Wikipedia as a Time Machine

Stewart Whiting and Joemon M. Jose
School of Computing Science
University of Glasgow
Scotland, UK
{stewh,jj}@dcs.gla.ac.uk

Omar Alonso
Microsoft Corp.
Mountain View
California, USA
omar.alonso@microsoft.com

ABSTRACT

Wikipedia encyclopaedia projects, which consist of vast collections of user-edited articles covering a wide range of topics, are among some of the most popular websites on internet. With so many users working collaboratively, mainstream events are often very quickly reflected by both authors editing content and users reading articles. With temporal signals such as changing article content, page viewing activity and the link graph readily available, Wikipedia has gained attention in recent years as a source of temporal event information. This paper serves as an overview of the characteristics and past work which support Wikipedia (English, in this case) for time-aware information retrieval research. Furthermore, we discuss the main content and meta-data temporal signals available along with illustrative analysis. We briefly discuss the source and nature of each signal, and any issues that may complicate extraction and use. To encourage further temporal research based on Wikipedia, we have released all the distilled datasets referred to in this paper.

Categories and Subject Descriptors: H.4.0 [Information Systems Applications, General]
Keywords: Wikipedia; Time; Temporal; Events

1. INTRODUCTION

Wikipedia has surged in popularity to become the seventh most visited website on the internet[1]. Since its creation in early 2001, the founding English language encyclopaedia project[2] has grown to include over 4.2M articles covering a wide range of topics, with an average of 20.35 revisions per article. For most countries, the local language Wikipedia domain is among the most visited websites. As well as the Wikipedia encyclopaedia in numerous languages, further spin-off 'Wiki' projects such as *Wikinews* and *Wiktionary* have also become increasingly popular.

Wikipedia's success is likely down to its open collaborative model of article management and organisation. Although there is a small dedicated administrative team to ensure consistency of high-profile articles, the vast majority of articles are open for immediate editing by anyone. This has encouraged a large-scale crowd-sourced effort by users to continuously create and edit articles to include their knowledge and understanding. Each article contains information and multimedia related to a significant topic, such as people, places or events. In many cases, authoring activity is triggered by users reporting ongoing real-world events, often very soon after they occurred, leading to an ever-evolving large-scale source of temporal information [16, 18, 14, 6].

In this paper we are interested in how Wikipedia reflects events through several temporal signals – which can be utilised for discovering and understanding events and their related topics. In particular, we are interested in how this knowledge can be modelled and exploited in diverse areas of time-aware information retrieval.

Events can be loosely defined as a 'significant happening or occurrence', delimited in scope by time and place, with an identifiable set of participants [2]. We consider both historic and recent events. Many events mentioned in Wikipedia article content may have occurred well before Wikipedia was started, or, were not reflected in the related articles shortly after their occurrence. Conversely, more recent major events will likely be reflected both in the content and meta-data signals, for instance, increased article viewing.

In the field of information retrieval (IR), many recent studies have utilised various Wikipedia temporal signals for understanding events. A number of demo papers [4, 18, 14] and a recent evaluated study [6] used Wikipedia content and meta-data for on- and off-line topic detection and tracking (TDT) [2]. All used a combination of signals with various heuristic and machine learning techniques to filter noise and identify important events. [11] took a different tack and used Wikipedia events to filter noise from Twitter-based TDT. Event summarisation using Wikipedia article content has seen some interest [7, 18], with further work likely to arise from the 2013 TREC Temporal Summarization track. Many tools have been proposed for exploring article history [17]. The temporal variation of ambiguous [20] and multi-faceted [19] information needs has been quantified using Wikipedia signals.

In Section 2 we discuss some of the characteristics that make Wikipedia suitable for time-aware research. Although this paper does not present any novel research findings, in Section 3 we provide insight into the many temporal signals available from Wikipedia for researchers to exploit.

2. WIKIPEDIA TEMPORAL CHARACTERISTICS

Wikipedia is the subject of a vast body of research across diverse fields, including social sciences, psychology and information sciences. In this section we outline some of the key aspects relevant to understanding time through Wikipedia.

[1] http://www.alexa.com/topsites (June, 2013)
[2] http://en.wikipedia.org

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579048.

