Exploratory chronotopic data analysis

Benjamin Adams and Mark Gahegan

Centre for eResearch
The University of Auckland
Auckland 1142, New Zealand
{b.adams,m.gahegan}@auckland.ac.nz

Abstract. The intrinsic connection between place, space, and time in narrative texts is the subject of *chronotopic* literary analysis. We take the notion of the chronotope and apply it to exploratory analysis of unstructured big data. Exploratory chronotopic data analysis provides a data-driven perspective on how place, space, and time are connected in large, crowdsourced text collections. In this study, we processed the English Wikpedia text to find all co-occurrences of named places and dates and discovered that times are linked to places in a large majority of cases. We analyzed these millions of connections between places and dates and discovered a number of interesting trends. Because of the scale of the data involved, we suggest that chronotopic data analysis will lead to the development of new data models and methods for geographic information science and related fields, such as digital humanities.

Keywords: place, time, chronology, historical geographic information science, big data, volunteered geographic information

1 Introduction

Although human history is a continuum of events and processes happening over time and space, when writing about history people structure historical information using discrete times and places as anchors. Wars are fought between countries, cities specialize in industries, historical eras are described at the granularity of centuries, and decades are characterized particular cultural or social movements. In popular historical writing it is common to talk about places having golden ages like Athens, Greece in the 4th century BCE or important seminal events in the history of place, such as the D-Day invasion in Normandy. How we refer to places and times together helps to create a conceptual framework for history. But how do we refer to places and times? There is scant research on this question from a data-driven perspective, looking at the integrated dynamics of spatial and temporal references in a large corpora of text. The availability of many such corpora, improvements in geographic and temporal parsing of natural language, and the ability to support the associated algorithms and data structures on high-performance computing infrastructure means we have an unprecedented opportunity to explore this topic in new ways.

The deep-rooted connection between representations of time and space in literature has been a focus of literary narrative analysis. The Russian literary theorist, Mikhail Bakhtin, introduced the concept of the *chronotope* to describe how literary genres are characterized by modes of language, which reflect specific spatio-temporal configurations [4]. For example, ancient Greek romances operate on "adventure-time" and are characterized by highly abstract, interchangable representations of times and places in an "alien world" that is not connected with a concrete, familiar landscape and historical timeline. Other works in contrast have more concrete and substantial spatial and temporal structure based on the life course of an individual. In later works there was an effort to merge historical time sequences describing the life of cities, nations, and other social organizations with individual life sequences, though the two sequences are not fused in the sense that they focus on different types of events. The changing ways that people have represented time and space in literature reflect changing conceptualizations of how people live their lives, and shifting cultural attitudes and ideas about the role of the individual and society [5]. Fundamentally, what differentiates chronotopic analysis from other kinds of investigations of place or time in literature is that it is predicated on the idea that spatial and temporal relations and structures in narrative texts are *intrinsically connected*. Thus in chronotopic analysis time and space are not analyzed independently and neither takes precedence over the other. The term chronotope, being an amalgamation of the Greek words for time and space, was inspired by the space-time theories developed in relativity physics in the early 20th century. Although Bakhtin first wrote about chronotopes in 1937, his essay on chronotopes was not published until the 1970s and not translated into English until 1981. But since that time chronotopic analysis has flourished into a broad and heterogeneous field of literary theory.

The development of data models, e.g., space-time prisms, and geographic information systems designed to enable analysis of spatio-temporal phenomena has also been an ongoing research area in GIScience for quite a while [18, 31]. Conventionally, these models extend existing spatial models to include time ('three-plus-one' representations), though there has been some exploration of fully four-dimensional models as well (see [7]). One of the key application areas for such systems is the representation and understanding of human activities and interactions [30]. The application of geographic information science to analyze and represent history has primarily focused on using existing GIS technologies to create historical snapshots of geographic information, e.g., a representation of the boundary of an ancient civilization and the cities within [14]. The use of integrated historical and geographic context can also be used to support geovisual analytics and sensemaking of unstructured information sources [27].

The emergence of new kinds of crowdsourced geographic information (e.g., social media data), which is primarily referenced in terms of named places rather than spatially, has led to research on how to model place-based information [28, 8, 26]. In GIScience this recent interest in modeling place (in contrast to space) has included the notion of representing places in terms of their temporal signatures [29]. And there are examples of using machine learning to infer spatio-temporal patterns in the themes that people write about in social media, for example to detect events [16, 23]. However, most of the research on place in GIScience has focused on gazetteer development as well as the spatial and thematic (or affordance-based) elements of their representation, not in an integrated way that combines space and time [12, 10, 1]. An analytic approach that incorporates the intrinsic connectedness between time and place (or space) in collections of unstructured texts remains largely underdeveloped.

