TimeMines: Constructing Timelines with Statistical Models of Word Usage

Russell Swan and David Jensen

Department of Computer Science
University of Massachusetts
Amherst, Massachusetts USA
{swan,jensen}@cs.umass.edu
http://www.cs.umass.edu/{~swan,~jensen}

ABSTRACT

We present a system, TimeMines, that automatically generates timelines from date-tagged free text corpora. TimeMines detects, ranks, and groups semantic features based on their statistical properties. We use these features to discover sets of related stories that deal with a single topic.

TimeMines requires free text with explicit date tags. We have used the system to generate overview timelines, indicating the most important topics in the corpus, how much coverage they receive, and their timespans.

Evaluation on two different news corpora show that non-random temporal patterns exist within the data, and the topics found by TimeMines correspond to the top news stories.

1. INTRODUCTION

Timelines are a logical, intuitive interface for time-dependent data. We have been investigating methods for automatically constructing overviews of text corpora suitable for browsing using timelines. We have built a system, TimeMines, that takes a time-tagged collection and generates an interactive overview timeline, showing the major topics covered by a corpus, and the dates of coverage of the topics[16]. We refer to these as "overview timelines". Figures 1 and 2 are sample timelines generated by TimeMines from a 1995 and a 1998 news corpus.

TimeMines makes few assumptions about the data – the data are free text, and explicit date tags are associated with each document. While some databases exist where each document has manually assigned keywords from a known hierarchy (e.g., medical journal articles with the Medical Subject Headings (MeSH), the Reuters news collections), most text collections do not supply information of this type. Natural language processing techniques and information extraction techniques induce produce data from text, but most systems either perform detailed analysis on an extremely narrow domain, or a simple analysis on a broad domain. We are interested in unrestricted free text, so TimeMines uses the more robust IE and NLP techniques for performing simplified data gathering. Along with this we are willing to accept some noise.

TimeMines uses robust, flexible techniques for determining significant keywords for documents, and judging the temporal significance of these keywords in the context of the corpus. To account for the limited attention span of human users, TimeMines restricts the amount of information that would be presented to the user, and only presents the most significant and important information. TimeMines uses statistical techniques, based on classical hypothesis testing, that are simple, extremely fast, and work surprisingly well at gisting a corpus and finding the most salient temporal features.

We have presented initial research results at information retrieval conferences[15, 16]. We present our research here because we feel our results are interesting from a knowledge discovery perspective — our selection of features is based on statistical techniques to find hidden patterns implicit within the data. The statistical model we propose is not novel; it is based on classical hypothesis testing. However, we have developed a novel approach to applying simple statistical tests to identify time-dependent features that identify important topics in text documents.

Our research addresses two hypotheses. They are

- A simple statistical ranking of term occurrence and co-occurrence can identify and group relevant documents into coherent time-dependent stories.
- These stories will prove comprehensible and useful to human users.

The paper is organized as follows: Section 2 presents our approach and discusses the building of our system. Section 3 presents our evaluation results on test corpora. We present new, previously unpublished research showing that our first

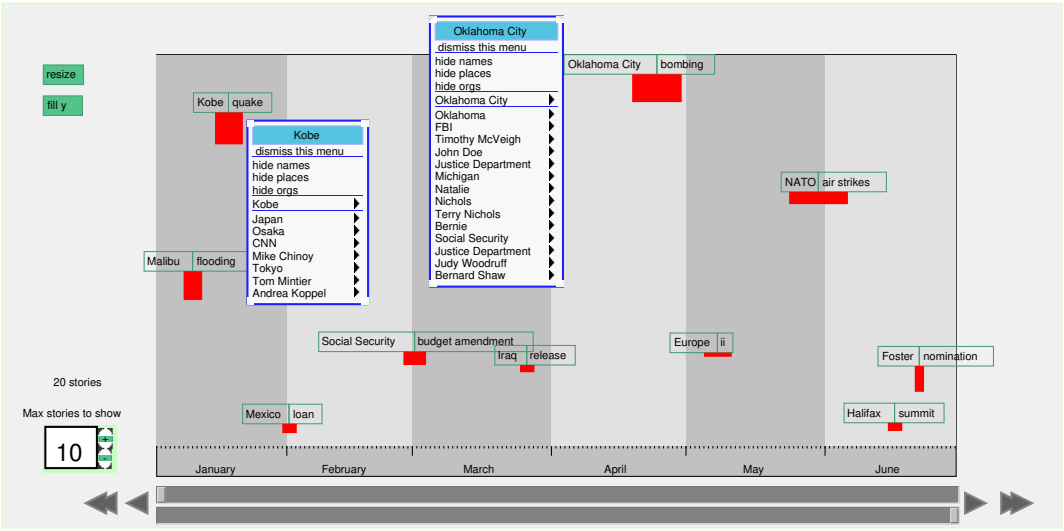

Figure 1: Overview of January - June, 1995. The topic labeled *Oklahoma City bombing* is the highest ranked topic, and the topic labeled *Kobe quake* is the second highest ranked. The pop-up on *Oklahoma City* shows significant named entities of *Oklahoma*, *FBI*, *Timothy McVeigh*, *John Doe*, *Justice Department*, etc. The other pop-up shows the terms associated with the Kobe earthquake.

