Automatic Generation of Overview Timelines

Russell Swan and James Allan

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, Massachusetts USA
{swan,allan}@cs.umass.edu
http://www.cs.umass.edu/{~swan,~allan}

Abstract We present a statistical model of feature occurrence over time, and develop tests based on classical hypothesis testing for significance of term appearance on a given date. Using additional classical hypothesis testing we are able to combine these terms to generate "topics" as defined by the Topic Detection and Tracking study. The groupings of terms obtained can be used to automatically generate an interactive timeline displaying the major events and topics covered by the corpus. To test the validity of our technique we extracted a large number of these topics from a test corpus and had human evaluators judge how well the selected features captured the gist of the topics, and how they overlapped with a set of known topics from the corpus. The resulting topics were highly rated by evaluators who compared them to known topics.

Keywords: UIs/visualization for collection overviews, text data mining, statistical/probabilistic models, event detection and tracking.

1 Introduction

We are interested in finding statistical techniques that enable more efficient organization of information. The world is awash in information, much of it in free text with little or no metadata, and there is a tremendous need for tools to help organize, classify, and store the information, and to allow better access to the stored information.

Over the years there has been much research in IR regarding implicit information in documents that can be obtained during indexing. Examples of this type of information are co-occurrence data, clustering of terms within a corpus, and clustering of documents within a corpus. While there has been initial research in these areas, the statistics used in IR systems are almost exclusively *tf* and *idf*.

There has also been substantial research in developing methods to categorize the type of information in a corpus or sub-corpus, for machine-made decisions about which corpora are fruitful for further exploration. In addition there has been much research on browsing interfaces for exploring information collections, which can be viewed as a method to allow humans to obtain a sense of the type of information stored in a corpus in order to perform their own categorization on the corpus.

We are interested in using timelines as a browsing interface to a document collection (see Figure 1). Timelines are a well known interface and are simple and intuitive for most people to use. We feel that if timelines were widely available to a variety of corpora they would increase the accessibility of the information to a large number of people. In order to make timelines readily available we need to find ways to construct them automatically.

We have developed a technique for determining the relative importance of the occurrence of extracted features within text. Our technique requires an explicitly time tagged corpus. With this technique we are able to analyze extracted features (named entities and noun phrases) and explicitly rank how likely these features are to be high content bearing. We are then able to group these features into clusters that correspond strongly with the notion of "topic" as defined in the Topic Detection and Tracking (TDT) study. This technique can be run at indexing time and introduces a small additional overhead to the classical indexing performed—in fact, the basic data structure required for our technique is the inverted list, so our process can be performed during or immediately after the creation of inverted lists.

This process produces a ranked list of groups of features that correspond to significant events in the news that are discussed in the corpus. For each group we get a relative ranking of importance, a range of dates when it was important, and an indication of the amount of coverage in the corpus. For each group we also have a measure of how distinctive or surprising it is and how many distinctive terms are associated with this topic. This information can be used to provide an overview timeline of a corpus, which is exactly what is desired for browsing. An example is shown in Figure 1. The x-axis represents our time scale, and the position on the y-axis represents the importance of the story, with the most important stories appearing at the top. The number of terms used in describing the story is represented by the area of the block, so that two stories each represented by 20 terms would have equal areas, indicating roughly equal coverage

