Describing Story Evolution from Dynamic Information Streams

Stuart Rose[*] Scott Butner[†] Wendy Cowley[‡] Michelle Gregory[§] Julia Walker[¶]

Pacific Northwest National Laboratory

Abstract

Sources of streaming information, such as news syndicates, publish information continuously. Information portals and news aggregators list the latest information from around the world enabling information consumers to easily identify events in the past 24 hours. The volume and velocity of these streams causes information from prior days to quickly vanish despite its utility in providing an informative context for interpreting new information. Few capabilities exist to support an individual attempting to identify or understand trends and changes from streaming information over time. The burden of retaining prior information and integrating with the new is left to the skills, determination, and discipline of each individual. In this paper we present a visual analytics system for linking essential content from information streams over time into dynamic stories that develop and change over multiple days. We describe particular challenges to the analysis of streaming information and present a fundamental visual representation for showing story change and evolution over time.

Index Terms: H.3.1 [Content Analysis and Indexing]: Abstracting methods—; H.3.3 [Information Search and Retrieval]: Information Filtering—; H.3.7 [Digital Libraries]: User Issues—; I.3.3 [Computer Graphics]: Picture/Image Generation—

1 Introduction

Most information consumers interact with snapshots of an information space that is continually changing. Sources of information streams, such as news syndicates and information services, provide a variety of mechanisms for feeding the latest information by region, subject, or by user-defined search interests. However new information that arrives eclipses prior information. Although there are many interfaces for conveying the latest information, few if any capabilities provide a temporal context longer than 24 hours. Accurately identifying and intelligently describing change in an information space requires a context that relates *new* information with *old*.

Fundamental tasks, such as tracking down initial precursors to an ongoing story can be problematic, particularly when stories span multiple documents and time intervals. Stories develop, merge, and split as they intersect and overlap with other stories over time. With a search interface to archived information, a user may formulate an effective query that recalls all documents related to a given story. However, identifying and exploring potential story connections using an archive search depends on each users prior knowledge of a related story and ability to formulate effective queries. Because it is human nature to prioritize and retain information that supports specific preferences, theories, or hypotheses, relying on individual abilities to acquire and maintain full coverage of a story's related context may actually diminish the objectivity and comprehensiveness of the coverage. We therefore consider the research presented here to provide a more effective means of discovering and maintaining coverage of information for a given analytic task as stories develop.

This paper presents three contributions to the field of visual analytics that together enable computation and presentation of a lengthy temporal context in which new information can be situated with thematically related previous events. The first is a novel text analytic framework that extracts essential content from individual documents and computes significant themes occurring within documents collected into corpora or published within information streams. Applied to information streams, the analytic framework computes themes within time intervals enabling accurate detection of stories and their temporal dynamics. The second is a data model for describing and archiving the temporal context of significant themes in a manner that readily adapts to story changes and enables tracking of story evolution and dynamics. The third is a fundamental visual representation of story flow over time that conveys the temporal dynamics for all stories in an information space, enabling observation of story evolution, interaction, and flow over time.

We consider the visual representation presented to be fundamental in nature, representing a basis for conveying the historic context and precursors to developing and ongoing stories. The data model presented exists independently of any visual representation and was developed to accomodate the dynamic nature of real stories. Stories are identifiable only as repeated joint membership of documents in the same sets. If documents do not repeatedly join up in the same set then they are not part of the same story.

2 Related Work

Our work relates to research on visual representations and analysis of structure and change in flows of information over time. Of particular interest is prior work related to temporal analysis of information which we explore in the next two sections.

2.1 Visual Methods

In his book, Envisioning Information, Tufte [15] describes a graphic time series produced in 1977 by Reebee Garofalo that tracks artistic genealogy and production among recording artists from 1955 to 1974. In Garofalo's graphic, names of recording artists are drawn on a two-dimensional cumulative trend plot, the horizontal axis representing time, the vertical axis generally representing market share. Artists are grouped by style of music and over time new artists are placed near those of similar styles. Tufte writes:

> *The multiple, parallel flows locate music-makers in two dimensions linking musical parents and offspring, and listing contemporaries for each year. The illustration presents a somewhat divergent perspective on popular music: songs are not merely singles rather, music and music-makers share a pattern, a context, a history.*

Bearing some similarity to Garofalo's graphic, ThemeRiver [7, 6] provides a visual representation of the temporal rise and fall

[*] e-mail: stuart.rose@pnl.gov
[†] e-mail: scott.butner@pnl.gov
[‡] e-mail: wendy.cowley@pnl.gov
[§] e-mail: michelle.gregory@pnl.gov
[¶] e-mail: hyunjoo.walker@pnl.gov

IEEE Symposium on Visual Analytics Science and Technology
October 12 - 13, Atlantic City, New Jersey, USA

of major themes in a document corpus. While ThemeRiver automatically detects themes and their trends in order to show their temporal characteristics, individual items that constitute those trends are not visible, in contrast to the detailed labeling of artists in Garofalo's graphic. As a result ThemeRiver has a much lower information density. Although direct relationships between themes may exist in the corpus, they are not evident in ThemeRiver. Observation of the ThemeRiver visualization does not provide insight into how or to what extent themes interrelate. While correlation among themes may be inferred through observation of similar temporal patterns in ThemeRiver, direct influence between elements is not indicated.

Representing the history of an individual document, HistoryFlow [17] presents a visualization of revision patterns made over the life of a document. The visualization mirrors the linear structure of the document, graphic representations of user revisions are shown in their location relative to the entire document and in the larger context of the document's history. The authors note the aid that the history flow visualization provides in understanding patterns of revisions and collaboration over time. Although focused on a single document, history flow effectively shows progression of all parts of a document in relation to each other.

