Detecting Events with Date and Place Information in Unstructured Text

David A. Smith
Perseus Project, Tufts University
Medford, MA 02155
dasmith@perseus.tufts.edu

ABSTRACT

Digital libraries of historical documents provide a wealth of information about past events, often in unstructured form. Once dates and place names are identified and disambiguated, using methods that can differ by genre, we examine collocations to detect events. Collocations can be ranked by several measures, which vary in effectiveness according to type of events, but the log-likelihood measure ($-2 \log \lambda$) offers a reasonable balance between frequently and infrequently mentioned events and between larger and smaller spatial and temporal ranges. Significant date-place collocations can be displayed on timelines and maps as an interface to digital libraries. More detailed displays can highlight key names and phrases associated with a given event.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces*

General Terms

Design

Keywords

event detection, geographic visualization, phrase browsing

1. INTRODUCTION

Digital libraries of historical documents provide a wealth of information about past events in an unstructured form. Natural questions about particular periods and places are "What happened then?" and "What happened here?", but they may not be best answered by ad hoc queries typed into search forms. Simply by restricting our queries to certain collections catalogued by time or place, we can exclude many irrelevant events, but questions of relevance, in a broad sense, remain. What events will different users find relevant when browsing four thousand years of history, or the nineteenth century, or 1862? What events are significant, in some sense, at global, national, and local scales? Of particular interest to digital libraries, dates, places, and events can provide general interfaces for access to diverse collections. Automatically detected events can also augment manually produced metadata, particularly for long documents that cover many topics.

The Perseus Digital Library Project (`http://www.perseus.tufts.edu`) has focused on developing automatic methods for structuring large document collections, especially in the humanities. Generalizing tools we first built for ancient Greek literature, art, and archaeology, we have built testbeds on English Renaissance literature, ancient and early modern science, the history and topography of London, and United States history in the nineteenth century. We have previously worked on named-entity, term, and date identification [3] and on place name disambiguation [9]. Especially in the United States, where there are a Springfield and several Middletowns in every state, place names have to be disambiguated before they can be plotted on maps.

Building on this work with individual terms, names, and dates, we have exploited co-occurrences of dates and place names in our testbeds to detect and describe likely events in a digital library. We use statistical measures to determine the relative significance of various events. We have also built interfaces that help users preview likely regions of interest for a given range of space and time and that identify key phrases associated with each possible event.

2. PRIOR WORK ON NEWS TEXTS

Although our testbeds are primarily in the humanities, it is useful to compare applications for historical digital libraries with the Topic Detection and Tracking (TDT) study. As with similar competitive evaluations, such as TREC for information retrieval, TDT seeks to advance the state of the art by concentrating research around a quantitatively evaluated task. TDT aims at developing techniques for "discovering and threading together topically related material from streams of data such as newswire and broadcast news" [12]. Topics are defined as specific events, "something (non-trivial) happening in a certain place at a certain time" [13] although some researchers use *event* to mean a single happening within a larger *topic* story [6]. Due to its focus on news data, TDT possesses "an explicitly time-tagged corpus". Although not part of the TDT task, systems such as

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.

[8] for visualizing news broadcasts on maps also take advantage of a time-tagged data stream.

TDT systems, by design, will aggregate stories over a span of several days, even with some gaps, into single event topics. Despite the definition of an event, however, as occurring in a certain place, most TDT systems do not directly take geographical location into account. Geographical names, rather, are treated just like other named entities, such as personal and company names, or even as single words. Although some TDT systems perform retrospective event detection across an entire corpus, many are designed to handle the more difficult task of classifying stories into topics in the order in which they come in. Applications to historical documents should be able to take advantage of less error-prone retrospective methods.

The most significant problem in adapting TDT methods to historical texts is the difficulty of handling long-running topics. For the mid-1990s events in the second TDT study, systems had trouble treating the O. J. Simpson case or the investigation of the Oklahoma city bombing as a single event [11, 13]. Many historical documents discuss long-running events, and many users will wish to browse digital libraries at a scale larger than events of a few days' length.