2.1 Freshness and Timeliness

The latency of events being reflected in Wikipedia temporal signals is important for time-aware research. An ever increasing amount of editing activity is triggered by users reporting ongoing real-world events, often very soon after they occurred [16, 9]. For mainstream events, [11] state that Wikipedia lags behind Twitter by about two hours on average, based on hourly page view statistics. However, [14] estimate lag time using the real-time article edit stream to be within 30 minutes, with major global news usually reflected in minutes (although initial edits are typically small and incremental). Worst-case scenarios are less studied. While anecdotal examples do not reflect overall timeliness, they do give insight into the temporal dynamics across all news sources.

Whitney Houston's death was first reported on Twitter by the niece of the hotel worker who found her at 00:15 UTC on the 12th February 2012 [18]. After spreading through Twitter, at 00:57 that the Associated Press verified and broke the news on their Twitter feed. The first edit to Whitney Houston's Wikipedia page to reference her death ("*has died*") was at 01:01 (UTC). High-frequency editing of unfolding details followed, citing available sources [18]. [14] uses the example of the resignation of Pope Benedict XVI, noting that the English and French Wikipedia articles were first edited at 10:58 and 11:00, respectively, which is impressive given that Reuters broke the news on their Twitter feed at 10:59, following the Vatican's public announcement at 10:57:47.

2.2 Topic Coverage

The Wikipedia project is available in 285[3] languages, with each having language/location-specific and translated articles. As of June 2013, Wikipedia English has over 4.2M articles, followed by Wikipedia Dutch, German and French, with 1.4-1.6M articles each. Although the large quantity of articles suggests extensive topic coverage, a 2008 study [8] quantified topic coverage by measuring the similarity of the topical distribution of articles on Wikipedia to the topical distribution of books in print, based on established library classification systems. A wide disparity was observed, with subjects such as science and geography better represented in Wikipedia than in books, and conversely, subjects such as medicine and literature much less represented in Wikipedia. Further work would need to extend this research to understand the nature of how events related to different topics (e.g. celebrity, politics, news, etc) are reflected over time in Wikipedia.

Event Coverage. Major predictable and unpredictable events typically have their own dedicated articles (e.g. '*39th G8 Summit*', and '*2013 North India Floods*'), with the most important events having multiple articles discussing different aspects (e.g. timeline, comparison to similar events, or reactions). Less prominent events (including those that occurred before Wikipedia began) may appear as a sentence or section in a related article, or be mentioned in the Current Events portal, with a brief summary referencing the date of the event and links associating entities.

Many mainstream recent international events (including sports events) appear in the Current Events portal[4], categorised by date and topic (e.g. ongoing events, deaths, conflicts, elections and trials). Archived versions provide a vast almanac of daily events since January 1900[5] (although earlier dates are more sparse and less structured).

Aside from content, the impact of an event that occurred during Wikipedia's lifetime may be understood through a number of temporal meta-data signals. For instance, increased article editing or article viewing frequency may suggest recent events. Additionally, the temporal link graph may contains bursts to other event-related articles. A combination of these signals with the content stream offers an informative view of what is happening, and it's impact over time.

2.3 Content Quality/Correctness

Many policies and guidelines govern Wikipedia editing in an attempt to maintain encyclopaedic consistency [1]. In essence, article content should be written from a neutral point of view, based on reliable sources and kept objective. A side-effect of Wikipedia's editing openness is that it sometimes leads to inaccurate reporting, deliberate vandalism or more subtle abuse [12]. The community reviewing mechanism often corrects obvious issues relatively quickly, with the aid of bots that watch recent changes and apply automated machine learning and heuristics to immediately flag issues (and occasionally instantly revert article revisions). High profile articles (e.g. celebrities, well-known politicians and currently prominent events) are often locked so that only administrators or established editors may change their content, reducing the volume of article edits, but ensuring accuracy. In the spirit of the fundamental policies and guidelines, article editing with current news must be backed by references, hence Wikipedia is not intended to be a primary source for news [1]. The 'Talk' page accompanying every article often contains commentary related to recent or necessary changes in the article, especially if there is concern about the content. Temporal discourse can be detected through the presence of discussion. An extensive discussion of the impact of Wikipedia editing policies on news reporting can be found in [9].

