Detecting Filter Bubbles in Ongoing News Stories

Giang Binh Tran[1], Eelco Herder[1]

L3S Research Center, Hannover, Germany
gtran,herder@L3S.de

Abstract. In this paper, we analyze differences in perspective between timelines created by news agencies from various countries. By employing methods for date and headline selection, which have been extensively evaluated in previous work, we show several types of bias that exist in the media landscape. As users typically select only a small number of news sources to follow, they necessarily experience at least some 'tunnel vision', which is commonly associated with the so-called 'filter bubble'. By recognizing and emphasizing the peculiarities of the users' self-selected news sources, we can help them to break out of the bubble.

1 Introduction

One of the tasks of a journalist is to monitor, gather, curate and contextualize the relevant information for the target audience. He needs to go through an enormous amount of records with information of very diverse degrees of granularity, in order to put information into context and tell his story from all significant angles, and, at the same time, he needs to reduce the noise of irrelevant content.

Many newspapers regularly publish manually created timelines, which allow the reader to gain a quick overview over events that span a longer time period - such as the Egypt revolution - and to answer questions such as: how and when did the event start? What were the main consequences of the initial events? What happened to the main protagonists in the event?

Though convenient for the reader, creating a timeline is expensive, as it requires substantial expertise. Moreover, creating a timeline is a very subjective task and therefore huge differences can be found in expert-generated timelines, even if they are on the same topic [3]. In addition, there exist different models of reporting, in the Western world varying from the Anglo-Saxon tradition of just reporting the facts, via the Northern-European model of combining news reporting with opinions, to the 'polarized Mediterranean' type of reporting [2]. Moreover, it is a known fact that newspapers in different countries provide different perspectives and points of focus, reflecting differences in national ideologies, priorities and opinions [5].

On the Internet, the filter bubble, as coined and popularized by Eli Pariser [4], is a well-known phenomenon that explains why personalization leads users to mainly encounter products, news articles and other types of content that match the users' own interests and viewpoints. The tunnel vision that the filter bubble is said to create, is not just a recent online phenomenon, but inherent to journalism: editors need to be selective and therefore necessarily focus on matters that are of highest interest to their readers.

These are some of the reasons why summaries of events, such as timelines, differ wildly, depending on the nationality, audience, political viewpoints and other sources of

subjectivity associated with a newspaper. At the same time, readers are known to select a small number of newspapers - or even just one - that best match their personality, viewpoints and interests. As argued by Paul Resnick in his keynote at UMAP 2011[1], emphasizing differences in perspective is an important instrument for allowing users to break out of their 'filter bubbles'.

In this paper, we investigate methods for automatically recognizing and emphasizing differences between various timelines, and evaluate them by comparing timelines from well-known news agencies and newspapers. These methods are based on our previous work on date and headline selection for news event summarization; by employing these methods and the Shannon Diversity Index, we are able to recognize dates that were considered important in one timeline but not in others, and differences in news coverage on dates that were considered important in all timelines.

The work is carried out in the context of the EUMSSI[2] project, in which cross-modal analysis techniques are developed for analysing news articles, videos and audio reports. This will allow journalists - and media consumers - to relate these messages with one another and to understand the underlying events.

2 Related Work

Many studies specific to timeline summarization, such as [7, 10, 8], focus on the extraction of salient sentences or headlines for generating the textual content of timelines. They assume either that the dates are given in advance or they use simple measures such as burstiness for date selection [1].

Prior approaches dedicated specifically to date selection include work by Tran et al. [8] and Kessler et al. [3]. They use supervised methods that score dates independently of each other, making use of frequency-based, temporal and topical features that are extracted from a corpus of event-related newspaper articles. We, however, score dates jointly, making use of interactions between dates in a graphical model.

Additionally, by using the Shannon diversity index [6] on the top of the graphical model, our approach can highlight dates on which important events happened, but that are likely to be ignored by many news agencies. This aspect has not been considered so far in the previous work.

3 Approach

In order to recognize and emphasize differences between timelines, it is important to first find events and corresponding dates that are 'subjective', in other words dates that are not considered important by all news agencies.

Our approach for subjective date detection makes use of a corpus of news articles (more details in Section 4) and consists of two steps. First, we use a random walk model, which we proposed in [9], to rank the dates based on their importance with regard to their impact on the future events[3]. After that, we re-rank the top selected dates based on

[1] https://presnick.wordpress.com/2011/07/17/personalized-filters-yes-bubbles-no/

[2] http://www.eumssi.eu/

[3] A demo of the WikiTimes system for automatic timeline creation is available at http://wikitimes.l3s.de/

their Shannon Diversity Index (SDI) scores. SDI is widely used in ecology and biology to measure the diversity of species in a community [6]. We use the SDI to express the rarity and commonness of events, as reported by different news agencies. When an (important) event is commonly reported by many news agencies, it is less subjective and thus less likely to be filtered out. For this reason, our approach here is to give a high rank to the date that has a low SDI score.

