Expert Systems with Applications 39 (2012) 12143–12157

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Visualization of temporal text collections based on Correspondence Analysis

Artur Šilić [a,*], Annie Morin [b], Jean-Hugues Chauchat [c], Bojana Dalbelo Bašić [a]

[a] *University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia*
[b] *IRISA, Université de Rennes 1, 35042 Rennes Cedex, France*
[c] *ERIC-Lyon2, Université de Lyon, 5 av. Pierre Mendès-France, 69676 Bron Cedex, France*

ARTICLE INFO

Keywords:
Information visualization
Singular Value Decomposition
Clustering
Text analytics

ABSTRACT

In this paper, we present CatViz—Temporally-Sliced Correspondence Analysis Visualization. This novel method visualizes relationships through time and is suitable for large-scale temporal multivariate data. We couple CatViz with clustering methods, whereupon we introduce the concept of final centroid transfer, which enables the correspondence of clusters in time. Although CatViz can be used on any type of temporal data, we show how it can be applied to the task of exploratory visual analysis of text collections. We present a successful concept of employing feature-type filtering to present different aspects of textual data. We performed case studies on large collections of French and English news articles. In addition, we conducted a user study that confirms the usefulness of our method. We present typical tasks of exploratory text analysis and discuss application procedures that an analyst might perform. We believe that CatViz is general and highly applicable to large data sets because of its intuitiveness, effectiveness, and robustness. We expect that it will enable a better understanding of texts in huge historical archives.

1. Introduction

In the age of information overflow, media analysts and others have to be able to access high-level information summaries. Consider the following question: what were the main news events in the past, and what are they today? To answer this important question in our information-centric society, interested readers might try to find a summary written by historians or experienced journalists. Another approach would be to analyze the data itself using statistical methods, as most of the important text media sources are digitalized, and vast computing power is available. This kind of purely objective information analysis will never achieve the quality of a historian's well-written essay, but it will enable independent exploration of trends in huge document collections with practically no prior bias caused by human interpretation. Additionally, little starting knowledge is needed, and the quantity of texts processed in this way can be much higher than the quantity of texts manually retrieved and read.

This paper presents CatViz—**T**emporally-Sliced **C**orrespondence **A**nalysis **Vi**suali**z**ation. This novel method visualizes relationships through time and is suitable for large-scale temporal multivariate data. CatViz uses Correspondence Analysis (CA) (Greenacre, 2007), a dimension-reduction method based on Singular Value Decomposition (SVD). CatViz executes quickly, which makes it well suited for the interactive exploration of very large text collections.

CA was used on textual data prior to this work. Early works include an analysis of research reports (Morin, Kerbaol, & Bansard, 2000). Kerbaol, Bansard, and Coatrieux (2006) briefly introduced the idea to see overall year trends while they were using CA to analyze texts of IEEE journals. By the same means, they studied the evolution of each IEEE journal in the biomedical field and made inferences about content change through years of those publications (Bansard, Kerbaol, & Coatrieux, 2006). We extend the idea of visualizing temporal change in many aspects. We couple CatViz with clustering and we use more advanced linguistic resources. We present a detailed analysis on two large corpora and show which different temporal patterns can be detected. Among other things, in our case studies, we present the important finding that distributional similarities of different corpora which are written in different languages can be easily seen in plots. Additionally, specific to text collections, we show that it is useful to distinguish between names and other words in the text. Finally, we describe a user study that confirms the usefulness of CatViz.

The work is structured as follows: We present related work in Section 2, and we formally describe our method in Section 3. We explain the application procedures in Section 4. Case and user studies are described in Sections 5 and 6. Section 7 concludes the paper.

2. Related work

As described in a recent survey (Šilić & Dalbelo Bašić, 2010), there are two main approaches to the visualization of

[*] Corresponding author.
E-mail address: artur.silic@gmail.com (A. Šilić).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.04.040

time-stamped text collections. In the following subsections we describe these approaches and compare them to our method.

2.1. Term trend approach

The most straightforward way to visualize trends in temporal text collections is to plot the frequencies of important terms (or their groups) during a given window of time. Feature selection is employed to reduce the number of signals and prevent visual occlusion. The first and extremely popular visualization technique using the term trend approach is ThemeRiver (Havre, Hetzler, & Nowell, 2000). NewsRiver, a similar, somewhat refined method, is employed in (Ghoniem, Luo, Yang, & Ribarsky, 2007). These methods can be useful when the user is concentrating on a shorter period of time and when the terms are known in advance or are generally very frequent. Next, the EventRiver technique (Luo, Yang, Krstajic, Ribarsky, & Keim, 2010) is oriented toward event detection and visualization. Events are labeled with certain words, which can be regarded as a sophisticated method of term selection. The MemeTracker (Leskovec, Backstrom, & Kleinberg, 2009) plots trends by efficiently detecting memetic phrases, which can be regarded as an advanced phrase selection.

Although very useful, the term trend approach is not suited for systematic visualization of similarities among time intervals. Overcoming this gap was one of the incentives for the research and design of our method.

2.2. Semantic space approach

Within this approach, each part of the view corresponds to some semantic category from the text collection. The texts are usually represented in a Vector Space Model (VSM) (Salton & McGill, 1983). The vectors representing texts are of high dimensions because textual features are numerous. Thus, dimensionality reduction techniques are employed to map these vectors to 2D or 3D space, and to obtain the most informative components by certain measures and to reduce noise. Dimension reduction is often based on Singular Value Decomposition (SVD), Multidimensional Scaling (MDS), or Force Directed Placement (FDP), but other approaches also exist (Šilić & Dalbelo Bašić, 2010).

Petrović, Dalbelo Bašić, Morin, Zupan, and Chauchat (2009) used CA on news texts and found the most suitable features, but did not perform any temporal analysis.

IN-SPIRE (Wise et al., 1995) and VxInisight (Davidson, Hendrickson, Johnson, Meyers, & Wylie, 1998) use the galaxy and landscape visual metaphors to depict strong thematic clusterings of a text collection in 2D space. These classic methods do not include temporal information, but the authors of (Davidson et al., 1998) indicate that trend discovery can be achieved by time slicing—the prevalent modality of trend discovery by methods that use the semantic space approach. Time slicing constrains a series of views to a series of time intervals, so the visualization method itself does not need to include the temporal information. By analyzing differences among plots, the user gains insight into changes in the text stream. Time slicing has been explicitly noted in a number of articles on visualization (including Davidson et al., 1998; Albrecht-Buehler, Watson, & Shamma, 2005), but it has also been criticized on account of limitations in human memory and change blindness in perception (Luo et al., 2010).

TextPool (Albrecht-Buehler et al., 2005) and STREAMIT (Alsakran, Chen, Zhao, Yang, & Luo, 2011) are text stream visualizations based on FDP. TextPool visually presents the relationships among terms, and STREAMIT depicts similarities of whole documents. Temporal relations are not explicitly shown.

Wong, Foote, Adams, Cowley, and Thomas (2003) used a combination of MDS and wavelets to project documents into a low-dimensional space. The main distinction of this is the ability to process streams by reusing the subspace calculated on previous documents and projecting new incoming data on the low-dimensional space. Although the temporal information is not directly seen on the plot, temporality is used to scale the calculation.

A technique similar to ours has been proposed by Kaban and Girolami (2002). Its target is to complement topic detection and tracking applications. The underlying methods are Hidden Markov Models (Roweis, 1999) and Latent Semantic Indexing (LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). In contrast to our work, this method is oriented toward particular documents. Therefore, it can introduce high computational costs when analyzing large document collections.

The Document Atlas system (Fortuna, Mladeni, & Grobelnik, 2005) combines LSI and MDS to perform dimension reduction and generate a semantic space. Although the original work was not concerned with time, an extension (Fortuna, Mladeni, & Grobelnik, 2009) makes it possible to visualize contexts of named-entities in news and their evolution over time. Our work differs in two aspects: we aim to visualize the relationships among time intervals and trends of all entities and words in the texts, not only the contexts of individual names. By doing so, we enable a more general tracking of trends and interaction among many entities and time intervals. Our choice of underlying method is different (CA), enabling a symmetric analysis of time intervals and features.

An approach similar to ours is described in (Mao, Dillon, & Lebanon, 2007). Sequential Document Visualization aims to visualize a single text, and the authors mainly concentrated on a smoothed text representation because low word number in a single document makes certain dimensions noisy. PCA is used to produce the final plots. Conversely, we visualize many texts at once. In addition, we include research on the utility of linguistic resources.

A novel approach to the visualization of time series is presented in (Alencar, de Oliveira, Paulovich, Minghim, & Andrade, 2007). The main idea was to introduce a similarity measure between time-varying variables and then to position a point representing each variable on a projection and try to preserve the original distances. This is an interesting approach, but in terms of the matrix that is to be visualized, the plot concentrates on relationships among columns, and the temporal data and individual row data are not directly shown on the plot.

Each of the above cited works differs from our work: our method explicitly includes temporal information; is oriented toward huge numbers of documents enabling high-level overviews; and provides a systematic and symmetric analysis of temporal correlations of time intervals and terms occurring in the texts. The last property stems from the fact that CA is the only dimension reduction technique that enables a symmetric analysis of relationships among the rows and columns of a matrix that is visualized.

3. Visualization

The visualization is produced in two steps. First, the data are represented using a Vector Space Model. Second, CatViz, a novel visualization method is used which enables analysis of any categorical data that is time-stamped. Essentially, it consists of representing each time interval with one or more profiles and then using Correspondence Analysis to reduce dimensionality and produce a visualization.

3.1. Vector space construction

In information retrieval, a text is often represented as a *bag-of-words* (BOW), an instance of the VSM. A BOW vector is constructed by counting all word occurrences and assigning them to their corresponding dimensions. Because the semantic value of two word forms from the same lexeme is very similar, and in order to strengthen word occurrences, morphological normalization is often employed.

3.1.1. Feature choice

Baseline feature choice includes all words found in the text, (*words*). However, some words are more informative than others. In the case of newspaper articles, names of persons, organizations, countries, and other entities carry the most information regarding the Who and Where questions. This fact inspired us to keep only these features, (*names*), when constructing the BOW vector. By choosing to visualize only names, the generated plot will concentrate only on entities that are actors, and words that describe events and trends in more detail will be discarded. Complementary to this selection, all words that are not names are kept to yield a representation, (*no-names*), that focuses the visualization on the underlying meaning irrespective of the subjects and places involved.

