Newsblaster Subsystems

ABSTRACT
Descriptions of clustering and summarization systems.

1. INTRODUCTION

Over the past five years, our research has focused on technology to classify and summarize multiple documents and to track events across time using clustering. We have integrated these technologies in a news browsing system XXX which aims to help a user keep abreast of current news using these novel techniques. Our system crawls the web every morning for news and classifies its findings into five broad categories of interest (US, World, Entertainment, Finance, Sports). Inside each topic, related news articles are clustered into events, related events are linked with a second clustering level, and articles about the same event are summarized.

Typically, summaries are generated on clusters of between five and 30 news stories, but in cases where an event generates a high level of interest, XXX creates a much larger cluster; for example, on January 17th, XXX generated a summary of 60 articles all of which described the Enron scandal. The system is available on the web (links will be given after blind review) and is used by a wide community of users daily.

The key features of our system include:

- Categorization of news into five broad top-level categories, using a novel smoothing approach for category selection based on binning [?].
- Dynamic organization of articles into two hierarchical levels, one corresponding to events and one to sets of related events, using state-of-the-art clustering technology augmented with learned linguistic features [?].
- Multi-document summarization using different strategies depending on input article type (e.g., biographical articles vs. event-based articles) where summary structure and wording is automatically determined. [?, ?, ?, ?].
- Automatic augmentation of summaries with related images [?].

We have developed a robust, online system which mirrors the news interfaces manually developed at sites such as Yahoo News. Recent DARPA sponsored conferences (e.g., TREC, DUC) show that language technology performs well in simulated, test settings. An intriguing question is whether this technology is useful in real-world applications and whether systems such as ours can really help users in their everyday information needs. For example, will users really find a summary more useful than a set of key words describing the article? Does organizing and presenting articles according to events help users find needed information?

In this paper, we overview our news browsing system, aiming to answer the question of real-world applicability. We first provide an overview of XXX and then we describe in detail the different evaluations that we carried out. While popularity of the system, as revealed by the number of regular visitors, is evidence that it is useful, we wanted to pinpoint which features are useful and why. We used a variety of approaches to evaluation, testing usability, user satisfaction and user preferences. Our evaluation shows that users strongly prefer the advanced functionalities offered by the system and are satisfied with the performance of the summarization and clustering components. User comments that we collected during the evaluation provide us with insights on system usability and contribute suggestions for improvement.

2. SYSTEM DESCRIPTION

The XXX system follows a pipeline architecture, shown in 1. First, the system crawls the web for news articles, followed by a pre-processing phase which normalizes the text into a standard form and extracts images. The news articles are first categorized into one of the five top-level categories. Within each category, they are clustered into event clusters, then these clusters are grouped at a higher level, putting together related events. A multidocument summary is created for each event cluster and augmented with pictures extracted from the articles.