hypothesis is valid, and review previously published evidence in favor of the second hypothesis. Section 4 discusses previous work on text data mining and timeline construction and Section 5 presents conclusions and future work.

2. SYSTEM OVERVIEW

The basic approach used by our system is diagrammed in Figure 3. The initial text corpus is assumed to contain explicit date tags, and explicit boundaries between documents or equivalent units. We choose a set of features to extract from our documents and propose a simple default model for the temporal occurrence of these features. The default model we propose is a stationary random model: the occurrence of a feature depends only on its base rate, and does not vary with time.

Any feature that matches our default model contains no new information relative to our context (the rest of our corpus) and should not be presented to the user. We statistically test for features that violate this model, and use this (greatly reduced) set in further processing. Our default model is then used again to statistically test whether these features should be grouped. The output is a collection of features related to topics in the news which can be visualized by a tool we have written, shown in Figures 1 and 2.

2.1 Extracting Features

We analyze our corpus by selecting a set of input features that are easily recognizable with high accuracy by current methods, and have strong semantic content. For a collection of text, these features could be words, word stems, nouns, verbs, noun phrases, named entities, or any other feature. We chose noun phrases and named entities as our features, though other choices are possible. Named entities are extracted from the text, where a named entity refers to a named person, location, or organization (e.g., Oklahoma City is a location). The text is also run through a shallow part-of-speech tagger and we label as noun phrases any groups of words of length less than six which matched the regular expression (Noun|Adjective)*Noun. Each document is then represented as a "bag of features" (similar to the "bag of words" model standard in Information Retrieval), and each document is explicitly marked with its date. We can then generate time series of the appearance of features over time, and perform statistical analyses on these time series. All the processing to this point involves off-the-shelf software and standard information retrieval and information extraction techniques.

2.2 Finding Significant Features

A disimple statistic for discrete events—the presence or absence of a specified feature—is the number of documents that both contain a feature and occur during a specified time interval. The model for the arrival of these features is a random process with an unknown binomial distribution. We assume that 1: the random processes generating the features are stationary, meaning that they do not vary over time, and 2: the random processes for any pair of features are independent. With this default model of word usage we define our interestingness function as the amount of deviation from our default model.

	f_0	$\overline{f_0}$
$t \in t_0$	a	b
$t \notin t_0$	c	d

If the process producing feature f_0 is stationary, then for an arbitrary time period t_0 the probability of seeing the feature is the same as the probability of seeing the feature at other time periods of equal duration. Specifically, looking at the number of documents within which we see f_0 during time t_0 (a in table), the number of documents where we do not see f_0 during t_0 (b in table), the number of documents containing f_0 when $t \notin t_0$ (c in table), and the number of documents not containing f_0 when $t \notin t_0$ (d in Table), gives a 2×2 contingency table.