These two kinds of features describe different aspects of text data. They also have different statistical properties and produce different plots; see Fig. 7, where names in sports fluctuate more intensively than other words, which reoccur in cycles. The conducted experiments confirmed our conjecture that basic feature filtering is important during text visualization.

3.2. CatViz method

3.2.1. Correspondence Analysis

Correspondence Analysis (CA) (Greenacre, 2007; Benzécri, 1973) is a dimensionality reduction method that satisfies the condition of minimum sum of squared χ^2 distances of points from the subspace on which the points are projected. Additionally, an alternative condition is that χ^2-weighted point projections are separated from one another as much as possible. An exact solution to this condition is obtained with SVD. CA can also be formulated using inertia—a measure of point dispersion. The calculated principal components, (PCs), are sorted in descending order by the inertia that they account for. That means that the first component accounts for the most variance.

During the same calculation, matrix columns are mapped onto the resulting subspace. These points are also called *vertices in standard coordinates*. A symmetric analysis of rows and columns is enabled by comparing their mutual distances on the projection. The closeness of two profile projections indicates their closeness in the original space, whereas a similar direction of profile and column points indicates their high correlation.

3.2.2. Representation of temporal changes

We are interested in changes of word usage in text collections during a given period of time. To visualize these changes, we divide the whole period T into smaller intervals t_i:

$$T = \bigcup_i t_i, \quad t_i = [t_{iS}, t_{iE}\rangle \subset \mathbb{R} \tag{1}$$

Each text a has a vector $\boldsymbol{x}(a)$ and a time stamp $t(a) \in T$ adjoint. Each interval t_i is represented as a sum of feature frequency vectors representing texts whose time stamp falls within the interval t_i:

$$\boldsymbol{n}_i = \sum_{\{\boldsymbol{x}(a) | t(a) \in t_i\}} \boldsymbol{x} \tag{2}$$

Representing each time interval with an aggregate vector has a dual purpose. First, it enables a simple representation and an intuitive interpretation of the interaction among different time intervals. Second, it enables the CA to be calculated because CA is based on SVD, whose calculation on a full document-term matrix would be unfeasible in the case of hundreds of thousands of documents. By *unfeasible*, we mean *computationally too demanding* for the equipment readily available to an average user. A good example of SVD calculation cost measurement is given in (Martin, Martin, Berry, & Browne, 2007), where experiments are done on textual term-document matrices. As we will see later, single texts can be projected on the principal space, preserving the functionality of particular text visualization.

CA is performed on the frequency matrix $\boldsymbol{N}$ whose rows $\boldsymbol{n}_i$ denote time intervals and whose columns denote text features. The following three equations describe the simple preprocessing steps.

$$\boldsymbol{P} = \frac{1}{\sum_{i,j} n_{ij}} \boldsymbol{N} \tag{3}$$

$$r_i = \sum_{j=1}^{J} p_{ij}, \quad \boldsymbol{D}_r = diag(\boldsymbol{r}) \tag{4}$$

$$c_j = \sum_{i=1}^{I} p_{ij}, \quad \boldsymbol{D}_c = diag(\boldsymbol{c}) \tag{5}$$

The following equation describes the crucial step in CA calculation. The expression to the right of the equality sign is the result of SVD performed on the matrix evaluated by the expression on the left:

$$\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}^T)\boldsymbol{D}_c^{-\frac{1}{2}} = \boldsymbol{U}\boldsymbol{D}_\alpha \boldsymbol{V}^T \tag{6}$$

Thus, the diagonal matrix $\boldsymbol{D}_\alpha$ containing sorted singular values is obtained. In addition, the orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are obtained.

The rows and columns of the original matrix $\boldsymbol{N}$ are represented in the principal space by rows of matrices $\boldsymbol{F}$ and $\boldsymbol{G}$, respectively:

$$\boldsymbol{F} = \boldsymbol{D}_r^{-\frac{1}{2}} \boldsymbol{U} \boldsymbol{D}_\alpha \tag{7}$$

$$\boldsymbol{G} = \boldsymbol{D}_c^{-\frac{1}{2}} \boldsymbol{V} \boldsymbol{D}_\alpha \tag{8}$$

Thus, we will find the projections of each time interval vector $\boldsymbol{n}_i$ as the ith row of F. In addition, if needed, a particular text vector $\boldsymbol{x}$ can also be projected onto the new space:

$$r_{\boldsymbol{x}} = \frac{1}{N} \sum_{j=1}^{J} x_j \tag{9}$$

$$\boldsymbol{f}_{\boldsymbol{x}} = r_{\boldsymbol{x}}^{-1} \left(\frac{1}{N} \boldsymbol{x} - r_{\boldsymbol{x}} \boldsymbol{c} \right) \boldsymbol{D}_c^{-\frac{1}{2}} \boldsymbol{V} \boldsymbol{D}_\alpha^{-1} \tag{10}$$

On the projection, the aggregation points of neighboring time intervals are connected with a line to emphasize their temporal closeness and to enable the user to continuously track the semantic orientation of the text collection through time.

By visualizing transformed vectors of time intervals and text features, we can see their mutual distances, allowing us to visually analyze their similarities, groupings, and associations.

3.2.3. Component selection

Matrices $\boldsymbol{F}$ and $\boldsymbol{G}$ are of sizes $I \times K$ and $J \times K$, respectively. The dimensionality of the lower-dimensional space is equal to $K = \min(I,J) - 1$, so to plot a 2D map, a choice (p,q) among the new components has to be made. A natural choice is to select the first two components, because they account for most of the explained variance. Apart from the first two, more combinations of foremost components are useful because different general trends are captured. For example, in the case study with Le Monde sports texts, first two components are seasonal, describing yearly cycles in sport coverage, whereas the third component captures the shift that occurred during the 1998 football World Championship; see Fig. 5 in Section 5.3.

3.2.4. Aggregation choice

Data can be visualized with an arbitrary granularity, enabling plots from high-level overviews down to very detailed views. Granularity is controlled by stretching or shrinking the length of time slice intervals $|t_i| = t_{iE} - t_{iS}$. In the captions of the presented plots, interval length in months is denoted with the symbol m.

One convenience of using CA is in that it adheres to the distributional equivalence property (DEP), which states the following (Greenacre, 2007): If two row profiles are identical (distributionally equivalent), then the corresponding two rows of the original data matrix may be replaced by a single row of their summation without affecting the geometry of the column profiles. The following inference explains why, in most cases, changing the granularity minimally affects the plot's geometry:

- Because of DEP, if some vectors are very similar, their aggregation will affect the plot minimally.
- If time-close vectors are generally more similar than time-distant vectors, aggregating time-close vectors will minimally affect the geometry of the plot.
- Aggregating time-close vectors is equivalent to changing the granularity.

The second proposition usually holds for nonseasonal data such as "World affairs", so in such cases, aggregation presents a sound method to continuously balance plot detail and computational performance. This can be seen on Fig. 1, where high- and low-aggregation points are close to each other and yield similar trajectory shapes.

3.2.5. Significance filtering

On a pair of chosen components (p,q), each original feature j contributes to the total inertia:

$$\mathrm{Inertia}(j;(p,q)) = (\boldsymbol{G}(j,p)^2 + \boldsymbol{G}(j,q)^2) \cdot c_j \qquad (11)$$

Similarly, an article vector $\boldsymbol{x}$ that has not been directly used to calculate CA can be given a weight based on the inertia calculation for the time profiles:

$$\mathrm{Inertia}(\boldsymbol{x};(p,q)) = \left(f_{\boldsymbol{x}p}^2 + f_{\boldsymbol{x}q}^2\right) \cdot r_{\boldsymbol{x}}, \qquad (12)$$

where $\boldsymbol{f_x}$ is the projection vector defined in (10). Features and articles with high contributions to inertia are considered to be of high significance. This valorization allows the user to visualize only the most important features and articles above some contribution threshold.

3.3. Coupling with clustering

Because a single text stream can discuss more than one significant topic at a time, representing each time interval with an aggregation vector that points in one direction of the semantic space might not be descriptive enough.

Single-aggregation representation is limited because an average of more vectors can be directed toward a part of the space not related to any of the original vectors and where no significant number of vectors exists. Other sources of limitation include the following. On the final plot with one aggregation vector per time interval, the vectors are projected to a subspace, and some subgroups of a time interval cannot be close to other similar subgroups of other time intervals because they are *chained* to other stronger subgroups of their own time interval. By introducing a semantic division of texts within a single time interval, the subgroups gain more freedom to *mingle* with semantically similar subgroups irrespective of their time-close neighboring subgroups.

A solution to this problem and a natural extension of the CatViz method is to have, for each time interval, more aggregation vectors that point in different directions of the semantic space. This is achieved with clustering, where a set of vectors from a single time interval is divided into a number of subsets on the criterion of semantic similarity.

