Digital History meets Microblogging: Analyzing Collective Memories in Twitter

Yasunobu Sumikawa*
University Education Center,
Tokyo Metropolitan University
ysumikawa@acm.org

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

Marten Düring
Luxembourg Centre for
Contemporary and Digital History
marten.during@uni.lu

ABSTRACT

Microblogging platforms can offer good opportunities to study how and when people refer to the past, in which context such references appear and what purposes they serve. However, this area still remains unexplored. In this paper we report the results of a large scale exploratory analysis of history-focused references in microblogs based on 11-months long snapshot of Twitter data. Besides understanding the nature of history-focused content sharing in microblogs, the results of this study can be used for designing content recommendation systems and could help to improve time aware search applications.

KEYWORDS

social media analysis; history; collective memory; Twitter

ACM Reference Format:
Yasunobu Sumikawa, Adam Jatowt, and Marten Düring. 2018. Digital History meets Microblogging: Analyzing Collective Memories in Twitter. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3197026.3197057

1 INTRODUCTION

Having good knowledge and comprehension of history is important for a variety of reasons: History is commonly believed to play significant roles in our society: First, to help us understand the processes which shape the present and thereby enable us to actively take part in contemporary society. Second, to be the basis for the development of coherent identities. Third, as a means to give meaning and orientation with regard to the past. Others argue that history opens up perspectives for the future and provides support for decision making [1, 10]. As such, history remains one of the fundamental subjects taught from elementary schools onwards.

Social media has been commonly utilized to study public attitudes towards real time events such as the US American elections [19]. Yet, similarly to other media, microblogs are also used for sharing and finding information related to the past, sometimes to distant past. This offers unique opportunities for computational studies on explicit references to past events and opens up novel perspectives for the study of collective memories as well as the pursuit of public history.

While computational collective memory studies have already been conducted on news articles [2, 6] and Wikipedia [8, 9, 15, 16], with regard to microblogging, so far we are only aware of one ongoing project which focuses on the First World War [5] and compares commemorative cultures across countries. We then attempt at filling this gap by focusing on Twitter as a common social media platform frequently used in the computational social science. Our analysis has exploratory character and aims to provide initial and broad investigation of history-related content sharing in social networks.

Among others our study is guided by the following questions:

(1) How do people refer to history in microblogs?
(2) What is the time horizon of history-related references?
(3) How are collective memories expressed in Twitter?
(4) What are the key remembered events and entities?

We approach these and other questions by investigating portions of tweet messages generated from March 2016 to February 2017 and tagged with hashtags which indicate a strong relation to historical events. In particular, we apply a bootstrapping approach to discover related hashtags, which allowed us to collect close to 1 million tweets with explicit references to past events and entities.

Based on the collected data we investigate the distinguishing characteristics of tweets which are related to the past. Among the aspects we research are time horizons, concerned entities and popularities of hashtags. Furthermore, we propose a novel categorization scheme of hashtags as well as perform deeper investigation of the characteristics of the hashtag categories (Sec. 5). By this we try to organize and provide structure to the portion of user activities that relate to referencing, appreciating and sharing information on historical events and entities in social network services.

Besides being a novel approach towards the general question on how the past relates to our lives, our investigations can be beneficial to several applications. First, the *construction of specialized content detection and recommendation systems* can be informed by the reported results. The objective of such systems would be to facilitate sharing of historical knowledge. Historical content recommendation based on social media offers an interesting and informal environment for learning history. Understanding the types of popular shared content and the context of sharing can be helpful for designing effective recommendation systems. Indeed several projects already use social platforms like Twitter to stimulate interest in history and for teaching exercises [1].

*This research was performed while the first author was at Tokyo University of Science, Japan.

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06…$15.00
https://doi.org/10.1145/3197026.3197057

[1] https://twitter.com/RealTimeWWII, https://twitter.com/civilwarwp, https://twitter.com/1948War, https://twitter.com/samuelpepys

Automatic content creation, especially, in the form of *conversing history-focused chatbots* could be another appealing idea. Tweets thanks to their short content and the effective means of measuring their popularity or attractiveness (e.g., retweet counts and analysis of user responses) are useful source of data for such systems.

Many history- and memory-focused tweets are triggered by current events or entities. Studying their formation and fluctuations of their popularity could be useful for understanding the conditions and circumstances for enabling "*historification*" of present texts. This means offering *access to relevant historical references and grounding for the present events and topics.*

To sum up we make the following contributions in this paper:

(1) To the best of our knowledge we are the first to study how users refer to history in microblogging based on large scale analysis.
(2) We provide novel findings which offer a better understanding of how collective memories are maintained and formed in microblogging.
(3) We propose new categorization of historical references in Twitter.
(4) We outline novel research directions and potential applications that can utilize history-related content in microblogs.

The remainder of this paper is structured as follows. In the next section we present related work. In Section 3 we describe the data collection and processing. Section 4 provides the findings of the general analysis, while the following section proposes our novel categorization of hashtags and details the results of the related study. We then include discussions in the Section 6. Finally, the last section concludes the paper and outlines future work.

2 RELATED WORK

The concept of *collective memory* (*social memory*) popularized by Halbwachs [11, 12] describes the shared reflection of the past within social groups. Collective memory can be contrasted with the concept of *collective amnesia* defined by Jacoby [14] as forceful or unconscious suppressions of memories, especially, those related to disgraceful or inconvenient events. In a similar fashion to personal memory [7], social memory is known to thin out over time and to be subject to temporal variations following the occurrence of memory triggers such as sudden events or anniversaries [2, 16, 17]. Studies of collective memory can help us to understand the mechanisms of forgetting and remembering as well as explain the role of the history and the past in our lives. In addition, they have direct implications on the archival selection by memory institutions such as national or dedicated archives [17]. Traditionally, research on collective memory has been based on small-scale investigations of personal accounts and the activities of political and cultural institutions. There is little literature on the use of computational approaches for the quantification of the characteristics of social memory over large text datasets. Cook *et al.* [6] investigated the decay of fame over time on the basis of the collection of news articles that spans 20th century. Au Yeung and Jatowt [2] have studied memory decay and the way in which past years are remembered based on the dataset of English news articles spanning 90 years. When it comes to other document genres, Ferron and Massa [8] and Kanhabua *et al.* [16] proposed to use Wikipedia as a global memory space. However, differently to our work they focused on memory triggers that cause forgotten or vaguely remembered events to be brought back into social attention. Anniversaries are natural examples of memory triggers. In another case, current events may also serve as triggers of the memories of similar, past events. Our work can be seen as complementary to the above mentioned researches.