3. THE HISTORICAL DOMAIN

Since a precise dateline heads each story, modern news texts are of course explicitly time-tagged. Indexing schemes can associate every term — be it a word, phrase, or named entity — with that date. Most historical texts do not fit this model for three reasons: *discursiveness*, *digression*, and *scale*. First, historical texts tend to be discursive, not broken into discrete date units. While some genres, such as chronicles and diaries, do fit this format, they do not make up a very sizable portion of most digital library collections. Domain-specific formatting cues, such as the title and dateline in news stories, can be used to segment such texts, but we need to automatically discover which documents should be so segmented in order for the solution to be scalable. Most documents, however, although not neatly segmentable, still contain a large amount of date information, but the association of each date in a text and the terms around is not one of simple "aboutness". Second, historical documents tend to be more digressive than news stories. Even if there is a main linear narrative, a historian will often digress about events from before or after the main period, or taking place in another region. These digressions, of course, may themselves provide information about other events. Henry Wheatley, in his 1891 survey of London streets, mentions that "Quebec Street commemorates the capture of Quebec by General Wolfe in 1759." Finally, many historical documents are simply on a larger scale than news stories. Not only are books, and even chapters, orders of magnitude longer than newspaper pieces, but the ranges of time and space covered are often much larger.

In addition to problems of interpretation, historical documents present obstacles merely to identifying relevant dates. First of all, many scholarly works are strewn with bibliographic citations. Bibliographic dates can be useful in their own right; it would be interesting to see, for example, that a work published in the 1990s cited works mostly from the 1960s. Bibliography is not, however, directly related to historical narrative and distracts from most information needs. News stories seldom make citations and current academic practice relegates much bibliography to a separate section, but older works often mix citations with narrative. In general, accurately identifying bibliographic references has been an active area of research with varying success [1]; nevertheless, as McKay and Cunningham point out [7], identifying bibliographic dates is easier than identifying (and linking) entire citations.

Further problems arise when older documents use dating schemes other than the modern, Western Gregorian calendar. Simultaneous events may have different dates on different calendars, as when the Russian revolution in Orthodox, Julian October took place in Western, Gregorian November. Even more involved are the problems with ancient systems that dated by the years in which various magistrates — such as Athenian archons or Roman consuls — served. At present, Perseus often avoids these problems by acquiring texts already annotated, in footnotes or headings, with modern date equivalents. Also, older texts with more involved and uncertain dating systems tend, unfortunately for historians, to contain many fewer dates.

4. RANKING COLLOCATIONS

Once dates and other features have been identified and, if necessary, disambiguated, they can be used to detect events in documents. Our initial experiments have focused on associations of dates and places. To cite one precedent, Swan and Allan report better event detection when associating named entities, rather than simple phrases, with dates[10]. Unlike other projects, we have privileged place names over other named entities since we can identify multiple names referring to a single place and detect the use of the same name for different places.

Since we cannot depend on our source documents to have marked or easily detectable story divisions, we must define some sort of window of association. Given the discursive and digressive properties of our documents mentioned above, we have chosen sentences and paragraphs. We count, for example, the number of sentences that contain each date or place and the number of times each date and place occur in the same sentence. For each date-place pair, we can thus build a contingency table where a is the number of times date D and place P occur in the same sentence, b the number of times D occurs without P, c the number of times P occurs without D, and d the number of sentences in which neither D nor P occur. These counts can be used to calculate several different measures of association between the date and place. Widely used measures are mutual information (MI) [2], chi-squared (χ^2), and phi-squared (ϕ^2), which is χ^2 normalized on the number of association windows. Dunning argued that the assumption that text tokens are normally distributed overestimated the significance of rare statistical events and proposed the log-likelihood test ($-2\log\lambda$) based on the binomial or multinomial distributions [4].

We have experimented with these statistics to test their effectiveness at detecting events. Without a definitive list of events in our testbeds, we have concentrated on relative ordering of events by significance rather than absolute relevance or irrelevance. As described below, users can select the amount of event information they want to see, and we hope this will effectively take them from short, highly precise lists, to total recall of all events in the corpus. As an example, we compare the twenty top-ranked events by each test for all world events of the nineteenth century (tables 1–4).

