Temporal Information Retrieval

Nattiya Kanhabua
Department of Computer Science
Aalborg University, Denmark
nattiya@cs.aau.dk

Avishek Anand
L3S Research Center
Leibniz Universität Hannover, Germany
anand@L3S.de

ABSTRACT

The study of temporal dynamics and its impact can be framed within the so-called *temporal IR approaches*, which explain how user behavior, document content and scale vary with time, and how we can use them in our favor in order to improve retrieval effectiveness.

This half-day tutorial will outline research issues with respect to temporal dynamics, and provide a comprehensive overview of temporal IR approaches, essentially regarding processing dynamic content, temporal information extraction, temporal query analysis, and time-aware retrieval and ranking. The tutorial is structured into two sessions. During the first session, we will explain the general and wide aspects associated to temporal dynamics by focusing on the web domain, from content and structural changes to variations of user behavior and interactions. We will begin with temporal indexing and query processing. Next step, we will explain current approaches to time-aware retrieval and ranking, which can be classified into different types based on two main notions of relevance with respect to time, namely, recency-based ranking, and time-dependent ranking.

In the latter session, we will describe research issues centered on determining the temporal intent of queries, and time-aware query enhancement, e.g., temporal relevance feedback, and time-aware query reformulation. In addition, we present applications in related research areas, e.g., exploration, summarization, and clustering of search results, as well as future event retrieval and prediction. To this end, we conclude our tutorial and outline future directions.

This tutorial targets graduate students, researchers and practitioners in the field of information retrieval. The goal is to provide an overview as well as an important context that enables further research on and practical applications within this area.

1. INTRODUCTION

The time dimension has strong influence in many domains, e.g., Topic Detection and Tracking (TDT) [1], and Emerging Trend Detection (ETD) [12]. In our context, we focus on the impact of time on search processes seen by the evolution of the Web [30] that can be categorized with respect to its changes of 1) content and structure, and 2) user querying behavior. The content of the Web changes constantly over time, e.g., documents are added, modified or deleted continuously. Similarly, the link structure of the Web also evolves [16]. Content and structure changes affect basic processes like crawling and indexing, but also the computation of graph-based authority measures that typically are used for document ranking.

Another web evolution aspect is the change of user querying behavior, which can be observable at least in two ways. First, search traffic for particular queries varies over time and presents certain temporal patterns, such as, spikes, periodicity (e.g., weekly or monthly), seasonality and trends [32]. Second, many queries are time-sensitive, which contain underlying temporal information needs that do not exhibit a temporal pattern in search streams [38]. Understanding temporal search intent is a challenging task before applying an appropriate time-aware ranking method.

Temporal IR has received a large share of attention in the last decade. For example, there are at least two main research challenges related to temporal IR that have been organized recently: the TREC Temporal Summarization Track and Temporalia (Temporal Information Access) [24]. In this tutorial, we will give a comprehensive overview of the important aspects of temporal information retrieval, which intends to support participants with basic knowledge who want to get acquainted with the research area, and for more advanced researchers aiming to understand in more detail this field of research. This tutorial complements previous work, such as, an overview of challenges and opportunities in temporal IR [3], a tutorial [41], and a recent survey [14] as follows:

- We categorize temporal IR research challenges into three main topics, i.e., temporal indexing and query processing, temporal query analysis, and time-aware ranking;
- We provide a review of most recent state-of-the-art and best practices in the topics related to the research challenges;
- We present a number of interesting application areas in order to open new horizons for research and innovations;
- We provide detailed description of the approaches in order to make it suitable as a teaching material.

2. TOPIC

This section outlines the topics that our tutorial will cover.

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07. . . $15.00
DOI: http://dx.doi.org/10.1145/2911451.2914805

2.1 Temporal Indexing and Query Processing

Evolving web content has an impact on several components of a search engine. In this section, we revisit three main building blocks of the search system, namely *crawling*, *indexing*, and *query processing*.

We will explain how to index versioned archives, which requires even more complex indexing and maintenance procedures. Several works have been proposed in the area of versioned indexing and query processing [7, 10, 13, 21, 37, 53]. Two general approaches to text indexing, which have been employed, are either based on index compression (by exploiting the redundancy between consecutive versions) [20, 21, 22], or based on index partitioning for faster temporal selections [5, 6, 7, 10]. We will present recent works on indexing, query processing and index maintenance efforts towards supporting temporally qualified queries, i.e., [5, 6, 7, 10], in more detail.

