the number of times it occurs in the different versions of a document gets higher. A decay factor is included so that the terms in older versions of the document get a higher value. RHA is then incorporated into BM25 and statistical language models, so that documents get ranked based on the importance of the terms in the past.

More recently, Kanhabua and Nørvåg [2012] proposed a new approach by applying a time-sensitive ranking model based on learning-to-rank techniques for explicit temporal queries. To learn the ranking model, they applied two classes of features: temporal and entity based. For the temporal-based ones, both the document focus time and the timestamp are combined. Entity-based features, on the other hand, are used for inferring semantic similarity (such named entities as person, location, or organization). An unseen <document, query> pair is then ranked by the weighted sum of the feature scores. The results show that the SVM^{MAP} learning-to-rank model outperforms the proposed method of Berberich et al. [2010]. Chang et al. [2012] reranked the search results based on user intents at different times of day using the user temporal click information obtained from query logs.

Other works have explored the temporal dimension in specific types of temporal ranking. For example, Pasca [2008] proposed a temporal question answering system that defines regular expressions to detect dates that meet certain requirements. The dates found in the document content should provide direct answers for the user's query (e.g., "*When was the Taj Mahal built?*"). Documents are processed offline, and their content is stored in a fact repository. Whenever a new query is issued, the system matches the query through a Boolean search against the text stored in the repository, scores each match individually upon text heuristics, and aggregates the texts

associated to the same date by combining the scores of the matching function. Dates with the highest score are then retrieved. In the particular context of temporal clustering (Section 4.5), Alonso et al. [2009a] proposed a measure to rank documents within a cluster based on the number of times a query occurs in a sentence with explicit, implicit, and relative temporal expressions. Kanhabua et al. [2011] also proposed a ranking model for future predictions using a learning-to-rank algorithm trained over a set of labeled query/prediction pairs. Many features are used to measure the similarity between a news article query (which is automatically generated) and the prediction. The query/prediction pair is ranked by the weighted sum of such feature scores as term similarity, entity-based similarity, topic similarity, and temporal similarity. Finally, Strötgen and Gertz [2013] presented a novel ranking approach that takes into account the proximity of text, temporal, and geographic query terms in documents to answer queries with a temporal and spatial information need, that is, queries of the form “when and where did something happen?”

4.5. Temporal Clustering

In this section, we focus on the temporal clustering of web page results, a relatively new subfield of T-IR. To the best of our knowledge, only few studies have been proposed so far. The first [Alonso and Gertz 2006] represents a document as a vector of temporal attributes extracted from its metadata (e.g., creation time) and from its content (by applying Annie temporal tagger; see Section 2.5). Documents are then clustered using a complete-link hierarchical clustering algorithm which results in a set of hierarchical clusters with two possible views: topical and temporal. Alonso et al. [2009a] introduced *TCluster*, an overlapping clustering algorithm, where each document is associated with a temporal document consisting of a list of 3-tuples, $\langle E, C, P \rangle$, where E is the set of all temporal expressions detected (explicit, implicit, and relative) within a document, C is the set of respective normalized time units (day, week, month, and year), and P is the set of their positions. Clusters are formed by a set of documents sharing the same year such that the more frequently the query occurs close to the set of temporal expressions, the more relevant the document is in the cluster.

Unfortunately, none of these works filtered out temporal patterns, which may lead to the selection of noisy information. A possible solution to this problem was first introduced by Campos et al. [2012b, 2014] who identified relevant temporal expressions in web snippets by applying a clustering solution in which documents were grouped into the same cluster if they shared a common relevant year. The underlying methodology is based on GTE (Section 4.3.2): a temporal similarity measure that identifies the top relevant dates within a document while filtering out the non-relevant ones. The obtained results show that the introduction of GTE improves the quality of generated clusters by retrieving a higher number of relevant dates than previous approaches that consider all the temporal expressions found as relevant dates. However, since this proposal simply clusters documents on the basis of common dates, documents may not be topically related.

Temporal clustering has also been the subject of study in a number of diverse temporal applications. For example, Shaparenko et al. [2005] tracked events over time by clustering. Jatowt et al. [2009] introduced a clustering approach to summarize future-related information using text content, the content dates, and the timestamps of future predictions. Also, Jatowt and Yeung [2011] proposed a model-based clustering algorithm to detect future events based on the information extracted from a reference text corpus. Each instance extracted from the corpus is generated by first picking a topical cluster with a particular probability from which the terms and the temporal expression of the instance from it are generated. Each final cluster corresponds to a forecasted event with a certain expected period of its occurrence in the future.

4.6. Temporal Text Classification

In this section, we examine the process of determining a document's time dimension, which may be useful for improving several T-IR tasks, including crawling, indexing, query understanding, and the ranking of web search results. As observed by Kanhabua and Nørvåg [2008], this process can be divided into two categories: (1) determining a document's timestamp (i.e., the time when the document was created, published or last-modified) and (2) determining its focus time (i.e., the time to which its contents refer). As previously stated in Section 3.1, these two times may differ significantly.

Determining a document's timestamp was first studied by Toyoda and Kitsuregawa [2006] and Nunes et al. [2007]. Both used neighboring pages to estimate the document creation date or the last-modified date, assuming that temporal information can be extracted from the web structure. In particular, Toyoda and Kitsuregawa [2006] proposed a measure to estimate the document creation date based on the scores of linking web pages. Nunes et al. [2007] employed link structure analysis by considering three types of features derived from the web, incoming links, outgoing links, and HTML src attributes (e.g., `<img src = "URL"`), to date web documents for which the last-modified date is not available. Such an approach is limited however by two main constraints: (1) the need to have a large set of external documents and (2) the requirement that the last-modified dates exist in a set of external sources, which is not guaranteed. Moreover, as stated in Section 3.1, it remains unclear whether this information is reliable. Jong et al. [2005] and Kanhabua and Nørvåg [2008] approached this problem by determining the time of nontimestamped documents, namely, news articles, resorting to temporal language models, which suffer from some drawbacks (see Section 4.3). Jatowt et al. [2007] proposed using web archives to compare the content of the current version of a web page with its past versions in order to estimate approximate age of separate content elements on the page. SalahEldeen and Nelson [2013a] combined multiple evidences including the Last-Modified HTTP response header of the resource, the time of the last crawl by Google, the time when it first appeared in any public web archive, the time when the page was first tweeted, and the time when its URL was first time shortened. Related to the concept of resource combination and sharing [SalahEldeen and Nelson 2013b] is the *Memento* project for developing the Memento Protocol that adds a time dimension to the HTTP protocol [Van de Sompel et al. 2009]. The protocol introduces the notion of datetime negotiation that makes it possible for a client to request a past version of a resource at or around a specified time in the past.

Finally, for a more general document genre, Garcia Fernandez et al. [2011] proposed a system to automatically determine the publication date of French historical documents based on unsupervised and supervised algorithms. For the former, chronological methods supported by clues in the text (a person's name, newly created or old words, or spelling reforms) and external resources (Wikipedia, Google books unigram, or a French dictionary) are used to determine the document publication date. For the latter, classification methods, such as support vector machines (SVMs) compute temporal similarities between the document and a training reference corpus (Google books Ngrams).

While some methods determine a document's timestamp, few tackled the problem of determining the document focus time, especially in the case of the lack or the scarcity of temporal expressions within the document content. Clearly, a more generic solution is needed than just relying on the presence of temporal expressions which may be scarce or weakly related to the main document theme. Such an approach was first addressed by Jatowt et al. [2011]. In their study, the focus time of page p is the set of time periods resulting from the occurrence of events covered by the content of p . Events are detected by applying a clustering algorithm to the related news articles. The resulting clusters

are then compared to the content of the web pages. The event's occurrence time is estimated as the average timestamp of the news articles belonging to the underlying cluster. More recent work on document focus time estimation [Jatowt et al. 2013a] takes a statistical approach based on data derived from large news collections. First, the association between a term and any year is calculated using the sets of sentences that contain explicit temporal expressions. For example, their study could determine that the term "*hitler*" has the strongest association with the time period between 1939 and 1945, or the term "*atomic*" has a strong association with 1945 while having weaker associations with other years. Next, the terms strongly associated with only a few years (e.g., names of events or entities specific to a single year or to short time periods) are found using temporal entropy and temporal kurtosis measures. Such terms are weighted to reflect their discriminative characteristics to estimate the document focus time. Finally, the focus time of a target document is calculated as the weighted average association of document terms with years.

