PhC: Multiresolution Visualization and Exploration of Text Corpora with Parallel Hierarchical Coordinates

K. SELÇUK CANDAN, Arizona State University
LUIGI DI CARO and MARIA LUISA SAPINO, University of Torino

The high-dimensional nature of the textual data complicates the design of visualization tools to support exploration of large document corpora. In this article, we first argue that the Parallel Coordinates (PC) technique, which can map multidimensional vectors onto a 2D space in such a way that elements with similar values are represented as similar poly-lines or curves in the visualization space, can be used to help users discern patterns in document collections. The inherent reduction in dimensionality during the mapping from multidimensional points to 2D lines, however, may result in visual complications. For instance, the lines that correspond to clusters of objects that are separate in the multidimensional space may overlap each other in the 2D space; the resulting increase in the number of crossings would make it hard to distinguish the individual document clusters. Such crossings of lines and overly dense regions are significant sources of visual clutter, thus avoiding them may help interpret the visualization. In this article, we note that visual clutter can be significantly reduced by adjusting the resolution of the individual term coordinates by clustering the corresponding values. Such reductions in the resolution of the individual term-coordinates, however, will lead to a certain degree of information loss and thus the appropriate resolution for the term-coordinates has to be selected carefully. Thus, in this article we propose a controlled clutter reduction approach, called *Parallel hierarchical Coordinates* (or *PhC*), for reducing the visual clutter in PC-based visualizations of text corpora. We define visual clutter and information loss measures and provide extensive evaluations that show that the proposed PhC provides significant visual gains (i.e., multiple orders of reductions in visual clutter) with small information loss during visualization and exploration of document collections.

Categories and Subject Descriptors: H.3.1 [**Information Search and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Search and Retrieval**]: Text Mining; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; I.7 [**Document and Text Processing**]:

General Terms: Algorithms

Additional Key Words and Phrases: Document set visualization, parallel coordinates, clutter reduction

ACM Reference Format:
Candan, K. S., Di Caro, L., and Sapino, M. L. 2012. PhC: Multiresolution visualization and exploration of text corpora with parallel hierarchical coordinates. ACM Trans. Intell. Syst. Technol. 3, 2, Article 22 (February 2012), 36 pages.
DOI = 10.1145/2089094.2089098 http://doi.acm.org/10.1145.2089094.2089098

1. INTRODUCTION

Today, text is being produced and consumed in a wide variety of applications, including science, news, e-commerce, blogs, and social networking sites. This flood of data in

This work was partially supported by NSF grant no. 1016921 "One Size does not Fit All: Empowering the User with User-Driven Integration".
Authors' addresses: K. S. Candan (corresponding author), Department of Computer Science and Engineering, Arizona State University, AZ; email: candan@asu.edu;rec; L. Di Caro and M. L. Sapino, Department of Computer Science, University of Torino, Italy.

DOI 10.1145/2089094.2089098 http://doi.acm.org/10.1145/2089094.2089098

(a) a term- or tag-cloud sample (b) a tag-flake sample[Di Caro et al. 2008]

Fig. 1. Examples of existing text collection vizulization schemes: (a) a (term or) tag cloud; (b) tag-flake visualization proposed in Di Caro et al. [2008].

textual form brings forth a need for effective ways to visualize and analyze large text collections.

The need for informed navigation within large text collections has been highlighted in the literature [Bates 1989], but effective solutions are elusive. Most commonly used text visualization tools, such as term- or tag-clouds (Figure 1) have significant limitations. A tag, whether provided by the user or extracted from the textual content itself, provides an easy way to search and index blogs and other online media and documents. For example, most visualizations of tag (or keyword) clouds vary the sizes of the fonts to differentiate most important tags from those that are less important (Figure 1(a)). While this helps users quickly observe the most frequent terms in a text collection, this representation falls short in making the context in which these terms/tags co-occur apparent. While some existing text collection visualization schemes, such as tag-flakes [Di Caro et al. 2008] and ContexTour [Lin et al. 2010], attempt to visually organize the terms extracted from the text collection in a way that highlights co-occurrences of tags in that collection, or reflects the underlying community clusters, even these fail to make it possible for a user to understand how the documents are distributed in the inherently multidimensional space with respect to a set of relevant terms.

1.1. Visualization of Multidimensional Text Collections Using Parallel Coordinates

Complex, multidimensional data do not fit well into 2D screens. Consequently, a key challenge in multidimensional data visualization is to map data on 2D graphical displays in a way that preserves the underlying information and is easy to view and explore. Existing techniques, including Parallel Coordinates (PC) [Inselberg and Dimsdale 1990], radial visualization [Hoffman et al. 1997], circle segments [Ankerst et al. 1996], heat maps [Eisen et al. 1998], and treemaps [Shneiderman 1992], all address this challenge differently.

The Parallel Coordinates (PC) technique [Inselberg and Dimsdale 1990], for example, is based on the following data mapping strategy: the dimensions of the data (e.g., the attributes of a data relation) are represented as (often equispaced) vertical parallel lines or *parallel coordinates*. Each distinct value in the domain of a given data dimension corresponds to a point on the corresponding vertical line. Given a dataset, each multidimensional data element (e.g., a tuple in a relation) is converted into a poly-line or a continuous curve which passes through the corresponding value points on the vertical coordinate lines. By mapping data elements onto the 2D space in such a way that similar elements are represented as similar poly-lines or curves on this space, PC is able to help users discern dominant patterns in the multidimensional data. For

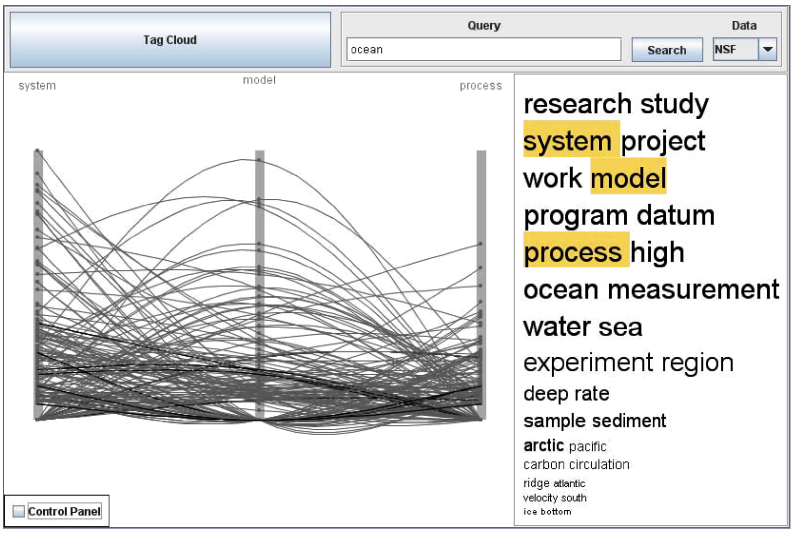


Fig. 2. PC-based visualization of a set of NSF award abstracts matching the query "ocean"; the visualization is with respect to the three user-selected terms "system", "model", and "process". This parallel coordinates visualization uses curves instead of poly-lines.

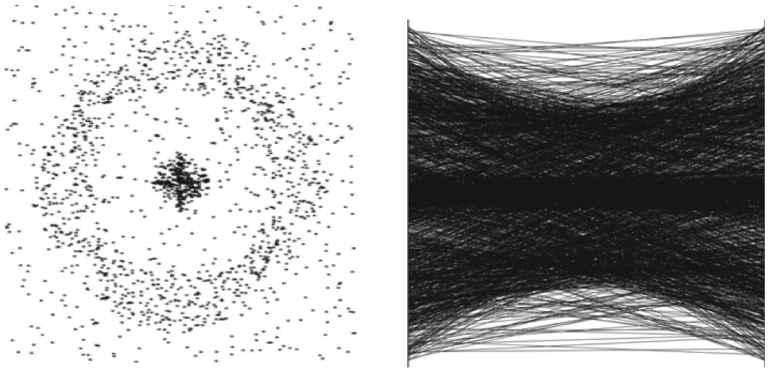
instance, a cluster of objects in the multidimensional space will appear as a dense cluster of lines that have similar flows in the 2D space.

In this article, we first note that parallel-coordinates-based visualization is suitable for helping the users observe the distribution of the data within a text collection and observe the underlying patterns. Figure 2 presents an example. Here, the user is analyzing a collection of NSF award abstracts, within the context of the query keyword "ocean". At the right-hand side of the interface, the underlying tag cloud is presented to the user. In this example, the user has selected three of the terms, "system", "model", and "process" for further analysis and the system mapped the matching award documents onto the parallel coordinates based on corresponding term frequencies[1]. This visualization, for instance, helps the user observe that, within this context, those documents that contain the term "system" frequently contain the term "model" very rarely.

1.2. Challenge: Visual Clutter in Parallel Coordinates

We, however, also note that, despite its ability to map multidimensional data to 2D visualization space, PC-based visualization also faces some challenges. In particular, the reduction in the dimensionality during the mapping of many-dimensional points to 2D lines often implies visual clutter and this visual clutter is obvious in Figure 2. Visual clutter in PC occurs because data clusters that are separate in the original space may end up overlapping with each other in the 2D space, making them harder to distinguish. These overlaps often appear as crossings of the edges in the visualization or unnecessarily dense regions of the 2D space (Figure 3). When overly dense, such

[1]Note that, while in the rest of the article, we simply refer to "*terms*" and "*term frequencies*", in practice a latent analysis algorithm, such as the standard LSA [Deerwester et al. 1989, 1990] or Latent Dirichlet Allocation [Song et al. 2009], may be used to identify latent semantic dimensions or significant topics that can be used as visualization coordinates. In these cases, instead of term frequencies, we would use topical document relevance (e.g., document vectors' similarities to the latent semantic dimensions or the selected topics) as coordinate values.



(a) Points in the original space (b) PC visualization

Fig. 3. (a) The two clusters that are separable in the original space (b) overlap with each other in PC, making them harder to distinguish (example taken from Yuan et al. [2009]).
.

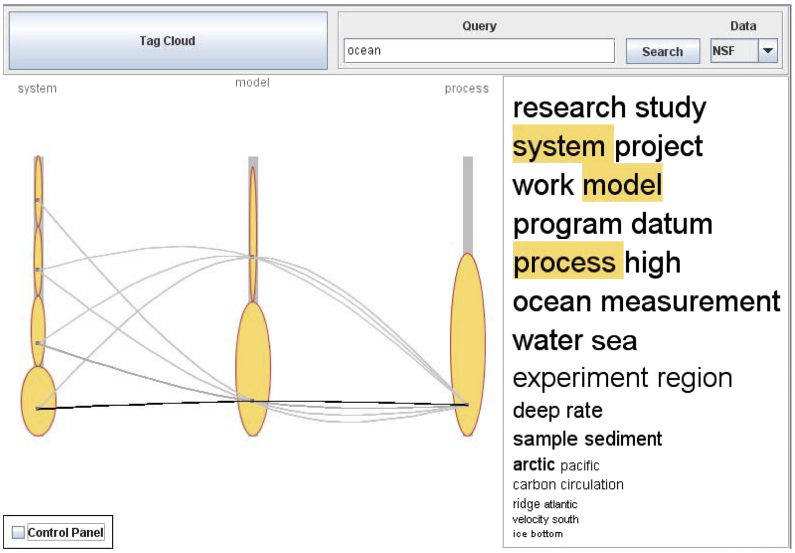


Fig. 4. Reducing the visual clutter in the PC visualization in Figure 2.

crossings can create significant visual clutter, rendering the patterns in the data hard to discern.

In this article, we focus on the reduction of the visual clutter in PC for obtaining cleaner and easier to interpret visualizations of text collections. Figure 4 shows that one way to achieve this is to cluster the values along the parallel coordinates: this helps reduce the visual clutter in Figure 2 and helps highlight the data patterns. In this visualization, each ellipse corresponds to a cluster of values along a tag-coordinate and the width of the ellipse corresponds to the amount of curves (i.e., documents) that fall into that cluster of values. The curves cross the clusters (not at the centers of the ellipses, but) at the cluster centroids. Note that, in the alternative visualization in Figure 4, the data patterns are visible with ease and this visualization imposes less visual load than the one in Figure 2.

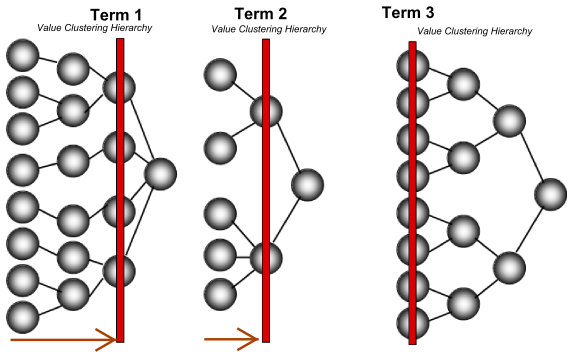


Fig. 5. For each term dimension selected by the user, PhC selects a visualization granularity (or *degree of value clustering*).

It is, however, important to note that reducing the detail of the data in the individual coordinates may lead to certain degree of information loss. In fact, given the same data while some value clusterings may help improve the visualization, some others may collapse distinct patterns and make them impossible to distinguish. Thus, which values are clustered and how much clustering is applied to each dimension have to be selected carefully.

1.3. Parallel Hierarchical Coordinates (PhC): Controlling the Information Loss in the Visualization of Large Text Collections

In this article, we present the *Parallel hierarchical Coordinates* (*PhC*) approach to multiresolution visualization of large text corpora. PhC relies on a parallel coordinates (PC)-based mapping, where user-selected visualization terms are represented as vertical parallel coordinates and each document in the dataset is drawn as a continuous curve which passes through the corresponding term frequency values on the vertical coordinate lines. In the resulting visualization, patterns in the document collection appear as dense regions.

While we argue that PC-based visualizations can be effective in text document collection visualization and analysis, in this article, we also note that naive PC-based visualizations often suffer significant visual clutter and propose the *Parallel hierarchical Coordinates* (*PhC*) approach to alleviate this problem.

