Visualization of Text Streams: A Survey

Artur Šilić
Department of Electronics, Microelectronics, Computer and Intelligent Systems
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia

***Abstract*— This work presents related areas of research, types of data collections that are visualized, technical aspects of generating visualizations, and evaluation methodologies. Existing methods are structured and explained from the aspect of visualization process. Successful applications are noted and some future trends in the field are anticipated.**

***Keywords*— Information Visualization, Visual Analytics, Topic Detection and Tracking, Text Mining, Trend Discovery, Visualization Evaluation, Dimensionality Reduction, Text Representation, Information Extraction, User Interaction.**

I. Introduction

From the earliest days of human civilization, visualizations have been an important way to record, communicate or analyze real and abstract entities. Over the past two decades, digitalization of textual data has greatly increased the accessibility of text documents in all areas of human society. This has introduced a strong need for efficient storage, processing, retrieval and analysis of texts, since the number of available documents has become quite large. Visual exploration is an efficient and natural way to analyze large text collections. This work presents a survey of methods to visualize text streams.

The article is structured as follows. First, research areas of wider scope and research areas that interact with text visualization are discussed in Section 2. Next, data collection types are structured in Section 3 and visualization process is explained in Section 4. Evaluation of visualizations is presented in Section 5 and successful applications are noted in Section 6. Section 7 concludes the article.

II. Related research areas

S. Owen defines visualization as "a tool or method for interpreting image data fed into a computer and for generating images from complex multi-dimensional data sets" [1]. Information visualization is a subfield that studies how to visually represent large collections of non-numeric information [2]. Texts carry information encoded in natural language, hence text visualization is a subfield of information visualization.

Visual analytics is a closely related area of research which has been described as "the science of analytical reasoning facilitated by interactive visual interfaces" [3]. Information visualization enables *visual analysis* which is a subsequent step in knowledge discovery [4]. The importance of moving from confirmatory to explorative data analysis is emphasized in [5], [6]. Integrating automatic analysis and exploratory analysis with the aid of visualization techniques will enable more efficient and effective knowledge discovery [5].

Another field of study that relates to text visualization is Topic Detection and Tracking (TDT) [7], a DARPA-sponsored initiative to investigate the state-of-the-art in finding and following new events in a stream of broadcast news stories. Pilot phase started in 1998 and final phase of the project ended in 2004. The developed methods for detection and tracking of stories are elaborate and data-specific. They can be used as a background tool for text stream visualization in which case the foreground drawing methods need not be complex while the visualizations can still be highly informative. An example of such method is EventRiver [8].

Analogous to text and data mining methods, text visualization heavily depends on algebraic algorithms and statistical methods. Dimensionality reduction and feature selection is employed on matrices that represent text collections. When representing texts, all sorts of information extraction and processing methods can be employed. These are mostly developed within the computational linguistics community.

In practice, designers of visualization systems have to be aware how their visualizations are used and question if user's needs are satisfied while performing exploratory or confirmatory analysis. Due to human factor involved in evaluation of text visualization methods, methodologies from social sciences are used. Information visualization has common tasks of interest with empirical research in human-computer interaction, perceptual psychology, and cognitive reasoning. Although some mutual research has been done, more integration of these communities is encouraged [9]. At last, visualization researchers and developers can refer to design principles to improve readability, interaction and general usability. For such advice, see E. Tufte's book [10].

III. Data Types

Visualization methods operate on three types of text data:

1) Collection of texts (col)
2) Single text (single)
3) Short interval of a text stream (flow)

The latter is used to visualize trends in texts arriving in real time [11], [12], [13]. Such visualizations can be dynamic meaning that they include animations [11], [13].

Essentially, text streams are characterized as collections of texts that arrive through time. Let us comprehend the following: A text collection is composed of documents which are composed of paragraphs. Finally, paragraphs are broken down to sentences which are the basic carriers of messages in a language. If we observe the paragraph sequence of a single

text as a sequence of independent texts, we will come to the conclusion that the former established partition of methods is superficial and that any text stream method can be used to visualize a single text (and vice versa).

