Watch the Story Unfold with TextWheel: Visualization of Large-Scale News Streams

WEIWEI CUI and HUAMIN QU, Hong Kong University of Science and Technology
HONG ZHOU, Shenzhen University
WENBIN ZHANG and STEVE SKIENA, State University of New York at Stony Brook

Keyword-based searching and clustering of news articles have been widely used for news analysis. However, news articles usually have other attributes such as source, author, date and time, length, and sentiment which should be taken into account. In addition, news articles and keywords have complicated macro/micro relations, which include relations between news articles (i.e., macro relation), relations between keywords (i.e., micro relation), and relations between news articles and keywords (i.e., macro-micro relation). These macro/micro relations are time varying and pose special challenges for news analysis.

In this article we present a visual analytics system for news streams which can bring multiple attributes of the news articles and the macro/micro relations between news streams and keywords into one coherent analytical context, all the while conveying the dynamic natures of news streams. We introduce a new visualization primitive called TextWheel which consists of one or multiple keyword wheels, a document transportation belt, and a dynamic system which connects the wheels and belt. By observing the TextWheel and its content changes, some interesting patterns can be detected. We use our system to analyze several news corpora related to some major companies and the results demonstrate the high potential of our method.

Categories and Subject Descriptors: H.4.0 [**Information Systems Applications**]: General; I.7.5 [**Document and Text Processing**]: Document Capture—*Analysis*

General Terms: Documentation, Design

Additional Key Words and Phrases: Document analysis, text visualization, macro-micro relation

ACM Reference Format:
Cui, W., Qu, H., Zhou, H., Zhang, W., and Skiena, S. 2012. Watch the story unfold with TextWheel: Visualization of large-scale news streams. ACM Trans. Intell. Syst. Technol. 3, 2, Article 20 (February 2012), 17 pages.
DOI = 10.1145/2089094.2089096 http://doi.acm.org/10.1145/2089094.2089096

1. INTRODUCTION

News articles are a major source of information. For topics such as major companies, famous people, and major events, the volume of news articles is enormous and reading them one by one becomes impossible. News visualization turns news streams into visual forms and shows them to users so they can use their prior knowledge and high-bandwidth visual processing capacity to gain insight into the data.

There are three major issues facing news stream visualization. First, news streams are time-varying high-dimensional data. It is a classic hard problem to develop

This work is supported by HK RGC grant GRF 619309.

Authors' addresses: W. Cui (corresponding author) and H. Qu, Department of Computer Science and Technology, Hong Kong University of Science and Technology, Hong Kong; email: weiwei@cse.ust.hk; H. Zhou, Department of Computer Science, Shenzhen University, Hong Kong; W. Zhang and S. Skiena, Department of Computer Science, State University of New York at Stony Brook, Stony Brook, NY.

DOI 10.1145/2089094.2089096 http://doi.acm.org/10.1145/2089094.2089096

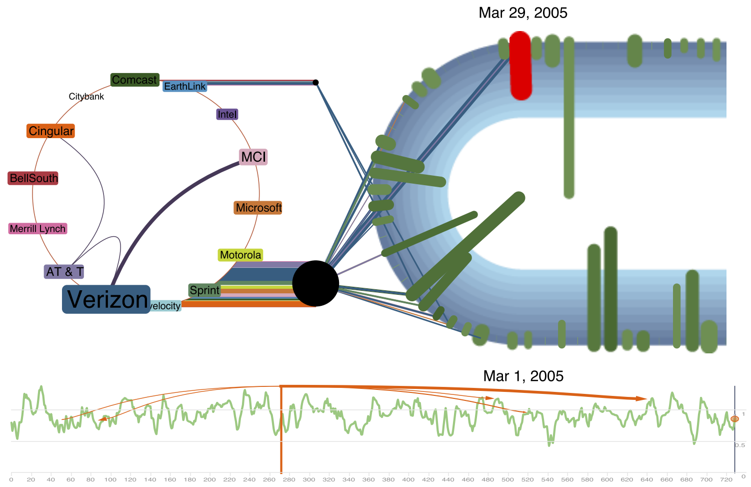


Fig. 1. Visualization of news streams with the TextWheel to reveal multiple attributes of news articles and the macro/micro relations between articles and keywords.

visual encoding schemes for keywords and articles to show their multivariate and time-varying nature. Second, there may exist complex macro/micro relations between keywords and articles. At the micro level, keywords (e.g., Bill Gates and Microsoft) have various relations. At the macro level, text articles may be also related (e.g., dealing with the same topic). Meanwhile, each article contains multiple keywords and each keyword likely appears in many articles; these relations may change with time. These complex macro/micro relations may be very useful for text analysis. However, it is very difficult to visually encode these macro/micro relations. Third, text data can be extremely large and this poses special challenges for the scalability of visual encoding schemes.

In this article we develop a visual analytics system for news streams and address the aforementioned issues facing news visualization. We focus on the multiple attributes of news articles and the dynamic relations between articles and keywords. Our visualization system is built on top of a text mining preprocessing, and consists of three components: significance trend chart generation, entity encoding, and relation encoding. During the preprocessing stage, keywords are extracted with various attributes (e.g., sentiment and frequency), and the relations between keywords and news articles are also established. With the help of modern text mining techniques, we may find a number of interesting or important keywords as candidates for further explorations. In return, our visualization system could also help users verify the keywords picked by an automatic algorithm, and further gain insights into the content of the news streams. After that, all the sentiment and word frequency information is collected and summarized as a line chart, which we call the significance trend chart, to provide an overview of the sentiment evolution over time. At the same time, we use a concise glyph to encode each article whose multiple high-level attributes are first extracted and then encoded using different visual channels of the glyph to provide a succinct overview of the article.

To deal with the complex macro/micro relations existing among keywords and articles, we introduce a novel visual primitive called TextWheel which consists of one or multiple keyword wheels, a document transportation belt, and chains to connect the belt and wheels (see Figure 1). By observing the TextWheel and its content changes, interesting patterns can be detected. Designed after some familiar objects in our life, the TextWheel is intuitive to use. We also provide a set of interaction tools to help users better use our system to analyze large-scale news corpora. The major contributions of this article are as follows.