2.1 Gathering News

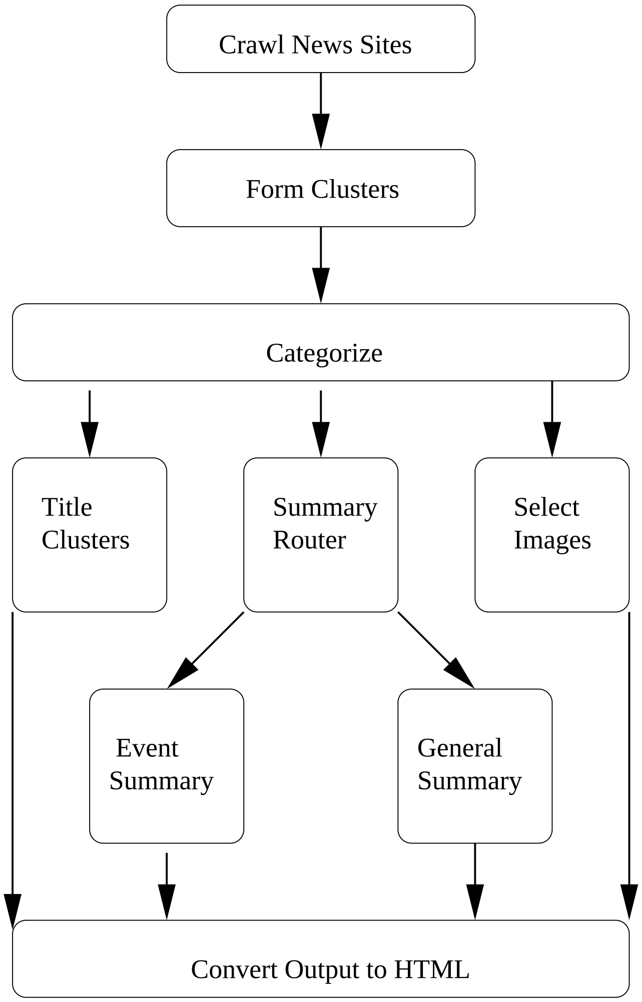

Figure 1: XXX System Architecture

XXX currently crawls 13 news sites including those of CNN, Reuters, Fox News, NY Post, and USA Today. The list of sites is stored in a text file and may change over time. Each site is traversed up to some maximum depth (currently 4) and only links within the site are considered. For each page examined, if the amount of text in the largest cell after stripping tags and links is greater than some particular constant (currently 512 characters), it is assumed to be a news article, and this text is extracted.

2.2 Organizing News

XXX hierarchically classifies the news stories gathered by the crawler into three levels. At the top level, it uses text categorization to determine whether the story falls into one of five pre-determined categories: US, World, Financial, Entertainment or Sports. Sample results for two of these categories are shown in Figure 5. Text categorization uses an approach based on smoothing bins [?] to determine the category of all the articles in the cluster. The cluster itself is assigned a category based on the majority of the articles.

Within each broad category of articles, XXX further organizes the news stories into two hierarchical levels. The lowest level corresponds to articles on the same event, while the higher level groups together related events. For example, on a recent day (??/??/??) the first group of articles in the World category had three clusters on different events related to developments in Afghanistan—one about the prisoners from Afghanistan, one about the hunt for bin Laden, and one about a Taliban surrender. While related, each of the clusters is a different event which has its own summary page which contains the articles about the event, and a summary like the one shown in Figure 5.

For both of these clustering steps, we use a hierarchical clustering system [?]. The system uses agglomerative clustering with a groupwise average similarity function. Groupwise-average performed significantly better in tests conducted on TDT-2 data than the single-link and complete-link methods and better than the single-pass clustering algorithm.

The system is distinguished by its use of not only the usual TF*IDF weighted words, but also linguistically motivated features, such as noun phrase heads and proper nouns. These are likely to correlate with events and not simply topically related stories. The features are automatically computed from the output of two text-processing systems, LinkIt [?] and Nominator [?].

A log-linear statistical model [?] operates on the extracted features to automatically adjust the relative weights of the different features. The model first calculates a linear internal predictor based on the weighted sum of predictors and then produces a final response via the logistic transformation. The log-linear model is quite similar to linear regression, which has been used successfully for combining features for text classification before [?]. The log-linear extension guarantees that the final response will lie in the interval $(0, 1)$, with each of the endpoints associated with one of the outcomes. Given modest assumption about the distribution of the predictors, the optimal set of weights can be calculated efficiently using the iterative reweighted nonlinear least squares algorithm [?]. The approach seeks to optimize the similarity function, rather than the evaluation metrics over the clustering, but our experiments showed that the weights assigned by this procedure were often better than those produced by extensive searching.

For its use within XXX, we have empirically determined two thresholds for the clustering at each of the two levels (single event and group of related events).

To facilitate the user's interaction with the categorized stories, we also provide labels for each cluster. For the lowest (event) level, where the articles are closely related in content, we use heuristics to select the article that is most related to the other articles in each cluster, and label the entire cluster with that article's title (underlined on the main page as in Figure 5). For the second level (related events) we extract all proper names and terms from the articles in the cluster, then weigh the terms according to their total frequency and inverse document frequency, and select the top five terms that are most representative of the related events (shown in bold, heading each group of events in Figure 5).

2.3 Summarizing News

The XXX summarizer is a composite summarization system that uses different summarization strategies depending on the type of documents in each cluster; this contrasts with other systems which typically perform one type of summarization only [?]. A router automatically determines the type of documents in each cluster and invokes the appropriate summarization subcomponent.