2.2 Temporal Query Analysis

Understanding search intent behind a user's query is the first step before applying a suitable retrieval and ranking model. In the temporal search realm, a system should be able to identify a query that contains an underlying temporal information need, so-called a *temporal query* [24]. We categorize temporal queries into two main classes. The first class of temporal queries can be observed through *temporal patterns* of user search behavior in query logs, whereas the second class of temporal queries may not exhibit such temporal patterns in a query stream, but can be referred to specific time or a particular event.

In this tutorial, we will explain current approaches to understanding temporal query intent [32], as well as methods for detecting and predicting time-sensitive queries, such as, spiky, periodic or seasonal queries [43, 45]. In addition, we will explain how to determine the underlying temporal information needs for a given temporal query by analyzing temporal document collections [25, 29, 54].

We will address the problem of *multi-faceted temporal queries* and its impact on result diversification [36, 46, 51, 56]. Finally, we will describe query enhancement techniques applied to temporal queries, which include *time-based pseudo-relevance feedback* [31, 40], and time-aware query reformulation [18, 26, 49].

2.3 Time-aware Ranking

Existing works on time-aware ranking can be classified into different types based on two main notions of relevance with respect to time: 1) recency-based ranking [16, 17, 19, 33], and 2) time-dependent ranking [11, 29]. Recency-based ranking methods promote documents that are recently created or updated. The preference of freshness is quite common in a general web search, where a user looks for recent information, e.g., about breaking news. On the other hand, time-dependent ranking methods takes into account the relevant time periods underlying a given temporal query, which this type of ranking explicitly adjusts the score of a document with respect to the time of queries, for example, by assuming that documents with creation dates close to the query's time are more relevant and thus must be ranked higher.

In addition to the aforementioned classification of time-aware methods, we also outline existing works that explicitly model entities and events into retrieval and ranking processes [9, 27, 44]. Note that, entity and event-based retrieval returns events instead of documents, whereas time-aware ranking methods use a very simple notion of events.

2.4 Applications of Temporal IR

Some applications have taken time-based exploration of textual archives beyond just searching over time, e.g., Time Explorer [34], and time-aware exploration search [39]. Filtering and displaying information might benefit from presenting time information conveniently in some domains, for instance, generating a timeline summary for a specific news story [35, 52, 55] and for a given entity [50]. Arguably, the most widespread summarization technology is the (query) focused summaries produced by search engines, or search results snippets using temporal expressions, which include the most frequent units of time appearing in search result pages [2, 15, 48]. There exists another form of temporal summary in terms of landmark documents, authors, and topics [47].

A series of temporal applications are the ones based on *prediction of events*, e.g., [8, 23, 27]. The most challenging task is the actual *forecast* of forthcoming events. A method to turn a standard IR engine into a future event predicting machine is proposed in [4], whereas an approach to forecasting real-life events by extracting information from a corpus comprised of 22 years of news stories is presented in [42].

3. DETAILED SCHEDULE

A tentative schedule, which is aimed to meet a high-quality presentation within the chosen time period, is as follows.

- 9:00 - 10:30 Part I (**1.5 hour**)
  - Introduction to Temporal IR (*20 minutes*)
  - Temporal Indexing and Query Processing (*35 minutes*)
  - Time-aware Retrieval and Ranking (*35 minutes*)
- 10:30 - 11:00 Coffee break
- 11:00 - 12:30 Part II (**1.5 hour**)
  - Temporal Query Analysis (*40 minutes*)
  - Applications of Temporal IR (*40 minutes*)
  - Conclusions and Future Directions (*10 minutes*)

4. INTENDED AUDIENCE AND MATERIAL

This tutorial targets graduate students, junior researchers and practitioners in the field of information retrieval. Our prospective participants should have a basic knowledge of search processes, essentially regarding web crawling, document indexing, query analysis, and retrieval and ranking. A prerequisite skill about other research areas, e.g., natural language processing, is required but not mandatory. The participants can download slides and a survey on temporal information retrieval [28] from the authors' homepages.

5. BIOGRAPHY

Nattiya Kanhabua is an assistant professor at the Department of Computer Science, Aalborg University, Denmark. Her research interests are information retrieval, data mining, machine learning, and spatial and temporal analytics. She did her PhD at the Department of Computer and

Information Science, Norwegian University of Science and Technology (NTNU). She was a postdoctoral researcher at the L3S Research Center Hannover, Germany. At L3S, she worked in several research projects, e.g., 1) EU Project ForgetIT: Concise Preservation by Combining Managed Forgetting and Contextualized Remembering, 2) ALEXANDRIA, an ERC Advanced Grant Project on Foundations for Temporal Retrieval, Exploration and Analytics in Web Archives, and 3) Medical Ecosystem: Personalized Event-based Surveillance. She has published her research work in top-tier conferences, e.g., SIGIR, WSDM, CIKM, JCDL and ECIR.

