Studying How the Past is Remembered: Towards Computational History through Large Scale Text Mining

Ching-man Au Yeung*
ASTRI
3/F Bio-informatics Centre
2 Science Park West Avenue, Hong Kong
albertauyeung@astri.org

Adam Jatowt
Graduate School of Informations, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto
606-8501, Japan
adam@dl.kuis.kyoto-u.ac.jp

ABSTRACT

History helps us understand the present and even to predict the future to certain extent. Given the huge amount of data about the past, we believe computer science will play an increasingly important role in historical studies, with computational history becoming an emerging interdisciplinary field of research. We attempt to study how the past is remembered through large scale text mining. We achieve this by first collecting a large dataset of news articles about different countries and analyzing the data using computational and statistical tools. We show that analysis of references to the past in news articles allows us to gain a lot of insight into the collective memories and societal views of different countries. Our work demonstrates how various computational tools can assist us in studying history by revealing interesting topics and hidden correlations. Our ultimate objective is to enhance history writing and evaluation with the help of algorithmic support.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous; I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms, Languages, Experimentation

Keywords

computational history, news analysis, temporal analysis

1. INTRODUCTION

George Santayana, the famous Spanish American philosopher, once wrote, "Those who cannot remember the past are condemned to repeat it." It is hard to overestimate the importance of history of societies. History helps to bridge the present and the past, understand the present and even to predict the future to some extent. Studying the past and maintaining an awareness of national and international history serve many purposes such as educational, patriotic or political ones. Images of the past also commonly serve to legitimate a present social order [8].

Traditional historical methods and guidelines, according to which historians study the past and write histories, rely to a large extent on manual effort in amassing and investigating historical evidences. However, with the proliferation of digitalized historical sources such as newspaper archives, scanned books and other digital artifacts, it has become possible to employ a wide range of computational techniques in historical studies. Such an approach to historical studies, what we call *computational history*, would assist historians in analyzing massive amounts of data to obtain evidences that support various hypotheses and different correlations among events in the past. This will open up new research areas very much like computational linguistics [17] and, more recently, computational social science [15] have done.

Besides knowledge creation, computational history can also be used as a verification tool for evaluating the credibility of the existing historical knowledge. According to the meta-history view, history should not always be considered credible but rather requires a constant process of revision. This is because in many cases a "usable past" is often created to serve political and identity needs of nations. Since there is no "universal history", the remembered past, as being an intellectual product, calls for stringent analysis and criticism, and should remain in permanent evaluation [20]. The power of computational tools in processing huge amount of historical data will be desirable in this context.

In this paper, we focus on one important aspect of history, which is the perception of the past, or how the past is remembered. We achieve this by first collecting a large dataset of news articles about different countries from Google News Archive[1], and analyzing the data using computational and statistical tools such as topic modeling. We show that analysis of references to the past in news articles allows us to gain a lot of insight into the collective memories and societal views of different countries, such as what events or periods are of great importance to a country, and what were the reasons that triggered the remembering of the past at a particular time. The study of the collective remembering of

*Au Yeung contributed to this paper while he was at the NTT Communication Science Laboratories, Kyoto, Japan.

[1] http://news.google.com/archivesearch

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

the past using this methodology will help to understand the way in which societies retain their memories, modify them over time by recalling or eradicating particular events, and keep the memory of others. We also show the possibility of investigating the selectivity of social memories, as reflected in the differences between what have actually happened in the past with what is currently remembered.

Michel *et al.* [19] introduce "culturomics," an approach of studying the evolution of human culture by applying computers to analyze a huge dataset of digitalized books. We extend this idea to investigate history and collective memory. Our work is novel because we do not only focus on a particular period in the past (e.g. by analyzing documents created at that time), but attempt to study the *image* or *perception* of the past by analyzing references to the past in recent news articles. This generates even much more information and knowledge about our history.

This paper is structured as follows. In Section 2 we discuss works related to our project. In Section 3, we give an overview of our methodology. In Section 4, we present the results of statistical analysis and examination of retrospective views of the past using topic models. We discuss some applications of our work, its limitations and possible future research directions in Section 5. Finally, we conclude this paper in Section 6.

2. RELATED WORK

The use of computer to assist data analysis and mining hidden patterns in social sciences has attracted substantial attention in recent years. Lazer *et al.* write about computational social science [15], which aims at leveraging computational power to collect and analyze large datasets to reveal patterns of individual and group behaviors. This trend is in particular propelled by the explosive growth of digital data in recent years, ranging from digitalized historical archives, user-generated data on the Web, to huge amount of records of user interactions on social networking sites.

Some examples can be found in the literature recently. Michel *et al.* [19] have shown that analysis of n-grams from a huge corpus of digitalized books can be used to reveal trends in the development of the English language and usage of vocabulary over time. The recent work by Takahashi *et al.* [26] is an example of how one can measure the impact of historical persons using Wikipedia. Shahaf and Guestrin [23] propose an algorithm for discovering hidden connections and chains of events in news articles. There are also works on finding "across-time" synonyms [14][4] and comparing word senses over time [25] to support searching in document archives.

Historians and sociologists propose the concept of *collective memory* as the notion of societal remembrances of the past in contrast to individual memories composed of references to one's personal history [11]. Studies of collective memories have so far been limited to anecdotal analysis and interrogations of subjects. Given now the availability of huge amount of data about the past, we believe that our understanding of history can also benefit from the use of various computational tools. However, to the best of our knowledge, we are the first to attempt to apply computational methods on text corpora in order to study how the past is remembered and other related topics.

Several techniques in the field of computer science are actually relevant to computational history. These techniques include topic detection and tracking (TDT) [1], and temporal analysis. Topic detection and tracking focuses on developing methods for tracking changes in the popularity of topics over time given a text corpus. For example, Blei *et al.* [5] and Wang and McCallum [28] extend latent Dirichlet allocation (LDA) [6] to model topic evolution over time. Blei *et al.* assume that the topics in one year are dependent on the topics in the previous year, while Wang and McCallum assume that each topic has its own distribution over time.