Technically, clustering can be done with various distance functions and clustering algorithms, but for the purpose of demonstration, in our work we employed the simplest technique—K-means clustering (Hartigan, 1975) with a cosine similarity function. A standard setting of the K-means algorithm has a random factor

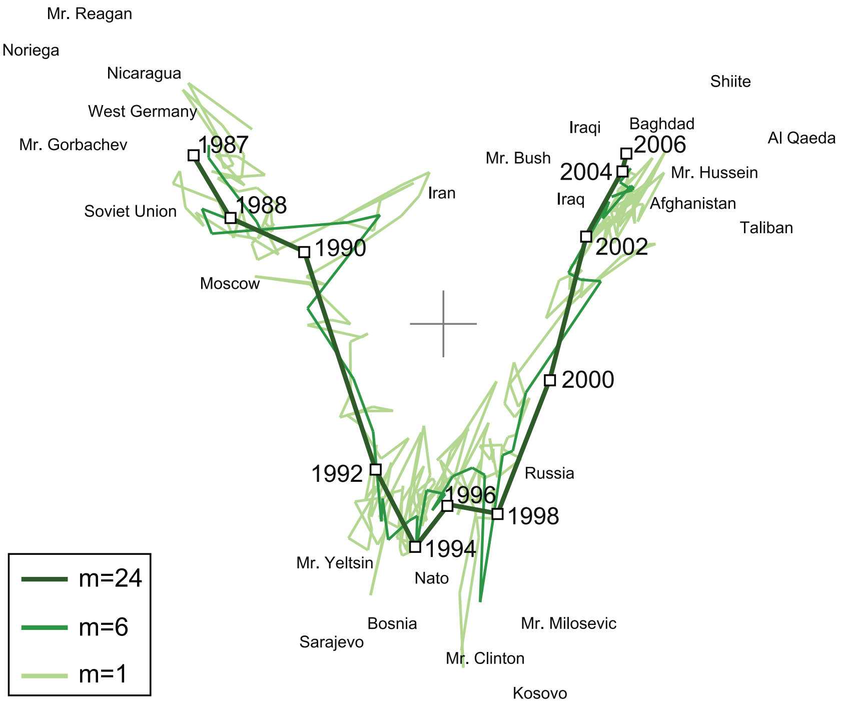


Fig. 1. NYT—World, 1987–2007, showing 124 934 texts, varying the aggregation length in months ($m = \{1,6,24\}$), features are names (f = names), 1st and 2nd principal components (PCs).

involved when choosing the initial centroids, and the result depends on this choice. Formally, for each interval t_i, the K-means algorithm clusters the set of vectors $V_i = \{\boldsymbol{x}(a) | t(a) \in t_i\}$ by trying to optimize the following criterion function Q_i:

$$Q_i = \sum_{l=1}^{K} \sum_{\boldsymbol{x} \in C_{i,l}} (1 - \cos(\boldsymbol{x}, \boldsymbol{c}_{i,l}))^2, \quad \boldsymbol{c}_{i,l} = \frac{1}{|C_{i,l}|} \sum_{\boldsymbol{x} \in C_{i,l}} \boldsymbol{x} \tag{13}$$

where $C_{i,l}$ is the lth cluster of the ith time interval, and $\boldsymbol{c}_{i,l}$ is its centroid.

3.3.1. Final centroid transfer

We introduce the final centroid transfer to let clusters of text evolve through all time intervals. For the first time interval, a standard K-means procedure is used to obtain clusters. For all subsequent time intervals, K-means is calculated using the final centroids from the preceding time interval as initial centroids.

$$\boldsymbol{c}_{i,l}^{(0)} = \boldsymbol{c}_{i-1,l}, \quad \text{for } i > 1 \tag{14}$$

The purpose of final centroid transfer is to associate clusters in neighboring time intervals. Alternatively, simple connecting strategies could be employed (e.g., using criteria such as nearest-first or minimum line crossings of connecting lines), but by using the final centroid transfer, this step is elegantly eliminated.

Similarly to the described algorithm, there are many data-stream clustering algorithms available. In particular, for text stream clustering see (Moerchen, Brinker, & Neubauer, 2007; Liu, Cai, Yin, & Fu, 2008). These algorithms calculate a continuous evolution of clusters by sequentially processing elements to be clustered. In our case, we work in larger discrete steps—the time intervals. In this light, the clustering with final centroid transfer can be regarded as a classic stream clustering algorithm, but one that makes updates to the centroids after longer periods.

3.3.2. Clustering choice

In general, the division of data based on its semantic orientation should be done with automatic or manual categorization prior to the visualization process. However, in the case when there are many documents per category and per time interval, clustering can be useful. The clustering has a dual function. First, to create substreams from a single stream and introduce a more detailed semantic division of the texts. Second, to give more significance to terms that have a constant flux in the original stream. An explanation follows. CA finds a subspace that best separates the projection of original points. In that way, the plot produced will give high inertia to the terms whose occurrence is concentrated in specific time intervals. Terms with almost constant flux (low variance) will not be of high inertia. By introducing clustering, some terms with constant flux that were equally distributed in all time slice vectors, will now be present in some time slice vectors but absent in others. This means that these terms will have greater variance and higher inertia on the plot.

4. Application procedures

4.1. Signal change

In this work, we classify temporal changes of a text collection into three types: bumps, shifts, and cycles. Bumps are short-term semantic changes where the stream quickly swings back to the starting orientation. Shifts are long-term changes in semantic orientation. There is no fundamental difference between a short bump and a long shift that eventually returns to its starting orientation. It is just a matter of definition between *long* and *short*. Similarly, in the field of time-series analysis, a signal is often described using an additive model (Falk, 2006):

$$Y_t = T_t + Z_t + S_t + R_t, \tag{15}$$

where T_t is a trend—a monotonous function, Z_t—a nonrandom long-term cyclic influence, S_t—a short-term cyclic influence such as seasonality, and R_t—a residual component describing the original signal's deviation from the sum of the previously introduced components.

Using the CatViz method is similar to using an additive model in two respects. First, in CA, the original signals can be obtained as linear combinations of the principal components, which themselves are linear combinations of terms. Second, our experiments have shown that the principal components obtained with the CatViz method largely realize T_t, Z_t or S_t. The first few components account for the changes that are of the most significance. Usually, those are long-term trends, very abrupt short changes or strong seasonal movements. Low-rank components describe more specific events of lower text coverage. We can achieve seasonal adjustment by omitting strong seasonal components. In contrast, by regarding only the seasonal components, we achieve trend adjustment.

The difference between CatViz and the simple additive model is in that when utilizing CatViz, the significance of the components is obtained.

4.2. Temporal analysis

The goal of temporal analysis is to identify and describe temporal changes of important features within a stream. More detailed questions that visualization users need to answer are as follows.

- How does the set of the most important words in a collection change during a period? The features most correlated to certain time intervals deviate from the center in the same direction. In addition, the notion of importance is introduced as the inertia of each point, so filtering is employed to give an arbitrary detailed projection.
- How is a time interval similar to other time intervals and in what respect? Does a semantic orientation repeat—meaning does history repeat? This is answered by comparing distances of projected time-interval profile points. Close time-interval points indicate similarity of the original feature points.
- How do the features relate in time? Similar to the previous question, this is answered by analyzing the mutual distances of features and time-interval points. If the feature points are close, that means that they have a similar distribution among time intervals.
- What types of change are present—bump, shift, or cycle? This question addresses more abstract properties of a text collection, and is answered by analyzing the temporal trajectories—a cyclic curve indicates an existing cyclic component in the original collection. Likewise for the bumps and shifts.
- When does the bump or shift begin to surge, how long does it last, and what strength does it have in comparison to other changes? The life cycle of a temporal change is indicated by the beginning and end of a curve. The curve's acuteness indicates its strength. A more transparent and exact approach is taken with further data processing on the obtained principal components. For example, a measure of the change strength is obtained by calculating the Euclidean distances of neighboring projected points, but only for the first Z principal components:

$$\boldsymbol{f}_i^{(Z)} = \langle \boldsymbol{F}(i,j) \rangle_{j=1..Z} \qquad (16)$$

$$\boldsymbol{d}^{(Z)}(i) = \| f_i^{(Z)} - f_{i-1}^{(Z)} \| \qquad (17)$$

The selection of the most important PCs makes it possible to clearly distinguish between significant changes and lesser fluctuations. The same method could be used on the original term-frequency matrix, but in that case, unfiltered noise would prevent a meaningful result. An example of Euclidean distances $\boldsymbol{d}^{(Z)}$ plot is given for $Z \in \{1,2,3\}$ in Fig. 2.

- Does seasonality exist in the given collection? The property of seasonality can be addressed with various methods from the domain of time-series analysis, but within the CatViz method, it emerges by itself as a useful side effect of the method. If significant seasonality exists in the text collection, the plotted trajectory will be circular on at least one pair of the principal components.

4.3. Zoom exploration

One goal of our method is to enable both overviews and detailed views of changes in a collection. A way to do this is by constraining the views by time. In general, a subperiod will be qualitatively different when visualizing it alone in comparison to regarding it on a plot where it is visualized as a part of a larger interval. This is because CA attempts to preserve all variance of time intervals from the subperiod alone, in contrast to a situation where it has to preserve most of the variance of time intervals from the whole period. Data constraining by time is a simple and effective exploration technique. An example of zoom exploration is given in a case study described in subSection 5.2.

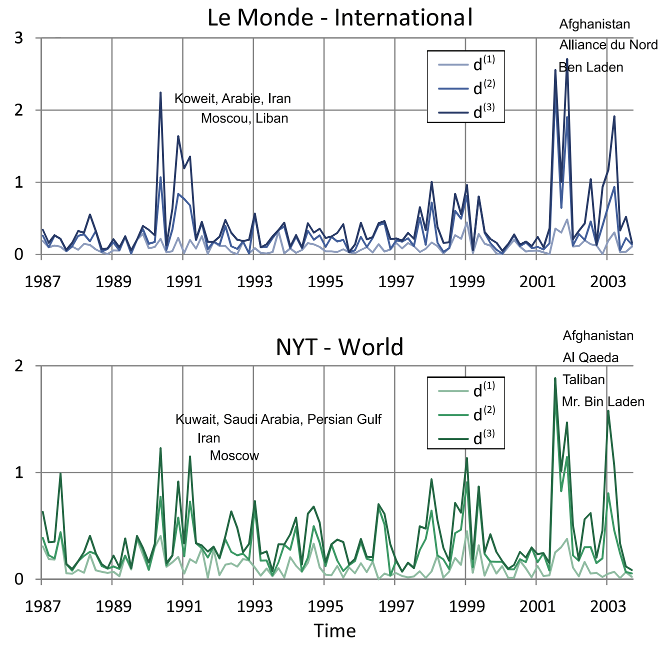


Fig. 2. Euclidean distances of neighboring time interval vectors ($\boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)}, \boldsymbol{d}^{(3)}$), 1987–2003, $m = 2$, f = names; features with most contribution to inertia associated to fluctuations from each month to the sequent month around years 1991 and 2001 are labeled on the plot.