Parallel to this, Kawai et al. [2010] presented an approach to filter out noisy year expressions y from web snippets s that are temporally irrelevant to query q by applying machine-learning techniques trained over a set of labeled $\langle s, q, y \rangle$ triplets, where each triplet is represented by a set of text features. Although the incorporation of a date filtering process is novel, their proposal does not determine the degree of relevance for each temporal pattern. An improved solution to this problem was done by Strötgen et al. [2012], who proposed the first approach to identify the most relevant temporal expressions in text documents. Each temporal expression, which is extracted by applying the HeidelbergTime tagger (Section 2.5), is represented by a set of document and corpus-based features. The relevance of the temporal expressions is combined into a single relevance function based on a set of predefined heuristics.

Similarly, the works of Alonso et al. [2009a] and Campos et al. [2012b] introduced the notion of temporal clusters, which can also be used to associate each document with a given time span. Future work, however, must focus more on the identification of the most relevant expressions within longer web documents. These pose a few more challenges, mostly due to the possibility of diverse topics discussed in documents. A potential solution is to segment the text into different pieces based on the different topics discussed. Each part of the text can then be assigned a different time period.

4.7. Temporal Search Engines

With the growth of research in temporal information retrieval, search engines have started to exploit time to improve their search processes. One of the first uses of time, as pointed out by Manica et al. [2012], was to push to the top list of the results the most recent web pages based on freshness metrics that take into account the document timestamp (publication time or last-modified date). Though still used by popular search engines such as Google, Yahoo, and Bing in news domains or with spike phenomena, this method is of little relevance to users interested in historical information. For instance, a user may type a query "*Football World Cup Brazil*" and be more interested in the competition held in Brazil in 1950 than in 2014. State-of-the-art search engines however, will likely retrieve more recent web pages about the competition that took place in 2014 instead of information about the 1950 tournament. Indeed, for the aforementioned query example, it is necessary to wade through more than 60 web pages in the Google engine search results to find the first reference to the 1950 event.

A more elaborate mechanism gives users the possibility to specify a point in time or a temporal interval of their interest to filter out results outside that time period. For such approaches, the time attribute is again the publication time or the last-modified date of the web page. Over the years, commercial search engines have adopted this solution.

For example, Google has a time feature that allows users to filter their search results. Yahoo, on the other hand, has been experimenting with basic temporal refinement in their web search engine to filter results by the publication time of the document (past day, past week, or past month). Although this solution may be very effective to filter in more detail recently published documents, it may prove inefficient when the user is seeking timely information about a given topic. For example, a user searching for information about “*Blaise Pascal*” will hardly obtain relevant data about the date of his death or about his well-known works when applying such a filter since recent information about him will tend to be scarce. Moreover, the fact that the user has to specify a given time period naturally represents a shortcoming in terms of the user experience. Other problems, as discussed in Section 3.1, are related to the gap that exists between the document timestamp and the time to which the document contents refer as well as the inherent difficulty of extracting timestamp information from such unstructured documents as web pages.

Obviously, users will greatly benefit if a search engine system can explore the temporal information within web pages. However, most popular search engines do not explicitly consider the use of the temporal information extracted from web pages, and the construction of an effective end-to-end temporal system remains to be seen. This is particularly evident for implicit temporal queries (e.g., “*Haiti earthquake*,” “*BP oil spill*,” or “*Madagascar*”) for which one would benefit if a comprehensive temporal contextualization of the topic is offered.

To tackle this problem several research works have been proposed, leading to the emergence of a number of search engine temporal applications. We next present a list of academic prototypes and focus on research works that offer a framework fully dedicated to T-IR, including indexing, query processing, and the ranking of web search results to answer user queries with temporal information needs. To the best of our knowledge, a rather limited number of prototypes has been proposed. Alonso and Gertz [2006] described a prototype that provides users with an alternative presentation of the results by a hit list of documents clustered by temporal attributes. Alonso et al. [2007a] presented an exploratory search interface that uses timelines to explain and explore search results. Berberich et al. [2007b] implemented FluxCapacitor, a time-travel text search prototype that extends the inverted file index structure to deal with successive versions of the same document (e.g., searches on web archive collections). Jin et al. [2008] introduced TISE, a temporal search engine that supports content time retrieval for Chinese web pages. Vicente-Diez and Martinez [2009] proposed a temporal expression recognition and normalization system for Spanish contents that has been integrated into a web search engine prototype. Another work [Alonso et al. 2009a] outlined a prototype implementation as a web interface where users can explore results by clusters returned in response to a query. Matthews et al. [2010] proposed Time Explorer, a timeline search tool that enables analysis within a news archive collection about how news topics change over time. Kawai et al. [2010] demonstrated an on-demand search engine called ChronoSeeker, which allows users to find past/future events. Finally, Campos et al. [2012b] presented GTE-Cluster, an online temporal search interface, which consistently allows searching for topics in a temporal perspective by clustering relevant temporal web search results. GTE-Cluster results can be graphically explored by a demo search engine interface that is publicly available for research purposes. In Figure 9, we show an example of the GTE-Cluster interface [Campos et al. 2014] for the query “*Margaret Thatcher*.” Examples of different timelines were already presented in the Introduction section.

In Table I, we summarize the contributions made by different research works. We categorize each one by considering the temporal expression taggers used to extract temporal information, the extraction methodology, the base collection, the type of interface,

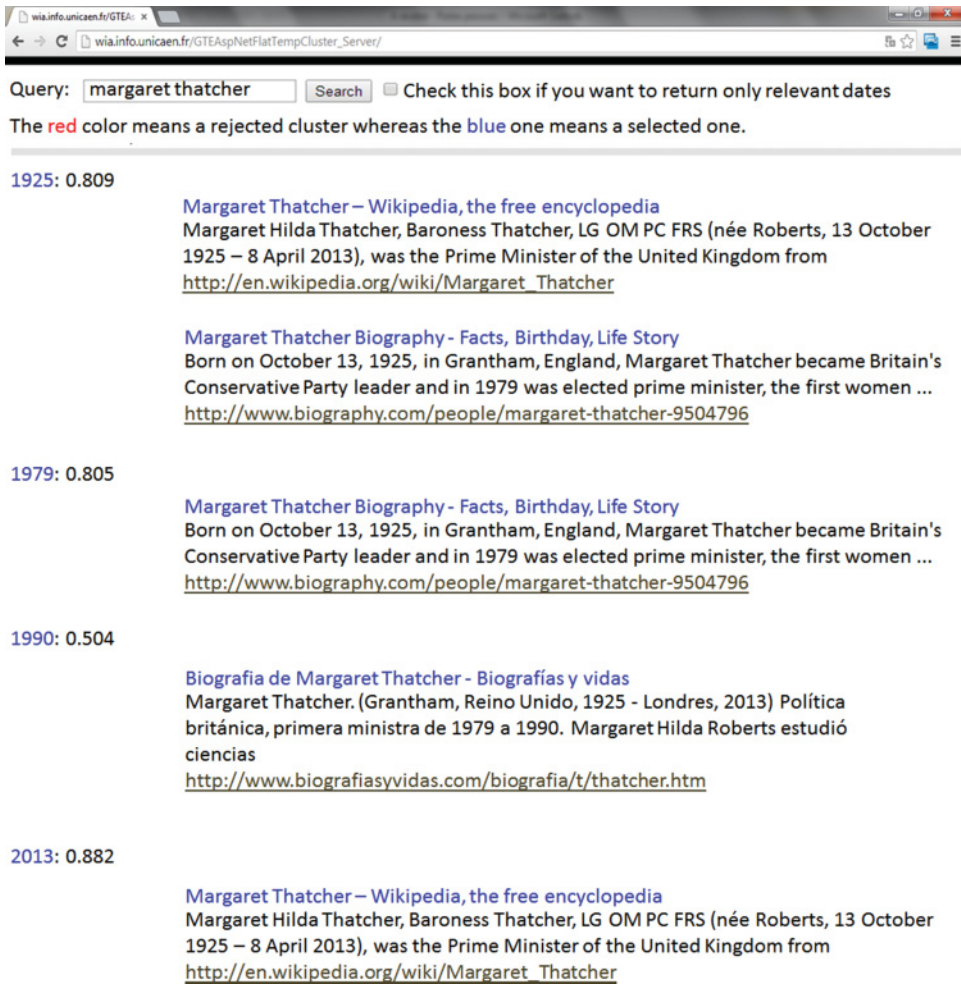


Fig. 9. GTE-Cluster interface for query “Margaret Thatcher,” extracted from http://wia.info.unicaen.fr/GTEAspNetFlatTempCluster_Server.

and whether an explicit temporal query is required. Details of these approaches have already been given throughout the text.