2.4 Comparison with Other Event Sources

Twitter has become popular for monitoring real-time events because of its immediacy and volume of citizen reporting with '*tweets*' about ongoing events. Compared to Wikipedia it poses challenges, including: the scale and volume of tweets, limited document size, slang vocabulary, misspellings and spam. Similarly, Twitter provides a 'soapbox' platform where user-generated content is often '*hearsay*', or non-neutral.

Twitter is undoubtedly an excellent source for detecting breaking new stories, especially instaneous event such as spreading earthquakes [13]. However for understanding events, the quantity and quality of conflicting and redundant information created by users discussing and speculating event-based topics makes it difficult to monitor and organise key event details. In contrast, news wires offer a stream of professionally curated, and in most cases, high quality news stories. Consequently, news agency articles may be shortly delayed due to the fact they need to be written and edited prior to publication.

Furthermore, Facebook has become an increasingly prominent means of sharing and discussing event-related information through user's social networks [3]. Vast information cascades caused by users sharing information, often reflecting temporal trends, have been extensively studied [5]. However, little work has publicly studied to what extent Facebook reflects events over time – most likely due to the difficulty of obtaining large-scale Facebook data.

Wikipedia is unlikely to reflect events as quickly as Twitter, except in the case of the most heavily discussed topics. However, Wikipedia trades a lag in reporting time for the sake of reporting accuracy. Likewise, in comparison to news wires, although Wikipedia doesn't have rigorous pre-publication editorial validation like well-respected publications, content can evolve quickly through editing. Users exposed to different media outlets aggregate and distill multi-

[3] http://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed: June 2013

[4] http://en.wikipedia.org/wiki/Portal:Current_events

[5] http://en.wikipedia.org/wiki/January_1900

ple sources of event detail into a single Wikipedia location through citations. Furthermore, Wikipedia offers unique structural characteristics for understanding many events. As it forms a linked knowledge base structure of articles, event details become encyclopaedic and hierarchically structured, as all related topics become associated.

3. TEMPORAL WIKIPEDIA SIGNALS

Events over time are reflected by many signals in Wikipedia, with combinations of signals providing informative clues of temporal details, related entities, sequencing and impact. Aside from understanding recent events, Wikipedia also serves as a valuable source of information on many past events.

In this section we discuss Wikipedia's main temporal dataset availability, concentrating on providing an overview of the scale, value and challenges of using these datasets rather than rigorous analysis. We use the 'Arab Spring' event to anecdotally illustrate the temporal data characteristics.

Data Availability. The majority of Wikipedia article text, structure and meta data (for all languages and public projects) is available under an open license, making it available for research purposes. A subset of data is available through real-time feeds, with full historic archives available to download. As many of the raw datasets are difficult to work with, we have made all the datasets distilled from the raw Wikipedia English dump (up to April 2013) which we refer to in this section available for download[6] (around 50GB uncompressed).

The Wikimedia dump mirrors[7] host raw XML, SQL and CSV-formatted datasets. The main temporal dataset: '*All pages with complete edit history*' monthly dump, was used to create many of the datasets discussed in this section. De-compressed, this dump is about 7TB of XML-formatted edit revision history for all articles, starting January 2003. Real-time monitoring of Wikipedia article creation and edit activity is provided through an internet relay chat (IRC) channel and syndication feeds. The Wikipedia application programming interface (API) can provide occasional access to specific article data. The IRC channels report almost all activity along with meta-data (e.g. article name, author, etc), however does not include any text change detail. Syndication feeds reflect only a subset of changes (≈30-50%, shown in Figure 1), but do include `diff` snippets of changed text. Hourly statistics of article views is available from December 2007.

