Information Processing and Management 000 (2015) 1–5

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Editorial

Time and information retrieval: Introduction to the special issue

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.

Over the last few years, temporality has been gained an increasing importance within the field of information retrieval. Research on time and information retrieval covers a large number of topics (Alonso, Strötgen, Baeza-Yates, & Gertz, 2011; Campos, Dias, Jorge, & Jatowt, 2014). These include: the extraction of temporal expressions and events; temporal representation of documents, of collections and of queries; temporal retrieval models; temporal and event-based summarisation; temporal text similarity; temporal query understanding; clustering of search results by time; temporality in ranking; visualisation and design of temporal search interfaces; and so on.

This intense research is just in time to meet increasing demand for more intelligent processing of growing amounts of data. For example, social media data represents a sampling of all human discourse, and is temporally annotated with a document creation date. The historical (i.e., longitudinal) and emerging aspects of social media data are as yet relatively untapped (Derczynski, Yang, & Jensen, 2013), and although they present challenging indexing and retrieval problems, they can support powerful search and analysis applications. For example, combining time series analysis on social media messages with effective processing of emerging data can predict voting intent (Lampos, Preotiuc-Pietro, & Cohn, 2013) or outbreaks of West Nile virus (Sugumaran & Voss, 2012).

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

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

2. A brief survey of temporal information retrieval research

The need for integrating temporal information was quickly recognised after the emergence of information retrieval systems at scale (Belkin & Croft, 1992). One of the first initiatives in temporally-aware information systems is the Internet Archive project (Kahle, 1997), which aimed to build a digital library of Web-sites. This successful longitudinal system inspired work on other ways to access information which including the temporal dimension, especially for exploration and search purposes.

http://dx.doi.org/10.1016/j.ipm.2015.05.002

Later, research that integrated temporality into retrieval rankings become mature (Jensen & Snodgrass, 1999). Today, major search engines have experimented bringing control of temporal search to the everyday user, with basic temporal refinement in their web search engine enabling filtering of results according to the publication time of the document.

Concurrently, standardisation of the temporal semantics within documents developed, and formal definitions for "temporal expression" and "event" were prototyped. In the case of TimeML, a temporal expression is a sequence of words that represents a particular time or period of time, and an event is a single-word reference to an eventuality, be it a change, an action, a state and so on. This proposal later developed into the now-widely adopted ISO standard (Pustejovsky, Lee, Bunt, & Romary, 2010). Nowadays, mature temporal taggers are available for multiple languages (Strötgen, Armiti, Van Canh, Zell, & Gertz, 2014).

Later work has identified and approached different sub-areas of temporal information retrieval. The work of Baeza-Yates (2005) defined foundations of temporal information retrieval. This was expanded into several parallel topics, such as understanding user queries (Campos, Dias, Jorge, & Nunes, 2012), generating summary snippets accounting for temporality (Alonso, Baeza-Yates, & Gertz, 2009), including time in result order (Kanhabua & Nørvåg, 2010), temporal clustering (Campos, Jorge, Dias, & Nunes, 2012), future retrieval (Radinsky & Horvitz, 2013) and temporal image retrieval (Dias, Moreno, Jatowt, & Campos, 2012). These lay the foundations for powerful analysis applications, with both general advances applicable across many areas and also tools and knowledge specific to certain domains.

Development in temporal information systems is driven by a constant flow of challenges and exercises. The well-known KBP (Knowledge Base Population) challenge has run a successful and popular temporal bounding task (Ji, Grishman, & Dang, 2011), where assertions found in text (e.g. "Benjamin Harrison – is_president_of – USA") are constrained to specific start and end dates – a difficult problem, but an important one; almost everything we know has finite start and end points (Derczynski & Gaizauskas, 2013; Rula et al., 2014). Finally, in 2015 we saw not one but four different temporal shared challenges at SemEval (Bethard, Derczynski, Pustejovsky, & Verhagen, 2015; Llorens et al., 2015; Minard et al., 2015; Popescu & Strapparava, 2015).

Most recently, focus has turned to our interactions with temporality, including our behaviour and how to present information that has temporal parts. Graphical representations of temporal information are hard to create, confused by imperfect metaphors and underspecification (Plaisant, Shneiderman, & Mushlin, 1998; Verhagen, 2005). In terms of visual information access, Google NGram Viewer has been released as basic tool for mining the rise and fall words used in five million books over selected years. MIT has developed SIMILE Timeline Visualisation, a Web widget prototype for visualising temporal data. Visualisation remains a big challenge to temporal information systems.

