Analysis of Temporal and Web Site References in History-related Tweets

Yasunobu Sumikawa
Dept. of Information Sciences,
Tokyo University of Science
ysumikawa@acm.org

Adam Jatowt
Dept. of Social Informatics,
Kyoto University
adam@dl.kuis.kyoto-u.ac.jp

Marten Düring
Institute for Contemporary History,
University of Luxembourg
marten.during@uni.lu

ABSTRACT

Having good knowledge and comprehension of history is believed to be important for a variety of reasons. In this paper we report the initial results of an exploratory analysis of history-focused references in microblogs based on 11-months' long snapshot of Twitter data. In particular we focus on the time periods referred in history-related tweets and on web sites shared in such tweets.

CCS CONCEPTS

•**Information systems** →**Specialized information retrieval;**

KEYWORDS

social media analysis; history; collective memory; Twitter

1 INTRODUCTION

Social media has been commonly utilized to study public attitudes towards current events such as the US American elections [7]. Yet, similarly to other media, microblogs are also used for sharing information related to the past, sometimes to the distant past. This offers unique opportunities for computational studies on explicit references to past events and opens up novel perspectives for the analysis of collective memories as well as for the pursuit of public history.

While computational studies of collective memory have already been conducted on news articles [1] and Wikipedia [3–5], with regard to microblogging, still little investigation has been done. For example, the authors of [2] analyzed tweets related to the First World War and compared commemorative cultures across different countries. We then try to fill in this gap by focusing on Twitter as a common social media platform that is also frequently used in the computational social science. Our analysis has exploratory character and aims to provide initial investigation of selected aspects about history-related content sharing in social networks.

2 DATA COLLECTION

To collect tweets that refer to the past or are related to collective memory of historical events/entities, we performed hashtag based crawl together with bootstrapping procedure. At the beginning, we gathered several historical hashtags selected by experts (e.g. `#history`, `#HistoryTeacher`, `#WmnHist`) [1]. In addition, we prepared several hashtags that are commonly used when referring to the past: `#throwbackthursday`, `#historicalevent`, `#onthisday`, `#otd`, `#thisdayinhistory`. We then collected tweets that contain these hashtags by using Twitter's official API[2].

The collected tweets were issued from 8 March 2016 to 24 February 2017. To increase the coverage, during this period, we searched for other hashtags that are frequently used with the seed hashtags by applying bootstrapping. These were iteratively included into the seed set after manual inspection as for their relation to the past and the history. In total, we gathered 147 history-related hashtags which allowed us to collect 888,251 tweets.

3 ANALYSIS

Temporal Reference Analysis. We first analyze which time periods users are interested in. For this we extract *time-references* from tweet content. We use both absolute and relative temporal references. To extract both types of time references, we use Heideltime [6], which is a temporal tagger with a specialized option for tweet processing. We then convert all relative temporal expressions to the absolute ones. In total, we have found over 357k tweets with temporal expressions.

We next map all the extracted temporal expressions on timeline as shown in Fig. 1 to reflect the strength of the collective attention of users towards different time periods of history. To plot the remembering curve, we convert the extracted temporal references to probability distributions over their corresponding timespans using year level granularity. We then combine all the probability distributions for every year.

Looking at Fig. 1 (especially, at the zoomed out plot in the inner graph), we can observe that the number of time references is rapidly increasing towards the present time. In general, the recent past matters more than the distant one and the memory decay is fastest in the recent years. This is intuitive and correlates with the corresponding study conducted on news articles about several countries [1]. When comparing the tweet line (blue color) and retweet line (green color) we also notice relatively strong difference between the two over the course of the recent years. For the distant past however the two lines have more similar values.

There are also *several significant peaks visible in Fig. 1 which represent two key events in the last century: WWI and WWII*, and the last year (2016). Two dates common for WWII are: 1941 denoting the Pearl Harbor attack and the subsequent participation of USA in the war, and 1945 which is related to the Normandy landing and the end of the war. Fig. 2 shows the most common hashtags used with content containing the peak years found in Fig. 1. We can notice that 1916 and 1941, 1945 have strong connection with hashtags #ww1 and #wwii as the two events were held during these respective years. Interestingly, Fig. 2 shows there are many mentions of 2016 with #ww1. This is because 2016 marked the 100th anniversary of the Battle of Verdun which is well-remembered due to the estimated 1 million casualties.

