



This is a repository copy of *Time and information retrieval: Introduction to the special issue*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/92765/>

Version: Accepted Version

Article:

Derczynski, L., Strötgen, J., Campos, R. et al. (1 more author) (2015) Time and information retrieval: Introduction to the special issue. *Information Processing and Management*, 51 (6). 786 -790. ISSN 0306-4573

<https://doi.org/10.1016/j.ipm.2015.05.002>

Article available under the terms of the CC-BY-NC-ND licence
(<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

Reuse

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>



Time and Information Retrieval: Introduction to the Special Issue

Leon Derczynski

University of Sheffield

Jannik Strötgen

Heidelberg University

Ricardo Campos

Polytechnic Institute of Tomar / LIAAD-INESC TEC

Omar Alonso

Microsoft Corporation

Abstract

Time intrinsically structures the information around us, and has deep and constant impacts on our information needs. Thus, time is a very important dimension of the Information Retrieval (IR) process, and, as a result, it has been used in many creative ways to improve the way documents and queries are handled to satisfy human information needs. Since researchers became drawn to its importance, temporal information retrieval techniques have seen a striking level of attention in both academia and industry, leading to the emergence of an increasing array of research and product innovation. In this introduction, we briefly summarise the area and outline the contributions included in this special issue.

© 2015 Published by Elsevier Ltd.

Keywords: Information Retrieval, Time

1. Introduction

With the rapid growth of digitised document resources, both on and off the web, and increased variety in types of document collections, future information systems will face growing difficulties in providing reliable, useful, and timely results. Time is a ubiquitous factor at many stages in the information-seeking process, with users having temporally-relevant information needs, and collections having temporal properties at collection, document metadata, and document content levels. This issue aims to explore opportunities and novel research on the intersection of time and information retrieval.

Existing work in temporal information retrieval has aimed to account for time in document collections and relevance measurements. In contrast to classical information retrieval, temporal information retrieval aims to improve user experience by augmenting document relevance with temporal relevance.

Email address: leon@dcs.shef.ac.uk (Leon Derczynski)

11 Over the last few years, temporality has been gained an increasing importance within the field of information
12 retrieval. Research on time and information retrieval covers a large number of topics [1, 2]. These include: the extrac-
13 tion of temporal expressions and events; temporal representation of documents, of collections and of queries; temporal
14 retrieval models; temporal and event-based summarisation; temporal text similarity; temporal query understanding;
15 clustering of search results by time; temporality in ranking; visualisation and design of temporal search interfaces;
16 and so on.

17 This intense research is just in time to meet increasing demand for more intelligent processing of growing amounts
18 of data. For example, social media data represents a sampling of all human discourse, and is temporally annotated
19 with a document creation date. The historical (i.e., longitudinal) and emerging aspects of social media data are as
20 yet relatively untapped [3], and although they present challenging indexing and retrieval problems, they can support
21 powerful search and analysis applications. For example, combining time series analysis on social media messages
22 with effective processing of emerging data can predict voting intent [4] or outbreaks of West Nile virus [5].

23 The recognition of temporal information need and presentation of temporal information are very challenging
24 problems. Expressions of time in documents are typically underspecified and vague (c.f. “*I’ll see you later*” - when?
25 or, “*As I was brushing my teeth.*” - one needs to know how often teeth are brushed to guess when this was). Indeed,
26 as humans experience time in the same way, temporality is not often expressed, instead remaining implicit. Further,
27 understanding how to present data such that change, duration, order and other temporal aspects are clear is an area in
28 which progress beyond the embryonic is just starting. This difficulty - of conveying temporality between system and
29 user - places a demand on builders of information systems to account for and model our understanding of time.

30 We argue that the interaction between time and information retrieval is broader than simply adding temporal
31 constraints to retrieval. Indeed, this is only the first step. To build the information systems of the future, one must
32 understand more about both the part that time plays in human information need and expectations, and also what is
33 already expressed by time in existing data collections. Doing this serves to provide better information access and
34 powerful analyses of untapped resources.

35 2. A brief survey of temporal information retrieval research

36 The need for integrating temporal information was quickly recognised after the emergence of information retrieval
37 systems at scale [6]. One of the first initiatives in temporally-aware information systems is the Internet Archive
38 project [7], which aimed to build a digital library of Web-sites. This successful longitudinal system inspired work
39 on other ways to access information which including the temporal dimension, especially for exploration and search
40 purposes.

41 Later, research that integrated temporality into retrieval rankings become mature [8]. Today, major search engines
42 have experimented bringing control of temporal search to the everyday user, with basic temporal refinement in their
43 web search engine enabling filtering of results according to the publication time of the document.