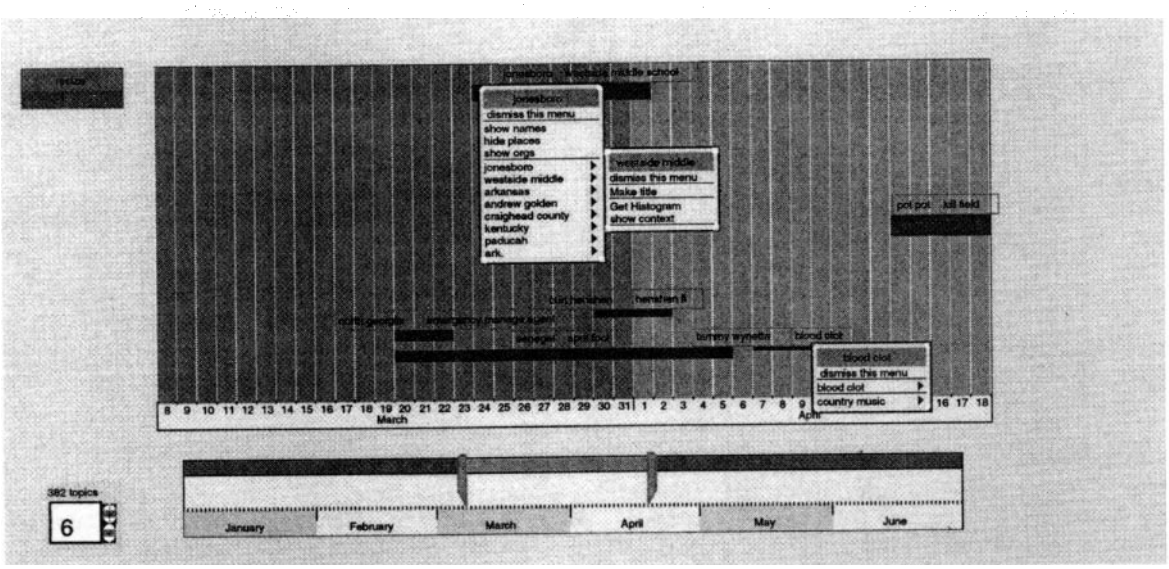

Figure 1: Overview of the topics from March 8 - April 18. The highest ranked topic is the school shooting at the Westside Middle School in Jonesboro, Arkansas. The second highest ranked topic is the death of Pol Pot. The pop-up menu below *jonesboro* displays the prominent places in the story (Andrew Golden was tagged as a location by the named entity extractor), and the sub-menu attached to *westside middle* allows the user to make that feature a title, or get more information about that feature. The pop-up menu adjacent to the topic on the death of Tammy Wynette shows the phrases *blood clot* and *country music* as being the most significant for that topic.

within the corpus. Clicking on a label term (named entity or noun phrase) pops up a menu of all the associated features of that type within the story, and a sub-menu option allows the user to choose this feature as the label, or to obtain more information about the feature. This timeline display provides a very fast graphical overview of the information a corpus contains.

In this paper, we describe a system that automatically extracts the most interesting and significant topics covered in a time-tagged corpus and displays the results to the user with a highly interactive user-directed interface. We believe we have developed statistical techniques that are effective at finding interesting topics, and that timelines are an effective form of presentation. This paper is concerned with our first hypothesis, that we have developed effective statistical techniques. Only after we are satisfied that we know *what* to display will we begin investigating *how* to display it. As such, this paper is primarily concerned with the methodology of automatically selecting features from a corpus for display. The timeline examples given are just one possible use of the extracted information, and we will not address the usefulness or the usability of the timeline in this paper.

The remainder of the paper is organized as follows: Section 2 discusses our model of feature occurrence, Section 3 discusses similarities and differences of other systems to ours, and Section 4 describes our experiments. Our conclusions and plans for future research are presented in Section 5.

2 Model

A simple statistic for discrete events—the presence or absence of a specified feature—is the number of documents containing a feature occurring during a specified time interval. The model for the arrival of these features is a random process with an unknown binomial distribution. We assume two hypotheses as defaults. The assumptions are 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.

	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 times. 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.

There are a wide range of statistics that can be used to characterize a 2×2 contingency table, for example 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. Our default assumption is that there is *no* association between features—the appearance of a feature in a given document, or the co-occurrence of two features, we assume to be devoid of meaning until it is shown to be statistically very unlikely. Only then do we consider features as being related. The χ^2 statistic,

while not as effective as the others for measuring strength of association, is an excellent statistic for distinguishing random association from true association.

	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 $P(f_k) = P(f_j|f_k)$. 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 Prior Work

There have been a large number of systems built for the purpose of browsing the information within text collections. Examples include I^3R[16], Kohonen Maps[18], *Themescapes* and *Galaxies*[17], and *Galaxy of News*[13]. 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, but none of these systems makes explicit use of time.

There has been recent research on the use of timelines as a front end for collections of data with explicit time tags, where the data is from a database of personal histories[12] or as an alternative to a hierarchical file system[9]. There has also been research on the use of timelines specifically for document collections[2, 11], 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[6, 1, 20]. 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. This work 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 this system runs in a retrospective fashion.

Our initial system was built in early 1999 and our preliminary research is reported elsewhere[14]. The preliminary research used the original TDT1 corpus. The results can be summarized as follows: Named entities and noun phrases produce interesting groupings of words. Named entity groupings were of higher quality, but noun phrases were more descriptive. Both should be used. Due to the larger number of noun phrases a higher threshold should be used for selecting noun phrases in order to reduce some of the noise. Our clustering in the preliminary research was very simple and produced surprisingly good results for such a crude system, but needed refinement. We also had problems with our evaluation, as we used a listing of the top news stories from Facts on File as our "truth" judgments, and the stories did not correlate precisely with our corpus.

