Sequential Document Visualization

Yi Mao, Joshua V. Dillon, and Guy Lebanon

Abstract— Documents and other categorical valued time series are often characterized by the frequencies of short range sequential patterns such as n-grams. This representation converts sequential data of varying lengths to high dimensional histogram vectors which are easily modeled by standard statistical models. Unfortunately, the histogram representation ignores most of the medium and long range sequential dependencies making it unsuitable for visualizing sequential data. We present a novel framework for sequential visualization of discrete categorical time series based on the idea of local statistical modeling. The framework embeds categorical time series as smooth curves in the multinomial simplex summarizing the progression of sequential trends. We discuss several visualization techniques based on the above framework and demonstrate their usefulness for document visualization.

Index Terms—Document visualization, multi-resolution analysis, local fitting.

1 Introduction

Categorial valued time series $y = \langle y_1, \ldots, y_N \rangle$, $y_i \in V$ such as text documents or protein sequences are difficult to visualize due to their discrete categorical nature. This difficulty is especially severe when the set V representing the range of possible nominal or categorical values is large. For example, the set V contains dozens of amino acids in the case of proteins and hundreds of thousands, if not millions, of possible words in the case of text documents. Due to the categorical nature of the data, visualization techniques designed for numeric time series data, e.g. [23, 9], are not applicable. A popular alternative is to use dimensionality reduction techniques such as principal component analysis (PCA) or multidimensional scaling (MDS) (e.g., [4]) for vectorized data based on the sequence histogram

$$\gamma^{\text{hist}}(y) \stackrel{\text{def}}{=} \left(\frac{1}{N} \sum_{j=1}^{N} \delta_{1,y_j}, \ldots, \frac{1}{N} \sum_{j=1}^{N} \delta_{k,y_j} \right) \in \mathbb{R}^{|k|} \tag{1}$$

where we assume without loss of generality that $V = \{1, \ldots, k\}$ and

$$\delta_{a,b} \stackrel{\text{def}}{=} \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}. \tag{2}$$

While conveniently representing categorical sequences as k-dimensional numeric vectors, the histogram representation, also known as bag of words, completely ignores word ordering. For example, assuming $V = \{1, \ldots, 5\}$ we have

$$\gamma^{\text{hist}}(\langle 1,4,3,1,4 \rangle) = \gamma^{\text{hist}}(\langle 4,4,3,1,1 \rangle) = \left(\frac{2}{5}, 0, \frac{1}{5}, \frac{2}{5}, 0 \right).$$

The histogram representation may lead in some cases to effective visualization of the semantics characterized by the entire document. However, it lacks the ability to capture and visualize the semantic transition between different parts of the document. This lack of a within-document locality is particularly problematic when the visualization focus is on visualizing the contents of a single document (or a small number of documents) rather than a large corpus.

- *Yi Mao is with School of Elec. and Computer Engineering, Purdue University - West Lafayette, E-mail: ymao@ece.purude.edu.*
- *Joshua V. Dillon is with School of Elec. and Computer Engineering, Purdue University - West Lafayette, E-mail: jvdillon@ece.purude.edu.*
- *Guy Lebanon is with Department of Statistics, and School of Elec. and Computer Engineering, Purdue University - West Lafayette, E-mail: lebanon@stat.purude.edu.*

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 27 October 2007.

For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

A partial solution to the above problem is to represent y by the histogram of length-n patterns called n-grams [14]. For example, n-grams for $n = 3$, also called trigrams, are sequential patterns $\langle v_1, v_2, v_3 \rangle$, $v_1, v_2, v_3 \in V$ and their histogram is simply the vector of normalized counts of length-3 pattern appearances. For example, the relative frequency of $\langle 3,1,3 \rangle$ in a sequence y is

$$\frac{1}{N-3+1} \sum_{i=1}^{N-3+1} \delta_{y_i,3} \delta_{y_{i+1},1} \delta_{y_{i+2},3}.$$

Increasing the pattern length n invokes the well known bias-variance tradeoff in statistics. High n has the potential of capturing longer range sequential information. On the other hand, the dimensionality of the n-gram representation increases exponentially with n making the n-gram histogram a poor estimator of the underlying expectation parameters. Reducing n improves the accuracy of the n-gram histogram as a statistical estimator, but prevents the representation from capturing long range interactions. Since for text documents $N \ll |V|$, most words and phrases in V will never occur at the same document. As a result, the pattern length n in n-grams is usually restricted to small sizes such as $n = 1, 2,$ or 3.

Unfortunately, the n-gram representation does not adequately address the histogram shortcomings associated with the lack of sequential information. Frequencies of short phrases do not reflect long or medium range sequential information, nor do they reflect the positions in the document at which the different phrases occur. It is fair to say that the n-gram representation may be viewed as offering a tokenization of short phrases rather than words which is useful for capturing linguistic ambiguity. It does not capture any sequential information beyond that and remains poorly suited for sequential visualization.

We present a new framework for sequential visualization of documents and other categorical time series which generalizes n-grams in a robust way and maintains both long and short range sequential information as well as position information. The framework, initially explored in [11] in the context of classification, is based on the ideas of local smoothing and multi-resolution analysis from non-parametric statistics and signal processing.

By locally averaging the word histograms at different document locations, we obtain a local version of the global histogram $\gamma^{\text{hist}}(y)$ describing the local word distribution. Viewed geometrically, the local averaging embeds the original sequence as a smooth curve in the space of histograms over V. This curve summarizes the progression of semantic and statistical trends within the document. The smooth nature of the curve enables the use of a wide range of tools from differential geometry and smooth analysis. By varying the amount of local averaging we obtain a family of sequential representations possessing different sequential resolutions or scales. Low resolution representations capture topic trends and shifts while ignoring finer details. High resolution representations capture the fine sequential details but make it difficult to grasp the general trends within the document.

The rest of the paper is organized as follows. Section 2 reviews related work and Section 3 presents the locally weighted bag of words framework for smooth embeddings of categorical time series. In Section 4 we explore various visualization techniques based on the new framework and present a publicly available visualization toolkit. We then proceed in Section 5 to describe a user study demonstrating the usefulness of the new visualization techniques, followed by a discussion in Section 6. We focus in this paper on demonstrating our ideas in the context of document visualization. However, both the visualization techniques and the toolkit are also applicable to visualizing biological sequences such as proteins or DNA sequences.

2 Related Work

Document visualization has recently gained considerable interest due to the recent increase in the size, number and accessibility of text archives. A partial list of references is [19, 5, 6, 3, 22, 2] with additional references available in [21]. A selection of software systems for visualizing text corpora are IN-SPIRE[1], Enron-exploration tools[2], Thomson's refviz[3], and the Science topic browser[4]. Most of the methods mentioned above as well as other ones mentioned in [21] are designed for non-sequential visualization of a corpus of documents by considering the distance relation between individual documents. Such visualization methods are useful tools in browsing vast textual archives and nicely augment automatic search methods.