44 Concurrently, standardisation of the temporal semantics within documents developed, and formal definitions for
45 “temporal expression” and “event” were prototyped. In the case of TimeML, a temporal expression is a sequence of
46 words that represents a particular time or period of time, and an event is a single-word reference to an eventuality, be
47 it a change, an action, a state and so on. This proposal later developed into the now-widely adopted ISO standard [9].
48 Nowadays, mature temporal taggers are available for multiple languages [10].

49 Later work has identified and approached different sub-areas of temporal information retrieval. The work of Baeza-
50 Yates [11] defined foundations of temporal information retrieval. This was expanded into several parallel topics, such
51 as understanding user queries [12], generating summary snippets accounting for temporality [13], including time in
52 result order [14], temporal clustering [15], future retrieval [16] and temporal image retrieval [17]. These lay the
53 foundations for powerful analysis applications, with both general advances applicable across many areas and also
54 tools and knowledge specific to certain domains.

55 Development in temporal information systems is driven by a constant flow of challenges and exercises. The well-
56 known KBP (Knowledge Base Population) challenge has run a successful and popular temporal bounding task [18],
57 where assertions found in text (e.g. “Benjamin Harrison - is_president_of - USA”) are constrained to specific start and
58 end dates - a difficult problem, but an important one; almost everything we know has finite start and end points [19, 20].
59 Finally, in 2015 we saw not one but four different temporal shared challenges at SemEval [21, 22, 23, 24].

60 Most recently, focus has turned to our interactions with temporality, including our behaviour and how to present
61 information that has temporal parts. Graphical representations of temporal information are hard to create, confused by
62 imperfect metaphors and underspecification [25, 26]. In terms of visual information access, Google NGram Viewer
63 has been released as basic tool for mining the rise and fall words used in five million books over selected years. MIT
64 has developed SIMILE Timeline Visualisation, a Web widget prototype for visualising temporal data. Visualisation
65 remains a big challenge to temporal information systems.

66 Organising, searching over and mining past information in terms of events has proven a difficult and interesting
67 challenge, and making headway is yielding interesting results [27, 28]. Commercial products have focused not only
68 on historical search, but also search over future information, such as Recorded Future and Yahoo!’s Time Explorer
69 application [29]; this promising direction is fueled by current research, such as Radinsky and Horvitz’s system - a
70 system that found, for example, that floods which occurred about a year after a drought in the same area often led to
71 cholera outbreaks [16].

72 Demanding as these challenges are, advances in being temporally aware while presenting, mining and analysing
73 data have led to extremely powerful results.

74 **3. Research in this issue**

75 This special issue includes the following research papers on the intersection between time and information re-
76 trieval.

77 In “Evaluating Document Filtering Systems over Time”, Tom Kenter and Krisztian Balog propose a time-aware
78 way of measuring a system’s performance at filtering documents [30]. This is designed to complement traditional
79 metrics. Their assumption is that current metrics do not capture all the relevant aspects of the systems being evalu-
80 ated, particularly those from the temporal dimension. The main idea of this work is to estimate a trendline by dividing
81 the timeline into subsets, performing overall evaluation in each subset, and project performance at the end of the
82 evaluation period based on the trendline. They show that traditional macro-averaged true-positive-based metrics, like
83 precision, recall and F-measure fail to capture essential information when applied in a batch setting. To overcome this,
84 a new metric, aptness, is presented, and we see how this is readily incorporated into F-measure. Finally, extrinsic ex-
85 perimental results are presented in a real-world setting, where the ability of aptness to represent temporal performance
86 is demonstrated.

87 Manika Kar, Sérgio Nunes and Cristina Ribeiro present interesting methods for summarising changes in dynamic
88 text collections over time in their paper “Summarization of Changes in Dynamic Text Collection using Latent Dirichlet
89 Allocation Model” [31]. The goal here is to obtain a summary of the most significant changes made to a document
90 during a specific period of time. Various extractive summarisation approaches are proposed. First, individual terms are
91 scored. Then, this information is used to rank and select sentences in order to produce a final summary. Evaluation
92 over a set of Wikipedia articles shows that a method based on Latent Dirichlet Allocation achieved strong above-
93 baseline performance.