There has also been work on the search and the retrieval of combined geographic and temporal information. Strötgen and Gertz [2010b], for example, presented a prototype system called TimeTrails for the extraction, querying, storage, and exploration of the spatiotemporal information stored in text documents. YAGO2 [Hoffart et al. 2011] provided a search interface to seek temporal and spatial knowledge facts in knowledge bases. The concept of terminology evolution has also recently received attention from researchers. In particular, Holzmann et al. [2012] developed *fokas*, which is a search engine that offers user query suggestion terms of the original query based on a named entity evolution procedure.

4.8. Future-Related Information Retrieval

The task of supporting searches for future-related information is critical since thinking about and predicting the future are basic human activities and a significant number

Table I. Summary of Research in Temporal Search Engines

Name	Extraction Methodology	Temporal Tagger	Temporal Queries	Evaluation Dataset	Interface
Alonso and Gertz [2006]	Metadata Content	Annie	Implicit	N/A	Clustering
Alonso et al. [2007a]	Metadata Content	N/A	Implicit	DBLP	Timeline
Jin et al. [2008]	Metadata Content	TempEx	Explicit	N/A	List
Berberich et al. [2007b]	Metadata	N/A	Explicit	English Wikipedia	List
Vicente-Díez and Martínez [2009]	Content	N/A	Explicit	Newswire Articles	List
Alonso et al. [2009a]	Metadata Content	GUTime	Implicit	DMOZ TimeBank	Clustering
Kawai et al. [2010]	Content	Regular Expressions	Implicit	Web Snippets	Timeline
Matthews et al. [2010]	Metadata Content	TARSQI	Implicit	N/A	Timeline
Campos et al. [2012b, 2014]	Content	Regular Expressions	Implicit	Web Snippets	Clustering

of users are looking for content about future events [Joho et al. 2013]. Indeed, future information retrieval is a promising T-IR trend that offers many advantages, especially for supporting decision makers. To name a few, imagine a person who wishes to buy a Toyota car and needs to know whether the company is planning to release a new model. In another example, a prospective house buyer might like information about key urban changes scheduled for housing areas.

The study of the retrieval and the processing of future-related information from text collections has only recently begun. Baeza-Yates [2005] was the first to suggest a future search engine and *future-related information retrieval*. He proposed to extract future temporal expressions from news articles and represented documents using tuples of time segments and the confidence probabilities of future events. Jatowt et al. [2009] proposed two methods to summarize future-related information in web pages and news articles. The first extracts future-related information about any entity by issuing queries containing future dates and the entity name to search engines and clusters the returned results. The second method focuses on the periodicity analysis of recurring events in news article collections to forecast future occurrences. Jatowt and Yeung [2011] extended these concepts by taking into account the uncertainty of a piece of future-related information. In particular, the proposed clustering approach considered not only the textual but also the temporal similarities between sentences referring to future events.

From the information retrieval viewpoint, Matthews et al. [2010] proposed Time Explorer, a search engine that lets users search in the future and analyze the future evolution of topics. Kawai et al. [2010] analyzed effective ways to automatically categorize future-related information in documents using supervised learning, and Kanhabua et al. [2011] proposed a learned ranking model for news predictions that considers the weighted sum of a number of feature scores.

Some research [Jatowt et al. 2010, 2013b; Jatowt and Yeung 2011; Dias et al. 2011; Campos et al. 2011a] has also been conducted to understand the characteristics of the future-related information in news articles and on the web. For instance, Jatowt et al. [2010] analyzed future-related information on the web by showing the distribution of hit counts obtained from web search engines for queries containing future dates as well as by listing terms that appear frequently with different future years. Their

work compares the amount of information and the typical topics related to the near or distant future and finds that a significant amount of near-term future-related information refers to the events scheduled to happen until the end of current calendar year. Their study was later extended by the cross-lingual comparison and sentiment analysis of future-related information on the web as well as topical comparison with the future-related content in news articles [Jatowt et al. 2013b]. Jatowt and Yeung [2011] studied the time range to which future references refer on average in news articles and the granularity of these temporal expressions as a function of the temporal distance from the article creation date (Figure 7). Dias et al. [2011] and Campos et al. [2011a] discussed whether web snippets can be used to understand the future temporal nature of text queries and described the results of applying classification and clustering algorithms to group informative, schedules, and rumors texts. Note that these works were based on time-referenced future-related information. This means the analyzed data was associated with explicit time references that point to the future. Obviously, the majority of future-related references do not contain any concrete time points and are thus difficult to be mapped on the timeline. Kanazawa et al. [2011] estimated in a simple experiment that only about 30% of future-related references in news articles are time referenced.

The methods advocated by Weerkamp and de Rijke [2012] and Radinski and Horvitz [2013] tackled the problem of predicting future activities. Weerkamp and de Rijke [2012], for example, explored the use of Twitter to predict upcoming events that users may perform in the near future, based on tweet messages referring to a future time (e.g., “*Excited for bodypump class tonight!*”). The extraction of time references from Twitter messages, however, can be a particularly difficult task mostly due to its informal communication style nature and short message length (e.g., “2nite” instead of “tonight”). Radinski and Horvitz [2013], instead, aims to predict future events by mining 22 years of news stories from the NYT archive corpus toward the goal of identifying significant increases in the likelihood of disease outbreaks, deaths, and violence.

5. PROMISING RESEARCH DIRECTIONS

Quite a few challenges remain to be explored in T-IR. In the following we describe some future trends in information retrieval including references to studies that have been developed so far. We note that by no means this discussed list of directions meant to be complete, and there must be many more exciting proposals to be discovered.

Credibility. As the world continues to change, time-sensitive information can rapidly become invalid. Particularly, future-related information is inherently uncertain in contrast to past-related information. One problem is the validity of future-related information, which arises from the gap between the timestamp (e.g., creation time) and its reading time. For example, imagine a sentence about Toyota planning to establish a new plant in Thailand. Suppose that, actually, this prediction soon afterward became outdated (e.g., the company decided to cancel the previous plan). An unsuspecting reader might be easily misled when reading the sentence. Users would thus benefit from automatic warnings when encountering future-related information that has become invalid. To filter out obsolete future-related information and to eliminate invalid predictions one can compare such information of the reports occurred with newer, related predictions [Kanazawa et al. 2011].

Other solutions that measure the trustworthiness of temporal information could be based on the document type, for example, putting more emphasis on news articles from major and reputable newspapers and less on articles published in less credible blogs or documents of unknown authorship. A further alternative is to compare this information against more reliable sources such as Wikipedia [Döhling and Leser 2014; Whiting et al. 2014]. In general, credibility estimation can be improved by considering

the source of the article, its linguistic style, citation count, and so forth. In addition, paying attention to the timestamp of predictions is critical because newer information tends to be more reliable than old information.

Modal expressions such as “*might*,” “*could*,” or “*is likely to*” are often used when news articles mention future events to indicate different levels of the certainty of events or different levels of confidence put into the predictions by the document’s author. Naturally, weighting instances by the modal expressions found near temporal expressions might improve accuracy. Moreover, often events are not totally independent from one another. It is not uncommon to see sentences in the form of “A will occur if B and C happen.” In other words, the probability of one event may be dependent on the probabilities of other related events.

The uncertainty of a future event influences the precision of temporal expressions in news articles. When an event is very likely to happen, usually the date and the time mentioned in news articles are more exact (finer granularity). Considering the distance to the event’s occurrence date and the actual granularity of the temporal expressions used to describe it might provide additional evidence. Note that credibility issue is also important for other types of temporal information besides the future-related information, especially information about recent events or fast-changing phenomena.

Memory Studies and Computational History. The temporal information in text collections can also provide a wealth of information for historical and memory studies. *Collective memory* [Halbwachs 1992] can be analyzed in a similar way as the analysis of collected predictions discussed earlier. With current text mining techniques, it has now become possible to measure society’s attention and focus when it comes to remembering past events and topics. One way to do this is by extracting the context of temporal expressions that refer to the past, whether recent or distant, from large-scale collections that reflect current concerns and interests of society, such as book-based Ngram datasets (Google Books Ngram), blog datasets, web page collections or news article collections [Au Yeung and Jatowt 2011]. This line of research can be extended in many ways, for example, by capturing more implicit remembrances of the past (e.g., historical events or person names) rather than explicit ones in the form of dates. One can study sentiment levels associated with the past, with certain events, or with historical entities or compare the collective images of the past in different document genres (e.g., blogs, books, news) as well as conduct cross-country comparisons as done in Au Yeung and Jatowt [2011]. In another related work, Kanhabua [2013] studied the problem of collective forgetting, especially the notion of forgetting employed as a means for making archival decisions on what should and should not be preserved.

