Detecting and Browsing Events in Unstructured Text

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

ABSTRACT

Previews and overviews of large, heterogeneous information resources help users comprehend the scope of collections and focus on particular subsets of interest. For narrative documents, questions of "what happened? where? and when?" are natural points of entry. Building on our earlier work at the Perseus Project with detecting terms, place names, and dates, we have exploited co-occurrences of dates and place names to detect and describe likely events in document collections. We compare statistical measures for determining the relative significance of various events. We have built interfaces that help users preview likely regions of interest for a given range of space and time by plotting the distribution and relevance of various collocations. Users can also control the amount of collocation information in each view. Once particular collocations are selected, the system can identify key phrases associated with each possible event to organize browsing of the documents themselves.

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

visualization, interactive IR, information extraction

1. INTRODUCTION

As digital libraries and the Internet grow in size and complexity, users have a greater need to get a sense of the scope and contents of information resources. In the terms of Greene et al. [7], we need *previews* to help us quickly grasp the overall relevance of documents and collections and *overviews* to exhibit their structure and highlight possible subsets of interest.

Historical documents provide a wealth of information about past events in an unstructured form.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 there?", but they may not be best answered by ad hoc queries. Simply by restricting our question to a certain time or place, of course, we exclude many 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? If these problems can be addressed, however, users will be able to browse document collections by the common and well-understood dimensions of time and space.

The Perseus Digital Library Project (`http://www.perseus.tufts.edu`) has focused on developing automatic methods for structuring large document collections from many genres, subjects, and historical periods [3, 4]. We have previously worked on named-entity, term, and date identification and on place name disambiguation [15]. 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 our 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 compare statistical measures for determining the relative significance of various events. We have built interfaces that help users preview likely regions of interest for a given range of space and time by plotting the distribution and relevance of various collocations. Users can also control the amount of collocation information in each view. Once particular collocations are selected, the system can identify key phrases associated with each possible event to facilitate browsing the documents themselves.

2. PRIOR WORK

Although our testbeds contain primarily unstructured historical texts, it is useful to compare our approach with the Topic Detection and Tracking (TDT) study. TDT aims at developing techniques for "discovering and threading together topically related material from streams of data such as newswire and broadcast news" [18]. Topics are defined as specific events, "something (non-trivial) happening in a certain place at a certain time" [19] although some researchers use *event* to mean a single happening within a larger *topic*

SIGIR'02, August 11-15, 2002, Tampere, Finland.

story [10]. Due to its focus on news data, TDT possesses "an explicitly time-tagged corpus". 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 [17, 19]. Many historical documents discuss long-running events — e.g., wars in addition to battles —, and many users will wish to browse digital libraries at a scale larger than events of a few days' length.

3. DOMAIN DEPENDENCIES

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 a scalable solution would automatically discover which documents should be so segmented. Dividing documents into passages about a single time period would be a special case of such automatic topic partitioning systems as TextTiling [9] and automatic theme generation [14].

Most documents, however, although not neatly divisible into time-stamped chunks, 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. Finally, many historical documents are simply on a larger scale than news stories. Not only are books, and even chapters, often orders of magnitude longer than newspaper articles, 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; one could 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 [13], identifying bibliographic dates is easier than identifying (and linking) entire citations. Distinctions between the document creator's context and the context of the subject are not limited to date information. A place name associated with an author, such as an address, may have very little to do with the setting or topic of a document [12].

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, we often avoid 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 using χ^2 statistics [16]. Unlike other projects, we have privileged place names over other named entities since we can identify multiple names referring to a single place and distinguish uses of the same name for different places.

Since we cannot depend on our source documents having 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-place pair occurs 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 [5].