Our model, described in the previous section, is very closely related to a model put forward by Conrad and Utt that formed the basis of their *Association System*[4]. They selected named features and measured associations between pairs using both EMIM and ϕ^2. They found the strongest results with ϕ^2. We instead use χ^2. For the contingency table in the prior section, the equations for ϕ^2 and χ^2 are

$$\phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

Thus we see that ϕ^2 is defined as χ^2/N, where N is the number of documents or text windows in the corpus. Since N is constant within a corpus, and they use ϕ^2 as a sort key, and we use χ^2 as a sort key, the statistics are identical. However, our use of χ^2 allows us to use the language of classical hypothesis testing to explicitly state our null hypotheses.[1]

The *Association System*, like the previously mentioned systems, does not make explicit use of time. The associations that are found tend to be implicit ones throughout the corpus, whereas our system finds associations that have a significant time locality. For example, in our initial study, the name *Boris Yeltsin* was strongly associated with the locations *Halifax* and *Nova Scotia*, when Boris Yeltsin attended the G-7 summit there. Using the *Association System* we would expect to find a correlation with *Moscow* or *Russia*, but not *Halifax*.

Dagan and Feldman built the *Knowledge Discovery from Text*[5, 8] 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, and data mining algorithms were used on the distribution of keywords across articles, and the co-occurrence information of keywords. 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.

[1] As a rule of thumb, χ^2 is an appropriate statistic for a 2×2 table only if the expected values for all cells is at least 5. With the very rare events we see here, $E(a) < 1.0$ over 90% of the time. At these values, the χ^2 statistic understates the true probability. We use Yates Continuity Correction on χ^2 for all our calculations, which provides higher accuracy for rare events, but this too is biased for extremely rare events. The only statistic that gives correct probabilities for these rare events is the Fisher Exact Test, but the test is expensive, and since it involves large numbers of factorials it tends to overflow in C. However, sorting based on both χ^2 and Fisher's Exact Test give the same results, so χ^2 is perfectly appropriate as a ranking function.[7]

4 Experiment

4.1 Goals

Our system finds clusters of terms that we believe are highly indicative of major news topics, enabling the automatic construction of a "Year in Review" type of interface. We had also noticed in our prior research that concatenating the top ranked named entity and the top ranked noun phrase often produced highly descriptive labels (e.g., *Oklahoma City bombing*, *Kobe quake*), and when the initial label was not good the second ranked noun phrase often was (e.g., *Mexico good morning* to *Mexico loan*). We are interested in performing a formal evaluation, both to empirically establish the validity of this approach, and to provide a means of evaluating the choice of design parameters. The initial questions we would like answered are: "Are the clusters of terms *good*? Are they representative of some news topic?" and "Can we label the clusters automatically?"

The majority of the evaluations performed in IR consist of document retrieval tasks. The task we are performing here is not document retrieval, but *is* information retrieval in the broad sense of the term, and requires some method of evaluating the quality of the snippets and chunks of information our system generates.

In a classic IR corpus there are a collection of queries and a collection of relevance judgments. The relevance judgments require human evaluators and are expensive to obtain, but once obtained, the relevance judgments can be used repeatedly, and systems can be run and evaluated automatically. Ultimately we would like to be able to find or construct a similar corpus for our task. This would consist of a corpus with an exhaustive judgment of all topics that are contained in the corpus. Definitions of how much coverage are required to make a topic would be difficult, and accurate recall numbers would be difficult to obtain. We would also need some method of determining whether or not a cluster matched a topic. We do not know of an automatic method for this. In the current experiment we employed human assessors to answer our questions.