Meanwhile, in recent years there has been growing interest in the use of corpus studies and the exploration of big data to understand broad cultural and sociological trends through the written word and other kinds of media. The Google n-grams project which looks at trends in word use in millions of published books has shown that data-driven analysis can uncover shifts in language use over time and examples of social forces acting to change how people write because of policies, such as censorship [17]. Spatial analysis has also grown in prominence in digital humanities [19].

A research program on chronotopic analysis of large text corpora would provide great value, helping us understand the varied ways in which people conceptualize place and time in an integrated way, which in turn can be used to help us organize historical geographic information. In this paper we carve off a preliminary slice of this research. We report on an exploratory analysis of the millions of references to places and times that are found to co-occur in the English Wikipedia corpus. This analysis provides a window into understanding how the semantics of time are structured in the context of one kind of historical content (crowdsourced, encyclopaedic) about places. This work can be viewed as a first step toward developing a broader methodology of data-driven chronotopic analysis of unstructured text.

In the following section we describe our data processing workflow to match place and temporal references in Wikipedia. In Section 3 we discuss patterns around the use of temporal references alone, and in Section 4 look at patterns in how place and time references co-occur. In Section 5 we discuss the larger implications of this exploratory study for GIScience research and point to future research directions in exploratory chronotopic data analysis.

2 Data processing methods

In this section we describe our methods for identifying place and temporal references and how we matched these references in the text. We leveraged existing open source tools to accomplish this task, but due to the large size of the data, custom analytic scripts were developed to explore the results. For our experiments we used the August 8, 2015 dump of the English Wikipedia, which consists of 7,131,349 articles of which 4,659,056 are actual article pages (i.e., not category, image, or disambiguation pages). The numbers of place and temporal references are both of the same order (in the tens of millions) – see below for more information.

2.1 Temporal tagging

The narrative-style HeidelTime temporal tagger was used to identify temporal tags in the articles [25]. In total **68,657,749 temporal references** were identified within all the main article pages of the English Wikipedia. The existing methods for matching of temporal entities in text are not perfect. There are some false positives that we noticed. For example, references to AM radio station frequencies are often identified as dates. We endeavored to identify and isolate these incorrectly classified entities, but no doubt some noise is still present in the results because of misclassified entities.

2.2 Place tagging

In order to find place references in Wikipedia we used DBpedia data to find all *place* pages and used the links to those pages to identify georeferences in other articles [3]. DBpedia organizes place references into classes, including Country, City, and Administrative Unit as well as other feature types like Museums and Parks. We identified all these place types in the texts, but for the analysis performed in this study we focused on two main categories of places: 1) **Countries** and 2) **Populated places**, corresponding to City, Town, Village, and Administrative Unit features in DBpedia. Table 1 shows the statistics on number of matched places by type, with **31,922,923 place references** identified in total. Since it is customary to make only one link to a referenced page within an entire Wikipedia article, we matched all additional references to place names that were linked at least once in an article. For example, if a page contains a link to the "Rome" page in the abstract, then we also find all other references to Rome in other paragraphs in the article and match those as well. Once these links were identified we removed all Category pages to focus on references in the narrative text of actual article pages.

Place type	**Instances**	**References**	**Articles**	**Pct. articles**
Country	255	6,330,851	1,998,273	42.9%
Populated places (cities, towns, etc.)	273,329	12,450,520	2,527,910	54.3%
Other place types (DBpedia)	351,453	13,141,552	1,900,407	40.8%
Any place types	625,037	31,922,923	3,480,667	74.7%

Table 1: Summary statistics on the occurrences of named place references in the English Wikipedia. The *Instances* column is the count of distinct named places, and the *References* column list the count of how many times a reference of that type is made in the corpus. *Number of articles* shows the total count of articles that reference at least one instance of the place category in the text, and *Pct. articles* is the percentage of all articles that contain a reference of that type.

2.3 Matching places and times

Although Wikipedia articles are crowdsourced and thus can vary in terms of writing style, in most cases the format of the writing in Wikipedia is fairly standardized. In particular, paragraphs tend to be self-contained to a degree that we can use the simple heuristic to match places with times if they are found in the same paragraph.In addition to these matches based on co-occurrence in paragraphs, we also matched temporal references to places when found anywhere within an article about that place (e.g., all dates within the main Wikipedia page for New Zealand are matched to New Zealand). While this will undoubtedly include some false positives in the sense that a place and time might be considered connected even if they are unrelated in the text, given the massive size of the data set these matches will be inconsequential in the overall statistical results. Using this method, 29,265,607 or **42.6% of all temporal references in the English Wikipedia are associated with some named place**, and 19,998,504 or **62.6% of all place references are associated with a temporal reference**. It is clear that place and time are connected concepts across a wide variety of encyclopaedic content. These statistics alone lend credence to the idea that data-driven analysis of time and place references in large text corpora in an integrated manner has the potential to lead us to a richer understanding of the semantics of place and time. In addition, it shows that temporality is at least as important, if not more so, for understanding and representing place as place is for understanding and representing time.