- We present a visual analytics system for users to visually summarize, organize, abstract, and analyze news streams and other text documents. Our system brings multiple attributes of text documents, dynamic relations between text documents and keywords, micro relations among keywords, and macro relations among documents into one coherent analytical context.
- We design text glyphs to visually encode keywords and documents. The glyph can intuitively and succinctly summarize various nominal and categorical features of keywords and documents.
- We develop a novel visualization primitive called TextWheel representing the complicated macro/micro relations among keywords and documents.

2. RELATED WORK

Text visualization has received considerable research interest in recent years. Various visualization approaches have been proposed to help people effectively analyze and understand large document archives. These techniques can be generally grouped into two categories: document visualization and semantic visualization.

2.1. Document Visualization

Document visualization techniques concentrate on visualizing large document corpora and illustrating relations between documents. Most of the techniques are designed to analyze similarity relations among documents by transforming each document into a feature vector (usually a keyword vector) and then measuring distances between these vectors. Based on the resulting similarity, these documents are clustered and then displayed either in a hierarchical way [Granitzer et al. 2004] or in a unstructured way [Wise 1999]. For example, the SPIRE system [Wise 1999] displayed documents as dots on a 2D plane and clustered them based on their keywords. Documents that are close in the high-dimensional space will also be close on the 2D plane. In 2005, Hetzler et al. [2005] further extended this system to the temporal domain and used it to analyze dynamic document flows. InfoSky [Granitzer et al. 2004] uses a similar approach to visualize hierarchical document collections on a plane space. Users can zoom in and out, just like using a telescope, to explore the documents at different levels of detail. However, InfoSky requires that the documents have already been organized into a hierarchical structure. Different from these two approaches, HiPP [Paulovich and Minghim 2008] can automatically cluster documents into a multi-level structure and allow users to dynamically change the hierarchical structure during the exploration.

2.2. Semantic Visualization

Semantic visualization techniques focus on revealing and analyzing the semantic patterns in documents. Based on the different types of patterns they want to explore, these techniques can be categorized into three groups: literal patterns, keyword patterns, and temporal patterns.

Literal Patterns. Word Tree [Wattenberg and Viégas 2008] uses trie-like structures to explore word collocation patterns in real documents. In the tree structure, nodes represent words with size encoding the word frequency. An edge linking two nodes indicates that these two words are concatenated in the original document. Phrase Nets [van Ham et al. 2009], which can be considered as a follow up work of Word Tree, also uses links to indicate collocation relations. However, it connect words as a graph instead of a tree to convey more sophisticated relations. Mao et al. [2007] developed a technique to visualize the sequential semantic progress in documents. They used statistical methods to identify patterns within the input document data, and then fitted the patterns to a curve. Oelke et al. [2008] applied text fingerprinting on the input text document and provided users with a loopback framework to evaluate and improve the visualization results.

Keyword Patterns. BlogPulse [Glance et al. 2004] monitors over 5.5 million Web blogs and records more than 450K posts every day. To assist users, it provides a Web search interface (http://www.blogpulse.com/) for keywords. It uses a line chart to show the keyword frequency so that users can see how one or multiple keywords appear and disappear over time. Wong et al. [1999] combined the strength of text mining techniques and 3D bar charts to visualize the association rules within multiple items. They used the x-axis to display rules and the y-axis to list items. Then, an association rule can be visualized as a 3D bar standing on the x-y plane. The Jigsaw system [Stasko et al. 2008] provides a visual interface for users to search different keywords in a large collection of documents. It focuses on the relations between different types of keywords, such as people, organizations, and places.

Temporal Patterns. Temporal patterns are also a very active research field in document analysis. Some papers try to track the trend of text flows and identify their changes over time. For example, Wong et al. [2003] proposed a method to visualize dynamic data streams by using animated scatterplots. Allan et al. [2005] also developed a technique to identify evolving stories and new stories by analyzing a growing set of news articles. ThemeRiver [Havre et al. 2000] uses stacked area charts to visualize a collection of documents according to their themes. In the chart, each color stripe represents a theme and is curved smoothly to make it look like a river. Some other papers focus on the cluster or entity relation evolutions in document streams. For example, Erten et al. [2004] combined graph visualization and clustering techniques to analyze how the coauthor relations evolve over time in scientific literature. TextPool [Albrecht-Buehler et al. 2005] clusters document contents on the screen and uses carefully designed animations to help users understand the content changes. Compared with these previous approaches, our article addresses a quite different problem, that is, how to visualize the macro/micro relations widely existing in news streams and other text data.

3. DATA COLLECTION AND PROCESSING

3.1. Data Collection

All experiments in this article were conducted on a 1.5 gigabyte corpus of 333,289 news articles published between 2004 and 2006, with an interesting pedigree. Our experiments in this article are run over subsets of documents selected from this corpus on the basis of a single keyword, say *Merck* or *Verizon* and the resulting set of several thousand articles visualized. We believe this procedure is quite representative of typical tasks concerning corpus understanding.

3.2. Named Entity Recognition

Named entity recognition is a natural language processing problem where one seeks to detect every named entity mentioned in a document. It serves as our feature extraction system for documents, identifying the topics of likely potential interest.

Named entity recognition is a well-studied problem with an extensive literature (e.g., Chieu and Ng [2002] and McDonald [1993]). We primarily employ rule-based techniques that are not vastly different from those in the literature, although they require substantial engineering to achieve good performance.

Here we present the most important phases of our named entity recognition procedure, namely part-of-speech tagging, syntactic tagging, proper noun classification, rules processing, alias expansion and geographic normalization.

Part of Speech Tagging. To extract proper noun phrases from text, we tag each input word with an appropriate part-of-speech tag (noun, verb, adjective, etc.) on the basis of statistical predication and local rules. Part-of-speech tagging requires that the input text be properly partitioned into sentences. This can be readily done using cues from capitalization and punctuation. We employ a popular Part-Of-Speech (POS) tagger [Brill 1994] in our analysis. Such taggers are based on large vocabularies and collections of tagging rules, and require a substantial amount of time to initialize internal data structures.