We have experimented with these statistics to test their effectiveness at ranking possible events. We have concentrated on relative ordering of events by significance rather than deciding on absolute relevance or irrelevance. As described below, users can select the amount of event infor-

Collection	Docs.	Words (millions)
London	53	13.0
California	186	12.8
Upper Midwest	140	16.2
Chesapeake	142	6.9
South	908	35.4
Civil War	237	56.4

Table 1: Collections on 19th c. history

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	995.16
Spotsylvania, Virginia	May 12 1864	66	994.90
Virginia, United States	1860	264	963.19
Pittsburg Landing, Tennessee	1862	124	881.62
Walcheren, Netherlands	1809	53	860.89
Gettysburg, Pennsylvania	1863	154	749.54
Chancellorsville, Virginia	May 3 1863	49	618.33
Crimea, Ukraine	1854	65	608.43
Atlanta, Georgia	1864	138	568.38
Huntsville, Alabama	1862	88	561.24
Great Britain, United Kingdom	1812	86	536.69
California, United States	1850	131	521.70
United States	1861	245	503.16

Table 2: 19th c. events: Ranked by log-likelihood

mation they want to see, and we hope this will effectively take them from short, highly precise lists, to total recall of all candidate events in the corpus.

4.1 Example Rankings

As an example, we compare the twenty top-ranked events by each test from a corpus of nineteenth-century historical documents (tables 2–4). 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 (table 5). Using a common rule of thumb in contingency table analysis, we exclude date-place pairs with fewer than five occurrences. Our collections for this period focus on British and U.S. history: a collection on the history and topography of London; one each on California, the Upper Midwest, and the Chesapeake region from the Library of Congress' American Memory project; a collection on the American South from ibiblio; and a collection of memoirs and official records of the U.S. Civil War (table 1). As one can infer from the table, many of these documents are quite long books; the London collection has an average of 245,000 words per document.

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 3: 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 4: 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 5: Ranked by raw association count

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. Interestingly, all of the χ^2 scores in these top twenty 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

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	160.520
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 6: Events in the 6th and 5th centuries BC, ranked by log-likelihood

list contain 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. Similarly, neither mutual information nor χ^2 highlight any collocations with places larger than individual towns or cities. All but seven events in the raw count list involve an entire state or nation. The log-likelihood list contains mostly specific towns or cities as well as six larger areas: three states (California twice and Virginia), two geographical regions (Great Britain and the Crimea), and one nation (the United States).

Even outside the scope of precise dates, log-likelihood ranking can perform well. Beyond the nineteenth century, there are fewer dates precise to the day. Tables 6 and 7 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 their effect by including the incorrect disambiguation of "Lade" for the United Kingdom instead of Greece. Still, decisive moments in Greek history are clear with the end of the Peloponnesian war at Aegispotami and of the Persian wars at Plataea. Our testbed does not contain any resources specifically for medieval history, but enough allusions are made in the 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. At these lower frequencies, the χ^2 measure seems to detect more spurious events (table 8).

4.2 Evaluating Rankings

Having some intuition about the characteristics of these ranking schemes, we can now try to quantify the differences among them. For the U.S. Civil War, Dyer's *Compendium of the War of the Rebellion* [6] includes a complete tabulation of all "battles, engagements, actions, skirmishes, etc." in that conflict. Each entry consists of a date range, a geographic name, an indicator of severity (e.g. "battle", "skirmish", "affair"), information on units engaged, and casualties. For this evaluation, we identified and disambiguated the toponyms in Dyer's list. Removing those places, such as "Bole's Farm", that could not be readily identified, we were left with 7602 distinct events.

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 7: Events in the 13th and 14th centuries

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 8: Events in the 13th and 14th centuries ranked by chi-squared

Although the primary goal of our system is visualizing the content of digital collections, any evaluation of various methods against such an *a priori* list risks saying more about the corpus than about the methods themselves. Our test documents simply have more to say about the battle of Shiloh than the battle of Springfield, Missouri. By contrast, the TDT list of topics only contains events that are represented in the corpus of news stories. Even so, events of greater significance had a greater chance of being included in the collection: 68% of all events listed by Dyer as "battles" were detected by our system, against 10% of all other events.

Using the mean reciprocal rank method employed by TREC evaluations, we compared the collocation-ranking systems. For each day of each event in Dyer, we checked whether Dyer's geographic location matched any of our detected date-place collocations for that day. If there was a match, we then recorded how far down the list of collocations for that date the match fell. The reciprocals of the ranks for each ranking scheme were then averaged.