Related is the emerging field of *computational history* (see, e.g., Michel et al. [2011], Hoffmann [2013], and Au Yeung and Jatowt [2011]) that uses digital historical texts or other artifacts to provide new types of knowledge or information interpretation either for general purposes or for supporting historians. New computer science techniques can be proposed to verify and validate historical assumptions. Some examples are exploratory interfaces over long-term document collections for supporting the work of history and social scientists [Odijk et al. 2012; Reinanda et al. 2013; Michel et al. 2011; Matthews et al. 2010] or data mining approaches for large-scale data analysis [Cook et al. 2012; Au Yeung and Jatowt 2011; Huet et al. 2013]. Several interdisciplinary events have also started to appear, such as Digital Humanities Conference,³⁸ Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,³⁹ or the Workshop on Histoinformatics.⁴⁰

³⁸<http://adho.org/conference> [May 26, 2014].

³⁹<http://sighum.science.ru.nl/latech2013/> [May 26, 2014].

⁴⁰<http://www.histoinformatics.org> [May 26, 2014].

Temporal Text Similarity. An interesting way to look at the present is to compare it to the past. In particular, when long periods of time are concerned, it may be important to understand the relation between old concepts and recent ones, for example, helping users choose appropriate queries when searching for collections of documents written in the distant past (the vocabulary mismatch problem in web archive searches). For instance, at the beginning of the 19th century, coaches were the cars of the 20th century. Currently, the word “coach” is usually associated with trainers, whether in athletic contexts or in business, showing that text similarity cannot rely on words but rather on concepts. A new research task can be defined to measure word or text semantic similarity over long time periods to find similar words, entities or events over time. This research direction presents many interesting challenges at the frontier of natural language processing and information retrieval. Already some computational advances have been made. For example, Berberich et al. [2009] developed a technique to reformulate user queries that rests on a novel measure of across-time semantic similarity, contributing to minimize the problem of terminology evolution when searching through web archives. Radinsky et al. [2011] proposed a new semantic relatedness model, temporal semantic analysis (TSA), which constructs a time series for each word of the NYT collection on the assumption that two words are highly related if their time series are related as well. In another work, Tahmasebi et al. [2012] introduced NEER, an unsupervised method for the named entity evolution recognition independent of external knowledge sources. Odijk et al. [2012] demonstrated the environment for visualizing term evolution for understanding how the meaning of words changes over time.

Time-Focused Visual Search Interfaces. A relatively significant research focus has been put on using temporal information for exploration and search purposes, as previously stated in Section 1. Many visualization systems have been proposed to display temporal information in text. One of the first efforts in this field was proposed by Cousins and Kahn [1991] through the Time Line Browser, which provides a basis for the development of further models. That work was followed by Karam [1994] and Plaisant et al. [1996] with LifeLines, a general visualization environment for portraying the summaries of personal histories in the health and legal justice fields. SIMILE⁴¹ is an example of an end-user visualization and navigation tool for temporal document collections. This widget is relatively easy to use and works with XML data. Further research on this topic should focus on answering which is the best way to display temporal information in documents: document lists, timelines, temporal or/and topic clusters? Or using term clouds with encoded temporal information, for example, by adding tiny time series plots under each term [Lee et al. 2010]? This issue remains unanswered as the temporal visualization of documents is still underexploited.

Temporal IR and Social Network Services. Microblog collections, like tweets or Facebook posts usually include rich temporal information. This new type of data, however, poses some new problems, mostly due to short message length, noisy or novel expressions and the uncertain trustworthiness of information.

Within the overall context of Twitter, various types of research efforts have been conducted recently on temporal issues ranging from real-time event detection [Sakaki et al. 2010; Whiting and Alonso 2012; Osborne et al. 2012] to tweet classification [Takemura and Tajima 2012] or tweet-based timeline generation [Alonso and Shiells 2013]. For example, Sakaki et al. [2010] monitored tweets to detect earthquakes in real time. Whiting and Alonso [2012] identified events based on Twitter hashtags to construct timelines. Osborne et al. [2012], on the other hand, used Wikipedia logs to improve event detection from Twitter streams. They also point out that there is a

⁴¹<http://simile-widgets.org/timeline/> [May 26, 2014].

delay of 2 hours before events breaking on Twitter begins to be searched on Wikipedia. Another approach is the work of Takemura and Tajima [2012], who categorized tweets into different classes based on their information value and decay over time. In this way, the highly dynamic characteristic of tweets can be better assessed to promote tweets with high informational value at given time points. For instance, the “it is raining outside” message has little informational value on the next day, but the “Hawaii is beautiful” message will retain its value for a long time. Alonso and Shiells [2013] demonstrated the concept of using Twitter to automatically summarize such events as World Cup matches through timeline generation by finding important or interesting time periods within these events.

Popular social networks like Facebook or Twitter include temporal evidence in their timelines and message posts. These SNSs are known to produce event-related information almost instantaneously. Thus, it is important to estimate the trustworthiness of the data since event-related information may change quickly during the event development. One possibility is to confirm its credibility by comparing with external sources, such as Wikipedia [Whiting et al. 2014]. Another important issue concerns the privacy control of users over their posts. Some users want to make their data more private while others prefer to keep their contributions more visible [Bauer et al. 2013; Ayalon and Toch 2013]. In the future, more effective use of social graph data (such as those enabled by Facebook social graph API access) is expected to improve search effectiveness, including the temporal aspects of retrieved information [Ugander et al. 2011; Bakshy et al. 2012; Bannur and Alonso 2014]. A further interesting aspect concerns social-based search and social-based recommendations, which may be seen as an essential part of the leading search engines in the coming years. For example, the temporal aspects of Amazon’s suggestion “*other users like you*” or similar systems should be considered as users interests and trends change continuously.

T-IR Standardized Tasks. Research in information retrieval and extraction is fostered through the availability of standardized, open test collections and task challenges, thanks to which research communities can compare diverse approaches to the same problems. Likewise, in the area of temporal IR and temporal information extraction, several research challenges have been proposed. One of the most well known is the TREC Temporal Summarization⁴² (TempSum) task, which is composed of two subtasks: Sequential Update Summarization and Value Tracking. The former requires finding timely, sentence-level, reliable, relevant, and non-redundant updates about a given developing event. The latter subtask tracks the values of event-related attributes with high importance to the event. Examples include the number of fatalities or the financial impact of an event. Both subtasks have clear temporal characteristics because the discovered updates have to be timely and relevant.

TREC Knowledge Base Acceleration⁴³ (KBA) is a challenge for filtering large streams of text to find documents that can support the updating of knowledge bases like Wikipedia, Facebook, LinkedIn, and so forth. Its subtask, called Streaming Slot Filling, seeks techniques to track the values of attributes and relations of a selected entity over time. Similar to the Temporal Summarization task, the recency of information is critical. Both TempSum and KBA are constrained to the information on a given past event or entity and focus on a particular type of information, such as event attributes or event relations. On the other hand, the Temporal Information Access⁴⁴ (Temporalialia) task [Joho et al. 2014] hosted by NTCiR⁴⁵ asks participants to categorize queries into

⁴²<http://www.trec-ts.org> [May 26, 2014].

⁴³<http://trec-kba.org/trec-kba-2013.shtml> [May 26, 2014].

⁴⁴<https://sites.google.com/site/ntcirtemporalialia/> [May 26, 2014].

⁴⁵<http://research.nii.ac.jp/ntcir/index-en.html> [May 26, 2014].

predefined sets of classes: atemporal queries, past-related or future-related queries, or recency-related queries. Temporalia also introduces a second subtask of time-sensitive ranking of news articles for different classes of temporal queries based on its own document collection composed of about 3.1 million news articles and blogs created from May 2011 to March 2013. Finally, the GeoTime⁴⁶ challenge answers mixed geo-temporal information and needs represented by such questions as “When and where did George Kennan die?” or “When and where were the last three Winter Olympics held?” The temporal component of the answers was in the form of a date or a period/interval type variable.

Temporal Aspects of Web Snippets. Other recent research is also related to constructing effective query-based summaries of results. For example, Alonso et al. [2009b, 2011] introduced the notion of temporal web snippets, where the text is partly replaced by a number of relevant temporal expressions. Svore et al. [2012], on the other hand, include recent temporal content in web snippet texts. Their results suggest that for trending queries, displaying new temporal content can be quite useful for users.

Temporal Web Image Retrieval. Another important topic is temporal image retrieval, which is defined as a process that retrieves sets of web images based on temporal intent behind text queries. As in document retrieval, image retrieval queries may sometimes have an explicit or implicit temporal intent. A temporal query can be used to obtain images of past or future objects (e.g., car).