However, the partition based on collection type is useful due to the following reasons. First, the statistical properties of text parts on single text level and on collection level differ significantly. For example, probabilities of word occurrence in a single text are noisier so smoothing has to be used, see [14]. Second, computational demands differ since a single text can be represented within a matter of kilobytes in comparison to large corpora like the *New York Times* [15] corpus that contains more than 1.8 million news articles from a period of over 20 years and has size measured in gigabytes. Even larger corpora are processed—in [16], a collection of over 90 million texts is reported.

However, it is useful to present methods at all levels of text streams since basic concepts are synonymous—to discover and analyze trends in texts. The relativity in comprehension of what is a text collection has already been noted in [17], [18].

Recently, some works were published on visualization of search results, human dialogue, and software code. Although these are similar data types, such works will not be discussed. Also, there are some works that relate social network media analysis and text mining. Discussion about these will be left for future work.

IV. Visualization process

In general, the visualization is composed of three steps. First, the texts processed in a representation more suitable for sequent operations. Second, in order to draw a view, a mapping onto a 2D or 3D space is performed. Documents, groups of documents, terms, events, relationships, summaries or other entities are drawn and labeled on the view. Third, user interaction is enabled. This process is refined in more detail in [4].

This section discusses a number of existing methods from the aspect of visualization process. A brief survey with an emphasis on historical aspect is available in [4]. A brief, but well-written review of methods is given as a related work in [13]. Text visualizations methods in relation to TDT and temporal text mining are structured in [19]. An extensive survey of trend detection which includes visualization methods and commercial products is described in [20]. Also, a comprehensive survey oriented on commercial patent analysis tools is available in [21].

A. Text representation

Unstructured text is not suitable for visualization, so a text is usually represented in *vector space model* (VSM) [22]. This concept is very popular in information retrieval. *Bag-of-words*, an instance of VSM, is the most widely used model for text representation. The method consists of counting word occurrences in text. This produces a vector in the space of features, which correspond to the words found in the text. Doing so, word order is abandoned. Often, the vectors are weighted using *Term Frequency Inverse Document Frequency* method (TFIDF) [23].

Alternatives to VSM exist—a text can be represented using a language model where a text is treated as a set of probabilities, for example refer to [24].
Another example of novel approach to text representation is employed within the MemeTracker framework [16] in which a collection of documents is represented as a *phrase graph*. For this method VSM is not suitable since information on word order is needed for extraction of phrases from text.

Other information apart from bare words can be extracted from text. Some relations can be trivial to extract from texts (e.g. co-occurrence of names in texts) while others are not expected to be feasible to extract in the foreseeable future (e.g. author's stance on some entity not directly written in the text, obtainable only by precise logic reasoning and background knowledge). However, as research in the computational linguistics field progresses, information extraction techniques will be advanced and information visualization will have more data to feed on. The following list gives methods of feature extraction that are currently used or are foreseen to be used in the near future.

1) *Bag-of-words*—this is a baseline used by many methods. It can be improved with stop-word filtering, morphological normalization and word sense disambiguation. Apart from single words, collocations and other significant word n-grams can be extracted.
2) *Entity recognition*—these techniques automatically identify names of people, organizations, places, or countries [25]. Moreover, relationships among entities found in text can be introduced. Events can be extracted and visualized, see [8].
3) *Summarization*—these techniques include keyword extraction, keyword assignment, thematic categorization, and fact extraction. The aim is to shorten texts and present only the most relevant information, hence the visualization is able to filter the clutter even more effectively. An example of extracted facts visualization can be found in a commercial product ClearForest [20].
4) *Document structure parsing*—these relatively simple techniques are used for automatically identifying structural elements such as title, author names, or publication date. Structural information can be used in the visualization process as well.
5) *Sentiment and affect analysis*—these techniques are used to emotionally characterize the content of texts. An example of affect visualization for single documents can be found in [26].

By using these feature extraction methods, relative importance of each feature present in a text is obtained by general or method-specific means. For example, simple occurrence frequency is a common numeric value of a feature in the *bag-of-words* model. In contrast, relations among entities can be the result of a very complex model that involves syntactic parsing of sentences.

B. Drawing a view

Table I summarizes this subsection by listing existing visualization methods and their underlying methods. The table also notes publication year and if the method has an inherent temporal orientation apart from time slicing. It is easily recognized that methods with inherent temporal orientation are researched only in recent years.