Table 9 shows that log likelihood reliably fell at the top of the list; its advantage is especially marked for low-frequency

Excl. < 5	$-2 \log \lambda$	MI	χ^2	Count	α
Battles	0.780	0.724	0.751	0.757	0.50
All events	0.751	0.716	0.733	0.722	0.05
Incl. < 5					
Battles	0.542	0.174	0.327	0.507	0.03
All events	0.326	0.240	0.273	0.259	0.00

Table 9: Mean reciprocal rank in detecting Civil War events

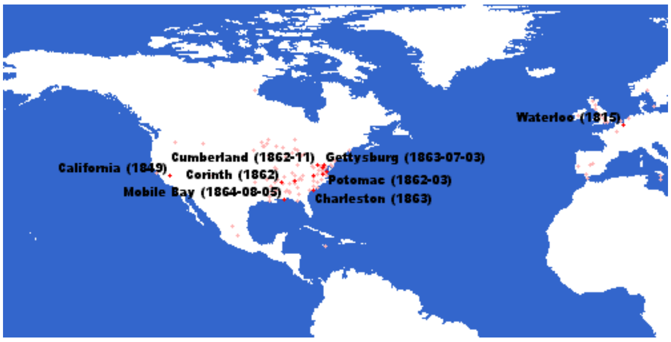

Figure 1: Significant collocations in the 19th c. Highlighted, labeled items include the battle of Waterloo, the eastern U.S. in the 1860s, and California in 1849.

collocations. The relative performance of raw counts, chi-squared, and mutual information also confirms our intuition. The table also shows that events of high importance — the major battles — tend to rank higher than all events in general. Although raw counts performed quite well, sign test statistics show that log-likelihood held a significant advantage over counts in all tests except the small number of battles with five collocations for each day. The sign test was computed by counting the number of cases where log-likelihood outranked raw counts and then calculating the probability that this was due to chance. As shown in the last column, significance (α) for higher frequency battles was inconclusive: there was merely a 50% chance that log-likelihood was the better ranking. In all other cases, log-likelihood outperformed counts with small probabilities of error.

5. BROWSING EVENTS

5.1 Map Browsing

We have developed an interface to explore these associations with a combination of graphical and tabular display. In addition to lists or timelines of significant events, we also generate global or regional maps. When the user selects a particular range of space or time — whether a century, decade, or year — the map is updated to show the sites of significant events in that range. 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 lighter in color. The top-ranked events are also listed, with date, place, and the number of times they co-occur in the digital library. The user can adjust the number of events that are listed and labeled on the map.

A high-level view, such as figure 1, can serve as a preview, informing the user of the geographic and temporal range of a collection or document. It can also be seen as an overview that guides the user to particular concentrations of data — in this case, for instance, to the California Gold Rush and the Civil War. Figures 2–4 cover a more restricted geographic area and only one year each. By sequentially browsing the time dimension, the user can gain a sense of the ebb and flow of information about this part of the world. Users could further analyze event data in a temporally-aware GIS such as TimeMap (`http://www.timemap.net`).

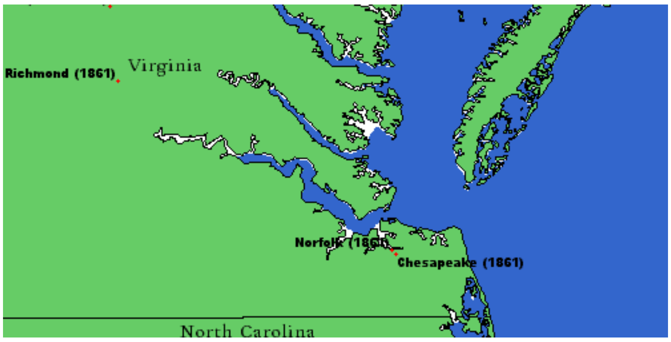

Figure 2: The Virginia Tidewater in 1861.

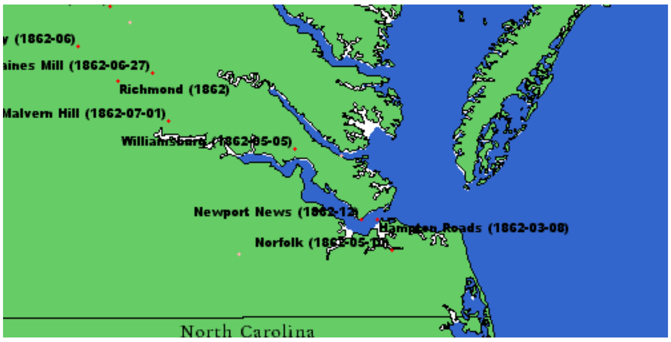

Figure 3: The Virginia Tidewater in 1862, showing action in the Peninsular Campaign.