Organising, searching over and mining past information in terms of events has proven a difficult and interesting challenge, and making headway is yielding interesting results (Strötgen & Gertz, 2012; Talukdar, Wijaya, & Mitchell, 2012). Commercial products have focused not only on historical search, but also search over future information, such as Recorded Future and Yahoo!'s Time Explorer application (Matthews et al., 2010); this promising direction is fueled by current research, such as Radinsky and Horvitz's system – a system that found, for example, that floods which occurred about a year after a drought in the same area often led to cholera outbreaks (Radinsky & Horvitz, 2013).

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

3. Research in this issue

This special issue includes the following research papers on the intersection between time and information retrieval.

In "Evaluating Document Filtering Systems over Time", Tom Kenter and Krisztian Balog propose a time-aware way of measuring a system's performance at filtering documents (Kenter, Balog, & de Rijke, 2015). This is designed to complement traditional metrics. Their assumption is that current metrics do not capture all the relevant aspects of the systems being evaluated, particularly those from the temporal dimension. The main idea of this work is to estimate a trendline by dividing the timeline into subsets, performing overall evaluation in each subset, and project performance at the end of the evaluation period based on the trendline. They show that traditional macro-averaged true-positive-based metrics, like precision, recall and *F*-measure fail to capture essential information when applied in a batch setting. To overcome this, a new metric, aptness, is presented, and we see how this is readily incorporated into *F*-measure. Finally, extrinsic experimental results are presented in a real-world setting, where the ability of aptness to represent temporal performance is demonstrated.

Manika Kar, Sérgio Nunes and Cristina Ribeiro present interesting methods for summarising changes in dynamic text collections over time in their paper "Summarization of Changes in Dynamic Text Collection using Latent Dirichlet Allocation Model" (Kar, Nunes, & Ribeiro, 2015). The goal here is to obtain a summary of the most significant changes made to a document during a specific period of time. Various extractive summarisation approaches are proposed. First, individual terms are scored. Then, this information is used to rank and select sentences in order to produce a final summary. Evaluation over a set of Wikipedia articles shows that a method based on Latent Dirichlet Allocation achieved strong above-baseline performance.

In contrast to the two previous articles, Hideo Joho, Adam Jatowt and Roi Blanco report on the temporal information searching behaviour of users and their strategies for dealing with searches that have a temporal nature in "Temporal Information Searching Behaviour and Strategies", a user study (Joho, Jatowt, & Blanco, 2015). In controlled settings, thirty participants are asked to perform searches on an array of topics on the web to find information related to particular time scopes. A large number of valuable observations that have considerable implications for the future design of temporal search mechanisms and search interfaces is presented. Of particular interest is that participants expressed difficulty in finding past and future-related information, in contrast to conducting recency-related search.

Finally, the last two works investigate techniques to detect content time within documents. Adam Jatowt, Ching-man Au Yeung and Katsumi Tanaka present a "Generic Method for Detecting Content Time of Documents" (Jatowt, Au Yeung, & Tanaka, 2015). The authors propose several methods for estimating the focus time of documents, i.e. the time a document's content refers to. They take a three-step statistical approach: (1) determine the strengths of word-time associations by exploiting external document sources; (2) estimate the temporal weights of words; and (3) calculate the text focus time. They evaluate the approach over three different test collections. Interestingly, unlike many prior attempts, this method does not require temporal expressions. So, focus time can still be estimated if a document lacks explicit dates – a major advantage.

In contrast to determining time per document, Xujian Zhao, Peiquan Jin and Lihua Yue present an approach to determining the time of the underlying topic or event in their article entitled "Discovering Topic Time from Web News" (Zhao, Jin, & Yue, 2015). They propose an approach consisting of temporal expression normalisation and topic time extraction. For normalisation, a new approach to determine the referential time for implicit temporal expressions and an algorithm to resolve vague temporal expressions are presented. Topic time is extracted by modelling the dependency between news topics and temporal information. Two models are proposed, one dependent upon position, the other upon topic. These are evaluated over two news datasets, with good results.