Themail [16] comprises a system for visualizing and exploring relationships based on interaction patterns as revealed through an individual's email archive. The authors note two distinct user interaction modes with the visualization, exploration of big picture trends and themes (haystack) and more detail-oriented exploration (needle). Based on user studies, the authors note that displaying large collections of keywords fostered user's insight into the evolution of relationships over time as users' past email exchanges were placed within a meaningful context. They also note that each users' familiarity with their respective email archives is essential in enabling efficient insight from the visualization. Single keywords in the visualization bring back memories of events and associated context. A user unfamiliar with the data would likely need to dig deeper into email content in order to resolve ambiguity associated with single keywords and to understand the context for particular keywords, relationships, and trends.

A framework for tracking the popularity of quoted phrases (memes) through on-line text is described in [11]. The authors describe a framework for meme-tracking and demonstrate its utility in representing the news cycle and unique temporal patterns for news media and blogs. A stacked plot is presented for displaying the popularity of particular memes over time.

A system for visualizing tags is described in [3]. The system characterizes the most interesting tags from an online image sharing community based on a sliding interval of time and provides an interactive visualization that allows the user to observe the temporal interestingness of tags as they evolve, at arbitrary scales in real time. Information about inter-relatedness of tags is not explored as the authors' focus is to show the latest, most interesting tags.

A common characteristic across this related research is a focus on providing overview and generalization to a user through computational analysis and reduction of unstructured information to a set of specific features whose characteristics are conveyed in a high-level overview. Although HistoryFlow differs by focusing on a structured set of information, and representing changes to that structure, it effectively provides a high-level overview in which patterns of activity and dynamic changes are easily observed.

2.2 Analytic Methods

The analysis of unstructured information in order to select a set of representative features is critical, as features that are not detected can not be shown or measured. Previous research in the visual analytics and information retrieval community has focused on methods of selecting features through the application of corpus-oriented methods that rate candidate features according to their ability to statistically discriminate between sets of documents within the corpus. [13] and [9] describe positive results of selecting statistically discriminating words across a corpus to define an index vocabulary for the corpus.

While these approaches are useful in many applications, features that are significant for individual documents may be lost by such corpus-oriented methods. In the context of time, stories that span a week may be overshadowed by major news trends that span a year. Because new features are likely to be less common than established features, corpus-oriented methods are likely to overlook features as they first appear.

A set of features based on corpus-oriented methods must be reevaluated when the corpus changes. Because corpora based on information streams change frequently, this incurs significant performance hurdles with real information streams. Reprocessing the entirety of the corpus may also wipe out mental models that a user has developed to understand, explore, and explain an information space.

Several of these challenges to analysis of streaming data were addressed in [8]. The authors describe a visual analytics system that enables analysis of dynamic corpora. The system provides multiple visual analytic tools for investigating the current state of a document corpus continually aging out old increments and adding new increments for analysis. The system avoids scalability issues by maintaining a sliding window over time, removing old documents as time moves onward. As this limits how far back in time a user can look, a feature is provided that allows a user to save a copy of any current increment of an analysis. A limitation of that system is that it only represents a single state, and provides no facilities for tracking changes across multiple time increments.

The set of features for any increment are selected through corpus-oriented methods, relying on overlap across and gradual change between increments to transition the user across increments. As a result, if many increments pass between a user's last interactions with the system, they may spend significant time reorienting to the new information and any change in context.

To avoid the limitations of corpus-oriented methods of feature selection, we focus our interest on methods of keyword extraction that operate on individual documents. Such document-oriented methods extract the same keywords from a document regardless of the current state of a corpus. Document-oriented methods therefore provide context-independent document features and are suited to corpora that change

As commonly defined in the recent literature, keywords may be single terms or comprise multiple words and phrases that reflect a document's content. Keywords provide a compact summary of a document and have been applied to improve many information retrieval (IR) systems, such as digital libraries, by linking documents via keywords, hyperlinks, citations, quotations, and key phrases.

Themail [16] provides an example of such a system as it enables exploration of email archives through keywords due to their ability to trigger user associations to a larger context. Another system, Phrasier [10], lists documents related to a primary document's keyphrases, and supports the use of keyphrase anchors as hyperlinks between documents, enabling a user to quickly access related material. Keyphind [4] uses keyphrases from documents as the basic building block for an IR system. As described in [14], key ideas may also be extracted and used to augment a system for browsing digital libraries of books using key terms extracted from popularly quoted passages.

Linking documents through essential content such as keywords enables exploration of associations between documents independent of a static and pre-defined corpus.

3 Analytic Framework

Within this section we first define our method of feature selection, then we describe our method for computation and analysis of significant themes (CAST) and show application to a static document corpus for clarity, then describe how the analytic framework applies CAST to an information stream or dynamic corpus.

3.1 Feature Selection

Many text analysis techniques focus on what distinguishes a text from its encompassing document corpus. When the information is streaming, the corpus is dynamic and can change significantly over time. Techniques that evaluate documents by discriminating features are only valid for a snapshot in time.

In order to eliminate the influence of a transient context or state in the larger information space we apply computational methods for characterizing each document individually. Such methods produce information on what a document is about, independent of its current context. Analyzing documents individually also further enables analysis of massive information streams as multiple documents can be analyzed in parallel or across a distributed architecture.

In order to extract content that is readily identifiable by users, we apply the technology Rapid Automatic Keyword Extraction (RAKE) described in [12] to extract keywords from individual documents. RAKE takes a simple set of input parameters to extract keywords from a single document. Figure 1 shows extracted keywords as keywords of a news article from Voice of America.

We consider keywords to include one or more words such as *Pakistan Muslim League-N leader Nawaz Sharif* and *criticized President Pervez Musharraf*.

Keywords provide an advantage over other types of signatures as they are readily accessible to a user and can be easily applied to search other information spaces. The value of any particular keyword can be readily evaluated by a user for their particular interests and applied in multiple contexts. Furthermore, the direct correspondence of extracted keywords with the document text improves the accessibility of a user with the system.