To the best of our knowledge only one work focuses on microblogging scenarios in which commemoration of the First World War [5] is studied in relation to diverse countries. In contrast, we use relatively large size data (at least, when it comes to history-related studies) and we investigate many unknown aspects ranging from the types of references, intensity of remembering, key entities, dates, temporal patterns and so on.

Table 1: Dataset statistics.

Number of history-related hashtags	147
Number of tweets	888,251
Period of timestamps	8 Mar. 2016 – 24 Feb. 2017
Period of time references	8156 BC – 2029
Number of tweets with time references	357,682
Number of users	390,106
Number of URLs	204,075
Number of tweets with URLs	404,136

3 DATA COLLECTION

This section describes the data collection and preprocessing procedures as well as general statistics of the dataset used for analysis. The key statistics of the collected data are summarized in Tab. 1.

Collecting hashtags and tweets. 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`, `#WmnHist`, `#HistoryTeacher`) [2]. In addition, we prepared several hashtags that are commonly used when referring to the past: `#onthisday`, `#throwbackthursday`, `#historicalevent`, `#thisdayinhistory`, `#otd`. We then collected tweets that contain these hashtags by using Twitter's official API[3].

The collected tweets were issued from 8 March 2016 to 24 February 2017. To increase the coverage, we applied bootstrapping to search for other hashtags frequently used with the seed hashtags. The tweets tagged by such hashtags were then included into the seed set after the manual inspection of all the discovered hashtags as of their relation to the history. In total, we gathered 147 history-related hashtags which allowed us to collect 888,251 tweets [4]. We show the complete list of the hashtags in Tab. 6.

Extracting time-references. We extracted *time-references* from tweet content. Two categories of temporal references are typically distinguished based on the common distinction of temporal expressions in IR and NLP [4]: *explicit* and *implicit* temporal expressions. The former one is a concrete time point or time period, such as "1945" or "1980s", while the latter is a relative temporal expression such as "yesterday" or "two years ago". We use both types of temporal references in our study and, so, we convert all implicit (relative)

[2] http://blog.historians.org/2013/08/history-hashtags-exploring-a-visual-network-of-twitterstorians/
[3] https://dev.twitter.com/rest/public
[4] Ids of tweets analyzed in this paper are available in http://tk2-222-20713.vs.sakura.ne.jp/jcdl_2018_target_tweet_ids.txt

temporal expressions to the explicit (absolute) ones. To extract both the types of time references, we use Heideltime [18], which is a temporal tagger with a specialized option for tweet processing. Heideltime outputs normalized temporal expressions according to the TIMEX3 annotation standard. In total, we have found over 357k tweets with temporal expressions which represents 40% of the dataset. [5]

4 GENERAL ANALYSIS

In this section, we first investigate characteristic features of history-related tweets based on three data types: time expressions, entities and hashtags.

4.1 Temporal Analysis

Q. What is the character of memory decay?
Q. Which years are strongly remembered in particular?

We first analyze which time periods users are interested in by investigating time references included in tweets. We map all the extracted temporal expressions on timeline as shown in Fig. 1. We call the curve in Fig. 1, the *remembering curve* as it reflects the strength of the collective attention of users towards different time periods of history. To plot such a curve, we converted the extracted temporal references to probability distributions over their corresponding timespans using year level granularity. In other words, for a given time reference (e.g., 1960s) with t_b denoting its start year (1960) and t_e indicating its end year (1969) we set the probability distribution with zero values for $t < t_b$ and for $t > t_e$ (e.g., before 1960 and after 1969) and with non-zero values for $t_b \leq t \leq t_e$ that sum to 1 (e.g., 1/10 for each year from 1960 to 1969). We then combined for every year all the computed probability distributions based on all the tweets in our dataset. The formal definition of the probability distribution for a year y is given in Eq. (1).

$$S(y) = \sum_{[t_b, t_e] \in T} \delta(y, [t_b, t_e]) * \frac{1}{t_e - t_b + 1} \quad (1)$$

where T includes all the extracted time references, and the function δ returns 1 if the first argument is included in the second argument; otherwise, it returns 0.

Looking at Fig. 1 (especially, at the zoomed out plot in the inner graph), we can see that the number of time references is usually rapidly increasing towards the present (neglecting short-term disturbances caused by key events to be discussed later). In general, *the recent past is referred to more than the distant past, and the memory decay is fastest in the recent years.* This is intuitive and correlates with the corresponding study conducted on news articles related to different countries [2].

Several significant peaks are visible in Fig. 1 which represent two key events in the last century: WWI and WWII, and 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. While

[5] Note that some tweets contain abbreviated temporal expressions (e.g., 6/11/16 or 3/19/88). As Heideltime adds for such expressions "00" at the head of year information (e.g., 0016, 0088) we converted "00" to "19" or to "20" depending on whether the last two digits are less (conversion to 20) or more than 17 (conversion to 19).

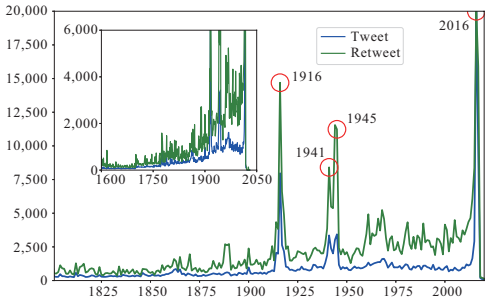


Figure 1: Distribution of time references in tweets. The peaks are in years 1916, 1941, 1945 and 2016.

WWII started in fact earlier with the Nazi invasion on Poland in 1939, many history-related tweets in our dataset originate from USA and Canada due to the chosen English hashtags resulting in the focus on the North American involvement in the war. This can be confirmed when looking at Fig. 2 that shows the most common entities corresponding to the peaks and at Fig. 4 which lists the top entities in our dataset. In this work we employ AIDA [13] - an annotation tool linking phrases in short text with their corresponding Wikipedia articles - for detecting entities.

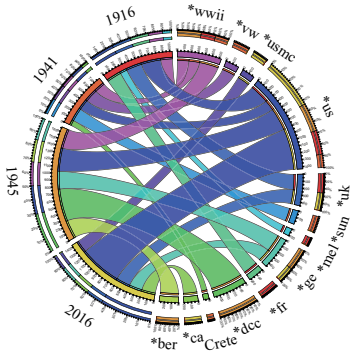


Figure 2: Top entities associated with the four peaks of Fig. 1. "*" is used to denote abbreviations made for saving space (*wwii: World War II , *vw: Virginia Woolf, *usmc: United States Marine Corps, *us: United States, *uk: United Kingdom, *sun: The Sun News-Pictorial, *mel:Melbourne, *ge: Germany, *fr: France, *dcc: Dachau Concentration Camp, *ca: Canada, * ber: Berlin).