Place	Date	Count	$-2 \log \lambda$
Corinth, Mississippi	1862	320	2745.31
Gettysburg, Pennsylvania	July 3 1863	164	2076.08
Mobile Bay, Alabama	August 5 1864	110	1870.14
Mobile Bay, Alabama	August 6 1864	80	1375.46
California, United States	1849	227	1219.85
Malvern Hill, Virginia	July 1 1862	76	1113.22
Knoxville, Tennessee	1862	170	1078.49
Waterloo, Belgium	1815	82	995.161
Spotsylvania, Virginia	May 12 1864	66	994.899
Virginia, United States	1860	264	963.186
Pittsburg Landing, Tennessee	1862	124	881.619
Walcheren, Netherlands	1809	53	860.891
Gettysburg, Pennsylvania	1863	154	749.540
Chancellorsville, Virginia	May 3 1863	49	618.326
Crimea, Ukraine	1854	65	608.433
Atlanta, Georgia	1864	138	568.375
Huntsville, Alabama	1862	88	561.238
Great Britain, United Kingdom	1812	86	536.693
California, United States	1850	131	521.704
United States	1861	245	503.163

Table 1: 19th c. events: Ranked by log-likelihood

Place	Date	Count	χ^2
Wakulla county, Florida	January 7 1859	9	2193820
Mobile Bay, Alabama	August 5 1864	110	935482
Mobile Bay, Alabama	August 6 1864	80	736456
Queretaro, Mexico	May 1848	10	576247
Dooly, Georgia	December 17 1860	7	498001
Crisfield, Maryland	September 1874	5	491228
Broad Creek, Massachusetts	September 1874	5	439518
Walcheren, Netherlands	1809	53	290660
Spotsylvania, Virginia	May 12 1864	66	262641
Waynesboro, Georgia	December 4 1864	16	255647
Jeffersonville, Ohio	March 13 1862	5	255635
Mayo, Cape Verde	March 12 1835	5	246335
Malvern Hill, Virginia	July 1 1862	76	232525
Puerto Cabello, Venezuela	July 26 1861	6	191783
Gettysburg, Pennsylvania	July 3 1863	164	152491
Mobile Bay, Alabama	August 8 1864	20	141363
Pocomoke, North Carolina	September 1874	7	139885
Five Forks, Maryland	April 1 1865	5	138559
Appomattox county, Virginia	January 31 1863	6	137580
Greenwich, Connecticut	May 30 1848	7	125128

Table 2: Ranked by chi-squared

Place	Date	Count	MI
Wakulla county, Florida	January 7 1859	9	17.8951
Crisfield, Maryland	September 1874	5	16.5841
Broad Creek, Massachusetts	September 1874	5	16.4237
Dooly, Georgia	December 17 1860	7	16.1185
Queretaro, Mexico	May 1848	10	15.8144
Jeffersonville, Ohio	March 13 1862	5	15.6418
Mayo, Cape Verde	March 12 1835	5	15.5884
Puerto Cabello, Venezuela	July 26 1861	6	14.9642
Five Forks, Maryland	April 1 1865	5	14.7583
Appomattox county, Virginia	January 31 1863	6	14.4851
Greenbrier county, West Virginia	March 1858	5	14.3862
Abingdon, United Kingdom	March 22 1860	6	14.3106
Pocomoke, North Carolina	September 1874	7	14.2867
Greenwich, Connecticut	May 30 1848	7	14.1258
Ashley River, South Carolina	December 7 1864	5	14.0987
Waynesboro, Georgia	December 4 1864	16	13.9639
Pocotaligo, South Carolina	December 20 1864	7	13.7488
Washington, Georgia	May 4 1865	8	13.7094
Drummond Island, Michigan	March 1816	7	13.6673
Nantucket, Massachusetts	August 1841	5	13.6232

Table 3: Ranked by mutual information

Place	Date	Count
Corinth, Mississippi	1862	320
Virginia, United States	1860	264
United States	1861	245
California, United States	1849	227
Richmond, Virginia	1862	171
Knoxville, Tennessee	1862	170
Gettysburg, Pennsylvania	July 3 1863	164
Gettysburg, Pennsylvania	1863	154
United States	1812	152
United States	1860	146
Atlanta, Georgia	1864	138
Georgia, United States	1864	136
United States	1862	134
California, United States	1850	131
Virginia, United States	1861	131
Virginia, United States	1862	128
United States	1864	128
Pittsburg Landing, Tennessee	1862	124
Washington, United States	1862	124
United States	1848	122

Table 4: Ranked by raw association count

The ϕ^2 measure would produce the same ranking as χ^2 and is not listed. We have also included place-date pairs ranked by raw association counts. Using a common rule of thumb in contingency table analysis, we exclude date-place pairs with fewer than five occurrences. Perseus collections for this period focus on British an U.S. history: the Bolles collection on the history and topography of London; three collections on California, the Upper Midwest, and the Chesapeake region from the Library of Congress' American Memory project; and a collection of memoirs and official records of the U.S. Civil War.