Syntactic Tagging. Here we employ regular-expression patterns to markup certain classes of important text features such as dates, numbers, and unit-tagged quantities.

Proper Noun Phrase Classification. Each proper noun phrase in a text belongs to some semantic class, such as *person*, *city*, or *company*. In this phase of our pipeline we employ gazetteer (e.g., popular first and last names provided by the U.S. Census Bureau)[1] and Bayesian method [Mitchell 1997] to classify each identified phrase.

Rules Processing. Compound entities are difficult to handle correctly. For example, the entity name *State University of New York, Stony Brook* spans both a comma and an uncapitalized word that is not a proper noun. By comparison, *China, Japan, and Korea* refers to these three entities. Our solution was to implement a small set (~ 60) of hand-crafted rules to properly handle such exceptions.

Alias Expansion. A single entity is often described by several different proper noun phrases, for example, *President Kennedy*, *John Kennedy*, *John F. Kennedy*, and *JFK*, even in the same document. We identify several common classes of aliasing and take appropriate steps to unify such representations into a single set.

Geographic Normalization. Geographic names can be ambiguous. For example, *Albany* is both the capital of New York State and a similarly-sized city in Georgia. However, auxiliary information can be useful in resolving this ambiguity. We developed a geographic normalization routine that identifies where places are mentioned, resolves any ambiguity using population and location information, and replaces the name with a normalized, unambiguous representation.

3.3. Sentiment Analysis

Sentiment analysis of natural language texts is a large and growing field, surveyed in Pang and Lee [2008]. Previous work falls naturally in two groups. The first relates to techniques to automatically generate sentiment lexicons. The second relates to systems that analyze sentiment (on a global or local basis) for entire documents.

Newspapers and blogs express opinion of news entities (people, places, things) while reporting on recent events. We have developed a method [Bautin et al. 2010; Lloyd et al. 2005] that assigns scores indicating positive or negative opinion to each distinct

[1] http://www.census.gov/geneology/names/names_files.html

entity in the text corpus. The method is also used in this system for estimating sentiment scores for entities, and evaluates entity polarity as

$$entity\_polarity_i = \frac{positive\_sentiment\_references_i}{total\_sentiment\_references_i}.$$

3.4. Co-Occurrence Analysis

To determine the significance of the co-occurrence of two entities, we bound the probability that two entities co-occur in more articles if occurrences were generated by a random process. To estimate this probability we use a Chernoff Bound. We have

$$P(X > (1+\delta)E[X]) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{E[X]},$$

where δ measures how far above the expected value the random variable is. If we set $(1+\delta)E[X] = F$ = number of co-occurrences, and consider X as the number of randomized juxtapositions, we can bound the probability that we observe at least F juxtapositions by calculating

$$P(X > F) \leq \left(\frac{e^{\frac{F}{E[X]}-1}}{\left(\frac{F}{E[X]}\right)^{\left(\frac{F}{E[X]}\right)}}\right)^{E[X]},$$

where $E[X] = \frac{n_a n_b}{N}$, N = number of sentences in the corpus, n_a = number of occurrences of entity a, and n_b = number of occurrences of entity b, as the juxtaposition score for a pair of entities.

4. SIGNIFICANCE TREND CHART

After the data collection and processing, all the sentiment and word frequency information is obtained. However, directly going through them one by one seems unwise because there are so many keywords in the document corpus. In this section, we introduce the significance trend chart, which is inspired by entropy and information theory, to analyze and visually summarize changes of sentiment and word frequency information along the whole document stream. According to entropy and information theory, if an object contains more exclusive information, it is more significant. Following this idea, we design a novel method to measure the significance value for a document in the document sequence. we define that a document is more significant if it has more exclusive sentiment and word frequency information compared with its neighboring (preceding or succeeding) documents in the document stream.

4.1. Entropy, Mutual Information, and Conditional Entropy

First of all, we briefly introduce several important information theory concepts which are used in our significance trend chart method. For more details about entropy and information theory, interested readers can refer to the book Cover and Thomas [2006].

The first concept is information entropy. For a discrete random variable X, its entropy value can be calculated as

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i),$$

where X has n possible values $\{x_1, \cdots, x_n\}$, and $p(x)$ is the marginal probability distribution function of X. The higher the entropy value, the more uniform the marginal probabilities. The entropy is maximized as $\log n$ when every marginal probability x_i equals to $\frac{1}{n}$, which means X has maximum information.

The second concept is mutual information, which is the measure of the dependence (i.e., the shared information in the information theory point of view) between two discrete random variables, for example X and Y. The mutual information $H(X;Y)$ between X and Y is calculated as

$$H(X;Y) = \sum_{i=0}^{n} \sum_{j=0}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)},$$

where X has n possible values $\{x_1, \cdots, x_n\}$, and Y has m possible values $\{y_1, \cdots, y_m\}$. $p(x, y)$ is the joint probability function. $p(x)$ and $p(y)$ are the marginal probabilities of X and Y respectively. $H(X;Y)$ equals zero if and only if X and Y are independent, which means X and Y share no information.

The last concept is conditional entropy. The conditional entropy $H(X|Y)$ of a random variable X given another random variable Y is the measure of the amount of information contained by X but not by Y. It can be calculated as

$$H(X|Y) = H(X) - H(X;Y).$$

From the definition, we can see that $H(X|Y)$ is nonnegative. It equals zero if and only if X and Y share everything (i.e., they are identical). On the other hand, it has maximum value $H(X)$, when X and Y are independent (i.e., $H(X;Y) = 0$).

4.2. Information-Theoretic Significance Calculation

To estimate the significance for every document in the stream, we need to calculate its conditional entropy value given its neighboring documents. First, we calculate the average sentiment towards every tagged word in the document. After that, a histogram is built to summarize the sentiment information distribution. For example, we find the maximum and minimum average sentiments in a document X, then set 32 intervals between these two values. Each interval in the histogram counts the number of words that fall into it. Then the information entropy $H(X)$ of the document can be computed as

$$H(X) = -\sum_{i=1}^{N} \frac{\text{count}_i}{\text{count}_X} \log \frac{\text{count}_i}{\text{count}_X},$$

where N is the bin number, and count_i and count_X indicate the word count in the ith bin and the whole document X, respectively.