3.2 Computation and Analysis of Significant Themes

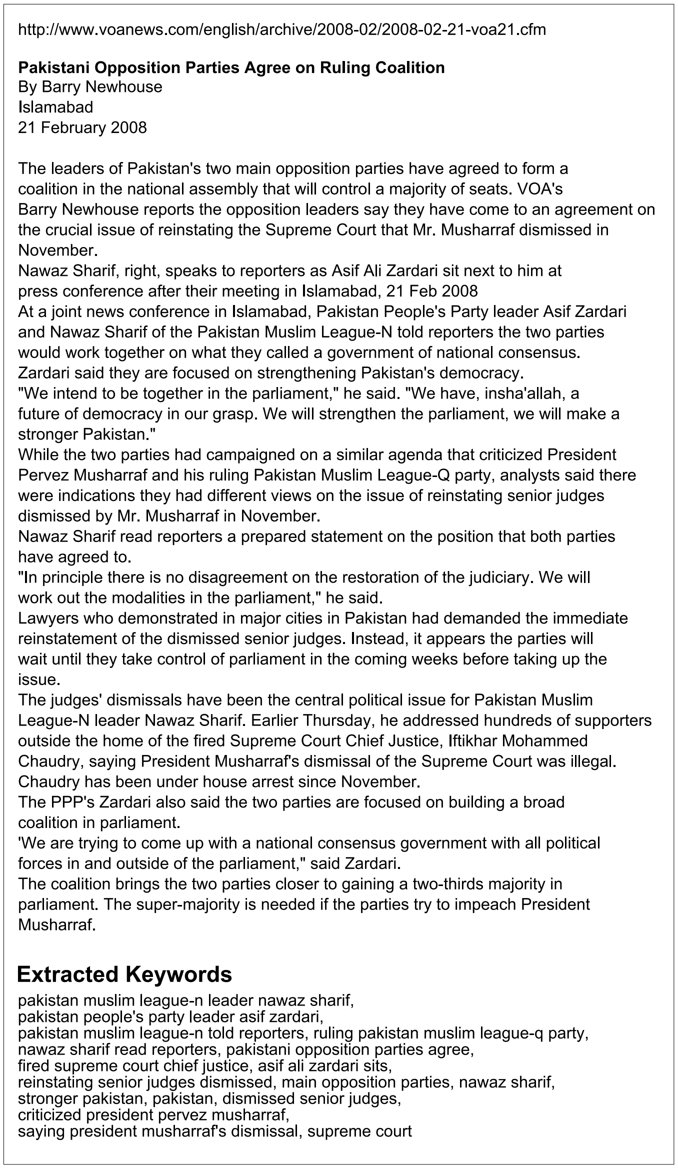
For a given corpus, whether static or representing documents within an interval of time, the extracted keywords are grouped into coherent themes by applying a hierarchical agglomerative clustering algorithm to a keyword similarity matrix based on keyword document associations in the corpus.

The association of each keyword to documents within the corpus is measured as the document's response to the keyword, which is obtained by submitting each keyword as a query to a Lucene index populated with documents from the corpus. The query response of each document hit greater than 0.1 is accumulated in the keyword's document response vector. Lucene calculates document similarity according to the vector space model described in [13]. We refer the reader to [5] for more complete implementation details. In most cases the number of document hits to a particular keyword query is a small subset of the total number of documents in the index. Keyword document response vectors typically have fewer entries than there are documents in the corpus and are very heterogenous.

The similarity between each unique pair of keywords is calculated as the Sorensen similarity coefficient of the keywords' respective document response vectors. The Sorensen similarity coefficient [13] is used due to its effectiveness on heterogeneous vectors and is identical to 1.0 - Bray-Curtis distance [14], shown in equation (1).

$$BC_{ab} = \frac{\sum |a_i - b_i|}{\sum (a_i + b_i)} \qquad (1)$$

Coherent groups of keywords can then be calculated by clustering keywords by their similarity. Because the number of coherent groups may be independent of the number of keywords extracted, we apply Ward's hierarchical agglomerative clustering algorithm [18] which does not require a pre-defined number of clusters.



http://www.voanews.com/english/archive/2008-02/2008-02-21-voa21.cfm

Pakistani Opposition Parties Agree on Ruling Coalition
By Barry Newhouse
Islamabad
21 February 2008

The leaders of Pakistan's two main opposition parties have agreed to form a
coalition in the national assembly that will control a majority of seats. VOA's
Barry Newhouse reports the opposition leaders say they have come to an agreement on
the crucial issue of reinstating the Supreme Court that Mr. Musharraf dismissed in
November.
Nawaz Sharif, right, speaks to reporters as Asif Ali Zardari sit next to him at
press conference after their meeting in Islamabad, 21 Feb 2008
At a joint news conference in Islamabad, Pakistan People's Party leader Asif Zardari
and Nawaz Sharif of the Pakistan Muslim League-N told reporters the two parties
would work together on what they called a government of national consensus.
Zardari said they are focused on strengthening Pakistan's democracy.
"We intend to be together in the parliament," he said. "We have, insha'allah, a
future of democracy in our grasp. We will strengthen the parliament, we will make a
stronger Pakistan."
While the two parties had campaigned on a similar agenda that criticized President
Pervez Musharraf and his ruling Pakistan Muslim League-Q party, analysts said there
were indications they had different views on the issue of reinstating senior judges
dismissed by Mr. Musharraf in November.
Nawaz Sharif read reporters a prepared statement on the position that both parties
have agreed to.
"In principle there is no disagreement on the restoration of the judiciary. We will
work out the modalities in the parliament," he said.
Lawyers who demonstrated in major cities in Pakistan had demanded the immediate
reinstatement of the dismissed senior judges. Instead, it appears the parties will
wait until they take control of parliament in the coming weeks before taking up the
issue.
The judges' dismissals have been the central political issue for Pakistan Muslim
League-N leader Nawaz Sharif. Earlier Thursday, he addressed hundreds of supporters
outside the home of the fired Supreme Court Chief Justice, Iftikhar Mohammed
Chaudry, saying President Musharraf's dismissal of the Supreme Court was illegal.
Chaudry has been under house arrest since November.
The PPP's Zardari also said the two parties are focused on building a broad
coalition in parliament.
'We are trying to come up with a national consensus government with all political
forces in and outside of the parliament," said Zardari.
The coalition brings the two parties closer to gaining a two-thirds majority in
parliament. The super-majority is needed if the parties try to impeach President
Musharraf.