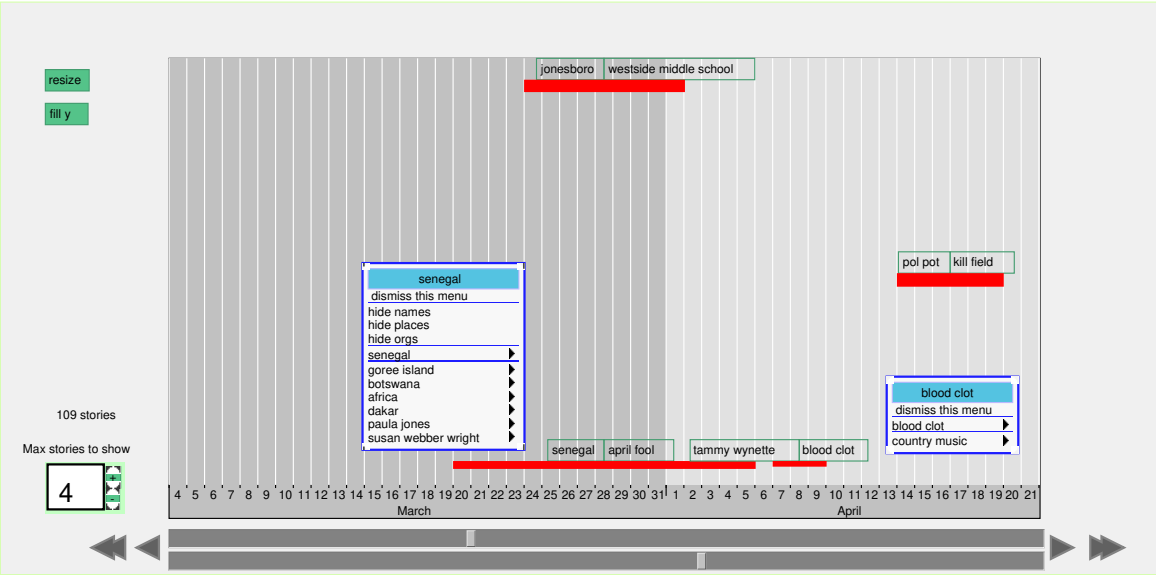

Figure 2: Detail, March and April, 1998. The pop-up menu displays features from news stories when the Paula Jones suit was dismissed. The shooting at Jonesboro high school was the top news event during this period.

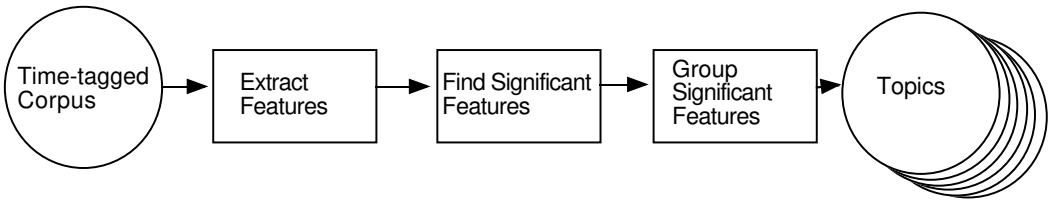

Figure 3: Process steps to discover features in text.

Many statistics can be used to characterize a 2×2 contingency table, including χ^2, ϕ^2, Expected Mutual Information Measure (EMIM) or Kullback-Leibler (KL). We chose χ^2 as the most appropriate statistic for the model we are using. KL, EMIM, and ϕ^2 are primarily used for measuring the strength of associations between features. The χ^2 statistic, while not as effective as the others for measuring strength of association, can easily be used to test the statistical significance of an association. Our default model is that most occurrences of features are uninteresting, and a feature's appearance is interesting for a sample only if it is drawn from a different distribution within the sample period. This hypothesis is explicitly tested by χ^2, and not by any of the other statistics frequently used on 2×2 contingency tables.

Implementing hypothesis tests is difficult, both due to both the nature of the data and the type of statistical inferences we wish to make. Tabulated values of the chi-square distribution are inaccurate for several reasons. The most fundamental reason is our use of a multiple comparisons procedure [11]. For a given feature and a given span of time, the results of a χ^2 test can be compared to some fixed threshold. If we select a threshold value corresponding to $p = 0.05$ this correspond to the probability of obtaining this result (or a better result) if the null hypothesis were true. If we have a years worth of data and perform a χ^2 test on every day, using the 5% significance level, then on purely random data we would expect 18 days to be falsely marked as significant.

Our system searches through individual days to initially find significant values. It then combines contiguous timespans and recalculates the statistic over these new timespans. Given multiple values of a test statistic x, the maximum of several values of that statistic $x_{max} = max(x_1, x_2, \ldots, x_n)$ will be distributed differently than any single value. That is, the sampling distribution of any one value will be a poor approximation to the sampling distribution of x_{max}. Thus, significance tests must account for the number of values compared to obtain x_{max}. This causes the sampling distributions for χ^2 values obtained by the system to be wildly different from the sampling distributions appropriate for a single χ^2 value. Determining the appropriate sampling distribution analytically is difficult due to the correlation among the χ^2 values. The χ^2 still ranks correctly, but we need a method for determining a selection threshold. For this reason we performed randomization tests (described below) to empirically approximate the correct sampling distributions and select thresholds.