4.2 Corpus

We used the TDT-2 corpus for our experiment. Though not matching our wish list, the TDT-2 corpus has some desirable properties. The TDT-2 corpus has 192 topics with known relevance judgments. Of these, 96 have exhaustive relevance judgments where an evaluation was done for every document in the collection, and the other 96 have a pooled evaluation, with many documents tagged, but not all documents have been compared to those topics. Each document was marked as being relevant, not relevant, or "brief", meaning that the document contained a brief mention of the topic, but was not primarily on that topic. These 192 topics are not an exhaustive listing of all the news topics contained in the corpus, nor do they represent the most strongly reported topics—some topics had only one or two articles in the corpus. As such, we cannot use the judged topics as a relevant set for our topic extraction, and generate precision and recall by comparing our topics with the judged topics. The judged topics represent a collection of news topics that the assessors at Linguistic Data Consortium (LDC) felt were a coherent describable event that was reported in the news, and for each topic there is a brief description of what comprises the topic, and a list of documents that refer to that topic. While comparing our generated topics with the LDC judged topics does not allow us to generate precision/recall scores, it does allow us to compare some of our generated topics with some judged topics. There should be a significant overlap between these two sets, and there should be some correlation between features extracted by our system and human assigned descriptors.

The TDT-2 corpus contains text transcripts of broadcast news in English and Chinese spanning from January 1, 1998 to June 30, 1998. This corpus was used in the 1998 TDT 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 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. We used only the English language portion of the TDT-2 corpus, consisting of articles from ABC News, CNN, Public Radio International, Voice of America, the New York Times, and the Associated Press newswire. BBN supplied us with an annotated version of the English language portion, where named entities were marked by the Nymble tagger[3], which identified 184,723 unique named entities. We then extracted noun phrases by running a shallow part of speech tagger[19], and labeling as a noun phrase any groups of words of length less than six which matched the regular expression (Noun|Adjective)*Noun. This led to a set of 1,188,907 unique noun phrases. The corpus BBN supplied us with contained 56,974 articles.

With the choice of the TDT-2 corpus and its known topics, we added a third question for our evaluation: "Does this cluster of phrases correspond to any of the TDT-2 topics?"

4.3 System

The system generates inverted lists for both the named entities and the noun phrases. It then divides the corpus into days and calculates the number of documents containing a feature on each day. Knowing the number of documents from a given day, the number of documents on that day containing f_i, the total number of documents in the corpus, and df for the feature, we can calculate the χ^2 value. We only calculate χ^2 if the occurrence on that day is higher than would be predicted by random, otherwise we assign the value 0. Figure 2 shows the number of documents per day containing the phrase "air power" and the χ^2 values calculated from the occurrences.

Our χ^2 value is compared to our threshold and runs of consecutive days above the threshold are combined into a single range. For "air power" we get two ranges, one from June 11–12, and one from June 14–15. Prior experience has taught us that frequently there are single day gaps in the range when a feature is in the news, so we combine any ranges separated by no more than a day; here, we combine the two ranges to get "air power" being a significant feature in the news from June 11–15.

We need a measure of how distinctive this feature was at its peak value. To derive this, we calculate the χ^2 values for every subrange of the selected range, and

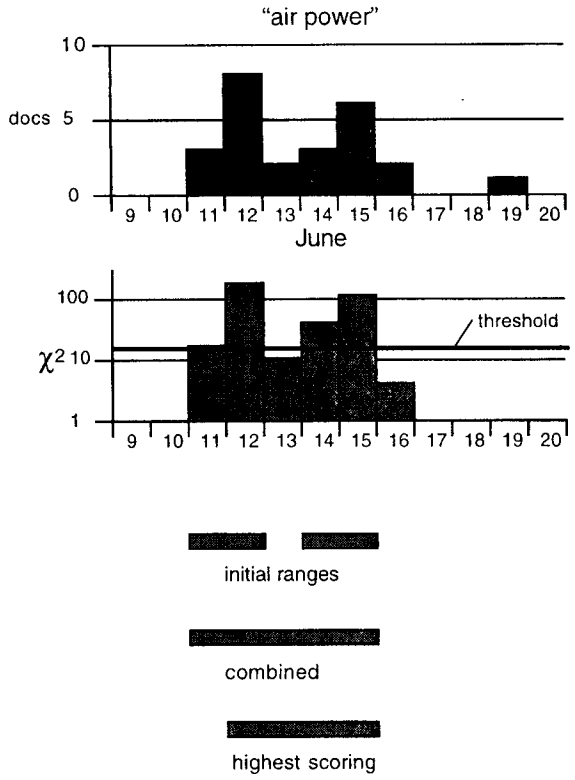

Figure 2: Determination of the time range and χ^2 score for the noun phrase *air power*. The top graphic shows the number of documents containing the phrase for a period of 12 days. The second graphic shows the χ^2 values for these occurrences.

choose the highest value. In this case, the 19 occurrences from June 12–15 have a χ^2 value of 387.94, so we choose that as our score.