1) Term trend approach: The most straightforward way to visualize trends in text streams is to plot frequencies (FP) of important terms found in texts at a given window of time. Feature selection is employed to reduce the number of dimensions and prevent visual occlusion. Feature selection methods range from simple frequency criteria like in [34], to more complex statistical measures of feature importance such as information gain or χ^2 [23] used in [48].

The term trend approach is used in [34], [40], [41]. In MemeTracker [16] trends are plotted by efficiently detecting memetic phrases what can be regarded as an advanced phrase selection. EventRiver plots frequencies of documents relating to certain events. Event labeling can be regarded as a sophisticated term selection.

2) Semantic space approach: In semantic space approach each part of the view corresponds to some semantic category found in the text collection. This approach is used by [27], [28], [29], [17], [18], [31], [32], [33], [35], [36], [37], [38], [14], [11], [13], [30], [39], [42], [43]. Often, the vectors representing texts are of high dimensions because textual features are numerous, so dimensionality reduction techniques are employed in order to map these vectors to 2D or 3D space.

Used dimensionality reduction techniques include Latent Semantic Indexing (LSI) [49], Principle Component Analysis (PCA) [50], and Correspondence Analysis (CA) [51] which are all methods based on Singular Value Decomposition (SVD). A very good blog entry on this subject is encouraged for reading [52]. Before performing these linear algebra operations, a distance function between vectors has to be defined—often the Euclidean distance is chosen among others.

PCA method yields the best low-level approximation of the original term-document matrix with regard to variance and classification. It tries to preserve as much of the space's variance in the remaining dimensions. LSI differs from PCA in the fact that it applies SVD to the covariance matrix instead to the original matrix. It is noted that LSI is can automatically solve problems of synonymy and polysemy in a language. Using CA, an importance of rows and columns is obtained by regarding inertias. Also, matrix rows and columns and can be represented both in the same lower dimensional space enabling a consistent and symmetric analysis of relations among terms and documents.

Multidimensional scaling (MDS) [53] is another popular method to reduce dimensions. MDS attempts to find an embedding from a set of objects into R^n such that the defined distances among objects are preserved as much as possible. Only distances among objects need to be given and whole problem is solved as an optimization task. The obtained embedding is non-linear and new dimensions lack straightforward interpretation. Sammon's mapping [27] is a version of MDS.

Since MDS-based and SVD-based methods used on large text collections can be too computationally demanding, some tricks are employed. Usually, the transform to the low dimensional space is calculated using a smaller set of vectors which can be a subset of the original set or centroids produced by clustering of the original set. These methodologies are similar to concept indexing [54] or, more generally, to coresets [55]. Examples of such methods are Anchored Least Stress (ALS) [56], Boeing Text Representation Using Subspace Transformation (TRUST) [33], Least Square Projection (LSP) [57], and Projection by Clustering (PROJCLUS) [39]. ALS combines PCA and MDS and makes use of the result of data clustering in the high dimensional space so that it can handle very large data sets.

Kohonen's Self Organizing Maps (SOM) [58] is a neuro-computational algorithm to map high-dimensional data to a lower dimensional space through a competitive and unsupervised learning process.

Force-Directed Placement (FDP) [59] is an algorithm that has an analogy in a physical system of masses and springs. Masses have inertia, so their velocity only changes after a force has exerted the mass for some time. The forces are a result of springs between masses being stretched or compressed. By simulation, the algorithm finds a stable point. FDP algorithms can be problematic because of their complexity of $O(n^3)$. IDMAP [60] is an adapted FDP-based method used to improve computational complexity. Other general or custom graph-drawing (GD) methods are also used, for examples see [44], [45], [19].

The TreeMap [61] visualization technique used in [12] makes use of all available display space, mapping the full hierarchy onto a rectangular region in a space-filling manner. The hierarchy among texts can be defined in various ways starting from simple hierarchical clustering. A similar placement method called *rectangle packing* is used in [47].

Other specific methods that are used to draw visualizations include: Wavelet filters used in [18], *Hidden Markov Models* (HMM) used in [35], *Associative Relation Network* (ARN) used in [30], *Voronoi Tessellations* used in [36], and *Random Projections* used in [32].