Extracted Keywords

pakistan muslim league-n leader nawaz sharif,
pakistan people's party leader asif zardari,
pakistan muslim league-n told reporters, ruling pakistan muslim league-q party,
nawaz sharif read reporters, pakistani opposition parties agree,
fired supreme court chief justice, asif ali zardari sits,
reinstating senior judges dismissed, main opposition parties, nawaz sharif,
stronger pakistan, pakistan, dismissed senior judges,
criticized president pervez musharraf,
saying president musharraf's dismissal, supreme court

Figure 1: Sample document from Voice of America and its keywords extracted by RAKE.

Ward's hierarchical clustering begins by assigning each element to its own cluster and then successively joins the two most similar clusters into a new, higher-level, cluster until a single top level cluster is created from the two remaining, least similar, ones. The decision distance dd_{ij} between these last two clusters is typically retained as the maximum decision distance $dd_{max}$for the hierarchy and can be used to evaluate the coherence cc_n of lower level clusters in the hierarchy as shown in equation (2).

$$cc_n = 1 - \frac{dd_n}{dd_{max}} \qquad (2)$$

Clusters that have greater internal similarity will have higher coherence. Using a high coherence threshold prevents CAST from selecting clusters that include broadly used keywords such as *president* that are likely to appear in multiple unrelated stories. Clusters

Table 1: MPQA Topics and CAST Themes. Each row contains the defined topic and description within the Defined Topics column and the labeled theme and top keywords within the CAST Themes column.

Defined Topics	CAST Themes
mugabe 2002 presidental election in Zimbabwe	*mugabe's re-election* (46) zimbabwe's election, mugabe, mugabe's, president mugabe, mugabe's government
guantanamo U.S. holding prisoners in Guantanamo Bay	*guantanamo prisoners* (43) detainees as prisoners, detainees prisoners, prisoners of war, war prisoners, prisoners
kyoto ratification of Kyoto Protocol	*kyoto protocol on climate* (40) kyoto protocol, ratification of the kyoto protocol, 1997 kyoto protocol
venezuela presidential coup in Venezuela	*venezuela's president hugo chavez* (37) venezuelan president hugo chavez, president hugo chavez, president chavez, hugo chavez
settlements Israeli settlements in Gaza and West Bank	*israeli* (34) palestinian, israel, occupied palestinian, israeli occupation, palestinian people's
taiwan relations between Taiwan and China	*china under which taiwan* (31) taiwan, united states and taiwan, taiwan issue, taiwan policy, taiwan affairs, taiwan strait
axisofevil reaction to President Bush's 2002 State of the Union Address	*axis of evil* (28) north korea as axis, iran and north korea, iran or north korea, iraq and north korea,
humanrights reaction to U.S. State Department report on human rights	*human rights* (28) human rights report, annual human rights report, human rights violations, rights
spacestation space missions of various countries	*space station* (28) international space station, space, international space, space shuttle, russian space mission
argentina economic collapse in Argentina	*argentina* (12) argentina's, help argentina, argentine government, argentine, worried argentina

with a coherence threshold of 0.65 or greater are selected as candidate themes for the corpus.

Each candidate theme comprises keywords that typically return the same set of documents when applied as a query. These keywords occur in multiple documents together and may intersect other stories singly or together.

We select the final set of themes for the corpus by assigning documents to their most highly associated theme. After all documents in the corpus have been assigned, we filter out any candidate themes for which no documents have been assigned. Keywords within each theme are then ranked by their associations to documents assigned within the theme. Hence the top ranked keyword for each theme best represents documents assigned to the theme and is used as the theme's label.

Before elaborating on how CAST is applied to multiple time intervals, we briefly show application of CAST to a static corpus. The following section presents results on application of RAKE and CAST to the Multi-Perspective Question Answering (MPQA) Corpus [1].

3.2.1 CAST Themes for the MPQA Corpus

The MPQA Corpus consists of 535 news articles provided by the Center for the Extraction and Summarization of Events and Opinions in Text (CERATOPS). Articles in the MPQA Corpus are from 187 different foreign and U.S. news sources and date from June 2001 to May 2002.

We applied RAKE to extract keywords from the title and text fields of documents in the MPQA Corpus and selected keywords from those extracted that occured in at least 2 documents. We then applied CAST to compute themes for the corpus. Of the 535 documents in the MPQA Corpus 327 were assigned to 10 themes which align with the 10 defined topics for the corpus as shown in Table 1. The number of documents that CAST assigned to each theme is shown in parentheses. As defined by CERATOPS:

> *The majority of the articles are on 10 different topics, but a number of additional articles were randomly selected (more or less) from a larger corpus of 270,000 documents.*

The majority of the remaining themes computed by CAST have fewer than four documents assigned, an expected result given the random selection of the remainder of documents in the MPQA Corpus.