The log-likelihood measure achieves a balance between events at a very specific place and time — such as the battles of Gettysburg (specifically the third day, July 3, 1863), Mobile Bay, Malvern Hill, Spotsylvania, and Waterloo — and larger regions of concentration — such as the California Gold Rush of 1849 and 1850 or the Crimean War. Civil War battles are well represented, probably because several different memoirs, diaries, and official histories will discuss the same event, while events in other corpora are less likely to receive repeat coverage. The chi-squared and mutual information scores highlight associations of rarer dates and places; for example, January 7, 1859 in Wakulla county, Florida, is singled out as the day that the offices of Tax Assessor and Collector and Sheriff were combined. Since this particular day and place are not mentioned except when together, the chi-squared and mutual information scores overestimate the significance of these nine occurrences. Similarly, Crisfield, Maryland, in September, 1874, is singled out with only five collocations due to a murder that occurred there. Although these are undoubtedly events, they are not very useful for a user wishing to get a sense of the contents of the digital library. Interestingly, all of the χ^2 scores in these top twenty in table 2 are far above the significance threshold of 10.83 for 99.9% confidence; while the statistic may be useful for determining absolute significance, it may not be as useful for establishing rank among significant collocations.

On the whole, mutual information shows a greater bias for rare events: in the top twenty ranked by MI, no event is represented by more than 16 passages. Log-likelihood and χ^2 exhibit a greater range in the number of passages supporting each event. Although ranking by raw counts privileges whole years and larger regions such as states and countries, such a result may also be appropriate at scales of the whole world and a century.

Finally, note that the raw count list contains only one event with a month and day — the heavily covered battle of Gettysburg. All events in the mutual information list contain at least a month, and χ^2 only shows one event without a month or day: the half-hearted Walcheren expedition of 1809 that is mentioned in many British officers' biographies. The log-likelihood measure, again, shows a balance of specific and more general dates.

Even outside the scope of precise dates, log-likelihood ranking can perform well. Beyond the nineteenth century, fewer dates are recorded precisely to the day. Tables 5 and 6 show events in the sixth and fifth centuries BC, and the thirteenth and fourteenth centuries AD. The digital library contains substantial material on the ancient period. As noted above, however, there are fewer dates to exploit in older documents, and the lower counts bear this out. The low numbers show up in a bogus disambiguation of "Lade" for the United Kingdom instead of Greece. Still, decisive moments in Greek history are clear with the end of the Peloponnesian

Place	Date	Count	$-2\log\lambda$
Aegospotami, Turkey	405 BC	24	467.124
Plataea	479 BC	17	241.044
Salamis, Greece	480 BC	20	211.093
Delium, Greece	424 BC	11	203.543
Lade, United Kingdom	494 BC	9	174.566
Athens, Greece	431 BC	18	160.52
Samos, Greece	440 BC	14	151.662
Olynthus	432 BC	9	146.786
Tanagra, Greece	457 BC	8	136.139
Sybaris	510 BC	9	129.891
Greece	480 BC	20	128.819
Athens, Greece	480 BC	22	125.905
Mantinea, Greece	418 BC	7	116.546
Athens, Greece	404 BC	14	114.052
Syracuse, Italy	485 BC	8	106.041
Amphipolis, Greece	422 BC	6	101.548
Sparta, Greece	404 BC	10	99.4967
Sardes, Turkey	481 BC	6	96.6489
Thurii	443 BC	5	96.5052
Sicily, Italy	415 BC	9	91.6774