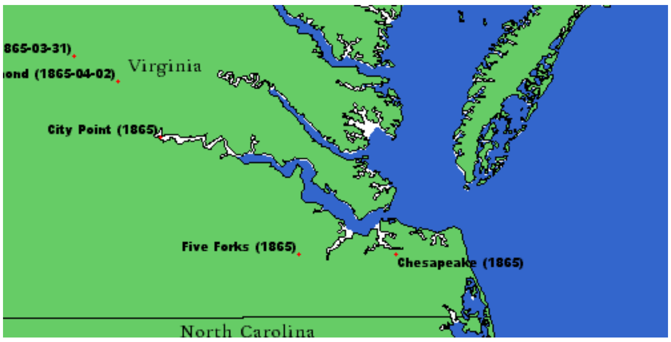

Figure 4: The Virginia Tidewater in 1865. Richmond fell on April 2. The map is empty for the remainder of the 1860s.

5.2 Phrase Browsing

If a user wishes to explore a candidate event more closely, he can click on the date-place collocation and call up a display of the individual passages from the digital library. Since our system disambiguates toponyms in texts, these searches are for the unique geographic identifiers, not string searches for the names themselves.

When searching for a combination of date and place, the default results display organizes retrieved passages by phrases common to two or more sentences. This display, first of all, takes advantage of the cluster hypothesis: result documents in one cluster are more likely to be relevant to the same topic [11, 8]. Thus, if more than one event has occurred in the given time and place, clustering can help to separate the documents pertaining to different events. The phrases that head each cluster can also provide a useful description of the event(s) contained in the retrieved documents. The retrieved passages could also be organized in other ways: by the document from which each passage comes or by personal names that co-occur with the event.

We produce the clusters at run time using a suffix-tree algorithm. As in [20], we can use this data structure to create a polythetic classification of search result passages as they are returned. For each sentence in the search results, we level case, remove punctuation, divide each sentence into words, and then strip off each suffix of a sentence: the first through last words, the second through last words, and so on. Once all these suffixes are inserted into a tree (specifically, a trie data structure), we can then easily determine all of the sentences that contain a given subsequence of words. These phrases are ranked by a score s that combines the number of words w in the phrase with the number of passages p in the cluster, using a cluster-constant c, usually set to 0.5 (equation 1; e is Euler's constant).

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

When cluster sizes are small, this formula favors longer common phrases, but for clusters with more documents, the cluster size will outweigh the phrase length. In normal display, we suppress clusters whose documents are a subset of a higher-ranked cluster. The examples show clusters for London, 1666, the date of the Great Fire (table 10); for California, 1849, the Gold Rush (table 11); for Atlanta, 1864, when a Union army captured the city (table 12); and, from the TDT3 corpus, for Libya, 1998, during the trial for the Pan Am bombing over Lockerbie, Scotland (table 13). Phrases containing dates are removed since they mostly show variations like "fire in 1666" and "fire in the year 1666".

These phrases can characterize events by listing associated people or places, such as the opposing generals Sherman and Johnston; Kofi Annan and Moammar Gadhafi; San Francisco; or Cape Horn, around which many sailed to California. Phrase clusters may also be more descriptive: "rebuilding of the city", "gold fever", "march to the sea", or "pan am bombing". The example from news texts also shows the extent to which multiple accounts of the same event can duplicate phrases such as "panel of scottish judges in the netherlands" (10 documents) or "libya has confirmed its seriousness and readiness to find a solution to the lockerbie problem" (7 documents, this from a quote by Kofi Annan). The user can also group passages by the book or collection

Phrase	Count	Score
fire of london	21	13.34
great fire	21	9.70
city of london	8	5.08
charles ii	6	2.77
act of parliament	4	2.54
duke of york	4	2.54
christ church oxford	3	1.91
house of commons	3	1.91
dreadful fire	3	1.39
rebuilding of the city	2	1.52
college oxford	3	1.39
privy council	3	1.39
view of london	2	1.27
burning of london	2	1.27
church of st	2	1.27