2.3 Grouping Significant Features

The results of our first set of hypothesis tests is a greatly reduced set of features (the majority of features are statistically significant so this step causes a large reduction in the number of features tracked and ultimately displayed to

the user) and a timespan for each feature indicating when the coverage of that feature was significant. These features represent a small subset of the original corpus, and are representative of major news topics covered by the corpus. There are usually many features associated with any given topic, so to aid comprehension, we group related features together and display them to users as a topic.

	f_j	$\overline{f_j}$
f_k	a	b
$\overline{f_k}$	c	d

To group these features we invoke our second assumption. The assumption that two features f_j and f_k have independent distributions implies that $P(f_k) = P(f_k|f_j)$. We test this for the timespans where features f_j and f_k are significant. The resulting counts also form a 2×2 contingency table where a is the number of documents in the timespan where f_k and f_j co-occur, b is the number of documents where f_j occurs without f_k, c is the number of documents where f_k occurs without f_j, and d is the number of documents in the timespan containing neither feature. Note that in the first table, N, the total count, is equal to the total number of documents in the corpus, whereas in the second case it is equal to the number of documents occurring in the time window, a far smaller number.

3. EVALUATION

We have performed several experiments with TimeMines attempting to test our first two hypotheses[15, 16]. We have evaluated our system using two corpora from the Topic Detection and Tracking studies[2, 13].

3.1 Test Corpora

The first corpus was derived from the Topic Detection and Tracking (TDT) pilot study's corpus[2]. That material consists of manually transcribed news articles from CNN broadcast news and Reuters newswire from July 1, 1994, to June 30, 1995.

We used JTAG[20] as our part-of-speech tagger, and a named entity extraction system called Badger IE[10]. Badger parsed the text to find locations, organizations, and names of people. The corpus was enhanced to include these named entities with markups. Unfortunately, the extraction system was built and tuned for another collection and it was too fragile to work well on all of our test corpus.

The original TDT corpus included 15,683 news stories. Failures of the Badger system forced us to use a subset of those stories. Specifically, we used stories 9001 through 15683—that is, 6683 stories over 175 days spanning January 7 through June 30, 1995.

We used the TDT-2 corpus for our second corpus. The TDT-2 corpus contains text transcripts of broadcast news in English and Chinese. We used only the English language portion, consisting of articles from ABC News, CNN, Public Radio International, Voice of America, the New York Times, and the Associated Press newswire, spanning from January 1, 1998 to June 30, 1998. This corpus was used in the TDT-2 task, and is divided into three sections: training (January and February), development (March and April), and evaluation (May and June). The corpus was divided in this way

Feature	Date Range
Oklahoma City (loc)	April 20 - April 29
Kobe (loc)	Jan 16 - Jan 20
Oklahoma (loc)	April 20 - April 27
FBI (org)	April 20 - April 27
Timothy McVeigh (pers)	April 21 - April 28
NATO (org)	June 2 - June 5
John Doe (pers)	April 21 - April 27
Japan (loc)	Jan 16 - Jan 20
Osaka (loc)	Jan 16 - Jan 18
NATO (org)	May 25 - May 27

Table 1: Top 10 named entities in TDT-1 by χ^2 value

Feature	Date Range
oklahoma	April 20 - April 29
oklahoma city	April 20 - April 29
f-16	June 2 - June 5
kobe	Jan 16 - Jan 20
bosnia	May 25 - June 8
bombing	April 20 - April 29
quake	Jan 16 - Jan 20
bosnian serbs	May 25 - June 8
serbs	May 24 - June 6
bosnian	May 25 - May 26

Table 2: Top 10 noun phrase features in TDT-1 by χ^2 value

in order to allow a holdout validation, with initial model development on the training data, model refinement being performed on the development data, and final system evaluation being performed on the evaluation sub-corpus. BBN tagged the TDT-2 corpus for us using the Nymble tagger[5], and supplied us with this marked up corpus, consisting of 56,974 articles. We used JTAG as our part-of-speech tagger. We identified 184,723 unique named entities and 1,188,907 unique noun phrases.

3.2 Top Ranked Topics

We tried three different methods of selecting cut-off values for our χ^2 tests, specifically, treating the statistic as though it came from a standard χ^2 distribution with one degree of freedom[15]; dividing our corpus into training and evaluation sets, and varying the values during runs of the training

Story	Date Range
Oklahoma City Bombing	April 20 - April 29
Earthquake in Kobe, Japan	Jan 16 - Jan 20
F-16 shot down over Bosnia	June 2 - June 5
NATO forces in Bosnia	May 25 - May 27
Flooding in California	Jan 10 - Jan 11
NATO forces in Bosnia	May 29 - May 31
Senate debates Balanced Budget	Feb 28 - Mar 2
Russia/US Summit	May 6 - May 10
Two Americans Sentenced in Iraq	Mar 25 - Mar 27
Henry Foster rejected by Senate as Surgeon General	June 21 - June 22

Table 3: Top 10 stories as calculated by named entity statistics (labels manually assigned)