94 In contrast to the two previous articles, Hideo Joho, Adam Jatowt and Roi Blanco report on the temporal in-
95 formation searching behaviour of users and their strategies for dealing with searches that have a temporal nature in
96 “Temporal Information Searching Behaviour and Strategies”, a user study [32]. In controlled settings, thirty partic-
97 ipants are asked to perform searches on an array of topics on the web to find information related to particular time
98 scopes. A large number of valuable observations that have considerable implications for the future design of temporal
99 search mechanisms and search interfaces is presented. Of particular interest is that participants expressed difficulty in
100 finding past and future-related information, in contrast to conducting recency-related search.

101 Finally, the last two works investigate techniques to detect content time within documents. Adam Jatowt, Ching-
102 man Au Yeung and Katsumi Tanaka present a “Generic Method for Detecting Content Time of Documents” [33]. The
103 authors propose several methods for estimating the focus time of documents, i.e. the time a document’s content refers
104 to. They take a three-step statistical approach: (1) determine the strengths of word-time associations by exploiting
105 external document sources; (2) estimate the temporal weights of words; and (3) calculate the text focus time. They
106 evaluate the approach over three different test collections. Interestingly, unlike many prior attempts, this method does
107 not require temporal expressions. So, focus time can still be estimated if a document lacks explicit dates - a major
108 advantage.

109 In contrast to determining time per document, Xujian Zhao, Peiquan Jin and Lihua Yue present an approach to
 110 determining the time of the underlying topic or event in their article entitled “Discovering Topic Time from Web
 111 News” [34]. They propose an approach consisting of temporal expression normalisation and topic time extraction.
 112 For normalisation, a new approach to determine the referential time for implicit temporal expressions and an
 113 algorithm to resolve vague temporal expressions are presented. Topic time is extracted by modelling the dependency
 114 between news topics and temporal information. Two models are proposed, one dependent upon position, the other
 115 upon topic. These are evaluated over two news datasets, with good results.

116 4. Conclusion

117 Temporal information retrieval is an exciting area which offers the research and the industrial communities several
 118 challenging opportunities which remain unsolved. By organising this special issue we wanted to capture a diverse
 119 range of problems and potential solutions on the intersection of temporality and IR. Unlike existing work that focuses
 120 exclusively on the interesting problems related to adding time to established methods of information retrieval (such
 121 as, e.g., how to incorporate temporal relevance in ranking of retrieved results), we sought to encourage discussion on
 122 new or powerful uses of temporality in all kinds of information systems.

123 Our call stimulated the submission of twenty manuscripts, with topics ranging over Document Representation and
 124 Content Analysis, Queries and Query Analysis, Retrieval Models and Ranking, Users and Interactive IR, Filtering
 125 and Recommending, Search Engine Architectures, and Evaluation. This issue forms a valuable source of material that
 126 communicates new research regarding the clear impact that temporality has on building information systems. This
 127 gives us a diverse and interesting snapshot of the field, which promises to be exciting to readers and valuable to the
 128 research community.

129 5. Thanks

130 The editors of this special issue thank the authors for their valuable contributions. In addition, we are deeply
 131 grateful to the international review board, without whose sustained and thorough efforts and timely responses this
 132 issue would not have been possible:

- 133 • Nikos Aletras (University of Sheffield / University College London)
- 134 • Ayser Armiti (Heidelberg University)
- 135 • Krisztian Balog (University of Stavanger)
- 136 • Klaus Berberich (Max Planck Institute for Informatics)
- 137 • Roi Blanco (Yahoo! Labs)
- 138 • António Branco (University of Lisbon)
- 139 • Matteo Brucato (University of Massachusetts Amherst)
- 140 • Michael Gertz (University of Heidelberg)
- 141 • Daniel Gomes (FCCN Lisboa)
- 142 • Min-Yen Kan (National University of Singapore)
- 143 • Nattiya Kanhabua (L3S Research Center)
- 144 • Oleksandr Kolomiyets (KU Leuven)
- 145 • Adam Jatowt (Kyoto University)
- 146 • Alípio Jorge (University of Porto / LIAAD - INESC TEC)
- 147 • Qi Li (Rensselaer Polytechnic Institute)
- 148 • Hector Llorens (Nuance)
- 149 • Sérgio Nunes (University of Porto / ICGS - INESC TEC)
- 150 • Kjetil Nørsvåg (Norwegian University of Science and Technology)
- 151 • Simone Ponzetto (Mannheim University)
- 152 • Ian Ruthven (University of Strathclyde)
- 153 • Ismail Sengor Altingovde (Middle East Technical University)
- 154 • Ian Soboroff (National Institute of Standards and Technology)
- 155 • Partha Pratim Talukdar (Carnegie Mellon University)