Table 10: Clusters for London, 1666

Phrase	Count	Score
san francisco	19	8.78
discovery of gold in california	8	6.79
discovery of gold	10	6.35
gold rush	9	4.16
united states	9	4.16
gold fields	7	3.23
trip to california	5	3.18
gold fever	6	2.77
cape horn	6	2.77
california gold	6	2.77
california during the years	3	2.28
early in the year	3	2.28

Table 11: Clusters for California, 1849

Phrase	Count	Score
military division of the mississippi	13	11.03
atlanta ga	19	8.78
atlanta georgia	18	8.32
atlanta campaign	14	6.47
march to the sea	5	3.81
major general	8	3.70
general sherman	7	3.23
sherman's army	5	2.31
effective strength of the army	3	2.54
advance on atlanta	4	2.54
battle of atlanta	4	2.54
capture of atlanta	4	2.54
general joseph e johnston	3	2.28
maj gen	4	1.85
kenesaw mountain	4	1.85

Table 12: Clusters for Atlanta, 1864

Phrase	Count	Score
secretary general kofi annan	31	23.61
secretary general	43	19.87
kofi annan	32	14.79
united states	29	13.40
lockerbie scotland	29	13.40
trial in the netherlands	13	9.90
panel of scottish judges in the netherlands	10	9.41
trial in a third country	11	9.33
moammar gadhafi	20	9.24
libyan leader	20	9.24
libya s foreign minister	12	9.14
foreign minister	19	8.78
hand over the two suspects	10	8.48
security council	18	8.32
pan am bombing	12	7.62
libya s official news agency jana	8	7.24
libya has confirmed its seriousness and readiness to find a solution to the lockerbie problem	7	7.00
travel to libya	11	6.99
libyan suspects	15	6.93
news agency	14	6.47
united nations	14	6.47

Table 13: Clusters for Libya, 1998

from which they come. The number of distinct documents recording a date-place collocation could be useful in deciding an event's significance.

6. CONCLUSIONS

Although historical documents do not often exhibit 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, date-place collocations can provide a useful generic interface to information 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 user can choose the amount of information to browse, we have concentrated on ranking events using statistical measures and have found evidence that the log-likelihood measure achieves a balance among spatial and temporal scope and frequency of occurrence. These ranking methods may also be useful for interpreting other kinds of collocations in text, such as co-occurrences of technical terms. We have built a browsing interface so that users can see regions of concentration within a document corpus and explore names and phrases associated with a given event. In future work, we hope to incorporate other document features into the event detection system. Since the distance between two places or dates is measurable, and not arbitrary as in some term models, we can work on grouping the data to minimize the aggregation effects of using individual days, years, or places as terms of association.

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, Jim French, Anne Mahoney, and the anonymous reviewers for their comments.

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] Gregory R. Crane, Robert F. Chavez, Anne Mahoney, Thomas L. Milbank, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman. Drudgery and deep thought: Designing a digital library for the humanities. *Communications of the Association for Computing Machinery*, 44(5):35–40, 2001.

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

[6] Frederick H. Dyer. *A Compendium of the War of the Rebellion.* Thomas Yoseloff, New York, repr. of 1908 edition, 1959.

[7] Stephan Greene, Gary Marchionini, Catherine Plaisant, and Ben Shneiderman. Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. *Journal of the American Society for Information Science*, 51(4):380–393, 2000.

[8] Marti Hearst and Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pages 76–84, Zurich, 1996. ACM Press.

[9] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual ACM SIGIR Conference*, pages 59–68, Pittsburgh, PA, June 1993.

[10] 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.

[11] Anton Leuski. Evaluating document clustering for interactive information retrieval. In *Proceedings of the 2001 ACM CIKM*, pages 33–40, Atlanta, GA, 5–10 November 2001.

[12] Kevin S. McCurley. Geospatial mapping and navigation of the web. In *Proceedings of the Tenth International WWW Conference*, pages 221–229, Hong Kong, 1–5 May 2001.

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

[14] Gerard Salton, Amit Sighal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207, 1997.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.