(1) PhC takes as input a document collection and produces a tag cloud from which the user selects the set of terms to be used for analysis.
(2) Next, for each term a *value clustering hierarchy* is created using a hierarchical value clustering algorithm, such as Ward [1963], applied to the (term frequency) values corresponding to the coordinate.
(3) Given a value clustering hierarchy for each of the term-coordinates and a user-provided target resolution, PhC, then uses these value clustering hierarchies to decide how much detail to maintain for each of the term-coordinates (Figure 5) to reduce the visual clutter while maintaining as much information as possible.
(4) The user can then navigate over the resulting multiresolution parallel coordinate space to understand the distribution of the available documents within the term space. To support this, PhC provides support for OLAP-like navigational operators, such as *drill-down*, where the user can navigate from a more general view to a more detailed view, by stepping down on a given hierarchy, and *roll-up*, which lets the user increase the amount of aggregation and clustering by climbing up a hierarchy.

The article is organized as follows: In Section 2, we discuss the related work. In Section 3, we describe the utilization of Parallel hierarchical Coordinates (PhC) for visualizing large text documents. In Section 4, we describe how PhC helps control the information loss during visual clutter reduction. In this article, in addition to proposing a novel multiresolution PC technique for visual clutter reduction in visualization of text document collections, we also present novel visual clutter and information loss measures for PC-based visualizations (Section 5). Evaluation results, presented in Sections 6, show that PhC provides significant visual gains (i.e., clutter reduction) with small information loss.

2. RELATED WORK

In this section, we present an overview of the related work in the areas of document collection visualization as well as discuss prior work on parallel coordinates.

2.1. Related Work on Document Collection Visualization

A critical aspect of the knowledge discovery process is the presentation layer, that is, information visualization and browsing. Effective and interactive designs can empower users, on the other hand, an ineffective design may cause the user be lost in a sea of information.

The image hosting Web site Flickr[2] was one of the first systems that used *tag clouds* for visualizing lists of words, associated to a given media object, where word importance is represented with font sizes. Hassan-Montero and Herrero-Solana [2006a] propose a graphical visualization of tag clouds, where the tags are selected on the basis of their frequency of use. Relationships among tags are defined in terms of their similarity, quantified by means of the Jaccard coefficient. K-means clustering is then applied on the tag similarity matrix, with an a priori chosen number of clusters and fixed number of selected relevant tags. Hassan-Montero and Herrero-Solana [2006b] further the work in Hassan-Montero and Herrero-Solana [2006a] by applying Multi Dimensional Scaling, *MDS*, (using Pearson's correlation as the tag similarity function) to create a bidimensional space, which is then visualized through a fish-eye system. We note, however, that while preserving distances, MDS does not preserve the energies of the input tags. Similarly, using only two dimensions can be overly lossy. Research on effective use of 2D spaces for multidimensional data visualization focuses on careful selection of the relevant dimensions [Seo and Shneiderman 2004] and organizing data in hierarchical visualization structures, such as TreeMaps, along the relevant dimensions, and mapping these two 2D spaces [Chintalapani et al. 2004].

PhaseTwo's goal is to create *visually pleasant* tag clouds, by presenting tags in the form of seemingly random collections of circles with varying sizes[3]: the size of the circle denotes its frequency. Each tag circle is first placed in the center of the cloud and then fired from the center along a random angle. The tag circle stops when it collides with another circle. This visualization scheme intentionally randomizes the placement of the tags with the hope of projecting a more pleasant (if not highly informative) feeling.

In Fortuna et al. [2005] the authors describe a system to visualize the semantics contained in a textual corpus. They rely on a Latent Semantic Analysis technique to extract the information about the principal dimensions emerging from the text by means of Singular Value Decomposition (SVD) applied on the term-document frequency matrix [Eckart and Young 1936]. They automatize the choice of the concept space dimensions, by choosing the minimum k such that $\frac{\sum_{i=1}^{k} S_i}{\sum_{i=1}^{n} S_i} > \theta$, where S_i are the

[2]http://www.flickr.com/

[3]http://phasetwo.org/post/a-better-tag-cloud.html

singular values on the diagonal of the singular value matrix, and θ is a system parameter threshold, set for example to 0.5. After identifying the relevant dimensions, in Fortuna et al. [2005], the authors apply MultiDimensional Scaling (*MDS*) [Cox and Cox 2001] to reduce the dimensionality all the way down to two dimensions, to allow the visualization on the 2D plane.

In Don et al. [2007] the authors address the problem of making text mining results more comprehensible to humanities scholars, journalists, intelligence analysts, and other researchers, using the *FeatureLens* system, that visualizes a text collection at several levels of granularity and enables users to explore interesting text patterns. The implementation has panels in which different document features could be focused and analysed. First of all, the authors focus on frequent itemsets of n-grams, so they capture the repetition of exact or similar expressions in the collection. Users can find meaningful co-occurrences of text patterns by visualizing them within and across documents in the collection. The system allows users to identify the temporal evolution of usage such as increasing, decreasing, or sudden appearance of text patterns. In order to provide repeated expressions that are exact repetitions as well as repetitions with slight variations, in Don et al. [2007] the authors propose to use frequent closed itemsets of n-grams, according to a specific definition of pattern: a set of n-grams X is a pattern if there exists no set of n-grams X' such that X' is a proper superset of X, and every paragraph containing X also contains X'. Don et al. [2007] also provide tooltips that show the paragraphs where some selected pattern occurs, making explicit their context of usage.

2.2. Related Work on Parallel Coordinates

As we mentioned in the Introduction, PC-based visualization faces visual clutter due to dimensionality reduction. In this section, we provide an overview of the prior attempts to tackle this challenge.

2.2.1. Dimension Ordering. In Peng et al. [2004], Peng et al. introduce a visual clutter measure for PC. They define visual clutter as the ratio of outliers in the visualization and recognize that the number of outliers can be reduced by careful ordering of the dimensions on the parallel coordinates visualization. Yang et al. [2003] also recognize that careful clustering and ordering of the dimensions (based on similarities) can help improve visualization effectiveness. In general, the optimal ordering of the dimensions is an NP-complete problem [Ankerst et al. 1998]. In this article, we assume that the appropriate order of dimensions is selected in advance, either by the user or by a dimension-ordering algorithm.

2.2.2. Data Filtering. An alternative approach for reducing visual clutter is to reduce the data themselves. Zhou et al. [2009] propose a *splatting*-based noise removal method which eliminates outlier poly-lines. The algorithm randomly selects a set of poly-lines and augments its and its neighbors intensities. Since outliers are not neighbors to many other poly-lines, the intensities of the outliers gradually become relatively lower and they disappear from the visualization. Artero et al. [2004] use data frequency for selecting the data to be presented. Instead of applying the reduction on the entire data visualization, Ellis and Dix [2006] propose a *sampling lens*, a movable region within which the data are presented in a down-sampled or reduced manner.

2.2.3. Data Clustering. Visual clutter can also be reduced by clustering the poly-lines as opposed to filtering some of them out. Siirtola [2000] proposes a *poly-line averaging* technique where a set of poly-lines, selected by the user, are represented by using the

corresponding *average* poly-line. In the *visual abstraction* scheme presented in Novotn [2004], clusters of poly-lines are visualized using their bounding polygons.

In Cui et al. [2006], the authors introduced measures of quality for sampling and clustering, in order to reduce the data to be visualized.

Edge bundles [Holten 2006] and Wavelets [Wong and Bergeron 1996] visualize datasets that contain hierarchical information or adjacency relations for reducing visual clutter.

In Fua et al. [1999a, 1999b], all the poly-lines are displayed; however, the elements belonging to different data clusters are differently colored to help the user discern them better. The colors are selected in a way that reflects the similarities and distances between the clusters. Then, the authors select the data clusters using a hierarchical clustering scheme and thus enable a multiresolutional view of the data (i.e., the user can select how much detail has to be preserved in the visualization).

In this article, we also consider clustering of the data. However, instead of using multidimensional clustering techniques that ignore the impact of the data clustering on the projections of the data on the given data dimensions, we primarily focus on controlling the reductions of the resolutions of the individual data dimensions when the data is clustered.

2.2.4. Other Visualization Enhancements. Graham and Kennedy [2003] introduce various refinements to the PC-based visualization. One of these enhancements is to replace the poly-lines with smooth B-spline curves (as we do in this article[4]). The advantage of using B-spline curves is that poly-lines that would partially overlap in space are differentiated from each other when plotted as smooth curves and this allows discerning individual data from each other even if they have shared values [Bartels et al. 1995]. This reduces ambiguity. As a downside, however, this increases the number of distinct curves on the visualization space, potentially worsening the visual clutter. To reduce clutter, Zhou et al. [2008] control the curvatures in a way that maximizes *parallelism* of related data curves. Wong and Bergeron [1997] augment the PC with a low-dimensional overview of the data obtained using principal component analysis. While not reducing the visual clutter, this often helps the user locate and track data clusters more effectively.

3. PHC FOR MULTIRESOLUTION VISUALIZATION OF LARGE TEXT DOCUMENT COLLECTIONS

As we mentioned in the Introduction, in this article, we present the Parallel hierarchical Coordinates (PhC) technique for visually analyzing large document collections. Unlike traditional tag- or term-clouds, PhC aims to help the user observe not only the frequently occurring terms, but the underlying patterns within this term space (Figure 6). In order to map multidimensional document data on 2D graphical displays in a way that preserves the underlying information and is easy to view and explore, PhC relies on a Parallel Coordinates (PC)-based mapping, where user-selected visualization terms are represented as vertical parallel *coordinates* and each document in the dataset is drawn as a continuous curve across these coordinate lines. In order to further help the user, PhC eliminates visual clutter relying on value clustering hierarchies obtained by analyzing the input set of documents. Based on the user input, the system suggests a clutter reduction strategy in a way that maintains as much information as possible; the user can then use OLAP-like navigational operators, such as drill-down and roll-up, to increase or decrease the hierarchical resolution to better observe data patterns.

[4]We used cubic B-splines.

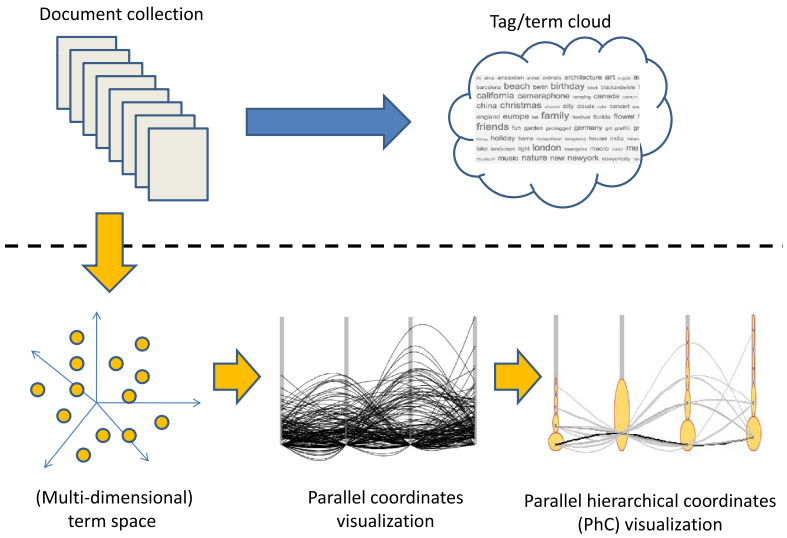
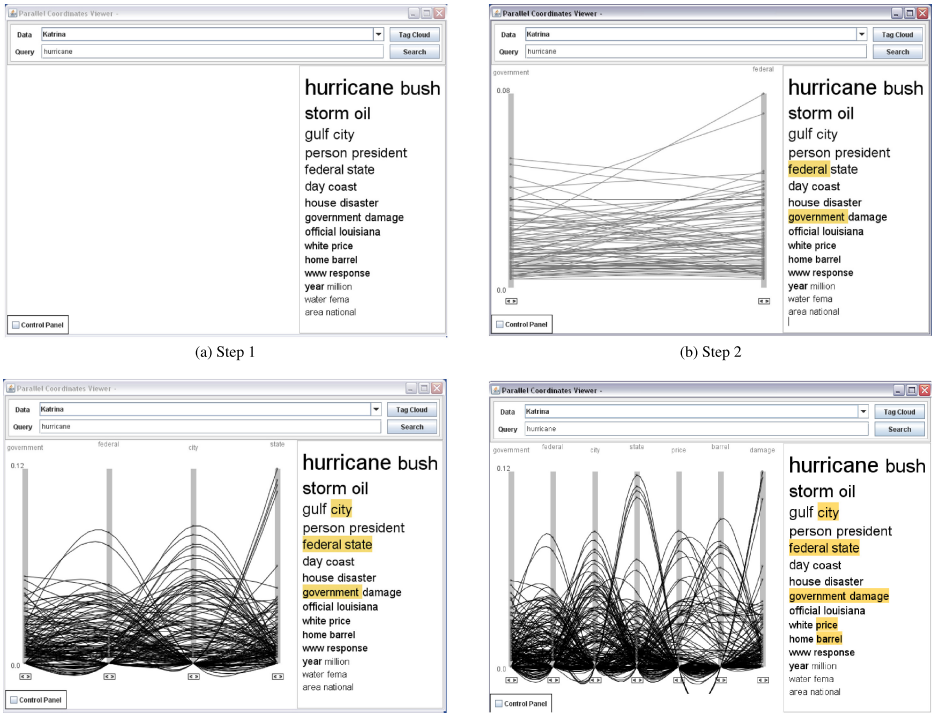


Fig. 6. Unlike traditional tag clouds, PhC helps visually analyze the given document collection for relationships between the tags.

In Section 4, we will discuss how PhC minimizes information loss during visual clutter reduction. Before that, however, in this section, we present an example user interaction sequence which describes how PhC helps the user visually analyze a document collection.

— *Step 1: Data selection.* In the first step, the user selects a data source and further focuses the analysis by providing a filter condition[5].
In our example shown in Figure 7(a), the user has selected the "Katrina" dataset and provided "hurricane" as the filter condition; the system then created and presented the user a term-cloud consisting of frequent terms in the documents that match this filter condition.
— *Step 2: Coordinate selection.* Given the term-cloud, the user can select any of these terms for analysis or can provide additional terms that do not occur in this term-cloud. In Figure 7(b), the user has selected the terms "government" and "federal" and the system created a PC visualization with two parallel coordinates.
Here, each document is represented as a curve that crosses each term coordinate at a point corresponding to the term's frequency in the document (in order to help the visualization, by default only the relevant frequency range is visualized, though the user can optionally select any frequency range for visualization).
— *Step 3: Coordinate selection (cont.).* In Figure 7(c), the user has selected two more terms for inclusion in the analysis: "city" and "state". As a result, the PC visualization now contains four parallel coordinates.
— *Step 4: Coordinate selection (cont.).* The selection can be very large, even though the displayed information becomes difficult to manage. In Figure 7(d), the user has selected an additional three terms for a total of seven terms.