Similarly, given another document Y, the mutual entropy $H(X|Y)$ of document X can be calculated using a two-dimensional histogram. Finally, we estimate the significance value $S(X_t)$ of document X_t at time t to be the average value of its conditional entropy values given its preceding document X_{t-1} and succeeding document X_{t+1}, that is,

$$S(X_t) = \frac{1}{2}(2H(X_t) - H(X_t; X_{t-1}) - H(X_t; X_{t+1})).$$

5. SYSTEM OVERVIEW

Figure 1 shows the interface of our TextWheel visualization system with all its main visual components: a significance trend chart, a document transportation belt, and one or multiple keyword wheels. These three components provide users with three levels of views of details in the document stream.

The line chart at the bottom is the significance trend chart. It shows the highest level of view by depicting the sentiment changes extracted from a document stream. The x-axis encodes the time and the y-axis encodes the significance value of the documents.

On the upper right, a U-shape document transportation belt shows users a small portion of documents in the whole stream. Each glyph on the belt represents one or multiple documents. When users interact with our system, the document glyphs are transported along the belt. When a new document glyph enters the focus region (the semicircular part of the belt), it is highlighted in red. We also put a sliding bar on the significance trend chart to show the location of the highlighted document glyph in the whole document stream. Users can also drag the sliding bar to synchronize the transportation belt and the significance trend chart. For example, when users interact with the transportation belt (e.g., speeding up the transportation or reversing the direction of the transportation), the bar can show users which part of the data stream the belt is currently showing. On the other hand, users can also directly drag the sliding bar on the chart to a chosen time of the document stream; the belt will automatically roll forward or backward to show the documents at that time. We also encode the macro relations between documents on the chart. By drawing arcs from the sliding bar, all the documents which are most related to the highlighted document are pointed out on the whole document stream.

On the upper left are one or multiple wheels that show different keywords users are interested in. To encode the micro relations between keywords, we put keywords uniformly on a circle and then use lines to indicate the relations between keywords. The keyword wheels also interact with all the documents in the focus area (the semicircular part of the transportation belt) by connecting them with some chains.

6. TEXTWHEEL

Our TextWheel system consists of several visual components, which are all inspired by some everyday objects that users are very familiar with. For example, our keyword wheel is inspired by Ferris wheels. News articles move along the transportation belt just like baggage moving along the conveyor belt. By drawing experience from our everyday life, our system has obvious metaphors and is intuitive to use. In this section, we introduce the design details for each component.

6.1. Entity Encoding with Glyphs

There are two major entities in our data: keyword and document. Both of them have multiple attributes.

Keyword Glyph. As keywords have been widely used in Web pages, some visual encoding schemes have been well established. For example, the Google Visualization API can allow users to use size and color to encode various attributes of keywords. Usually, the font size of a keyword represents the frequency of the keyword appearing in a document. We adopt this scheme in our system. Figure 2(a) shows the word cloud generated by the Google Visualization API. These keywords are later arranged into a circular frame called *keyword wheel* (see Figure 2(b)) in our system.

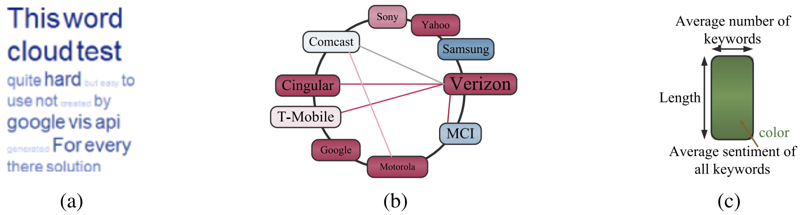


Fig. 2. Glyphs: (a) Keyword cloud generated by the Google Visualization API; (b) keyword wheel; (c) the document glyph for one article.

Document Glyph. We adopt a simple rectangular shape glyph in our current system, though other more complicated glyphs can also be used. The width of the rectangle encodes the average number of keywords while the height indicates the article length. The color of the glyph encodes the average sentiment expressed in an article. Figure 2(c) illustrates the document glyphs used in our system.

6.2. Document Transportation Belts

Layout. We adopt a U-shape layout for our belt because the U-shape belt can provide more space for the layout of documents and better convey the “endless” feeling about news streams, considering there usually are much more documents than keywords. In addition, the U-turn can naturally divide the document belt into three parts and give a nice focus+overview view. The curved part serves as a focus region, in which the keywords are connected to the documents that fall into this region, while the top and bottom straight parts give users a bigger picture about what have just left and what are coming.

Speed. The speed of the belt is highly controllable. We can directly drag the sliding bar on the significance trend chart to roll the transportation belt forward or backward to any time points we are interested in. We also set a uniform or automatically computed speed for the transportation belt, so we can see the belt rolling automatically. If we find something interesting, we can stop the belt or lower the speed to allow more time for inspection.

Order of documents. The documents can come into the transportation belt with different orders. By default, they are arranged by time. Other orders are also possible. For example, the documents from the same sources can come together.

6.3. Keyword Wheels

The keyword glyphs are put in the keyword wheels. The relation between two keywords is encoded by simply connecting them with a line. Meanwhile, the sentiment change of the keywords can be naturally revealed by the rotation of the wheel.

Keyword selection. We select the most relevant keywords according to user interest or recommended during data preprocessing. Every keyword glyph on the same wheel is assigned with one unique background color. Therefore, users are not suggested to choose more than 20 keywords for one wheel. Otherwise, the colors may become difficult to differentiate. In case users have an actual need to show more than 20 keywords, they can put them in different wheels.

Keyword position. Keywords will be uniformly positioned in the circular frame of the keyword wheel. Their positions are computed based on their inter-relations. We use a greedy algorithm to position keywords and the keyword appearing most frequently in the documents will be put at the center of the circular segment falling into the focus window. Then other keywords can be positioned accordingly.

Keyword update. As the documents in the focus window change gradually some keywords may become more or less frequent or even disappear from the documents in the focus window. When a keyword becomes more frequent, the glyph size becomes bigger, and vise versa. When a keyword disappears from the documents in the focus window, it is still kept on the wheel, but its background color will disappear, so that it will not cause much distraction to users.