3 Dynamics of date references

In this section we begin the analysis by looking at patterns found in the temporal information on its own. Temporal taggers capture some of the diversity of ways that times are referenced in text. In the TIMEX3 format generated by HeidelTime, a temporal reference type can be DATE, TIME, DURATION, or SET [22]. A TIME reference refers to a time in a day, e.g., 3:45 pm. A DURATION refers to a length of time, such as "for 2 hours". A SET reference is a collection of dates, such as the second Thursday of every month or "annual". A DATE reference is a relative or absolute date based on the Gregorian calendar. The temporal granularity of DATE references ranges from centuries to decades to years through to seasons, months and weeks to individual days and days of the week. In this work our analysis focuses on DATE references, which make up the vast majority of all the temporal references found in Wikipedia. Table 2 shows the summary statistics for these different granularities of date references in the text.

3.1 Decade, year, month, and day patterns

Figure 1 shows a log scale plot of references to decades from the year 1000 to the 2010s. A remarkable feature of this is the identification that the 10s

Temporal Type	Count	Number of articles	Pct. articles	Avg. per article
DATE	59,225,232	4,282,056	91.9%	12.71
TIME	1,029,268	422,923	9.1%	0.22
DURATION	6,867,967	1,876,934	40.3%	1.47
SET	2,102,917	978,907	21.0%	0.45
Any type	68,657,749	4,343,050	93.2%	14.74

Table 2: Summary statistics on the temporal references in the English Wikipedia.

decade of every century is referenced on an order of magnitude fewer times than other decades are. The first decade (00s) of the century is referenced more so than others, however that is likely an artifact of the parser not being able to distinguish between century and decade references in those cases. A plausible reason for the reduction in the 10s is that it reflects the common use of phrases like “the early 1900s” for the first two decades of the century; however, that remains to be evaluated. Ignoring the first two decades of the century, from the early 18th century on there is a steady increase in references to decades, which matches the overall trend for more fine-grained dates as well. In the 20th century a reduction in decade references is found in references to the 1940s as well, which appears to be a result of the events of World War II dominating the structure of temporal references, so that there are more single year references in that decade than others. This is corroborated by Figure 2. That Figure illustrates that the U.S. Civil War and the two World Wars are such dominant topics in Wikipedia, that events are described in finer grained (at the level of days and months) detail for those years. Since 2000 the ratio of day references has increased substantially, so that it is on a trend to eclipse year references. It remains to be seen whether this increase is due to the recency of the dates or whether there is a genuine shift in how we are writing about history due to changes in digital technology and our ability to record temporal events at increasing granularity and precision.

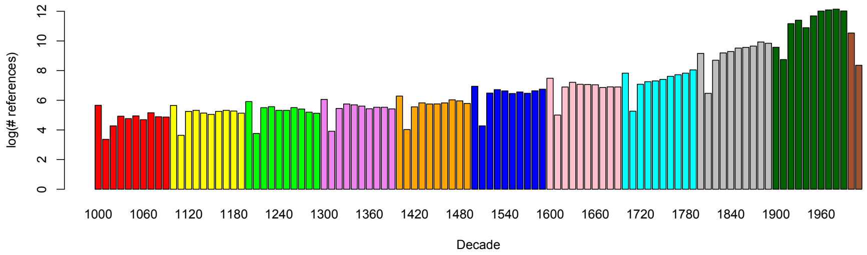

Fig. 1: Log scale plot of the number of references to decades (e.g., “1960s”) from 1000s to 2010s.

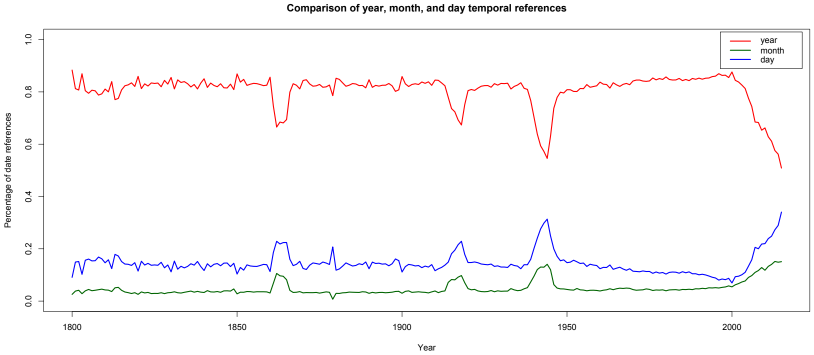

Fig. 2: A comparison of types of temporal references. The U.S. civil war and the two world wars are described in much finer temporal granularity than other years from 1800-2000. For most other years there are approximately 5 times as many year references as single day references.