[5]In this example, we are using a set of 750 hurricane Katrina-related news articles, also used in Di Caro et al. [2008]: the event has a multitude of, now well understood, facets, including geographic, humanitarian, economic, and political (local, regional, and federal) aspects.



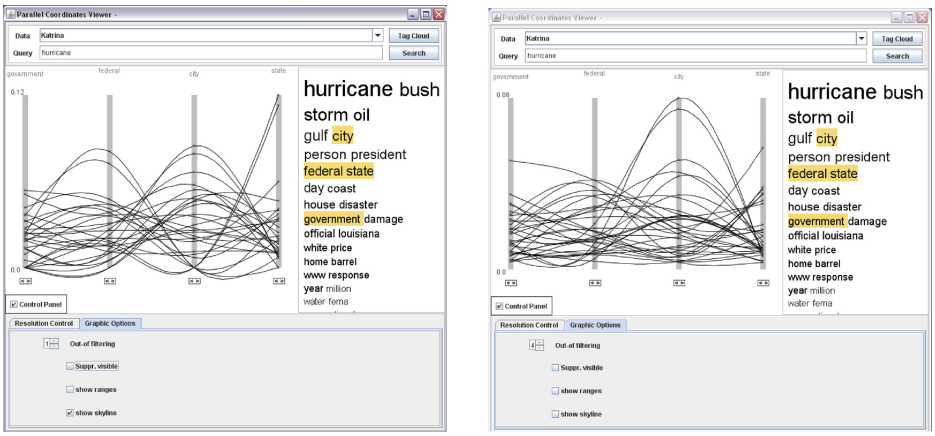
(a) Step 1 (b) Step 2

(c) Step 3 (d) Step 4

Fig. 7. First user interaction steps: selection of parallel coordinates.

At this point, the user can continue the iteration with the system using several available functionalities.

— *Data reduction through "skyline" selection.* At this point, the user may be uncomfortable with the amount of curves in the visualization and may want to reduce the clutter.
One way to achieve this is to keep only those documents that are not *dominated* by any others in the dataset. This set is also known as the *skyline* set [Borzsonyi et al. 2001].
  — No document in the skyline has a higher value than any other one in the skyline set with respect to all the dimensions of the space, and
  — no document in the dataset is in the skyline if there is at least one other document that has a higher score in all visualization dimensions.
Intuitively, for text exploration this implies that the skyline set contains documents that are the best representatives of different weight combinations. For example, if the document with weights $d_1 = < 0.8, 0.6, 0.1 >$ is in the skyline, then the documents $d_2 = < 0.7, 0.6, 0.1 >$ or $d_3 = < 0.3, 0.3, 0.1 >$ cannot be in the skyline. The document $d_4 = < 0.3, 0.3, 0.4 >$ on the other represents a document which is relatively more dominant in terms of the 3rd dimension in contrast to the first text document, d_1, and therefore can be in the skyline along with the initial document. Figure 8(a) shows the skyline set of the documents in the dataset selected by the user (see Figure 7(c)).
— *Data reduction through coordinate filtering.* Alternatively, the user can reduce the documents in the visualization shown in Figure 7(c) by putting a lower and/or upper



(a) skyline reduction (b) dimension filtering

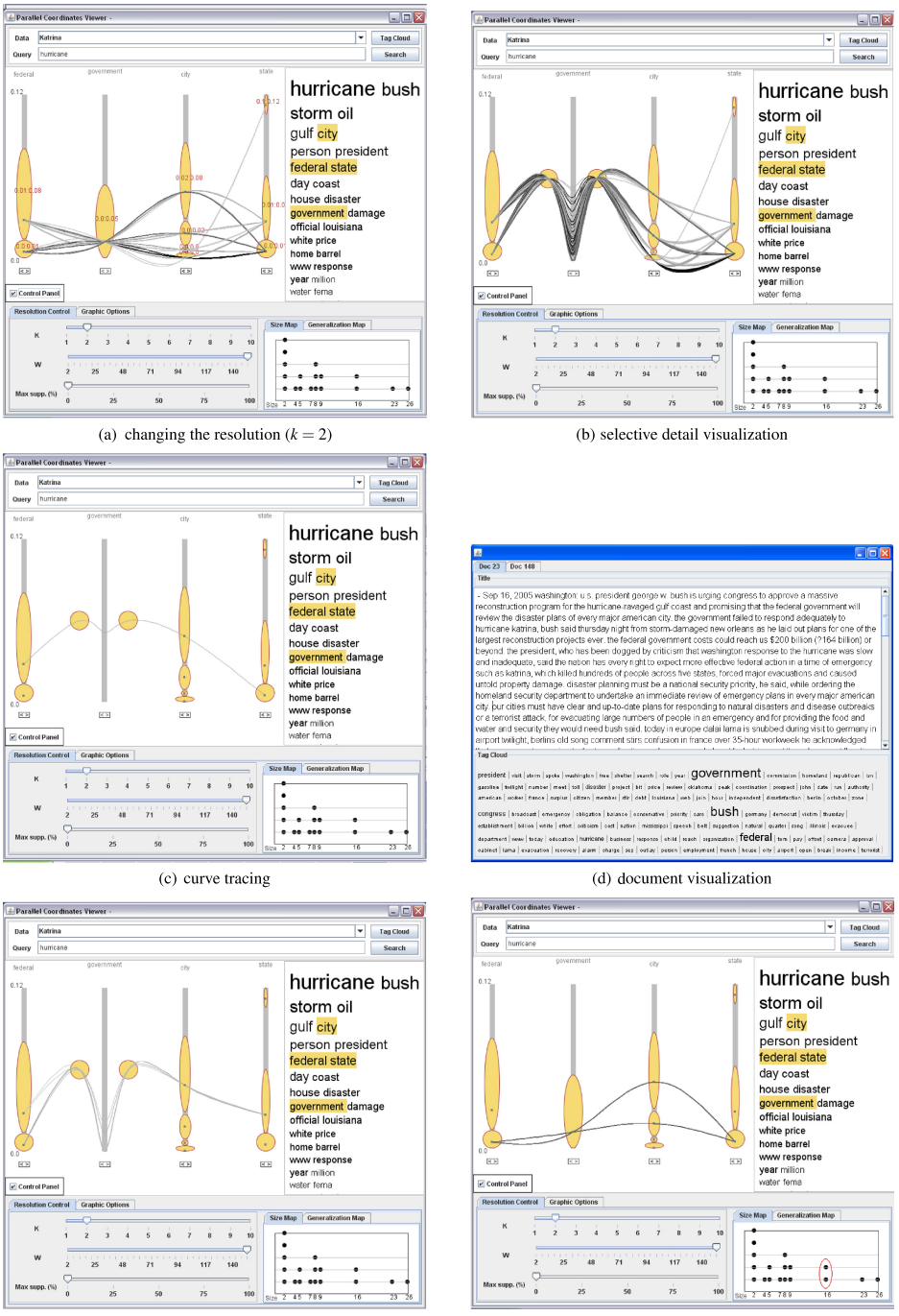
Fig. 8. Options for data reduction: skyline reduction and dimension filtering.

bound on the frequency values of the coordinates of the visualization space. In Figure 8(b), for example, the user instructed the system to eliminate any documents that have a 0.0 value along any of the visualization dimensions (i.e., missing any of the four visualization terms).

— *Reduction of the visualization resolution by document clustering (k = 2).* Another way to reduce the clutter in the visualization in Figure 7(c) is to cluster *similar* documents together. In Figure 9(a), the user instructs PhC to reduce the amount of details in Figure 7(c) by ensuring that each curve in the display corresponds to at least two documents (the user achieves this by bringing the sliding bar widget named "K" to 2). The system selects a value clustering strategy that achieves this effect with the minimal amount of information loss (see Section 4 for more details). As can be seen in the *size map* at the lower right corner of the interface in Figure 9(a), the system achieves this by creating document clusters of varying sizes, the smallest having 2 and the largest having 26 documents. In the main window, each of these document clusters are represented using a single curve and the shade of the curve is used to denote the size of the corresponding cluster (the darker the curve, the larger the corresponding document cluster).
Note that, in order to achieve the target document clustering, PhC needed to cluster some of the values along the coordinates. The visualization interface shows the user the clustering strategy selected by the system by marking each value cluster using an ellipse spanning the corresponding value range. The document cluster curves cross the value clusters at the cluster centroids and the width of each ellipse represents the number of document clusters passing through the corresponding value cluster.

— *Selective detail visualization.* At any point in time, the user can peek into any of the resulting value clusters to see how the individual curves are distributed within the value cluster. Figure 9(b) shows an example where the user selected the term coordinate "government" and instructed the system to display how precisely the document cluster curves are distributed inside the corresponding value cluster.

— *Curve tracing.* The user can select and remove any of the uninteresting document clusters from visualization or can focus onto any of the interesting ones. For example, as shown in Figure 9(c), at any point, the user can click onto any interesting curve and ask the system to separate it from the crowd for better visualization.



(a) changing the resolution ($k = 2$)

(b) selective detail visualization

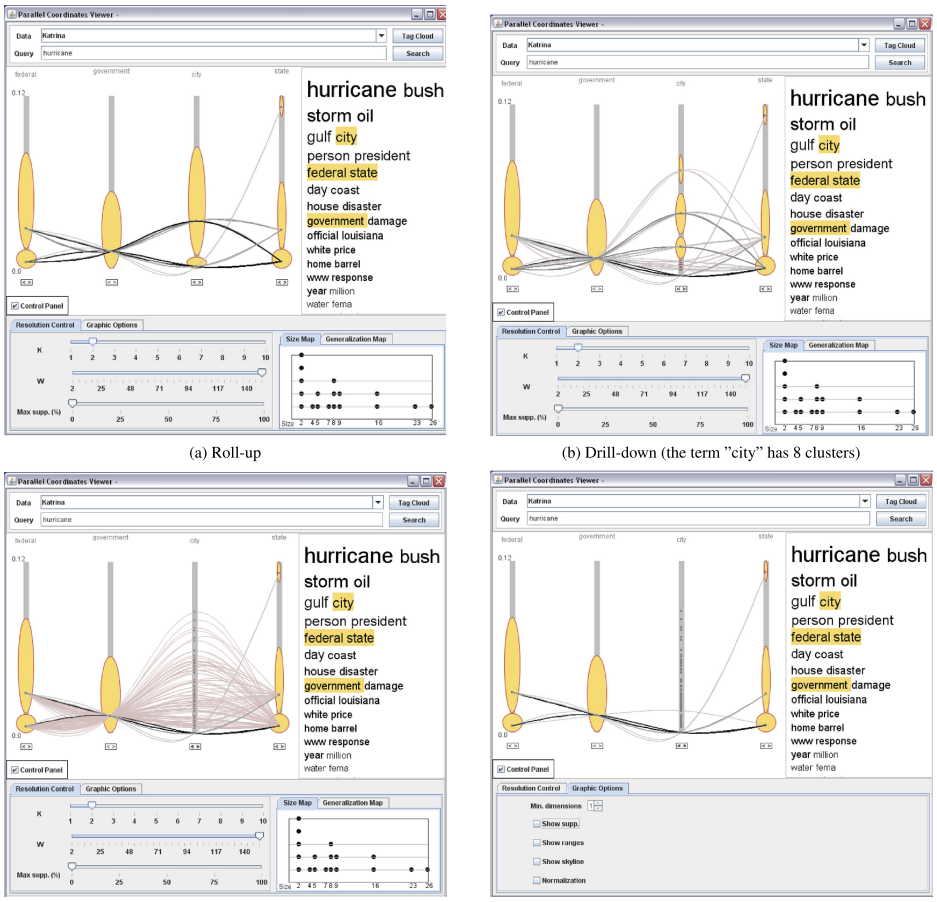
(c) curve tracing

(d) document visualization

(e) curve group tracing

(f) curve tracing by size map

Fig. 9. Options of interaction with PhC.



(a) Roll-up

(b) Drill-down (the term "city" has 8 clusters)

(c) Drill-down

(d) Elimination of outliers

Fig. 10. Roll-up/Drill-down on PhC, and elimination of outliers.

— *Document visualization.* The user can also instruct the system to provide more details about the selected document cluster; for example, the system can compute and return a new term-cloud for this document cluster or simply open the documents in the cluster for the user's browsing (Figure 9(d)).
— *Curve group tracing.* Alternatively, as shown in Figure 9(e), the user can select a set of document cluster curves to be separated from the crowd for better comparison.
— *Curve tracing by size map.* The user can also highlight document cluster curves by interacting with the document cluster size map at the bottom right corner. In Figure 9(f), the user clicked on the dots corresponding to clusters with 16 documents and the corresponding curves are highlighted to the user by the system.
— *Roll-up.* Figure 10(a) shows the roll-up operation. Here, the user selected the "city" coordinate and instructed the system to climb up on the corresponding hierarchy. As a result, the three value clusters that PhC has selected for visualization are further clustered into two value clusters, corresponding to the clusters at a higher level of the value clustering hierarchy.
— *Drill-down.* In contrast, in Figures10(b) and (c), the user drills-down on the "city" coordinate, thus obtaining smaller and smaller value clusters. In fact, in Figure10(c),

the user has drilled-down in the hierarchy all the way to the level of individual data values: in the figure, there are no value clusters on the "city" coordinate.

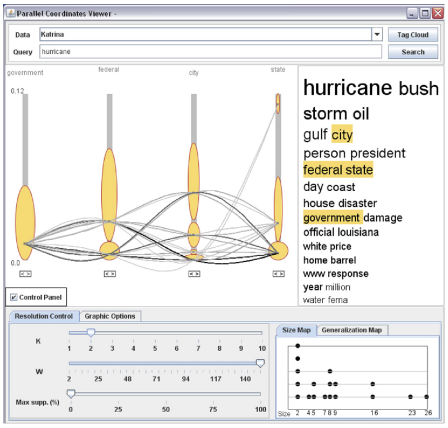
— *Elimination of outliers.* Note that not all curves on Figure10(c) are of the same color: while some curves are shades of gray as before, there are also some curves that are reddish in color. These reddish curves are those that violate the user-provided document clustering lower bound constraint "$k = 2$", which requires that all curves correspond to at least 2 documents in the document base. When the user drills-down along a coordinate, she is increasing the resolution along that coordinate and this may lead to the dissolution of some of the document clusters originally selected by PhC. This may consequently result in outlier documents that cannot be clustered with the rest of the documents at the selected resolution.
In Figure10(d), the user instructs the system to eliminate these outliers from the visualization, keeping in the visualization only those curves that correspond to document clusters with at least two documents in them.