In contrast to the above methods, our approach aims at visualizing the sequential semantic progression within a single document. Related work concerning sequential analysis of a single document includes TextTiling [7, 8] which partitions the data into multi-paragraph segments based on the local word histogram. Similarity scores between local word histograms at different document locations are used to segment the document into "text tiles" which were found to correspond well to subtopic boundaries according to human judgement. Similar ideas are explored in [17, 18, 24], in the statistical text segmentation literature e.g. [1] and in the text summarization literature e.g. [20].

Multi-resolution analysis provides a convenient mechanism for the sequential visualization of documents at several levels of granularity. The modern multi-resolution analysis of data is inspired by the short time Fourier transform and wavelet representation [13]. An interesting application of multi-resolution wavelet analysis to document browsing is the Topic-Islands technique [15]. Topic-Islands constructs a digital signal that corresponds to the text document and then proceeds to compute its discrete wavelet transform. Multi-resolution visualization is then carried out by visualizing the energy of the various wavelet coefficients.

Our approach has several unique properties that distinguish it from the related research described above and other studies. The categorical sequences are represented as smooth curves in the histogram space which may be conveniently studied in several resolutions by varying the degree of smoothing. The representation includes the original categorical sequence $y = \langle y_1, \ldots, y_N \rangle$ and the word histogram $\gamma^{\text{hist}}(y)$ as special cases. Interpolating between these two extremes provides the flexibility of viewing sequential details at the desired resolution in a simple and effective manner.

The smoothness of the representation enables the use of tools from differential geometry and smooth analysis such as gradient, curvature, differential operators, and phase diagrams. These tools are used to create new visualization techniques for sequential trends in documents. Another interesting aspect of the proposed framework is that it draws interesting connections between document visualization and the considerable literature of local fitting [12] and functional data analysis [16] in statistics.

[1] http://in-spire.pnl.gov
[2] http://jheer.org/enron, http://enron.trampolinesystems.com
[3] http://www.refviz.com
[4] http://www.cs.cmu.edu/~lemur/science

3 The Locally Weighted Bag of Words framework

The locally weighted bag of words (lowbow) representation was introduced in [11] for the purpose of improving text document classification. In this section, we review its definition and some of its important properties before proceeding to discuss its applications to visualization in the next section. The description below applies to 1-gram or word histogram and is a slightly simplified version of the original definition in [11]. Its extension to the locally weighted *n*-grams is straightforward.

Since there is no clear ordinal relationship between words in the vocabulary, it is natural to represent words as basis vectors $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^{|V|}$ rather than integers. Viewed in this way, a document corresponds to a sequence of vectors in $\mathbb{R}^{|V|}$ representing its sequential contents. For example, a document $y = \langle 2, 2, 1, 1, 2 \rangle$ over a vocabulary $V = \{1, 2, 3\}$ corresponds to

$$\langle e_2, e_2, e_1, e_1, e_2 \rangle = \left\langle \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\rangle. \quad (3)$$

The locally weighted bag of words representation assigns local word histograms to different locations in the document. It is obtained from (3) by turning the discrete document location quantity into a continuous quantity $\{1, \ldots, N\} \mapsto [0, N]$ and smoothing the vector sequence (3) over the continuous quantity.

Definition 1 *Given a sequence* $y = \langle y_1, \ldots, y_N \rangle$, $y_i \in V$, *the continuous-time analog of* (3) *is the function*

$$\delta_y : [0, N] \times V \to [0, 1] \qquad \delta_y(t, j) \stackrel{\text{def}}{=} \delta_{j, y_{\lceil t \rceil}} \quad t \in [0, N], j \in V \quad (4)$$

where $\lceil t \rceil$ *is the smallest integer larger than or equal to* t *and* $\delta_{a,b}$ *is given by* (2).

Locally averaging temporal variations in δ_y provides sequential smoothing across the continuous document locations. It is obtained by convolving δ_y with a smoothing kernel whose shape determines the amount of smoothing and the resulting sequential resolution.

Definition 2 *A smoothing kernel is a positive valued function* $K_{\mu,\sigma} : [0, N] \to \mathbb{R}_+$ *parameterized by a location parameter* $\mu \in [0, N]$ *and a scale parameter* $\sigma > 0$. *We also assume that* $K_{\mu,\sigma}(t)$ *is smooth in* μ, t *and is normalized i.e.,* $\int K_{\mu,\sigma}(t)\, dt = 1$.

A standard example for a smoothing kernel $K_{\mu,\sigma}(t)$ is the Gaussian or normal distribution with mean μ and variance σ^2: $K_{\mu,\sigma}(t) = C \exp(-(t-\mu)^2/(2\sigma^2))$ where C ensures normalization over the range $[0, N]$. The parameter μ represents the mode or maximum and σ represents the spread or concentration around μ which determines the amount of local smoothing.

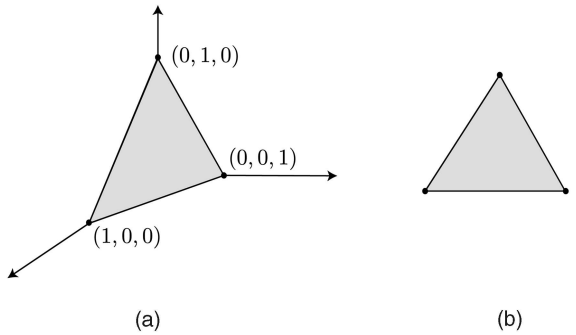
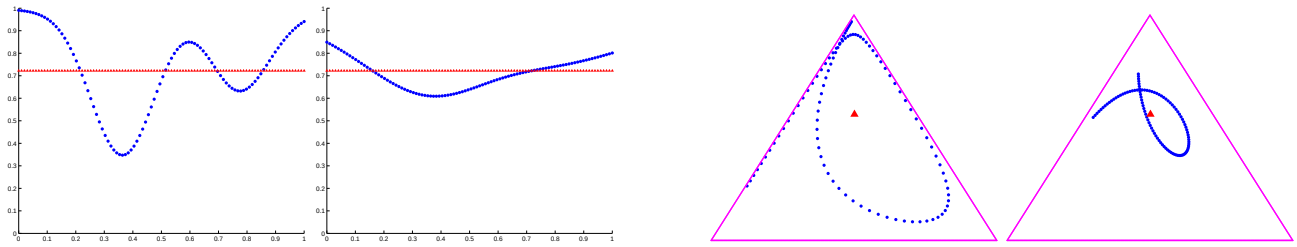


Fig. 1. The set $\mathbb{P}_{\{1,2,3\}} \subset \mathbb{R}^3$ may be visualized as a surface in $\mathbb{R}^3$ (left) or as a triangle in $\mathbb{R}^2$ (right).