Using the training corpus provided for the recent Document Understanding Conference (DUC) [?], we manually derived a typology of document sets. *Single-event* documents center around one single event happening at one place and at roughly the same time, involving the same agents and actions. *Person-centered* (or "biography") documents deal with one event concerning one person and include background information about that person. *Multi-event* documents describe several events occurring at different places and times, and usually with different protagonists. There is a common theme to these events, e.g., a cluster might collect many fire incidents on unrelated cruise ships. The time span covered is unpredictable, but longer than in the single-event case. *Other* clusters contain even more loosely related documents and don't fit any of the categories above.

To summarize documents on the same event, the XXX summarizer uses an enhanced version of MultiGen [?, ?]. MultiGen integrates machine learning and statistical techniques to identify similar paragraphs [?, ?], intersection of similar phrases within paragraphs [?], ordering of the selected themes [?] and language generation to reformulate the wording of the summary. For biographical documents, it uses an alternate system, DEMS (Dissimilarity Engine for Multidocument Summarization) [?], tuned to the biographical task; and for sets of loosely similar documents, it uses DEMS with a more general configuration. DEMS selects sentences that contain information important or interesting enough to be included in a summary by combining several features that are critical for new-information detection with some traditional heuristics used in single-document summarization [?].

The articles collected are stripped of HTML coding, and the router uses a group of simple heuristics to decide which summarizer to use for the articles. These include the overall time span between publication dates, the proportion of articles published on the same day to the total number in the set, the frequency of capitalized words to determine if one named entity dominates the set, and the frequency of the personal pronouns "he" and "she" to determine if the named entity is a person.

2.4 Multigen

MultiGen summarizes a specific type of input: news articles presenting different descriptions of the same event. Repeated information is a good indicator of importance, and can be used for summary generation. Our approach is unique in its integration of machine learning and statistical techniques to identify similar paragraphs, intersection of similar phrases within paragraphs, and language generation to reformulate the wording of the summary.

The analysis, or similarity computation component, takes as input a set of articles. It breaks the article into sentence-sized units for comparison, and then extracts a set of linguistic and positional features, which serve as input into the similarity algorithm. These features include primitive features such as word, stem and WordNet overlap as well as composite features, which aim to capture matches on the syntactic level such as subject-verb and verb-object relations. We use a log-linear regression model to convert the evidence from the various features to a single similarity value. The model was trained on a set of 10,535 pairs of paragraphs which were manually marked for similarity. The output of the model is a listing of real-value similarity values on sentence pairs. These similarity values are fed into a non-hierarchical clustering algorithm [?], producing clusters of closely related sentences that we term *themes*.

Once similar paragraphs are identified, they are passed to the generation component which further identifies and selects information to be reformulated as coherent text. The generation component consists of an ordering component, an intersection component, and a sentence generator. The goal of the ordering component is to order themes into coherent text which respects the chronological order of the main events. To implement this strategy, we identify blocks of themes which talk about the same event and apply chronological ordering on blocks of themes using information about publication date.

Once the themes are ordered, the intersection component identifies similar phrases across sentences of each theme by computing predicate-argument structure for each sentence and comparing arguments. The sentence generator then determines which phrases should be combined into a single, more complex sentence and the resulting constituent structure is used to generate an English sentence for each theme with a non-empty intersection. In order to produce summaries of consistent length, MultiGen ranks the themes on the basis of size, similarity score and significance. The first two of these scores are produced by the similarity component, while the significance score of the theme is computed using *lexical chains* [?]. Lexical chains are sequences of semantically related words; a theme with many sentences linked by lexical chains is given a higher significance score for the multi-document summary.

2.5 DEMS

Since Multigen operates over clusters of highly-related documents, we use a different approach for input articles which are not as related. The DEMS summarizer specifically attempts to extract sentences that contain the most salient new information from each of the input documents. It uses four categories of features to determine which sentences to extract: word importance, semantic classes, location, and style. It also measures prominence of one NP over others for biographical summaries.

Word importance is measured in part by using a lexicon of key terms, i.e. nouns, verbs and adjectives that are more likely to appear in the first paragraph of a news article than in the entire article. DEMS also measures informativeness using a study of verbs that was done for the BioGen [?] system. Verbs associated with a large number of subjects are unlikely to express important content by themselves and sentences with these words are weighted less. For example, the verb "arrest" is strongly associated with the subject "police," but not with a large number of other nouns. Thus, the verb "arrest" conveys some contextual information that a verb like "happen" would not. Semantic features measure importance within a cluster of articles. But instead of word based metrics, we count semantic classes of words identifying the class for each word based on WordNet synonyms, hypernyms, and hyponyms, excluding words with more than five senses. One feature in this group is based on frequency within the entire cluster and the other is based on frequency within each document.