6.4. Dynamic System

We further connect the keyword wheel and document transportation belt with chains to form a dynamic system. A keyword is likely contained in multiple documents while every document holds different sentiments towards it. Once a document enters the focus window, a chain is connected to the document with all the related keywords in the keyword wheel. The width, color, and opacity of the chain can encode various attributes of the relation between the document and the keyword. For example, we can use the width to encode the strength of the sentiment and use the color to indicate what word this sentiment is about. There are two hubs between the keyword wheel and the transportation belt. Every chain will go to one of them first. All the chains indicating positive sentiments go to the lower hub, while all the chains indicating negative sentiments go to the upper hub. At each hub, all the chains with sentiments towards the same keywords (i.e., all the chains with the same color) are bundled together. Then the bundled chains connect to the keyword wheel to drag the wheel to rotate. We assume both bundled chains have attractive forces on the wheel, and the force values are proportional to the bundle size. Therefore, if the sum of negative sentiments towards to all the keywords is stronger than the sum of positive sentiments, the wheel will rotate clockwise in our system, and vise versa.

6.5. User Interaction

We provide a set of interaction tools to help users better use our system to deal with a very large number of news articles. Some of these tools are summarized as follows.

- *System configuration*. Our system is highly configurable. The size of the wheel and belt, the order of the documents in the belt, and the encoding schemes are all configurable. Our system allows users to deploy multiple keyword wheels in the display. Each wheel may represent a separate group of keywords.
- *Speed control.* Users can pause, fast forward, or rewind the belt. Users can also set a time for the whole stream and then the speed will be automatically computed.
- *Coordinated view.* Users can link document glyphs in the TextWheel view with the real documents by simply clicking the glyph icons. Users can also click on a single edge connecting two keyword glyphs, then a line chart will pop up to show the correlation evolution between these two keywords over time.
- *Filtering.* Filtering is available to show a small set of articles that users are interested in. There is a control panel beside the visualization display for users to filter out some news articles by date, size, source, etc.
- *Clustering and highlighting.* Users can also cluster the documents or cluster the chains in the window region to reduce clutter. If users still cannot see individual chains in the focus region, they could also click a document glyph, and all the related chains and keywords will be highlighted accordingly.

6.6. Alternative Designs

We have considered some other designs for our TextWheel system. For example, the transportation belt can also be a wheel or a straight belt. We have also considered putting the keywords in another straight or U-shape belt. The wheel-shape transportation belt does not work in our system, since not all documents can be positioned on the wheel. The straight belt is a better choice. But the U-shape belt looks more consistent with the keyword wheels. Moreover, since we may occasionally want to explore relations between document glyphs, it is much easier to draw edges inside a U-shape belt than on a straight belt. However, the movement of the keywords looks strange compared with the rotation of the keyword wheel. The wheel could also encode the micro relation between keywords more effectively. In addition, using different primitives for keywords and documents may avoid confusing these two different entities.

For the dynamic system, we have considered directly connecting the keyword glyphs with document glyphs, or to directly use a physical spring model on the keyword wheel. However, they are too distracting and hard for users to focus on what they are really interested in. Thus we decided to simplify the design, such as grouping all the chains first before they connect to the wheel, and fix all the keyword glyphs on the wheel before further exploration.

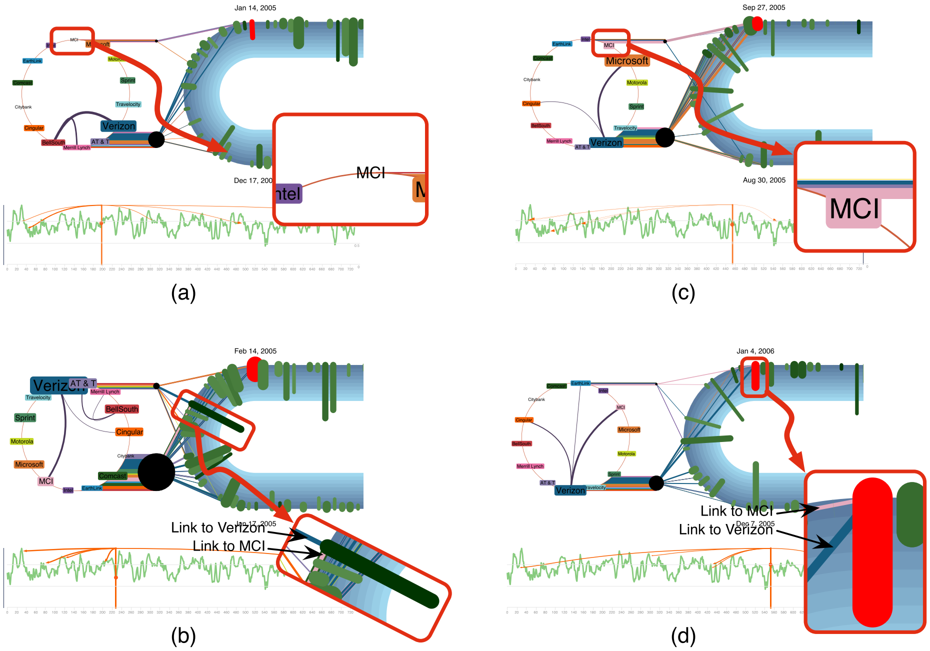
7. CASE STUDY

We have applied our system to the news streams mentioned in Section 3.1. The news streams are related to six major topics: Microsoft, Sony, NYSE, Merck, China, and Verizon. Each topic contains thousands of news articles from various sources. We first ran our system for each topic and did some initial screening. Once we found interesting patterns, we configured the system and fine-tuned the visual displays to bring out more details for analysis.