Fig. 3 shows the most common hashtags used with content containing the peak years of 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. 3 shows that there are many mentions of 2016 with `#ww1`. This is because 2016 marked the 100th anniversary of the Battle of Verdun which is especially remembered due to its estimated nearly 1 million casualties. In the same year (1916) another remarkable event occurred - the British Army suffered its worst day with the loss of 19,240 men in the Battle of Somme (hence, the hashtag `#somme100`).

These events together with the Easter Rising of Irish republicans against British occupation in Dublin are ones of the primary causes for the strong remembrance of 1916. The round anniversary in 2016 especially amplified this remembering; hence, the corresponding peak in Fig. 1 is even higher than the subsequent peaks associated with WWII.

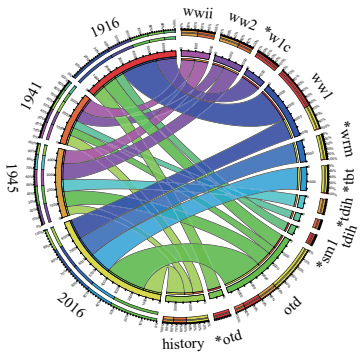


Figure 3: Top hashtags associated with the four peaks of Fig. 1. "*" is used to indicate abbreviations made for saving space (*w1c: #ww1centenary, *wrm: #weremember, *tbt: #throwbackthursday, *tdih: #thisdayinhistory, *sm1: #somme100, *otd: #onthisday).

Finally, as shown in Fig. 3 #otd and #onthisday are commonly used labels for indicating historical content that occurred on the same calendar day in the past. Past-to-present connection through the reference to a calendar day is in fact a popular way of recalling past events and is typically used in newspapers (e.g., sections about 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.

4.2 Entity Analysis

> Q. Which types of entities and which entities in particular tend to be remembered?
> Q. Which past and present entities are compared or mentioned together?

We now look into entities referred to in tweets as, often, a particular entity such as a person or an event is what society remembers strongly from the past.

4.2.1 Entity Popularity and Type. We first count the number of times a given entity is mentioned in tweets. Fig. 4 lists the top frequent 30 entities overall. In these top 30 entities, there are 22 countries, regions, or cities, two historical events (WWI and WWII), three persons (Adolf Hitler, Abraham Lincoln, and Donald Trump), and three other kinds of entities. Location entity type tends to be then most frequently mentioned within the group of the top common entities. This is because places are key constructs helping to locate the occurrence of events, indicate locations of historical buildings or areas where famous people lived, as well as they form a kind of a "bridge" between the past and the present by existing across long time periods.

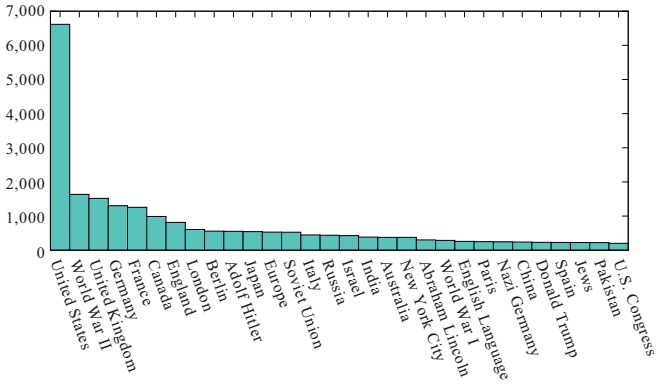


Figure 4: Top 30 entities mentioned in the dataset.

To thoroughly investigate entities and their types, we next automatically map all the entities into DBpedia [3] to obtain their type assignments. We then divide all the entities into five major types (Person, Group, Place, Event, and Others), and show their rates in Fig. 5. It can be noticed that *persons, places and groups tend to be frequently mentioned in history-focused tweets and the person category is especially common.* Note that while places were the most common entity type in the set of the top frequently mentioned entities as indicated in Fig. 4, they are actually less common than persons when all the entities in our dataset are concerned.

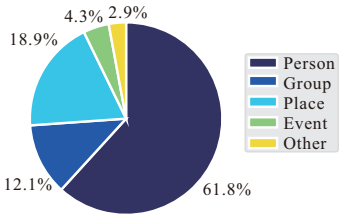


Figure 5: Rate of different types of entities.

To give some examples of entities, Tab. 2 lists the top 10 entities for the Person, Group and Event types[6]. As most of the tweets were posted in 2016 when the United States presidential election was held, the names of the five US presidents (Abraham Lincoln, Donald Trump, Bill Clinton, Barack Obama, and George Washington) appear within the top 10 persons. As for the events, wars and battles are the prevailing type. Interestingly, groups include many military units (e.g., U.S. Army, Royal Air force), *which also suggest the significant focus on wars and conflicts from the past.*

4.2.2 Connection of Past and Present Entities. The issue of connecting present and past entities is especially interesting as it relates to the notion of "usable history". Historical entities can be used for a variety of reasons, for example, for comparison with present entities or present context, for emphasizing analogy, making predictions and so on. To analyze the way in which past entities are utilized in connection to the present ones we first need to distinguish present from past entities, which is of course not straightforward. We apply

[6]Location examples can be seen in Tab. 5.

Table 2: Top 10 entities of persons, groups, and events.

Person	Group	Event
Adolf Hitler	Jews	World War II
Abraham Lincoln	U.S. Congress	World War I
Donald Trump	Royal Navy	Vietnam War
Sharon Corr	U.S. Army	Battles of Saratoga
Bill Clinton	U.S. House of Representatives	Battle of Verdun
Alan Evans	Luftwaffe	American Revolutionary War
Barack Obama	Royal Air Force	Korean War
Napoleon	U.S. Navy	Omaha Beach
George Washington	BBC	Cinco de Mayo
Jerrard Tickell	Federal Bureau of Investigation	Battle of Gettysburg

a simple division rule, according to which, an entity is regarded as a *past entity* if the end of its lifetime[7] (e.g., life, event duration) falls within the last millennium. We are aware that this may be seen arbitrary, yet, we had to set some threshold to carry out the analysis, and the turnover of the millennium seemed like a good choice for a temporal landmark. Entity lifetimes were collected from DBpedia [8].

First, in Tab. 3, we compare the sizes of the past and present entity sets extracted from our dataset. We can observe that the number of unique past entities extracted by AIDA and typed by DBpedia is relatively large constituting roughly half of that of the present entities. This confirms that our dataset is specifically focused on history.