— *Reducing the resolution of the visualization by increasing the lower bound of the document cluster sizes.* In Figures 11(a) to (c), the user varies the value of k (i.e., the lower bound of the document cluster sizes) from 2 to 6. As a result of the reduction in the visualization resolution, the numbers of document cluster curves as well as the value clusters along the visualization coordinates drop.

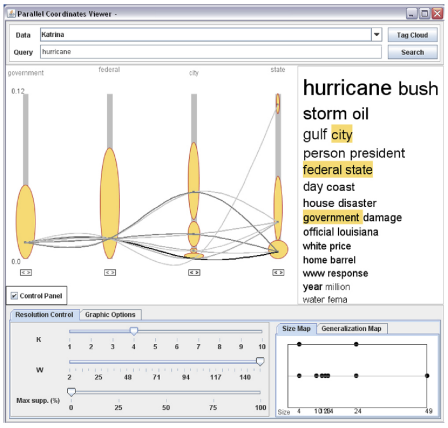
— *Decreasing the upper bound of the document cluster sizes.* The user can also place an upper bound on the number of documents in each document cluster created by PhC and displayed as a curve on the screen. Noticing from the "size map" corresponding to $k = 2$ (lower right corner in Figure 11(a)) that there is one curve representing 26 documents, in Figure 11(d) the user instructs the system (using the sliding bar widget marked as "W") to find an alternative clustering strategy where there are no document cluster curves representing more than 20 documents.
As a result, the system selects an alternative clustering strategy which repartitions the values along the "government", "city", and "state" coordinates into finer value clusters and identifies and plots new document cluster curves. The new "size map" on the lower right corner of the interface in Figure 11(d) shows that, now, the maximum document cluster size is 18.
One consequence of placing an upper bound on the sizes of document clusters along with a lower bound is that (as later explained in Section 4) this may result in clusters that have less documents than the user-selected lower bound ($k = 2$ in this example). In our example, the lower and upper bounds selected for the document cluster sizes by the user necessitate $\sim 60\%$ of the documents being identified as outliers that cannot be clustered with the rest. This is communicated to the user with the slider bar at the lowest left corner of the interface.

— *Increasing the maximum suppression rate.* Alternatively, instead of providing an upper bound on the document cluster sizes, the user can allow the system to mark up to a specific portion of the documents as outliers. In Figure 11(e), the user allows the system to consider up to 20% of the documents in the dataset as outliers. This results in a clustering strategy which maintains more details in the visualization while eliminating some of the documents from the visualization as outliers (compare the value clusters along the dimensions in Figures 11(a) and (e)).

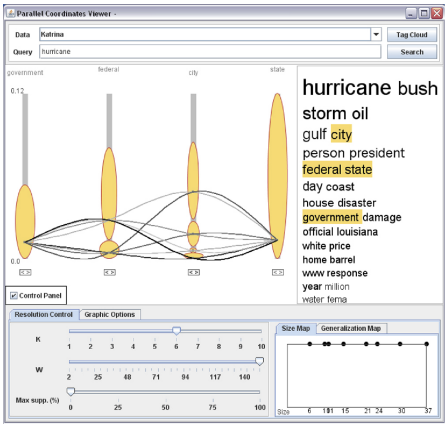
— *Changing the visualization coordinates.* At any point, the user can drop any of the current visualization coordinates and/or add one or more new terms. In Figure 12(a), we see that the user has dropped the coordinate "government" from the visualization and has added the new coordinate, "president". After this change of coordinate, the user can continue exploring the patterns in the document set along this new set of dimensions.



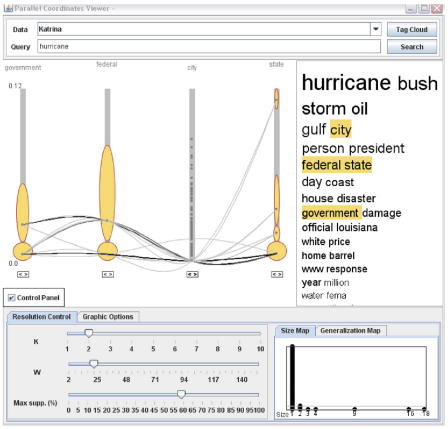
(a) changing the resolution ($k = 2$)



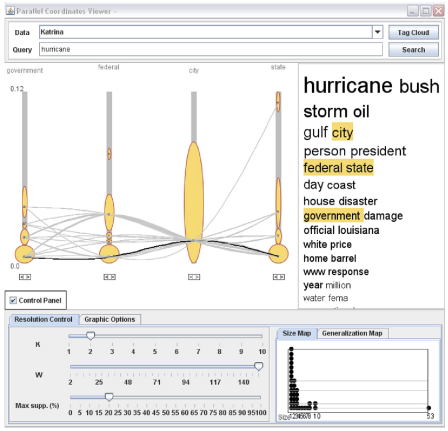
(b) changing the resolution ($k = 4$)



(c) changing the resolution ($k = 6$)

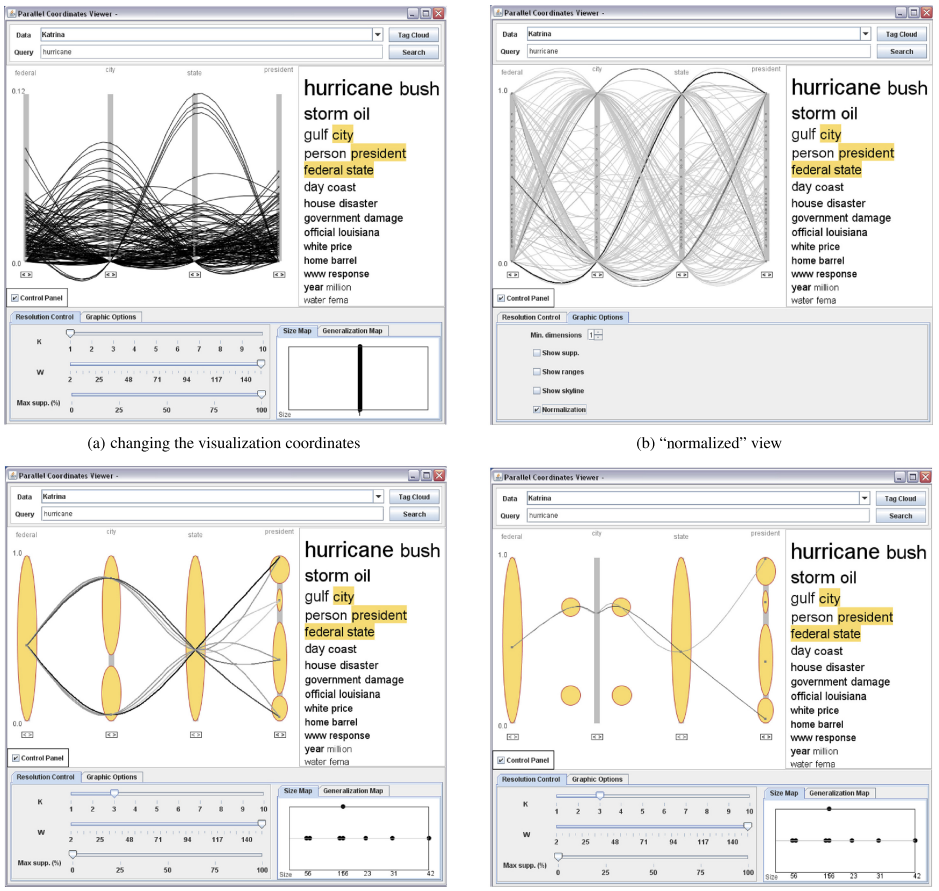


(d) changing the upper bound ($w = 20$)



(e) changing the maximum suppression rate

Fig. 11. Use of parameters k and w.



(a) changing the visualization coordinates

(b) "normalized" view

(c) "normalized" view, changing resolution k

(d) "normalized" view and curve tracking

Fig. 12. Steps 21 through 24: use of hierarchical coordinates.

— *"Normalized" view.* In Figure 12(b), the user has decided to investigate the documents in a normalized document vector space, instead of in the original document vector space. In other words, each document vector $\vec{v}$ in the 4D visualization space is normalized into $\vec{v}' = \frac{\vec{v}}{|\vec{v}|}$. Consequently, any two document curves that have similar keyword compositions, but of are different length (such as $\langle 0.4, 0.2, 0.1, 0.3\rangle$ and $\langle 0.2, 0.1, 0.05, 0.15\rangle$) are now overlapping in the visualization space and can easily be clustered with each other. Note that this implies that the document clusters obtained by varying the resolution of the visualization space (Figure 12(c) and (d)) will likely contain documents that are similar to each other in terms of relative keyword composition (i.e., in terms of *cosine similarity*).

Note that in the preceding example sequence, each parallel coordinate correspond to a single term. In general, however, multiple terms may be grouped by the user into a combined concept and assigned to a visualization coordinate. Alternatively, as mentioned earlier, a latent analysis algorithm, such as the standard LSA Deerwester [1989, 1990] or Latent Dirichlet Allocation [Song et al. 2009], may be used to identify significant term vectors or topics to be visualized as coordinates.

4. MINIMIZATION OF THE LOSS DURING CLUTTER REDUCTION IN PHC

As we discussed in Section 2.2, there have been various attempts to leverage clustering for visual clutter reduction in parallel-coordinates-based visualizations. It is important to note that tackling visual clutter through filtering or clustering may lead to information loss and what is clustered and how much clustering has been used have to be decided carefully. In fact, a common deficiency of the existing approaches, such as Fua et al. [1999a], is that they focus on clustering of the dataset, ignoring the characteristics of the individual data dimensions/coordinates. In this article, we note that starting from the (hierarchical) clusterings of the values along the individual coordinates can help better control the amount of loss in resolution along the individual dimensions of the data.

Thus, to tackle visual clutter, PhC relies on hierarchical clusterings of the values along the visualization coordinates (Figure 5). During the user's exploration of the document collection, PhC often decides for each dimension the most appropriate resolution based on a user-provided target document clustering rate (see Figures 4, 5, and 9(a)). As described in the earlier section, the user can then interact with the system to selectively roll-up or drill-down along the hierarchies corresponding to different visualization coordinates to explore the patterns in the document set.

In the rest of this section, we describe how the Parallel hierarchical Coordinates (PhC) controls the degree of value and document clustering in such a way that the information loss is minimized.

4.1. Value-Clustering Hierarchies

Let A be a user-selected term (or a "concept" consisting of a set of terms). The corresponding value clustering hierarchy (identified using a hierarchical clustering algorithm) is a tree $H_A(V, E)$, where

— each $v = (nodeid : value) \in V$ is a node in the tree and $v.value$ is either the weight of a document in the dataset for A or is the value range using the value clustering algorithm, and
— $e = v_i \rightarrow v_j \in E$ is a directed edge denoting that the value encoded by the node v_j can be clustered under the value encoded by the node v_i.

Given a value hierarchy H_A, a tree node v_i is a clustering of a tree node v_j, denoted by $v_j \prec v_i$, if there exists a path $p = v_i \hookrightarrow v_j$ in H. For example, if we consider ranges of values between 0.0 and 1.0, $[0.2, 0.3) \prec [0.0, 0.4)$ and $[0.0, 0.4) \prec [0.0, 0.5)$.

Note that these value-clustering hierarchies are computed for each of the user-selected visualization coordinates (i.e., terms) before the PhC visualization for the given set it created.

4.2. k-Clustering of Documents

In the rest of this section, we present the algorithms PhC uses for reducing the visual clutter in PC, given a target document clustering resolution. The idea is to cluster the values in each dimension in a controlled manner using the given value-clustering hierarchies, in such a way that each poly-line or curve in the resulting PC visualization will group at least k individual documents in the document collection. We refer to this as the k-clustering of the documents.

Intuitively, k is the parameter controlling the degree of resolution of the PC visualization. Note that a given document set can be k-clustered in many different ways. For example, a simple (but obviously unacceptable) k-clustering strategy would be using a single poly-line to represent the whole document collection. The challenge is thus to identify a value clustering strategy (i.e., a level in the corresponding hierarchy for

each term) in a way that achieves the k-clustering goal, yet loses as little information as possible (see Section 5 for more details on information-loss measures).

To obtain good k-clusterings of the data, we build on the privacy-preserving data publication approaches presented in the literature [LeFevre et al. 2005; Li and Li 2007; Machanavajjhala et al. 2007; Samarati and Sweeney 1998]. Given a table to be published consisting of sensitive attributes and their values, the goal in these approaches is to limit the amount of data leaked by replacing the specific entries in the database table with more general cluster labels. In k-anonymization[6] problems, for example, the acceptable degree of hiding is defined as the generalization in which each row in the published table is indistinguishable from at least $k-1$ other rows [Ciriani et al. 2007] This k-anonymization approach eliminates the possibility of linkage attacks by ensuring that, in the disseminated table, each value combination of attributes is matched to k others. To achieve this, these techniques rely on a priori knowledge about acceptable value generalizations. Thus, most of these algorithms assume that there is a taxonomy associated to each sensitive attribute and that this taxonomy can serve as the value-clustering hierarchy for that attribute. Given that there may be many publishable tables, each providing the same level (k) of row hiding, most approaches aim to locate a publishable table that also preserves as much information in the original table as possible to make sure that the published data table will be of use to its recipient after its generalization. In a low-resolution representation of the data, an internal node of the hierarchy can be used to cluster all the values (i.e., leaves). On the other hand, more general cluster labels will also cause a higher degree of information loss; thus, among all possible clusterings that put each tuple with $k-1$ other ones, Samarati and Sweeney [1998] and many others aim to find those that require minimal generalizations.