Table 5: Events in the 6th and 5th centuries BC, ranked by log-likelihood

Place	Date	Count	$-2\log\lambda$
Poitiers, France	1356	19	357.045
Lewes, United Kingdom	1264	19	314.943
Crecy, France	1346	16	309.233
Bannockburn, United Kingdom	1314	15	305.789
Neville's Cross, United Kingdom	1346	11	235.198
Gascony, France	1264	14	233.708
Lewes, United Kingdom	1265	13	222.948
Sluys, Netherlands	1340	11	217.536
Lewes, United Kingdom	1263	12	208.978
Montfort, France	1264	11	201.241
Flanders, Belgium	1297	14	193.794
Gascony, France	1265	11	193.198
Gascony, France	1297	11	190.275
Epsom, United Kingdom	1265	11	183.179
Lewes, United Kingdom	1258	11	182.392
Halidon Hill, United Kingdom	1333	8	177.775
Montfort, France	1263	9	176.772
Gascony, France	1253	10	176.184
Montfort, France	1265	9	172.843
Bannockburn, United Kingdom	1313	9	172.033

Table 6: Events in the 13th and 14th centuries

war at the battle of Aegispotami and the climax of the Persian wars at Plataea. The Perseus Digital Library does not contain any resources specifically for medieval history, but enough allusions are made in the Bolles London collection to detect some significant events in medieval England. The battles of Poitiers, Lewes, Crecy, and Bannockburn, at the top of the list, are decisive events in the Hundred Years War, the unrest in the reign of Henry III, and the Scottish struggle with the English. When working with small numbers of passages, however, the different ranking strategies appear to make less difference (table 7).

5. BROWSING EVENTS

5.1 Geo-Temporal Overview

Place	Date	Count	χ^2
Neville's Cross, United Kingdom	1346	11	821941
Halidon Hill, United Kingdom	1333	8	821624
Bannockburn, United Kingdom	1314	15	786028
Boroughbridge, United Kingdom	1322	8	626645
Bretigny, France	1360	6	593667
Crecy, France	1346	16	530521
Poitiers, France	1356	19	483353
Sluys, Netherlands	1340	11	449818
Codnor, United Kingdom	1241	5	430686
Montfort, France	1263	9	363850
Montfort, France	1265	9	296822
Bannockburn, United Kingdom	1313	9	287064
Bannockburn, United Kingdom	1306	9	275102
Poitou, France	1214	7	267580
Crecy, France	1342	9	264700
Neville's Cross, United Kingdom	1341	5	262741
Neville's Cross, United Kingdom	1338	5	236020
Sluys, Netherlands	1344	6	228297
Montfort, France	1264	11	227686
Crecy, France	1356	9	215066

Table 7: Events in the 13th and 14th centuries ranked by chi-square

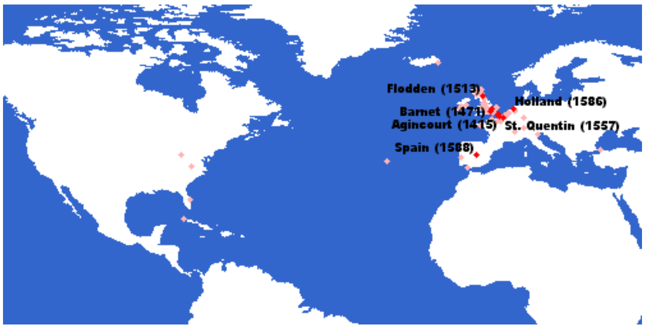

Figure 1: Map of top events from 1400–1600: for this period, the DL primarily deals with British history. Sites in Europe are English expeditions.

We have developed an interface to explore these associations with a combination of graphical and tabular display. This display is useful not only for browsing the results of our event detection but also as a generalized interface to many heterogeneous digital libraries. In addition to lists or timelines of significant events, we also generate global or regional maps. When the user selects a particular range of time — whether a century, decade, or year — the map is updated to show the sites of significant events in that range. Users can also zoom in on particular regions to see events in a specific area. The locations of top-scoring events in any given space-time range are brighter in color and labeled on the map; lower-scoring events are fainter in color. The top-ranked events are also listed below the map, with date, place, and the number of times they co-occur in the digital library.