In this section, we describe two findings. The following encoding schemes were used in all experiments: the keyword size encodes the frequency (larger size means higher frequency); the document glyph represents all the articles in one day; the document glyph height encodes the average length of all the articles in that day (larger height means longer length); the document glyph width encodes the number of articles (larger width means more articles); the document glyph color encodes the average number of keywords mentioned in all the articles (darker color means larger number); the width of the lines in the keyword wheel encodes the strength of the co-occurrence (thicker line means higher co-occurrence); the arcs in the significance chart indicate the places having the documents most similar to the document at the sliding bar (thicker arc means higher similarity); the color of a line linking a wheel and a document encodes which keyword the document has sentiment towards (connecting to the upper part of a wheel means the sentiment is negative, and connecting to the lower part of a wheel means the sentiment is positive). All the experiments are conducted on a Macbook Pro with Intel Core 2 Duo 2.2 GHz CPUs and 2GB memory. With some preprocessing, our system can handle thousands of news articles in real time.

7.1. Verizon's Acquiring MCI

During the initial screening of the news streams related to Verizon, we noticed some interesting relations between Verizon and MCI. Sometimes these two keywords are quite large in the display and dominate the view. Meanwhile, there is a thick line linking them (see Figures 3(b) and 3(d)). Sometimes, the MCI keyword totally disappears from the documents in the focus window (i.e., MCI background color disappears) (see Figures 3(a) and 3(c)). Thus, we decided to focus on this situation and try figure out what happens. Figure 3 shows some screen shots of the exploration.



(a) (c)

(b) (d)

Fig. 3. The story of Verizon's merger with MCI.

From Figure 3(a), we can see that the MCI keyword has not appeared in the keyword wheel before January 2005. Then around February 2005, MCI starts to appear in the keyword wheel and the line linking the Verizon glyph and the MCI glyph is quite thick (see Figure 3(b)). In all the document glyphs which have links with both MCI and Verizon keywords, we noticed that there is one that has green color, high height, and very thin width (highlighted by a red rectangle in Figure 3(b)), which indicates that it only contains one long article mentioning both keywords many times. We opened that glyph and read the document. After quickly going through this article, we found this paragraph: "MCI Inc. confirmed an agreement to be acquired by Verizon Communications in a deal with a total value of $6746000000." This explains the reason that these two companies became hot topics and were frequently mentioned together. Then around October 2005, the link connecting MCI and Verizon disappears (see Figure 3(c)). We believed that the merging of these two companies was no longer a hot topic. However, around December 2005, the thick link between these two glyphs shows up again (see Figure 3(d)). So we paused, and chose the only glyph that has links to both MCI and Verizon (highlighted by a red rectangle in Figure 3(d)). There are only two documents in that glyph, and one of them mentioned that "Verizon Communications, Inc., has completed its acquisition of MCI. MCI's assets will be folded into a new Verizon unit called Verizon Business."

From this case study, we can see that our system can help users quickly identify the relations between keywords and then narrow down to some articles to find the reasons for these relations.

7.2. Merck and Its Troublesome Drug Vioxx

For the news streams related to Merck, we divided the keywords into two groups, that is, company and drug. During the initial exploration, we noticed that a drug called Vioxx appears frequently in the display and the sentiment towards it changes

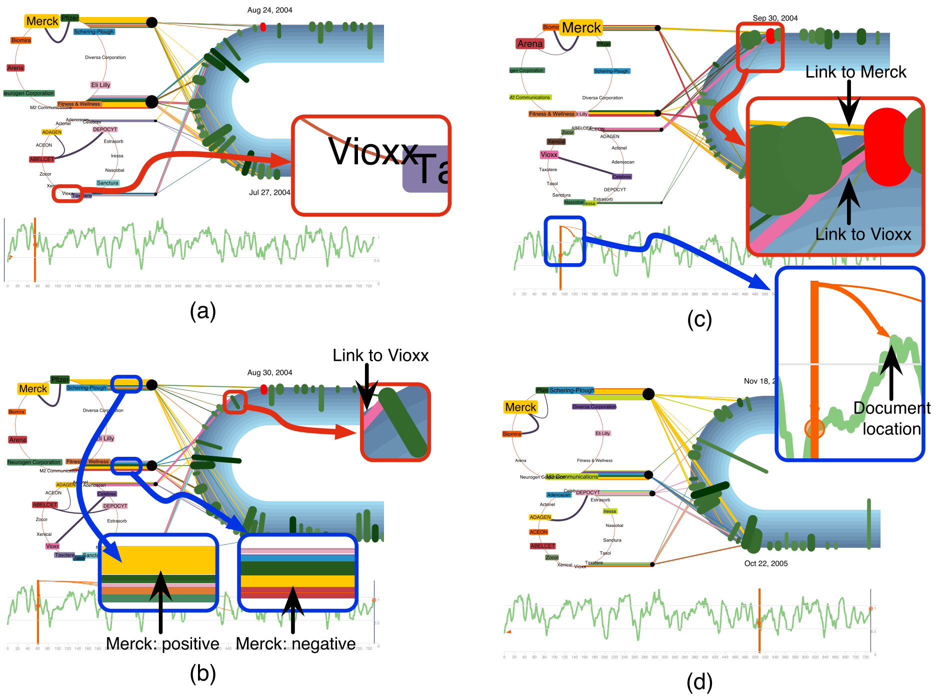


Fig. 4. The story of Merck's troublesome drug Vioxx.

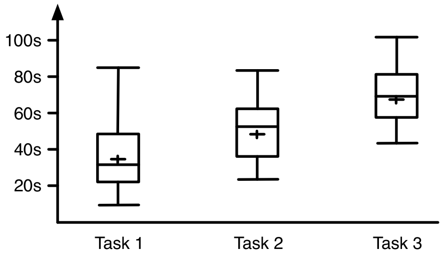
dramatically in 2004 from neutral to bad. Then we reran the system and paid special attention to this drug. Figure 4 shows some screen shots.

According to Figure 4(a), Vioxx has not appeared in the drug keyword wheel around August 23, 2004. Then on August 30, Vioxx shows up and the sentiment towards it is quite negative (see Figure 4(b)). We followed the thickest link from the Vioxx glyph and identified a document glyph representing articles on August 26 (highlighted by a red rectangle in Figure 4(b)). This document glyph is slim because it only has one article from the AFX UK Focus on that day. This article says: "Analysis of a study on the safety of COX-2 inhibitors found that Vioxx doses above 25 milligrams per day tripled the risk of cardiovascular events...". Therefore, we can see that the negative sentiment towards Merck, the maker of Vioxx, is slightly stronger than the positive sentiment (highlighted by blue rectangles in Figure 4(b)).