(a) (b)

Fig. 2. (a) The curve $\gamma^{(\sigma)}(z)$ in $\mathbb{P}_{\{1,2\}}$ is visualized by graphing $[\gamma_\mu^{(\sigma)}(z)]_1$ as a function of μ/N. (b) The curve $\gamma^{(\sigma)}(w)$ in $\mathbb{P}_{\{1,2,3\}}$ is visualized by graphing it on the 2-D triangular space representing $\mathbb{P}_{\{1,2,3\}}$ (see Figure 1). In both cases $\sigma/N = 1/10$ (left) and $\sigma/N = 2/10$ (right) were used. Increasing σ causes the curves to be more smooth and to shrink towards the red line (a) or red triangle (b) which denote the histograms or the degenerate curves $\gamma^{(\infty)}(z), \gamma^{(\infty)}(w)$.

Definition 3 *The locally weighted bag of words (lowbow) representation of a sequence y, at time μ and scale σ, is the vector $\gamma_\mu^{(\sigma)}(y) \in \mathbb{R}^{|V|}$*

$$[\gamma_\mu^{(\sigma)}(y)]_j \stackrel{\text{def}}{=} \int_0^N \delta_y(t,j)\, K_{\mu,\sigma}(t)\, dt \qquad j = 1,\ldots,|V|. \tag{5}$$

It is easy to verify that the vector $\gamma_\mu^{(\sigma)}(y)$ is a probability distribution over V representing the locally weighted word histogram at the document location μ. Furthermore, it can be shown that $\gamma_\mu^{(\sigma)}(y)$ is a continuous and smooth function in μ (for details see [11]). As a result of these two observations, we conclude that the lowbow representation $\gamma^{(\sigma)}(y) \stackrel{\text{def}}{=} \{\gamma_\mu^{(\sigma)}(y) : \mu \in [0,N]\}$ of y is in fact a smooth curve in the space $\mathbb{P}_V$ of histograms or probability distributions over V

$$\mathbb{P}_V \stackrel{\text{def}}{=} \left\{ \theta \in \mathbb{R}^{|V|} : \quad \forall i\ \theta_i \geq 0, \quad \sum_{j=1}^{|V|} \theta_j = 1 \right\}.$$

By increasing σ, more local smoothing is applied and the resulting curve $\gamma^{(\sigma)}(y)$ becomes smoother and more slowly varying. In the limit of $\sigma \to \infty$, the kernel becomes a constant function $\forall \mu$, $\lim_{\sigma\to\infty} K_{\mu,\sigma}(t) \equiv 1/N$ and the lowbow curve contracts into a single point which turns out to be equal to the word histogram $\gamma^{\text{hist}}(y)$ defined in (1)

$$\lim_{\sigma\to\infty} \gamma_\mu^{(\sigma)}(y) = \gamma^{\text{hist}}(y) \qquad \forall \mu \in [0,N]. \tag{6}$$

On the other hand, reducing σ reduces the amount of sequential smoothing making the lowbow curve more wiggly. At the extreme case of $\sigma \to 0$ the lowbow curve loses its continuity and reverts to δ_y which is equivalent to the original sequence $y = \langle y_1,\ldots,y_N\rangle$. In general, varying σ changes the spread of $K_{\mu,\sigma}$ around its mode μ thus controlling the desired amount of smoothing or sequential resolution. Choosing a large σ reveals the general sequential behavior of the document while smoothing away the finer details. Choosing a small σ increases the sequential resolution which makes finer details available - but doing so may obscure the more general and important sequential trends. The choice of sequential resolution reflects the bias-variance statistical trade-off and ultimately depends on the precise data analysis goal. As we demonstrate in later sections, in some cases there is a need to analyze a document by considering its lowbow curves at multiple scales or resolutions.

We illustrate the construction of the lowbow curve by considering a couple of artificial documents with a small vocabulary: $z = \langle 1,1,1,2,2,1,1,1,2,1,1\rangle$ and $w = \langle 1,3,3,3,2,2,1,3,3\rangle$ over $V = \{1,2\}$ and $V = \{1,2,3\}$ respectively. $\gamma^{(\sigma)}(z)$ is a curve in the space of distributions $\mathbb{P}_{\{1,2\}}$. Since in this case $[\gamma_\mu^{(\sigma)}(z)]_2 = 1 - [\gamma_\mu^{(\sigma)}(z)]_1$ we can visualize the curve $\gamma^{(\sigma)}(y)$ by graphing $[\gamma_\mu^{(\sigma)}(y)]_1$ as a function of μ/N. Figure 2(a) illustrates these graphs for $\sigma/N = 1/10$ and $\sigma/N = 2/10$. In the second case, $\gamma^{(\sigma)}(w)$ is a curve in the space of distributions $\mathbb{P}_{\{1,2,3\}}$ which is a 2-D triangular subset of $\mathbb{R}^3$ demonstrated in Figure 1. Figure 2(b) visualizes the curve $\gamma^{(\sigma)}(w)$ for the same σ/N values in $\mathbb{P}_{\{1,2,3\}}$ after embedding it in two dimensions (see Figure 1). Notice how in both cases increasing σ causes the curve to be smoother and to contract into the red line (for z) or red triangle (for w) which denotes the degenerate curve $\gamma^{(\infty)}(z)$ or $\gamma^{(\infty)}(w)$ corresponding to the bag of words or global histogram $\gamma^{\text{hist}}(z), \gamma^{\text{hist}}(w)$.

Graphing the smoothed representation $\gamma^{(\sigma)}$ as in Figures 2(a), 2(b) is a convenient technique to visualize sequential variation at a particular resolution. By selecting an appropriate smoothing scale σ, important sequential patterns stand out and are uncluttered by patterns at lower resolutions. For example, in the case of Figure 2(a) and word sequence $\langle 1,1,1,2,2,1,1,1,2,1,1\rangle$ selecting a resolution of $\sigma/N = 1/10$ reveals the fact that there are actually two local minima representing high local concentration of 2 and one local maximum in between representing high local concentration of 1. Switching to $\sigma/N = 2/10$ we obtain only one broad local minimum and the additional details that were revealed before in the case of $\sigma/N = 1/10$ disappear.

The curve $\gamma^{(\sigma)}$ is easily visualized in the case of $V = \{1,2\}$ or $V = \{1,2,3\}$ as shown by the 1-D or 2-D figures described above. In the more interesting cases of text or biological sequences, V is a much larger set making $\gamma_\mu^{(\sigma)}(y)$ a high dimensional vector and $\gamma^{(\sigma)}(y)$ a curve in a high dimensional space. The next section deals with various techniques to visualize such high dimensional cases.

4 Visualization Techniques

As mentioned in the previous section, when $|V|$ is large, $\gamma^{(\sigma)} = \{\gamma_\mu^{(\sigma)} : \mu \in [0,N]\}$ is a smooth curve in a high dimensional space which cannot be directly visualized. In this section we introduce several techniques for visualizing such curves and demonstrate their effectiveness in the context of document visualization.