Table 3: Sizes of the past and present entity sets.

	Total	Person	Group	Place	Event
Size of the past entity set	8,262	6,368	1,257	746	637
Size of the present entity set	16,355	10,858	4,723	340	774

We plot in in Fig. 6 conditional probabilities based on the studied entity types to understand how often the different types appear given another entity of a given type. We can observe that *present places tend to co-occur with entities of any other type. Especially, many past entities tend to appear together with present places.* Furthermore, $P(PresentPerson|PresentEvent)$ has relatively high value which is understandable. Finally, if a past place, such as "Nazi Germany" and "Holy Roman Empire", is in a tweet, past persons or past events tend to occur as well. This is also expected. Interestingly, past places are often mentioned with present places as well (supposedly for emphasizing place continuity, spatial relations or for place-oriented comparisons).

Next, in Tab. 4 we take the top 5 common present persons (column “Entity" in the top part of the table), the top 5 common past persons (column “Entity" in the middle part of the table), and the top 5 common past events (column “Entity" in the bottom part of the table). We then output in columns “1”, “2” and “3” their top 3 most often co-occurring entities from the opposite time frame. In particular, if the “Entity" column contains past entities then in columns “1”, “2” and “3” we show their top co-occurring present entities. Otherwise, we show past entities. When looking at past persons and past events, one can observe that indeed present places commonly co-occur with them. For example, for the top 5 past persons, their most common co-occurring entities contain 8 countries and 1 city, while for the past events the corresponding number of locations is 12. Apparently, when recalling past persons and past events users tend to also mention where the persons lived or where the events occurred, thus, “grounding” them in spatial dimension.

Finally, Tab. 5 lists past entities co-occurring with the top frequent locations, from where again the significance of WWII in relation to collective memories of many countries can be observed.

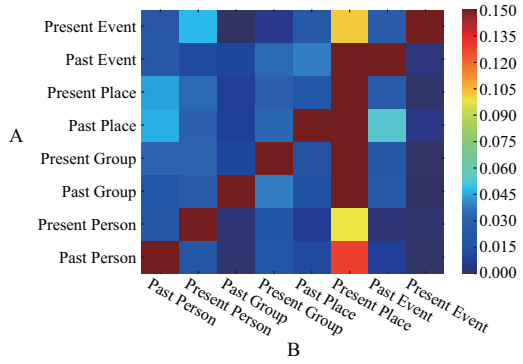


Figure 6: Conditional probabilities $P(B|A)$ of entity type on x axis (in column) given the presence of entity type on y axis (in row).

Table 4: Top 5 present, top 5 past persons and top 5 past events (given in “Entity” column) with their top 3 co-occurring past/present entities of any type. "-" denotes cases when no corresponding entity can be found.

	Rank	Entity	1	2	3
Present person	1	Donald Trump	Adolf Hitler	Abraham Lincoln	Nazi Germany
	2	Sharon Corr	-	-	-
	3	Bill Clinton	Charles Frankel	Nazi Germany	World War II
	4	Barack Obama	Muhammad Ali of Egypt	Adolf Hitler	-
	5	Elizabeth II	James II of England	George Washington	-
Past person	1	Adolf Hitler	Germany	Donald Trump	United States
	2	Abraham Lincoln	United States	Donald Trump	Israel
	3	Alan Evans	-	-	-
	4	Napoleon	France	United Kingdom	Italy
	5	George Washington	United States	Philadelphia	Martin Scorsese
Past event	1	World War II	United States	United Kingdom	France
	2	World War I	United States	Canada	United Kingdom
	3	Vietnam War	United States	CBS	Canada
	4	Battles of Saratoga	United States	-	-
	5	Battle of Verdun	France	Germany	United Kingdom

4.3 Hashtag Analysis

Q. What are the most popular hashtags?

In this section we investigate popularity patterns of hashtags. Hashtags are commonly used to indicate specific themes of tweets allowing others to find them. First, Fig. 7 (a) shows the top frequently used hashtags based on their tweet counts. We also list the top hashtags ranked by the retweet count in Fig. 7 (b) and by the number of Twitter accounts from which the tweets originate in Fig. 7 (c). From these data, we can observe that `#throwbackthursday`,

[7] For currently valid entities such as currently alive persons the end of their lifetimes is set to the current year.

[8] We use “birthDate” and “deathDate” for person entities, “formationDate” and “dissolutionDate” for groups, and “foundingDate” and “dissolutionDate” for locations.

Table 5: Top 10 locations and their top 3 co-occurring past entities of any type. The abbreviated names of entities are for: Helen Farnsworth Mears (H. F. Mears) and Dominion of Newfoundland (Dom. of Newfoundland).

Rank	Present locations	1	2	3
1	United States	World War II	Battles of Saratoga	Soviet Union
2	United Kingdom	World War II	James II of England	Soviet Union
3	Germany	Adolf Hitler	John Cudahy	Soviet Union
4	France	World War II	Napoleon	Louis XIV of France
5	Canada	World War II	World War I	Dom. of Newfoundland
6	Japan	World War II	H. F. Mears	Hideki Tojo
7	Italy	Holy Roman Empire	World War II	Battle of Monte Cassino
8	Russia	World War II	Catherine the Great	Nicholas I of Russia
9	Israel	World War II	Nazi Germany	Vichy France
10	India	Vasco Da Gama	Edwin Lutyens	World War II

which is representative for a trend among social media sites including Twitter and Facebook to post own past photographs (often from one's childhood), gains most attraction across all these dimensions of popularity. The tweets with hashtag `#throwbackthursday` and similar ones predominantly refer to personal experiences and these hashtags are used by large number of users. Another observation is that *Fig. 7 (a) shows that* `#onthisday`, `#otd` *are present in a relatively large number of tweets and retweets, yet, they are used by fewer accounts. Unlike* `#throwbackthursday`, *these hashtags tend to be used by specialists (often historians and scientists from related areas) who select and disseminate interesting content about the past.* This content then triggers relatively *high engagement from other users as evidenced by the high retweet popularity of* `#onthisday`, `#otd` *in Fig. 7 (b).* We note that the demographics analysis should provide many more interesting insights regarding typical profiles of posting users, and is going to form the part of our future work.

5 CATEGORY BASED ANALYSIS

Q. How can past-related tweets be categorized and arranged?
Q. What kinds of semantic categories hashtags can be grouped into?
Q. What are the characteristics of these categories?