Figures 1, 2, and 3 show three snapshots of the North America and Europe, the primary focus of the Perseus Digital Library collections. Users browsing at these two hundred year intervals can clearly see the shift in coverage from Europe — primarily Britain — in the early modern period to North America in the nineteenth century. Events on the continent of Europe tend to relate to English and British wars: Holland (1586), Blenheim (1704), Fontenoy (1745), and Waterloo (1815). As we observed with the tabular data above, battles stand out particularly well, since they are memorable, and heavily documented, events that occur at a specific place and time. An error in figure 2 is instructive: the town of Monmouth, in Wales, is associated with 1685. This collocation highlights the rebellion of the Duke of Monmouth, Charles II's illegitimate son, against James II. Many of the references in the DL to the Duke could be construed as ambiguous: e.g., "commanded regiment of horse against Monmouth" or "summoned to join royalist forces against Monmouth". The collocation, nevertheless, points to an important event in Britain in 1685.

5.2 Phrase Browsing

If users wish to explore the detected events more closely, they can click on the date-place collocation and call up a display of the individual text passages from the digital library. Since the Perseus system disambiguates toponyms in texts, these searches are for the unique toponym identifiers, not for the names themselves as strings.

The default display organizes these passages by phrases common to two or more sentences. This clustering feature is

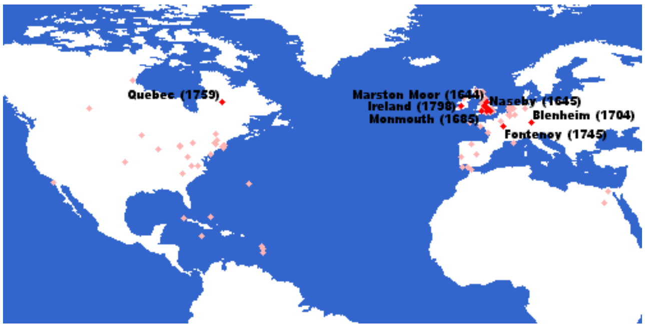

Figure 2: Map of top events from 1600–1800: the DL continues its focus on Britain. Some North American information, particularly the capture of Quebec, is present. The strong association of 'Monmouth' with 1685 refers to the Duke of Monmouth's uprising.

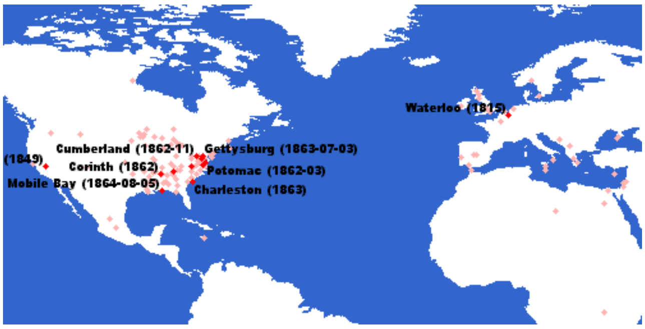

Figure 3: Map of top events from 1800–1900: collections on pioneering in the Upper Midwest and California (note the 1849 at the extreme west) combine with a Civil War collection to give a North American focus. The battle of Waterloo holds out for British history.

Phrase	Count
fire of london	21
great fire	21
city of london	8
charles ii	6
act of parliament	4
duke of york	4
christ church oxford	3
house of commons	3
dreadful fire	3
rebuilding of the city	2
college oxford	3
privy council	3
view of london	2
burning of london	2
church of st	2

Table 8: Clusters for London, 1666

Phrase	Count
san francisco	19
discovery of gold in california	8
discovery of gold	10
gold rush	9
united states	9
gold fields	7
trip to california	5
gold fever	6
cape horn	6
california gold	6
california during the years	3
early in the year	3

Table 9: Clusters for California, 1849

available for all searches, not just these date-place searches, in the Perseus Digital Library. We produce the clusters at run time using a suffix-tree algorithm similar to [14]. The phrases are ranked by a score s that combines the number of words in the phrase w with the number of passages in the cluster p, using a cluster-constant c, usually set to 0.5 (equation 1). Clustering is polythetic: each search result may belong to one or more clusters. The clustering and ranking are fast enough to be used interactively without any offline computation, as in [5].

$$s = p \cdot \frac{1 - e^{-cw}}{1 + e^{-cw}} \tag{1}$$

The examples show clusters for London, 1666, the date of the Great Fire (table 8); for California, 1849, the Gold Rush (table 9); and for Atlanta, 1864, when a Union army captured the city (table 10). Phrases containing dates are removed since they mostly show variations like "fire in 1666" and "fire in the year 1666". Note that the cluster head phrases need not contain the search terms.