4.1 Visualizing Semantic Variation

Since the lowbow curve $\{\gamma_\mu^{(\sigma)} : \mu \in [0,N]\} \subset \mathbb{P}_V$ describes the movement of the local word histogram it may be used to visualize subtopic boundaries and other sequential shifts. Portions of the lowbow curve that remain more or less in the same region of $\mathbb{P}_V$ correspond to a semantically homogenous region within the document. On the other hand, portions of the lowbow curve containing drastic movements in $\mathbb{P}_V$ correspond to abrupt topic changes.

The first approach we take to visualize the movement of the lowbow curve is to summarize it using linear differential operators. Linear differential operators $L(f) = \sum_i \alpha_i D^i f$, commonly used in functional data analysis [16], represent a function f by a linear combination of its derivatives of varying orders. They are well defined on the lowbow curve due to its smoothness and are well suited to capture the instantaneous behavior of the curve.

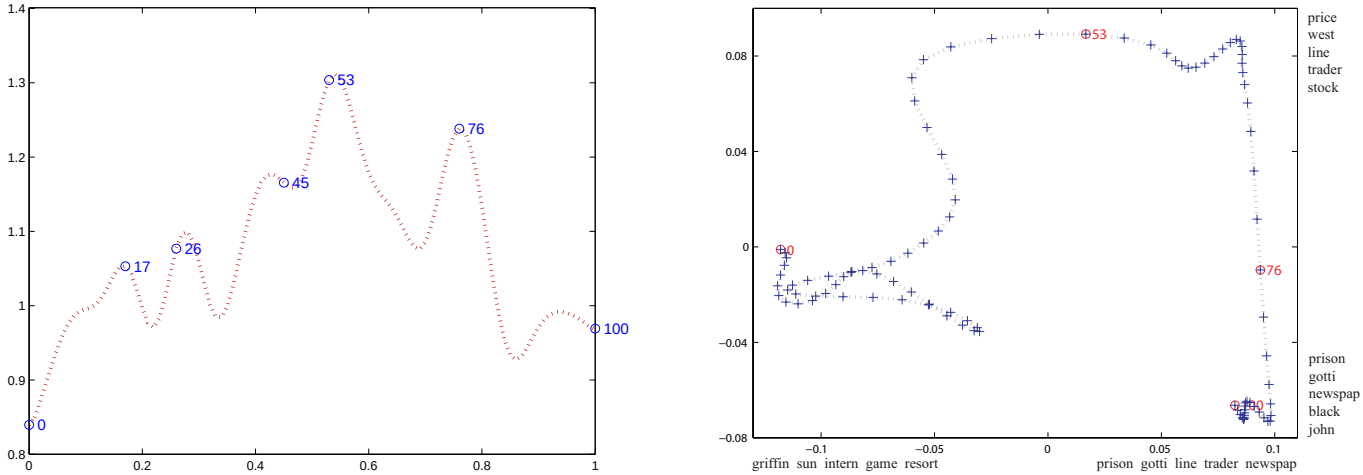


Fig. 3. Speed $\|\dot{\gamma}_\mu^{(\sigma)}\|$ (left, $\sigma/N = 0.064$) and 2D embedding using PCA (right, $\sigma/N = 0.026$) of the lowbow curve representing the three successive RCV1 stories of varying lengths as a function of μ/N. The words appearing to the right and bottom of the right panel correspond to the words whose projection on the principle components give the highest (or lowest) values providing a convenient visual summary for the user.

The simplest differential operator is the gradient vector whose components measure the rate of change in different dimensions. Applied to the lowbow curve, we define it as the following vector

$$\dot{\gamma}_\mu^{(\sigma)} \stackrel{\text{def}}{=} \left(\frac{d}{d\mu}[\gamma_\mu^{(\sigma)}]_1, \ldots, \frac{d}{d\mu}[\gamma_\mu^{(\sigma)}]_{|V|} \right) \in \mathbb{R}^{|V|}$$

whose components reflect the instantaneous change in the frequency of different words. The gradient vector is defined separately for each location μ and may be considered as a vector field along the lowbow curve pointing at the direction the curve is currently progressing in. It is easy to visualize the rate at which the lowbow curve moves at different locations μ by plotting the gradient norm $\|\dot{\gamma}_\mu^{(\sigma)}\|$ as a function of μ at different scales σ. Doing so reveals the presence of topic boundaries as well as areas representing homogenous and heterogenous semantic contents. In contrast to complex tools from the segmentation literature [1, 20], the gradient norm is simple, intuitive, and easy to visualize.

To illustrate the role of the gradient in document visualization we examine its behavior for documents containing clear and predetermined segments. Following [1], we consider a document created by concatenating individual news stories which resembles the continuous transcription of online news. We created it by randomly picking a news-wire chunk containing three successive news articles from the RCV1 collection[5] (document id: 4078, 4079 and 4080). The two internal segment borders occur at $\mu/N = 0.53$ and $\mu/N = 0.76$ with the first story obviously longer than the next two stories. The left panel of Figure 3 displays the gradient norm as a function of μ/N. The curve has several local maxima, but the largest two local maxima correspond almost precisely to the segment borders at $\mu/N = 0.53, 0.76$. The first three local maxima correspond to internal segment boundaries within the first story. Indeed, the first news story begins with the announcement of Sun International Hotels Ltd.'s acquisition of Griffin Gaming & Entertainment Inc.; it then switches to discuss the influencing factor behind Sun's decision at point $\mu/N = 0.17$ before switching again at $\mu/N = 0.26$ to talk about the benefit of the acquisition to Griffin. The story moves on at $\mu/N = 0.45$ to discuss the deal in detail. As with the different news story boundaries, the internal segment boundaries of the first story closely match the local maxima of the gradient norm.

As demonstrated above the first derivative of the curve conveys important information for the purpose of detecting areas of fast and slow semantic transition. In a similar way, the second derivative $\ddot{\gamma}^{(\sigma)}$ contains information concerning the curvature or the rate of change of the curve velocity. Regions in the curve corresponding to high or low second derivative or other linear differential operators could provide additional visual aid to a human observer.

The curve itself $\gamma^{(\sigma)}$ can be visualized by projecting it into its principal components or via multidimensional scaling. We compute the principal components (PCA) or the multidimensional scaling (MDS) of the curve $\gamma^{(\sigma)}$ based on a finite sample of points from the curve $\{\gamma_{\mu_1}^{(\sigma)}, \ldots, \gamma_{\mu_s}^{(\sigma)}\}$. Such visualization techniques would provide the user with smooth curves in 2-D or 3-D whose behavior reveals local semantic information. Portions of the curve corresponding to similar semantic components will typically contain similar local word histograms and therefore will naturally cluster together in both the high dimensional space $\mathbb{P}_V$ and in the low dimensional projections obtained by PCA or MDS.