Cell generalization schemes [Aggarwal et al. 2005] treat each cell in the data table independently. Thus, different cells for the same attribute (even if they have the same values) may be generalized in a different way. This provides significant flexibility in anonymization, while the problem remains extremely hard (NP-hard [Meyerson and Williams 2004]) and only approximation algorithms are applicable under realistic usage scenarios [Aggarwal et al. 2005]. Attribute generalization schemes [LeFevre et al. 2005; Samarati 2001] treat all values for a given attribute collectively; that is, all values are generalized using the same unique domain generalization strategy. While the problem remains NP-hard [Meyerson and Williams 2004] (in the number of attributes), this approach saves a significant amount of time in processing and may eliminate the need for using approximation solutions, since it does not need to consider the individual values. Most of these schemes, such as Samarati's original algorithm [Samarati 2001], however, rely on the fact that, for a given attribute, applicable generalizations can be put into a total order of information loss; that is, the higher you go in the hierarchy, the more you lose information. More specifically, if there is a generalization at depth d that puts all tuples into clusters of size k, then any other generalization at level $d' \leq d$ will also group all tuples into clusters of size at least k, but it will have more loss; conversely, if one can establish that there is no generalization at level d that is a k-clustering, then there is no other clustering of level $d' > d$ that can cluster all tuples into clusters of size at least k. LeFevre et al. [2005] leverages this to develop an algorithm which achieves attribute-based k-anonymization one attribute at a time, while pruning unproductive attribute generalization strategies. Samarati [2001], on

[6]Note that k-anonymization is not the only criterion used in privacy-preserving data publication. Machanavajjhala et al. [2007] and Li and Li [2007] present l-diversity and t-closeness criteria, respectively. However, for the goals of our article, the k-clustering property of the k-anonymization approaches provides the best match.

the other hand, leverages this to develop a binary search scheme to efficiently identify the most specific generalization which guarantees clusters of size at least k: Let D be a dataset and $GA = A_1, ..., A_m$, be the set of generalization attributes. For each attribute in A_i, the algorithm takes a value-clustering hierarchy (a taxonomy, T_i) which describes the generalization/specialization relationship between the possible values in the domain of the attribute A_i. Let the height of T_i be h_i

(1) The algorithm first computes the maximum amount of information loss possible: this happens when all m attributes of the table are generalized all the way to the root; that is, the maximum amount of generalization is $H = h_1 + h_2 + \ldots + h_m$.
(2) $Max = H$ and $Min = 0$ (0 corresponds to the original table, where none of the values has been generalized).
(3) The algorithm then considers all possible generalizations that involve a total of $L = \lceil \frac{Max - Min}{2} \rceil$ steps.
   — If a k-clustering generalization is found at this level, then any k-clustering generalizations that requires more generalizations will have extra information loss; however, there may be k-clustering solutions that require less generalizations. Therefore, the algorithm next considers the range $Max' = L - 1$ and $Min' = Min$.
   — On the other hand, if a k-clustering generalization is not found at this level, then we can be sure that there is no generalization at any level below L. However, there may still be a k-clustering solution that requires more generalization steps than L. Therefore, the algorithm next considers the range $Max' = Max$ and $Min' = L + 1$.
(4) Step 3 is repeated is repeated until $Max = Min$ and the k-clustering with the smallest amount of generalization found in the process is returned.

In this article, we note that a similar strategy can be used for reducing the visual clutter in parallel coordinates-based visualization of text documents. In particular, given a document set and a value-clustering hierarchy for each of the terms selected by the user for visualization, we can locate a k-(document)-clustering strategy that requires the least amount of value generalizations[7]. This way, the algorithm would identify the appropriate resolution for each dimension that collectively minimizes the information loss while ensuring that each document cluster will contain at least k documents. Figure 9(a) shows the application of k-clustering applied to the PC visualization of document sets.

The computational complexity of the PhC visualization is determined by the underlying k-clustering process. In our implementation, we use the k-clustering strategy presented in Ciriani et al. [2007], which performs binary search on the levels of the input hierarchy. The k-clustering problem and the specific algorithm [Ciriani et al. 2007] we use are known to be exponential in the number of attributes (i.e., the visualization dimensions), but only quadratic in the number of data entries. Moreover as k increases, the complexity of the problem tends to drop as it is easier to find clusters that satisfy the given clustering target. Since the number of visualization attributes are often small and since the quadratic processing of the data entries (to compute a so-called pairwise distance table can be done offline as a preprocessing step), in practice the runtime cost of PhC is not a major obstacle. When the number of visualization attributes is large, the clusterings may need to be precomputed and cached to support

[7]Note that, while we do not explicitly consider this in this article, we can also associate a degree of information loss to each internal node in the hierarchy (for example, representing the span of the corresponding range; e.g., the range $[0.0, 0.4)$ is twice less precise than the range $[0.0, 0.2)$) and thus minimize the total amount of information loss measures not in terms of generalization steps, but in terms of this information loss measure.

interactive exploration. PhC also leverages caching of k-clustering solutions to ensure that once the clustering is computed, it can be reused throughout the user interaction process.

4.3. Outlier Suppression and [k,w]-Clustering of Documents

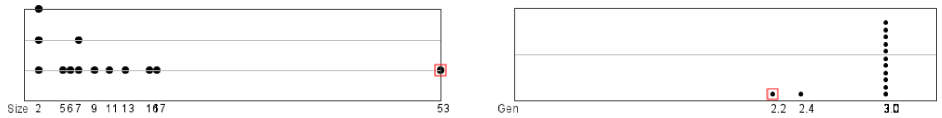
One difficulty with k-clustering is that clustering the outlier documents with the rest of the document to obtain the lower bound cluster size, k, may necessitate high degrees of value clusterings, which in turn would cause document clusters much larger than the desired lower bound, k. This explains the large document cluster with 26 documents in Figure 9(a). Given such large value or document clusters, the user can either drill-down along a selected dimension explicitly or ask the system to try to suppress the outliers from the visualization.

4.3.1. [k, w]-Clustering of Documents. With the basic k-clustering scheme, the user specifies the lower bound clustering constraint k, but no constraints are imposed on the maximum level of clustering. This means that in the visualization we can have polylines or curves that correspond to $\gg k$ data elements; more importantly, the number of elements represented by different curves in the visualization may differ significantly from each other. While this variation in cluster sizes may be communicated to the user with visual cues, such as line thickness, PhC also allows the user to place an upper bound constraint on the number of elements captured by each curve. We refer to this as the [k, w]-clustering of the data, where k represents the lower bound and w is the upper bound of the clustering rate.

One difficulty with placing an upper bound on the cluster sizes is that there may be situations in which there are no generalizations that can satisfy both lower bound and upper bound constraints. This situation occurs especially when the document set has outliers: Let d be an outlier document, with one or more of the dimension weights significantly different than the rest. Clustering the document d with $k-1$ others may require increasing the ranges of some of the value clusters so much that, inadvertently, many other documents may fall into this range resulting in a document cluster with a size $\gg k$. Therefore, given k and w, if no appropriate [k, w]-clustering is found, then PhC identifies the minimum number of outlier documents whose *suppression* will ensure [k, w]-clustering of the remaining documents. This is achieved by, if needed, varying the degree of suppression using binary search until a [k, w]-clustering is found.

To search for suppressions, we build on a variation of the k-anonymization problem where the user is allowed to specify the maximum number (*maxsup*) of suppressions allowed when an appropriate generalization cannot be found. When the *maxsup* is specified by the user, the step 3 of the algorithm in the previous step is modified in such a way that the system searches not for k-clusterings of the table, but k-clusterings which have at most *maxsup* suppressions. Samarati [2001], for example, presents a dynamic programming-based algorithm that can verify if a given generalization strategy provides a k-clustering with at most *maxsup* submissions or not in quadratic time.

Given a [k, w]-clustering target, PhC first locates a k-clustering and checks the size of the maximum cluster size, if the maximum cluster size is $\leq w$, then this solution is returned. If the maximum cluster size is greater than w, then the algorithm searches for a k-clustering using binary search, where each iteration a different *maxsup* rate is considered. Starting from $\lceil \frac{N}{2} \rceil$, where n is the number of documents, the algorithm considers different suppression rates, each time halving the range and decreasing the target suppression rate when a [k, w]-clustering is located and increasing the suppression rate when a [k, w]-clustering is not found. Note that this scheme differs from naive binary search in that the algorithm does not stop immediately when a [k, w]-clustering is identified; instead it tries to see whether there exist [k, w]-clusterings



(a) document cluster size map (b) document cluster generalization (loss) map

Fig. 13. The cluster size map in (a) shows that one of the document clusters in the dataset is much larger than the rest; in the corresponding cluster generalization map shown in (b), the same cluster is seen to have low average loss. This implies that this cluster represents a dominant pattern (i.e., the large number of documents with low loss in value clustering).

with lower suppression rates. Overall, the algorithm continues $O(logN)$ iterations, where N is the size of the document set or until the $[k, w]$-clusterings converge.

Figure 11(d) in the previous section presents an example $[2, 20]$-clustering, which results in a maximum document cluster size of 18. Note that, in addition to reducing the size of the largest document cluster, this also helps further partition the values along the visualization dimensions, resulting in more detailed visualizations. Note also that to achieve the desired lower and upper bounds, the algorithm has selected a suppression rate of 60% of the documents. Of course, suppression does not mean that these documents are lost, but only that they are visualized differently (as outliers) than the rest: suppressed documents can either be hidden from the visualization or, as shown in Figure10(c) , they can optionally be included in the visualization in red.

4.3.2. Direct Selection of Suppression Rate. Alternatively, PhC can take the acceptable suppression rate, *maxsupp*, directly as an input from the user. Given *maxsup* > 0, the algorithm would identify a value-clustering strategy which would require as little clustering as possible by suppressing up to *maxsup* many documents.

4.4. Document Cluster Sizes and Loss in Precision

As described in this section, PhC allows the user to explore the data at different resolutions, specified by a clustering lower bound k and an upper bound w (as well as the outlier ratio *suppr*). Within these bounds, the number of documents included in different document clusters may differ from each other. Similarly, the number of generalization steps required for obtaining document clusters may also vary from cluster to cluster. Thus, PhC also provides visualization mechanisms to help the user isolate individual document clusters represented based on their sizes/generalizations and study the relationships between these two cluster properties.

One of these mechanisms is a tool called *document cluster size map* which allows the user to explore document clusters based on their sizes. The cluster size map can be seen at the lower right corner of the user interfaces in Figure 13. For example, in Figure 13(a), the sizes of the document clusters range from 2 to 53 documents. A related exploration tool PhC provides is the *document cluster generalization (or loss) map*, where each document cluster is visualized in terms of the number of value generalization steps required to obtain this cluster (i.e, the amount of loss in precision).

Figure 13(a) presents an example where one cluster has significantly more documents than the others; as mentioned earlier, this may be due to an outlier or may simply be because there are a lot of entries that have similar values. If the cluster is large and the corresponding loss is also high, this may be due to an outlier in the document set which may call for a high degree of clustering. In contrast, a cluster with a lot of documents, but a low degree of loss would indicate a strong pattern in the dataset (as is the case in the example in Figure 13).

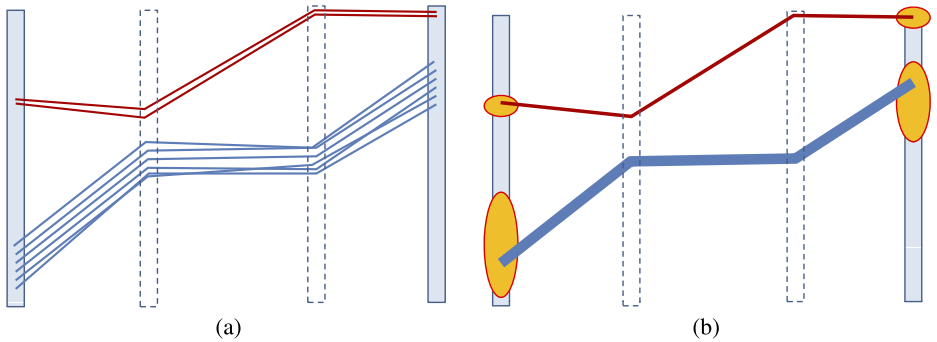


Fig. 14. (a) The visual clutter due to line density (b) can be reduced through document clustering.

4.5. Selection of Parameters *k* and *w*

For an unknown text document collection without any a priori statistics, there is no principled way to pick the very initial k/w values. However, the visualization interface provides exploration widgets and so "cluster size" and "cluster generalization" maps to help the user change k and w values in a way that maintains sufficient detail, with the least visual clutter. While different users can use these tools differently and different applications may put different emphasis on patterns versus outliers, the common approach is to vary k values until sufficient detail versus clutter trade-off is achieved. If during the process, one recognizes (using the "cluster size" map) that some of the clusters remain very large clusters and this makes it hard to investigate the collection, such clusters are further partitioned by putting a tighter limit on the minimum document cluster size w.

5. MEASURING VISUAL CLUTTER AND LOSS

As we stated in the Introduction, the goal of PhC is to reduce visual clutter, while preventing information loss as much as possible. In this section, we formalize visual clutter and loss and present alternative quantifiable measures for assessing different PC-based visualization strategies.

5.1. Visual Clutter Measures

In Yang et al. [2003], Yang et al. define visual clutter in terms of the number of outliers in the visualization. While this measure is attractive when focusing solely on outliers, in this article we focus on two alternative sources of visual clutter; namely, (a) the line density and (b) crossings of poly-lines or curves in the visualization.

5.1.1. Clutter Due to Line Density. In Tufte [2001], Tufte argues that when visualizing information, for each active point in the screen, there is a visual cost associated. Therefore, the number of active points/pixels should be proportional to the amount of information being represented in the visualization. Thus, given two visualizations that communicate the same information, the one with less active points/pixels is preferable. Based on this intuition, the first measure we define for visual clutter is the number of poly-lines or curves in the visualization. That is, among two PC visualizations that are able to communicate the same patterns (e.g. clusters and outliers) to the user, the one which uses the smaller number of lines or curves is more desirable (Figure 14).