Avishek Anand is a postdoctoral researcher at the L3S Research Center in Hannover, Germany. His research interests lay in the intersection of retrieval, mining, and data management aspects of temporal Web collections like Web archives, Wikipedia and news collections. He did his PhD at the Department of Databases and Information Systems, Max Planck Institute for Informatics, Saarbruecken, Germany, where he worked on indexing and query processing approaches for supporting temporal text workloads. Currently, he is working on retrieval models for historical intents, tag-based search over Archives and mining methods for enriching Wikipedia using news collections. He has published his research in several top-tier conferences, such as, SIGIR, WSDM, CIKM, ICDE and EDBT.

6. REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 37–45, 1998.

[2] O. Alonso, M. Gertz, and R. A. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 97–106, 2009.

[3] O. Alonso, J. Strötgen, R. A. Baeza-Yates, and M. Gertz. Temporal information retrieval: Challenges and opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW 2011)*, 2011.

[4] G. Amodeo, R. Blanco, and U. Brefeld. Hybrid models for future event prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1981–1984, 2011.

[5] A. Anand, S. Bedathur, K. Berberich, and R. Schenkel. Efficient temporal keyword search over versioned text. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 699–708, 2010.

[6] A. Anand, S. Bedathur, K. Berberich, and R. Schenkel. Temporal index sharding for space-time efficiency in archive search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 545–554, 2011.

[7] A. Anand, S. Bedathur, K. Berberich, and R. Schenkel. Index maintenance for time-travel text search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 235–244, New York, NY, USA, 2012. ACM.

[8] C.-m. Au Yeung and A. Jatowt. Studying How the Past is Remembered: Towards computational history through large scale text mining. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1231–1240, 2011.

[9] R. A. Baeza-Yates. Searching the future. In *Proceedings of SIGIR workshop on mathematical/formal methods in information retrieval MF/IR*, SIGIR '05, 2005.

[10] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 519–526, 2007.

[11] K. Berberich, S. J. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. In *Proceedings of the 32nd European Conference on IR Research on Advances in Information Retrieval*, ECIR '10, pages 13–25, 2010.

[12] M. Berry. *Survey of Text Mining: Clustering, Classification, and Retrieval.* Springer, Sep. 2003.

[13] A. Z. Broder, N. Eiron, M. Fontoura, M. Herscovici, R. Lempel, J. McPherson, R. Qi, and E. J. Shekita. Indexing shared content in information retrieval systems. In *Proceedings of the 10th International Conference on Extending Database Technology*, EDBT '06, pages 313–330, 2006.

[14] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt. Survey of temporal information retrieval and related applications. *ACM Comput. Surv.*, 47(2):15:1–15:41, Aug. 2014.

[15] R. Campos, A. M. Jorge, G. Dias, and C. Nunes. Disambiguating implicit temporal queries by clustering top relevant dates in web snippets. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 1–8, 2012.

[16] N. Dai and B. D. Davison. Freshness matters: in flowers, food, and web authority. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 114–121, 2010.

[17] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. Towards recency ranking in web search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 11–20, 2010.

[18] M. Efron. Query representation for cross-temporal information retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 383–392, 2013.

[19] J. L. Elsas and S. T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 1–10, 2010.

[20] J. He and T. Suel. Faster temporal range queries over versioned text. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 565–574, 2011.

[21] J. He, H. Yan, and T. Suel. Compact full-text indexing of versioned document collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 415–424, 2009.

[22] J. He, J. Zeng, and T. Suel. Improved index compression techniques for versioned document collections. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, CIKM '10, pages 1239–1248, 2010.

[23] A. Jatowt, É. Antoine, Y. Kawai, and T. Akiyama. Mapping temporal horizons: Analysis of collective future and past related attention in twitter. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 484–494, 2015.

[24] H. Joho, A. Jatowt, and B. Roi. A survey of temporal web search experience. In *Proceedings of the 22nd International Conference on World Wide Web (Companion)*, WWW '13, pages 1101–1108, 2013.

[25] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25, July 2007.

[26] A. C. Kaluarachchi, A. S. Varde, S. Bedathur, G. Weikum, J. Peng, and A. Feldman. Incorporating terminology evolution for query translation in text retrieval with association rules. In *Proceedings of the 19th ACM*

international conference on Information and knowledge management, CIKM '10, pages 1789–1792, 2010.

[27] N. Kanhabua, R. Blanco, and M. Matthews. Ranking related news predictions. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 755–764, 2011.