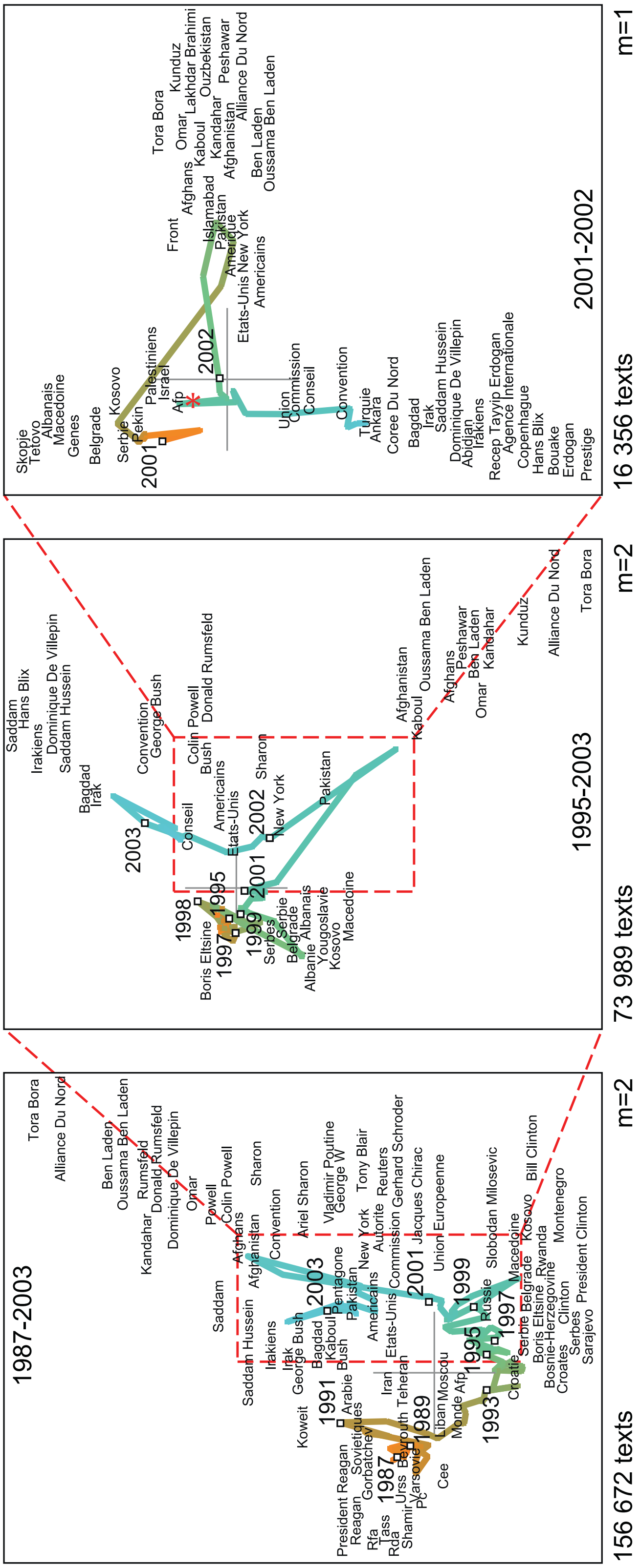


Fig. 3. An example of zoom exploration, Le Monde—International, 1987–2003, f = names, 1st and 2nd PCs.

Table 1
Corpora used.

Corpus	Articles	Time period	Language	Size
Le Monde	833 845	01/1987—12/2003	French	2.5 GB
NYT	1 855 658	01/1987—06/2007	English	13.7 GB

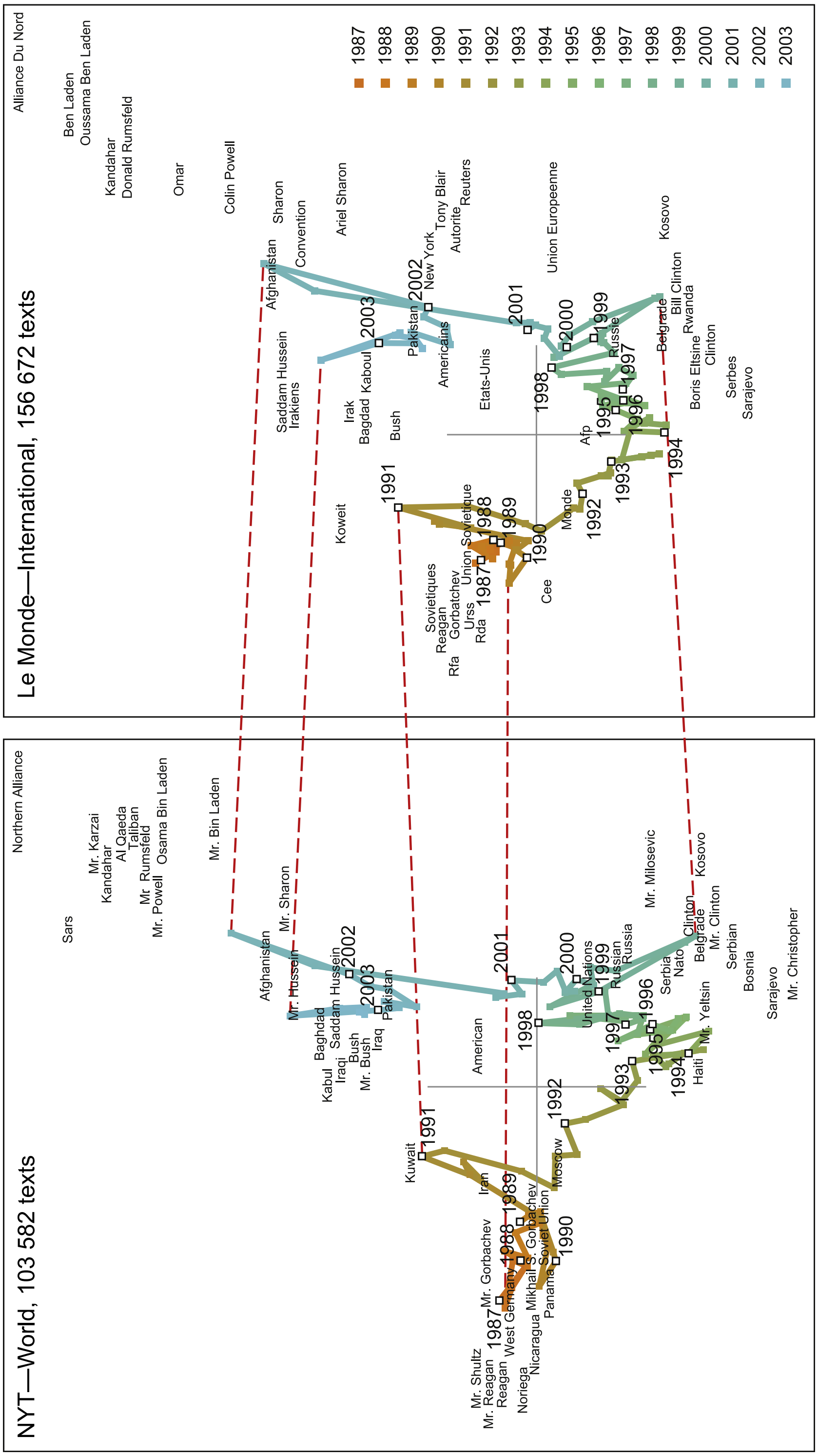


Fig. 4. An example of news comparison, NYT & Le Monde, 1987–2003, $m = 2$, f = names, 1st and 2nd PCs.

4.4. Source comparison

Source comparison is another important application of text visualization. Different text sources can be compared to show their common trends and subtle differences. For now, source comparison is made possible by visual inspection and not by automated procedures, which will be a subject of future work. Our idea was to compare two different sources that discuss the same topics. We chose news articles from two different cultural backgrounds, written in two different languages, over a period of 17 years. A very interesting result showed that temporal CA identifies the same key features and plots qualitatively equal trajectories, which confirms not only that the news sources have similar distributions of key themes in time, but also that in such a case CatViz will be able to identify this similarity. Moreover, we claim that CatViz is robust from the viewpoint of different analyzed data.

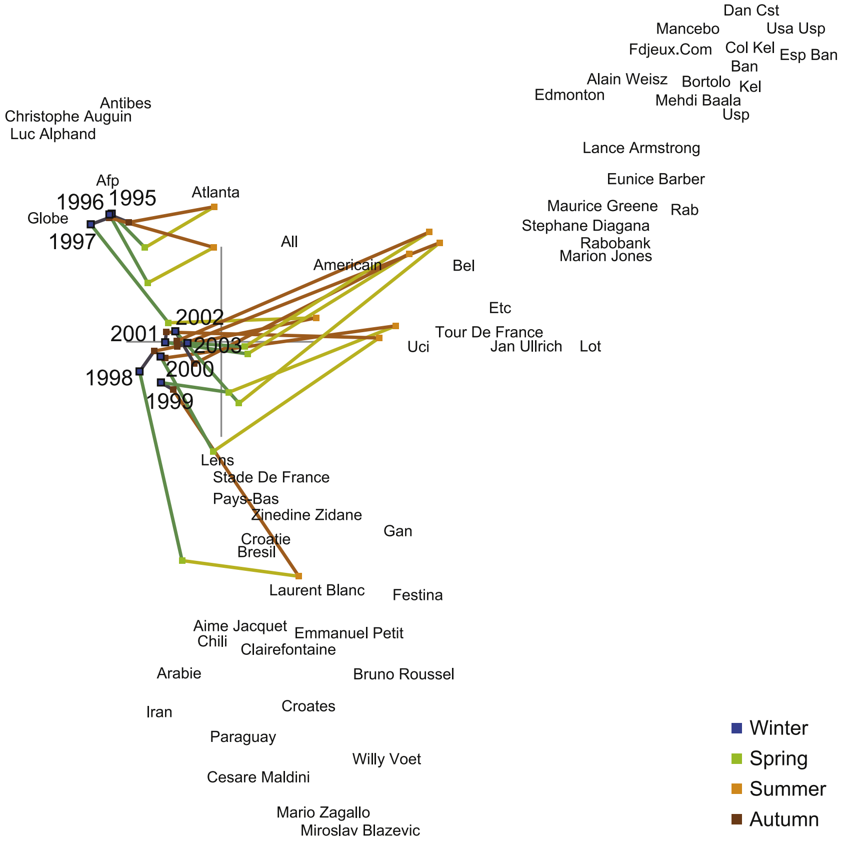


Fig. 5. Le Monde—Sport, 1995–2003, 26 413 texts, $m = 3$, f = names, 1st and 3rd PCs.