Story	Date Range
Oklahoma City Bombing	April 20 - April 29
F-16 Shot down in Bosnia	June 2 - June 5
Earthquake in Kobe, Japan	Jan 16 - Jan 20
NATO air strikes in Bosnia	June 2 - June 3
Senate debates Balanced Budget	Feb 28 - Mar 2
Flooding in California	Jan 10 - Jan 11
???(*march, kuwait*)	Mar 21 - March 30
Scott O'Grady rescued	June 7 - June 10
???(*june, saturday*)	June 8 - June 17
U.N. Peacekeepers in Bosnia	June 5 - June 7

Table 4: Top 10 stories found from noun phrase features (labels manually assigned)

Feature	Date Range
Iraq (loc)	Jan 26 - Feb 25
Jonesboro (loc)	March 24 - March 31
Iraqi Foreign Ministry (org)	Feb 21 - Feb 23
Nagano (loc)	Jan 31 - Feb 23
Davos (loc)	Jan 31 - Feb 2
Barry Goldwater (pers)	May 29 - May 30
Pol Pot (pers)	March 15 - March 19
Ross Rebagliati (pers)	Feb 11 - Feb 14
Duisenberg (pers)	May 2 - May 3
v.o. (loc)	May 20 - May 22

Table 5: Top 10 named entities in TDT-2 by χ^2 value

sets, while looking at the topics generated[16]; and performing randomization tests on the corpus in order to select a threshold sufficiently high to make any random occurrences extremely unlikely. The randomization results are new and are reported in Section 3.3; the other results have been previously published, and are presented in Section 3.4. We found that no matter which method was chosen the most highly ranked features stayed the same, as did the most highly ranked topics, and differences between the methods were observed only in lower ranked items. Tables 1 and 2 show the top features found in the TDT-1 corpus.

For the TDT-2 corpus the most highly ranked items were due to an artifact of processing. A large number of articles contained the following copyright statement:

> COPYRIGHT 1998 BY WORLDSOURCES, INC., A JOINT VENTURE OF FDCH, INC. AND WORLD TIMES, INC. NO PORTION OF THE MATERIALS CONTAINED HEREIN MAY BE USED IN ANY MEDIA WITHOUT ATTRIBUTION TO WORLDSOURCES, INC.

For the dates April 22 through April 24, this copyright statement appeared within the <TEXT> </TEXT> sgml tags. On all other dates this copyright appeared before the <TEXT> tag. This copyright was treated as text for the three days in question and was included in our bag of features model. This resulted in the copyright notice being indexed 83 times during those three days (961 documents) and never being indexed outside those days (56013 documents), which led to χ^2 values > 4500.

Disregarding the copyright, the top entities and noun phrases in the TDT-2 corpus are shown in Tables 5 and 6. The tenth

Feature	Date Range
henshen	March 30 - March 31
easter	May 9 - May 13
three-hour meeting	Feb 22 - Feb 23
arat	March 9 - March 11
iraq	Jan 26 - Feb 25
clock saturday	Feb 22 - Feb 23
st patrick	March 15 - March 17
jonesboro	March 24 - March 31
naval armada	Feb 22 - Feb 23
westphele	April 14 - April 15

Table 6: Top 10 noun phrase features in TDT-2 by χ^2 value

feature in the named entity list, v.o., is an artifact from Voice of America (V.O.A.) tags. Henshen, the highest ranked of the noun phrases, is a reporter for V.O.A.

We have generated our topics separately for nouns and named entities, and by combining them. We find that the combined method works better. Table 3 shows the top 10 topics detected in TDT-1 by grouping named entities, and Table 4 shows the top 10 topics detected by grouping noun phrases. The entries labeled "???" in Table 4 correspond to topics found by the system that we were unable to determine what the topic might be. The terms associated with each topic are:

> march, kuwait
>
> june, saturday;

The top 10 topics TimeMines generated from TDT-2 are presented in Table 7 and in Figure 4.

The top news topic of the first half of 1998 was of course the Monica Lewinsky story. TimeMines did not select this topic as among the top ten. TimeMines measures how distinctive a feature's occurrence is against the background of the remainder of the corpus. Monica Lewinsky and the Starr investigation were mentioned so frequently throughout the entire period of the corpus that even days of frequent mentions did not deviate from the average values by very large amounts. If the corpus had comprised a longer period, or the background was taken as a different time period, the Monica Lewinsky story would almost certainly have appeared as the most highly ranked story. Even though the Monica Lewinsky story was heavily mentioned throughout the corpus, there was still enough variation in the amount of coverage for TimeMines to choose it as the 12th most important topic.