The χ^2 calculation is fast. These steps (initial calculation of χ^2, tagging, aggregation, and calculating of maximum χ^2) take 12 seconds on an alpha workstation for the evaluation sub-corpus (21,255 documents, 287,472 initial χ^2 calculations) and 13 seconds for the full corpus (56,784 documents, 802,593 initial χ^2 calculations). (Before the calculations can be performed, the inverted list is first fetched from disk. The measured times include the time for the data fetches, which overwhelms the calculation time.)

After selecting terms with significant appearances in the news and associated ranges we sort on the maximum χ^2 value. This gives us a sorted list of the most significant features in the corpus and their dates. We then cluster these features into topics. The method we use is to take the highest ranked unclustered feature, and compare the time ranges with all lower ranked features. If the dates overlap, we perform a χ^2 calculation, and if the value is above a (fairly low) threshold we mark this feature as a potential member of the cluster. When we finish processing the list, we perform a standard hierarchical agglomerative clustering on the marked features. We then cut the dendrogram at a prespecified threshold, and take as our valid cluster the one containing the original central element, provided there was at least one named entity and two noun phrases.

Parameter	Threshold
Named entity	6.635
Noun phrases	15.827
Initial clustering	3.841
Final clustering	7.879

Table 1: Threshold values used in χ^2 tests for system.

We knew from prior experiments that we would need different thresholds for the named entities and the noun phrases, and that we would also need to investigate different clustering techniques. Our evaluation method would be to compare the generated clusters with the known topics, which was the final evaluation our assessors would be performing. In order to avoid fitting our data for the final evaluation, we set aside the evaluation section of TDT-2 (May and June) and built and trained our system on the training and developments sets.

Three different clustering schemes were investigated: single link, average link ,and complete link. These clustering operations are expensive ($O(n^3)$), however, our preprocessing of the possible matches on both dates and initial matches with the leading feature reduced the potential clusters to sizes on the order of a few hundred elements. Sorting and clustering the features took 23 seconds on the evaluation corpus and 91 seconds on the full corpus. Since these operations can all be performed at indexing time the additional overhead is small.

We found complete link clustering to be too restrictive. If there are a single pair of phrases that do not show a high correlation within an otherwise good cluster, this pair will split the cluster into two clusters. Single link also does not work well. We had some poor clusters in our previous work which were due to a single link clustering. If a single noise word links strongly to two disjoint clusters, single link clustering will combine them. With single link, we often saw one big cluster such as "Saddam Hussein, Monica Lewinsky, Richard Butler, Hillary Clinton, Davos, Nagano, Lillehammer". Average link clustering tended to produce uniformly good results, and was tolerant of minor weighting errors.

Our final parameters are presented in Table 1.

4.4 Evaluation

Our final run on the evaluation portion of TDT-2 produced 146 clusters. We believe that the clusters of features found are indicative of the major news stories that were covered by the news organizations during the time spanned by the corpus and provide a good summation of these topics. To test this, we hired four students (three undergraduates and one graduate student) to evaluate the clusters. A list of hyperlinks to stories that these features were extracted from was provided in a separate frame. The evaluator was also given a list of LDC-provided topics that overlapped in time with our cluster, and a title for each topic. Each topic contained a hyperlink to the LDC-supplied topic description which opened in a separate frame, along with a list of hyperlinks for relevant stories. Clicking on a story hyperlink brought up a new browser window containing the story.

We provided a list of TDT-2 judged topics, with checkboxes, where the evaluator could check off which topic(s) was represented, as well as text boxes where the evaluator could enter their own topic descriptor, if none matched. We used javascript to perform consistency checking, requiring that the number of named topics (0, 1, or >1) matched the number of topics (0, 1, or many) that had previously been identified.

For each cluster, a cluster label was automatically assigned, consisting of the highest ranked named entity followed by the highest ranked noun phrase, eg, "karla faye tucker execution", "daimler-benz industrial merger". The purpose of this label was to allow a very brief encoding for display. For example, in Figure 1 the block labeled *jonesboro westside middle school* is clearly about the Jonesboro school shooting, and if a user had no interest in the topic they do not need to read more than those four words. On the other hand, the label *senegal april fool* is unclear, and the user would need to look at more of the cluster to discern the topic. The evaluator was asked to rate how helpful these labels were, on a six point Likert scale.