We divide the problem of returning images, which satisfy temporal text queries, into the following subtasks: (1) detecting and recognizing the temporal intent of a user query, (2) finding relevant images, and (3) returning relevant images that satisfy the temporal intent in the query. Step (1) resembles the task of temporal intent detection within queries for searching textual documents, although it may need to be adapted for image searches. Step (2) has been well studied so far, and many successful methods have been proposed. Of particular interest is step (3) that filters out images obtained from step (2) which are not representative of the required time period. When assuming the existence of annotated image collections with timestamp metadata, this step essentially contains any of the temporal ranking methods used for text documents described in Section 4.4. However, in unstructured collections such as the web, many images do not contain explicit metadata that can be easily retrieved. Thus additional methods must be proposed that automatically estimate the timestamps of images and can be scaled to large-size collections. Solving this step would also help satisfy users who are interested in the evolution of entities by generating timeline-like overviews that contain representative images for significant years.

Dias et al. [2012] proposed two approaches for solving this problem. In the first approach, they used ephemeral clustering (postretrieval clustering) to cluster web search results on the fly as they are returned by a web search engine. Web image results are retrieved by temporal query expansion (e.g., the query “*Olympic Games*” is expanded to “*Olympic Games 2012*,” “*Olympic Games 2008*,” etc.). In the second method, they estimate the approximate age of images by SVMs trained over a collection of temporally annotated old images. A similar method was also investigated by Palermo et al. [2012]. The authors extended their research by comparing the results obtained by classification task and those of a user survey where untrained humans classified the same set of photos into corresponding decades. Lastly, an approach for satisfying time-sensitive queries was also recently proposed for image retrieval [Kim and Xing 2013]. More specifically, the authors extracted temporal patterns from Flickr datasets (e.g., the time when the photo was taken) to rank images when an explicit temporal query is issued.

⁴⁶<http://metadata.berkeley.edu/NTCIR-GeoTime/> [May 26, 2014].

Temporal Query Similarity. Can two queries be considered similar based on the temporal features shared by the documents they return? The central idea here is to infer if two queries are semantically related based mainly on their temporal information. This issue can be illustrated by such queries as “war” and “peace” that are related over time, although they usually appear in different documents. A possible application in this scope is query expansion.

Time Period Query Expansion. Predicting a query’s temporal intent is a critical step to decide appropriate ranking. Thus it is of high importance to develop temporal predictive models that identify queries that may benefit from personalized time-sensitive results (see Section 4.3). However, none of the research studies has proposed time period query expansion, which is mostly due to the fact that systems continue to adopt a simplistic approach that reduces temporal expressions to a single point in time rather than to a time span. Within this context, detecting periods for entities is certainly an interesting challenge that may receive attention in the next few years. For example, the query “Obama” might suggest a set of period queries “Obama 1961 – 2003,” “Obama Illinois senator 1997–2004,” or “Obama president 2008–2012.”

Temporal Diversity. Another challenge is developing an approach that provides users with diversified results depending on query intents. Within this scope, we should consider different dimensions, such as topicality, spatiality, and of course, temporality. Gathering all these dimensions into a single model seems a promising research area both for web search and the visualization of web search results. A recent work [Berberich and Bedathur 2013] explores the concept of temporal diversification and proposes an approach in which search results are composed of documents that were published at diverse times of interest to the query.

6. CONCLUSION

Time is obviously one of the key dimensions of our lives, and timeliness is a fundamental feature of information quality. In recent years, time-aware document processing has been gaining increased importance in information retrieval and in a large number of its subareas. However, despite the fact that documents are full of temporal expressions and many have strong temporal characteristics, current IR systems still do not sufficiently exploit this information. As an example, when a user’s information need includes temporal aspects, traditional IR systems may fail because they continue to treat temporal expressions as normal text terms.

Consider the following elucidative example. For the query “*FIFA World Cup*,” a traditional IR system barely returns a document concerning “*FIFA World Cup in 1994*,” yet it would probably retrieve more current results, such as 2014 or even 2018. Another example is the query “*FIFA World Cup Germany*” which mostly returns results related to 2006 as opposed to 1974, due to the typically high importance of recency, thus downplaying the subject’s historical perspective. What these two examples show is that neglecting the temporal dimension is a key search signal that some content has been omitted.

In this survey we overviewed the important advances in T-IR, a new IR subfield. We first outlined crucial concepts related to the notion of time, calendar systems, handling temporal expressions in texts, and the different types of sources of temporal information on the web. We then surveyed existing research that deals with the temporal aspects of both search queries and documents and the diverse ways of generating temporally enhanced search results. Finally, we provided a list of promising research directions.

Despite the growing importance of the area, this recent research trend is still without immediate or at least visible effects for average users since most of the research developed so far has a rather specific scope or is still to be commercialized. Thus, a

number of significant advances must be made before search engines can entirely understand the temporality of a query and correctly reflect it in the returned results. We particularly emphasize the detection of the implicit intents inherent in temporal queries, the development of retrieval models that include temporal features extracted from web documents, and the presentation of the results based on the query type. A further problem is related to the difficulty of evaluating research proposals, since in many cases the community still lacks a gold standard to which most of the approaches can be compared.

REFERENCES

- Advanced Research Projects Agency. 1993. Software and intelligent systems technology office. In *Proceedings of the 5th Conference on Message Understanding*. Morgan Kaufmann.
- E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. 2009. The web changes everything: Understanding the dynamics of web content. In *Proceedings of the WSDM'09*. ACM, 282–291.
- A. Aji, Y. Wang, E. Agichtein, and E. Gabrilovich. 2010. Using the past to score the present: extending term weighting models through revision history analysis. In *Proceedings of the CIKM'10*. ACM, 629–638.
- J. Allan, J. Carbonell, G. Doddington, and J. Yamron. 1998. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. 194–218.
- F. J. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26, 11, 832–843.
- O. Alonso and M. Gertz. 2006. Clustering of search results using temporal attributes. In *Proceedings of the SIGIR'06*. ACM, 597–598.
- O. Alonso, R. Baeza-Yates, and M. Gertz. 2007a. Exploratory search using timelines. In *Proceedings of the Workshop on Exploratory Search and Computer Human Interaction Associated to CHI'07: SIGCHI*. ACM.
- O. Alonso, M. Gertz, and R. Baeza-Yates. 2007b. On the value of temporal information in temporal information retrieval. *SIGIR Forum* 41, 2, 35–41.
- O. Alonso, M. Gertz, and R. Baeza-Yates. 2009a. Clustering and exploring search results using timeline constructions. In *Proceedings of the CIKM'09*. ACM, 97–106.
- O. Alonso, R. Baeza-Yates, and M. Gertz. 2009b. Effectiveness of temporal snippets. In *Proceedings of the WSRSP'09 Workshop associated to WWW'09*. ACM.
- O. Alonso, M. Gertz, and R. Baeza-Yates. 2011. Enhancing document snippets using temporal information. In *Proceedings of the ISSPIR'11*. Springer-Verlag, 26–31.
- O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. 2011. Temporal information retrieval: challenges and opportunities. In *Proceedings of the TAW'11 Workshop Associated to WWW'11*.
- O. Alonso and K. Shiell. 2013. Timelines as summaries of popular scheduled events. In *Proceedings of the TempWeb'13 Workshop Associated to WWW'13*. 1037–1044.
- A. Anand, S. Bedathur, K. Berberich, and R. Schenkel. 2012. Index maintenance for time-travel text search. In *Proceedings of the SIGIR'12*. ACM, 235–243.
- I. Arikian, S. Bedathur, and K. Berberich. 2009. Time will tell: Leveraging temporal expressions in information retrieval. In *Proceedings of the WSDM'09*. ACM.
- C.-M. Au Yeung and A. Jatowt. 2011. Studying how the past is remembered: Towards computational history through large scale text mining. In *Proceedings of the CIKM'11*. ACM, 1231–1240.
- S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. DBpedia: A nucleus for a web of open data. In *Proceedings of the ISWC'07*. Springer-Verlag, 722–735.
- O. Ayalon and E. Toch. 2013. Retrospective privacy: Managing longitudinal privacy in online social networks. In *Proceedings of the SOUPS'13*. ACM, 1–13.
- R. Baeza-Yates, F. Saint-Jean, and C. Castillo. 2002. Web structure, dynamics and page quality. In *Proceedings of the SPIRE'02*. Springer-Verlag, 117–130.
- R. Baeza-Yates. 2005. Searching the future. In *Proceedings of the Mathematical/Formal Methods in Information Retrieval Workshop Associated to SIGIR'05*. ACM.
- E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the WWW'12*. ACM, 519–528.
- S. Bannur and O. Alonso. 2014. Analyzing temporal characteristics of check-in data. In *Proceedings of the WWW'14*. ACM, 827–832.