Both Merck and Vioxx remain stable until September, when Vioxx becomes much more negative (see Figure 4(c)). We found the glyph with the thickest links to Merck and Vioxx (highlighted by a red rectangle in Figure 4(c)) and retrieved the corresponding articles (two documents in total). One article from AFX International Focus mentioned that "Merck said Merck was withdrawing its Vioxx arthritis drug from shelves worldwide, resulting in a 50 cent to 60 cent reduction in per-share earnings." We then followed the thickest arc on the significance trend chart to reveal similar news articles from the same source (highlighted by a blue rectangle in Figure 4(c)). This article also said something negative about Merck: "Merck slumped more than 5 percent and was the biggest percentage loser among Dow Jones Industrial Average components." Finally, Vioxx is no longer a hot topic because it becomes small in size and often without any background color (see Figure 4(d)), which means it is not mentioned by any documents in the focus window.

Task	Time (second)	
	Mean	SD
Task 1	35	21
Task 2	49	18
Task 3	68	16

(a)



(b)

Fig. 5. User study result for evaluating the efficiency of the encoding schemes.

This case study demonstrates that the sentiments expressed by the keyword glyphs are very useful in news analysis and with the macro/micro relation information provided by our system we can quickly identify the sources of these sentiments.

8. USER STUDY

In addition to the case study, we also conducted an informal user study consisting of 12 college students. They were all year two students having no prior information visualization knowledge, each of whom was asked to use our system and answer three questions. Before attempting the study, the users were briefly introduced to our system. Meanwhile, they were also encouraged to play with it using different configurations such that they could get more familiar with our visual encodings.

To perform the case study, we chose a fraction (718 documents in a half year) of the data used in the second case study as the testing data. Then we processed the data and picked the top 12 frequently mentioned companies from the documents. After that, the document corpus, along with all 12 company names, was loaded into our system. The documents on the transportation belt were grouped by day, and all 12 companies were placed in the same wheel. Then, we presented this system to the users and asked them to finish three tasks. These tasks are mainly designed to test the efficiency of the encoding scheme of the macro/micro relations. (The effectiveness of macro/micro relation encoding scheme is demonstrated in the case study.)

The first task is designed to test the efficiency of our sentiment encoding scheme. In this task, the users need to find out the time when a specific company on the wheel reaches its largest positive sentiment. The second task is designed to test the efficiency of the relation encoding scheme. In this task, they are asked to discern which two companies have the strongest relationship in the whole document stream. The last task combines these two previous tasks. In this task, we challenge them to find out the company that has a very positive sentiment and still has a relatively strong relation with a specific keyword.

In each task, we recorded their answers and the response time they needed to finish it. The results are shown in Figure 5. For the first task, ten users generally found the correct time (with average error of 5.5 days), while the remaining two users found the time that was also a local peak of positive sentiment with the second largest value in the whole stream. The average response time is 35 seconds. However, we also noticed that the standard deviation is as big as 21 seconds (see Figure 5(a)), which is probably because some of the users were still not quite familiar with our system. The first task is designed to warm up the subjects. It does not demonstrate the advantages of our system, since a classical line chart may be better for this task. Therefore, we asked the subject to finish two more complicated tasks, in which our system may better show its advantages. The second task is a little harder than the first one, since users may need to track multiple keywords at the same time. This time, eight users found the correct pair with an average response time of 49 seconds, which is a little longer than that in the first task. However, the standard deviation is reduced to 18 seconds. The final task is the most difficult one, because it requires users to synthesize two different visual encodings and keep the results in mind such that they could find the most appropriate one in the whole stream. However, the result seems quite satisfactory. Since our system includes a sliding bar for quickly rolling the transport belt forward and backward, users can freely examine the documents at any speed they like. Ten users found the correct one with an average response time of 68 seconds. On the other hand, the standard deviation is further reduced to 15 seconds. All of the three examples demonstrate that, with a little training, most users can use our system to explore large news streams and correctly find some patterns in the testing data for the three tasks. On the other hand, classical line charts are not suitable for the remaining two tasks. Since it needs to generate a curve for each pair of keywords, users may be easily overwhelmed when there are many keywords to explore.

After they finished all three tasks, we also asked the users about their general feeling about our design. The biggest concern we had before conducting the user study was the visual clutter and distraction in the interface. However, the responses to our system from users are quite enthusiastic. Overall, they feel the system is informative, intuitive, and visually appealing. In particular, they think that the visual interface is well organized, and each of the visual component has a very clear purpose, which makes the exploration easy.

We also asked their opinions for each visual component. Most of them appreciate the familiar metaphors, such as wheels, belts, and chains, to intuitively represent document streams and sophisticated relations involving documents, keywords, and sentiment values. Among all the visual elements, the users appreciate the wheel most, because it combines multiple aspects of information in a natural and intuitive way and they do not need to think too much to understand the meanings during the fast exploration. In addition to the wheel, they also agree that the significance curve is a very important feature, because without the overview, they will have no choice but to wildly explore the whole document stream, which could make them feel insecure. For example, during the free exploration, they were all interested in those time points when the curve is very steep, and spent some time there trying to figure out why. Therefore, they also feel that the accuracy of the curve is extremely important. Otherwise, the curve is misleading.

In addition to those positive comments, they also raised a few constructive suggestions. For example, five users feel that using thickness to perform quantitative comparison is not very efficient. Therefore, they suggested we provide an option to show actual numbers when the thicknesses are too similar to compare. Another suggestion is regarding the system control. For example, some of the users would like to control the sensitivity of the wheel to the summarized keyword sentiments, so they can ignore those minor fluctuations and focus more on big patterns. Some of them would like to be able to add or remove keywords during exploration other than choosing them before exploring. We think these suggestions are all very valuable, and intended to explore them in future work. Furthermore, we also plan to deploy to a news Web site to reach more audiences and further improve our system based on their comments.