The clusters produced by our system were randomized and the users were given different starting points in the system so we could guarantee coverage.

4.5 Results

Of the 146 clusters 79 were judged three times, and 67 had four judgments. The four evaluators found that the great majority of clusters were indicative of a single topic (71.2%, 79.4%, 82.2% and 90.2% of the clusters judged), and the pairwise overlap on the judgments of how many topics were contained in a cluster 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 poor agreement among evaluators and that the data are not reliable. This can also be seen by looking at the scores given individual clusters. Only twenty of the 146 were not judged to be a single topic by the majority of assessors, and of these twenty there were only three where the assessors unanimously agreed.

The poor agreement between assessors on what constitutes a topic is not very surprising, as debates on what topic means have occurred throughout the TDT research project. Our assessors were students whose instructions did not include a definition of the word "topic" nor explicit criteria for deciding what was or was not a topic. Each assessor felt that overall our system was good at finding what they considered to be a "topic", but they did not agree among themselves on what that meant.

We also asked the assessors to compare the generated clusters with the TDT-2 topics and indicate if they agreed. Here the results were stronger. The (pairwise) overlap in topic/cluster matches was 86.7%, and the six pairwise Kappa statistics ranged from 0.600 to 0.785, with an average value of 0.699, indicating very good agreement. This indicates that if a topic is defined, the features our system selects are sufficient for recognizing the topic.

Feature	Date Range
Barry Goldwater (pers)	May 29 - May 30
Duisenberg (pers)	May 1 - May 5
V. O. (loc)	May 20 - May 22
Jean-Claude Trichet (pers)	May 2 - May 3
Shanghai(loc)	June 28 - June 30
Daimler-Benz (org)	May 6 - May 10
Xi'an (loc)	June 27 - June 28
Wim Duisenberg (pers)	May 2 - May 3
Habibie (pers)	May 20 - May 26
Forbidden City (loc)	June 27 - June 28

Table 2: Top 10 named entities by χ^2 value

Feature	Date Range
phonetic	May 25 - May 27
g-8	May 15 - May 17
memorial day	May 22 - May 25
barry goldwater	May 29 - May 30
124th	May 1 - May 3
goldwater	May 29 - May 30
world cup	June 9 - June 30
nuclear test	May 11 - May 19
memorial day weekend	May 22 - May 25
habibie	May 20 - May 25

Table 3: Top 10 noun phrase features by χ^2 value

The clusters of terms were automatically labeled and our assessors were asked to rate the usefulness of the label on a six point Likert scale. In general our assessors felt that the labels were very poor, with an average rating of 2.8 (1 = poor, 6 = excellent). Our assessors were in good agreement on the ratings, with the average standard deviation equaling 1.0. The 10 labels that our assessors felt were the best were *volusia county wildfire, daimler-benz industrial merge, 91 fm june fundraise, discovery mir, air france strike pilot, iraq travel ban, new delhi underground nuclear test, shirley capaci powerball, ramon delgado france open tennis, lisa hackney lpga championship.* Eight labels were unanimously rated as the least descriptive (rating = 1). They were *steve crevice, v.o. v.o.a news, smiley militarism, susan jeans zawadzk, sarah williams williams in, medan live ammunition, paul westfeeling phonetic, cliff richard chri de.* The score given the label by the assessors correlates strongly with the rank of the cluster assigned by the system ($r_s = 0.889, p < 10^{-6}$). (The ordering of the clusters were randomized before being given to the assessors, and the assessors began at different points, so we can rule out order effects.)

The ten most highly ranked named entities and noun phrases from the evaluation sub-corpus are presented in Tables 2 and 3. The ten most highly ranked clusters, and the assigned labels, are presented in Table 4. Figure 1 displays part of the interactive timeline generated for the complete TDT-2 corpus, comprising six months of data.

5 Conclusions

We presented a technique for generating clusters of named entities and noun phrases that capture the information

Story (Label)	Date Range
Barry Goldwater dies (barry goldwater senate barry goldwater)	May 29 - May 31
Introduction of the Euro (duisenberg europe central bank)	May 1 - May 5
Riots in Indonesia (v.o. v.o.a news)	May 12 - May 27
Clinton visits China (shanghai faith call)	June 23 - June 30
Daimler-Benz / Chrysler Merger (daimler-benz industrial merge)	May 6 - May 10
PGA tour (payne stewart amateur matt kuchar)	June 16 - June 22
Indonesian president releases political prisoners (b.j. habibie b.j habibie)	May 18 - May 31
Frank Sinatra dies (frank sinatra wee small hour)	May 15 - May 19
India Tests Nuke (group of eight g-8)	May 11 - May 20
Pakistan tests nuke (pakistan nuclear test)	May 25 - June 2

Table 4: Top 10 topics from TDT-2 evaluation corpus. Story tags are manually generated, and the labels are system generated.

corresponding to major news topics covered in the corpus. This was evaluated with human assessors, who felt that clusters selected were very indicative of news topics.