4. Conclusion

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

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

5. Thanks

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

- Nikos Aletras (University of Sheffield/University College London)
- Ayser Armiti (Heidelberg University)
- Krisztian Balog (University of Stavanger)
- Klaus Berberich (Max Planck Institute for Informatics)
- Roi Blanco (Yahoo! Labs)
- António Branco (University of Lisbon)
- Matteo Brucato (University of Massachusetts Amherst)
- Michael Gertz (University of Heidelberg)
- Daniel Gomes (FCCN Lisboa)
- Min-Yen Kan (National University of Singapore)
- Nattiya Kanhabua (L3S Research Center)
- Oleksandr Kolomiyets (KU Leuven)
- Adam Jatowt (Kyoto University)
- Alípio Jorge (University of Porto/LIAAD – INESC TEC)
- Qi Li (Rensselaer Polytechnic Institute)
- Hector Llorens (Nuance)
- Sérgio Nunes (University of Porto/ICGS – INESC TEC)
- Kjetil Nørvåg (Norwegian University of Science and Technology)
- Simone Ponzetto (Mannheim University)
- Ian Ruthven (University of Strathclyde)
- Ismail Sengor Altingovde (Middle East Technical University)
- Ian Soboroff (National Institute of Standards and Technology)

- Partha Pratim Talukdar (Carnegie Mellon University)
- Hristo Tanev (JRC)
- Jaime Teevan (Microsoft Corporation)
- Christoph Trattner (Graz University of Technology)
- Bin Yang (Aalborg University)

Acknowledgments

Leon Derczynski received funding from the European Union's Seventh Framework Programme, through Grant No. 611233, PHEME. Ricardo Campos was partially 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, Fundação para a Ciência e a Tecnologia (FCT) within project UID/EEA/50014/2013. He was also financed by the Center of Mathematics, University of Beira Interior, within project "PEst-OE/MAT/UI0212/2014".

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

References

Alonso, O., Baeza-Yates, R., & Gertz, M. (2009). Effectiveness of temporal snippets. In *Proceedings of the workshop on web search result summarization and presentation (WSSP) at the world wide web conference*.

Alonso, O., Strötgen, J., Baeza-Yates, R., & Gertz, M. (2011). Temporal information retrieval: Challenges and opportunities. In *Proceedings of the 1st international temporal web analytics workshop (TWAW 2011)* (pp. 1–8).

Baeza-Yates, R. (2005). Searching the future. In *Proceedings of the mathematical/formal methods in information retrieval workshop associated to SIGIR05*.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM, 35*(12), 29–38.

Bethard, S., Derczynski, L., Pustejovsky, J., & Verhagen, M. (2015). SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the workshop on semantic evaluation*. ACL.

Campos, R., Dias, G., Jorge, A. M., & Jatowt, A. (2014). Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR), 47*(2), 15.

Campos, R., Dias, G., Jorge, A., & Nunes, C. (2012). GTE: A distributional second-order co-occurrence approach to improve the identification of top relevant dates in web snippets. In *Proceedings of the 21st ACM international conference on information and knowledge management (CIKM 2012)* (pp. 2035–2039). ACM.

Campos, R., Jorge, A. M., Dias, G., & Nunes, C. (2012). Disambiguating implicit temporal queries by clustering top relevant dates in web snippets. In *Proceedings of the IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology (WI-IAT)* (pp. 1–8). IEEE.

Derczynski, L., & Gaizauskas, R. (2013). Information retrieval for temporal bounding. In *Proceedings of the 2013 conference on the theory of information retrieval (ICTIR '13)* (pp. 129–130). ACM.

Derczynski, L. R., Yang, B., & Jensen, C. S. (2013). Towards context-aware search and analysis on social media data. In *Proceedings of the 16th international conference on extending database technology (EDBT 2013)* (pp. 137–142). ACM.

Dias, G., Moreno, J. G., Jatowt, A., & Campos, R. (2012). Temporal web image retrieval. In *String processing and information retrieval. Lecture notes in computer science* (Vol. 7608, pp. 199–204).

Jatowt, A., Au Yeung, C.-m., & Tanaka, K. (2015). Generic method for detecting content time of documents. *Information Processing & Management, 51*.

Jensen, C. S., & Snodgrass, R. T. (1999). Temporal data management. *IEEE Transactions on Knowledge and Data Engineering, 11*(1), 36–44.

Ji, H., Grishman, R., & Dang, H. T. (2011). Overview of the TAC 2011 knowledge base population track. In *Proceedings of the text analysis conference (TAC)*.

Joho, H., Jatowt, A., & Blanco, R. (2015). Temporal information searching behaviour and strategies. *Information Processing & Management, 51*.