5.1 Definitions

In this section we introduce our categorization scheme of hashtags. The objective is to determine key types of history references. Based on the proposed categories automatic classifiers could be built to allocate tweets into different classes. Automatically labeling tweets could be then used for improving content retrieval, recommendation or for further analysis that would lead to better understanding of history-related interest and content sharing. Based on manual investigation of a large sample of tweets, we propose the following categories:

1. **General History** : hashtags used in general to broadly identify history-related tweets that do not fall into any specific type (e.g., `#history`, `#historyfacts`).
2. **National or Regional History** : hashtags annotating tweets which describe national or regional histories, for example, `#canadianhistory` or `#ushistory` including also past names of locations (e.g., `#ancientgreece`).
3. **Facet-focused History** : hashtags which relate to particular thematic facets of history (e.g., `#sporthistory`, `#arthistory`).
4. **General Commemoration** : hashtags that contain tweets commemorating or recalling a certain day or period (often somehow related to the day of tweet posting), or unspecified entities, such as `#onthisday`, `#otd`, `#todayweremember`, `#4yearsago` and `#rememberthem`.
5. **Historical Events** : hashtags related to particular events in the past (e.g., `#wwi`, `#sevenyearswar`).
6. **Historical Entities** : hashtags denoting references to specific entities such as persons, organizations or objects (e.g., `#stalin`, `#napoleon`).

Tab. 6 shows our assignment of all the history-related hashtags found within the dataset into these categories. Note that hashtags

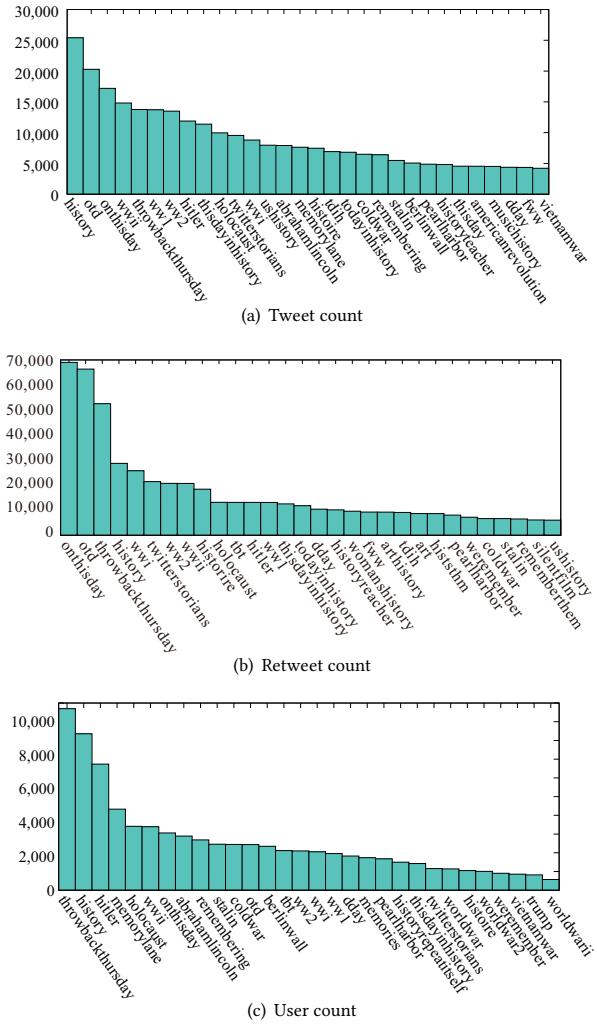


Figure 7: Top 30 popular hashtags by counts of tweets, retweets and users.

Table 6: Collected hashtags and their categories.

Category	Hashtags
General History	history, historyfacts, oldpicture, historyteacher, memorylane, histoire, twitterstorians, historicalcontext, colorization, memories, oldphoto, earlymodern, historicalevent, worldhistory, twitterstorian, historynerd, histedchat, historyfeed, archives, historymatters
National or Regional History	canadianhistory, ushistory, histoireducanada, jewishhistory, nazigermany, ottoman, cdnhistory, dchistory, cdnhist, thirdreich, tohistory, mdhistory, bchist, abhistory, vthistory, britishhistory, ancientchina, ancientegypt, ancientgreece, americanhistory, thisiscanadashistory, ottomanempire, ontariohistory, earlyamhistory, japanhistory, japanesehistory, chinesehistory, localhistory
Facet-focused History	blackfacts, histoiremiliterre, wmnshist, arthistory, sporthistory, womenshistory, navalhistory, presidentialhistory, musichistory, militaryhistory, blackhistory, envhist, histmed, wmnhist, todayintennishistory, todayinblackhistory, ibhistory, u2history, historythroughcoins, histSTM, silentfilm, historyscience, histsci, digitalhistory, foodhistory, histmonast, histnursing, histgender, histtech
General Commemoration	onthisday, otd, otdh, thisdayin, thisdayinhistory, todayinHistory, tdih, onthisdayinhistory, otdih, 100yearsago, thisday, lessthan100yearsago, todayweremember, titanicremembranceday, weremember, 100yearsago, remembering, wewillrememberthem, rememberthem, remembranceday, historyrepeatsitself, throwbackthursday, tbt
Historical Events	1ww, gulfwar, ColdWar, ww2, ww1, worldwar, worldwarii, vietnamwar, worldwar2, worldwartwo, veday, worldwarone, greatwar, battleofmidway, holocaust, frenchrevolutionarywar, wwii, wwi, sevenyearswar, firstworldwar, coldwarhist, gulfwar, battleofokinawa, dday, berlinwall, ddayoverlord, operationoverlord, fww, pearlharbor, americanrevolution, 6juin44, sww, june61944, victoryineuropeday, dday72, neverforget84, warof1812, ww1politics, ww1centenary,ww1economy, cw150
Historical Entities	stalin, hitler, abrahamlincoln, rudolfhess, napoleon

concerning particular dates or time periods such as `#june61944` could be considered as a separate category or could be made a part of **General Commemoration**. We decided however to place them under the **Historical Events** category as they tend to be used to refer to particular events by their occurrence dates (e.g., `#june61944` referring to the Normandy landings).

In Fig. 8 we show the rate of each category based on the count of tweets in our dataset. **Facet-focused History** appears to be quite common category followed by **General History** and **General Commemoration**. This suggests *relatively significant amount of specialized history-related content besides broad and general history related content.*

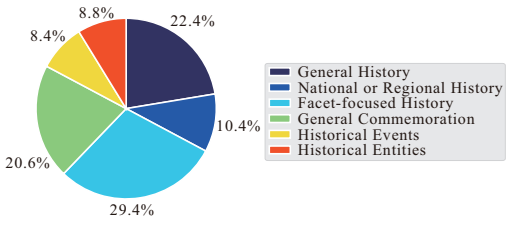


Figure 8: Distribution of categories.