- L. Bauer, L. F. Cranor, S. Komanduri, M. Saranga, L. Michelle, M. K. Reiter, M. Sleeper, and B. Ur. 2013. The post anachronism: The temporal dimension of Facebook privacy. In *Proceedings of the WPES'13*. ACM.
- K. Berberich, M. Vazirgiannis, and G. Weikum. 2005. Time-aware authority ranking. *Internet Mathematics* 2, 3, 301–332.
- K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. 2007a. A time machine for text search. In *Proceedings of the SIGIR'07*. ACM, 519–526.
- K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. 2007b. FluxCapacitor: Efficient time-travel text search. In *Proceedings of the VLDB'07*. 1414–1417.
- K. Berberich, S. Bedathur, M. Sozio, and G. Weikum. 2009. Bridging the terminology gap in web archive search. In *Proceedings of the WebDB'09 Associated to SIGMOD'09*.
- K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. 2010. A language modeling approach for temporal information needs. In *Proceedings of the ECIR'10*. Springer-Verlag, 13–25.
- K. Berberich and S. Bedathur. 2013. Temporal diversification of search results. In *Proceedings of the TAIA'13 Workshop Associated to SIGIR'13*.
- K. Bollacker, P. Tufts, T. Pierce, and R. Cook. 2007. A platform for scalable, collaborative, structured information integration. In *Proceedings of the IWeb'07 Workshop Associated to AAAI'07*. AAAI Press, 22–27.
- I. Bordino, P. Boldi, D. Donato, M. Santini, and S. Vigna. 2008. Temporal evolution of the UK web. In *Proceedings of the 1st International Workshop on AND'08 Associated to ICDM'08*. IEEE Computer Society Press, 909–918.
- S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the WWW'98. Computer Networks: The International Journal of Computer and Telecommunications Networking* 30, 1–7, 107–117.
- B. C. Bruce. 1972. A model for temporal references and its application in a question answering program. *Artificial Intelligence*, 3: Elsevier. 1–25.
- J. Callan and A. Moffat. 2012. Panel on use of proprietary data. In *ACM SIGIR Forum* 46(2), 10–18.
- R. Campos, G. Dias, and A. M. Jorge. 2011a. An exploratory study on the impact of temporal features on the classification and clustering of future-related web documents. In *Proceedings of the EPIA'11. Lecture Notes in Artificial Intelligence - Progress in Artificial Intelligence*, 7026/2011: Springer-Verlag. 581–596.
- R. Campos, G. Dias, and A. M. Jorge. 2011b. What is the temporal value of web snippets? In *Proceedings of the TWAU'11 Workshop Associated to WWW'11*.
- R. Campos, A. Jorge, and G. Dias. 2011c. Using web snippets and query-logs to measure implicit temporal intents in queries. In *Proceedings of the QRU'11 Workshop associated to SIGIR'11*. 13–16.
- R. Campos, G. Dias, A. M. Jorge, and C. Nunes. 2012a. GTE: A distributional second-order co-occurrence approach to improve the identification of top relevant dates. In *Proceedings of the CIKM'12*. ACM, 2035–2039.
- R. Campos, A. M. Jorge, G. Dias, and C. Nunes. 2012b. Disambiguating implicit temporal queries by clustering top relevant dates in web snippets. In *Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE Computer Society Press, 1–8.
- R. Campos, G. Dias, A. M. Jorge, and C. Nunes. 2014. GTE-Cluster: A temporal search interface for implicit temporal queries. In *Proceedings of the ECIR'14*. Springer-Verlag. 775–779.
- P.-T. Chang, Y.-C. Huang, C.-L. Yang, S.-D. Lin, and P.-J. Cheng. 2012. Learning-based time-sensitive re-ranking for web search. In *Proceedings of the SIGIR'12*. ACM, 1101–1102.
- A. Chang and C. Manning. 2012. SUTIME: A library for recognizing and normalizing time expressions. In *Proceedings of the LREC'12*.
- S. Cheng, A. Arvanitis, and V. Hristidis. 2013. How fresh do you want your search results? In *Proceedings of the CIKM'13*. ACM, 1271–1280.
- J. Cho and H. Garcia-Molina. 2003. Estimating frequency of change. *ACM Transactions on Internet Technology* 3, 3, 256–290.
- J. Cho, S. Roy, and R. Adams. 2005. Page Quality: In search of an unbiased web ranking. In *Proceedings of the SIGMOD'05*. ACM, 551–562.
- F. Costa. *Processing Temporal Information in Unstructured Documents*. PhD thesis, Universidade de Lisboa.
- J. Cook, A. Das Sarma, A. Fabrikant, and A. Tomkins. 2012. Your two weeks of fame and your grandmother's. In *Proceedings of the WWW'12*. ACM, 919–928.
- S. Cousins and M. Kahn. 1991. The visual display of temporal information. *Artificial Intelligence in Medicine* 3, 6, 341–357.

- H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of ACL02*. Association for Computational Linguistics, 168–175.
- N. Dai and B. Davison. 2010. Freshness matters: in flowers, food, and web authority. In *Proceedings of the SIGIR'10*. ACM, 114–121.
- N. Dai, M. Shokouhi, and B. D. Davison. 2011. Learning to rank for freshness and relevance. In *Proceedings of the SIGIR'11*. ACM, 95–104.
- W. Dakka, L. Gravano, and P. G. Ipeirotis. 2012. Answering general time sensitive queries. *IEEE Transactions on Knowledge and Data Engineering*, 24(2): IEEE Computer Society Press. 220–235.
- G. Dias, R. Campos, and A. Jorge. 2011. Future retrieval: what does the future talk about? In *Proceedings of the ENIR'11 Workshop Associated to SIGIR'11*.
- G. Dias, J. G. Moreno, A. Jatowt, and R. Campos. 2012. Temporal web image retrieval. In *Proceedings of the SPIRE'12*. Springer-Verlag, 199–204.
- F. Diaz. 2009. Integration of news content into web results. In *Proceedings of the WSDM'09*. ACM, 182–191.
- L. Döhling and U. Leser. 2014. Extracting and aggregating temporal events from text. In *Proceedings of the WWW'14*. ACM, 839–844.
- A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. 2010a. Towards recency ranking in web search. In *Proceedings of the WSDM'10*. ACM, 11–20.
- A. Dong, R. Zhang, P. Kolari, B. Jing, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. 2010b. Time is of the essence: Improving recency ranking using twitter data. In *Proceedings of the WWW'10*. ACM, 331–340.
- M. Efron and G. Golovchinsky. 2011. Estimation methods for ranking recent information. In *Proceedings of the SIGIR'11*. ACM, 495–504.
- M. Efron. 2012. Query-specific recency ranking: Survival analysis for improved microblog retrieval. In *Proceedings of the TALA'12 Workshop associated to SIGIR'12*.
- J. L. Elsas and S. T. Dumais. 2010. Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of the WSDM'10*. ACM, 1–10.
- D. Fetterly, M. Manasse, M. Najork, and J. Wiener. 2003. A large-scale study of the evolution of web pages. In *Proceedings of the WWW'03*. ACM, 669–678.
- A. Garcia-Fernandez, A.-L. Ligozat, M. Dinarelli, and D. Bernhard. 2011. When was it written? automatically determining publication dates. In *Proceedings of the SPIRE'11*. Springer-Verlag, 221–236.
- D. Gomes, J. Miranda, and M. Costa. 2011. A survey on web archiving initiatives. In *Proceedings of the TPDF'11*. Springer-Verlag, 408–420.
- D. Gomes, M. Costa, D. Cruz, J. Miranda, and S. Fontes. 2013. Creating a billion-scale searchable web archive. In *Proceedings of the TempWeb'13 Workshop associated to WWW'13*. 1059–1066.
- D. Graff, J. Kong, K. Chen, and K. Maeda. 2003. *English Gigaword*. Linguistic Data Consortium, Philadelphia.
- M. Halbwachs. 1992. *On Collective Memory*. University of Chicago Press.
- J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194, 28–61.
- L. Hoffmann. 2013. Looking back at big data. *Communications of the ACM* 56, 4, 21–23.
- H. Holzmann, G. Gossen, and N. Tahmasebi. 2012. Formerly known as—a search engine incorporating named entity evolution. In *Proceedings of the Coling'12*. Association for Computational Linguistics, 215–222.
- T. Huet, J. Biega, and F. M. Suchanek. 2013. Mining history with le monde. In *Proceedings of the AKBC'13 Workshop Associated to CIKM'13*. ACM, 49–53.
- Y. Inagaki, N. Sadagopan, G. Dupret, A. Dong, C. Liao, Y. Chang, and Z. Zheng. 2010. Session based click features for recency ranking. In *Proceedings of the AAAI'10*. AAAI Press, 1334–1339.
- A. Jatowt, Y. Kawai, and K. Tanaka. 2007. Detecting age of page content. In *Proceedings of the WIDM'07 Workshop associated to CIKM'07*. ACM, 137–144.
- A. Jatowt, H. Kawai, K. Kanazawa, K. Tanaka, and K. Kunieda. 2009. Supporting analysis of future-related information in news archives and the web. In *Proceedings of the JCDL'09*. ACM, 115–124.
- A. Jatowt, H. Kawai, K. Kanazawa, K. Tanaka, and K. Kunieda. 2010. Analyzing collective view of future, time-referenced events on the web. In *Proceedings of the WWW'10*. ACM, 1123–1124.
- A. Jatowt, Y. Kawai, and K. Tanaka. 2011. Calculating content recency based on timestamped and non-timestamped sources for supporting page quality estimation. In *Proceedings of the SAC'11*. ACM.
- A. Jatowt and C. M. Yeung. 2011. Extracting collective expectations about the future from large text collections. In *Proceedings of the CIKM'11*. ACM, 1259–1264.