3.3 Describing Stories

We developed our analytic framework to operate on streaming information; to extract essential content from documents as they are received and to calculate themes at defined time intervals. When the current time interval ends (one day in our evaluations), a set of keywords is selected from the extracted keywords and keyword document associations are measured for all documents published or received within the current and previous n intervals ($n = 7$ in our evaluations). Keywords are clustered into themes according to the similarity of their document associations, and each document occurring over the past n intervals is assigned to the theme for which it has the highest total response.

The set of themes computed for the current interval are persisted along with their member keywords and document assignments. Overlap with previous and future themes may be evaluated against previous or future intervals by comparing overlap of keywords and document assignments. Themes that overlap with others across time together relate to the same story.

Repeated co-occurrences of documents within themes across multiple intervals are meaningful as they indicate real similarity and relevance of content between those documents for those intervals.

In addition to the expected addition of new documents to an existing story and aging out of documents older than n intervals, it is not uncommon for stories to gain or lose documents to other stories. Documents assigned to the same theme within one interval may be assigned to different themes in the next interval. This feature of how we develop and apply the data model to define themes at each interval enables our analytic framework to automatically adapt to future thematic changes and accommodates the reality that stories often intersect, split, and merge.

In order to show its utility, we applied our analytic framework on documents within the TDT-2 corpus [2] tagged as originating from the Associated Press's (AP) *World Stream* program due to it's similarity to other news sources and information services of interest.

Table 2: CAST Themes for 01/12/1998 and their document counts

Assigned Docs		
< 01/12	01/12	Theme
1	5	chuan government
0	3	serena williams who is playing
0	3	men's match
0	2	stabilizing japan's shaken financial system
0	1	women's race
1	2	news agency
3	2	five-kilometer race
3	1	northern ireland
5	1	japanese prime minister ryutaro hashimoto
6	1	world cup
7	1	president suharto
9	3	world swimming championships
15	4	hong kong

Table 2 lists the CAST themes on January 12, 1998 for AP documents in the TDT-2 Corpus. The first column lists the count of documents assigned to each theme that were published before January 12. The second column lists each themes count of documents that were published on January 12. Comparing these counts across themes allows us to easily identify which stories are *new* for example *chuan government*, *serena williams who is playing*, *men's match*; which stories are the largest, for example *hong kong* and *world swimming championships*.

In the following section, we describe two fundamental visual representations that provide greater insight into the characteristics of themes and stories over time.

4 Visual Representation

The analytic framework we have described analyzes documents within time intervals; generating for each time interval a set of themes, each theme comprising a coherent set of keywords from that interval's documents, and having assigned documents from the previous *n* intervals. We describe two visual representations that yield insight to the temporal context of documents, themes, and stories for an information consumer.

The first view, a portion of which is shown in Table 3, represents the current time interval and its themes. The view presents each theme as a listing of its member documents in ascending order by date. This view has the advantage of simplicity. An observer can easily assess the magnitude of each theme, its duration, and documents that have been added each day. However, lacking from this view is the larger temporal context and information on how related themes have changed and evolved over previous days.

To provide a temporal context we developed the story flow visualization whose structure follows that of Garofalo's graphic mentioned previously and described in [15]. The story flow visualization, a portion of which is shown in Figure 2, shows for a set of time intervals, the themes computed for those intervals, and their assigned documents which may link themes over time into stories. The visualization places days across the horizontal axis and orders daily themes along the vertical axis by their assigned document count.

For a given interval, each theme is labeled with its top keyword in *italics* and lists its assigned documents in descending order by date. Each document is labeled with its title on the day that it is first published (or received), and rendered as a line connecting its positions across multiple days. This preserves space and reinforces the importance and time of each document, as the document title is only shown in one location. Similar lines across time intervals represent flows of documents assigned to the same themes, related to the same story. As stories grow over days, they add more lines. A document's line ends when it is no longer associated with any themes.

Table 3: Documents Assigned to CAST Theme *iraq* on 02/14/1998

Day	Document Title
02-07	U.N. to excavate weapons dumps in Iraq
02-07	Kuwaiti government to distribute gas masks in two days
02-07	U.N. to excavate dumps where chemical weapons and warheads buried
02-08	Saudi Arabia will not back military strike against Iraq
02-08	Egyptian Foreign Minister upbeat on diplomatic end to US-Iraq
02-08	Russian FM: Dispute over Iraq has not damaged U.S.-Russian
02-08	U.S. won't ask Saudis to allow attacks from their territory
02-08	Iraq says talks on weapons destruction making progress
02-09	Citing regional tensions, Annan cancels visit to the Middle East
02-09	Canada: No decision yet on commitment against Iraq
02-09	byline
02-09	control on Yeltsin-Annan remarks, other new material
02-09	Iraq launches campaign to rally Arab support
02-09	Iraq crisis at 'critical stage,' U.N. chief pushes for peaceful
02-10	Report: Lebanon tells Iraq to cooperate with weapons inspectors
02-10	Iraqi foreign minister seeks Syrian support in standoff
02-10	Precede MAJDAL SHAMS
02-10	Iraqi ambassador says Baghdad cannot meet Annan's proposed sales
02-10	Israel short on gas masks for children
02-11	Iraqi daily says Washington always determined to use force
02-11	Britain rejects latest Iraqi offer on weapons
02-11	Iraq puts oil export potential at 1.6 million barrels daily
02-11	Gulf foreign ministers meet to discuss unified stand on Iraq
02-11	BC-Iraq-Opposition
02-11	Santer says EU position will change if Iraq stays defiant
02-12	Iraqi foreign minister calls U.S. rejection 'a bluff'
02-12	UN advises vacationing staff to stay away from Iraq
02-13	American, Russian defense chiefs hold cordial talks at military
02-13	Iraq accuses U.S. of psychological warfare
02-13	U.S. envoy meets with Japanese leaders on Iraq
02-13	Foreign Office urges 'higher degree of caution' for travelers
02-13	Iraq dispute underlines strain in U.S.-Russia ties
02-14	U.S. envoy seeks China's support on Iraq
02-14	Demonstrators at U.S. Embassy in Japan urge no military attack
02-14	Iraq urges diplomacy, releases Egyptian prisoners