3.3 Randomization Test

Initial inspection of the results show TimeMines performs well—the features ranked highly tend to be related to events in the news, and the groupings of features into topics tend to be understandable. We are interested in performing formal evaluations of our first two hypotheses. Our first hypotheses states "A simple statistical ranking of term occurrence and co-occurrence can identify and group relevant documents into coherent time-dependent stories." To verify that the patterns we found are inherent in the data we performed a randomization test[8]. We shuffled the temporal order of the

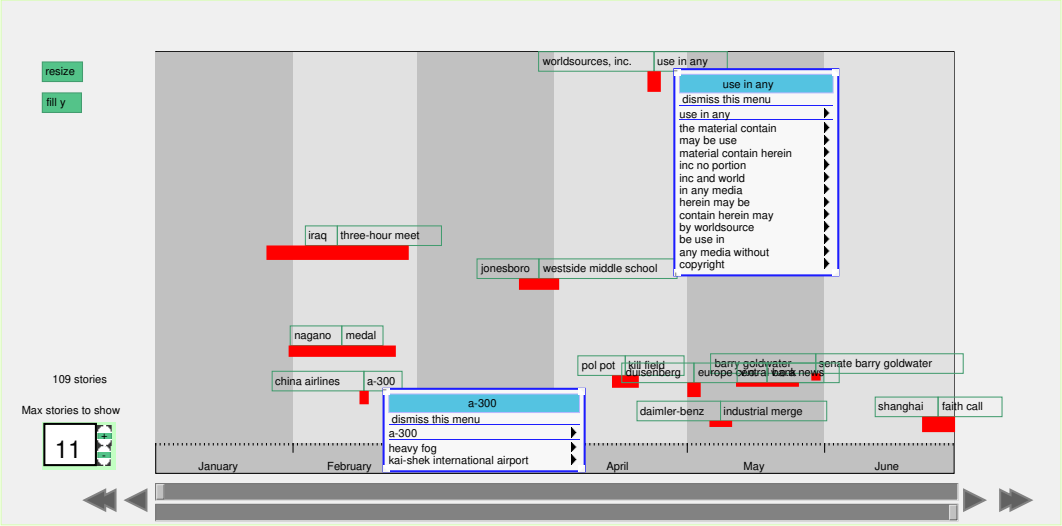

Figure 4: Top 10 topics (and copyright) for TDT-2. Detail is shown on noun phrases of copyright notice and China Airlines crash.

Topic	Date Range
U.S. Confrontation with Iraq *iraq, iraqi foreign ministry, three-hour meeting, ...*	Jan 26–Feb 26
Shooting at Westside Middle School, Jonesboro *jonesboro, westside middle, westside middle school, ...*	March 24–April 1
1998 Winter Olympics *nagano, ross rebagliati, medal, snowboarder, ...*	Jan 30–Feb 25
Barry Goldwater dies *barry goldwater, senator barry goldwater, arizona senator*	May 29–May 31
Pol Pot dies *pol pot, khmer rouge, killing fields, ...*	April 14–19
Introduction of the Euro *duisenberg, jean-claude trichet, european central bank, ...*	May 1–May 5
Unrest in Indonesia *habibie, indonesia, president suharto, ...*	May 12–May 27
Crash of China Airlines A-300 Airbus *china airlines, china airlines airbus, a-300, ...*	Feb 16–Feb 17
Clinton debates Jiang Zemin *shanghai, xi'an, faith call, ...*	June 23–June 30
Daimler-Benz/Chrysler merger *daimler-benz, daimler chrysler, industrial merger, ...*	May 6–May 10

Table 7: Top 10 stories from TDT 2 (labels manually assigned) Included are the top ranked named entities and noun phrases.

documents in the corpus 100 times and ran TimeMines on each of the shuffled corpora. Each document was left intact, but the putative date of each document was changed, and the frequency of documents on each date was kept fixed. We did this to remove any information that may be contained in the document dates. By keeping the documents fixed, we preserved the statistics that information retrieval (IR) systems use, *tf* (term frequency) and *idf* (inverse document frequency). The result of this is that each of the shuffled corpora has the exact same information content from the viewpoint of an IR system, and the different corpora would be indistinguishable to the user of an IR system.

For each shuffled corpus, we ran TimeMines and recorded the highest scoring named entity and noun phrase. We then sorted these scores and selected as our threshold the 10th highest ranked score. Using this threshold corresponds to a 10% probability of selecting at least one feature given that all features are random. The thresholds we obtained were 160.0 for named entities, and 194.0 for noun phrases. If our data were random, the most likely result from running our system with these thresholds would be 0 named entities selected, and 0 noun phrases selected. The second most likely statistic would be 1, then 2, etc. From the number of features selected by TimeMines we should be able to bound the probability that this number could be generated by random data. Using these thresholds TimeMines selected 979 named entities and 2689 noun phrases, that grouped into 109 topics. These numbers are far higher than could be explained by chance, and effectively represent a probability of 0. From this we conclude that the patterns we are discovering in our data are valid patterns.