3.2 Temporal references and human population

The number of temporal references in Wikipedia grows as a function of the date being referenced, which simply means that we've recorded more of our history over time. What is unclear is whether this growth is due to our technical ability to record history with better temporal precision, or if it perhaps reflects other factors as well. To explore this we plotted two ratios in Figure 3. The red line shows human population relative to the population at 1950, so there are approximately 3 times as many people living today as in 1950. The blue line shows the number of temporal references for each year in ratio to the 1950 count. Interestingly, both values grow at the same rate until around 1990, with exponential growth in the number of temporal references from that point on. While this is merely correlation it points to a hypothesis that as population grows the number of interesting events to record grows in the same way, barring any major technological change.[1] The explosion in temporal references is perhaps due to the advent of the Internet, which revolutionized our ability to record history digitally. Wikipedia was not founded until 2001 so long after this increase began. The drop off at the end is most likely due to a lag in recording contemporary events in Wikipedia (and the dump not including the full year of 2015).

4 Place and time together

Chronotopic analysis is based on the premise that there are characteristic space-time configurations that help us understand categories of written texts and their social context. The first step in approaching this process from a data-driven

[1] Or alternatively, we have increasing time and energy to devote to minutiae!

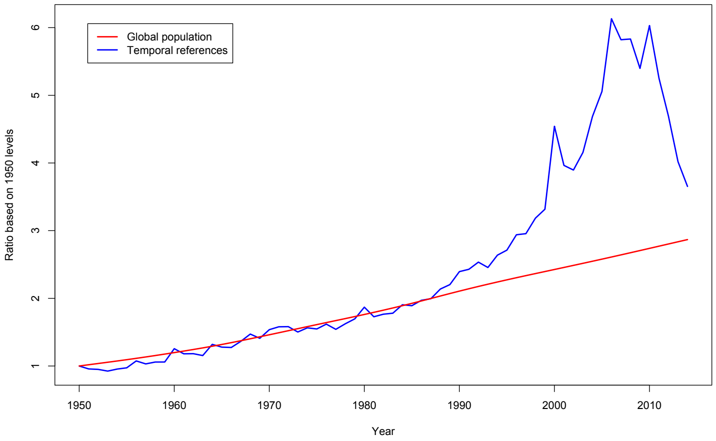

Fig. 3: Based on 1950 levels the number of temporal references for a year grows with global population until the late 1980's where it begins growing much faster.

perspective is to investigate how places and times are expressed together. That is, what are the configurations that exist? In this section we present an initial exploratory analysis of the connectedness of places and times in the English Wikipedia.

4.1 Historical trends for places

Some places have long recorded histories whereas others are more circumscribed due to a combination of factors, including not only the eurocentrism of Wikipedia but also the variations in quality of written historical records from around the world that have made it into the modern era. We can use the data we have collected to understand these differences in the historical record of places.

Looking at the changing number of temporal references for a place over time can show trends in how the history of that place has been recorded. We looked at these trends at the granularity of centuries, by aggregating all references to dates at finer granularities (year, day, etc.) into century bins. Then we looked at the average number of references for the countries per century and compared individual countries to that average. Figure 4 shows the results for four countries (Iraq, Greece, France, and China) from 3000 BCE to present day. This chart shows that the region of Iraq is of outsized importance in the 3rd and 2nd millennium BCE as it was the home of many of the earliest civilizations in the fertile crescent. China has a long recorded history, and in Greece there is a clear spike during the 4th century BCE. France in contrast has relatively low numbers of temporal references until after 1000 CE.

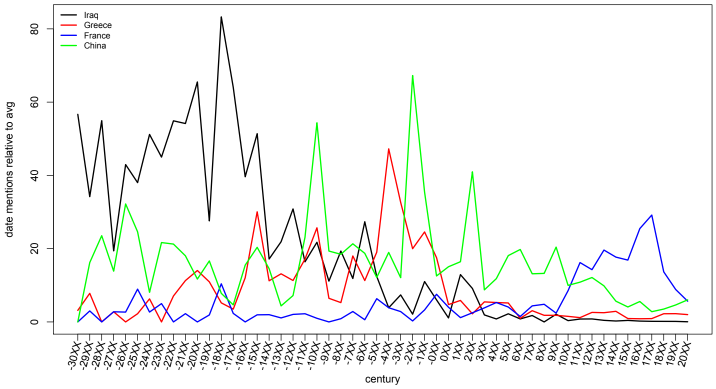

Fig. 4: Changes in the number of temporal references in proportion to the average shows historically important eras for countries.