A number of methods have also been proposed to cluster documents by temporal information. For example, Alonso *et al.* [2] propose to perform search result clustering by constructing a timeline for each document based on the temporal information extracted. Similarity of a pair of documents depends on how much their timelines overlap with each other. Qamra *et al.* [22] propose a time-sensitive and community-sensitive model for clustering blog posts, basing their model on a modified time-sensitive Dirichlet process [29]. In addition, Cooper *et al.* [9] describe a method for clustering photos based on their timestamps. Readers who are interested in an overview of the challenges and research directions related to temporal information extraction, processing and analysis should refer to the recent work by Alonso *et al.* [3].

In this work we rely on temporal information extraction from text collections, for which various methods have been proposed [16, 18, 24, 27]. Strötgen and Jannik [24] demonstrate a system for extraction, querying, storage, and exploration of spatio-temporal information stored in text documents. Our main focus is however not on the extraction of temporal infomarmation but on studying how it is used as a representation of collective memories of societies.

3. METHODOLOGY

Our aim is to apply computational methods to study history on a large scale, especially by analyzing huge corpora of documents that are dated and contain time expressions referred to the past. In this section, we describe our methodology in detail. Figure 1 shows an overview of the procedures we will follow in this paper.

3.1 Data Collection

To obtain data for our purpose of studying how the past is remembered, we first collected from Google News Archive[2] a dataset of news articles published in the period of 1990-2010. This period was chosen because for the majority of news articles in this period that are indexed in the archive, we were able to obtain the full texts in digital form from the Web, while most articles published earlier are only available in image format, which cannot be subject to text analysis.

News articles were collected by issuing 32 different country names as queries with the above time constraints to Google News Archive.[3] Table 1 shows the list of countries we used to collect our dataset. We decided to focus on countries because they are related to diverse topics and events. They also provide meaningful results when contrasted with one another. We note however that many other kinds of entities such as area names, company names and person names can

[2] `http://news.google.com/archivesearch`

[3] Our list includes names that are not countries (e.g. Hong Kong). However, to give a more concise discussion, we will refer to all these entities as countries in this paper.

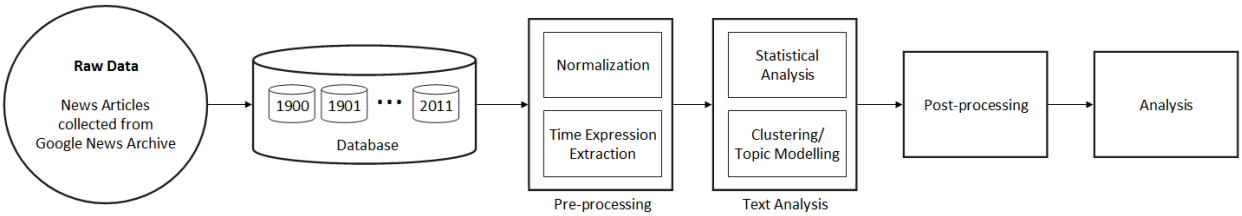


Figure 1: An overview of the methodology used in this study. Data are collected from an online news archive. Documents are then subjected to pre-processing and algorithms are applied to extract time expressions in the texts. Data mining techniques such as topic modeling are then applied to extract meaningful patterns from the data, which are used in further analysis.

Argentina	Australia	Austria	China
Egypt	France	Germany	Hong Kong
Iceland	India	Indonesia	Italy
Iran	Iraq	Ireland	Israel
Japan	New Zealand	Norway	Poland
Romania	Singapore	South Africa	South Korea
Soviet Union	Spain	Sweden	Switzerland
Taiwan	Tunisia	Turkey	United States

Table 1: The list of country names used in collecting news articles.

be used as the subjects of our investigation, and we plan to explore these possibilities in the future.

For each query, we gathered all the search results with links to the original articles, and downloaded their original Web pages. For a small percentage of news articles, we were unable to collect their full texts due to subscription restrictions and the lack of textual content. In these cases, we collected the article abstracts instead. On average each query resulted in 72K news articles. In total, our dataset consisted of 2.4 million news articles.

In addition to news articles published in the period of 1990-2010, we also created a reference dataset by collecting news articles in the period of 1900-1989. For each query listed in Table 1, we downloaded the snippets and titles of all articles published from 1900 to 1989 that are available in the Google News Archive. The purpose and use of this dataset will be further discussed in Section 4. All data collected were organized and indexed in a database for future analysis.

3.2 Pre-processing

News articles collected in the previous step were mainly in the form of Web pages. To extract useful parts from these pages, we processed each article such that HTML tags, JavaScript codes, and other non-content elements were removed. We then extracted the core part of the news articles based on identifying the largest chunk of text in each article. This heuristic worked well due to relatively simple layout of news articles. It allowed us not only to recover the main content of each news article, but also to remove many noisy temporal expressions such as copyrights dates or dates used for labeling archival content. We also discarded articles written in languages other than English using text categorization algorithm based on n-gram matching [7].

3.2.1 Extracting Temporal Expressions

A more important pre-processing task in our case is to extract temporal expressions from news articles. We need to know whether a document mentions something in the past and which time is mentioned. For example, we need to recognize the year 1945 in the sentence "The Second World War ended in 1945," in order to create an association between the year 1945 with the topic mentioned in the sentence.

To extract temporal expressions, we use the GUTime tagger [16], which is a temporal tagger for identifying and normalizing temporal expressions in text. It is one of the several components in the TARSQI toolkit [27], which can be used to extract events and their associations with temporal expressions. GUTime identifies temporal expressions in text and annotates them with the TimeML markup language which is ISO standard for robust specification of events and temporal expressions in natural language. Currently it is the most effective and state-of-the-art solution for the temporal information extraction and processing.

GUTime is able to detect both absolute and relative expressions. Absolute temporal expressions are defined as expressions that are unambiguously associated with a given time point or interval (e.g., 2nd November 1977, 1964). Relative temporal expressions, such as "last year" and "10 years ago", require a reference time expression called anchor in order to be converted into absolute time expressions. GUTime uses certain procedures to resolve relative temporal expressions such as using article timestamp and absolute dates that appear in the context of a relative expression. In this work we focus on both absolute and relative temporal references referring to past years. In our framework, for simplicity we use only temporal expressions of yearly granularity; we thus map references such as "2nd November 1977" or "March 1977" to "1977". However, we are aware that by this simplification some information is lost. In the future we plan to use temporal expressions with their original granularities.