Figure 3 (right) shows the 2D projection of the lowbow curve for the three concatenated RCV1 stories mentioned above. To embed the high dimensional curve in two dimensions we used principal component analysis. The blue crosses indicate the positions of the sampled points in the low dimensional embedding while the red circles correspond to the segment boundaries of the three documents. The sampled points in the figure $\{\gamma_{\mu_1}^{(\sigma)}, \ldots, \gamma_{\mu_s}^{(\sigma)}\}$ are naturally grouped into three clusters, indicating the presence of three distinct segments mentioned above. The distance between successive points near the segment boundaries is relatively large which demonstrates the high speed of the lowbow curve at these points. This is easily confirmed by examining the gradient norm in the left panel of Figure 3.

Another interesting visualization technique is to graph the projection of the curve $\gamma_\mu^{(\sigma)}$ on the top principal components separately as a function of the location μ/N. Figure 4 contains such plots for the concatenated document. The 3 panels correspond to the projection on the top three principal components (top panel represents the first component and bottom panel represents the third component). The words appearing to the right (left) of the panels correspond to the words whose projection on the components give the highest (lowest, respectively) values. The plots illustrate how the first principal component of the curve $\gamma^{(\sigma)}$ captures the variation between the first and the third stories and the second principal component of $\gamma^{(\sigma)}$ captures the variation between the second and the other two stories.

4.2 Visualizing Energy Transfer

An interesting visualization method that draws inspiration from harmonic motion in physics is to graph the first derivative vs. the second derivative of $\gamma_\mu^{(\sigma)}(y)$ with respect to μ. The resulting graph, called the phase-plane plot, describes the relationship between the velocity and

[5] http://trec.nist.gov/data/reuters/reuters.html

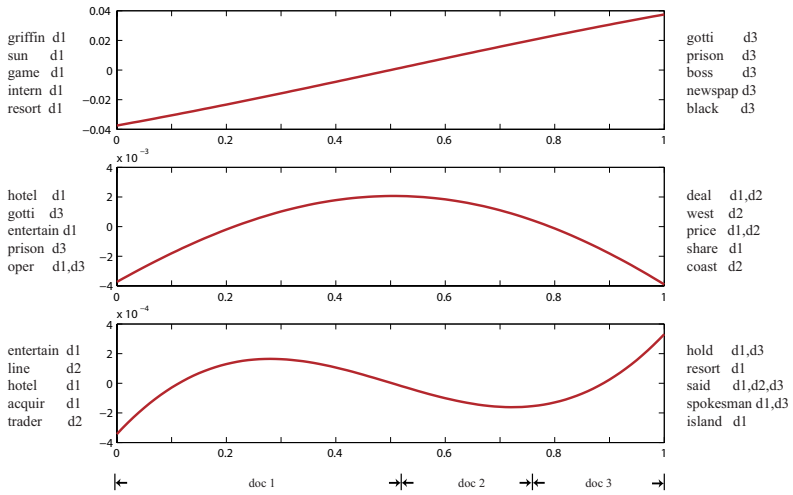


Fig. 4. Projection of the curve $\gamma^{(\sigma)}$ on its top three principal components for the three concatenated RCV1 stories. The numbers at the right hand side of a word give the word's position in the concatenated documents, e.g. the word "hotel" only appears in the first document.

acceleration of the curve $\gamma^{(\sigma)}$ as a function of the location parameter. Harmonic motion in physics, such as the movement of a simple pendulum or a spring behaves according to the differential equation

$$\ddot{x}(t) = cx(t). \tag{7}$$

It is easy to see that in this case, velocity $\dot{x}(t)$ and acceleration $\ddot{x}(t)$ behave as a sine and a cosine function respectively and drawing $\dot{x}(t)$ vs. $\ddot{x}(t)$ for various t results in the circle displayed in Figure 5 (left). The intersections of the circle with the x and y axes represent extreme points of the kinetic and potential energy. The sum of both energies is constant when no external force is present, but the relative proportion of them varies depending on the position $x(t)$ of the harmonic motion.

The curve in Figure 5 (right) corresponds to a phase-plane diagram of the first and second derivative of $\gamma_{\mu}^{(\sigma)}(y)$ projected on the third principal component. The striking resemblance between the two phase-plane plots indicates that the projection of $\gamma^{(\sigma)}$ on its third principal component satisfies the harmonic motion differential equation (7). Stated differently, we have that the curve $\gamma^{(\sigma)}$ approximately satisfies a partial differential equation $\langle D^2\gamma^{(\sigma)}, v_3\rangle = c\langle\gamma^{(\sigma)}, v_3\rangle$ where v_3 is the third principal component of the curve.

4.3 Multi-resolution Visualization of Data Content

We have already seen how the shape or resolution of the curve is influenced by different smoothing scales σ (recall Figures 2(a) and 2(b)). In this section we examine the simultaneous visualization of multiple curves $\gamma^{(\sigma_1)}(y),\ldots,\gamma^{(\sigma_r)}(y)$ representing the same document at varying resolutions or scales $\sigma_1 < \cdots < \sigma_r$. Such visualization provides a simple yet effective way to summarize a document at various levels of detail.

Our multi-resolution visualization is based on constructing a sequence of documents $y^{(\sigma_1)},\ldots,y^{(\sigma_r)}$ which are reduced resolution versions of y. The words of the reduced resolution document $y^{(\sigma)}$ are taken to be the most prominent components of the lowbow curve in the appropriate scale $\gamma^{(\sigma)}(y)$. More specifically, we define the reduced resolution document to be

$$y_{\mu}^{(\sigma)} \stackrel{\text{def}}{=} \arg\max_{j\in V} [\gamma_{\mu}^{(\sigma)}(y)]_j. \tag{8}$$

The construction $\gamma^{(\sigma)}(y) \mapsto y^{(\sigma)}$ in (8) may be considered as the approximate inverse to the lowbow mapping $y \mapsto \gamma^{(\sigma)}(y)$. In other words, $y^{(\sigma)}$ is the optimal estimate of the original document based on the lowbow curve $\gamma^{(\sigma)}(y)$ (perfect reconstruction is impossible since $y \mapsto \gamma^{(\sigma)}(y)$ is non-invertible). Increasing σ makes $y^{(\sigma)}$ more homogenous as it captures major sequential trends while discarding finer details. Reducing σ makes $y^{(\sigma)}$ more precise and detailed retrieving the original document y at the limit $\sigma \to 0$.

We demonstrate the above idea by constructing the multi-resolution representation $y^{(\sigma_1)},\ldots,y^{(\sigma_r)}$ corresponding to a recent news story from `reuters.com`.[6] The document was preprocessed by removing stop-words, punctuation marks and numbers, and stemming, resulting in a sequence of 272 lower-cased words. We arbitrarily set the length of the reduced resolution documents to be 51 and compute $y^{(\sigma)}$ for $\sigma/N = 0.04, 0.07, 0.5$.