Although plotting the timelines of individual places helps us understand the temporality of those places, similarity and clustering techniques for time series data can help to uncover larger trends across a set of places [13]. Figure 5 shows that among countries that have 500 or more century co-references, there is a stand-out group of nine countries that are distinctly different from the others: Egypt, Syria, Greece, Iraq, Iran, Italy, France, China, and India. The plot is a multidimensional scaling (MDS) of the century time series data based on Euclidean distance [6]. Note, that although these countries did not exist as such for much of this time, they are still used in reference to dates long before their founding. This demonstrates that present-day place names (such as Iraq) can operate as metonyms for historical places (e.g., Mesopotamia) in many cases. This has implications for spatio-temporal representation of place in a historical GIS, since we cannot assume that a place name should semantically be restricted to a founding (or ending) timestamp.

For different types of DATE references we can also construct histograms for each place, which indicate the distribution of dates for the place. We constructed two histograms of this type based on counts of individual century references from 3000 BCE to the 21st century. The first of these two histograms was built based on counts of pure century references, e.g. “the 14th century.” The second was based on counts of references to all century, year, and day binned by century. Therefore, a date like 1941 will be binned into the 20th century as will the day February 3, 1996. Based on these histograms we can calculate the entropy of temporal references for a place, which serves as a measure of how spread out the dates are over time vs. being focused on one or a few centuries. The Shannon entropy measure is shown in Equation 1, where $H(X)$ is the entropy value and

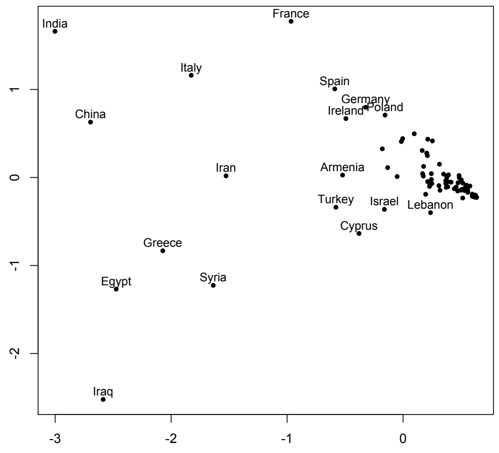

Fig. 5: MDS of countries based on Euclidean similarity of century time series.

$P(x)$ is the probability of date x in the histogram [24].

$$H(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_2 \mathrm{P}(x_i) \tag{1}$$

Figures 6 and 7 are choropleth maps showing the century reference entropy results for the countries of the world. There is a very strong spatial autocorrelation for the measure when it comes to specific century references (Fig. 6). The highest entropy values run in an east-west band from China through the Middle East to North Africa and southeastern Europe, indicating that references to many centuries at a coarse granularity are made in the context of these countries. This matches the spread of complex state societies out of the fertile crescent [20]. There is less historical record of the pre-Columbian states in the Americas, which is reflected here as well. When more fine-grained dates are included in the century counts (Fig. 7), Western Europe as well as Egypt and parts of the Middle East show the most spread of centuries represented. This would reflect more historical record across many centuries after around 1000 rather than before, when the recording temporal references became more precise.

In contrast to looking at how centuries are referenced, we can also examine the distribution of different individual years that are referenced in the context of a country. For this measure we look at all the years from 1000 to 2015 and make a similar choropleth map for the countries, shown in Fig. 8. In this case European countries have the most spread of years referenced and in strong contrast to the centuries mapped in Fig. 6 the Middle East is referenced in terms of a relatively small number of individual years.

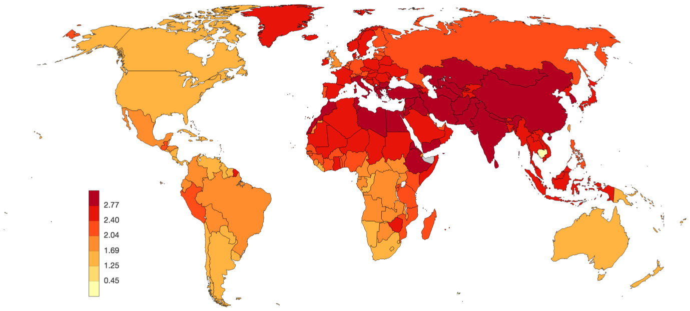

Fig. 6: Information entropy of dates per country by century reference only.

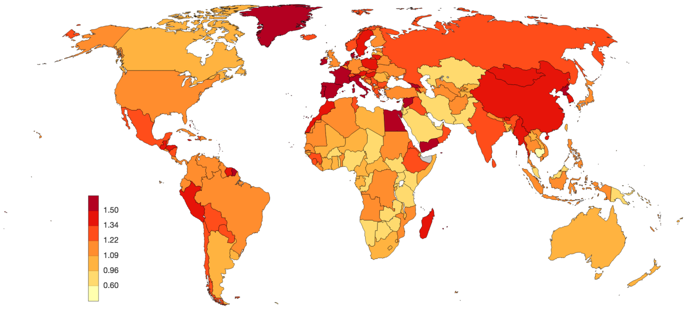

Fig. 7: Information entropy of dates per country by references to all dates aggregated by century.