Applying the GUTime tagger to our dataset, we extracted on average 19K temporal expressions that relate to the period 1900-1989 for a single country, and in total, 630K temporal expressions for all the countries. The period of 1900-1989 is the target time frame of our study.

3.3 Text Analysis

In order to obtain a summary of the topics mentioned in a corpus, latent Dirichlet allocation (LDA) [6] is commonly used. LDA is a generative model for obtaining the probability distributions of a chosen number of topics for both words and documents in the corpus. It assumes that each word in a document is generated by first picking a topic (from a set of finite number of topics) and then sampling a word from the word distribution of the selected topic. By training an

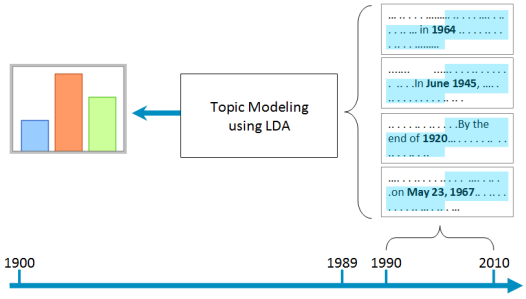


Figure 2: A diagram showing how we applied LDA on documents created by first extracting sentences containing temporal expressions. These temporal expressions were extracted from news articles published in the period of 1990-2010. The text around a temporal expression (the sentence containing it, the preceding and the next sentences) forms its context.

LDA model on a corpus, we will be able to obtain a set of topics with their word distributions, as well as the topic distribution of each document.

We are aware of some useful extensions of the original LDA model. For example, the Topic over Time (TOT) model described in [28] incorporates temporal information such that topic discovery is influenced not only by word co-occurrences but also the timestamps of the documents. However, since these more advanced models impose various time constraints when assigning words to topics, they are not suitable in our study. Instead, using a method similar to the one described in [12], we first use LDA to discover topics in news articles, and then perform post-hoc calculations to obtain different probability distributions related to the publication years of the news articles and the years in the past for further analysis.

We note an important consideration when applying LDA to our dataset. While a news article in general focuses on a specific topic, not all words are relevant when a particular temporal expression in the article is considered. Hence, to identify topics that are more likely to be about past events, we constructed a new document for each temporal expression identified in the dataset. Such a new document contains the sentence in which a temporal expression referring to the past was found, together with the preceding and following sentences. These sentences serve as the *context* of the temporal expression found. LDA was applied to these documents instead of the original news articles. Figure 2 depicts the idea of this procedure. Since LDA only returns the topic distribution of a document and the word distribution of a topic, we estimated other distributions, such as distributions of topics over the past years and over the recent years, for further analysis. The results are presented in the next section.

4. EXPERIMENTS AND ANALYSIS

4.1 Distribution of Past References

Firstly, we investigate to what extent different periods in the past are remembered. We expect to obtain a different picture for different countries. We also want to test and quantify the intuitive assumption that the distant past is on average remembered less than the recent past. From news articles published in the period of 1990-2010, we extracted all temporal expressions that refer to any year in the period of 1900-1989. Figures 3(a) and (b) show the distributions of these temporal expressions for some selected countries. We can see that depending on the country there are peaks at different years. These can be considered as years that are more remembered for the countries.

For example, if we look at the distributions of references for European countries in Figure 3(a), peaks can be found in the years during the Second World War for most countries. However, there are individual differences. For example, 1939 is the year of invasion of Poland by Germany, and it is particularly significant to Poland. On the other hand, 1933 marks the date of appointment of Hitler for Chancellor of Germany, and thus this year has a more special meaning to Germany. There are of course peaks referring to other events, although less prominent. For example, the peak in 1975 for Spain reflects the significance of the death of the dictator Francisco Franco. Sweden has a peak in 1958, when the country hosted the World Cup.

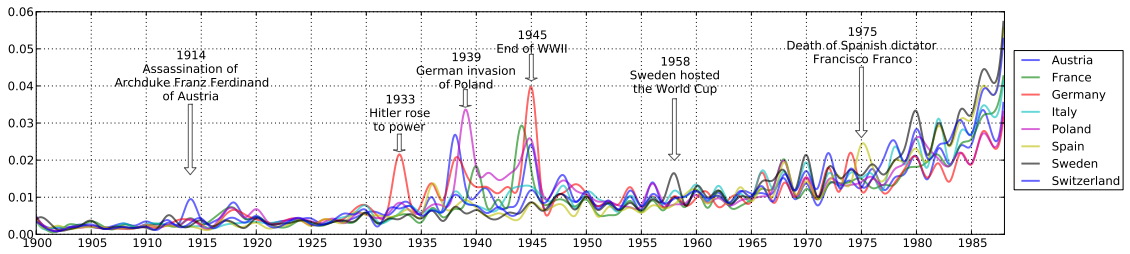
In Figure 3(b), we see a different picture. WWII is no longer a dominating event, although there is still a significant amount of references referring to its end. We can easily observe several important years in the Chinese history. For example, 1949 saw the end of the civil war in China, the founding of the People's Republic of China and the retreat of the Republic of China administration to Taiwan. In 1972, President Nixon visited China, which eventually led to the establishment of formal diplomatic relations between the US and China in 1979. The distribution of references for Hong Kong has a small peak in 1918, in which the Spanish Flu broke out. This is probably due to the fact that the city had several outbreaks of Avian Flu in 2003, and that the media referred to a related outbreak in the history when reporting these events.

Putting aside the occasional peaks in these graphs, we can see that they generally follow a similar shape. Figure 3(c) shows the graph averaged over all countries in our dataset. We notice that the decline in the number of past references resembles an exponential function. The graph can be compared to the well-known *forgetting curve* proposed by Ebbinghaus [10] to illustrate the decline of memory retention in time of individual persons.