5. Case studies

This section presents case studies of exploration by answering typical questions posed during a visual analysis in a real setting. To better depict the statements given in Section 4, we applied application procedures to show all effects of parameter choice and differences among signal change types.

5.1. Data sets and experiment settings

The analyzed data contains newspaper articles of the Le Monde corpus (European Language Resources Association, 2007) and the New York Times corpus (Linguistic Data Consortium, 2008). Basic corpus info is in listed in Table 1, where *Size* denotes the memory needed to store the corpora in their XML formats. Both corpora contain general topic labels such as: *Politics*, *World*, *Sport*, and *Culture*.

In the experiments that were conducted, each plot was constrained to show only one category. Document structure has been omitted, where title and article body were rewritten into a single chunk of text. Words in Le Monde articles were normalized with the French version of Porter's algorithm for stemming (Porter, 2001), whereas the words in the NYT texts were normalized with a lemmatizer for English (Atkinson, 2003). To segment names, two named-entity recognition systems were used: Nemesis (Fourour, 2002) for French and Stanford Named Entity Recognizer (Finkel, Grenager, & Manning, 2005) for English. Names were not morphologically normalized. We used *words*, *names*, and *no-names* as feature choices; see Section 3.1.1.

Pure term frequencies were chosen to represent articles in the vector space. Feature selection was based on counting occurrence, so only terms that appeared in at least 40 different documents were taken into account. The following interval lengths were used: 1 month ($m = 1$), 2 months ($m = 2$), 3 months ($m = 3$), 4 months ($m = 4$), 6 months ($m = 6$), 1 year ($m = 12$), and 2 years ($m = 24$).

CA was calculated in R using the `ca` package (Nenadić & Greenacre, 2007), and the final visualization was drawn with a proprietary online application—the CatViz System. For the purpose of presentation in this work, the labels on plots were manually

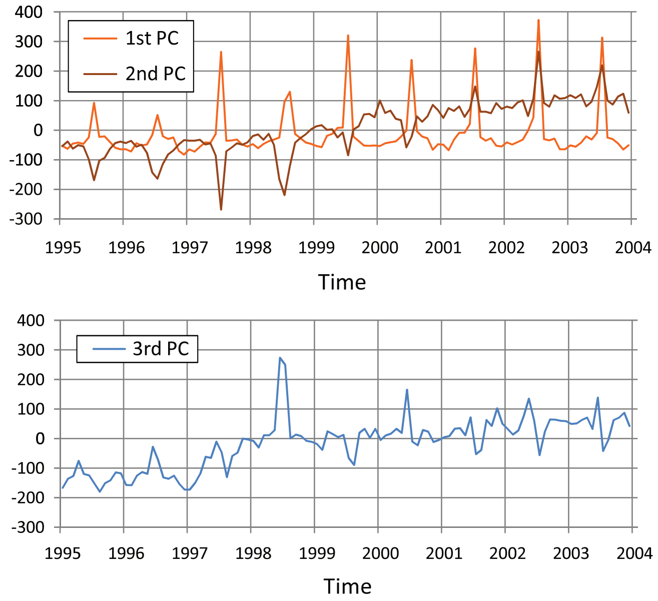


Fig. 6. Le Monde—Sport, 1995–2003, 26 413 texts, $m = 1$, f = names, 1st, 2nd, and 3rd PCs.

adjusted to prevent overlapping. Year labels on all graphs and plots denote the start of an interval.

5.2. Case study: World affairs 1987–2007

In this case, we studied global affairs over the period of 20 years. The data contains articles appearing in the Le Monde under categories Étrangère or International and in the New York Times under the category World. By choosing to visualize *names*, we analyzed trends of countries and leaders on which the news concentrated most. Figs. 1, 3, and 4 show plots for this case study. In the late 1980s, the main topics were concerned with the political crisis in Central America (”Nicaragua", “Mr. Reagan", and “Noriega"). After that, the news covered the fall of the Soviet Union (”Soviet Union", “Mr. Gorbachev", and “Moscow"). In the 1990s, Yugoslavia broke apart, giving rise to conflicts in Croatia, Bosnia, and Kosovo (”Sarajevo", “Bosnia", ”Mr. Clinton", “Mr. Milosevic", “Kosovo", and “Nato"). After the year 2000, articles focused on Middle East conflicts (”Afghanistan", “Iraq", ”Al Qaeda", “Mr. Hussein", “Mr. Bush"). We can see an important characteristic of the CatViz method—on a graph, mutual distances of time intervals indicate their semantic similarity. For example, time-slices around the year 1990 that cover the Gulf War conflict have a semantic bump toward the time slices of 2003 (“Iraq", “Iran", and “Mr. Bush"), indicating that these two periods in history discuss equal regions of conflict. A part of this similarity is accounted for by the presidents having the same family name. The choice of aggregation granularity discussed in subSection 3.2.4 can be seen on Fig. 1. Trajectories for three parameters are plotted: 1 month ($m = 1$), 6 months ($m = 6$), and 2 years ($m = 24$). It can be seen that by changing the granularity, a continuous transition from a detailed trajectory towards a coarse trajectory can be achieved.

As described in Section 4.3, zooming is done to better visualize a specific period. Fig. 3 shows Le Monde articles that are zoomed in from an overview of 17 years into a period of 2 years. On the plot that shows the period from 1995 to 2003, we can see three distinct groups: (1) “Kosovo", “Serbie", “Albanie"; (2) “Afghanistan", “Oussama Ben Laden", “Tora Bora", “Alliance du Nord"; (3) ”Saddam Hussein", “Irak", “Hans Blix". By further zooming in on the period 2001–2002, we can see more detail of news evolution. For example, on the rightmost plot of Fig. 3, the terms “Israel" and “Palestine" were associated with the starting months of 2001, and then reoccurred in April of 2002 when the Israeli–Palestinian conflict intensified. We can see this effect of reoccurrence as a backward bump on the time trajectory; see * on Fig. 3.

Next, an analyst might be interested in comparing text collections. Fig. 4 shows the period from 1987 to 2003—NYT on the left and Le Monde on the right. The trajectories are very much alike, although the NYT's has certain ridges more spread out. First, we imply that the two sources generally discuss equal subjects. It is unlikely that so similar plots could be generated on random if the main subjects of such large corpora were essentially different. The claim that the two newspapers discuss similar subjects with similar distributions is also justified by the fact they are open sources in democratic societies, and on the other hand that they are mainstream media from politically close countries. Second, undertaking the equality of sources, bearing that they are written in different languages over a period of 17 years, and seeing their plots are equal, we imply that our method can robustly show the underlying and most prominent text orientation over time of large collections even if noise is present.

We can see that the news in NYT's category *World* fluctuated more intensively in 1999 during the Kosovo crisis and in 2003 during the Iraq disarmament crisis. That can also be seen in Fig. 2, where Euclidean distances of time-neighboring vectors are plotted. Similarly, in Fig. 2, we can easily see that NYT had greater fluctuations in 1987, when it covered the Iran-Contra affair (“George P. Shultz", “Poindexter", “Oliver L. North", “McFarlane", “Mr. Casey") to a much greater extent than Le Monde.

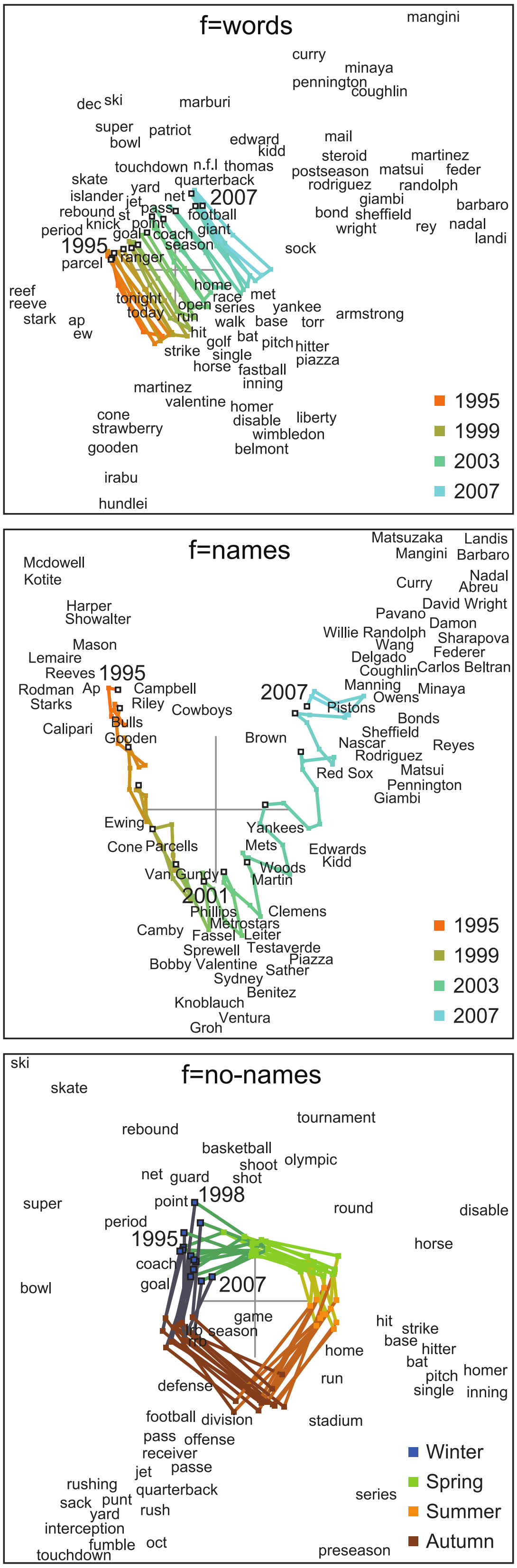


Fig. 7. An example of feature choice, NYT—Sports, 1995–2007, 98 247 texts, $m = 2$, 1st and 2nd PCs.