Figure 1 is a screen shot from a demonstration system we built in Squeak[10], which runs on Windows 95, Windows NT, Macintosh, IRIX, and OSF systems. Our system is discussed in more detail as a demonstration[15].

This study is highly promising, but it raises more questions than it answers. These questions merit further research.

We found that our automatically generated labels are of poor quality. However, the quality of the labels corresponds strongly to how highly ranked the topic is, and the most highly ranked documents tend to have reasonable labels. For a general overview of ten or twenty topics the labels may be adequate. We intend to perform further research on this point.

Perusal of the clustered features show them to be of high quality, but we have been unable to formally verify this with two different evaluations. We are actively investigating methods for evaluating content summarizing systems. As the scope of information retrieval increases non standard evaluations such as is needed here will become more important.

This work was motivated by the TDT study. TDT is an on-line task, and for simplicity we have restricted ourselves to a retrospective task here. While the retrospective task is highly useful (automatically supplying a timeline as a table of contents to an archived corpus increases the accessibility of the information) we have not yet tested these techniques in an on-line setting.

This system selects features based on how distinctive they are using the corpus as its own reference. The two corpora we have analyzed so far (TDT-1 and TDT-2) each had a long running story that was in the news for the entire period of the corpus (the O. J. Simpson trial and the Monica Lewinsky case, respectively). With a larger reference corpus, or a reference corpus from a different time period, these two stories clearly would have been the top topics. There was enough variation in the amount of coverage that these topics did appear, but not at a very high rank. These long running topics sometimes disappear and reappear with certain key features remaining constant and other features changing, but our system cannot currently collect these topics separated in time into long running topics. Altering the reference time period, modifying the clustering algorithm, or devising rules for linking features that appear in several different topics might allow us to discover long running topics.

The techniques presented in this study can make a significant contribution to the accessibility of information, as it allows the automatic generation of interactive overview timelines at modest cost. As archives of news, e-mails, historical newspapers, memos, and other such time based corpora become increasingly common in digital libraries we feel that this system, or one like it, will be a useful tool to allow broader access to electronic information.

6 Acknowledgments

We would like to thank Dan Bergeron, Troy Dube, Sarah Ford and Joseph Sullivan for their work in evaluating the clusters. We would like to thank Victor Lavrenko for doing much of the pre-processing work on the corpus, and Daniella Malin and David Jensen for their helpful comments. We would also like to thank the reviewers for their helpful comments which allowed us to clarify several parts of this paper.

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 also 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, Any opinions, findings and conclusions or recommenda-

tions expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] 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.

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

[3] 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.

[4] J. G. Conrad and M. H. Utt. A system for discovering relationships by feature extraction from text databases. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval*, pages 260–271, Dublin, 1994. Association for Computing Machinery.

[5] 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.

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

[7] B. S. Everitt. *The Analysis of Contingency Tables.* Chapman and Hall, London, 1977.

[8] 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.

[9] Scott Fertigand Eric Freeman and David Gelernter. Lifestreams: an alternative to the desktop metaphor. In *Human Factors in Computing Systems CHI '96 Conference Proceedings*, pages 410–411, Vancouver, Canada, 1996. Association for Computing Machinery.

[10] Dan Ingalls, Ted Kaehler, John Maloney, Scott Wallace, and Alan Kay. Back to the future: the story of squeak, a practical smalltalk written in itself. In *Proceedings of the 1997 ACM SIGPLAN conference on Object-oriented programming systems, languages and applications (OOPSLA'97)*, pages 318–326. ACM, November 1997.

[11] 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.

[12] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. Life lines: Visualizing personal histories. In *CHI'96 Conference Proceedings*, pages 221–227, Vancouver, BC, Canada, 1996.

[13] 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.

[14] 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.

[15] Russell Swan and James Allan. Timemine: Visualizing automatically constructed 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. demonstration.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.