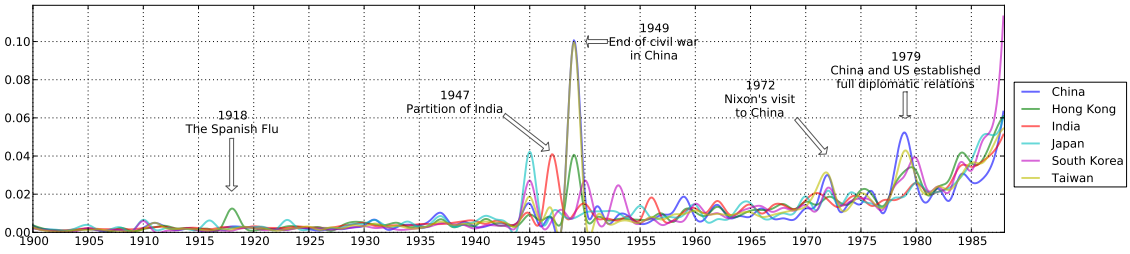
We next investigate the change of the remembering patterns over time. The objective is to check whether the shape of the average remembering curve will change when it is plotted using data created in different time frames. We thus divide our dataset into four bins depending on the article publication date to represent every five consecutive years and we plot the average frequency over all the countries in Figure 3(d). The results confirm that the remembering pattern that we showed in Figure 3(c) is universal, at least over the time frame of 20 years. Naturally, the plot obtained for the first bin (1990-1994) results in the highest values due to its proximity to the period under analysis. However on a relative basis, the average shape of the function that governs remembering does not change much for the data extracted from different bins.

4.2 Remembering Topics/Events in the Past

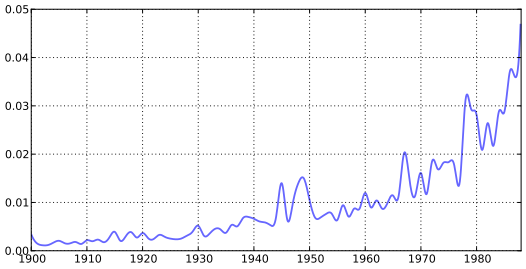
In the previous section, we described analysis that gave us



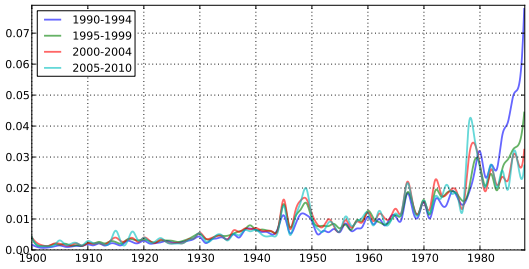
(a) European Countries



(b) Asian Countries



(c) Average across all countries



(d) Average for four different periods

Figure 3: The distribution of years mentioned in news articles that are related to different countries and are published in the period 1990-2010. (a) shows the graphs of some European countries, while (b) shows the graphs of some Asian countries. We observe a general decaying curve but also peaks at some eventful years for different countries. (c) shows the overall average distribution, and (d) shows the average distributions for four different periods of article publication dates.

a general picture of the characteristics of temporal references in news articles. However, we would like to go a step further to gain more insight into how the past is remembered. In particular, the graphs in Figure 3 only told us which years were mentioned more frequently than other years given a country, but they did not tell us why and when they were mentioned. We were able to name the events in Figure 3 only because we have some knowledge of the past. It would be very much desirable if we can identify automatically the topics associated with those years, and more interestingly, when and why were these topics and years mentioned in the news articles.

For this purpose, we now turn our focus to the topics or events mentioned in the news. We mainly rely on the results of topic modeling using LDA to perform post-hoc estimation of various probability distributions in order to study the relations between topics and years mentioned. Note that we use topics and events interchangeably here, although they are not strictly equivalent. However, since we choose a relatively large number of topics in running LDA, most topics do correspond to individual events. Incorporating event extraction analysis into our methodology will be considered in our future work.

LDA is a generative model of documents. Assume that we have a corpus $\mathcal{C}$ containing a set of documents, and each document d is represented by a bag of words. LDA assumes that every word in a document is generated by first picking a latent topic and then sampling from the distribution of words conditioned on the selected topic. Probability distributions obtained after training an LDA model include $P(w|z)$, the probability word w given topic z, and $P(z|d)$, the probability of topic z given document d.

In addition, we introduce the following notations. We use p to represent an article publication year in the period of 1990-2010, and y represent a year in the period of 1900-1989. For each country, we can estimate the distributions $P(y)$, $P(p)$ and $P(p, y)$ directly from the dataset by counting the occurrences of years in these periods. In the following, we combine these distributions with the distributions returned by LDA to perform various kinds of analysis.

We first focus on one country at a time and examine which years and topics are particularly important. Secondly, we investigate the differences between news articles in the past

and news articles in recent years that refer to the past. Finally, we explore how topic modeling can be used to reveal the similarities between the histories of different countries.

4.2.1 Significant Years and Topics

Every country has her own historic moments and events. The significance of a particular year or a particular event to a country can, to certain extent, be measured by how frequently they are mentioned in the news. In the following, we examine how these characteristics of remembering can be revealed by making use of the results of topic modeling.

Our first step is to determine the significant years for a given country. This can be estimated in a straightforward way, by counting how many times each year has been mentioned in news articles in the respective dataset. Furthermore, we would like to know the reason why a particular year has been mentioned frequently. For this purpose, we estimate the topic distribution of each year in the past. Let $P(z|d)$ be the topic distribution of each document returned by LDA, and let D_y be the set of documents mentioning the year y. We can estimate the topic distribution of each year, i.e. $P(z|y)$ using the following equation:

$$P(z|y) = \frac{1}{|D_y|} \sum_{d \in D_y} P(z|d). \quad (1)$$

Table 2 shows the results of this experiment performed on 10 selected countries. For each country, we show the three most significant years, as well as a list of top words from the topic with the highest probability conditioned on the respective years. An immediate observation is that years related to WWII, such as 1939, 1944 and 1945, are significant to quite a number of countries on the list. This is expected since WWII involved most of the world's nations at that time and can be considered as one of the most significant events in the 20th century. However, similar to what we have discussed in Figure 3, we also observe obvious variations across different countries. The end of the civil war in 1949 tops the list of China, while for Spain the year 1975 in which the dictator Francisco Franco died is the most significant.