Location features include the date of publication (recent documents more important) and the location of the sentence within the article (beginning sentences more important). Syntax and style features are based on the presence and location of pronouns and the length of the sentence. To avoid both cryptic, short sentences and extraneous information in long sentences, we set the length of an ideal sentence at 20 words. The presence of pronouns is also weighted negatively, to avoid dangling references.

The last group of features targets the biographical document sets, which were those that covered a sequence of events surrounding one individual. Such clusters have a stronger focus than many of the general clusters and the central characters are good candidates for inclusion in summaries. The main feature here is a binary value reflecting whether or not the target individual is found in the sentence, and a related feature of whether or not another individual is found in the sentence.

3. ADDING RELATED IMAGES

XXX selects and displays thumbnails of images that are related to an event on the same page where the summary of events is displayed. During the web crawling phase, in addition to looking for news articles, XXX also looks for embedded images in the articles. Since articles are taken from many different sources on the web, and these sources might change over time, it is important that the rules used by the system to find such images are general, and can be applied to multiple sites. The rules must extract most of the appropriate images without also taking advertisements and other inappropriate images. We weighted precision higher than recall, since users will not notice if certain images are not found, but inappropriate images would be visible. Some patterns we noticed when manually examining news sites were: (1) Images are almost always in the same cell as the article or in an embedded cell. (2) Images that are jpeg's tend to be appropriate, but other formats are more likely advertisements or link related. (3) Images with a word like "ad" or "advertisement" in the URL are probably not appropriate. By combining such rules, our system seems to achieve nearly perfect precision while still recalling the high majority of appropriate images.

4. EVALUATION

Our evaluation focuses on two questions: the usability of XXX as a tool for browsing news and the performance of the individual system components.

4.1 Experimental Design

The evaluation was performed through a mirror website identical to the continuously running XXX system, but with links to our evaluation materials (described below). The evaluation website was advertised to our colleagues in the Department of Computer Science as well as to a more general lay audience. All the transactions in this website were recorded, and user identities were tracked through a cookie mechanism, which allowed us to determine unique IDs for each computer system that accessed the site. The participants of the experiment were asked to check the news with XXX for a week, and to answer questions about the system and summaries several times during the week.

4.2 Methods

We used three different techniques to collect the evaluation data: a preference architecture experiment, quantitative analysis of user answers to our questionnaire and qualitative analysis of user comments about the system. In the preference architecture experiment, users were able to choose among different functionalities of the system, configuring it to their needs. A questionnaire was used to evaluate the system as a whole, as well as specific system components; we also included questions about the background of the evaluators. A comment box allowed evaluators to comment on all aspects of the system.

4.2.1 Subjects

During the week of the experiment, 94 subjects accessed the system with 2.87 average number of accesses per person. 48 users accessed the system only once and 46 users regularly used the system (number of accesses between 2 and 41, average number of accesses 4.8). During the first interaction with the system, users were presented with questions about their information needs. We asked how frequent an evaluator checks news and which media he uses to obtain news from. [1] Almost all our evaluators regularly check the news: 17 users check news daily, 7 read news several times a week and only 2 evaluators check the news rarely. The majority of the users (16) obtain their news mostly from on-line sources, the rest access news through print (8) and broadcast media (2). This data shows that the majority of the respondents fit the profile of target audience of our system — users who regularly follow the news through on-line sources. All of our evaluations were performed over judgments from humans different from the authors.

4.3 Usability Evaluation

The first issue we address in the evaluation is whether our tool is useful for the task it was designed for – efficient news access. More specifically, we wanted to examine whether the particular combination of system functionalities we have chosen for the publicly available XXX system is useful. Given the current state of information technology, it is an open question whether a user prefers advanced features or more simple, but robust ones. For example, the system can more reliably compute the keywords for a text, but the quality of a summary may vary from one input to another, although it should provide a better indication of the content of the text. Thus, to justify our selection of sophisticated features for our system (automatic categorization, grouping related

[1] Since these questions were optional, only 26 users answered them.

articles by topic, and natural-language summaries), we performed an experiment based on user preferences which allowed users to turn system features on or off, tailoring the system to their tastes. The users had the following options:

- Classification based on news type(US, worlds, Finance, Entertainment, Sport) vs. no classification.
- Clustering related events according to topic (e.g., military actions in Afghanistan) vs. a flat list of events.
- Summarization vs. a list of keywords.