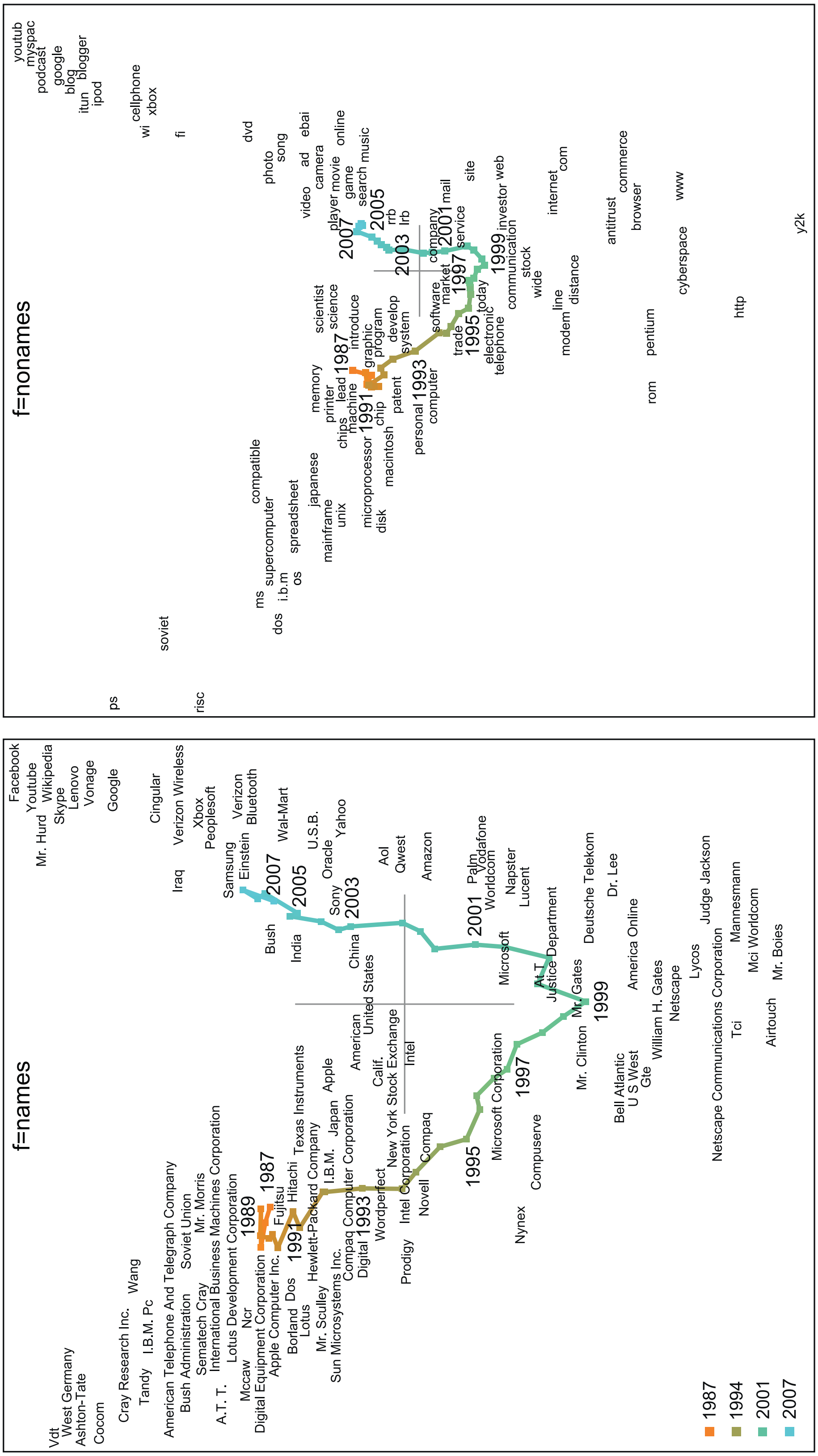


Fig. 8. NYT—Technology, 1987–2007, 51 760 texts, $m = 6$, 1st and 2nd PCs.

5.3. Case study 2: Sport trends 1995–2003

This case study aims to show the evolution of topics in the sports sections of Le Monde and the NYT. Fig. 5 shows the *names* plot from 1995 to 2003 on 1st and 3rd PCs. We can see a circular motion where summer time-slices point toward a distinctive part of the semantic space, where features of athletes performing during the summer are clustered (”Lance Armstrong”, “Eunice Barber”,

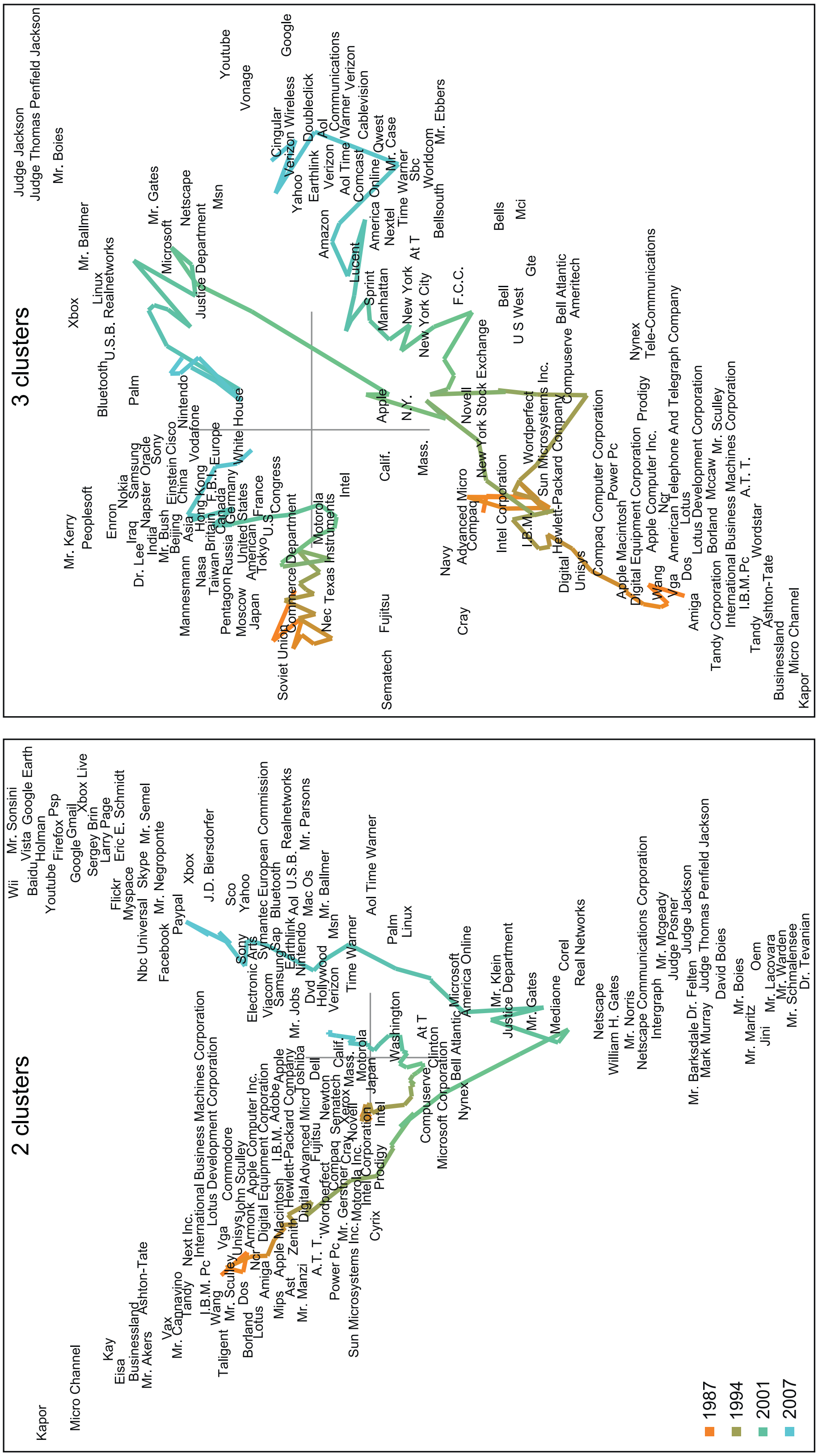


Fig. 9. NYT—Technology, 1987–2007, 51 760 texts, $m = 6$, f = names, 2nd and 3rd PCs.

"Marion Jones", "Mehdi Baala", etc.). In addition, the time-slices of 1998 point toward their own cluster of words, indicating a difference in themes that occurred that year. By inspecting this cluster, we realize that it is composed of features associated with the Football World Cup that occurred in 1998: "Croatie", "Bresil", "Zinedine Zidane", "Miroslav Blazevic", etc.

As an alternative to the scatter plot in Fig. 5, we can examine Fig. 6, where time plots of the principal components are drawn.

Table 2
Knowledge scale.

0	Do not know
1	Have heard something about it, but not sure
2	Can answer questions about who, where, and when
3	— Who, where, when, and what

Table 3
Event entry example.