Overall, we observe that topic modeling is very useful in this case since it tells us clearly why a particular year is considered significant to a given country. It clearly demonstrates the power of computational tools in assisting historical studies. While some of the results are not too surprising given our general knowledge of the histories of these countries, not all of them are immediately obvious to a person who is not familiar with a country's past. We also note that the consideration of using only words appearing close to a temporal expressions in topic modeling results in very coherent topics in many cases.

4.2.2 Triggers of Remembering the Past

As we have mentioned above, the significance of a historic event can be context-dependent. The above analysis with the help of topics tells us the reason why a particular year is significant to a country. However, it does not tell us *when* a year in the past was recalled, and *what* triggered the recalling. A way to look into this characteristic is to see in what particular period of time was the event mentioned in the news. We can study this by estimating the probability distribution of the publication dates of news articles conditioned on a particular year in the past and the corresponding significant topic. We represent this probability distribution by $P(p|y,z)$, where p refers to a publication year. In our case, $1990 \leq p \leq 2010$. We use the following equation to estimate $P(p|y,z)$:

$$P(p|y,z) = \frac{P(p,y,z)}{P(y,z)} \quad (2)$$

$$= \frac{P(z|p,y)P(p,y)}{P(z|y)P(y)} \quad (3)$$

where $P(z|p,y)$ is estimated using equation similar to Equation 1, and $P(p,y)$ and $P(y)$ can be easily obtained by counting the number of documents published in year p and the number of documents mentioning year y.

Figure 4 shows the probability distribution over the twenty-year period for the most significant years and topics for four selected countries. As we have hypothesized, the distribution is not uniform. In other words, a historical event is usually mentioned more frequently in a particular period of time than in other periods. For example, for Japan the year 1972 was mentioned very frequently in the year 1998 (with high probability at $p = 1998$). This is because in both years the Winter Olympics was held in Japan.

For Japan, Germany and Poland, we observe peaks in the years 1995 and 2005 for the year 1945; obviously this reflects the fact that the end of WWII was remembered and mentioned more frequently in its 50th and 60th anniversaries. For China, the year 1949 was remembered relatively more in and after the year 2004. In 2004, Taiwan held its presidential election, which probably invoked various discussion postulating the development of the mutual relationship between China and Taiwan, and thus triggered the remembering of the end of the civil war in 1949.

To better assist the analysis of correlations between years in recent times and years in the past in context, we can use the topic model to generate topic distributions conditioned on a pair of past-recent years. In other words, we estimate $P(z|p,y)$, the topic distribution given a year in recent time (the publication date of a news article) and a year in the past. This can be estimated using the following equation:

$$P(z|p,y) = \frac{1}{|D_{p,y}|} \sum_{d \in D_{p,y}} P(z|d). \quad (4)$$

where $D_{p,y}$ represents the set of documents that were published in the year p and contain a reference to the year y.

Table 3 shows four examples of the most significant topics (by probability) for frequent pairs for four different countries. The context of the connections between two years is immediately understood by studying the topic words. For example, for Indonesia, 1975 and 1999 are both significant years regarding its relationship with East Timor, which was annexed by Indonesia in 1975 but claimed independence in 1999. On the other hand, the Polish Pope John Paul II, who was originally known by the name of Karol Wojtyla, was elected as the Pope in 1978 and passed away in 2005. He is considered a significant figure in Polish history.

4.2.3 Events Remembered and Forgotten

So far we have examined which topics and years were frequently mentioned in news articles of recent times. However, it is not uncommon that references to the past can be selective. In other words, while a year in the past was eventful, very often only few events were singled out and remembered after a long period of time. It might even be possible that

Country	Year	Top Words
Argentina	1978	world, cup, first, team, maradona, win, tournament, home, won, years
	1976	military, said, rights, human, dictatorship, argentine, government, years, war, people
	1974	pinochet, court, chile, chilean, regime, prats, death, charges, supreme, opponents
Australia	1975	east, timor, indonesia, australian, indonesian, invasion, timorese, sea, killed, agreement
	1974	world, cup, team, first, won, played, final, win, africa, tournament
	1980	states, united, australian, open, year, title, wimbledon, slam, grand, won
China	1949	taiwan, mainland, war, island, beijing, civil, communists, nationalist, government, chinese
	1972	president, visit, nixon, minister, beijing, years, relations, first, kissinger, trip
	1978	million, year, economic, billion, percent, growth, world, economy, years, population
France	1944	war, world, army, american, battle, soldiers, legion, served, veterans, french
	1940	war, french, german, germany, hitler, occupation, world, resistance, nazi, britain
	1968	killed, people, group, spain, attack, eta, basque, region, year, police
Germany	1945	war, world, end, day, second, declared, allies, europe, first, empire
	1939	soviet, poland, union, europe, war, eastern, western, czechoslovakia, polish, invaded
	1974	world, cup, final, england, team, first, won, second, win, football
India	1947	pakistan, kashmir, wars, fought, war, region, countries, independence, nations, territory
	1971	pakistan, bangladesh, war, military, pakistani, army, said, east, forces, report
	1962	china, border, war, chinese, relations, countries, beijing, area, territory, recently
Japan	1945	nuclear, hiroshima, atomic, bomb, nagasaki, weapons, august, dropped, bombs, bombing
	1972	world, tokyo, games, gold, won, cup, olympics, olympic, event, asian
	1980	percent, year, oil, million, said, yen, economy, cent, rate, billion
Poland	1939	war, hitler, germany, invasion, britain, invaded, france, german, september, world
	1945	camp, concentration, auschwitz, camps, nazi, death, nazis, sent, january, prisoners
	1980	communist, solidarity, walesa, gdansk, workers, movement, union, leader, government, lech
Spain	1975	franco, death, francisco, war, country, democracy, spanish, dictator, civil, dictatorship
	1968	basque, people, eta, group, killed, northern, campaign, state, independent, said
	1978	world, cup, european, team, club, championship, final, group, real, won
USA	1979	iran, said, relations, iranian, embassy, government, revolution, islamic, diplomatic, tehran
	1980	team, first, world, cup, olympic, time, games, americans, medal, won
	1972	president, treaty, soviet, nuclear, weapons, said, missile, union, washington, first

Table 2: Frequently mentioned years and topics in the news articles for different countries. For each country, we show the three most frequently mentioned years in the news articles, and the corresponding top words from the topic with the highest probability conditioned on the given year.