5.1.2. Clutter Due to Line Crossings. Large numbers of line crossings can render patterns in the data hard to discern. Consider, for example, two clusters in Figure 3 that are separate in the original space; due to dimensionality reduction, the corresponding

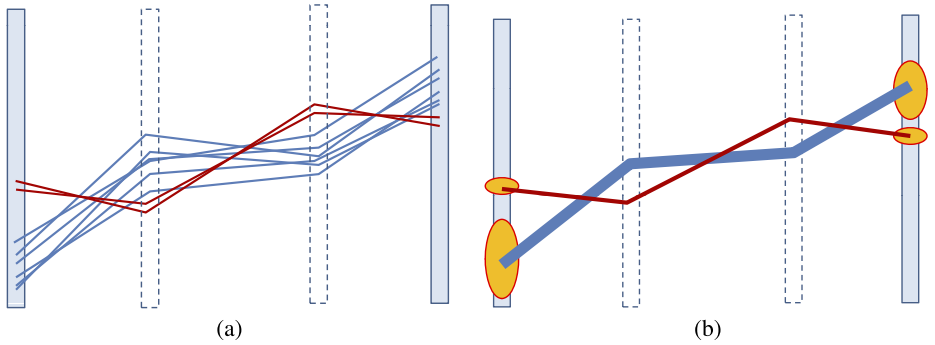


Fig. 15. (a) The clusters that are hard to discern due to too many line crossings (b) can also be made more apparent by clustering documents and thus reducing the number of crossings.

poly-lines or curves in the PC visualization may share the same visualization space resulting in large numbers of crossings. As a result, especially when the number of distinct clusters is large, these patterns may become increasingly harder to discern (Figure 15). Therefore, our second visual clutter measure focuses on the line crossings (excluding the crossings on the coordinate-lines due to value sharing). In particular, we use two different line-crossings-based measures.

— *Total number of crossings.* This simply counts the total number of crossings in the visualization. Since in the curve-based PC, a pair of curves may cross each other multiple times, we consider both *total* and *unique-total* measures (the latter including each crossing curve pair in the total only once).
— *Document cluster confusion.* Two document clusters would be easier to discern if their curves did not intersect in the visualization space. Therefore, given a document set with m a priori known document clusters, $C =< c_1, c_2, ..., c_m >$, we define the corresponding cluster confusion degree as the total number of line crossings, where the crossing lines belong to different document clusters.

Note that for a given k- or $[k, w]$-clustering, the line density and line-crossing measures of clutter are compared to the default PC visualization, which corresponds to $k = 1$.

5.2. Visual Loss Measures

In this section, we present measures that quantify the inadvertent visual loss that occurs during the visual clutter reduction process.

5.2.1. Visual Compression. A drop in the resolution of the visualization will affect the visual information communicated to the user. We measure this loss in the visualization in terms of the *average compression in the visual distances* among the curves.

Let doc_1 and doc_2 be two documents. The curves corresponding to doc_1 and doc_2 (along with the parallel coordinates that are at the end points) will define a closed space in the PC space. Relying on the observation that the amount of visual information conveyed by a shape on a 2D graph is proportional to the space (area) covered by it [Tufte 2001], we define the visual information conveyed by these two curves (i.e., $\Delta_{visual}(doc_1, doc_2)$) as the area in the visualization space between the two curves corresponding to doc_1 and doc_2 (Figure 16(a)).

Let D be a set of documents, $\Delta_{visual}()$ denote the function that returns visual distances in the original high-resolution (low clustering) visualization, and $\Delta'_{visual}()$ denote

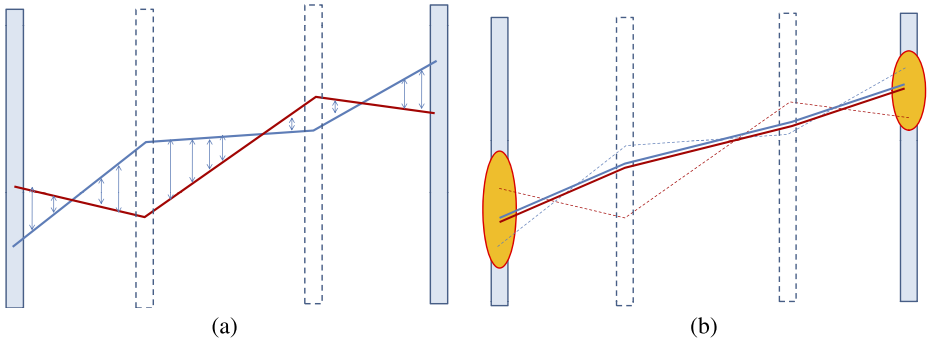


Fig. 16. (a) The visual distance between two curves (b) may be lost if the corresponding two documents are clustered together into a single curve.

the visual distances in the low-resolution (high clustering) visualization. Given these, the amount of visual compression in the visualization is defined as

$$comp_{visual}(D) = 1 - \frac{\sum_{doc_i, doc_j \in D(doc_i \neq doc_j)} \Delta'_{visual}(doc_i, doc_j)}{\sum_{doc_i, doc_j \in D(doc_i \neq doc_j)} \Delta_{visual}(doc_i, doc_j)}.$$

Intuitively, $X\%$ visual compression implies that curves are perceived to be $X\%$ closer in the average to each other than they really are after the clustering (Figure 16(b)).

5.2.2. Value-Normalized Visual Compression. Note that the preceding visual diversity measure does not consider the actual values that are being visualized to the user. One can argue that the visual diversity is more important when the documents are diverse; that is, if the document set does not contain diverse documents, then the visualization does not need to be diverse either. Let us define the value difference between two documents doc_1 and doc_2 as

$$\Delta_{val}(doc_1, doc_2) = \sum_{1 \leq i \leq m} \Delta_{val,i}(doc_1, doc_2),$$

where $\Delta_{val,i}(doc_1, doc_2)$ denotes the difference between doc_1 and doc_2 along the i^{th} term. Note that the term $\Delta_{val,i}$ can be defined in different ways. One possible definition is in terms of the absolute difference between the values of the corresponding terms (i.e., L1-distance).

$$\Delta_{val,i}(doc_1, doc_2) = |doc_1.term_i - doc_2.term_i|$$

Since it considers the absolute difference between the term values, this definition does not take into account the value distribution. Alternatively, one can consider the structure of the value-clustering hierarchies and define $\Delta_{val,i}$ as

$$\Delta_{val,i}(doc_1, doc_2) = \Delta_{hier,i}(doc_1, doc_2)$$

where $\Delta_{hier,i}(doc_1, doc_2)$ is the amount of loss in precision (as defined in Section 4.1) needed to cluster the values $doc_1.term_i$ and $doc_2.term_i$ into the range represented by their closest common ancestor in the value clustering hierarchy.

Alternatively, one can define $\Delta_{val}(doc_1, doc_2)$ directly using a dot-product-based interpretation of document similarity.

$$\Delta_{val}(doc_1, doc_2) = m - \left(\sum_{1 \leq i \leq m} doc_1.term_i \times doc_2.term_i \right)$$

Table I. Acronyms for Visual Clutter Measures

Abbreviation	Meaning
LD	Line Density
TCL	Total Crossing Lines
UCL	Unique Crossing Lines
CC	Cluster Compression

Given the appropriate definition of Δ_{val}, we define *value-normalized compression in visual diversity* as

$$comp_{norm}(D) = 1 - \frac{\sum_{doc_1,doc_2 \in D(doc_1 \neq doc_2)} \Delta'_{visual}(doc_1, doc_2) \times \Delta_{val}(doc_1, doc_2)}{\sum_{doc_1,doc_2 \in D(doc_1 \neq doc_2)} \Delta_{visual}(doc_1, doc_2) \times \Delta_{val}(doc_1, doc_2)}.$$

6. EVALUATION

In this section, we study the visual clutter and loss behaviors of the proposed PhC scheme. For this purpose, we use two datasets.

— *Katrina news dataset*, which was also used in our past work [Di Caro et al. 2008]: the dataset contains 750 hurricane Katrina-related news articles, published between August 25, 2005 and February 26, 2008. The reason why we chose this dataset as a case study earlier in the article and for evaluation in this section is that the event has a multitude of, now well-understood, facets, including geographic, humanitarian, economic (e.g., employment- and energy-related), and political (local, regional, and federal) aspects. In our experiments we used the filter and term context shown in Figure 7(c) (keyword "hurricane" as filter, with the coordinates "government", "federal", "city", and "state").
— *NSF abstracts dataset*, which contains 1000 abstracts of National Science Foundation (NSF) funded research[8]. This dataset contains abstracts that describe a diverse spectrum of scientific research topics. Furthermore, given the interdisciplinary nature of most NSF funded research, the dataset also provides opportunities to investigate interrelationships between different research areas. In our experiments we used filter and term context shown in Figure 2 (keyword "ocean" as filter, with the coordinates: "system", "model", and "process").

For each user-selected visualization coordinate (i.e., term), we have created the corresponding value-clustering hierarchy by recursively splitting the value range [0, 1] using EM clustering of the values along that coordinate in the selected dataset [Dempster et al. 1977], until no further splitting is possible or until a predetermined depth is reached[9].

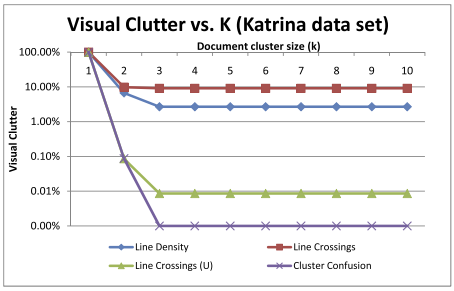
6.1. Visual Clutter Reduction

Figure 17 shows the drop in the values of the various visual clutter measures introduced in Section 5 (see Table I for the acronyms) as a function of the visualization resolution (i.e., minimum document cluster size k) selected by the user. As can be seen here, in both datasets, the amount of clutter, especially the number of line crossings, can be reduced multiple orders by using even relatively low k values, such as $k = 2$. While the absolute value of the reduction is data dependent, the results are very similar for both datasets and highlight the fact that one can achieve less cluttered visualizations of the data even with low reductions in resolution.

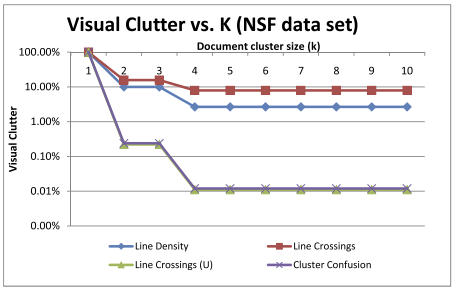
[8]http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html
[9]In the experiments, we used a maximum depth of 5 levels.

Visual clutter vs. k					
Data	*k*	*LD*	*TCL*	*UCL*	*CC*
Katrina	1 (default)	149	12115	11681	11462
	2	10	1188	10	10
	3-10	4	1102	1	0
NSF	1 (default)	150	11841	9054	8379
	2-3	15	1863	20	20
	4-10	74	936	1	1

(a) (absolute) visual clutter vs. k

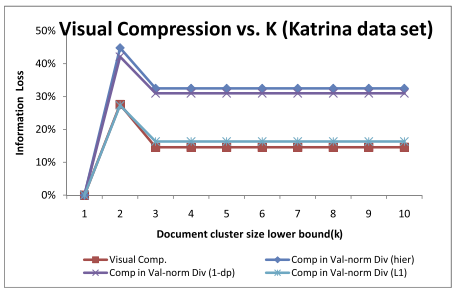


(b) relative visual clutter in the Katrina dataset

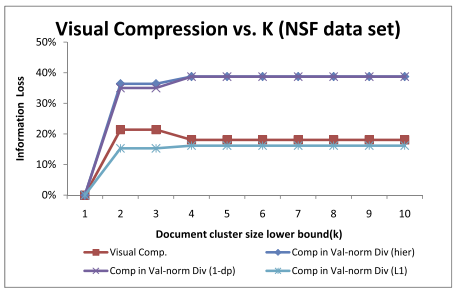


(c) relative visual clutter in the NSF dataset

Fig. 17. Visual clutter as a function of the resolution selected by the user; as can be seen here, the amount of clutter can be reduced multiple orders by using even low k values (e.g., $k = 2$ or 3). See Table I for the acronyms.



(a) visual compression in the Katrina dataset



(b) visual compression in the NSF dataset

Fig. 18. Visual compression as a function of the resolution selected by the user; as can be seen here, the visual compression behavior is consistent across data sources.

The quick flattening of the curves in Figure 17 indicates that a significant portion of the visual clutter can be eliminated using only small degrees (k) of clustering, as noninformative line crossings are quickly removed from the PhC visualization.

6.2. Visual Loss

In order to study the impact of low-resolution visualizations on the visual loss, we consider the loss measures presented in Section 5.2. Figure 18 shows the values of the various loss measures as function of the minimum document cluster size lower bound (k) selected by the user. As can be seen here, despite the multiple orders of reductions in visual clutter as k increases (see Figure 17), the amount of compression in visual distances grows much slower. This indicates that the generalization scheme presented in this article is able to remove visual clutter while maintaining the major patterns

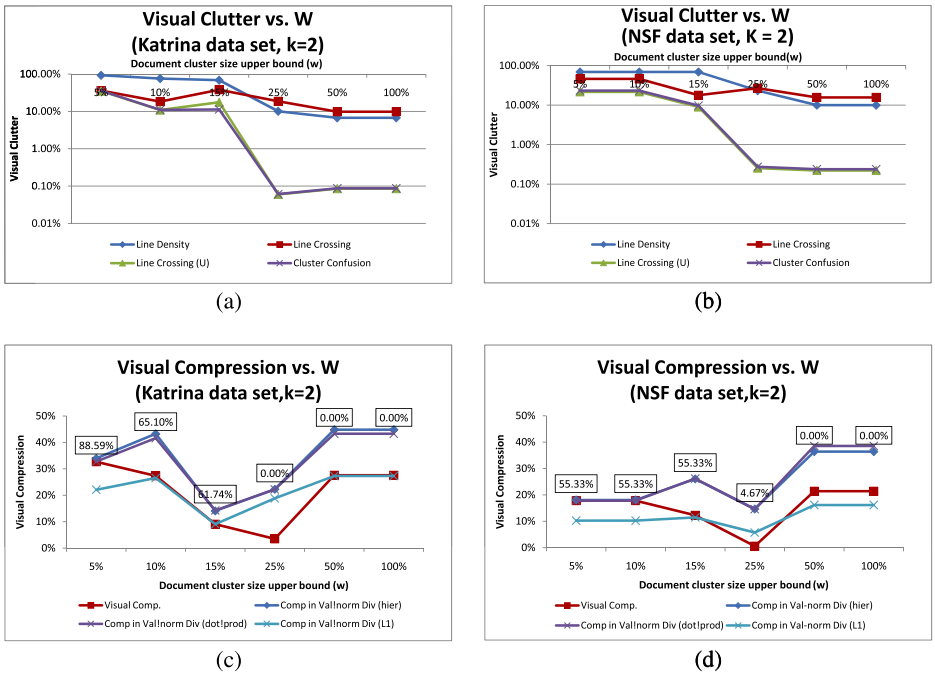


Fig. 19. The impact of the upper bound w (varied as a ratio of the total data size) on (a) (b) visual clutter and (c) (d) visual compression. Note that $w = 100$ corresponds to the default case where no upper bound is provided (the amount of suppression corresponding to different w settings are shown as data labels within boxes in (c) and (d)).

intact. Especially in terms of pure visual compression and compression normalized based on L1 interpretation of document distances, the amount of visual compression is only ~ 20; that is, on the average, curves are perceived to be 20% closer to each other than they really are. When hierarchy- or dot-product-based document distances are considered, the amount of value-normalized visual loss is relatively higher, but still on the average, curves are perceived to be only $\sim 30\%-40\%$ closer to each other than they really are. Most importantly, all four measures stabilize beyond a small level of k, indicating that the stable patterns with only limited information loss emerge as k increases.