Every user was initially assigned a random set of preferences with an equal probability over all configurations. The system prominently displays the option for chaning preferences on the front page and when selected, gives a brief description of the different options. When the user selects a new configuration, the system remembers the choice for future usage. However, the user can change his preferences on any session using a preference button; all the changes are recorded in the system log. From 46 users participating in this experiment, 35 users preferred summaries over keywords, 40 users preferred categories, and topical grouping was used by 31 users. (Our initial random settings for the 46 users had 24 users starting with the natural language summary option on, 29 users with categorization, and 17 users with topical grouping). These results provide strong evidence for user preference of advanced features of the system. In particular, there were eight possible system configurations; while only 10% started out in our "optimal" configuration, people who visited the site more than once chose this exact configuration 56twice 57%, more than three times 57%, and 63% for people with more than four visits.

We also evaluated user satisfaction through a questionnaire. The questions, answer scale and answer distributions are shown in Table 4.4.0.1. We asked two questions regarding the usage of the system: whether the system is useful for keeping up with current events, and whether the system alone is able to satisfy the users' information needs. Only two out of 27 users did not find the system useful for keeping up with current events. However, only seven users found that the system alone can satisfy their information needs, while the majority of users (16) found that the system can somewhat satisfy their needs. Several users commented on this question saying "no single news source would ever be enough". Two other questions focused on how well the system aided people in those tasks. 54% of the users found that the system provided them with a broader perspective on the news, and 80% of the users found that the system saved them time. This data clearly indicates that users found our system as an useful tool for news access.

4.4 Evaluation Per Component

We evaluated the quality of the main system functionalities – content organization and summarization. This evaluation was performed through a questionnaire which users accessed from a link on the main page. The users were asked to perform the evaluation after browsing the site. The questionnaire page opened in a new window, which had links referring the user to the specific page over which the questions were asked (either a link to the main index page, or a particular summary page.)

4.4.0.1 Summary evaluation

XXX generates 34.7 summaries per day on average; two summaries among them were randomly selected for evaluation every day. To ensure that the user carefully examined the summary, he was asked to evaluate one summary per day; on average, each summary was evaluated by 7.5 judges. Overall, we collected 61 judgments for 14 summaries during the week of the evaluation. Each summary in our test set summarized 25.8 articles on average. A summary was evaluated along three dimensions: content, organization and clarity. The results are shown in the first three questions of Table **??**. The majority of the summaries got the highest rating in all three categories, and only very few summaries received low marks. Interestingly, there is a high correlation between summary grading in all three categories — if a user rated the summary content high, he frequently rated the readability and organization of the summary high. We hypothesize that users were unable to look beyond their overall perception of the summary to give a more detailed breakdown of the different qualities of the summary.

4.4.0.2 Evaluation of Content Organization

Three components of the system contribute to the quality of information organization in our system: categorization, clustering, and keyword titling of the clusters of related articles. We decided to incorporate evaluation of categorization and clustering into one question because we thought lay users would be confused by references to the different types of grouping. This turned out to be problematic since we couldn't separate comments about categorization from comments about grouping of events. According to the data presented in Table **??**(the last three questions), the vast majority of evaluators found that XXX is easy to navigate, and the topical grouping of articles is good. The users were not as satisfied with our keyword labelling strategy as with other functions of the system. This suggests that we should explore alternative methods for generating cluster titles.

4.5 Comments

Analysis of comments provided us with interesting insights on user opinions about overall functionality of the system as well as performance of individual components. Overall, we collected 32 comments, which can be grouped in four main categories: comments on user satisfaction with the system (9), suggestions for system improvement(9), comments on the presentation style(9) and explanations on the ratings users gave to the system(9). For each category, we first list a representative comment and then briefly discuss comments in that group.

4.5.0.3 Comments on user satisfaction with the system

"XXX seems to be good at finding stories that get a lot of coverage. XXX is useful to me for condensing this coverage, and providing a summary."