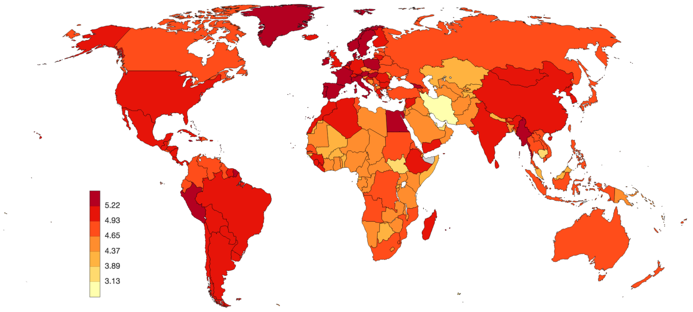

Fig. 8: Information entropy of dates per country by individual year from 1000 to 2015.

4.2 Times in terms of places, places in terms of times

Not surprisingly, countries on average have more associated temporal references than do populated places such as cities and towns. However, countries and other populated places are similar in that on average there are about equal numbers of century and decade references, on the order of ten times more day references, and about four times again more references to individual years (with no specific day reference). Table 3 breaks down how place types and date types are related in the texts. A remarkable 61.8% of all Wikipedia pages have a place and temporal reference that co-occur in a paragraph. Further, this means that out of all pages that reference a place (N=3,480,667), **82.7%** of those articles have a place and date reference co-occurring in a paragraph. This result points to the potential benefit of using place-time information as fundamental dimensions by which to organize information retrieval systems for large-scale text data, with the further implication that place-based GISs that intrinsically include a temporal dimension will open up significant opportunities for analysis that a spatial (only) GIS cannot[11, 2].

Place type + Temporal ref.	**Count**	**Avg. per type**	**Articles**	**Pct. articles**
Country + DATE	9,343,550	36,641.37	1,413,690	30.3
Country + Century	177,377	695.60	72,259	1.6
Country + Decade	204,481	801.89	94,109	2.0
Country + Year	4,951,018	19,415.76	1,073,744	23.0
Country + Day	1,418,277	5,561.87	588,293	12.6
Pop. place + DATE	22,687,527	83.00	2,029,940	43.6
Pop. place + Century	508,843	1.86	153,415	3.3
Pop. place + Decade	475,722	1.74	179,734	3.9
Pop. place + Year	12,301,207	45.01	1,616,998	34.7
Pop. place + Day	2,950,422	10.79	858,563	18.4
Other + DATE	36,626,672	104.21	1,865,095	40.0
Other + Century	571,218	1.63	154,708	3.3
Other + Decade	530,834	1.51	183,913	3.9
Other + Year	13,081,170	37.22	1,517,277	32.6
Other + Day	2,972,499	8.46	746,753	16.0
All place types + DATE	68,657,749	–	2,880,090	61.8

Table 3: Summary statistics on the co-occurrence of named place and date references in paragraphs of the English Wikipedia. The *Avg. per type* column shows the ratio of count to the number of instances of the place type (i.e., country, populated place, or other).

4.3 Wars and conflict: myths of creation and eschatology

In his essay on chronotopic analysis, Bakhtin wrote, “For a long time the central and almost sole theme of purely historical narrative was the theme of war” [4]. We examined the top-3 referenced single day pre-2000 dates for each of the 255 countries and found that 65% of the dates are related to a battle, declaration of war, or peace treaty. It is similar for large cities. This shows that, in the English Wikipedia at least, the theme of war still dominates how we talk about places. The other major category of event is the creation of a new geopolitical entity (often after a period of war). Table 4 shows a sample of the most cited days.

Table 4 also demonstrates that the recording of historical events in the English Wikipedia, no matter where the events have occurred, is heavily skewed to a United States, United Kingdom and commonwealth perspective. For example,