6.3. Impact of the Upper Bound (w) and Outlier Suppression

Figure 19 shows the impact of providing a w upper bound (i.e., using $[k, w]$-clustering instead of k-clustering; note that in these charts $w = 100$ corresponds to the default case where no upper bound is provided).

As Figures 19(a) and (b) show, providing looser (i.e, higher) upper bounds on the document cluster sizes tends to reduce the line density as well as the crossings, thus eliminating visual clutter. However, as can be seen in Figures 19(c) and (d), this also corresponds to relatively high visual compression rates.

One way to reduce the loss due to visual compression is to provide tighter document cluster size upper bounds. As Figures 19(c) and (d) show, an upper bound of $\sim 25\%$ is able to reduce the visual compression rate significantly (in some cases very close to 0.0%), without causing significant document suppression ($< 5\%$ in these experiments)). A quick look at Figures 16(a) and (b) also confirms that the amount of visual clutter is still very low at 25% upper bound rate.

Table II. The Scoring Rubric for the Subjective Assessments.

1	2	3	4	5
Not at all	Maybe no	Maybe yes	I think so	Yes, no doubt about it

Further reducing the document cluster size upper bound has three undesirable impacts: First of all, the rate of document suppression jumps to significant rates. Secondly, the resulting increase in the number of document clusters would imply that line density and other visual clutter measures, including the line crossings, would increase. Finally, pushing the document cluster size upper bound further down than 25% may in fact cause a rebound in the information loss (visual compression) as, after the resulting suppression of a large portion of outlier documents, the system would try to merge as many of the new clusters of documents as possible to resist the increase in visual clutter.

7. USER STUDIES

We conducted a series of user studies to understand whether (and why) our proposed system, Parallel hierarchical Coordinates (PhC) is effective in helping users observe and understand text corpora.

The participants that were involved in this study had different technical backgrounds. None of them was an expert in the data visualization domain and they were not part of the team that has developed PhC. The study is divided in four sets of experiments.

(1) subjective assessment of the usefulness of PC-based approaches for the exploration of text documents, in comparison with other schemes (21 participants);
(2) task-oriented assessment of the effectiveness of the proposed approach in communicating specific data patterns, in comparison with other schemes (18 participants);
(3) assessment of the effectiveness of our proposed scheme PhC with respect to the standard PC (19 participants);
(4) verification of the information loss measures used for the analytical evaluation (21 participants).

Depending on the particular evaluation goal, we have used both subjective and task-oriented evaluation strategies. In subjective studies, users were asked to respond to a series questions about PC, PhC, and other visualization schemes; Table II lists the rubric scale used in the subjective studies.

For these studies, again depending on the particular evaluation goal, we used both real and synthetic data. If specified otherwise, the studies were done on the Katrina news dataset described earlier.

7.1. Effectiveness of Parallel Coordinates in Visualizing Patterns in Data Collections

Although PC visualization [Inselberg and Dimsdale 1990] is not our contribution, our first goal was to verify whether using a PC-based scheme for visualizing data collections is indeed the right strategy.

7.1.1. Subjective Assessment. Therefore, this subjective user study aimed at assessing the effectiveness of PC-based visualizations for the exploration of collections. For this study, we constructed a synthetic dataset with 4 visualization attributes and 10 data entries. The data entries were created such that they form two different clusters: the 4-dimensional centroids of the clusters $c_i = < c_{i,1}, c_{i,2}, c_{i,3}, c_{i,4} >$ were randomly chosen according to a normal distribution with mean 0.5 and variance 0.5, while the

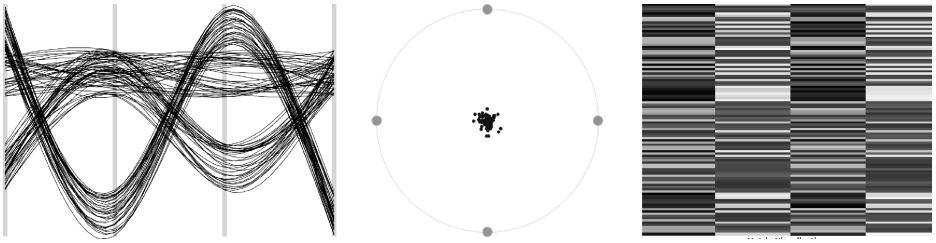


Fig. 20. Three different schemes visualinge three clusters: $c_1 = < 0.3, 0.7, 0.3, 0.7 >$, $c_2 = < 0.7, 0.7, 0.7, 0.7 >$, and $c_3 = < 0.9, 0.1, 0.9, 0.1 >$.

Table III. Average Participant Scores for Effectiveness of Parallel Coordinates in Visualizing Patterns in Data Collections (N indicates scores from the nonexperts and E indicates responses from the data management experts)

Question	Radviz (N)	Matrix (N)	PC-baed (N)	Radviz (E)	Matrix (E)	PC-based (E)
Q1	**3.07** / 5	2.92 / 5	2.31/ 5	2.57 / 5	2.86 / 5	**3.14** / 5
Q2	3.30 / 5	**3.61** / 5	2.89 / 5	3.57 / 5	3.00 / 5	**4.85** / 5
Q3	**3.46** / 5	3.07 / 5	2.69 / 5	2.57 / 5	2.29 / 5	**3.28** / 5

data entries were randomly generated around these centroids as $< v_{i,1}, v_{i,2}, v_{i,3}, v_{i,4} >$, where $v_{i,j}$ is a normally distributed random variable with mean $c_{i,j}$ and variance 0.2.

The participants were presented with three scenarios, each scenario were visualized using Radviz [Hoffman et al. 1997], matrix-oriented visualization[Keim 2002], and PC-based schemes (see Figure 20 for samples), and the participants were asked to respond to the following three questions for each case.

(*Q1*) Are you able to observe and discriminate the relationships (e.g., similarity, difference) among the selected terms?
(*Q2*) Are you able to observe and discriminate the relationships (e.g., similarity, difference) among the corresponding documents in the collection?
(*Q3*) How intuitive do you think this tool is for exploring the relationships among the terms and documents in the collection?

The participants were classified in advance into two groups based on their expertise in databases and data mining. Even though none of the participants was an expert in data visualization, 8 of the participants (which we refer to as *experts*) had experience in data management, while 13 users (which we call *nonexperts*) did not have any knowledge in the data management area.

The results, reported in Table III, indicate a major difference between nonexperts in data management and experts. According to these results, subjectively, nonexperts preferred Radviz or Matrix visualizations, whereas (again subjectively) data management experts did not prefer these visualization schemes. This, we believe, was the case because (while a PC-based visualization scheme looks less familiar to users at the first sight) data managements experts are able to perceive patterns in the data using a PC-based approach better than they do with Radviz or Matrix.

7.1.2. Task-Oriented Assessment. We next considered task-oriented verification of the effectiveness of the proposed visualization approach. In particular we focused on tasks involving identification of the numbers of clusters in the data. For this purpose, we again considered three multidimensional data visualization techniques: the

Table IV. Percentage of the Correct Guess of the Number of Clusters

Type of scenario	PC-based	Radviz	Matrix
single-shift scenarios	58.3%	16.7%	33.3%
polar equi-distance scenarios	91.7%	0%	8.3%
equi-ratio scenarios	91.7%	0%	0%
Avg	80.6%	5.6%	13.9%

pixel-oriented matrix visualization [Keim 2002], Radviz [Hoffman et al. 1997], and our proposal.

In order to better observe the advantages and disadvantages of different schemes in helping users identify the data clusters, we focused on challenging situations with similar, difficult to distinguish clusters coexisting in the dataset. In particular, we considered three types of data distribution scenarios.

Single-shift scenarios. This scenario consists of data clusters that differ from each other in only one dimension. For example, two data clusters randomly generated around $< 0.6, 0.6, 0.6, 0.7 >$ and $< 0.6, 0.6, 0.6, 0.5 >$ would give us a 4-dimensional single-shift scenario.

Equiratio scenarios. In this scenario, the data clusters have similar compositions along the different dimensions. For instance, a pair of clusters around points $< 0.6, 0.6, 0.6, 0.6 >$ and $< 0.3, 0.3, 0.3, 0.3 >$ is an example of this type of scenario.

Polar equidistance scenarios. This set of scenarios specifically focuses on the Radviz visualization; in particular, those dimensions that would be placed opposite to each other in the radial visualization are given similar values. A pair of clusters randomly generated around $< 0.3, 0.7, 0.3, 0.7 >$ and $< 0.7, 0.7, 0.7, 0.7 >$ would be an example. Figure 20 provides an example: here, the three visualization schemes visualize 3 clusters $c_1 = < 0.3, 0.7, 0.3, 0.7 >$, $c_2 = < 0.7, 0.7, 0.7, 0.7 >$, and $c_3 = < 0.9, 0.1, 0.9, 0.1 >$.

We constructed two scenarios (with two and three clusters respectively) for each of these types for a total of six datasets. Given a scenario with n cluster centroids and 4 visualization attributes, we generated $m = 50$ data entries for each centroid such that they form n different clusters around the 4-dimensional cluster centroids $c_i = < c_{i,1}, c_{i,2}, c_{i,3}, c_{i,4} >$. The data entries were computed as $< v_{i,1}, v_{i,2}, v_{i,3}, v_{i,4} >$, where $v_{i,j}$ is a normally distributed random variable with mean $c_{i,j}$ and variance 0.2.

We finally built three cases of six random visualizations each, and each of the 18 participants was asked to work only on one of them. The participants were presented with the three visualizations of different scenarios (sequentially in random order) and asked to identify the numbers clusters in the data. The random order ensured data participants could not guess the number of clusters for one visualization using the hints from a previous visualization for the same scenario. The participants were also not told which was our contribution.

Table IV shows the results of the study in terms of the percentage of cases where the number of clusters were correctly identified by the users. As these results demonstrate, PC-based visualizations helped the participants achieve 80% of accuracy on average, whereas the other visualizations were not effective on these difficult scenarios. The single-shift scenarios proved more difficult cases for PC-based schemes; but Radviz and Matrix faired worse even in those scenarios. In the polar equidistance scenarios, we were expecting Radviz to be ineffective and this was indeed the case. Interestingly, however, the matrix visualization also fared poorly for those scenarios. Furthermore, in equiratio scenarios users of neither the Radviz nor the Matrix visualizations could identify a single case correctly.

Table V. Qualitative Questions: Average Values

Qualitative questions	PC-based	Radviz	Matrix
Is it easy to use?	3.82 / 5	3.47 / 5	1.59 / 5
Is it effective?	4.12 / 5	2.82 / 5	1.71 / 5
Does it give a good overview of the distribution of documents and features?	83.4% of the users	17.6% of the users	0% of the users

Although, in this set of experiments, our focus was to objectively evaluate the users' preferences on difficult cases, we also asked each participant to provide subjective feedback through an exit questionnaire. The users were asked to give a rating from 1 to 5 (see Table II) to questions about the ease and the effectiveness of the three schemes. The results reported in Table V confirm our previous observations about the user perception and the effectiveness of the three schemes.

7.2. Effectiveness of k-Generalizations with PhC

After confirming the potential of PC for data collection visualization, our next goal was to see whether the proposed PhC approach provides additional benefits. For this purpose, we ran a second set of subjective studies where participants were asked to compare standard PC visualization with PhC using the same document collection. For this purpose, we selected three combinations of terms which express three different facets of the discussion on events related to hurricane Katrina.

Case 1 — terms: bush, state, louisiana. Analysis of the news about the relationships between the government and the management of the emergency in the state of Louisiana.

Case 2 — terms: hurricane, damage, louisiana, mexico. Analysis of the news about the geographic extent of the physical damages of the hurricane.

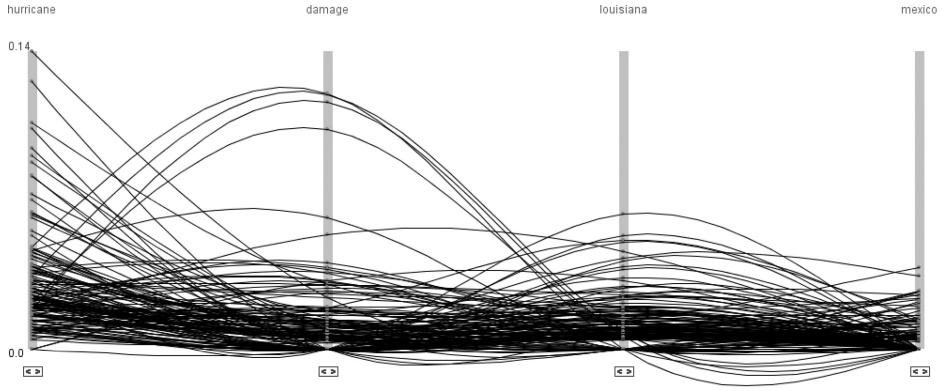
Case 3 — terms: hurricane, oil, price, production, gas, energy. Analysis of the news about the implications of the hurricane on oil and energy production in the region.

For each of the aforesaid cases, we presented to the participants the standard PC visualization of the data as well as the k-generalized version (where $k = 2$). Figure 21 shows an example. The participants were asked to respond to a set of questions in which they subjectively evaluated the effectiveness of these PhC approach.