Country	p	y	Topic
Indonesia	1999	1975	east, timor, indonesian, portuguese, invaded, timorese, colony, invasion, territory, annexed
Israel	2009	1967	palestinian, state, said, hamas, borders, end, solution, abbas, final, negotiations
Italy	2009	1980	people, southern, region, naples, town, europe, earthquake, south, killed, die
Poland	2005	1978	pope, john, paul, catholic, krakow, church, visit, first, wojtyla, karol

Table 3: Topics for frequent pairs of past-recent years for four selected countries.

some events that were widely reported in the past eventually were forgotten over time.

To explore this characteristic, we run topic modeling over a combined dataset that contains both news articles from the recent times (1990-2010) and from the past (1900-1980).[4] Firstly, we present three examples (Figure 5) that visually compare two topic distributions, namely $P_p(z|y)$, the topic distribution given a year in the past based on past data, and $P_r(z|y) = \sum_{y' \in Y_r} P(z|y')P(y')$, the distribution given a year in the past based on recent data, where Y_r is the set of years in the period of 1990-2010.

[4]We use the reference dataset described at the end of Section 3.1. Note that we purposely limited the past period to 1900-1980 so as to have more distinct separation between the data from the past and data from the present in the subsequent analysis.

We observe from Figure 5 that there is an obvious difference between the two distributions. For the distribution obtained by running topic modeling on recent news articles, we usually observe a few spikes. These represent a few important topics/events in that particular year as judged in the recent times. However, while we also observe relatively high probabilities in a few topics in the distributions obtained from the news articles in the past, the distributions are generally more uniformly distributed. For example, referring to Figure 5(a), most news articles considered the year 1978 an important year after the Cultural Revolution in China, while at that time the attention of the media seemed to have focused on other events, such as the signing of the Treaty of Peace and Friendship between China and Japan.

Another observation is the difference between distributions in the past and in the recent times actually changes as we move towards the present times. For each year y in the period 1900-1980, we compute the KL-Divergence between the topic distributions $P_p(z|y)$ and $P_r(z|y)$. The results for two countries, China and Japan, are shown in Figure 6. An interesting observation can be made from this figure. In both cases we observe a gentle downward trend. This suggests that the difference between two distributions is becoming smaller as we move along the timeline. One explanation of this result is that news articles are more selective when mentioning events in the distant past. For example, only one or two topics related to the year 1905 may be frequently mentioned in the news in the recent times, while news articles may refer to more diverse topics related to the year 1980. This result seems to support the idea that attention would eventually focus on only a few topics/events that are considered significant as time passes.

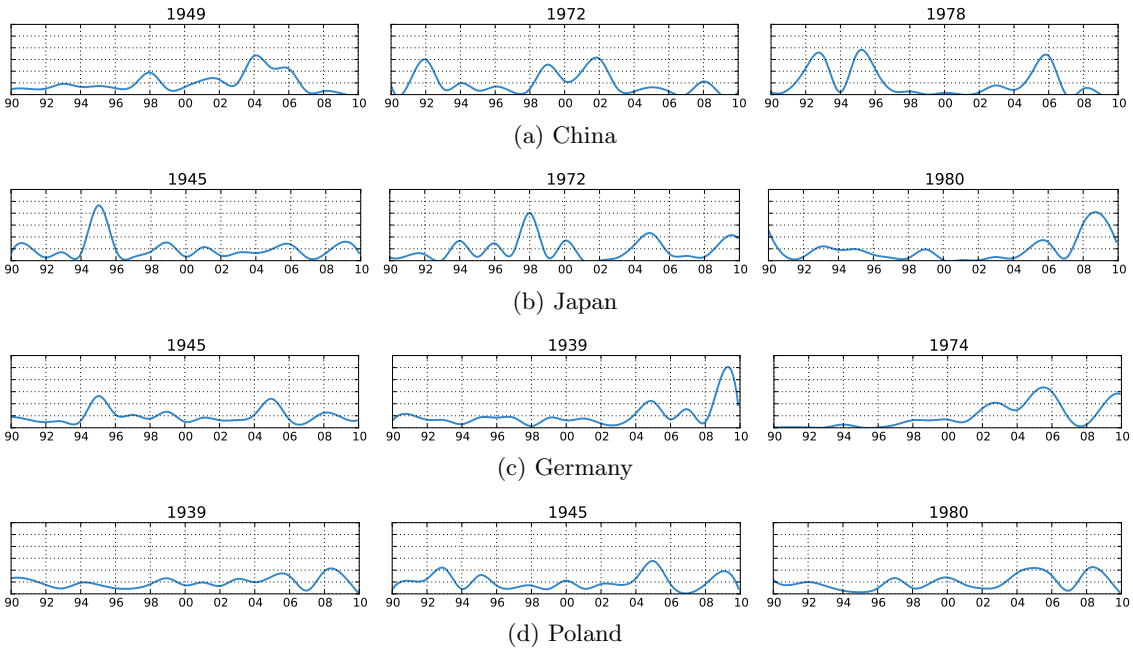


Figure 4: The distribution of the years of publication for selected countries. For each selected country, we pick the three most frequently mentioned years in the past and then the topic with the highest probability given a year. We then estimate $P(p|y,z)$, the probability of the years of publication conditioned on the year in the past and the topic. In other words, the above graphs show the extent to which each year in the past and the corresponding topic was mentioned in the period of 1990 to 2010 (c.f. Table 2).

From an exploratory perspective, we note that it is also possible to contrast how historical entities such as persons were mentioned in documents in the past and how they are remembered in recent times. Figure 7 shows a comparison of the distribution of the term "Hitler" in past and recent news articles, with the targeting years in the period of 1900-1989. We can observe that the year 1933 has higher significance in the distribution obtained from the present articles than from the past articles. This can be explained by the fact that the year is now commonly recognized as the major turning point when Hitler rose to power. Also some other events from his early life are reflected in the graph generated from the recent articles.