Country	Count	Date	Historical event
Argentina	32	1816-07-09	Argentine declaration of independence
Argentina	23	1982-04-02	Falklands War begins
China	101	1949-10-01	Mao speech creating People's Rep. of China
China	56	1997-07-01	Transfer of sovereignty of Hong Kong
Egypt	29	1915-04-25	Landing at Anzac cove (Gallipoli)
Egypt	23	1973-10-06	Yom Kippur War
France	73	1918-11-11	Armistice of 11 November 1918
France	52	1944-06-06	D-Day Normandy landings
Germany	109	1990-10-03	Reunification of Germany
Germany	70	1939-09-01	Invasion of Poland
Greece	34	1940-10-28	Ohi Day (Greco-Italian War)
Greece	31	1941-04-06	Germany invades Greece
India	156	1947-08-15	Independence day (India)
India	81	1950-01-26	Republic day (India)
Indonesia	19	1941-12-07	Dutch East Indies Campaign
Indonesia	16	1949-12-27	Proclamation of Indonesian Independence
Iran	30	1979-11-04	Iran hostage crisis
Iran	22	1988-07-03	Shooting of Iran Air Flight 655
Japan	145	1941-12-07	Pearl Harbor bombing
Japan	88	1945-08-15	Surrender of Japan (V-J Day)
Mexico	27	1848-02-02	Treaty of Guadalupe Hidalgo
Mexico	24	1994-01-01	NAFTA operational, Zapatista uprising
Russia	30	1991-12-25	Dissolution of the Soviet Union
Russia	26	1998-02-02	Russian financial crisis
South Africa	134	1910-05-31	South African independence
South Africa	102	1994-04-27	First democratic elections (Freedom day)
United Kingdom	45	1939-09-03	Britain declares war on Germany
United Kingdom	43	1910-05-31	South African independence
United States	461	2001-09-11	September 11 terrorist attacks
United States	131	1941-12-07	Pearl Harbor bombing
Paris	24	1792-08-10	Insurrection of 10 August 1792
Paris	19	1860-01-01	Annexation of 1860
Rome	19	1944-06-04	Liberation of Rome
Rome	17	1870-09-20	Capture of Rome (Risorgimento)

Table 4: Top-2 referenced days from 2001 and earlier for selected places.

the most highly referenced day for Egypt (29 references) is the date of the ANZAC landing during the Gallipoli campaign, which happened in Turkey, though the troops disembarked for the campaign from a station in Egypt. In contrast the beginning of the Yom Kippur War, a date presumably of more interest to the population living in Egypt, is referenced 23 times. This is further evidence of the eurocentric bias in Wikipedia content, which has been well-documented across all language editions [9].

5 Implications of chronotopic analysis for GIScience research

The chronotopic analysis we performed in this study point to many interesting relationships between place and time in very large unstructured data collections. Going forth with this kind of research there are a number of representational and algorithmic challenges to building general systems for chronotopic data analysis.

Better discovery of spatial and temporal references in text. Currently, methods to discover spatial and temporal entities in leave a lot of room for improvement. For example, place name disambiguation still relies on rough heuristics that could potentially be improved with machine learning classifiers.

Scaling of discovery methods. The document scraping or feature extraction stage of such work can require massive amounts of time and consume large amounts of storage. In the work we describe above, the temporal tagging and creation of the database of temporal references required approximately 50,000 core hours of processing in a single pass (equivalent to approximately 5 years on a single core computer). Fortunately, the tasks are embarrassingly parallel (the task can easily be decomposed into n smaller but separate tasks), so in our case we could make use of a local HPC service, utilizing 3000 compute cores and a GPFS parallel file system, bringing the elapsed time down to a couple of days. In our experience this stage often needs to be repeated many times to train and refine the extraction methods used, so such savings are critical.

Data structures and algorithms. The figures quoted in Table 3 suggest that both spatial and temporal dimensions are useful ways to organize this corpus. In fact a strong case could be made for a combined spatio-temporal index, given that this would cover over 60% of the documents. Within GIScience there has been some useful work on adding in the temporal dimension [15, 21], but less on the data structures and related algorithms that could scale to many millions of objects that each have complex, multi-valued relationships to both place and time.

Formalizing more complex spatial and temporal references. How do we describe the 'spatiality' or 'temporality' of a document more formally, again given that there may be multiple spatial and temporal references in a document, each taking different forms? How do these map onto human understandings of space and time? What kinds of query operators and interfaces are needed? How do we extend the current formal models of topology and spatial relations to address these more complex, multi-space, multi-time objects?

6 Conclusion

In this paper we introduced the notion of chronotopic data analysis as a methodology to study spatio-temporal structure in a large text corpora. As a preliminary example of this kind of analysis we examined the set of all place and date co-references in the English Wikipedia and found that millions of place references have a temporal association. We demonstrated that by exploring places and dates together we can uncover a number of unexpected patterns that shed light on the importance of the temporal dimension in understanding place.

We have just scratched the surface of chronotopic analysis of big data. Our investigation into place and time in Wikipedia was done by looking at statistics for the entire corpus. Chronotopic analysis in literature also looks at how the spatio-temporal configuration relates to other aspects of the narrative. Toward that end, there is much that can be done to extend the methodology, for example looking at how different types of articles within Wikipedia reference place-time differently. In addition, this type of exploratory data analysis can discover regularities or unique characteristics in the spatio-temporal patterns that manifest in different kinds of historical textual collections, such as novels, newspaper collections, and the literature of private life, e.g., diaries and letters.