In this set of comments, users listed the features of the system that they found helpful; these included wide coverage of events, alternative descriptions of a particular event and efficient organization of information through grouping and

Question	Answer Scale	Answer Distribution
Would XXX alone satisfy your needs for news?	no	3
	somewhat	16
	yes	7
Is XXX a useful vehicle for keeping up with events?	no	2
	somewhat	14
	yes	11
Did XXX save you time?	no	5
	yes	20
Did XXX provide a broader perspective?	no	12
	yes	14
Is the XXX site easy to navigate?	no	3
	mostly	12
	yes	43

Figure 2: User satisfaction questionnaire and answer distributions.

Question	Answer Scale	Answer Distribution
Does the summary content give you a good idea of what the articles are about?	no	9
	mostly	21
	yes	30
How would you describe the organization (e.g., order of information, flow of information) of the summaries?	poor	7
	adequate	22
	good	31
How would you describe the clarity of the summaries?	poor	5
	adequate	21
	good	34
Is the topical grouping of articles done well?	not	4
	mostly	17
	yes	37
Do the general topic headings (keywords) provide a good description of the articles?	no	7
	mostly	31
	yes	20

Figure 3: User satisfaction questionnaire and answer distributions.

summarization. Users stated that they will use the system to get an overview of the daily news, or as a source of detailed information for a specific event they are interested in.

4.5.0.4 Suggestions for the system improvement

The one big problem, I think, is the fact that this site isn't quick enough to keep us with breaking news.
The most frequent complaint was that the system is not updated frequently enough [2]. The users wanted to include breaking news as soon as it is reported by main news providers, such as CNN. Users also suggested that it would be useful to increase the coverage of the system by including translations of the foreign newspapers.

4.5.0.5 Comments on the presentation style

There should be a "return to main page" button on each subpage. There should be double space between paragraphs in the articles; as it is, it is difficult to read.
The pictures at top of summaries are distorted. Visually poorly designed. Color of logo is unsophisticated... it doesn't look at all elegant.
Many presentation related comments dealt with very specific suggestions on the interface, such as changes in background colors and frame configuration, addition of new navigation bars and proposals on the formatting of articles. Several users mentioned that they liked pictures next to the summaries.

4.5.0.6 Explanation for the ratings users gave to the system

Argentina's Peso Is Freed to Float, And Quickly Sinks (8 articles from 01/12/2002 — 01/15/2000) should be under World category, not US.
The summary for "shooting in NYC school" is kind of repetitive.
Comments in these categories included user justification for the grades they gave to a specific summary or analysis of mistakes in categorization and clustering. These comments were particularly useful to us. For example, we discovered that the majority of categorization mistakes happen between two categories – US and world news. Users also pointed to

[2] XXX is currently updated once a day

summaries which contain repetitions or have ordering problems.

5. RELATED WORK

Research presented in this paper relates to three areas of information access: multidocument summarization, event tracking, and categorization. While there have been major advances in each of these areas (see DUC [?] and TDT [?] proceedings), very little work has been done on integrating these three technologies into one system, a step which can improve user access to information in a real-world setting. A recent effort in this direction was undertaken by [?] who developed a system for temporal summarization operating over output of a TDT system. However, their summarization system runs over the TDT-2 data set, and has not been integrated with an on-line news tracking system.

The only other system (that we are aware of) running in a real-world environment [?] focuses on dynamically augmenting information on a particular news story provided by a user. The system accepts a URL from the user, at run-time finds similar stories by querying news sites, and provides a sentence-extraction based summary for the article set. The two systems satisfy different information needs: one is useful for users who want to explore a specific topic known in advance, while our system is useful for efficiently browsing the news.

6. REFERENCES

[1] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, August 1997. Association for Computational Linguistics.

[2] R. Barzilay, N. Elhadad, and K. R. McKeown. Sentence ordering in multidocument summarization. In *Proceedings of the 1st Human Language Technology Conference*, San Diego, California, 2001.

[3] R. Barzilay, K. R. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 550–557, College Park, Maryland, June 1999. Association for Computational Linguistics.

[4] D. M. Bates and D. G. Watts. *Nonlinear Regression Analysis and its Applications.* Wiley, New York, 1988.

[5] Document understanding conference, 2001.

[6] V. Hatzivassiloglou, L. Gravano, and A. Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-00)*, pages 224–231, Athens, Greece, July 2000.