- 156 • Hristo Tanev (JRC)
- 157 • Jaime Teevan (Microsoft Corporation)
- 158 • Christoph Trattner (Graz University of Technology)
- 159 • Bin Yang (Aalborg University)

160 6. Acknowledgements

161 Leon Derczynski received funding from the European Union’s Seventh Framework Programme, through grant
 162 No. 611233, PHEME. Ricardo Campos was partially funded by Project “NORTE-07-0124-FEDER-000059” which
 163 is financed by the North Portugal Regional Operational Programme (ON.2 - O Novo Norte), under the National
 164 Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national
 165 funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT) within project
 166 UID/EEA/50014/2013. He was also financed by the Center of Mathematics, University of Beira Interior, within
 167 project “PEst-OE/MAT/UI0212/2014”.

168 Finally, the editors thank their institutions: the University of Sheffield, UK; Heidelberg University, Germany; the
 169 Polytechnic Institute of Tomar, Portugal; the LIAAD / INESC TEC - INESC Technology and Science, Portugal; and
 170 the Microsoft Corporation.

171 7. References

172 References

- 173 [1] O. Alonso, J. Strötgen, R. Baeza-Yates, M. Gertz, Temporal Information Retrieval: Challenges and Opportunities, in: Proceedings of the 1st
 174 International Temporal Web Analytics Workshop (TAWAW 2011), 2011, pp. 1–8.
- 175 [2] R. Campos, G. Dias, A. M. Jorge, A. Jatowt, Survey of temporal information retrieval and related applications, *ACM Computing Surveys*
 176 (CSUR) 47 (2) (2014) 15.
- 177 [3] L. R. Derczynski, B. Yang, C. S. Jensen, Towards context-aware search and analysis on social media data, in: Proceedings of the 16th
 178 international conference on Extending Database Technology (EDBT 2013), ACM, 2013, pp. 137–142.
- 179 [4] V. Lampos, D. Preotiu-Pietro, T. Cohn, A user-centric model of voting intention from social media, in: Proceedings of the Annual Meetings
 180 of the Association for Computational Linguistics (ACL 2013), 2013, pp. 993–1003.
- 181 [5] R. Sugumaran, J. Voss, Real-time spatio-temporal analysis of West Nile Virus using Twitter data, in: Proceedings of the 3rd International
 182 Conference on Computing for Geospatial Research and Applications (COM.Geo ’12), ACM, 2012, p. 39.
- 183 [6] N. J. Belkin, W. B. Croft, Information filtering and information retrieval: two sides of the same coin?, *Communications of the ACM* 35 (12)
 184 (1992) 29–38.
- 185 [7] B. Kahle, Preserving the internet, *Scientific American* 276 (3) (1997) 82–83.
- 186 [8] C. S. Jensen, R. T. Snodgrass, Temporal data management, *Knowledge and Data Engineering, IEEE Transactions on* 11 (1) (1999) 36–44.
- 187 [9] J. Pustejovsky, K. Lee, H. Bunt, L. Romary, ISO-TimeML: An international standard for semantic annotation, in: Proceedings of the interna-
 188 tional Language Resources and Evaluation Conference (LREC 2010), ELRA, 2010, pp. 394–397.
- 189 [10] J. Strötgen, A. Armiti, T. Van Canh, J. Zell, M. Gertz, Time for more languages: Temporal tagging of Arabic, Italian, Spanish, and Vietnamese,
 190 *ACM Transactions on Asian Language Information Processing (TALIP)* 13 (1) (2014) 1–21.
- 191 [11] R. Baeza-Yates, Searching the future, in: Proceedings of the Mathematical/Formal Methods in Information Retrieval Workshop associated to
 192 SIGIR05, 2005.
- 193 [12] R. Campos, G. Dias, A. Jorge, C. Nunes, GTE: a distributional second-order co-occurrence approach to improve the identification of top
 194 relevant dates in web snippets, in: Proceedings of the 21st ACM international Conference on Information and Knowledge Management
 195 (CIKM 2012), ACM, 2012, pp. 2035–2039.
- 196 [13] O. Alonso, R. Baeza-Yates, M. Gertz, Effectiveness of temporal snippets, in: Proceedings of the Workshop on Web Search result Summariza-
 197 tion and Presentation (WSSP) at the World Wide Web Conference, 2009.
- 198 [14] N. Kanhabua, K. Nørvgå, Determining time of queries for re-ranking search results, in: *Research and Advanced Technology for Digital
 199 Libraries*, Vol. 6273 of Lecture Notes in Computer Science, Springer, 2010, pp. 261–272.
- 200 [15] R. Campos, A. M. Jorge, G. Dias, C. Nunes, Disambiguating implicit temporal queries by clustering top relevant dates in web snippets, in:
 201 Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, 2012,
 202 pp. 1–8.
- 203 [16] K. Radinsky, E. Horvitz, Mining the web to predict future events, in: Proceedings of the Sixth ACM International Conference on Web Search
 204 and Data Mining (WSDM), ACM, 2013, pp. 255–264.
- 205 [17] G. Dias, J. G. Moreno, A. Jatowt, R. Campos, Temporal web image retrieval, in: *String Processing and Information Retrieval*, Vol. 7608 of
 206 Lecture Notes in Computer Science, 2012, pp. 199–204.
- 207 [18] H. Ji, R. Grishman, H. T. Dang, Overview of the TAC 2011 Knowledge Base Population track, in: Proceedings of the Text Analysis Confer-
 208 ence (TAC), 2011.
- 209 [19] L. Derczynski, R. Gaizauskas, Information retrieval for temporal bounding, in: Proceedings of the 2013 Conference on the Theory of
 210 Information Retrieval (ICTIR ’13), ACM, 2013, pp. 129–130.