Formally, in the first step, we build a *date reference graph*, which is a *fully directed graph* $\mathbf{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of dates mentioned in any text in corpus C, including publication dates. The edges $\mathcal{E} = \{e(d_i, d_j)\}$ indicate that at least one text published on d_i refers to the date d_j.

We represent each link between events as a multi-value tuple

$$e(d_i, d_j) = (M_{ij}, freq(d_i, d_j), I_{temporal}(d_i, d_j), I_{topical}(d_i, d_j))$$

to integrate different measures of date importance. The first value, $M_{ij} = \frac{1}{N}$ expresses the prior transition probability between 2 dates where $N = |\mathcal{V}|$. The other values express the strength of the connection between d_i and d_j, modeled by the following aspects: frequency ($freq$), temporal influence ($I_{temporal}$) and topical influence ($I_{topical}$). Our random walk model uses these perspectives to rank the collection of dates. The intuition is that when a date d_j is referred to from either a past or future news article (published on d_i), it is likely involved in the events that are reported in that article.

In the second step, we compute the SDI for the top-K ranked dates, based on the distribution of news articles published on those dates. For example, a date that contains only news articles from the BBC is considered less diverse than a date that contains news articles from several large news agencies over the world. The computation of SDI is sketched as follows:

$$SDI(d_i) = -\sum_{i=1}^{R} p_i \ln p_i$$

where R is the number of news agencies from which we collected data and p_i is the proportion of news articles from news agencies p_i. That measure quantifies uncertainty in predicting the agency identity of a news article that is taken randomly from the dataset, hence, suggesting how subjective (or non-diversified) a date is.

4 Experiment

We present our preliminary results of two experiments on detecting subjective dates in the *Crisis data*[4] dataset, which contains ground truth (**GT**) timelines (written by professional journalists) as well as a corpus of around 12K news articles that cover events that happened in the context of four news stories: Egypt Revolution, Syria War, Libya War and Yemen Crisis. The dataset is suitable for our purpose for the following reasons: (1) it is a heterogeneous dataset that contains news articles and expert timeline summaries from 25 well-known news agencies and; (2) it covers long-term stories that have been happening since 2011, making the date selection problem non-trivial for any system.

[4] The Crisis data is currently available at `http://l3s.de/~gtran/timeline/`

4.1 Experiment 1: Subjective Dates Selected by One News Agency

In this experiment, we aim to detect dates that have been included in the timeline of only one news agency (and not in other timelines). We use a cut-off K = 50 to select the 50 most important dates, using our random walk model, and rank them by their SDI score. We then compare these selected dates with those in our ground-truth timelines. The performance of the subjective date detection process is measured as the proportion of the top-10 dates (this can be extended to a larger number) that are included in *exactly one* timeline.

The result is described in Table 1, which shows that a significant number of important dates and their events are reported by only one news agency. We took a closer look into those events (see examples in Table 2) and found evidence that the events that are included in only one timeline are typically not about the main theme of the ongoing news stories, but about related aspects, such as business and human rights. As we discussed in [9], timelines often contain substories or side-paths that involve major actors of the main story. Due to space limitations, journalists can only incorporate a limited number of substories and the decision which stories to incorporate is often subjective.

	Proportion (%)
Egypt	90%
Libya	50%
Syria	70%
Yemen	50%

Table 1: Performance of the *subjective date* detection process

	Date	Event
Egypt	29-12-2012	The Egyptian Central Bank announces that foreign reserves drained to $15 billion from $36 billion in 2010 have fallen to a " critical minimum " and starts new measures to stop a sharp slide in the value of the Egyptian pound . The pound 's decline slows but doesn't stop , and now stands at just over 7 to the dollar , compared to as strong as 5.5 to the dollar in 2010 . ... (AP)
Libya	18-08-2011	Libyan Foreign Minister Moussa Koussa says the country has decided on " an immediate cease-fire and the stoppage of all military operations . " But sources inside Libya say violence continues . ....(CNN)
Syria	14-08-2012	More than 23,000 people have been killed since the outbreak of the revolt , the Syrian Observatory for Human Rights says . Former prime minister Riad Hijab , who defected on August 5 , says that the regime is collapsing . Government forces concentrate their operations on the two main cities Damascus and Aleppo . ...(DailyStar)
Yemen	18-09-2011	Security forces open fire on tens of thousands of demonstrators in Yemen's capital , Sana , killing at least 26 protesters in one of the bloodiest days of the 9-month-old rebellion against President Ali Abdullah Saleh ...(L.A.Times)

Table 2: Example *subjective dates* and the corresponding events, which are included in only timeline in our ground truth.

