Modeling Temporal Evidence from External Collections

Flávio Martins
NOVA LINCS
School of Science and Technology
Universidade NOVA de Lisboa
Caparica, Portugal
flaviomartins@acm.org

João Magalhães
NOVA LINCS
School of Science and Technology
Universidade NOVA de Lisboa
Caparica, Portugal
jm.magalhaes@fct.unl.pt

Jamie Callan
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
callan@cs.cmu.edu

ABSTRACT

Newsworthy events are broadcast through multiple mediums and prompt the crowds to produce comments on social media. In this paper, we propose to leverage on this behavioral dynamics to estimate the most relevant time periods for an event (i.e., query). Recent advances have shown how to improve the estimation of the temporal relevance of such topics. In this approach, we build on two major novelties. First, we mine temporal evidences from *hundreds of external sources* into topic-based external collections to improve the robustness of the detection of relevant time periods. Second, we propose a formal retrieval model that *generalizes the use of the temporal dimension* across different aspects of the retrieval process. In particular, we show that temporal evidence of external collections can be used to (i) infer a topic's temporal relevance, (ii) select the query expansion terms, and (iii) re-rank the final results for improved precision. Experiments with TREC Microblog collections show that the proposed time-aware retrieval model makes an effective and extensive use of the temporal dimension to improve search results over the most recent temporal models. Interestingly, we observe a strong correlation between precision and the temporal distribution of retrieved and relevant documents.

KEYWORDS

Microblog search, social media, learning to rank, time-aware ranking models, temporal information retrieval

1 INTRODUCTION

A networked world and the increasing pervasiveness of Internet access enables the rapid adoption of new online communication mediums to discuss current events. Previous research has explored this symbiosis between Twitter and the news [17, 29] and linked the two mediums [12, 33]. Events are discussed on the Web as they happen and people following them can add to the conversation immediately. Hence, improving the *temporal relevance estimation* for searching such events became a significant research priority.

Nowadays, the state-of-the-art Web search systems are based on learning to rank feature models that combine multiple text retrieval functions as well as other features. Relevance on Twitter has many dimensions: authority, popularity, freshness, geographical context, and topical relevance. Previously, time-aware ranking research explored the assumption that fresh documents are more relevant [19]. Later models revised this assumption in line with what is observed in Twitter: for time-sensitive queries, documents tend to cluster temporally [9, 11]. Our approach is based on the intuition that discussions about a topic and its subtopics are likely to occur around the same time across multiple mediums.

The rationale is that newsworthy events trigger a cascade of activity on the Web and Twitter. This information can be useful for ranking and, in some cases, can be gathered with ease. The news often have a good coverage of current topics, clean journalistic language, and reliable timestamps. Thus, it is desirable to mine news sources to offer more context to the *tweets* as well as to the users' queries intent. In particular, we aim to explore the *crowd aggregation effect* to extract temporal evidence from news verticals. Temporal evidence is further used to refine the selection of query expansion terms and to estimate query topics temporal relevance. This approach is completed with the re-ranking of the final search results leading to improved precision. Hence, the proposed method brings a series of novel contributions:

- Explore the crowd effect by aggregating posts published by news sources into topic-based external collections;
- Mining of crowds' temporal evidence at different granularities (i.e., *verticals*, *documents*, and *terms*);
- A formal time-aware ranking model that unifies multiple temporal features into a single comprehensive retrieval model.

Including the temporal dimension at the different steps of the search engine pipeline, improves the accuracy of several retrieval tasks, leading to greater overall gains. This is possible because, the temporal dimension introduces stronger evidence in many decision tasks (e.g., selection of query expansion terms). Evaluation on the TREC 2013 and TREC 2014 Microblog Track datasets shows that the proposed retrieval model outperforms state-of-the-art methods.

This paper is organized as follows: in Section 2 we present the related work; in Section 3 the formal temporal ranking model is detailed and the following sections detail its implementation; evaluation is presented in Section 5; and a more fine-grained discussion of results in Section 6.

REFERENCES

[1] Robin Aly, Djoerd Hiemstra, and Thomas Demeester. 2013. Taily: Shard Selection Using the Tail of Score Distributions. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, New York, NY, USA, 673–682. https://doi.org/10.1145/2484028.2484033

[2] Jaime Arguello, Fernando Diaz, and Jamie Callan. 2011. Learning to Aggregate Vertical Results into Web Search Results. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 201–210. https://doi.org/10.1145/2063576.2063611

[3] Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2012. Effective Query Formulation with Multiple Information Sources. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 443–452. https://doi.org/10.1145/2124295.2124349

[4] Jaeho Choi and W. Bruce Croft. 2012. Temporal Models for Microblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2491–2494. https://doi.org/10.1145/2396761.2398674

[5] Jaeho Choi, W. Bruce Croft, and Jin Young Kim. 2012. Quality Models for Microblog Retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1834–1838. https://doi.org/10.1145/2396761.2398527

[6] Miguel Costa, Francisco Couto, and Mário Silva. 2014. Learning Temporal-Dependent Ranking Models. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 757–766. https://doi.org/10.1145/2600428.2609619

[7] Olga Craveiro, Joaquim Macedo, and Henrique Madeira. 2014. Query Expansion with Temporal Segmented Texts. In *Advances in Information Retrieval (ECIR '14)*. Springer, Cham, 612–617. https://doi.org/10.1007/978-3-319-06028-6_65

[8] Na Dai, Milad Shokouhi, and Brian D. Davison. 2011. Learning to Rank for Freshness and Relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 95–104. https://doi.org/10.1145/2009916.2009933

[9] W. Dakka, L. Gravano, and P.G. Ipeirotis. 2012. Answering General Time-Sensitive Queries. *IEEE Transactions on Knowledge and Data Engineering* 24, 2 (Feb. 2012), 220–235. https://doi.org/10.1109/TKDE.2010.187

[10] Fernando Diaz and Donald Metzler. 2006. Improving the Estimation of Relevance Models Using Large External Corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 154–161. https://doi.org/10.1145/1148170.1148200

[11] Miles Efron, Jimmy Lin, Jiyin He, and Arjen de Vries. 2014. Temporal Feedback for Tweet Search with Non-Parametric Density Estimation. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 33–42. https://doi.org/10.1145/2600428.2609575

[12] Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 239–249.