5.2 Inter-category Similarity

To better understand characteristics of the proposed categories, we now investigate inter-category affinity by measuring the co-occurrence values between the categories. We use Jaccard coefficient computed as follows:

$$Score(A, B) = \frac{| T_A \cap T_B |}{| T_A \cup T_B |} \quad (2)$$

where: $| \cdot |$ is the size of a set. T_A and T_B are the numbers of tweets that include hashtags classified in category A and B, respectively.

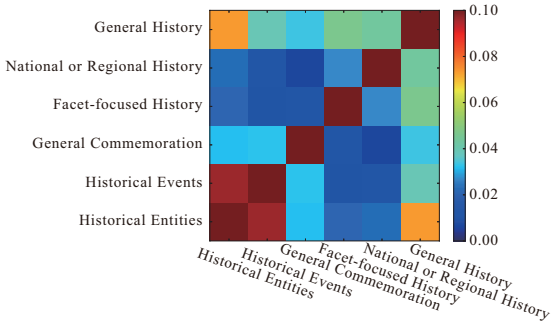


Figure 9: Category co-occurrence.

Fig. 9 plots the obtained co-occurrence values between the categories. We can see that the *category **General History** is truly "general" since the hashtags in this category tend to highly co-occur with hashtags in the other categories.* This is the only category that has relatively high similarity with every other category. Another observation is that the co-occurrence value between **Historical Events** and **Historical Entities** is quite high. This is because many famous entities in our dataset were involved in key events (e.g., Stalin, Hitler in WWII). Similarly, when users refer to the past as a general commemoration they tend to focus on particular well-known past events and key persons. Hence, **General Commemoration** and **Historical Entities**/**Historical Events** hashtags are sometimes used together. On the other hand, **National or Regional History** and **Facet-focused History** have rarely co-occurring hashtags with the hashtags of other categories (except for **General History**). This indicates that they tend to be assigned to relatively unique and specialized content.

5.3 Temporal Category Analysis

We now investigate temporal references in tweets in relation to their categories. Our interest is in understanding how similar are time references included within tweets annotated with the hashtags of the same category (or in other words, whether tweets under the same category tend to mention similar or rather different years).

We compute such temporal coherence for each category by comparing the vectors of hashtags in a given category. These are built based on temporal expressions associated with the hashtags. In particular, for each hashtag, we construct a vector representing year scores derived from temporal references within tweets labeled by this hashtag. We first map all the temporal references to the year level granularity and then compute year scores using Eq. 1. Such vectors reflect commonly mentioned years for each hashtag. Pairwise similarities of hashtags falling into the same category are then computed by using cosine similarity measure and are averaged to give the final scores displayed in Tab. 7.

Tab. 7 shows that the time-based similarities of hashtags in **Historical Entities** and **General History** are relatively high (when compared to the average value for the entire data shown in the last row). Though, we should keep in mind that many tweets under these hashtags lack time references as indicated by their ratio values

being lower than the average values (see the 4th column). Nevertheless, hashtags under these two categories tend to be relatively similar to each other in terms of the focused time periods. This actually is not surprising for the **Historical Entities** category since several its hashtags represent entities with overlapping lifetimes (at least this is the case in our dataset). Yet for **General History** it would mean that there is a good level of agreement in the question of the most important historical periods and events (e.g., WWI and WWII).

Table 7: Average cosine similarity for each category based on years (2nd column), standard deviation of the similarities (3rd column) and the rate of tweets including time references (4th column).

Categories	Similarity	Std. dev.	Ratio
General History	0.75	0.29	0.19
National or Regional History	0.58	0.35	0.20
Facet-focused History	0.57	0.32	0.32
General Commemoration	0.54	0.34	0.55
Historical Events	0.47	0.31	0.18
Historical Entities	0.81	0.12	0.09
Total	*0.56*	*0.33*	*0.29*

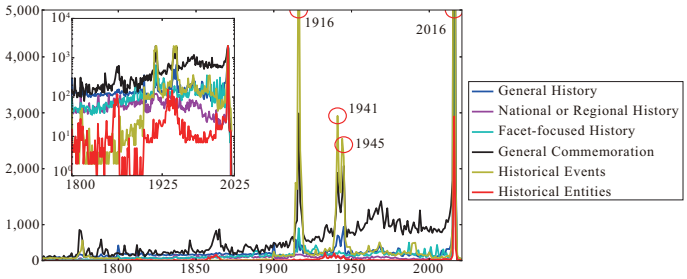


Figure 10: Distributions of time references in the categories (small inner graph shows the plots in log scale).

We next plot in Fig. 10 the distributions of time references extracted from tweets in each category. Naturally, all the categories have strong relation to the present (e.g., to the current events or present entities) since year 2016 is characterized by a high peak for all the plots. **Historical Events** category is strongly focused on the two critical events of the last century: WWI and WWII. While **General Commemoration** has also strong focus on the two wars, it features actually a very different time plot when compared to the ones from other categories. Since the common reason for commemorating events are their anniversaries (e.g., `#otd`) rather than external triggers such as ongoing events, the *tweets under **General Commemoration** relate to many diverse years in the past. The pattern of tweeting under this category reflects thus smaller selectivity (or higher diversity) of the collective attention towards time periods of history.*

5.4 Entity-focused Category Analyses

Next, we look into entity distributions to investigate their coherence in each category. In a similar way to the above-discussed temporal analysis, we first count how many times each entity is mentioned with a given hashtag in order to create hashtag's entity vector. We then calculate pairwise similarities between hashtags in each category by comparing their vectors and we average these similarities to give the final score per each category.

Table 8: Average cosine similarity for each category based on entities (2nd column), its standard deviation (3rd column) and the rates of: entities, past entities and present entities, displayed respectively in the last three columns.