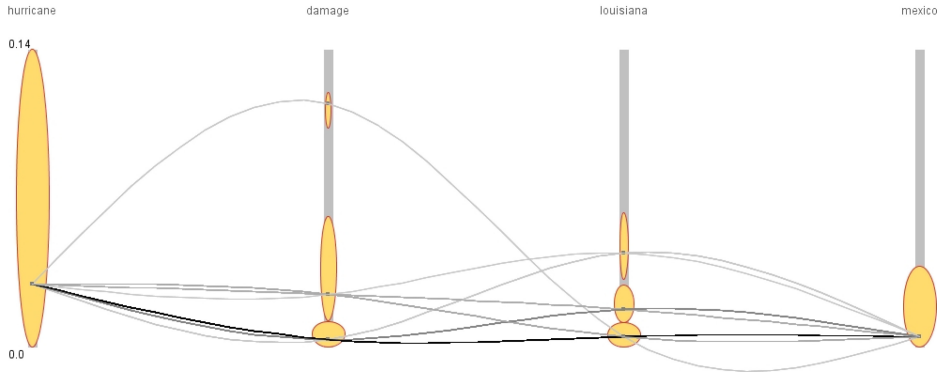
(*A.*) This clustering preserves the existing relationships between the terms and the documents.
(*B.*) Clustering makes the relationships between the terms and the documents easier to observe and understand.

Assertion **A** is to understand if the users feel that the PhC scheme preserves the main patterns in the data. The second assertion, **B**, provides feedback on how effective PhC is in visualizing such patterns with respect to the standard PC approach. Again, for both cases (or contexts), the users had to provide a score (from 1 to 5 as described in Table II).

The average user ratings for both assertions are shown in Table VI. These results indicate that, in all cases (from the simplest one with three parallel coordinates to the third case with six) the users were positive (between "Maybe yes" and "I think so") in the validities of both assertions. As expected, given that there is some information loss in the generalization process, it is not surprising that the participants were reluctant to give the rating of "Yes, no doubt about it". Therefore, in the next study we will assess whether the information loss versus visual clutter reduction results that showed the advantages of PhC over PC were based on valid loss measures.



(a) Parallel Coordinates Visualization for Case 2.



(b) Parallel hierarchical Coordinates Visualization for Case 2.

Fig. 21. Given a set of context keywords (case 2, in this figure), the users were asked to compare the standard PC visualization with the k-generalized version of the system PhC.

Table VI. User Ratings Values for PhC

Case	Assertion A	Assertion B	Avg
1	3.72 / 5	3.56 / 5	3.64 / 5
2	3.94 / 5	3.72 / 5	3.83 / 5
3	3.72 / 5	3.50 / 5	3.56 / 5
Avg	3.79 / 5	3.59 / 5	3.68 / 5

7.3. Appropriateness of the Quantitative Measures

In Section 6, we had observed experimentally that PhC was highly effective in reducing visual clutter while minimizing visual loss. These results were, on the other hand, based on loss measures presented in Section 5 and were based on the assumption (common in the literature [Koffka 1999]) that the area between the curves on the display reflected the distance perceived between the curves. In this set of experiments, we validate whether the distances between the curves on the PhC display are indeed correlated to the distances users perceive when provided with the original data entries.

In this study, we primarily focused on observing the correlation between: (a) the 21 participants' assessments of the curve similarities and (b) the similarity assessments of the curves based on the area-based distance measure described in Section 5. We

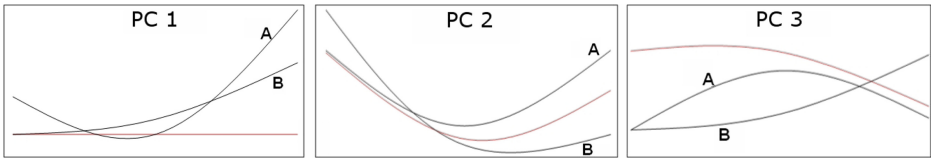


Fig. 22. Curve-based visualizations of the three tables in Figure 23. The participants were asked whether they think the data entry represented by curve A or curve B appears to be more similar to the entry corresponding to the curve highlighted in red.

table 1		
1	1	1
1	1.5	3
2	1	4

table 2		
3	1	2
3	1.5	3
4	1	1

table 3		
3	3	2
1	2.5	1.5
1	1.5	3

Fig. 23. The three tables used for validating the information loss measures defined in Section 5. In each table, the first row (in bold) corresponds to the target data entry; the second and third rows correspond to the candidate entries. The participants were asked to judge which candidate entry was more similar to the target entry.

provided each one of the participants the data curve plots shown in Figure 22 and asked them to select the curve (among the black ones) that appears to be more similar to the red curve. The results has shown that similarity assessments were very highly correlated (correlation $\sim$ 1.0). This supports the appropriateness of the distance (and consequently the loss) measures we used in our experiments.

Secondly, we also considered whether participants' similarity assessments for the curves are correlated well with their similarity assessments if they are provided the pure data. For this purpose, we also provided them the three data tables, shown in Figure 23, corresponding to the these curves and asked them to pick among the two candidate data entries the one that they think is most similar to the highlighted data entry. In this study, 20 out of the 21 participants made similar similarity selections when given the data tables and when given the corresponding curve plots. Only one participant made a different selection in one of the tables for the two schemes (the participant selected the second row for Table 3 and the curve *B* for the plot *PC 3*; see Figures 22 and 23). Along with the previous results showing the effectiveness of the curve-based visualizations, this supports the observation that the curves used in PhC have the potential to correctly capture the data similarity/distance judgments of users.

8. CONCLUSIONS AND FUTURE WORK

In this article, we presented a new visualization mechanism, called Parallel hierarchical coordinates (or PhC) for supporting the visualization and exploration of document collections. At its core, PhC relies on a parallel-coordinates-based approach, where multidimensional vectors are mapped onto a 2D space in such a way that documents with similar term frequencies are represented as similar poly-lines or curves in the visualization space. PhC associates a value-clustering hierarchy to each visualization coordinate (e.g., term provided by a user) and leverages these hierarchies to reduce visual clutter in the visualization with minimal information loss. The user can then interact with the system to selectively roll-up or drill-down along the different visualization coordinates to explore the patterns in the document set, without being overloaded with visual clutter.

In our future works, we will consider the selection of the dimensions to be used for generalization. We also plan to investigate the impact of the number of dimensions used for generalization and their ordering on the reduction of visual clutter and

information loss and carry out user studies designed to investigate the effectiveness of the proposed approach and its future extensions also in higher-level tasks (e.g., pattern seeking).

REFERENCES

AGGARWAL, G., TOMAS FEDER, K. K., MOTWANI, R., PANIGRAHY, R., THOMAS, D., AND ZHU, A. 2005. Approximation algorithms for k-anonymity. *J. Privacy Technol.*

ANKERST, M., KEIM, D. A., AND KRIEGEL, H.-P. 1996. Circle segments: A technique for visually exploring large multidimensional data sets. In *Proceedings of the Visualization Conference*.

ANKERST, M., BERCHTOLD, S., AND KEIM, D. A. 1998. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Proceedings of the IEEE Symposium in Information Visualization (INFOVIS)*.

ARTERO, A. O., DE OLIVEIRA, M. C. F., AND LEVKOWITZ, H. 2004. Uncovering clusters in crowded parallel coordinates visualizations. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*. IEEE Computer Society.

BARTELS, R., BEATTY, J., AND BARSKY, B. 1995. *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann.

BATES, M. J. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online Rev. 13*, 5, 407–424.

BORZSONYI, S., KOSSMANN, D., STOCKER, K., AND PASSAU, U. 2001. The skyline operator. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 421–430.

CHINTALAPANI, G., PLAISANT, C., AND SHNEIDERMAN, B. 2004. Extending the utility of treemaps with flexible hierarchy. In *Proceedings of the International Conference on Information Visualisation*. 335–344.

CIRIANI, V., DE CAPITANI DI VIMERCATI, S., FORESTI, S., AND SAMARATI, P. 2007. k-anonymity. In *Secure Data Management in Decentralized Systems*, T. Yu and S. Jajodia Eds., Springer, 323–353.

COX, T. AND COX, M. 2001. *Multidimensional Scaling*. Chapman Hall.

CUI, Q., WARD, M., RUNDENSTEINER, E., AND YANG, J. 2006. Measuring data abstraction quality in multiresolution visualizations. *IEEE Trans. Vis. Comput. Graph.* 709–716.

DEERWESTER, S. C., DUMAIS, S. T., FURNAS, G. W., HARSHMAN, R. A., AND LANDAUER, T. K., ET AL. 1989. Computer information retrieval using latent semantic structure. `http://www.mendeley.com/research/computer-information-retrieval-using-latent-semantic-structure-1/`.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *J. Amer. Soc. Inf. Sci. 41*, 6, 391–407.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. 39*, 1, 1–38. Series B.

DI CARO, L., CANDAN, K. S., AND SAPINO, M. L. 2008. Using tagflake for condensing navigable tag hierarchies from tag clouds. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, New York, 1069–1072.

DON, A., ZHELEVA, E., GREGORY, M., TARKAN, S., AUVIL, L., CLEMENT, T., SHNEIDERMAN, B., AND PLAISANT, C. 2007. Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*. ACM, New York, 213–222.

ECKART, C. AND YOUNG, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika 1*, 3, 211–218.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O., AND BOTSTEIN, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA 95*, 25, 14863–14868.

ELLIS, G. AND DIX, A. 2006. Enabling automatic clutter reduction in parallel coordinate plots. *IEEE Trans. Vis. Comput. Graph. 12*, 5, 717–724.

FORTUNA, B., GROBELNIK, M., AND MLADENIC', D. 2005. Visualization of text document corpus. *Informatica*, 497–502.

FUA, Y.-H., WARD, M. O., AND RUNDENSTEINER, E. A. 1999a. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of the 10th IEEE Visualization Conference (VIS'99)*. IEEE Computer Society.

FUA, Y. H., WARD, M. O., AND RUNDENSTEINER, E. A. 1999b. Navigating hierarchies with structure-based brushes. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'99)*. G. Wills and D. Keim Eds.

GRAHAM, M. AND KENNEDY, J. 2003. Using curves to enhance parallel coordinate visualisations. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'03)*.

HASSAN-MONTERO, Y. AND HERRERO-SOLANA, V. 2006a. Improving tag-clouds as visual information retrieval interfaces. In *Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies (InScit'06)*.

HASSAN-MONTERO, Y. AND HERRERO-SOLANA, V. 2006b. Interfaz visual para recuperacin de informacin basada en anlisis de metadatos, escalamiento multidimensional y efecto ojo de pez. *El Profesional de la Información 15*, 4.

HOFFMAN, P., GRINSTEIN, G., MARX, K., GROSSE, I., AND STANLEY, E. 1997. Dna visual and analytic data mining. In *Proceedings of the 8th Conference on Visualization (VIS'97)*. IEEE Computer Society Press.

HOLTEN, D. 2006. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Trans. Vis. Comput. Graph.*, 741–748.

INSELBERG, A. AND DIMSDALE, B. 1990. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *Proceedings of the Visualization Conference (VIS)*. 361–378.

KEIM, D. 2002. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Trans. Vis. Comput. Graph. 6*, 1, 59–78.

KOFFKA, K. 1999. *Principles of Gestalt Psychology*. Psychology Press.

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2005. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the ACM SIGMOD Conference on Management of Data*. 49–60.

LI, N. AND LI, T. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of IEEE International Conference on Data Engineering*.

LIN, Y.-R., SUN, J., CAO, N., AND LIU, S. 2010. Contextour: Contextual contour analysis on dynamic multi-relational clustering. In *Proceedings of the SDM Conference*. 418–429.

MACHANAVAJJHALA, A., KIFER, D., GEHRKE, J., AND VENKITASUBRAMANIAM, M. 2007. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data 1*, 1, 3.

MEYERSON, A. AND WILLIAMS, R. 2004. On the complexity of optimal k-anonymity. In *Proceedings of the PODS Symposium on Principles of Database Systems*. 223–228.

NOVOTN, M. 2004. Visually effective information visualization of large data. In *Proceedings of the Central European Seminar on Computer Graphics (CESCG)*.

PENG, W., WARD, M. O., AND RUNDENSTEINER, E. A. 2004. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization*. IEEE Computer Society, 89–96.

SAMARATI, P. 2001. Protecting respondents' identities in microdata release. *Trans. Knowl. Discov. Engin. 13*, 6, 1010–1027.

SAMARATI, P. AND SWEENEY, L. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*.

SEO, J. AND SHNEIDERMAN, B. 2004. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*. IEEE Computer Society, 65–72.

SHNEIDERMAN, B. 1992. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph. 11*, 1, 92–99.

SIIRTOLA, H. 2000. Direct manipulation of parallel coordinates. In *Extended Abstracts on Human Factors in Computing Systems (CHI'00)*. ACM, New York, 119–120.

SONG, Y., PAN, S., LIU, S., ZHOU, M. X., AND QIAN, W. 2009. Topic and keyword re-ranking for lda-based topic modeling. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management (CIKM)*. 1757–1760.

TUFTE, E. R. 2001. *The Visual Display of Quantitative Information* 2nd Ed. Graphics Press.

WARD, J. 1963. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc. 58*, 236–244.

WONG, P. C. AND BERGERON, R. D. 1996. Multiresolution multidimensional wavelet brushing. In *Proceedings of the 7th Conference on Visualization (VIS'96)*. IEEE Computer Society Press, 141–ff.

WONG, P. C. AND BERGERON, R. D. 1997. Multivariate visualization using metric scaling. In *Proceedings of the 8th Conference on Visualization (VIS'97)*. IEEE Computer Society Press, 111–ff.

YANG, J., PENG, W., WARD, M. O., AND RUNDENSTEINER, E. A. 2003. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of the IEEE Symposium on Information Visualization*.

YUAN, X., GUO, P., XIAO, H., ZHOU, H., AND QU, H. 2009. Scattering points in parallel coordinates. *IEEE Trans. Vis. Comput. Graph. 15*, 6, 1001–1008.

ZHOU, H., YUAN, X., QU, H., CUI, W., AND CHEN, B. 2008. Visual clustering in parallel coordinates. *Comput. Graph. Forum 27*, 3, 1047–1054.

ZHOU, H., CUI, W., QU, H., WU, Y., YUAN, X., AND ZHUO, W. 2009. Splatting the lines in parallel coordinates. *Comput. Graph. Forum 28*, 3, 759–766.

Received July 2010; revised February 2011; accepted April 2011