Referring to Figure 2, which is showing computed themes for four days of AP documents from the TDT-2 APW corpus, we can see that the top story for the first three days is initially labeled *pakistan and india* but changes to *nuclear tests* on the following two days. The theme *pakistan and india* loses two documents to other themes on the following day. These are likely documents that do not relate directly to the theme *nuclear tests* and therefore were assigned to other stories as the earlier theme *pakistan and india* became more focused on *nuclear tests*. No documents published on June 2 are assigned to the *nuclear tests* theme. Another story that is moving up over the days begins as *ethnic albanians* and quickly becomes labeled as *kosovo*. Stories can skip days, as shown by the documents related to the *broader tokyo stock price index* themes that appear on June 2 and June 4.

Ordering schemes that take into account relative positions of related groups across days may minimize line crossings at interval boundaries. However we have chosen to consistently order themes for each interval by their number of assigned documents so that the theme order for each day is unaffected by future days. This preserves the organization of themes in the story flow visualization across days and supports information consumers' extended interaction over days and weeks. An individual or team would therefore be able to print out each day's story flow column with document titles and lines, and post that next to the previous day's columns. Such an approach would be unrestricted by monitor resolution and support interaction and collaboration through manual edits and notes on the paper hard copies. Each foot of wall space could hold up to seven daily columns, enabling a nine foot wall to hold two months worth of temporal context along a single horizontal span.

On a single high-resolution monitor, seven days are easily rendered as each daily column is allocated 300 pixels which accomodates most document titles. Longer time periods can be made accessible through the application of a scrolling function.

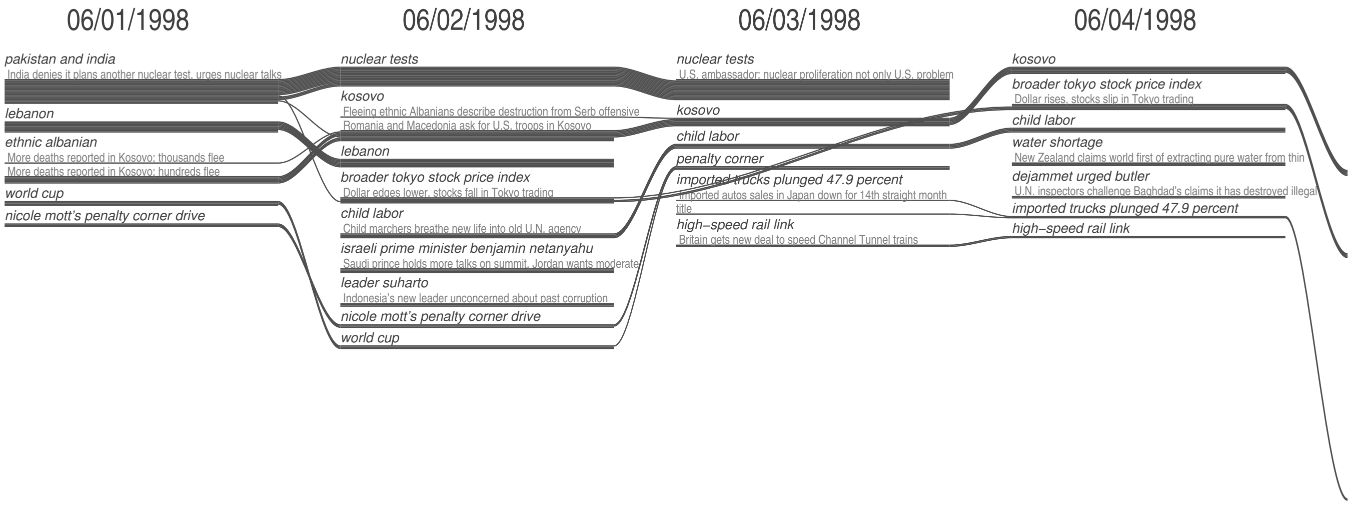


Figure 2: Story flow visualization on four low volume days from the TDT-2 APW Corpus.

5 Discussion

The following discussion explores the described analytic framework and story flow visualization on AP documents within the TDT-2 Corpus. For cases in which multiple documents with identical titles occurred on the same day, we included only the latest version within the analysis. The 10,671 analyzed AP documents span a 6 month period from January 4, 1998 to June 30, 1998.

The AP documents span a range of subjects and primarily cover international topics. Major stories include the olympics, the world cup, iraq, kosovo, human rights, nuclear tests, and many others.

Figure 3 shows the story flow visualization on four weeks of AP documents. At a macro level, it can be observed that a few stories dominate the first two weeks presented in Figure 3a. On February 28, 1998, the top story essentially ends and only small themes are visible until March 5, 1998, in Figure 3b when a new story starts to accumulate new documents. Over the next seven days, that story adds approximately 40 new documents. The motivated reader is encouraged to acquire an electronic version of this paper and zoom into Figure 3 at 500%. At this scale, the individual theme labels and document titles can be easily read and story dynamics observed.

Figure 4 shows the story flow visualization on one week of AP documents. The main stories include *world cup*, *kosovo*, *eu*, *english fans*, and *japanese yen*. On June 15, *kosovo* and *nato* are separate stories that later merge on June 17. On June 20 the story is relabeled as *kosovo albanians* suggesting a new development. Also on June 20, the *world cup game* story splits out across multiple themes, *iran*, *world cup finals*, *korea's choi yong-soo*, *players* only to rejoin again on June 21. This is likely due to an increase in the number of documents published on June 20 that individually relate to specific sub-themes of *world cup game* such as specific countries and players.