- A. Jatowt, C. M. Yeung, and K. Tanaka. 2013a. Estimating document focus time. In *Proceedings of the CIKM'13*. ACM, 2273–2278.
- A. Jatowt, H. Kawai, K. Kanazawa, K. Tanaka, K. Kunieda, and K. Yamada. 2013b. Multi-lingual, Longitudinal analysis of future-related information on the web. In *Proceedings of the ICCW'13*. IEEE Computer Society Press, 27–32.
- P. Jin, J. Lian, X. Zhao, and S. Wan. 2008. TISE: A temporal search engine for web contents. In *Proceedings of the ISIITA'08*. IEEE Computer Society Press, 220–224.
- R. Jones and F. Diaz. 2007. Temporal profiles of queries. *ACM Transactions on Information Systems* 25, 3, Article 14.
- H. Joho, A. Jatowt, and R. Blanco. 2013. A survey of temporal web search experience. In *Proceedings of the TempWeb'13 Workshop Associated to WWW'13*. 1101–1108.
- H. Joho, A. Jatowt, and R. Blanco. 2014. NTCIR temporalia: A test collection for temporal information access research. In *Proceedings of the WWW'14*. ACM, 845–850.
- F. Jong, H. Rode, and D. Hiemstra. 2005. Temporal language models for the disclosure of historical text. In *Proceedings of the AHC'05*. 161–168.
- B. Kahle. 1997. Preserving the internet. *Scientific American Magazine* 276, 3, 72–73.
- K. Kanazawa, A. Jatowt, and K. Tanaka. 2011. Improving retrieval of future-related information in text collections. In *Proceedings of the WI-IAT'11*. IEEE Computer Society, 278–283.
- N. Kanhabua and K. Nørvåg. 2008. Improving temporal language models for determining time of non-timestamped documents. In *Proceedings of the ECDL'08*. Springer-Verlag, 358–370.
- N. Kanhabua and K. Nørvåg. 2010. Determining time of queries for re-ranking search results. Determining time of queries for re-ranking search results. In *Proceedings of ECDL'10*. 261–272.
- N. Kanhabua, R. Blanco, and M. Matthews. 2011. Ranking related news predictions. In *Proceedings of the SIGIR'11*. ACM, 755–764.
- N. Kanhabua and K. Nørvåg. 2012. Learning to rank search results for time-sensitive queries. In *Proceedings of the CIKM'12*. ACM, 2463–2466.
- N. Kanhabua, C. Niederee, and W. Siberski. 2013. Towards concise preservation by managed forgetting: Research issues and case study. In *Proceedings of the iPRES'13*.
- M. Karam. 1994. Visualization using timelines. In *Proceedings of the ISSTA'94 associated to SIGSOFT*. ACM, 125–137.
- H. Kawai, A. Jatowt, K. Tanaka, K. Kunieda, and K. Yamada. 2010. ChronoSeeker: Search engine for future and past events. In *Proceedings of the ICUIMC'10*. ACM, 166–175.
- G. Kim and E. P. Xing. 2013. Time-sensitive web image ranking and retrieval via dynamic multi-task regression. In *Proceedings of the WSDM'13*. ACM, 163–172.
- M. Klein and M. L. Nelson. 2010. Evaluating methods to rediscover missing web pages from the web infrastructure. In *Proceedings of the JCDL'10*. ACM, 59–68.
- A. C. König, M. Gamon, and Q. Wu. 2009. Click-through prediction for news queries. In *Proceedings of the SIGIR'09*. ACM, 347–354.
- A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais. 2011. Understanding temporal query dynamics. In *Proceedings of the WSDM'11*. ACM, 167–176.
- G. Kumaran and J. Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the SIGIR'04*. ACM, 297–304.
- E. Kuzei and G. Weikum. 2012. Extraction of temporal facts and events from Wikipedia. In *Proceedings of the TempWeb'12 Workshop Associated to WWW'12*. ACM, 25–32.
- B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. 2010. SparkClouds visualizing trends in tag clouds. In *IEEE Transactions on Visualization and Computer Graphics* 16, 6, 1182–1189.
- X. Li and W. B. Croft. 2003. Time-based language models. In *Proceedings of the CIKM'03*. ACM, 469–475.
- H. Li, G. Xu, W. B. Croft, M. Bendersky, Z. Wang, and E. Viegas. 2012. A public dataset for promoting query representation and understanding research. In *Proceedings of the WSCD'12*.
- X. Li, B. Liu, and P. Yu. 2008. Time sensitive ranking with application to publication search. In *Proceedings of the ICDM'08*. IEEE Computer Society Press, 893–898.
- J. Makkonen and H. Ahonen-myka. 2003. Utilizing temporal information in topic detection and tracking. In *Proceedings of the ECDL'03*. Springer-Verlag, 393–404.
- E. Manica, C. F. Dorneles, and R. Galante. 2012. Handling temporal information in web search engines. In *Proceedings of the SIGMOD Record* 41, 3, 15–23.
- I. Mani and G. Wilson. 2000. Robust temporal processing of news. In *Proceedings of the ACL'00*. Association for Computational Linguistics, 69–76.

- J. Masanès. 2006. *Web Archiving*. Springer-Verlag.
- M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, and H. Zaragoza. 2010. Searching through time in the *New York Times*. In *Proceedings of the HCIR'10 Workshop*. 41–44.
- P. Mazur. 2012. *Broad-Coverage Rule-Based Processing of Temporal Expressions*. PhD thesis, Australia Macquarie University.
- F. McCown and M. L. Nelson. 2008. Recovering a website's server components from the web infrastructure. In *Proceedings of the JCDL'08*. ACM, 124–133.
- M. J. Metzger. 2007. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology* 58, 13, 2078–2091.
- D. Metzler, R. Jones, F. Peng, and R. Zhang. 2009. Improving search relevance for implicitly temporal queries. In *Proceedings of the SIGIR'09*. ACM, 700–701.
- J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331, 6014, 176–182.
- Google Books. 2014. Ngram Viewer. Available at: <https://books.google.com/ngrams>.
- M. A. Nascimento and M. H. Dunham. 1999. Indexing valid time via B+-trees—the MAP21 approach. *IEEE Transactions on Knowledge and Engineering* 11, 6, 929–947.
- A. Ntoulas, J. Cho, and C. Olston. 2004. What's new on the web?: The evolution of the web from a search engine perspective. In *Proceedings of the WWW'04*. ACM, 1–12.
- S. Nunes, C. Ribeiro, and G. David. 2007. Using neighbors to date web documents. In *Proceedings of the WIDM'07 Workshop Associated to CIKM'07*. ACM, 129–136.
- S. Nunes, C. Ribeiro, and G. David. 2008. Use of temporal expressions in web search. In *Proceedings of the ECIR'08*. Springer-Verlag. 580–584.
- D. Odijk, G. Santucci, M. de Rijke, M. Angelini, and G. Granato. 2012. Exploring word meaning through time. In *Proceedings of the TALA'12 Workshop Associated to SIGIR'12*.
- M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. 2012. Bieber no more: First story detection using Twitter and Wikipedia. In *Proceedings of the TAIA'12 Workshop Associated to SIGIR'12*.
- F. Palermo, J. Hays, and A. A. Efros. 2012. Dating historical color images. In *Proceedings of the ECCV'12*. Springer-Verlag, 499–512.
- M. Pasca. 2008. Towards Temporal Web Search. In *Proceedings of the ACM Symposium on Applied Computing*. ACM, 1117–1121.
- J. Perkiö, W. Buntine, and H. Tirri. 2005. A temporally adaptative content-based relevance ranking algorithm. In *Proceedings of the SIGIR'05*. ACM, 647–648.
- C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. 1996. LifeLines: Visualizing personal histories. In *Proceedings of the SIGCHI'96*. ACM, 221–227.
- J. M. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the SIGIR'98*. ACM, 275–281.
- J. Pustejovsky, J. Castaño, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, and G. Katz. 2003. TimeML: Robust specification of event and temporal expression in text. In *Proceedings of the IWCS'03*. 28–34.
- J. Pustejovsky, M. Verhagen, R. Sauri, J. Littman, R. J. Gaizauskas, G. Katz, I. Mani, R. Knippen, and A. Setzer. 2006. TimeBank 1.2. In *Proceedings of the Linguistic Data Consortium*.
- K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the WWW'11*. ACM 337–346.
- K. Radinsky and E. Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the WSDM'13*. ACM, 255–264.
- K. Radinsky, F. Diaz, S. Dumais, M. Shokouhi, A. Dong, and Y. Chang. 2013. Temporal web dynamics and its application to information retrieval. In *Proceedings of the WSDM'13*. Rome, Italy: ACM, 781–782.
- R. Reinanda, D. Odijk, and M. de Rijke. 2012. Exploring entity associations over time. In *Proceedings of the TAIA'12 Workshop Associated to SIGIR'12*.
- H. SalahEldeen and M. L. Nelson. 2013a. Carbon dating the web: Estimating the age of web resources. In *Proceedings of the WWW'13*. ACM, 1075–1082.
- H. SalahEldeen and M. L. Nelson. 2013b. Reading the correct history?: Modeling temporal intention in resource sharing. In *Proceedings of the JCDL'13*. ACM, 257–266.
- T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the WWW'10*. ACM. 851–860.