Finally, #otd and #onthisday as shown in Fig. 2 are commonly used labels for indicating historical content that occurred on the same calendar day in the past. Past-to-present connection by the calendar day is in fact a popular way of recalling past events and is typically used in newspapers (e.g., "Events reported in our newspaper on this day in the past"). #otd and #onthisday are hashtag-based mechanisms for indicating this type of connection in Twitter.
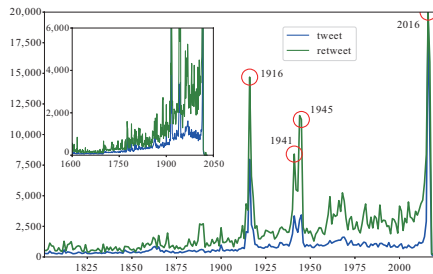

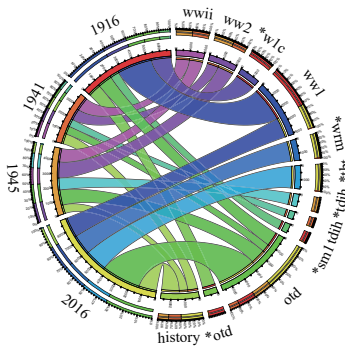**Figure 1: Distribution of time references in tweets. The peaks are in years 1916, 1941, 1945 and 2016.**

Figure 2: Top hashtags associated with the four peaks of Fig. 1. "*" is used to indicate abbreviations made for saving space (*w1c: #ww1centenary, *wrm: #weremember, *tbt: #throwbackthursday, *tdih: #thisdayinhistory, *sm1: #somme100, *otd: #onthisday).

URL Analysis. The past is commonly remembered by making use of diverse artifacts such as images, videos, or historical documents. Indeed, many web services offer such kinds of data, and

approximately half of tweets in our dataset include links to some kind of web services. We analyze then what kind of external data users refer to when they post history-related tweets. Fig. 3 lists the top 20 websites mentioned in the tweets in our dataset[3]. For obtaining the original URLs, we used twitter-text-python library[4]. As expected, *users often enrich tweets with images (Instagram) or videos (YouTube and vine).* In addition, users tend to sometimes link to shopping sites (e.g., Amazon, eBay) using past-related hashtags. Closer manual investigation of such tweets revealed that the posted links refer to history related items such as movies about historical events or entities.
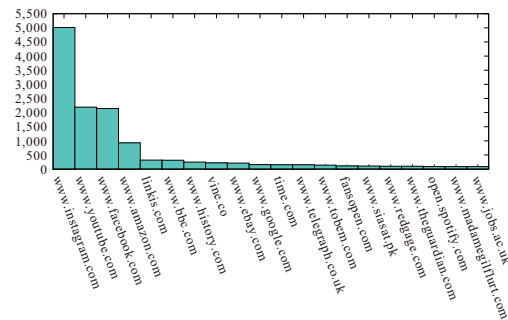

Figure 3: Top 20 web sites referred in tweets.

4 CONCLUSIONS

In this paper we have performed initial analysis of collective memory in Twitter focusing on temporal expressions mentioned in tweets which are related to the past and on associated URLs. Future work will shed more light on the way in which history-related content is used and shared in microblogging to encourage research and development of systems aiming at educating history.

REFERENCES

[1] Ching-man Au Yeung and Adam Jatowt. 2011. Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining. In *Proceedings of CIKM2011*. ACM, 1231–1240.
[2] Frdric Clavert, Benot Majerus, and Nicolas Beaupr. #ww1. Twitter, the Centenary of the First World War and the Historian. https://halshs.archives-ouvertes.fr/halshs-01148548/document
[3] Michela Ferron and Paolo Massa. 2011. Collective Memory Building in Wikipedia: The Case of North African Uprisings. In *Proceedings of WikiSym2011 (WikiSym '11)*. ACM, 114–123.
[4] Adam Jatowt, Daisuke Kawai, and Katsumi Tanaka. 2016. Digital History Meets Wikipedia: Analyzing Historical Persons in Wikipedia. In *Proceedings JCDL2016 (JCDL '16)*. 17–26.
[5] Nattiya Kanhabua, Tu Ngoc Nguyen, and Claudia Niederée. 2014. What Triggers Human Remembering of Events?: A Large-scale Analysis of Catalysts for Collective Memory in Wikipedia. In *Proceedings of JCDL2014 (JCDL '14)*. IEEE Press, 341–350.
[6] Erdal Kuzey, Jannik Strötgen, Vinay Setty, and Gerhard Weikum. 2016. Temponym Tagging: Temporal Scopes for Textual Phrases. In *Proceedings of WWW2016 (WWW '16 Companion)*. 841–842.
[7] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.. In *Proceedings of ICWSM 2010*.

---

[3] We have removed twitter.com focusing on external websites.
[4] https://github.com/edburnett/twitter-text-python