3.4 Topic Based Evaluations

Our second hypothesis states "These stories will prove comprehensible and useful to human users." Our experience confirms this, but we have not yet devised a formal evaluation. We have attempted two different evaluations but both had problems. Our first attempt was an IR-style evaluation, where we had a "truth" set that we compared our

results against to measure precision and recall scores. To get a "truth" set we used the 1995 Year-In-Review section of the January 1996 *Facts on File*[1]. There were 24 stories during the time period of our TDT-1 corpus that *Facts on File* identified as major stories, and our system found 28 stories. The overlap was only seven stories, giving recall of 29% and precision of 25%. Further analysis showed that many of the stories chosen by Facts on File were either not mentioned in our corpus or were only briefly mentioned, and many of the stories chosen by TimeMines and not in Facts on File were arguably as important as the chosen stories. Due to the misalignment between our corpus and our truth set we are not able to arrive at a valid result, positive or negative[15].

We attempted another evaluation where we tuned the χ^2 threshold by running TimeMines on a training corpus (the first four months of TDT-2), then did a final run on the evaluation portion (May and June). We had four students read the grouped features and judge whether they constituted one topic, multiple topics, or no topics. The students were allowed to read the documents from which the terms had been taken to help in making their decisions. The four evaluators found that the great majority of groups were indicative of a single topic (71.2%, 79.4%, 82.2% and 90.2% of the groups judged), and the pairwise overlap on the judgments of how many topics were contained in a group was 73.6%. However the overlap expected by chance was nearly 70%, and the pairwise Kappa statistics ranged from 0.045 to 0.315, with a (weighted) average value of 0.223. The Kappa statistic is a measure of inter-evaluator reliability, and a value of 0.0 indicates an overlap that would be expected by chance and a value of 1.0 indicates perfect overlap. A Kappa value of 0.233 indicates that the data are not reliable, and no conclusions should be drawn[16].

We feel that the information found by TimeMines is both interesting and useful, and we have no evidence to the contrary, but we have not yet developed a formal evaluation that can measure this.

4. RELATED WORK

There have been a large number of systems built for the purpose of browsing the information within text collections. Examples include I^3R [17], Kohonen Maps [19], *Themescapes* and *Galaxies*[18], and *Galaxy of News*[14]. These systems select significant words and phrases and display them in such a way as to allow the user to graphically gist what topics are contained in a system. All these systems are term centered rather than document centered, as is TimeMines, but none of these systems makes explicit use of time. Timelines as an interface for document collections are an active area of research[4, 12], but there has been little to no discussion of what kind of models of term usage are appropriate for automatically selecting and grouping terms for display in a timeline.

This work was motivated by and heavily influenced by the Topic Detection and Tracking study[7, 3, 21]. This study involves analyzing time tagged streams of broadcast news in order to detect the occurrence of a new topic, and to track stories on known topics as they unfold. TimeMines differs from TDT in that TDT is intended to run in near real time, and as such can only use information from prior articles, and TimeMines runs in a retrospective fashion, and TDT is intended to alert the user of the system upon the first appearance of a new topic, whereas TimeMines is designed to provide an overview to a corpus, including a ranking function of how important different topics within a corpus are.

Dagan and Feldman built the *Knowledge Discovery from Text*[6, 9] system (KDT). KDT performed KDD operations on a text corpus, specifically a news corpus (Reuters-22173) containing approximately 22,000 articles. Each article was tagged with a set of keywords describing its content, where the keywords were from a predetermined hierarchy (*Japan* and *Germany* are both examples of *countries* and *G-7 countries*). The data that were then investigated with data mining algorithms were the distribution of keywords across articles, and the co-occurrence information of keywords. Data about the co-occurrence of keywords was labeled as interesting if it deviated greatly from the expected value. For expected values, several different measures were used: how the distribution of the data conditioned on a specific keyword compared to the distribution compared to similar related keywords (sibling nodes); and how the distribution at a specified time compared with the distribution at a later time. The KDT system reports associations between keywords in different categories, for example, comparing the economic activity of South American countries to European countries shows that economic activity in South America is more concerned with agriculture and rare metals, while European economic activity is more concerned with financial instruments. The temporal data the system reports are simple trend analyses, such as the percentage of articles about OPEC countries that were about crude oil as a function of time. The graphical display is a trendline. The KDT system used the Reuters corpus, where all the articles had been hand tagged with keywords, and all the keywords were from a known pre-defined hierarchy that included concepts such as *car manufacturer*, *G-7 country*, *computer company*, and relationships such as *Germany is a G-7 country* and *Germany is a European country*. The system makes heavy use of the knowledge contained in the keyword categorization scheme, and performs mining on those associations.