[28] N. Kanhabua, R. Blanco, and K. Nørvåg. Temporal information retrieval. *Foundations and Trends in Information Retrieval*, 9(2):91–208, 2015.

[29] N. Kanhabua and K. Nørvåg. Determining time of queries for re-ranking search results. In *Proceedings of the 14th European conference on Research and advanced technology for digital libraries*, ECDL'10, pages 261–272, 2010.

[30] Y. Ke, L. Deng, W. Ng, and D.-L. Lee. Web dynamics and their ramifications for the development of web search engines. *Computer Networks*, 50(10):1430–1447, July 2006.

[31] M. Keikha, S. Gerani, and F. Crestani. TEMPER: A temporal relevance feedback method. In *Proceedings of the 33rd European Conference on IR Research on Advances in Information Retrieval*, ECIR '11, pages 436–447, 2011.

[32] A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais. Understanding temporal query dynamics. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining*, WSDM '11, pages 167–176, 2011.

[33] X. Li and W. B. Croft. Time-based language mmdels. In *Proceedings of the 12th international conference on Information and knowledge management*, CIKM '03, pages 469–475, 2003.

[34] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, and H. Zaragoza. Searching through time in the new york times. In *HCIR Workshop on Bridging Human-Computer Interaction and Information Retrieval*, HCIR '10, 2010.

[35] R. McCreadie, C. Macdonald, and I. Ounis. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 301–310, 2014.

[36] T. N. Nguyen and N. Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *Proceedings of the 36th European Conference on Advances in Information Retrieval*, ECIR '14, pages 222–234, 2014.

[37] K. Nørvåg and A. O. Nybø. Dyst: Dynamic and scalable temporal text indexing. In *Proceedings of the 13th International Symposium on Temporal Representation and Reasoning*, TIME '06, pages 204–211, 2006.

[38] S. Nunes, C. Ribeiro, and G. David. Use of temporal expressions in web search. In *Proceedings of the 30th European Conference on IR Research on Advances in Information Retrieval*, ECIR '08, pages 580–584, 2008.

[39] D. Odijk, G. Santucci, M. d. Rijke, M. Angelini, and G. Granato. Time-aware exploratory search: Exploring word meaning through time. In *SIGIR 2012 Workshop on Time-aware Information Access*, TAIA '12, 2012.

[40] M.-H. Peetz, E. Meij, and M. Rijke. Using temporal bursts for query modeling. *Information Retrieval*, 17(1):74–108, 2014.

[41] K. Radinsky, F. Diaz, S. T. Dumais, M. Shokouhi, A. Dong, and Y. Chang. Temporal web dynamics and its application to information retrieval. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 781–782, 2013.

[42] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 255–264, 2013.

[43] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 599–608, 2012.

[44] D. Shan, W. X. Zhao, R. Chen, B. Shu, Z. Wang, J. Yao, H. Yan, and X. Li. Eventsearch: A system for event discovery and retrieval on multi-type historical data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1564–1567, 2012.

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

[46] J. Singh, W. Nejdl, and A. Anand. History by diversity: Helping historians search news archives. In *Proceedings of the 1st International ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)*, 2016.

[47] R. Sipos, A. Swaminathan, P. Shivaswamy, and T. Joachims. Temporal corpus summarization using submodular word coverage. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 754–763, 2012.

[48] K. M. Svore, J. Teevan, S. T. Dumais, and A. Kulkarni. Creating temporally dynamic web search snippets. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1045–1046, 2012.

[49] N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmann, and T. Risse. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *Proceedings the 24th International Conference on Computational Linguistics*, COLING '12, pages 2553–2568. ACL, 2012.

[50] T. A. Tran, C. Niederée, N. Kanhabua, U. Gadiraju, and A. Anand. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1201–1210, 2015.

[51] S. Whiting, K. Zhou, J. Jose, and M. Lalmas. Temporal variance of intents in multi-faceted event-driven information needs. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 989–992, 2013.

[52] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 745–754, 2011.

[53] J. Zhang and T. Suel. Efficient search in large textual collections with redundancy. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 411–420, 2007.

[54] R. Zhang, Y. Konda, A. Dong, P. Kolari, Y. Chang, and Z. Zheng. Learning recurrent event queries for web search. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1129–1139, 2010.

[55] X. W. Zhao, Y. Guo, R. Yan, Y. He, and X. Li. Timeline generation with social attention. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 1061–1064, 2013.

[56] K. Zhou, S. Whiting, J. M. Jose, and M. Lalmas. The impact of temporal intent variability on diversity evaluation. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR '13, pages 820–823, 2013.