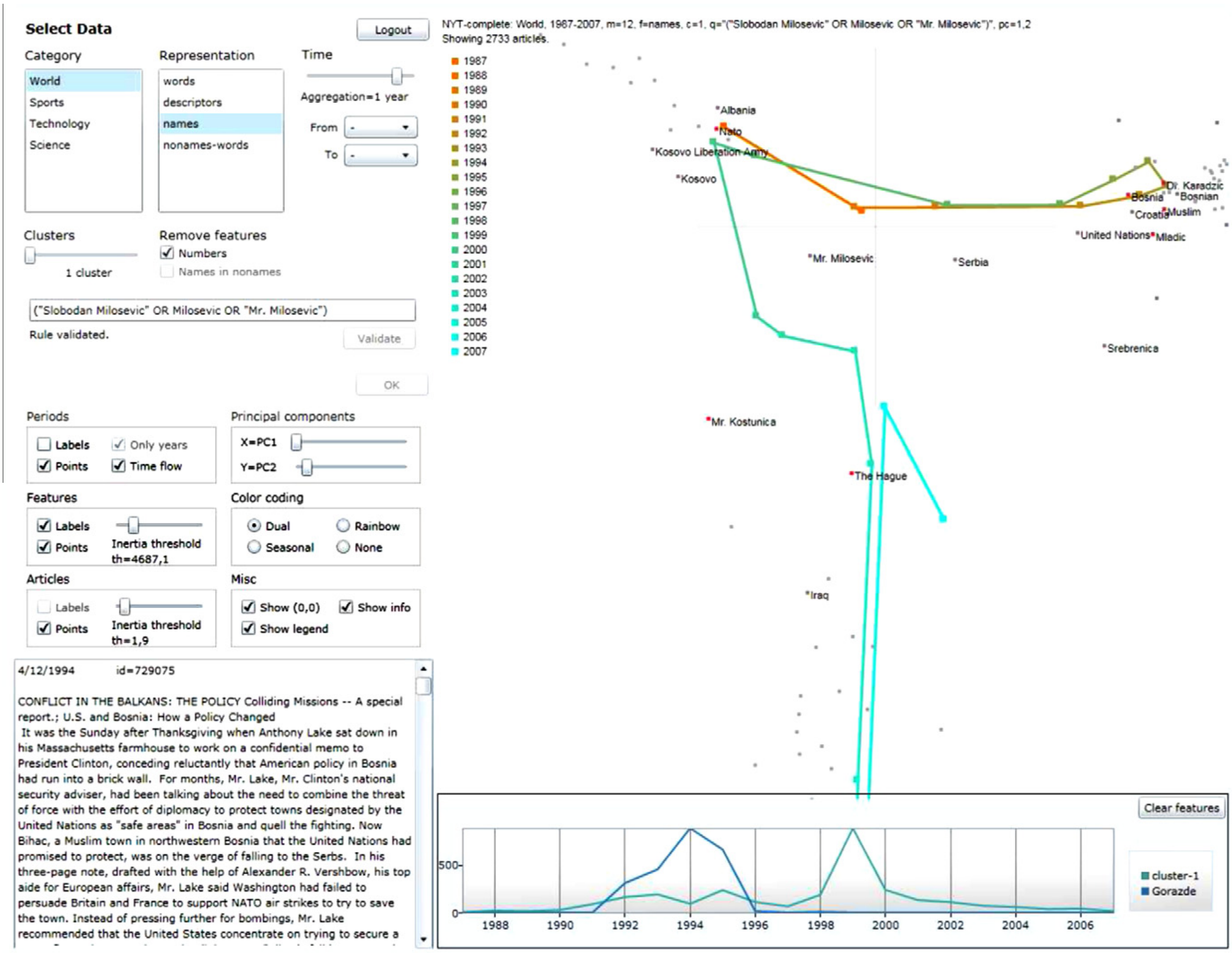
Time	1990
Place	Germany
Persons	Helmut Kohl, Mikhail Gorbachev
What	Unification of FRG and GDR
Knowledge before	2
Knowledge after	3
Article id	368677

We see that the first two components have a strong seasonality with a periodic change lasting one year. This is due to the seasonality of sport events and their news coverage. Apart from seasonality in 1st and 2nd PC, there is a constant drift due to the constant change of the athletes who compete in the sport events. The third principal component lacks seasonality, but among other things, it accounts for a strong semantic bump in 1998 due to the Football World Cup.

In Fig. 7, we can see the effect of feature choice. As already discussed, the names of persons do not alone explain the underlying trends, so other features also need to be used. The simplest choice is to use all words found in texts. When using the *words* features on the NYT Sports data, the first two components show both seasonality and shift. If we split the words in texts into *names* and *no-names*, and use these two sets independently, we obtain the middle and right plots in Fig. 7. We see that using *no-names* features, major sport seasons in the USA are clearly identified: football is major in autumn, winter sports in winter, basketball in spring, and baseball in summer. By plotting only *names*, the first two components mostly explain the shift, although there are traces of seasonality manifested as bumps. The seasonality of *names* can be better seen using combinations of lower rank components.

5.4. Case study 3: Technology 1987–2007

In this case study, we tried to summarize technology trends that were covered by 51 760 articles issued in NYT from 1987 to 2007. Fig. 8 shows the stream with *no-names* and *names* feature choice. In the right subfigure, we read technology trends: end of 1980s:



Select Data

Category: World (selected), Sports, Technology, Science

Representation: words, descriptors, names (selected), nonames-words

Time: Aggregation=1 year; From -; To -

Clusters: 1 cluster

Remove features: ☑ Numbers; □ Names in nonames

("Slobodan Milosevic" OR Milosevic OR "Mr. Milosevic")

Rule validated.

Periods: □ Labels ☑ Only years ☑ Points ☑ Time flow

Principal components: X=PC1 Y=PC2

Features: ☑ Labels ☑ Points Inertia threshold th=4687,1

Color coding: ● Dual ○ Rainbow ○ Seasonal ○ None

Articles: □ Labels ☑ Points Inertia threshold th=1,9

Misc: ☑ Show (0,0) ☑ Show info ☑ Show legend

4/12/1994 id=729075

CONFLICT IN THE BALKANS: THE POLICY Colliding Missions -- A special report.; U.S. and Bosnia: How a Policy Changed
It was the Sunday after Thanksgiving when Anthony Lake sat down in his Massachusetts farmhouse to work on a confidential memo to President Clinton, conceding reluctantly that American policy in Bosnia had run into a brick wall. For months, Mr. Lake, Mr. Clinton's national security adviser, had been talking about the need to combine the threat of force with the effort of diplomacy to protect towns designated by the United Nations as "safe areas" in Bosnia and quell the fighting. Now Bihac, a Muslim town in northwestern Bosnia that the United Nations had promised to protect, was on the verge of falling to the Serbs. In his three-page note, drafted with the help of Alexander R. Vershbow, his top aide for European affairs, Mr. Lake said Washington had failed to persuade Britain and France to support NATO air strikes to try to save the town. Instead of pressing further for bombings, Mr. Lake recommended that the United States concentrate on trying to secure a

NYT-complete: World, 1987-2007, m=12, f=names, c=1, q="("Slobodan Milosevic" OR Milosevic OR "Mr. Milosevic")", pc=1,2
Showing 2733 articles.

Fig. 10. Screenshot of the CatViz System – An example of analyzing all articles mentioning Slobodan Milosevic – his rise to power occurred with the ethnic clashes in Kosovo at the end of the 1980s; afterwards, the news concentrated on the conflicts in Croatia and Bosnia during the 1990s; the trajectory swings back to Kosovo showing that the end of the 1990s was marked by the NATO intervention; an extradition to the ICTY court in Hague occurred in 2001.

"supercomputer", "spreadsheet", "mainframe", "unix", "disk", "dos"; first half of 1990s: "personal", "computer", "software", "program"; second half of 1990s, beginning of 2000s: "communication", "modem", "pentium", "internet", "www"; 2005–2007: "youtube", "google", "ipod", "cellphone", "dvd", "blog".

A similar analysis of companies and leaders can be done by inspecting names. In the left subfigure of Fig. 8, we can see that the feature "Intel" is nearly in the middle of the plot, being relatively close to all time-slice points. This means that "Intel's" flux over time is relatively constant compared to the flux of "Google", which is more time-specific. We can choose to cluster time-sliced articles to better identify these kinds of features. On the left side of Fig. 9, the collection is clustered into two subcollections depicted with trajectories. The inner one discusses more constant companies whose flux is more stable, such as "Apple", "Hewlett–Packard", "Motorola", etc. The outer subcollection consists of articles discussing companies whose coverage is more period-specific—for example, "Microsoft" is closer to the outer subcollection because a large amount of articles discussed it in the second half of the 1990s during the antitrust trials. By choosing to cluster the collection in three subcollections, we obtain one whole subcollection dedicated to various trials ("Judge Jackson", "Mr. Boies", "Realnetworks", "Netscape", "Microsoft", etc.), which can be seen on the right side of Fig. 9.

6. User study

According to Carpendale (2008), intrinsic evaluations of information visualizations can be structured in the same way that evaluations are structured in social sciences. In that light, we classify this user study as a laboratory experiment with a quantitative methodology.

There were 11 respondents, of whom 3 were female and 8 were male. The respondents were aged from 23 to 55, and were not experts in the field of contemporary history or media analysis. The evaluation was conducted in three steps. First, the users were introduced to the CatViz System and their tasks in a 20-min presentation. They then used the visualization to solve four tasks. They were instructed not to use any external source of information. Complementary to the interactive plot, the users were given access to the original articles that correspond to their current plot. Finally, we collected and analyzed their answers.

We observed whether the users could quickly find and analyze important events from the NYT corpus in texts that were published in the section "World". The numbers of available texts were large, which can be seen in Table 4.

6.1. User tasks

The users solved their tasks by filling in a table of events. For each event entry, they were asked to subjectively assess their own knowledge of it—before and after using the visualization. The knowledge scale used is given in Table 2. The tasks were defined by the following text.

Task 1 According to the volume and importance of texts, list the five most important war zones or armed conflicts in the period from 1990 to 1995. If possible, for every event, write the involved countries, persons, approximate time interval, and type of event.

Task 2 List eight to ten of the most important events in the period from 1990 to 2005. In addition, list one to two articles that describe the event in good detail or are considered the climax of the given event and that you think are a good introduction for reading about that event. A real example of an event entry submitted for this task is given in Table 3.

Tasks 3 and 4 List the five most significant events in 2006 and 1999, respectively.

6.2. CatViz System

The users used the CatViz System—an online implementation of the CatViz method that allowed them to choose all parameters already described in the previous sections: time period, aggregation, text representation, number of clusters, and the combination of principal components. Moreover, the users could select only texts that satisfied a certain Boolean query. This allowed them to focus on specific subjects. Another important aspect of the CatViz System is that it supports interactivity through zooming and filtering of features and articles according to the measure of inertia. This proved very useful, as it enabled users to find relevant articles for certain subjects at a given time. On the plot, each article was represented as a dot, and by clicking on it, the text of the article was shown to the user. The users were also granted the option of plotting a temporal distribution of each feature's frequency. A screenshot of the system is presented in Fig. 10.