[13] Rosie Jones and Fernando Diaz. 2007. Temporal Profiles of Queries. *ACM Trans. Inf. Syst.* 25, 3 (July 2007). https://doi.org/10.1145/1247715.1247720

[14] Nattiya Kanhabua and Kjetil Nørvåg. 2012. Learning to Rank Search Results for Time-Sensitive Queries. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2463–2466. https://doi.org/10.1145/2396761.2398667

[15] Mostafa Keikha, Shima Gerani, and Fabio Crestani. 2011. Time-Based Relevance Models. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 1087–1088. https://doi.org/10.1145/2009916.2010062

[16] Anagha Kulkarni, Almer S. Tigelaar, Djoerd Hiemstra, and Jamie Callan. 2012. Shard Ranking and Cutoff Estimation for Topically Partitioned Collections. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 555–564. https://doi.org/10.1145/2396761.2396833

[17] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What Is Twitter, a Social Network or a News Media?. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 591–600. https://doi.org/10.1145/1772690.1772751

[18] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. ACM, New York, NY, USA, 120–127. https://doi.org/10.1145/383952.383972

[19] Xiaoyan Li and W. Bruce Croft. 2003. Time-Based Language Models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*. ACM, New York, NY, USA, 469–475. https://doi.org/10.1145/956863.956951

[20] Jimmy Lin and Miles Efron. 2013. Overview of the TREC-2013 Microblog Track. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, Ellen M. Voorhees (Ed.), Vol. Special Publication 500-302. National Institute of Standards and Technology (NIST).

[21] Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. 2014. Overview of the TREC-2014 Microblog Track. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. Special Publication 500-308. National Institute of Standards and Technology (NIST).

[22] Flávio Martins, João Magalhães, and Jamie Callan. 2016. Barbara Made the News: Mining the Behavior of Crowds for Time-Aware Learning to Rank. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. ACM, San Francisco, CA, USA.

[23] Flávio Martins, João Magalhães, and Jamie Callan. 2018. A Vertical PRF Architecture for Microblog Search. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '18)*. ACM, New York, NY, USA.

[24] Donald Metzler, Congxing Cai, and Eduard Hovy. 2012. Structured Event Retrieval over Microblog Archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 646–655.

[25] Brendan O'Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*. 384–385.

[26] Maria-Hendrike Peetz, Edgar Meij, and Maarten de Rijke. 2013. Using Temporal Bursts for Query Modeling. *Information Retrieval* (July 2013), 1–35. https://doi.org/10.1007/s10791-013-9227-2

[27] Jinfeng Rao and Jimmy Lin. 2016. Temporal Query Expansion Using a Continuous Hidden Markov Model. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. ACM, New York, NY, USA, 295–298. https://doi.org/10.1145/2970398.2970424

[28] Tetsuya Sakai. 2014. Statistical Reform in Information Retrieval? *SIGIR Forum* 48, 1 (June 2014), 3–12. https://doi.org/10.1145/2641383.2641385

[29] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09)*. ACM, New York, NY, USA, 42–51. https://doi.org/10.1145/1653771.1653781

[30] Milad Shokouhi. 2007. Central-Rank-Based Collection Selection in Uncooperative Distributed Information Retrieval. In *Advances in Information Retrieval (ECIR '07)*. Springer, Berlin, Heidelberg, 160–172. https://doi.org/10.1007/978-3-540-71496-5_17

[31] Milad Shokouhi and Luo Si. 2011. Federated Search. *Found. Trends Inf. Retr.* 5, 1 (Jan. 2011), 1–102. https://doi.org/10.1561/1500000010

[32] Luo Si and Jamie Callan. 2003. Relevant Document Distribution Estimation Method for Resource Selection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*. ACM, New York, NY, USA, 298–305. https://doi.org/10.1145/860435.860490

[33] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. 2011. Linking Online News and Social Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 565–574. https://doi.org/10.1145/1935826.1935906

[34] Yulu Wang and Jimmy Lin. 2017. Partitioning and Segment Organization Strategies for Real-Time Selective Search on Document Streams. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/3018661.3018727

[35] Wouter Weerkamp, Krisztian Balog, and Maarten de Rijke. 2012. Exploiting External Collections for Query Expansion. *ACM Trans. Web* 6, 4 (Nov. 2012), 18:1–18:29. https://doi.org/10.1145/2382616.2382621

[36] Stewart Whiting, Iraklis A. Klampanos, and Joemon M. Jose. 2012. Temporal Pseudo-Relevance Feedback in Microblog Retrieval. In *Advances in Information Retrieval (ECIR '12)*. Springer, Berlin, Heidelberg, 522–526. https://doi.org/10.1007/978-3-642-28997-2_55

[37] Tan Xu, Douglas W. Oard, and Paul McNamee. 2014. HLTCOE at TREC 2014: Microblog and Clinical Decision Support. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. Special Publication 500-308. National Institute of Standards and Technology (NIST).

[38] Chengxiang Zhai and John Lafferty. 2001. Model-Based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM '01)*. ACM, New York, NY, USA, 403–410. https://doi.org/10.1145/502585.502654

[39] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179–214. https://doi.org/10.1145/984321.984322