These phrases can characterize events by listing associated people or places, such as the opposing generals Sherman and Johnston, San Francisco, or Cape Horn, around which many sailed to California. Phrase clusters may also be more descriptive: "rebuilding of the city", "gold fever", or "march to the sea". The user can also group passages by the book or collection from which they come. The number of distinct

Phrase	Count
military division of the mississippi	13
atlanta ga	19
atlanta georgia	18
atlanta campaign	14
march to the sea	5
major general	8
general sherman	7
sherman's army	5
effective strength of the army	3
advance on atlanta	4
battle of atlanta	4
capture of atlanta	4
general joseph e johnston	3
maj gen	4
kenesaw mountain	4

Table 10: Clusters for Atlanta, 1864

documents recording a date-place collocation could be useful in deciding an event's significance.

6. CONCLUSIONS

Although historical documents cannot often benefit from the tight topic focus and reliable structure of news or scholarly articles, their broad scope and lack of structure can provide a useful testbed for building more scalable architectures for event detection and information extraction systems. Once detected and ranked, events can provide a useful generic interface to digital library systems through maps, timelines, and tabular displays.

Evaluating these and other methods of event detection requires attention to varying information needs. Does the user wish to gain a broad overview of a particular corpus or subcorpus or to focus on events that stand out from the rest of the corpus? Since the distance between places or dates is measurable, and not arbitrary as in many topic browsing systems, we can group the data to minimize the aggregation effects of using individual days, years, or places as terms of association. We have concentrated on ranking events using statistical measures, finding evidence that the log-likelihood measure achieves a balance among spatial and temporal scope and frequency of occurrence. Future work can concentrate on finding genre-specific cues for events in diaries, letters, encyclopedias, and biographical dictionaries. We have also built a browsing interface so that users can see regions of concentration within the digital library and explore names and phrases associated with a given event.

7. ACKNOWLEDGMENTS

The work presented in this paper has been supported by a grant from the Digital Library Initiative Phase 2 (NSF IIS-9817484), with particular backing from the National Endowment for the Humanities and the National Science Foundation. I would also like to thank Greg Crane and Jeff Rydberg-Cox for comments on this paper.

8. REFERENCES

[1] Donna Bergmark and Carl Lagoze. An architecture for automatic reference linking. In *Proceedings of ECDL 2001*, pages 115–126, Darmstadt, 4-9 September 2001.

[2] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[3] Gregory Crane, David A. Smith, and Clifford E. Wulfman. Building a hypertextual digital library in the humanities: A case study on London. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 426–434, Roanoke, VA, 24-28 June 2001.

[4] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[5] Steve Jones and Gordon Paynter. Topic-based browsing within a digital library using keyphrases. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 114–121, Berkeley, CA, 11-14 August 1999.

[6] Vikash Khandelwal, Rahul Gupta, and James Allan. An evaluation corpus for temporal summarization. In James Allan, editor, *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*, San Francisco, 2001. Morgan Kaufmann.

[7] Dana McKay and Sally Jo Cunningham. Mining dates from historical documents. Technical report, Department of Computer Science, University of Waikato, 2000.

[8] Andreas M. Olligschlaeger and Alexander G. Hauptmann. Multimodal information systems and GIS: The Informedia digital video library. In *Proceedings of the ESRI User Conference*, San Diego, California, July 1999.

[9] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of ECDL*, pages 127–136, Darmstadt, 4-9 September 2001.

[10] Russell Swan and James Allan. Extracting significant time varying features from text. In *Proceedings of the Eighth International Conference on Information Knowledge Management (CIKM '99)*, pages 38–45, Kansas City, MO, November 1999.

[11] Russell Swan and James Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56, Athens, Greece, July 2000.

[12] Charles L. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *LREC 2000: 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, June 2000.

[13] Yiming Yang, Tom Pierce, and Jaime Carbonell. A study on retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 28–36, Melbourne, Australia, August 1998.

[14] Oren Zamir, Oren Etzioni, Omid Madani, and Richard M. Karp. Fast and intuitive clustering of web documents. In *Proceedings of the 3rd ACM SIGKDD Conference*, 1997.