3.1 Temporal Expressions

Temporal taggers extract and resolve absolute (e.g. 'Mon 12th May 2013') and relative (e.g. 'yesterday') temporal expressions contained in text [15]. Using a rudimentary temporal tagger, we extracted all the `YEAR`, `MONTH-YEAR` and `DAY-MONTH-YEAR` temporal expressions from the current article revision text. 2001-09-11, 2000-09-24, 1960-09-24, 1999-09-09 and 1999-12-07 are the most mentioned dates, with 4,387, 2,064, 1,717, 1,525 and 1,497 appearances, respectively. Similarly, Jan 2011, Jan 2010, Jan 2009, Jun 2009 and May 2010 are the most mentioned months, with 17,791, 17,789, 17,418, 16,898 and 16,479 appearances, respectively. The distribution of year mentions are presented in Figure 2 (note the spike for 2001, and also the presence of future dates). Current dates in recent article changes strongly indicate ongoing events [7].

[6] `http://www.stewh.com/wiki-datasets`
[7] `http://dumps.wikimedia.org`

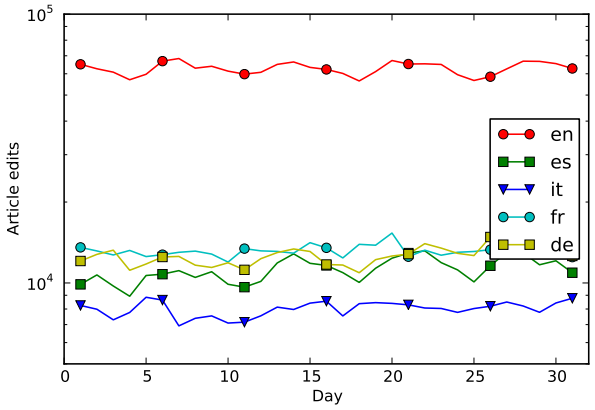

Figure 1: May 2013 daily article changes syndication feed volume (in logarithmic scale) for Wikipedia English (*en*), French (*fr*), Italian (*it*), German (*de*) and Spanish (*es*).

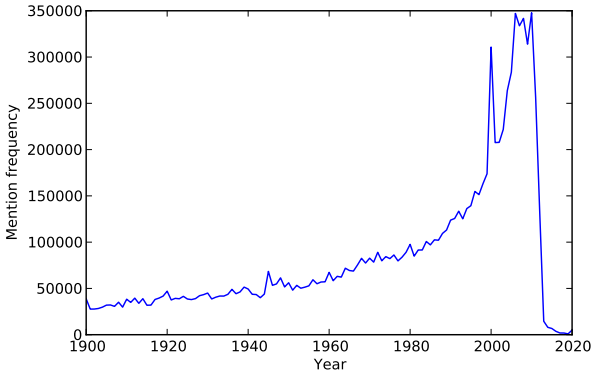

Figure 2: Year mentions in Wikipedia English from 1900 to 2020.

3.2 Temporal Link Graph

Wikipedia has a vast and amorphous link graph created by user's linking together entities appearing in text and citing external references. Intra-Wikipedia links refer to other articles in the Wikipedia project, cross-lingual article links (prefixed with '*fr:*', etc.), or to media (prefixed with '*image:*', etc.). In Figure 3 we present the cumulative temporal degree of in- and out-links for the 'Arab Spring' article. Many articles have a significant number of footnote/citation links to external web pages. In the case of events, many of these refer to major news outlet articles.

Links created over time can be extracted by parsing new `[link[|name]]` markup in each article revision. In many cases, editors create links that refer to a synonym for the actual article name. Extracting the article redirect pages (e.g. '*Barak Obama*' to '*Barack Obama*') allows link graph synonyms to be resolved correctly to the final article (note that redirects to other redirects are not permitted). In our dump we have also extracted the article section (if any) each out-link is contained in.

3.3 Page Edit Stream

Every Wikipedia page (including articles, talk and meta pages) has a full revision history available, including reverted (e.g. vandalised or unacceptable) versions. Care is needed to determine ac-

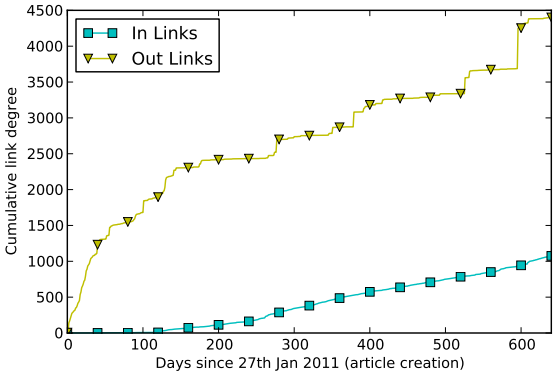

Figure 3: Cumulative 'Arab Spring' article in- and out-link degree since 27th January 2011.