Categories	Sim.	Std. dev.	All	Past	Present
General History	0.13	0.19	0.15	0.03	0.05
National or Regional History	0.05	0.16	0.16	0.03	0.08
Facet-focused History	0.12	0.16	0.19	0.05	0.07
General Commemoration	0.19	0.25	0.35	0.09	0.14
Historical Events	0.11	0.18	0.20	0.05	0.08
Historical Entities	0.12	0.12	0.09	0.02	0.03
Total	*0.10*	*0.17*	*0.14*	*0.05*	*0.08*

In Tab. 8 we show per each category the average similarities, their standard deviations as well as we output the rates of all the entities, the rate of past entities and the ones of present entities. The entity-focused similarities are in general relatively low indicating that hashtags of the same category tend to refer to different entities. Interestingly, when looking at Tab. 8 we can observe that all the values of **General Commemoration** are higher than the average values for the entire dataset (see the last row). Over one third of tweets under this category contain entities (see 4th column in Tab. 8), while more than half of its tweets contain time references as shown in Tab. 7. Thus, *compared with other categories, users tend to include more entity names and more temporal expressions into tweets tagged with the hashtags from **General Commemoration** category. This may be distinguishing characteristic of commemorating activity.*

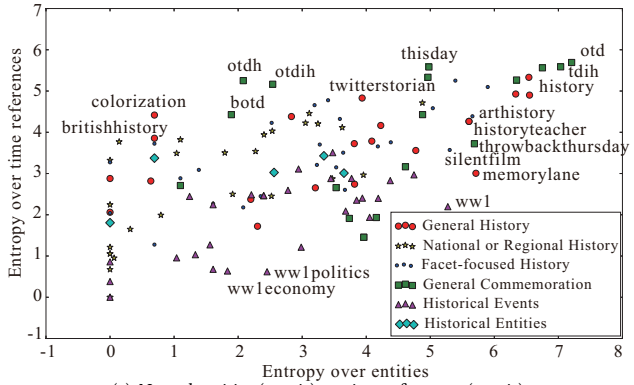
5.5 Analysis of Entity and Time Reference Dispersions

Finally, we study the dispersions of entities and time expressions for each category. Fig. 11(a) places hashtags according to their entropy values calculated over the distributions of contained entities and mentioned temporal expressions. We can observe that **General Commemoration** hashtags (e.g., `#otd`, `#thisday`, `#otdih`) and some of **General History** hashtags (e.g., `#twitterstorians`, `#history`, `#colorization`) are often characterized by high values of entropy over time references. Hashtags under the **General Commemoration** category tend to also have relatively high values of entropy over entities (e.g., `#otd`, `#throwbackthursday`, `#tdih`). On the other hand, **Historical Events** and **Historical Entities** categories have on average low values of the entropy over temporal references, which is understandable given that they focus on relatively short-lasting events (e.g., `#ww1politics`, `#ww1economy`, `#june61944`). For the same reason, these categories also achieve low values of entity-based entropy. Furthermore, **National or Regional History** hashtags (e.g., `#britishhistory`) tend to have less varying entities compared to other categories.

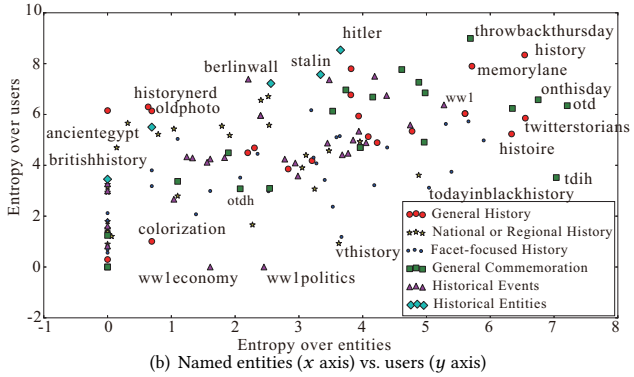
Lastly, in the remaining two figures (Figs. 11(b) and (c)) we look into the relation between the entropy of user distributions and the

entropy of entity distributions as well as the relation between the entropy of user distributions and the entropy of temporal references, respectively.

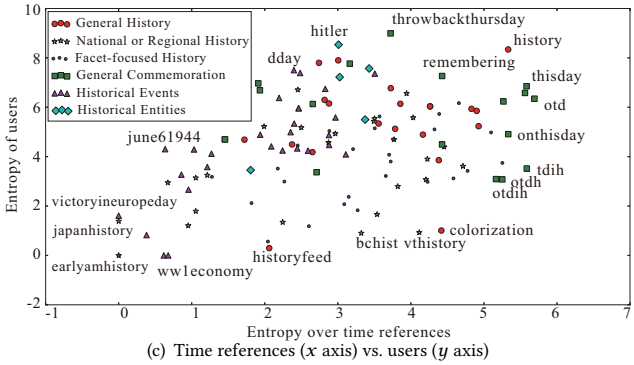
The right hand side corner of Fig. 11(b) is occupied by both ***General Commemoration*** *and* ***General History*** *hashtags suggesting that many different users tweet with these hashtags and the users tend to include references to many different entities.* **Historical entities** in our dataset like `#stalin` and `#hitler` tend to be referred to by large number of users and, interestingly, they also have relatively high values of entity entropies. In Fig. 11(c), **General Commemoration** tweets contain both many diverse dates and are issued by many users.



(a) Named entities (x axis) vs. time references (y axis)



(b) Named entities (x axis) vs. users (y axis)



(c) Time references (x axis) vs. users (y axis)

Figure 11: Entropies related to different categories.

6 DISCUSSIONS

6.1 Limitations

Data Collection. We note that the data collection method that we relied on naturally misses the portion of tweets not tagged by any history-related hashtags. We list here two other approaches that could be used to collect history-related data: (1) *collecting content with temporal expressions pointing to the past*, and (2) *collecting content that contains past entities*. Every approach is however not without its shortcomings. The first method was used in [2] for extracting past references in news articles and relied on the presence of temporal expressions in text. This however is not always guaranteed for history-related tweets. Indeed, as it can be seen in Tab. 1 the rate of tweets with temporal expressions is 40%, hence, about 4 tweets out of 10 contain tweets any time reference. Thus, this approach would miss many relevant tweets resulting in rather low recall. Similarly, history-related tweets may not mention any past entity. Indeed, from Tab. 8 (see the column "Past" and the last row) we can actually notice that the rate of tweets containing at least one past entity, which can be recognized by the state-of-the-art tools, is only 0.05 for our dataset. Furthermore, some entities, especially, more obscure ones may not be even present in any knowledge base or may not be detectable using standard tools.

We thus assumed in this work an approach that relies on extracting explicit history-focused hashtags and on subjecting them to the manual analysis of tagged content. While such a choice is likely characterized by a high precision, it may obviously suffer from lowered recall as discussed before. Yet, in the view of the reported statistics, we still believe it is superior than collecting tweets based on contained dates or historical entities. Future work should nevertheless explore more refined approaches for extracting implicit past-related content; ideally, ones making use of the combination of all the signals (hashtags, past entities, temporal expressions, etc.). We also note that while we put great effort into the manual verification of used hashtags as for their relevance to the history, there is always chance that some might not be fully devoted to the past, or, in general, their selection may cause certain biases.

Different Languages. In the current study we have mainly focused on English tweets. Further exploration should involve different languages (e.g., French, Japanese) as well as the cross-comparison of the obtained results. Our current aim is however not to look into particular aspects and specificities related to different countries but to rather uncover general tendencies. The focus on English was a natural decision for this initial study due to the international role and ubiquity of this language.