- 211 [20] A. Rula, M. Palmonari, A.-C. N. Ngomo, D. Gerber, J. Lehmann, L. Bühmann, Hybrid acquisition of temporal scopes for RDF data, in: *The*
212 *Semantic Web: Trends and Challenges*, Vol. 8465 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 488–503.
- 213 [21] A.-L. Minard, E. Agirre, I. Aldabe, M. van Erp, B. Magnini, G. Rigau, M. Speranza, R. Urizar, SemEval-2015 Task 4: TimeLine: Cross-
214 Document Event Ordering, in: *Proceedings of the workshop on Semantic Evaluation*, ACL, 2015.
- 215 [22] H. Llorens, N. Chambers, N. UzZaman, N. Mostafazadeh, J. Allen, J. Pustejovsky, SemEval-2015 Task 5: QA TempEval, in: *Proceedings of*
216 *the workshop on Semantic Evaluation*, ACL, 2015.
- 217 [23] S. Bethard, L. Derczynski, J. Pustejovsky, M. Verhagen, SemEval-2015 Task 6: Clinical TempEval, in: *Proceedings of the workshop on*
218 *Semantic Evaluation*, ACL, 2015.
- 219 [24] O. Popescu, C. Strapparava, SemEval-2015 Task 7: Diachronic Text Evaluation, in: *Proceedings of the workshop on Semantic Evaluation*,
220 ACL, 2015.
- 221 [25] C. Plaisant, B. Shneiderman, R. Mushlin, An information architecture to support the visualization of personal histories, *Information Process-*
222 *ing & Management* 34 (5) (1998) 581–597.
- 223 [26] M. Verhagen, Drawing TimeML relations with T-Box, in: *Proceedings of the Dagstuhl Seminar on Annotating, extracting and reasoning*
224 *about time and events*, Vol. 05151 of *Dagstuhl Seminars*, 2005, pp. 7–28.
- 225 [27] P. P. Talukdar, D. Wijaya, T. Mitchell, Coupled temporal scoping of relational facts, in: *Proceedings of the fifth ACM international conference*
226 *on Web Search and Data Mining (WSDM)*, ACM, 2012, pp. 73–82.
- 227 [28] J. Strötgen, M. Gertz, Event-centric search and exploration in document collections, in: *Proceedings of the 12th ACM/IEEE-CS Joint Con-*
228 *ference on Digital Libraries (JCDL'12)*, ACM, 2012, pp. 223–232.
- 229 [29] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, H. Zaragoza, Searching through time in the New York Times, in: *Proceedings of*
230 *the 4th Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2010) in association with IiX*, ACM, 2010, pp. 41–44.
- 231 [30] T. Kenter, K. Balog, Evaluating document filtering systems over time, *Information Processing & Management* 51.
- 232 [31] M. Kar, S. Nunes, C. Ribeiro, Summarization of changes in dynamic text collection using latent Dirichlet allocation model, *Information*
233 *Processing & Management* 51.
- 234 [32] H. Joho, A. Jatowt, R. Blanco, Temporal information searching behaviour and strategies, *Information Processing & Management* 51.
- 235 [33] A. Jatowt, C. man Au Yeung, K. Tanaka, Generic method for detecting content time of documents, *Information Processing & Management*
236 51.
- 237 [34] X. Zhao, P. Jin, L. Yue, Discovering topic time from web news, *Information Processing & Management* 51.