- E. Sandhaus. 2008. *The New York Times Annotated Corpus*. Linguistic Data Consortium. Philadelphia.
- F. Schilder and C. Habel. 2005. From temporal expressions to temporal information: Semantic tagging of news messages. In I. Mani, J. Pustejovsky, and R. Gaizauskas (eds.), *The Language of Time: A Reader*. Oxford University Press, 533–544.
- A. Setzer and R. J. Gaizauskas. 2000. Annotating events and temporal information in newswire texts. In *Proceedings of the LREC'00*.
- B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims. 2005. Identifying temporal patterns and key players in document collections. In *Proceedings of the TDM'05 Workshop Associated to ICDM'05*. IEEE Computer Society Press, 165–174.
- M. Shokouhi. 2011. Detecting seasonal queries by time-series analysis. In *Proceedings of the SIGIR'11*. ACM, 1171–1172.
- M. Shokouhi and K. Radinsky. 2012. Time-sensitive query auto-completion. In *Proceedings of the SIGIR'12*. ACM, 601–610.
- R. Snodgrass and I. Ahn. 1985. A taxonomy of time databases. In *Proceedings of the SIGMOD'85*. ACM, 236–246.
- S. Song and J. JaJa. 2008. *Archiving Temporal Web Information: Organization of Web Contents for Fast Access and Compact Storage*. Technical Report UMIACS-TR-2008-08, University of Maryland Institute for Advanced Computer Studies, Maryland, MD.
- J. Strötgen and M. Gertz. 2010a. HeideTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the IWSE'10 Associated to ACL'10*. 321–324.
- J. Strötgen and M. Gertz. 2010b. TimeTrails: A system for exploring spatio-temporal information in documents. In *Proceedings of the VLDB Endowment: VLDB Endowment*. 3, 1–2, 1569–1572.
- J. Strötgen and M. Gertz. 2012. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*. 1–30.
- J. Strötgen and M. Gertz. Proximity²-aware ranking for textual, temporal, and geographic queries. In *Proceedings of CIKM'13*. ACM, 739–744.
- J. Strötgen, O. Alonso, and M. Gertz. 2012. Identification of top relevant temporal expressions in documents. In *Proceedings of the TempWeb'12 Workshop associated to WWW'12*. ACM, 33–40.
- F. M. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the WWW'07*. ACM, 697–706.
- K. M. Svore, J. Teevan, S. T. Dumais, and A. Kulkarni. 2012. Creating temporally dynamic web search snippets. In *Proceedings of the SIGIR'12*. ACM, 1045–1046.
- R. Swan and J. Allan. 1999. Extracting significant time-varying features from text. In *Proceedings of the CIKM'99*. ACM, 38–45.
- R. Swan and D. Jensen. 2000. TimeMines: Constructing timelines with statistical models of word usage. In *Proceedings of the TM'00 Workshop Associated to KDD'00*. ACM, 73–80.
- A. Styskin, F. Romanenko, F. Vorobyev, and P. Serdyukov. 2011. Recency ranking by diversification of result set. In *Proceedings of the CIKM'11*. ACM, 1949–1952.
- N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmänn, and T. Risse. 2012. NEER: An unsupervised method for named entity evolution recognition. In *Proceedings of the Coling'12*. Association for Computational Linguistics, 2553–2568.
- H. Takemura and K. Tajima. 2012. Tweet classification based on their lifetime duration. In *Proceedings of the CIKM'12*. ACM, 2367–2370.
- P. P. Talukdar, D. Wijaya, and T. Mitchell. 2012. Coupled temporal scoping of relational facts. In *Proceedings of the WSDM'12*. ACM, 73–82.
- M. Toyoda and M. Kitsuregawa. 2006. What's really new on the web? Identifying new pages from a series of unstable web snapshots. In *Proceedings of the WWW'06*. ACM, 233–241.
- J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. 2011. The anatomy of the Facebook social graph. *Computing Research Repository Journal*. CoRR abs/1111.4503.
- N. UzZaman, H. Llorens, L. Derczynski, M. Verhagen, J. Allen, and J. Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating events, time expressions, and temporal relations. In *Proceedings of the SemVal'13 Workshop*. ACM, 1–9.
- H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. 2009. Memento: Time travel for the web. CoRR abs/0911.1112.
- P.-Y. Vandenbussche and C. Teissèdre. 2011. Events retrieval using enhanced semantic web knowledge. In *Proceedings of the DeRiVe'11 Workshop Associated to ISWC 2011*.

- M. T. Vicente-Diez and P. Martinez. 2009. Temporal semantics extraction for improving web search. In *Proceedings of the DEXA'09*. IEEE, 69–73.
- M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. 2004. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the ICMD'04*. ACM, 131–142.
- Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum. 2010. Timely YAGO: Harvesting, querying, and visualizing temporal knowledge from Wikipedia. In *Proceedings of the EDBT'10*. ACM, 697–700.
- Y. Wang, B. Yang, L. Qu, M. Spaniol, and G. Weikum. 2011. Harvesting facts from textual web sources by constrained label propagation. In *Proceedings of the CIKM'11*. ACM, 837–846.
- Y. Wang, D. Maximilian, M. Spaniol, and G. Weikum. 2012. Coupling label propagation and constraints for temporal fact extraction. In *Proceedings of the ACL'12*. Association for Computational Linguistics, 233–237.
- W. Weerkamp and M. de Rijke. 2012. Activity prediction: A Twitter-based Exploration. In *Proceedings of the TAIA'12 Workshop Associated to SIGIR'12*.
- S. Whiting and O. Alonso. 2012. Hashtags as milestones in time. In *Proceedings of the TAIA'12 Workshop Associated to SIGIR'12*.
- S. Whiting, J. Jose, and O. Alonso. 2014. Wikipedia as a time machine. In *Proceedings of the WWW'14*. ACM, 857–861.
- Yahoo. 2014. Yahoo!'s Time Explorer. Available at: <http://fbmya01.barcelonamedia.org:8080/future/>.
- Y. Yamamoto, T. Tezuka, A. Jatowt, and K. Tanaka. 2007. Honto? Search: Estimating trustworthiness of web information by search results aggregation and temporal analysis. In *Proceedings of the Joint 9th Asia-Pacific Web and 8th International Conference on Web-age Information Management Conference on Advances in Data and Web Management*. 253–264.
- R. Zhang, Y. Chang, Z. Zheng, D. Metzler, and J.-Y. Nie. 2009. Search result re-ranking by feedback control adjustment for time-sensitive query. In *Proceedings of the NAACL-HLT'09*. Association for Computational Linguistics, 165–168.
- R. Zhang, Y. Konda, A. Dong, P. Kolari, Y. Chang, and Z. Zheng. 2010. Learning recurrent event queries for web search. In *Proceedings of the EMNLP'10*. Association for Computational Linguistics, 1129–1139.
- J. Zobel and A. Moffat. 2006. Inverted files for text search engines. *ACM Computing Surveys* 38, 2, 1–56.

Received May 2013; revised September 2013; accepted May 2014