Fine-grained Analysis. The present study is quantitative aiming to provide first glimpse into the issue of historical-references in SNS and to conduct broad exploratory analysis. Qualitative exploration should be later carried for obtaining fine-grained comprehension of the way in which users refer to the past. For example, the future analysis could examine why some entities are (or are not) popular as well as could apply sentiment analysis to identify tendencies in polarity towards particular past events or entities. Another research direction could be the detailed analysis of contexts in which the past entities or past years are mentioned. These would necessarily require some sort of manual and qualitative exploration of tweet content.

User-focused Analysis. Our study has exploratory character and focuses on the shared content first. However, a very interesting question is about the type of users who share or are interested in historical content in social network services. Due to the time and space constraints, we have left however the user-focused analysis for the future work.

Studying Temporal Data Snapshots. Subsequent work should investigate portions of data collected over different time frames in order to verify which of the results are specific to the particular time frames of data generation and which are of general character. We are currently in the process of collecting more data to perform such comparative investigation (aiming at the comparison of results based on tweets from 2016 with ones based on tweets issued in 2017).

6.2 Potential Applications

Finally, we list here several example applications that could be potentially constructed based on the history-related content shared in Twitter:

(1) *Recommending past-related content for readers interested in studying history.* This could be, for example, popular and interesting content or the content that matches particular user interests such as tweets under hashtags of the Facet-focused category that a user is interested in. Through this analysis and other forthcoming ones we could better understand what kind of history-related content at what time periods is becoming attractive to many users.
(2) *Creating history-focused chatbots for disseminating historical knowledge and for entertaining users.*
(3) *Finding, summarizing and explaining past entities which are mentioned in relation with the popular present entities to provide analogy and a novel, potentially interesting context for the latter.*
(4) *Automatically summarizing and comparing history-related opinions and popular topics across different regions.*
(5) *Automatically suggesting hashtags for tweets based on included entities, years and based on the predicted hashtag categories.*

7 CONCLUSIONS & FUTURE WORK

In this paper we have for the first time studied how people explicitly refer to the history in microblogging and in which contexts such references occur. As mentioned before, history-focused content recommendation should offer opportunities for the dissemination of interesting and trendy recollections by pushing them to users.

Our analysis does an initial groundwork in establishing how microbloggers conceive of, share and refer to history-related content such as one on past events and persons. Through this promising study we hope to shed more light on the way in which history-related content is used and shared in microblogging, and by this to encourage subsequent research and development of systems aiming at educating history. We perform basic study on a coarse level, providing initial observations, identifying several interesting research directions and suggesting potential applications. Our analysis is nevertheless conducted from multiple perspectives.

Future work will identify (a) *differences between general and personal histories as well as will look into what makes tweets about personal history appear interesting*. Social media provide novel opportunities to create such personal connections which can help raise an interest in the significance of historical knowledge beyond the personal experience. Future work will also include (b) *geographical analysis of tweets which may point to different cultural practices with regard to references to the past.* As currently most of the tweets in our dataset originate from the English-speaking part of the world we should contrast these results with ones obtained on data collected in different languages. Next, we plan also to (c) *explore the interdependence between present-day events, their function as triggers for references to history and the latter's effect on the interpretation of the present.* Finally, as mentioned before, (d) *we plan to study in detail the characteristics of users sharing history-related content in Twitter* such as their demographics, characteristics of their followers and followees, and their interaction patters.

Acknowledgments. This work was supported in part by MIC SCOPE (#171507010) and MEXT Grant-in-Aids (#17H01828 and #17K12792).

REFERENCES

[1] R. P. Abelson and A. Levi. 1985. Decision Making and Decision Theory, Handbook of Social Psychology. 231–309.
[2] C.-m. Au Yeung and A. Jatowt. 2011. Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining. CIKM '11, Glasgow, Scotland, UK, 1231–1240.
[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. ISWC'07/ASWC'07, Busan, Korea, 722–735.
[4] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt. 2015. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)* 47, 2 (2015), 15.
[5] F. Clavert, B. Majerus, and N. BeauprÃĪ. [n. d.]. #ww1. Twitter, the Centenary of the First World War and the Historian.
[6] J. Cook, A. D. Sarma, A. Fabrikant, and A. Tomkins. 2012. Your Two Weeks of Fame and Your Grandmother's. WWW '12, Lyon, France, 919–928.
[7] H. Ebbinghaus. 1913. *Memory: A Contribution to Experimental Psychology.*
[8] M. Ferron and P. Massa. 2011. Collective Memory Building in Wikipedia: The Case of North African Uprisings. WikiSym '11, Mountain View, California, USA, 114–123.
[9] R. G.-Gavilanes, A. Mollgaard, M. Tsvetkova, and T. Yasseri. 2017. The memory remains: Understanding collective memory in the digital age. *Science Advances* 3, 4 (2017).
[10] T. Gilovich. 1981. Seeing the Past in the Present: The Effect of Associations to Familiar Events on Judgments and Decisions. *Journal of Personality and Social Psychology* 40, 5 (1981), 797.
[11] M. Halbwachs. 1950. *La Memoire Collective.* Les Presses universitaires de France, (in French).
[12] C. Hoerl and T. McCormack. 2001. *Time and Memory: Issues in Philosophy and Psychology.*
[13] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. 2012. KORE: Keyphrase Overlap Relatedness for Entity Disambiguation. CIKM '12, Maui, Hawaii, USA, 545–554.
[14] R. Jacoby. 1997. *Social Amnesia: A Critique of Contemporary Psychology.*
[15] A. Jatowt, D. Kawai, and K. Tanaka. 2016. Digital History Meets Wikipedia: Analyzing Historical Persons in Wikipedia. JCDL '16, Newark, New Jersey, USA, 17–26.
[16] N. Kanhabua, T. N. Nguyen, and C. Niederée. 2014. What Triggers Human Remembering of Events?: A Large-scale Analysis of Catalysts for Collective Memory in Wikipedia. JCDL '14, London, United Kingdom, 341–350.
[17] N. Kanhabua, C. NiederÃĪe, and W. Siberski. 2013. Towards Concise Preservation by Managed Forgetting: Research Issues and Case Study. iPres'13, Lisbon, Portugal.
[18] E. Kuzey, J. Strötgen, V. Setty, and G. Weikum. 2016. Temponym Tagging: Temporal Scopes for Textual Phrases. WWW '16 Companion, Republic and Canton of Geneva, Switzerland, 841–842.
[19] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM'10, Washington, DC, USA.