References

1. Adams, B., Janowicz, K.: Thematic signatures for cleansing and enriching place-related linked data. International Journal of Geographical Information Science 29(4), 556–579 (2015)
2. Adams, B., McKenzie, G., Gahegan, M.: Frankenplace: Interactive thematic mapping for ad hoc exploratory search. In: Proceedings of the 24th International Conference on World Wide Web. pp. 12–22. International World Wide Web Conferences Steering Committee (2015)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: The Semantic Web, Lecture Notes in Computer Science, vol. 4825, pp. 722–735. Springer Berlin Heidelberg (2007)
4. Bakhtin, M.: Forms of time and of the chronotope in the novel. In: The Dialogic Imagination, pp. 84–258. University of Texas Press, Austin (1981)
5. Bemong, N., Borghart, P., De Dobbeleer, M., Demoen, K., De Temmerman, K., Keunen, B.: Bakhtin's theory of the literary chronotope: Reflections, applications, perspectives. Academia Press (2010)
6. Borg, I., Groenen, P.J.: Modern multidimensional scaling: Theory and applications. Springer Science & Business Media (2005)
7. Galton, A.: Fields and objects in space, time, and space-time. Spatial cognition and computation 4(1), 39–68 (2004)
8. Goodchild, M.F.: Formalizing place in geographic information systems. In: Communities, Neighborhoods, and Health, pp. 21–33. Springer (2011)
9. Graham, M., Hogan, B., Straumann, R.K., Medhat, A.: Uneven geographies of user-generated information: patterns of increasing informational poverty. Annals of the Association of American Geographers 104(4), 746–764 (2014)
10. Hill, L.L.: Core elements of digital gazetteers: placenames, categories, and footprints. In: Research and advanced technology for digital libraries, pp. 280–290. Springer (2000)
11. Janowicz, K.: The role of space and time for knowledge organization on the semantic web. Semantic Web 1(1, 2), 25–32 (2010)
12. Jordan, T., Raubal, M., Gartrell, B., Egenhofer, M.: An affordance-based model of place in GIS. In: 8th Int. Symposium on Spatial Data Handling, SDH. vol. 98, pp. 98–109 (1998)
13. Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and knowledge discovery 7(4), 349–371 (2003)
14. Knowles, A.K., Hillier, A.: Placing history: how maps, spatial data, and GIS are changing historical scholarship. ESRI, Inc. (2008)
15. Langran, G.: Issues of implementing a spatiotemporal system. International Journal of Geographical Information Science 7(4), 305–314 (1993)
16. Mei, Q., Liu, C., Su, H., Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proceedings of the 15th international conference on World Wide Web. pp. 533–542. ACM (2006)
17. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176–182 (2011), http://science.sciencemag.org/content/331/6014/176
18. Miller, H.J.: Modelling accessibility using space-time prism concepts within geographical information systems. International Journal of Geographical Information System 5(3), 287–301 (1991)

19. Moretti, F.: Graphs, maps, trees: abstract models for a literary history. Verso (2005)
20. Peebles, C.S., Kus, S.M.: Some archaeological correlates of ranked societies. American Antiquity pp. 421–448 (1977)
21. Peuquet, D.J.: Making space for time: Issues in space-time data representation. GeoInformatica 5(1), 11–32 (2001)
22. Pustejovsky, J., Castano, J.M., Ingria, R., Sauri, R., Gaizauskas, R.J., Setzer, A., Katz, G., Radev, D.R.: Timeml: Robust specification of event and temporal expressions in text. New directions in question answering 3, 28–34 (2003)
23. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. pp. 851–860. ACM (2010)
24. Shannon, C.E.: The mathematical theory of communication. Bell System Technical Journal 27, 379–423, 623–656 (1948)
25. Strötgen, J., Gertz, M.: Multilingual and cross-domain temporal tagging. Language Resources and Evaluation 47(2), 269–298 (2013)
26. Sui, D., Goodchild, M.: The convergence of gis and social media: challenges for giscience. International Journal of Geographical Information Science 25(11), 1737–1748 (2011)
27. Tomaszewski, B., MacEachren, A.M.: Geo-historical context support for information foraging and sensemaking: Conceptual model, implementation, and assessment. In: Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on. pp. 139–146. IEEE (2010)
28. Winter, S., Kuhn, W., Krüger, A.: Guest editorial: Does place have a place in geographic information science? Spatial Cognition & Computation 9(3), 171–173 (2009)
29. Ye, M., Janowicz, K., Mülligann, C., Lee, W.C.: What you are is when you are: The temporal dimension of feature types in location-based social networks. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 102–111. GIS '11, ACM, New York, NY, USA (2011)
30. Yu, H.: Spatio-temporal GIS design for exploring interactions of human activities. Cartography and Geographic Information Science 33(1), 3–19 (2006)
31. Yuan, M.: Temporal GIS and applications. In: Shekhar, S., Xiong, H. (eds.) Encyclopedia of GIS, pp. 1147–1150. Springer (2008)