We develop two techniques for visualizing the multi-resolution representation. Figure 6(a) demonstrates the color bar approach which displays the reconstructed documents $y^{(\sigma_1)},\ldots,y^{(\sigma_r)}$ as vertical bars side by side ordered in the same order as the scale values. Each vertical bar is split into colored segments describing the word content of the corresponding $y^{(\sigma_i)}$. Each word is assigned a different color and its starting position is indicated by a number at the left side of the color box, along with the word itself at the right side. The second visualization method, displayed in Figure 6(b), is inspired by the source code folding tool in integrated development environments such as Microsoft Visual Studio, Netbeans, or Eclipse. It displays the reconstructed document $y^{(\sigma_i)}$ as a list of words where each word acts as a button which can be expanded in a tree-like manner into the corresponding portion of the original y.

Both the color-bar and text folding designs enable the user to traverse the multi-resolution representation $y^{(\sigma_1)},\ldots,y^{(\sigma_r)}$ starting from a high σ representing a coarse summary of the original document y to lower values of σ representing finer temporal details. For the specific example described above, at the large scale $\sigma/N = 0.5$ we obtain the keywords of `US` and `Iran` which demonstrate the general topic of the article. As σ decreases, more details are added to the reduced resolution document diagram. At $\sigma/N = 0.07$, we get the added details concerning oil and gasoline. At the finer level of $\sigma/N = 0.04$, we learn additional details such as the involvement of Europe in the news story. Despite being a lossy representation, the reduced resolution document idea represents a convenient way to visualize temporal summaries of y at increasing levels of detail. The text folding design helps to effectively browse a document and quickly locate regions of interest using a tree-like search rather than the slower linear scan.

[6] `http://today.reuters.com/news/articleinvesting.aspx?type=hotStocksNews&storyID=2007-03-28T004723Z_01_SP116282_RTRUKOC_0_US-MARKETS-OIL.xml`

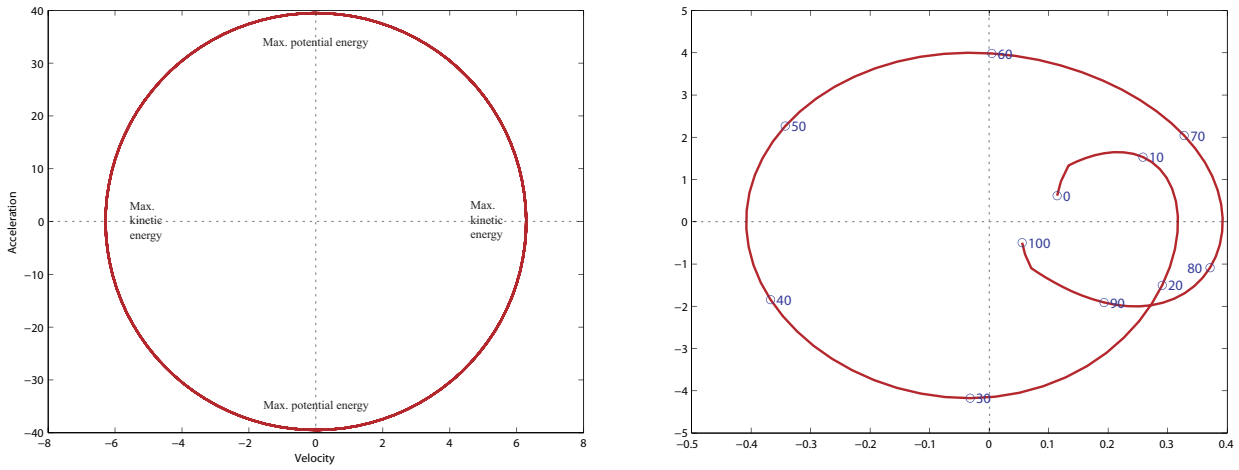


Fig. 5. Phase-plane plot for simple harmonic motion such as a simple pendulum or a spring (left) and for $\gamma_{\mu}^{(\sigma)}(y)$ representing two concatenated RCV1 stories (right).

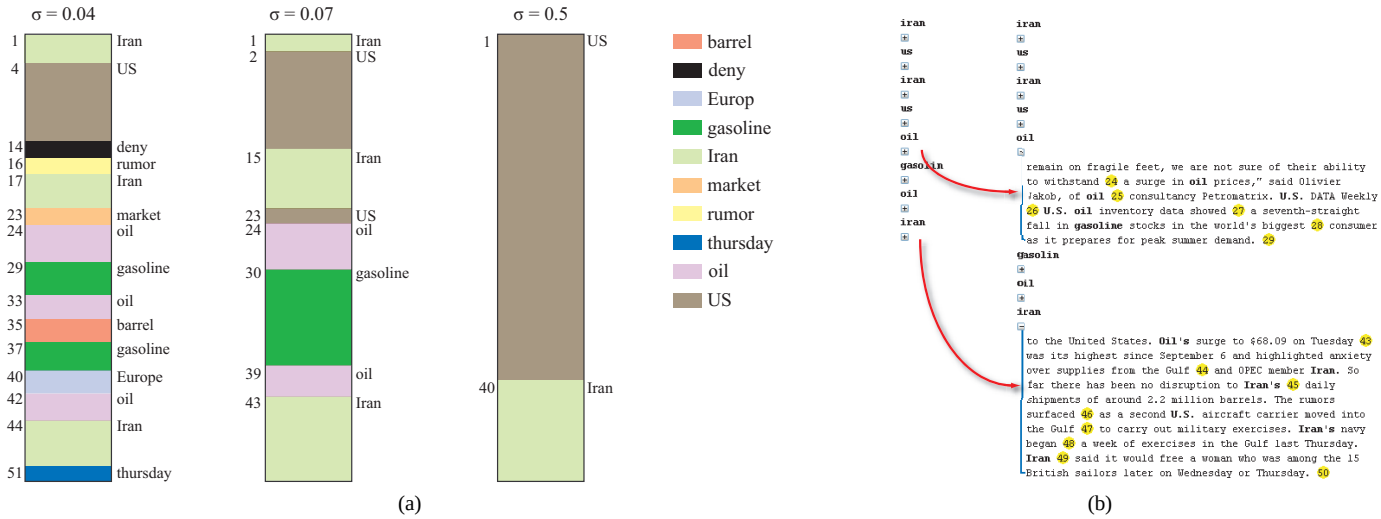


Fig. 6. Documents mapped from lowbow curves generated with three different values of σ/N. Some words are capitalized in the graph for improved readability. (a) Color bar approach. (b) Text folding approach.

4.4 Visualizing Multiple Documents

Thus far we were mostly concerned with visualizing a single sequence y. In this section we briefly explore visualizing the relative behavior of sequential trends in multiple sequences. As with other forms of sequential visualization, these differences are usually difficult to identify using the traditional n-gram approach. As an example, we compare the news story used in Section 4.3 with a somewhat similar news story.[7] The cosine similarity between the corresponding histograms or γ^{hist} vectors is very high - 0.9928. The projection of the two resulting curves $\gamma^{(\sigma)}(y_1), \gamma^{(\sigma)}(y_2)$ into the first two principal components is displayed in Figure 7. Note that in this case, in contrast to previous cases the PCA is computed based on sampled points from two curves rather than just one. The words that correspond most strongly and weakly to each principal component are displayed along the x and y axes.