While story evolution for large stories is easy to identify, single articles and stories that last two to three days are more difficult to spot. This is not unexpected, as the story flow visualization is designed to show story evolution over time and gives priority to the larger themes in the layout. Themes that do not develop into stories are not optimally represented and would likely benefit from a selective filtering operation or a second visualization in which the main stories play a less dominant role.

Anecdotal feedback from potential users suggest that a filtering mechanism would be welcome in order to minimize line crossings and enable investigation into specific stories and themes. It was also suggested that in application, the story flow visualization would likely be focused on a specific subject area, although the high-level overview was considered effective for users that only need to understand the major story dynamics.

6 Conclusions and Future Work

We have described a novel analytic framework for automatically computing themes and tracking stories as they develop over time and have presented a fundamental visual representation upon which user interface designers can iterate interaction refinements for specific applications.

We anticipate that systems applying these methods will enable information consumers to more effectively integrate new information and understand developments over time. Future work to build on the visual representation presented here will likely focus on identifying preferred interaction and selection capabilities to support information consumers' work flows.

In this paper we presented contributions to the field of visual analytics; a novel analytic framework for calculating themes based on automatically extracted keywords and documents from information streams, a data model that captures requisite temporal information for tracking and adapting to evolving stories, and a visual representation of story flow that conveys temporal context for an information space, enabling information consumers to rapidly identify change and investigate story evolution over time.

Acknowledgements

This work was supported by the National Visualization and Analytics Center (NVAC), a U.S. Department of Homeland Security Program operated by the Pacific Northwest National Laboratory.

References

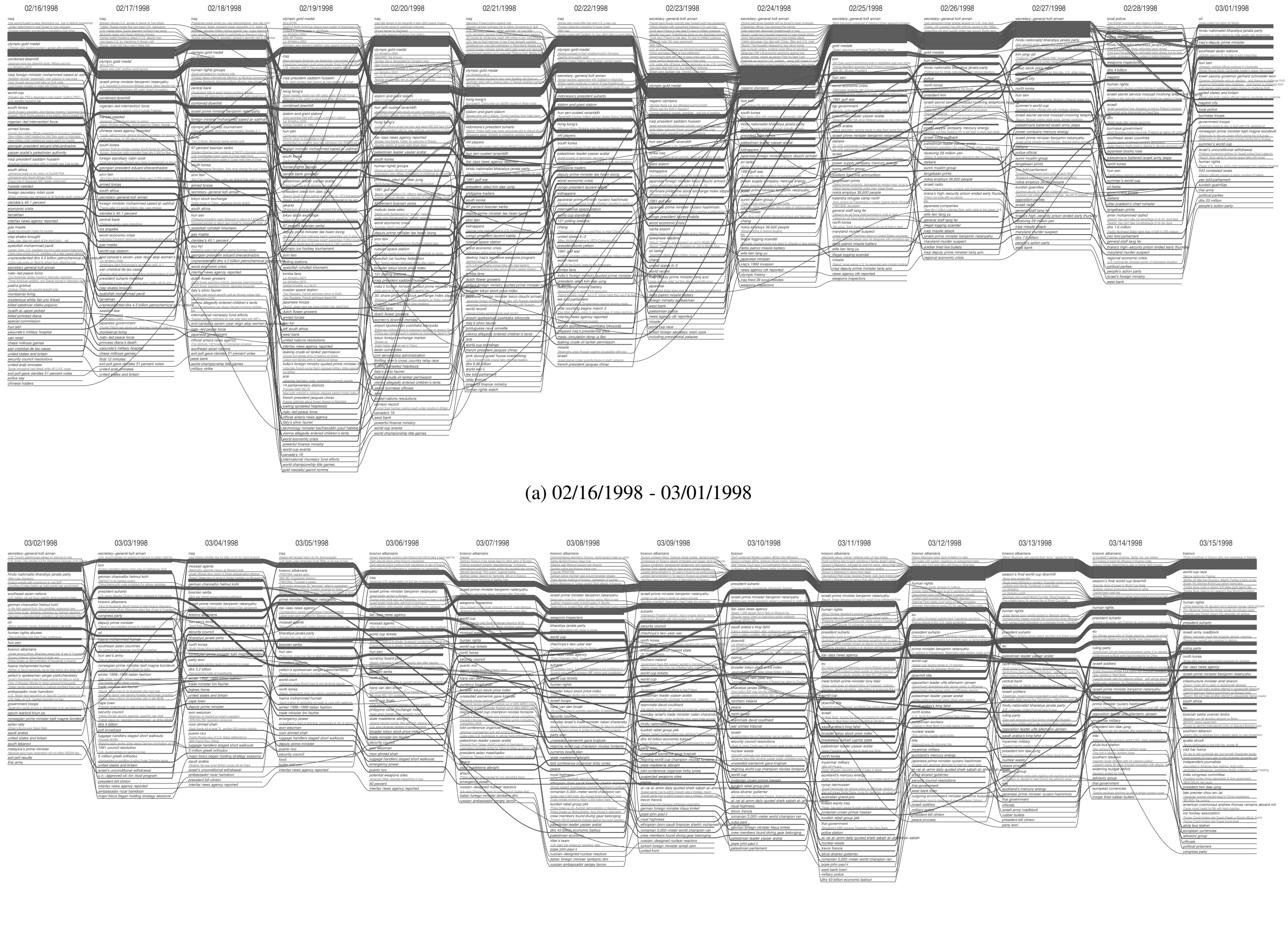
[1] CERATOPS. MPQA Corpus, 2009.

[2] C. C. David, D. Graff, M. Liberman, N. Martey, and S. Strassel. The tdt-2 text and speech corpus. In *in Proceedings of DARPA Broadcast News Workshop*, pages 57–60. Morgan Kaufmann, 1999.

[3] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 193–202, New York, NY, USA, 2006. ACM.