4.2.4 *Historical Similarity of Countries*

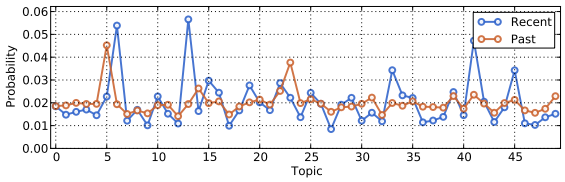
One may say that a nation or a country is defined by its history. For some countries, their histories overlap one another from time to time, while others find their histories run on parallel lines. Looking at the histories of different countries thus represents one of the many ways to study the similarities and the relations between two countries. In this section, we try to explore how topic modeling allows us to discover countries with similar "historical narratives". Firstly, we pool data of all countries together and perform topic modeling using LDA. Using the result of topic modeling, we characterize each country with $P(z,y|c)$, the joint distribution of topics and years in the past conditioned on a country c. Similarity is then defined as the reciprocal of the KL-Divergence between the distributions of two countries.

Figure 8 shows the similarity matrix obtained based on the above method. One intuitive observation is that if countries are geographically close to one another, their historical views, or more exactly their perceptions of the past, are likely to be more similar to one another. For example, we observe high similarity among Germany, France, Italy and Poland. However, we also find that the similarity matrix is dominated by major events that are common across many countries, most noticeably WWII. Thus, it becomes difficult to determine the similarity among smaller countries such as Singapore and Iceland. We believe that several procedures can be used to reduce the dominating effect of major historical events, such as by performing pairwise comparison, or by normalizing the importance of different topics. We plan to investigate these possibilities in the future. For now, we observe that this approach is potentially useful in exploring historical similarity among countries.

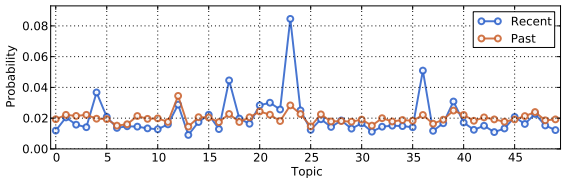
5. DISCUSSION

We note that news articles only represent one source of data, and we do not exclude the usefulness of other data sources. One may consider that since the objective of a news article is to report a certain event of the present time, it may not necessarily reflect how people perceive the past. However, news articles are produced by the media which can be strongly influenced by the interests and attention of the people. Thus, while a reference to the past in a news article is usually invoked by a certain event that happens when the article is written, the reasons that a certain event in the past is mentioned are also tightly connected to the collective memory of the people. Hence, news articles represent a suitable proxy for studying how the past is remembered.

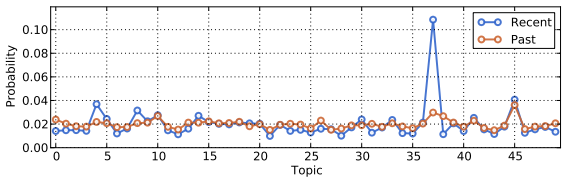
The application of computational tools such as topic modeling to assist historical studies can be used to answer many



(a) China 1978

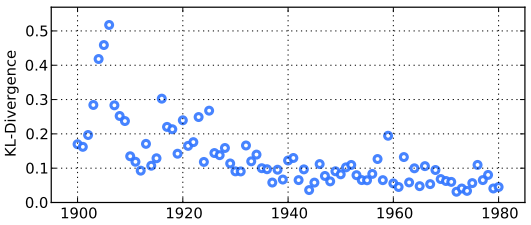


(b) Japan 1972

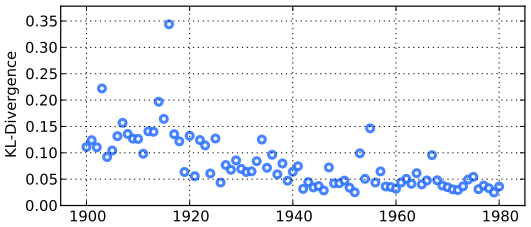


(c) Poland 1980

Figure 5: The distribution of topics for selected country-year pairs. For each country-year pair, we plot the distribution of topics conditioned on the recent years (1990-2010) and the particular year in the country-year pair. The graphs therefore reflect the differences between the topics/events happened in that particular year and the topics/events remembered in recent news articles when the particular year was mentioned.



(a) China



(b) Japan

Figure 6: Scatter plots of KL-Divergence between the distribution of topics in recent and past datasets for China and Japan. We observe in both cases that the difference between the two distributions decreases as we move towards the present time.

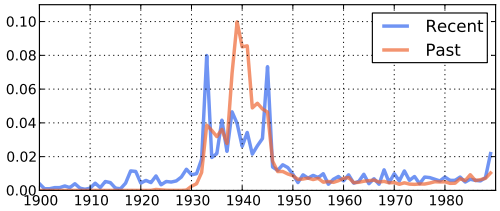


Figure 7: The distribution of occurrences of the word "Hitler" in articles about Germany. The "Recent" line refers to statistics obtained by analyzing the dataset we have collected, while the "Past" line refers to the normalized hit counts in each year returned by Google News Archive when "Germany" and "Hitler" are used together as a query.

more questions than those we have considered in this paper. We outline a few possible research directions.

Firstly, while news article provide the benefit of objectivity when compared to other genres like books or blogs, they only give part of the whole picture of the past and our perception of the past. To be more comprehensive, we plan to take other sources of data, such as magazines, textbooks, fictions, books and Web pages into consideration in the future.

As far as language is concerned, we only focused on news articles published in English this time. An even better way to probe the social memories of different countries is to analyze news articles and other datasets published in their own languages. It would be interesting to perform different kinds of analysis that compare and contrast the distributions of topics and periods obtained from different languages, thus revealing not only historical perspectives but also differences in cultural understanding.