The two curves share similar sequential trends most of the time, with the red curve diverging to a different region towards the end. Note also how both curves move from discussing political events concerning Iran and Britain at the left part of the x axis to discuss the effect on oil price and other economic factors at the right part of the x axis. The blue circles in Figure 7 indicate two overlapping regions between the documents. Actually, both documents at $\mu_1/N_1 \approx 0.5$ and $\mu_2/N_2 \approx 0.91$ talk about Iran's daily shipments of oil and at $\mu_1/N_1 \approx 0.95$ and $\mu_2/N_2 \approx 0.82$ talk about a strike by workers at the French Mediterranean oil terminal. Visualizing the relative sequential trends in more than 2 documents results in a cluttered graph containing many curves. An alternative is to use tools from functional data analysis such as functional principal component analysis and functional canonical correlation analysis [16].

4.5 Interactive Document Visualization Toolkit

To compare the different techniques and evaluate them we constructed an interactive document visualization and exploration toolkit. The toolkit, displayed in Figure 8 is implemented in Java and is publicly available[8]. It contains all the features described in this paper as well as a few additional ones.

As shown in Figure 8, the right panel of the toolkit displays a user specified text document, either with text folding or without. In its left panel the toolkit displays various graphical summaries of the resulting curve $\gamma^{(\sigma)}$ such as velocity or PCA and MDS projections. The

[7] http://today.reuters.com/investing/financeArticle.aspx?type=hotStocksNews&storyID=2007-03-28T184159Z_01_SP116282_RTRUKOC_0_US-MARKETS-OIL.xml

[8] http://web.ics.purdue.edu/~ymao/lowbow.htm

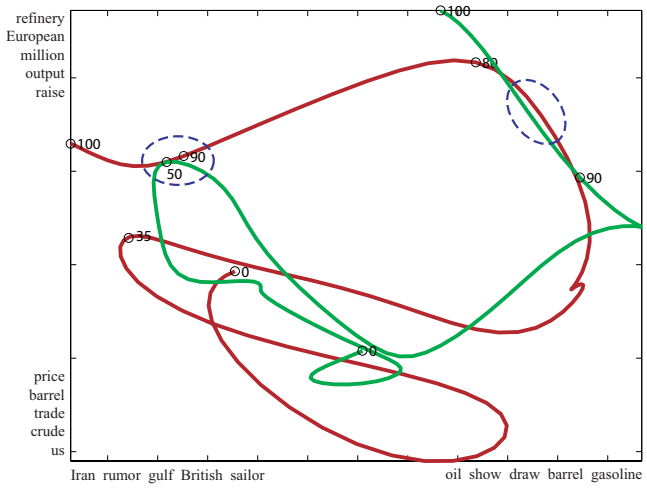


Fig. 7. Two dimensional PCA projection of the lowbow curves representing two similar news stories (see Section 4.4 for more details) using a VARIMAX rotation of PCA ($\sigma/N = 0.04$).

bottom panel contains various buttons that allow the user to change the resolution determined by the scale σ or to change the visualization technique. Some features of the lowbow curve such as the velocity can be visualized in the right panel by coloring the text words with appropriate colors. Other features included in the toolkit are visualization of linear differential operators and 3D graphical exploration of PCA and MDS projections using OpenGL.

Document preprocessing steps that need to be completed before using the toolkit are: removing stop-words, punctuation marks and numbers, removing casing, stemming, and constructing a vocabulary. The most computer intensive aspect of the preprocessing is the building of the vocabulary whose complexity depends on the length of the document. For all but extremely long documents, the vocabulary building as well as the rest of the preprocessing is extremely fast. Furthermore, all preprocessing tasks are standard and can be accomplished using a wide variety of publicly available text processing softwares. The complexity of constructing and maintaining the lowbow representation is completely determined by the number of sample points and the number of unique words in the document. More specifically it is equivalent to the number of sampled points times the complexity of compiling and maintaining the bag of words or word histogram.

5 User Study

To evaluate this toolkit we conducted an informal study consisting of 30 users each of whom were asked to use the toolkit to answer three types of questions. The users ranged in age from 18 to 55 years old with 1/3 of the users being female. Slightly more than 2/3 of the users were not native speakers of English with the majority of participants being graduate students from engineering or science departments.

Before attempting the study, a user was introduced to the toolkit by following a written, self-directed tutorial in which he or she was encouraged to experiment with varying scale values on a test document and observe its effect with respect to the velocity, the PCA projection and the reduced resolution document by text folding. The remaining features were disabled for the sake of brevity.

In the first of the three tasks, the users were asked to discern topic boundaries in eight passages containing two to four concatenated news stories from RCV1. In the second task, users were presented with four documents and a few topic phrases and were asked to find portions of a document that were best represented by the supplied phrases. The phrases were specifically chosen so that their constituent words either did not appear or rarely appeared within the documents. In the third and final task users were asked to identify five keywords for each of eight documents from a list of keyword candidates.

score	PCA	PCA components	velocity	text folding
task 1	1.60	1.63	4.40	3.07
task 2	1.63	1.60	1.9333	4.10
task 3	1.53	1.80	1.8667	3.40

Table 1. Average user scores representing the benefit users received from different visualization techniques with respect to the three types of questions they were asked. A score of 5 indicates the feature was very useful while 1 indicates not at all.

In all three tasks, half of the questions disallowed use of the toolkit while the remaining half encouraged, but did not require, its use. Users were required to finish each task within some time constraint, and indicate, at the end of the task, how useful they found the various visualization techniques in accomplishing the three tasks. The average score received by each feature-task pair is displayed in Table 1.

Our initial expectation was to receive high score for velocity in task 1, for PCA and PCA components in task 2, and for text-folding in task 3. Part of the user study results confirm our expectation, as indicated by the scores for (task 1, velocity) and (task 3, text folding) pairs. Surprising aspects of the results in Table 1 are that users found the text folding instructive in all three tasks, and they largely disregarded the PCA and PCA component plots.

Among all the visualization techniques, the velocity graph and the text folding technique are the most intuitive and simplest to understand. This explains the popularity of these two techniques in the user study. In particular, the text folding allowed users to quickly hone in on semantically meaningful words thus increasing the search efficiency as indicated by its effectiveness for tasks 2 and 3. Text folding is more powerful than a simple 'find' or pattern matching tool since the user does not have to specify all the synonyms of the words he is interested in. For example, in task 3, users using the text folding tool were able to quickly identify keywords relevant to a certain passage even if a keyword did not appear as one of the words in the passage itself. The benefit of the tree-like search in text folding and other multi-resolution methods over linear scan is similar to the complexity advantage of tree search $(O(\log n))$ over a linear search $(O(n))$.