[4] C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning, and E. Frank. Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems*, 27(1-2):81–104, 1999.

[5] E. Hatcher and O. Gospodnetic. Lucene in action. *Action series. Manning Publications Co., Greenwich, CT*, 2004.



(a) 02/16/1998 - 03/01/1998

(b) 03/02/1998 - 03/15/1998

Figure 3: Story flow visualization of calculated themes and document assignments for each day within a four week period from the TDT-2 APW Corpus.

[6] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*, pages 115–123, 2000.

[7] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: visualizing thematic changes in large documentcollections. *IEEE transactions on visualization and computer graphics*, 8(1):9–20, 2002.

[8] E. Hetzler, V. Crow, D. Payne, and A. Turner. Turning the Bucket of Text into a Pipe. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*, pages 89–94, 2005.

[9] K. Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[10] S. Jones and G. Paynter. Topic-based browsing within a digital library using keyphrases. In *DL '99: Proceedings of the fourth ACM conference on Digital libraries*, pages 114–121, New York, NY, USA, 1999. ACM.

[11] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM New York, NY, USA, 2009.

[12] S. Rose, D. Engel, N. Cramer, and W. Cowley. *Text Mining*, chapter Automatic Keyword Extraction from Individual Documents. John Wiley and Sons, Ltd, 2009. to appear.

[13] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[14] B. N. Schilit and O. Kolak. Exploring a digital library through key ideas. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 177–186, New York, NY, USA, 2008. ACM.

[15] E. Tufte and D. Robins. *Visual explanations*. Graphics Press Cheshire, Conn, 1997.

[16] F. Viégas, S. Golder, and J. Donath. Visualizing email content: portraying relationships from conversational histories. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988. ACM New York, NY, USA, 2006.

[17] F. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM New York, NY, USA, 2004.

[18] J. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, pages 236–244, 1963.

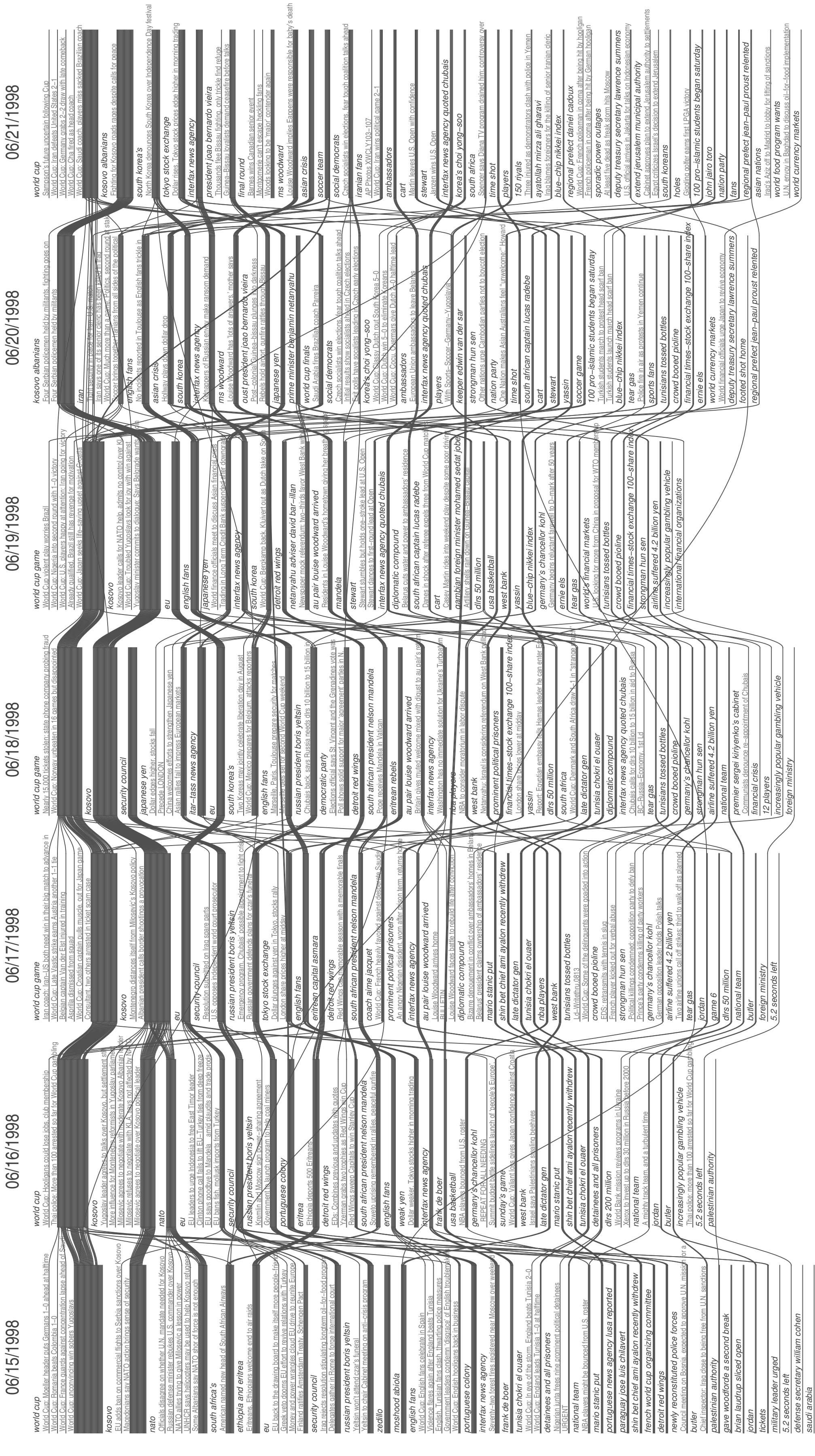


Figure 4: 06/21/1998