5. CONCLUSIONS AND FUTURE WORK

We have presented a system — TimeMines — that automatically selects the most significant topics from a time-tagged news corpus. TimeMines uses basic information extraction and natural language processing components in a black box fashion, allowing improved components to be quickly integrated into the system as they become available. The system only uses standard, well-defined technologies, and is able to extract the most salient temporal features using classical statistical methods. The statistical model is quite simple, though the implementation details are more involved. We have shown through a randomization test that the patterns discovered by TimeMings in the data are truly in the data, and could not be produced by chance.

Perusal of the grouped features selected by TimeMines shows them to be highly indicative of major topics covered by the corpus. We have attempted two formal evaluations of the quality of the groups, but have been unable to get conclusive results.

TimeMines is an interesting example of the usage of statistics to extract explicit knowledge from data. The model is quite simple, and produces surprisingly effective results for such a simple model.

6. ACKNOWLEDGMENTS

We would like to thank Bruce Croft for initially suggesting this area of research and supplying us with the name TimeMines. We would also like to thank James Allan and Victor Lavrenko for their numerous helpful suggestions.

This material is based on work supported in part by the Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623 and supported in part by the National Science Foundation under grant number IRI-9619117, and in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912, and also in part by DARPA/AFOSR contract F49620-97-1-0485. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] *Facts on File, 1996*. Facts on File, New York, 1997.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[3] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45, 1998.

[4] R. B. Allen. Timelines as information system interfaces. In *Proceedings International Symposium on Digital Libraries*, pages 175–180, Tsukuba, Japan, 1995.

[5] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a high-performance learning name-finder. In *Fifth Conference on Applied Natural Language Processing*, pages 194–201. ACL, 1997.

[6] Ido Dagan and Ronen Feldman. Keyword-based browsing and analysis of large document sets. In *Proceedings of the Symposium on Document Analysis and Information Retrieval (SDAIR-96)*, Las Vegas, Nevada, 1996.

[7] DARPA, editor. *Proceedings of the DARPA Broadcast news Workshop*, Herndon, Virginia, February 1999.

[8] E. Edgington. *Randomization Tests*. Marcel Dekker, New York, NY, 1995.

[9] Ronen Feldman and Ido Dagan. Knowledge discovery in textual databases (kdt). In *Proceedings of the First International Conference on Knowledge Discovery (KDD-95)*. ACM, August 1995.

[10] D. Fisher, S. Soderland, J. McCarthy, F. Feng, and W. Lehnert. Description of the umass systems as used for muc-6. In *Proceedings of the 6th Message Understanding Conference, November, 1995*, pages 127–140, 1996.

[11] David D. Jensen and Paul R. Cohen. Multiple comparisons in induction algorithms. *Machine Learning*, 38:1–33, 2000.

[12] Vijay Kumar, Richard Furuta, and Robert B. Allen. Metadata visualization for digital libraries: interactive timeline editing and review. In *Proceedings of the third ACM Conference on Digital libraries*, pages 126–133, Pittsburgh, Pennsylvania, July 1997.

[13] Ron Papka, James Allan, and Victor Lavrenko. Umass approaches to detection and tracking at TDT2. In *Proceedings of the DARPA Broadcast Workshop*, 1999.

[14] E. Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *UIST 94, ACM Symposium on User Interface Software and Technology*. ACM, 1994.

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

[16] 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 (these proceedings)*, Athens, Greece, 2000. Association for Computing Machinery.

[17] R. Thompson and W. Croft. Support for browsing in an intelligent text retrieval system. *International Journal of Man-Machine Studies*, pages 639–668, 1989.

[18] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visulaizing the non-visual: Spatial analysis and interaction with information from text documents. In Stuart Card, Jock Mackinlay, and Ben Shneiderman, editors, *Readings in Information Visualization: Using Vision to Think*, San Francisco, California, 1999. Morgan Kaufmann.

[19] L. Xia, D. Soergel, and G. Marchioni. A self-organizing semantic map for information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval*, pages 134–141, Chicago, August 1991. ACM.

[20] Jinxi Xu, J. Broglio, and W. B. Croft. The design and implementation of a part of speech tagger for english. Technical Report IR-52, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, 1994.

[21] Y. Yang, T. Pierce, and J. 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, 1998.