The relatively low scores given by users to the PCA and PCA components techniques illustrate their difficulty in grasping fully the PCA concept and effectively utilizing it after only a short introduction and experimentation. The users were given only several minutes before the questions to experiment with the toolkit and most of them could not comprehend its potential right away. Our conclusion is that to fully comprehend and effectively use the less intuitive PCA and PCA components, users need more guidance into how to use it and spend more time (perhaps up to 1 hour) experimenting with it.

6 Discussion

The smoothed representation $\gamma^{(\sigma)}(y)$ is a promising new direction for visualizing categorical sequences such as text documents or biological data. By varying the smoothing scale σ, $\gamma^{(\sigma)}(y)$ interpolates between the original sequence $\gamma^{(0)}(y)$ or y and $\gamma^{(\infty)}(y)$ or $\gamma^{\text{hist}}(y)$. The equivalence between the $\gamma^{(\sigma)}(y)$ representation and smooth curves enables us to use techniques inspired by differential analysis and to use techniques invented in the statistical area of functional data analysis. In contrast to n-gram, $\gamma^{(\sigma)}(y)$ captures topical trends and incorporates long range as well as position information at the required level of sequential detail. On the other hand, the novelty of our idea is orthogonal to n-gram as it is possible to combine the two by constructing smoothed representation over n-grams rather than the word set V.

We demonstrated a number of useful visualization techniques based on curve derivatives, principal component analysis, multi-dimensional scaling, phase-plane analysis, reduced resolution documents and multi-resolution analysis. Based on our experimentation with the interactive toolkit as well as a user study we conclude that the highly intuitive velocity curves and text folding provide valuable assistance for users in absorbing textual information. Less intuitive techniques

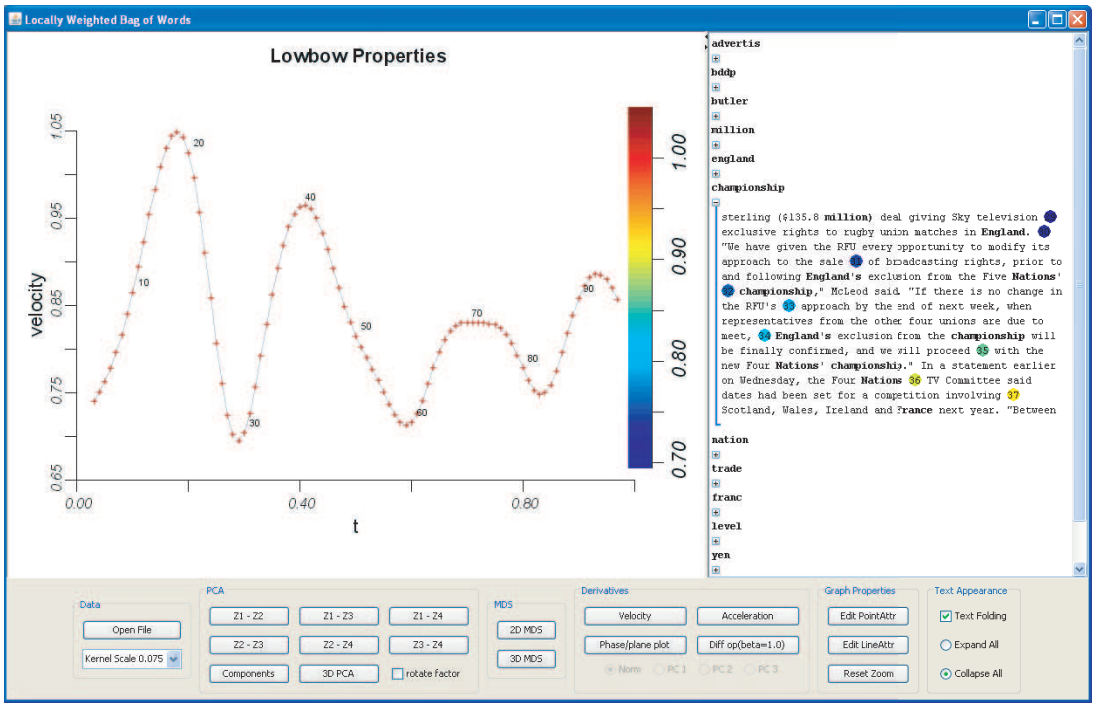


Fig. 8. Screen shot of the toolkit.

provide useful assistance to trained individuals. Other individuals, however, may require some amount of training or guidance.

Viewing the text sequence next to salient visual cues as in Figure 8, users are able to comprehend the text and localize their attention faster and more accurately. By controlling the sequential resolution of the visual cues users can efficiently traverse the document much like a binary search over a tree structure. While unlocking the full potential of the visualization framework may take some experimentation and guidance, it is worth it in the long run. People spend a decent portion of their time comprehending text and effective sequential visualization tools have the potential to make this process quicker and more efficient.

Acknowledgements

The authors wish to thank Jeff Bilmes for his interesting comments regarding the lowbow representation, and anonymous reviewers for their helpful suggestions on earlier drafts. This work was supported in part by NSF grant DMS-0604486.

References

[1] D. Beeferman, A. Berger, and J. D. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210, 1999.

[2] D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the Twenty-Third International Conference on Machine Learning*, pages 113–120, 2006.

[3] B. Fortuna, M. Grobelnik, and D. Mladenic. Visualization of text document corpus. *Informatica*, 29:497–502, 2005.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2003.

[5] S. Havre, E. Hetzler, K. Perrine, E. Jurrus, and N. Miller. Interactive visualization of multiple query results. In *IEEE Symposium on Information Visualization*, page 105, 2001.

[6] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[7] M. A. Hearst. Multi-paragraph segmentation of expository text. In *Association of Computational Linguistics*, pages 9–16, 1994.

[8] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[9] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets, timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[10] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163, 2005.

[11] G. Lebanon. Sequential document representations and simplicial curves. In *Proc. of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

[12] C. Loader. *Local Regression and Likelihood*. Springer, 1999.

[13] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[14] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[15] N. E. Miller, P. C. Wong, M. Brewster, and H. Foote. Topic islands - a wavelet based text visualization system. In *IEEE International Conference on Visualization*, pages 189–196, 1998.

[16] J. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, second edition, 2005.

[17] G. Salton, J. Allen, C. Buckley, and A. Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. *Readings in Information Retrieval*, pages 478–483, 1997.

[18] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic text decomposition using text segments and text themes. In *UK Conference on Hypertext*, pages 53–65, 1996.

[19] A. Spoerri. Infocrystal: A visual tool for information retrieval & management. In *International Conference on Information and Knowledge Management*, pages 11–20, 1993.

[20] S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[21] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path*. IEEE Computer Society, 2005.

[22] F. B. 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, 2006.

[23] M. Weber, M. Alexa, and W. Muller. Visualizing time-series on spirals. In *Proc. of IEEE Symposium on Information Visualization*, pages 7–14, 2001.

[24] P. C. Wong, W. Cowley, H. Foote, E. Jurrus, and J. Thomas. Visualizing sequential patterns for text mining. In *Proc. of IEEE Symposium on Information Visualization*, pages 1–5, 2000.