4.2 Experiment 2: Important Dates that Are Not Covered

In this experiment, we aim to find out how many important dates are not included in any of the news agencies' timelines. Similar to the previous experiment, we picked the top-50 important dates from the results of our random walk algorithm and calculated the proportion of dates that are not mentioned in any timeline. The results are shown in Table 3. The high number of 'missed dates' does not mean, however, that the ground truth timelines are of poor quality: as explained earlier, journalists need to be selective when creating timelines. The SDI values also confirm that journalists usually do a good job in selecting the dates to be included: for all stories, except for Libya, the average SDI of dates included in at least one timeline is higher than the average SDI of dates that did not make it into any of the timelines. In other words, dates that are included in at least one timeline are typically covered by more news agencies than dates that do not appear in any timeline.

	Proportion (%)	SDI included	SDI excluded
Egypt	62%	2.47	1.52
Libya	24%	2.34	2.25
Syria	74%	1.65	1.70
Yemen	56%	2.10	1.70

Table 3: Proportion of important dates that are not included in any timeline, along with the average SDI of included and not-included dates

To better understand which events were left out of the timelines, we analyzed several of those dates with a high SDI score but that were not included. Some example events are shown in Table 4. Except perhaps for the Libya article, these are events that readers may be interested in, but that they will not be able to find in the timelines of the news agencies that they are subscribed to.

	Date	Event
Egypt	11-02-2012	Three people including an Australian journalist, an American student and their Egyptian guide are arrested in the city of El-Mahalla El-Kubra for allegedly offering inducements to people to join. Labour activist Kamal al-Fayyumi is also arrested in El-Mahalla El-Kubra...
Libya	12-06-2012	Libyan leader Muammar Gaddafi plays chess with Kirsan Ilyumzhinov , the president of the international chess federation....
Syria	08-03-2012	Syrian dissidents reject a call by Kofi Annan to stop fighting and seek peace talks...
Yemen	29-11-2011	Opposition candidate Mohamed Salem Basindwah was chosen to lead the national unity government , and pledged on Nov. 29 to tackle issues including fuel shortages...

Table 4: Example of dates and corresponding events that did not make it to any of our ground-truth timelines.

5 Discussion and Conclusions

In this paper, we have shown that it is possible to automatically find dates and corresponding events in news stories that are likely to be covered in only one or a subset of manually generated timelines. Further, we have shown that there are important dates that are not incorporated in any timeline.

As we argued earlier, it is unavoidable that timelines are selective and, to a certain extent, subjective. Still, it would be very desirable to make readers aware of the peculiarities of the timeline that they currently inspect. This can be achieved by highlighting dates with large differences in coverage between news agencies or timelines, by adding links to other timelines for dates that are not incorporated in the current timeline, or, alternatively, by constructing an annotated 'timeline of timelines' for those who wish to contrast the bias in their personal selection of news sources with news sources that they usually do not visit.

Conversely, journalists will be able to create better balanced - or, alternatively, even more argumentative - timelines with feedback on their current selection of dates and events. Even though - as we have shown in previous work - it is possible to automatically construct timelines by selecting the most relevant dates and headlines, still manual processing and editing would be needed to enhance the communicative qualities of the timeline, and to adapt it to the needs of the readers.

6 Acknowledgments

The work was partially funded by the European Commission for the FP7 project EUMSSI (611057)

References

1. H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of SIGIR'04*, pages 425–432, 2004.
2. F. Esser and A. Umbricht. Competing models of journalism? political affairs coverage in us, british, german, swiss, french and italian newspapers. *Journalism*, 14(8):989–1007, 2013.
3. R. Kessler, X. Tannier, C. Hagege, V. Moriceau, and A. Bittar. Finding salient dates for building thematic timelines. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 730–739. Association for Computational Linguistics, 2012.
4. E. Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.
5. S. Seo. Hallidayean transitivity analysis: The battle for tripoli in the contrasting headlines of two national newspapers. *Discourse & Society*, 24(6):774–791, 2013.
6. C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, Jan. 2001.
7. R. C. Swan and J. Allan. Timemine: visualizing automatically constructed timelines. In *SIGIR*, page 393, 2000.
8. B. G. Tran, M. Alrifai, and D. Quoc Nguyen. Predicting relevant news events for timeline summaries. In *WWW*, 2013.
9. G. Tran, E. Herder, and K. Markert. Joint graphical models for date selection in timeline summarization. In *Proceedings of ACL*, 2015.
10. 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 SIGIR'11*, pages 745–754, 2011.