Other computational tools can also help us gain more sight. For example, we can use sentiment analysis [21] to study whether the sentimental view on some historical events have changed over time, that is, whether people regard certain events with the same sentiment as they were seen at the time of their occurrence. By considering the source of data separately, we can even investigate the biases of these sources (e.g. newspapers vs. books). It is also desirable to explore new methods of text mining that are particularly suitable for computational history. For example, we may need algorithms to extract not only explicit references to the past (as in the form of temporal expressions such as past years), but also implicit historical references embedded in a text (e.g. mentions of historical persons, events).

In terms of applications, topic models can be used to assist the estimation of the *focus time* of Web documents, thus improving temporal information retrieval [3, 13]. Focus time (in contrast to document timestamp) is defined as the union of time periods to which the content of the document refers. For example, the focus time of an article about the attack on Pearl Harbor would be December 7th, 1941. Given topic models for each past year, one can estimate the focus time of a document by inspecting its probability distribution over topics. With the estimated focus time documents matching implicit time frames behind user queries can be retrieved.

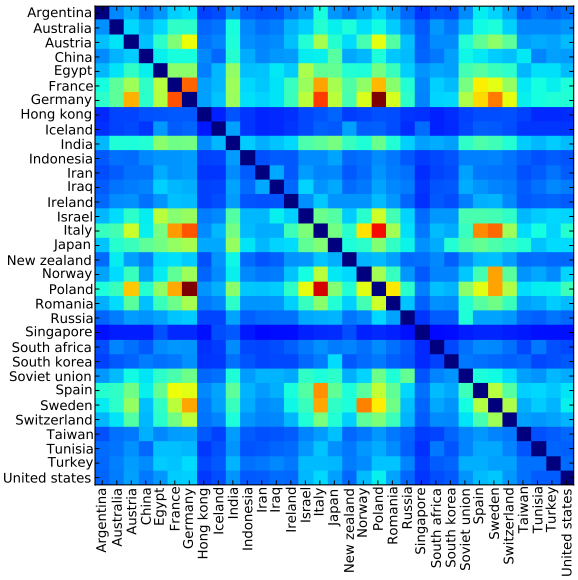


Figure 8: A similarity matrix of countries obtained by comparing their topic distributions in each year. A red cell indicates high similarity, while a blue cell indicates low similarity. Similarity is measured by the reciprocal of the KL-divergence between two topic distributions.

6. CONCLUSIONS

History plays an important role in our society. It is a subject that is taught since early stages of education in most nations. Computational history as a interdisciplinary field that aims at harnessing computational power to support history analysis is very appealing. In this paper, we described our effort in this direction in the context of collective social memories, and presented some interesting results obtained by applying text mining techniques to a large scale corpus of news articles. In the future, we plan to extend our work in the directions mentioned above.

7. ACKNOWLEDGMENTS

This research was partially supported by the MEXT Grant-in-Aid for Young Scientists B (#22700096) "Towards time-focused Web Search and Mining".

8. REFERENCES

[1] J. Allan. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[2] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *CIKM '09*, pages 97–106, 2009.

[3] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. Temporal information retrieval: Challenges and opportunities. In *TWAW '11*.

[4] K. Berberich, S. J. Bedathur, M. Sozio, and G. Weikum. Bridging the Terminology Gap in Web Archive Search. In *WebDB '09*, 2009.

[5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML '06*, pages 113–120, 2006.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[7] W. B. Cavnar and J. M. Trenkle. N-Gram-Based Text Categorization. In *SDAIR-94*, pages 161–175, 1994.

[8] P. Conenrton. *How Societies Remember*. Cambridge University Press, 1989.

[9] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM Trans. Multimedia Comput. Commun. Appl.*, 1:269–288, August 2005.

[10] H. Ebbinghaus. *Uber das Gedchtnis. Untersuchungen zur experimentellen Psychologie*. Duncker and Humblot, Leipzig, 1885.

[11] M. Halbwachs. *On collective memory*. The University of Chicago Press, 1992.

[12] D. Hall, D. Jurafsky, and C. D. Manning. Studying the history of ideas using topic models. In *EMNLP '08*, pages 363–371, 2008.

[13] A. Jatowt, Y. Kawai, and K. Tanaka. Calculating Content Recency based on Timestamped and Non-Timestamped Sources for Supporting Page Quality Estimation. In *SAC '11*, pages 1156–1163, 2011.

[14] N. Kanhabua and K. Nørvåg. Exploiting time-based synonyms in searching document archives. In *JCDL '10*, pages 79–88, 2010.

[15] D. Lazer *et al.* Computational social science. *Science*, 323(5915):721–723, 2009.

[16] I. Mani and G. Wilson. Robust temporal processing of news. In *ACL '00*, pages 69–76, 2000.

[17] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[18] B. Martins, H. Manguinhas and J. Borbinha. Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. In *ICSC '08*, pages 1–9, 2008.

[19] J.-B. Michel *et al.* Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, January 2011.

[20] P. Nora. Between Memory and History. In *Representations*, volume 26, pages 7–25, 1989.

[21] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, 2008.

[22] A. Qamra, B. Tseng, and E. Y. Chang. Mining blog stories using community-based and temporal clustering. In *CIKM '06*, pages 58–67, 2006.

[23] D. Shahaf and C. Guestrin. Connecting the dots between news articles. In *KDD '10*, pages 623–632, 2010.

[24] J. Strötgen and G. Michael. TimeTrails: a system for exploring spatio-temporal information in documents. In *VLDB '10*, pages 1569–1572, 2010.

[25] N. Tahmasebi, K. Niklas, T. Theuerkauf and T. Risse. Using Word Sense Discrimination on Historic Document Collections. In *JCDL '10*, pages 89–98, 2010.

[26] Y. Takahashi, H. Ohshima, M. Yamamaoto, H. Iwasaki, S. Oyama, and K. Tanaka. Evaluating significance of historical entities based on tempo-spatial impacts analysis using wikipedia link structure. In *HT '11*, pages 83–92, 2011.

[27] M. Verhagen and J. Pustejovsky. Temporal processing with the tarsqi toolkit. In *COLING '08*, pages 189–192, 2008.

[28] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD '06*, pages 424–433, 2006.

[29] X. Zhu, Z. Ghahramani, and J. Lafferty. Time-sensitive dirichlet process mixture models. Technical report, Carnegie Mellon University, 2005.