Kahle, B. (1997). Preserving the internet. *Scientific American, 276*(3), 82–83.

Kanhabua, N., & Nørvåg, K. (2010). Determining time of queries for re-ranking search results, *Research and advanced technology for digital libraries. Lecture notes in computer science: Vol. 6273* (pp. 261–272). Springer.

Kar, M., Nunes, S., & Ribeiro, C. (2015). Summarization of changes in dynamic text collection using latent Dirichlet allocation model. *Information Processing & Management, 51*.

Kenter, T., Balog, K., & de Rijke, M. (2015). Evaluating document filtering systems over time. *IInformation Processing & Management, 51*.

Lampos, V., Preotiuc-Pietro, D., & Cohn, T. (2013). A user-centric model of voting intention from social media. In *Proceedings of the annual meetings of the association for computational linguistics (ACL 2013)* (pp. 993–1003).

Llorens, H., Chambers, N., UzZaman, N., Mostafazadeh, N., Allen, J., & Pustejovsky, J. (2015). SemEval-2015 Task 5: QA TempEval. In *Proceedings of the workshop on semantic evaluation*. ACL.

Matthews, M., Tolchinsky, P., Blanco, R., Atserias, J., Mika, P., & Zaragoza, H. (2010). Searching through time in the New York Times. In *Proceedings of the 4th workshop on human–computer interaction and information retrieval (HCIR 2010) in association with IIiX* (pp. 41–44). ACM.

Minard, A.-L., Agirre, E., Aldabe, I., van Erp, M., Magnini, B., Rigau, G., et al. (2015). SemEval-2015 Task 4: TimeLine: Cross-document event ordering. In *Proceedings of the workshop on semantic evaluation*. ACL.

Plaisant, C., Shneiderman, B., & Mushlin, R. (1998). An information architecture to support the visualization of personal histories. *Information Processing & Management, 34*(5), 581–597.

Popescu, O., & Strapparava, C. (2015). SemEval-2015 Task 7: Diachronic text evaluation. In *Proceedings of the workshop on semantic evaluation*. ACL.

Pustejovsky, J., Lee, K., Bunt, H., & Romary, L. (2010). ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the international language resources and evaluation conference (LREC 2010), ELRA* (pp. 394–397).

Radinsky, K., & Horvitz, E. (2013). Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on web search and data mining (WSDM)* (pp. 255–264). ACM.

Rula, A., Palmonari, M., Ngomo, A.-C. N., Gerber, D., Lehmann, J., & Bühmann, L. (2014). Hybrid acquisition of temporal scopes for RDF data, *Lecture notes in computer science: Vol. 8465. The semantic web: Trends and challenges* (pp. 488–503). Springer.

Strötgen, J., Armiti, A., Van Canh, T., Zell, J., & Gertz, M. (2014). Time for more languages: Temporal tagging of Arabic, Italian, Spanish, and Vietnamese. *ACM Transactions on Asian Language Information Processing (TALIP), 13*(1), 1–21.

Strötgen, J., & Gertz, M. (2012). Event-centric search and exploration in document collections. In *Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries (JCDL'12)* (pp. 223–232). ACM.

Sugumaran, R., & Voss, J. (2012). Real-time spatio-temporal analysis of West Nile Virus using Twitter data. In *Proceedings of the 3rd international conference on computing for geospatial research and applications (COM.Geo '12)* (p. 39). ACM.

Talukdar, P. P., Wijaya, D., & Mitchell, T. (2012). Coupled temporal scoping of relational facts. In *Proceedings of the fifth ACM international conference on web search and data mining (WSDM)* (pp. 73–82). ACM.

Verhagen, M. (2005). Drawing TimeML relations with T-Box. In *Proceedings of the Dagstuhl Seminar on Annotating, extracting and reasoning about time and events. Dagstuhl Seminars* (Vol. 05151, pp. 7–28).

Zhao, X., Jin, P., & Yue, L. (2015). Discovering topic time from web news. *Information Processing & Management, 51*.

Leon Derczynski
University of Sheffield, United Kingdom

Jannik Strötgen
Heidelberg University, Germany

Ricardo Campos
Polytechnic Institute of Tomar, Portugal

LIAAD-INESC TEC, Portugal

Omar Alonso
Microsoft Corporation, United States

E-mail address: leon@dcs.shef.ac.uk

Received 2 April 2015