6.3. Results

First, we calculated the users' agreement on the most important events simply by calculating the overlap of event lists. Two events were considered equal if their place agreed and if their time interval overlapped. The pairing was also manually checked to make sure that entries were precise and that all equalities were correct. The results are given in Table 4. On average, the users agreed in 50% of cases on the most important events for given time periods. This was expected, and it shows that the users' impression of event importance differs. The reason for this is rooted in their foregoing bias, different interaction with the parameters, and varying perception of plots. The noted 50% agreement clearly shows the explorative property of visualization as a knowledge discovery tool. A higher agreement on the first two tasks might be in that a longer period has more extreme frequency distribution of the most frequent events. This makes these top events more prominent and influences users to agree slightly higher.

Next, we compared the users' answers with external sources that we regard as relevant. For Task 1, we considered the Armed Conflict Events Database (ACED).[1] For tasks 2, 3, and 4 we took English Wikipedia's lists of important events during the 1990s and 2000s.[2] We calculated the macro precision of user event entries with respect to the mentioned external sources. More precisely, the macro precision was calculated as follows: $P = n_M/n_T = n_M/(n_M + n_O + n_N + n_I)$, where n_M is the number of users' event entries appearing in the main lists of the external source, n_T is the number of all entries, n_O is the number of entries in other pages of the given source, n_N is the number of entries not existing in the given source, and n_I is the number of incorrect or imprecise entries. The precision ranges from a good 74.0% to a very high 97.6%, which can be seen in Table 5. These excellent results tell us that if a user tries to find important events from a news source using the CatViz method, it is highly likely that the events found will really be important. The assumption being made here is that if an important event was described in the NYT, then it will be included in Wikipedia. This bias can be tolerated if we consider that Wikipedia is open for editing and that its editors read the NYT. Recall was not calculated, not only because the numbers of events returned by the users were much smaller than the number of events contained in the main lists, but also because the given sources have an extended notion of event importance based on factors other than term frequency.

Next, based on our scale (Table 2), users' knowledge has increased by 3 grades in 18 cases, by 2 grades in 41 cases, by 1 grade

[1] <http://www.onwar.com/aced/>, accessed 28/02/2011

[2] e.g., <http://en.wikipedia.org/wiki/1990s>, acc. 28/02/2012

Table 4
Annotator agreement.

Task	Period	Articles	Agreement (%)
1	1990–1995	34867	54.0
2	1990–2005	92414	54.7
3	2006	5869	43.3
4	1999	5287	46.0

Table 5
Precision with respect to relevant event lists.

Task	Source	n_M	n_O	n_N	n_I	Precision (%)
1	ACED	45		2	3	90.0
2	Wikipedia	81	1	1		97.6
3	Wikipedia	37	12	1		74.0
4	Wikipedia	43	4	2	1	86.0

in 95 cases, and has not increased in 79 cases. This indicates two things. First, that the users were not completely acquainted with the subject before the experiment. Second, in 66% of cases, they gained knowledge. We believe that the procedure to plot and filter single articles helped the users in finding important texts and gaining knowledge from large amounts of texts.

7. Conclusion

In this work, we have presented CatViz—a novel visualization method based on Correspondence Analysis that is oriented towards exploration of temporal data. CatViz is efficient and intuitive, making it suitable for the exploration of large text collections. The method is formally presented, and the use of parameters during an analysis is explained. An extension of the CatViz with clustering is presented. The CatViz method is compared with other methods, and differences from each existing research are noted. We have set out case studies and a user study on large corpora that confirm the value of CatViz. We have shown that feature filtering based on automatic named-entity recognition is useful for information-visualization tasks. In addition, we have shown that overviews of two text collections written in different languages can be easily compared using CatViz. Instructions for analysts are given in order for them to easily interpret plots and answer typical analysis questions. A brief connection to the time-series analysis field is given.

We believe that CatViz will enable more efficient, unbiased access to knowledge hidden in huge data sets by a wide spectrum of users. The importance of CatViz is that it enables both high-level overviews and detailed inspection, giving the users a capability of exploring millions of texts at a time and bringing us closer to the objectification of history and contemporary affairs. Moreover, we expect that CatViz will be used on other temporal multivariate data.

The future work will include some statistical tests on principal components obtained with CatViz. In addition, multi-table analysis will be researched to enable the comparison of many tables at once.

Acknowledgments

This research has been jointly supported by the Croatian Ministry of Science, Education and Sports under the Grant No.036-1300646-1986. The authors would like to thank Stéphane Huet for preparing the Le Monde corpus.

References

Albrecht-Buehler, C., Watson, B., & Shamma, D. A. (2005). Visualizing live text streams using motion and temporal pooling. *IEEE Computer Graphics and Application, 25*(3), 52–59.

Alencar, A. B., de Oliveira, M. C. F., Paulovich, F. V., Minghim, R., & Andrade, M. G. (2007). Temporal-PEx: Similarity-based visualization of time series. In *Proceedings 20th Brazilian symposium computer graphics and image processing (SIBGRAPI)*.

Alsakran, J., Chen, Y., Zhao, Y., Yang, J., & Luo, D. (2011). Streamit: Dynamic visualization and interactive exploration of text streams. In *Pacific visualization symposium (PacificVis) 2011 IEEE* (pp. 131–138). IEEE.

Atkinson, K. (2003). Automatically generated inflection database (AGID). <http://wordlist.sourceforge.net/>, accessed June 2010.

Bansard, J.-Y., Kerbaol, M., & Coatrieux, J.-L. (2006). An analysis of IEEE publications in biomedical engineering. *IEEE Engineering in Medicine and Biology Magazine, 25*(5), 10–12.

Benzécri, J.-P. (1973). *L'analyse des correspondances*. Paris: Dunod.

Carpendale, S. (2008). Evaluating information visualizations. In *Information visualization: Human-centered issues and perspectives* (pp. 19–45). Springer-Verlag.

Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers, C. E., & Wylie, B. N. (1998). Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems, 11*(3), 259–285.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of American Society for Information Science*, 41.

European Language Resources Association. (2007). Text corpus of "Le Monde. <http://catalog.elra.info/product_info.php?products_id=438>.

Falk, M. (2006). A first course on time series analysis—examples with sas. <http://statistik.mathematik.uni-wuerzburg.de/timeseries/>.

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings 43rd annual meeting of the association for computational linguistics. The association for computer linguistics* (pp. 363–370).

Fortuna, B., Mladeni, D., & Grobelnik, M. (2005). Visualization of text document corpus. *Informatica (Slovenia), 29*(4), 497–504.

Fortuna, B., Mladeni, D., & Grobelnik, M. (2009). Visualization of temporal semantic spaces. In *Semantic knowledge management*. Springer.

Fourour, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entité nommées pour le francais. In *Actes de la* 9$^{\text{ème}}$ *Conférence Nationale sur le Traitement Automatique des Langues Naturelles* (Vol. 1, pp. 265–274). TALN, Nancy.

Ghoniem, M., Luo, D., Yang, J., & Ribarsky, W. (2007). Newslab: Exploratory broadcast news video analysis. In *IEEE symposium on visual analytics science and technology* (pp. 123–130).

Greenacre, M. J. (2007). *Correspondence analysis in practice* (2nd ed.). CRC, London: Chapman and Hall.

Hartigan, J. A. (1975). *Clustering algorithms*. Wiley.

Havre, S., Hetzler, E. G., & Nowell, L. T. (2000). Themeriver: Visualizing theme changes over time. In *Proceedings of the IEEE conference on information visualization* (pp. 115–124).

Kaban, A., & Girolami, M. (2002). A dynamic probabilistic model to visualise topic evolution in text streams. *Journal of Intelligent Information Systems, 18*(2–3), 107–125.

Kerbaol, M., Bansard, J.-Y., & Coatrieux, J.-L. (2006). An analysis of IEEE publications. *IEEE Engineering in Medicine and Biology Magazine, 25*(2), 6–9.

Leskovec, J., Backstrom, L., & Kleinberg, J. M. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 497–506).

Linguistic Data Consortium. (2008). The New York Times Annotated Corpus. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19>.

Liu, Y.-B., Cai, J.-R., Yin, J., & Fu, A. W.-C. (2008). Clustering text data streams. *Journal of Computer Science and Technology, 23*(1), 112–128.

Luo, D., Yang, J., Krstajic, M., Ribarsky, W., & Keim, D. (2010). Eventriver: Visually exploring text collections with temporal references. *Visualization and Computer Graphics IEEE Transactions on PP* (99).

Mao, Y., Dillon, J., & Lebanon, G. (2007). Sequential document visualization. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1208–1215.

Martin, D. I., Martin, J. C., Berry, M. W., & Browne, M. (2007). Out-of-core SVD performance for document indexing. *Applied Numerical Mathematics: Transactions of IMACS, 57*(11–12), 1230–1239.

Moerchen, F., Brinker, K., & Neubauer, C. (2007). Any-time clustering of high frequency news streams. In *Data mining case studies workshop, 13th ACM SIGKDD international conference on knowledge discovery and data mining*.

Morin, A., Kerbaol, M., & Bansard, J.-Y. (2000). Etude des résumés en francais des rapports de recherche d'un institut d'informatique publiés de 1989 à 1998. In *Proceedings of Journées Internationales d'Analyse Statistique des Données Textuelles – JADT*.

Nenadić, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The `ca` package. *Journal of Statistical Software, 20*(3), 1–13.

Petrović, S., Dalbelo Bašić, B., Morin, A., Zupan, B., & Chauchat, J.-H. (2009). Textual features for corpus visualization using correspondence analysis. *Intelligent Data Analysis, 13*(5), 795–813.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. <http://snowball.tartarus.org/texts/introduction.html>.

Roweis, S. T. (1999). Constrained hidden Markov models. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *NIPS* (pp. 782–788). The MIT Press.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw Hill.

Šilić, A., Dalbelo Bašić, B. (2010). Visualization of text streams: A survey. In *Proceedings of the 14th international conference on knowledge-based and intelligent information and engineering systems*.

Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). *Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In Proceedings of the Information Visualization, 1995*. IEEE, pp. 51–58.

Wong, P.C., Foote, H., Adams, D., Cowley, W., & Thomas, J. (2003). Dynamic visualization of transient data streams. In *Proc. IEEE Symposium on Information Visualization*.