[7] V. Hatzivassiloglou, J. L. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, College Park, Maryland, June 1999. Association for Computational Linguistics.

[8] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M.-Y. Kan, and K. R. McKeown. SIMFINDER: A flexible clustering tool for summarization. In *NAACL Workshop on Automatic Summarization*, pages 41–49. Association for Computational Linguistics, 2001.

[9] K. McKeown, R. Barzilay, D. K. Evans, V. Hatzivassiloglou, M.-Y. Kan, B. Schiffman, and S. Teufel. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of the First Document Understanding Conference*, pages 43–63, 2001.

[10] K. R. McKeown, J. L. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-99)*, pages 453–460, Orlando, Florida, July 1999.

[11] W. Nina, Y. Ravin, and M. Choi. Disambiguation of names in text, 1997.

[12] C. Sable and K. W. Church. Using bins to empirically estimate term weights for text categorization. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2001.

[13] T. J. Santner and D. E. Duffy. *The Statistical Analysis of Discrete Data.* Springer-Verlag, New York, 1989.

[14] B. Schiffman, I. Mani, and K. J. Concepcion. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 2001.

[15] H. Späth. *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples.* Ellis Horwood, Chichester, West Sussex, England, 1985.

[16] N. Wacholder. Simplex NPs clustered by head: A method for identifying significant topics within a document. In F. Busa, I. Mani, and P. Saint-Dizier, editors, *The Computational Treatment of Nominals*, pages 70–79, 75 Paterson Street, Suite 9 New Brunswick, NJ 08901 USA, August 1998. Coling-ACL, Association of Computational Lingustics.

[17] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR-99)*, pages 42–49, Berkeley, California, 1999.

U.S.

News on Enron, Arthur Andersen, President Bush, White House, Kenneth Lay

- Congressional Dems Call for Bush Officials to Recuse Themselves From Enron Probe (60 articles from 01/14/2002 - 01/17/2002)

News on Richard Reid, United States, Qaeda, Richard Colvin Reid, Boston

- 'Brother Abdul Ra'uff' (11 articles from 01/16/2002 - 01/17/2002)

News on Peter Odighizuwa, L. Anthony Sutin, law school, Jack Briggs, Grundy

- Student kills 3 during shooting spree at law school (10 articles from 01/16/2002 - 01/17/2002)

News on Bud Selig, John Henry, Boston Red Sox, Expos, Jeffrey Loria

- Owners OK Sale of Red Sox; Sales of Marlins, Expos Likely Next (9 articles from 01/15/2002 - 01/17/2002)

World

News on United States, Afghanistan, Qaeda, Taliban, Laden

- U.N.'s Robinson: Cuba Detainees Are Prisoners of War (13 articles from 01/14/2002 - 01/17/2002)
- U.S. to stop 'chasing shadows' of bin Laden (10 articles from 01/14/2002 - 01/16/2002)
- Man claiming to be Taliban elder surrenders, U.S. official says (6 articles from 01/14/2002 - 01/17/2002)

News on Pakistan, India, Colin Powell, United States, Afghanistan

- India Gears for Long Haul On Border With Pakistan (16 articles from 01/14/2002 - 01/17/2002)
- U.S. Hunts Bin Laden, Defends Rights Record in War (6 articles from 01/16/2002 - 01/17/2002)

Figure 4: XXX Main Page

Congressional Dems Call for Bush Officials to Recuse Themselves From Enron Probe

Summary:

Enron declared bankruptcy on Dec. 2 after investors lost confidence in the company, following disclosures that it had vastly overstated its earnings for three years. Congressional investigators sifting through Enron documents on Wednesday came across an internal Arthur Andersen memo showing that Sherron Watkins, an Enron vice president who warned last August of the firm's impending financial problems, had taken her concerns directly to Andersen.

Source Articles:

- Enron auditor fired by Arthur Andersen (Canadian Broadcasting Corporation 01/17/02)
- Enron Blame Game (CBS News 01/17/02)
- House Grills Fired Enron Auditor (FOX News 01/17/02)
- Andersen fires lead auditor in wake of Enron failure (USA Today 01/17/02)
- Congressional Cooperation (ABCNews 01/17/02)
- Andersen Alerted (ABCNews 01/17/02)
- Lead Auditor Fired (ABCNews 01/17/02)

Figure 5: XXX Summary Page