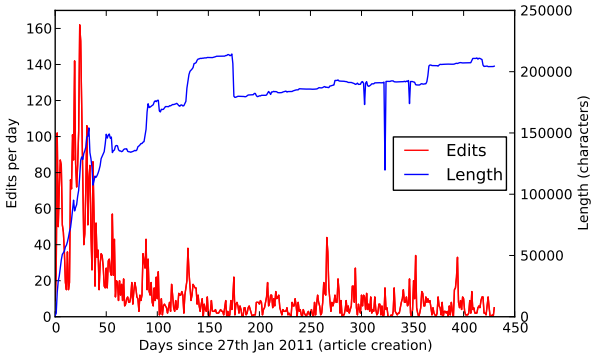

Figure 4: 'Arab Spring' daily article edit frequency and length (in characters) since 27th January 2011 (to 23rd March 2012).

cepted revisions in the revision history. Comments accompanying each revision often specify whether the revision is the result of a revert with revision reference (along with a reason). We filtered out articles by excluding pages which included an organisational namespace (e.g. '*Talk:*').

Many signals can be extracted from the raw article revision history, using markup for structure. A simple `diff` operation between the text of two revisions reveals changed text. The temporal editing activity and length for the 'Arab Spring' article is shown in Figure 4. A `patch` operation resolves changes to character locations, allowing resolution to the hierarchical section in which they happened (sections will also evolve over time). Section change activity is illustrated for the 'Arab Spring' article in Figure 5.

3.4 Current Events Portal

The curated Wikipedia Current Events portal includes major and mainstream international news, sports news and deaths. Each item links to the related entities, or a specific event page if available, and are categorised into a coarse taxonomy. This source of events has been exploited for Twitter-based TDT evaluation [10]. In May 2013, 475 news articles were presented in total, an average of 15.3 ($\pm$5.5) articles per day. 477 deaths were also included. Further work is needed to characterise coverage and speed of this source, as items have to be nominated before acceptance and inclusion.

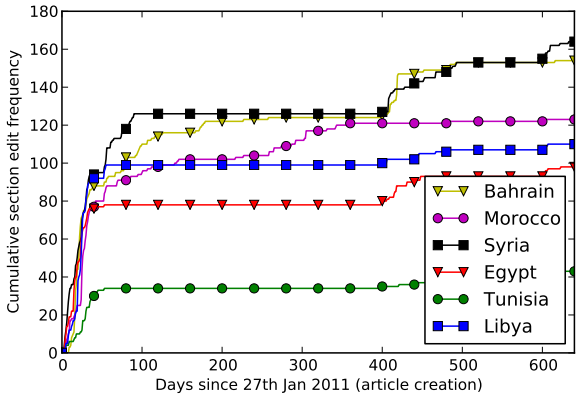

Figure 5: Cumulative 'Arab Spring' article section edit frequency since 27th January 2011.

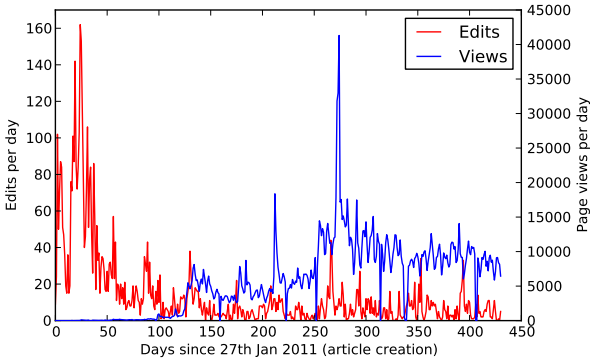

Figure 6: 'Arab Spring' article daily edit frequency and page views since 27th January 2011 (to 23rd March 2012).