9. DISCUSSIONS

From the experiments, we can see that our system has advantages and can encode a lot of information into one display for analysis such that some unexpected correlations may emerge. Meanwhile, we also identify some weaknesses of our system. Our system can handle thousands of documents and tens of keywords effectively. If the news streams contain too many articles, exploration may still take a very long time and reduce its effectiveness. Too many keywords may overwhelm the keyword wheels and cause clutter in the display. To deal with this problem, it is better to use our system together with some data mining techniques to first narrow down the document scope. Some well-established techniques in the visualization field, such as clutter reduction methods, can also be applied. As our system provides useful information into one display, it is possible that users are overwhelmed and lose their focus. Thus, we recommend that users turn off some features of the system and only focus on one feature at the beginning. After getting familiar with the data, more details can be brought into the display.

10. CONCLUSION

In this article, we have presented a visual analytics system for large-scale news streams. Our system aims at providing the multiple attributes of news articles and keywords, the dynamic relations between news articles and keywords, the micro relation among keywords, and the macro relation among documents simultaneously to users for analysis. We designed an original TextWheel which consists of a document transportation belt, one or multiple keyword wheels, and a chain system to connect the belt and wheel. Our system is based on some everyday objects which users are familiar with and thus the learning curve for our system should be low. We demonstrated the effectiveness of our system by applying it to several news corpora related to some major companies, obtaining some interesting findings. The application of our system is not limited to news streams. It can be used to analyze other data in text format (e.g., emails, blogs, and internal memos), and to reveal the macro/micro relations existing in other data formats, such as video clips.

In the future, we plan to further extend our system to encode more attributes of text documents. We believe the integration of our system with other data mining methods will make it more powerful. We also want to encode the uncertainty associated with the sentiment and co-occurrence computations. Although we conducted a user study which has suggested that our subjects agree that our design and system can help them with pattern huntings in macro/micro relations, it is still preliminary and informal. Therefore, we also plan to conduct a more thorough user study, which involves more subjects, comparison with other tools, and a formal questionnaire-based survey.

REFERENCES

ALBRECHT-BUEHLER, C., WATSON, B., AND SHAMMA, D. 2005. Visualizing live text streams using motion and temporal pooling. *IEEE Comput. Graph. Appl.* 52–59.

ALLAN, J., HARDING, S., FISHER, D., BOLIVAR, A., GUZMAN-LARA, S., AND AMSTUTZ, P. 2005. Taking topic detection from evaluation to practice. In *Proceedings of the Hawaii International Conference on System Sciences*. 101–101.

BAUTIN, M., WARD, C., PATIL, A., AND SKIENA, S. 2010. Access: News and blog analysis for the social sciences. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 1229–1232.

BRILL, E. 1994. Some advances in rule-based part of speech tagging. In *Proccedings of the 12th National Conference on Artificial Intelligence*.

CHIEU, H. AND NG, H. 2002. Named entity recognition: A maximum entropy approach using global information. In *Proceedings of the International Conference on Computational Linguistics*. 190–196.

COVER, T. M. AND THOMAS, J. A. 2006. *Elements of Information Theory* 2nd Ed. Wiley-Interscience.

ERTEN, C., HARDING, P., KOBOUROV, S., WAMPLER, K., AND YEE, G. 2004. Exploring the computing literature using temporal graph visualization. In *Proceedings of the Conference on Visualization and Data Analysis (VDA)*.

GLANCE, N. S., HURST, M., AND TOMOKIYO, T. 2004. Blogpulse: Automated trend discovery for weblogs. In *Proceedings of the Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

GRANITZER, M., KIENREICH, W., SABOL, V., ANDREWS, K., AND KLIEBER, W. 2004. Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In *Proceedings of the IEEE Symposium on Information Visualization*. 127–134.

HAVRE, S., HETZLER, B., AND NOWELL, L. 2000. Themeriver: Visualizing theme changes over time. In *Proceedings of the IEEE Symposium on Information Visualization*. 115–123.

HETZLER, E. G., CROW, V. L., PAYNE, D. A., AND TURNER, A. E. 2005. Turning the bucket of text into a pipe. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'05)*. IEEE Computer Society. 89–94.

LLOYD, L., KECHAGIAS, D., AND SKIENA, S. 2005. Lydia: A system for large-scale news analysis. In *Proceedings of the 12th Symposium of String Processing and Information Retrieval (SPIRE'05)*. Lecture Notes in Computer Science, vol. 3772. Springer, 161–166.

MAO, Y., DILLON, J., AND LEBANON, G. 2007. Sequential document visualization. *IEEE Trans. Vis. Comput. Graph. 13*, 6, 1208–1215.

MCDONALD, D. 1993. Internal and external evidence in the identification and semantic categorization of proper names. In *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*. 32–43.

MITCHELL, T. 1997. *Machine Learning*. McGraw-Hill.

OELKE, D., BAK, P., KEIM, D., LAST, M., AND DANON, G. 2008. Visual evaluation of text features for document summarization and analysis. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. 75–82.

PANG, B. AND LEE, L. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers.

PAULOVICH, F. V. AND MINGHIM, R. 2008. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Trans. Vis. Comput. Graph. 14*, 6, 1229–1236.

STASKO, J., GÖRG, C., AND LIU, Z. 2008. Jigsaw: Supporting investigative analysis through interactive visualization. *Inf. Vis. 7*, 2, 118–132.

VAN HAM, F., WATTENBERG, M., AND VIEGAS, F. B. 2009. Mapping text with phrase nets. *IEEE Trans. Vis. Comput. Graph. 15*, 6, 1169–1176.

WATTENBERG, M. AND VIÉGAS, F. B. 2008. The word tree, an interactive visual concordance. *IEEE Trans. Vis. Comput. Graph. 14*, 6, 1221–1228.

WISE, J. A. 1999. The ecological approach to text visualization. *J. Amer. Soc. Inf. Sci. 50*, 13, 1224–1233.

WONG, P. C., WHITNEY, P., AND THOMAS, J. 1999. Visualizing association rules for text mining. In *Proceedings of the IEEE Symposium on Information Visualization*. 120–123.

WONG, P. C., FOOTE, H., ADAMS, D., COWLEY, W., AND THOMAS, J. 2003. Dynamic visualization of transient data streams. In *Proceedings of the IEEE Symposium on Information Visualization*. 97–104.

Received July 2010; revised October 2010; accepted October 2010