3.5 Page View Stream

Article page views are a direct visitor-driven measure of a topic's temporal popularity in Wikipedia. In Figure 6 we present the edits and page views per day for the 'Arab Spring' article. Interestingly, page views are not always strongly correlated with increased editing activity.

The latest data is available with a one hour lag, however there is occasionally corruption or empty data, for which smoothing or extrapolation can be helpful. Individual article views can be visualised and downloaded in JSON format using a 3rd-party tool[8].

4. CONCLUSION

In summary, Wikipedia is undoubtedly an extensive, valuable and accessible source of temporal signals for time-aware research. Most pertinently, it has: (1) historic and real-time data availability, (2) large-scale topic coverage, (3) multi-lingual (international) versions, (4) quick collaborative/iterative event reflection, (5) a vast evolving link graph, and (6) multiple rich levels of structure (e.g. article sections, taxonomy and lists) which also evolve over time.

While some of Wikipedia's characteristics make it challenging to work with, it has many structural aspects that offer temporal research prospects beyond what is readily available in many other event sources, such as Twitter and news streams.

[8] `http://stats.grok.se/`

5. REFERENCES

[1] Wikipedia: Wikipedia is not a newspaper. http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_newspaper.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Research and Development in Information Retrieval*, pages 37–45, 1998.

[3] B. Baresch, L. Knight, D. Harp, and C. Yaschur. Friends who choose your news: An analysis of content links on facebook. In *ISOJ: The Official Research Journal of International Symposium on Online Journalism, Austin, TX*, volume 1, 2011.

[4] M. Ciglan and K. Nørvåg. Wikipop: personalized event detection system based on wikipedia page view statistics. In *CIKM '10*, pages 1931–1932, 2010.

[5] P. A. Dow, L. A. Adamic, and A. Friggeri. The anatomy of large facebook cascades. In *ICWSM*, 2013.

[6] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer. Extracting event-related information from article updates in wikipedia. In *ECIR '13*, pages 254–266, 2013.

[7] M. Georgescu, D. D. Pham, N. Kanhabua, S. Zerr, S. Siersdorfer, and W. Nejdl. Temporal summarization of event-related updates in wikipedia. WWW '13 Companion, pages 281–284, 2013.

[8] A. Halavais and D. Lackaff. An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440, 2008.

[9] B. Keegan, D. Gergle, and N. Contractor. Hot off the wiki: Structures and dynamics of wikipedia's coverage of breaking news events. *American Behavioral Scientist*, 2013.

[10] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. Building a large-scale corpus for evaluating event detection on twitter. CIKM '13, pages 409–418, New York, NY, USA, 2013. ACM.

[11] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First Story Detection using Twitter and Wikipedia. *SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012)*, 2012.

[12] M. Potthast, B. Stein, and R. Gerling. Automatic vandalism detection in wikipedia. In *ECIR*, pages 663–668, 2008.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[14] T. Steiner, S. van Hooland, and E. Summers. Mj no more: using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection. WWW '13 Companion, pages 791–794, 2013.

[15] J. Strötgen and M. Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 2012.

[16] F. Vis. Wikinews reporting of hurricane katrina. In *Citizen Journalism: Global Perspectives*, Global Crises and the Media. Peter Lang, 2009.

[17] M. Wattenberg, F. B. Viégas, and K. Hollenbach. Visualizing activity on wikipedia with chromograms. INTERACT '07, pages 272–287, Berlin, Heidelberg, 2007. Springer-Verlag.

[18] S. Whiting, K. Zhou, J. Jose, O. Alonso, and T. Leelanupab. Crowdtiles: presenting crowd-based information for event-driven information needs. CIKM '12, pages 2698–2700, New York, NY, USA, 2012. ACM.

[19] S. Whiting, K. Zhou, and J. M. Jose. Temporal variance of intents in multi-faceted event-driven information needs. SIGIR '13. ACM, 2013.

[20] K. Zhou, S. Whiting, J. M. Jose, and M. Lalmas. The impact of temporal intent variability on diversity evaluation. ECIR '13, pages 820–823, Berlin, Heidelberg, 2013. Springer-Verlag.