Different clustering methods are used in many of the described methods as an intermediate step with the aim to convert a large set of texts into a smaller set of elements, to find groups of texts with specific properties, or to summarize a number of texts.

For now, most of methods that use semantic space approach enable trend discovery in text streams by time slicing. The time slicing method constrains a series of views to a series of time intervals. By analyzing differences in the views, the user gains insight into changes in text stream. Time slicing has explicitly been noted for [17], [31], [33], [13], [12] although it can be performed with any method since it consists only of constraining the text set to be visualized. Time slicing has been criticized for being limited in discovery since human memory is limited and change blindness is present [8].

An interesting approach to time series visualization is presented in [43]. Entities are visualized using points that have an associated time signal. The similarity between time

TABLE I
LIST OF VISUALIZATION METHODS

Method name	Basic underlying methods	Data type	Temporal	Year	Reference
Sammon	Sammon's mapping	col	-	1969	[27]
Lin et al.	SOM	col	-	1991	[28]
BEAD	FDP	col	-	1992	[29]
Galaxy of News	ARN	col	-	1994	[30]
SPIRE / IN-SPIRE	MDS, ALS, PCA, Clustering	col/single	-	1995	[17]
TOPIC ISLANDS	MDS, Wavelets	single	N\A	1998	[18]
VxInsight	FDP, Laplacian eigenvectors	col	-	1998	[31]
WEBSOM	SOM, Random Projections	col	-	1998	[32]
Starlight	TRUST	col	-	1999	[33]
ThemeRiver	FP	col	+	2000	[34]
Kaban and Girolami	HMM	col	+	2002	[35]
InfoSky	FDP, Voronoi Tessellations	col	-	2002	[36]
Wong et al.	MDS, Wavelets	col	-	2003	[37]
NewsMap	Treemapping	flow	˜	2004	[12]
TextPool	FDP	flow	˜	2004	[13]
Document Atlas	LSI, MDS	col	-	2005	[38]
Text Map Explorer	PROJCLUS	col	-	2006	[39]
FeatureLens	FP	col	+	2007	[40]
NewsRiver, LensRiver	FP	col	+	2007	[41]
Projection Explorer (PEx)	PROJCLUS, IDMAP, LSP, PCA, Sammon's m.	col	-	2007	[42]
SDV	PCA	single	N\A	2007	[14]
Temporal-PEx	IDMAP, LSP, DTW, CDM	col	+	2007	[43]
T-Scroll	GD, Special clustering	col	+	2007	[44]
Benson et al.	Agent-based clustering	flow	˜	2008	[11]
FACT-Graph	GD	col	+	2008	[45]
Petrović et al.	CA	col	-	2009	[46]
Document Cards	Rectangle packing	single	N\A	2009	[47]
EventRiver	Clustering, 1D MDS	col	+	2009	[8]
MemeTracker	FP, Phrase clustering	col	+	2009	[16]
STORIES	GD, Term co-occurence statistics	col	+	2009	[19]

signals is introduced by the following measures: Euclidean, *Dynamic Time Warping* (DTW) [62], and *Compression-based Dissimilarity Measure* (CDM) [63]. The DTW distance is used for time series with different sizes or distortions in the time axis. The CDM distance is used to detect structural differences and only works in long time series. Having defined a distance between time series, LSP or IDMAP enable to find projections of points to a low dimensional space.

C. User interaction

User interaction is concerned with the way user generates the view and how he advances to a next one while gaining insights into the analyzed data. Usually, approaches to interaction are the following [64]:

1) *Brushing and linking*—a change in perspective on one view of the data set affects other views as well. For example, refer to [65].
2) *Panning and zooming*—concerns with physically constraining the view in order to enable the user to analyze more general or more specific part of the visualized space. This is the most prevalent method of user interaction, by default enabled in most of visualization methods.
3) *Focus-plus-context*—enables to enlarge a specific part of the view while simultaneously shrinking the context, making distant objects smaller. One example of this approach is the *fisheye view*. In contrast to zooming, focus-plus-context keeps the context visible.
4) *Magic lenses*—filters that can transform the view of a visualized object in order to emphasize some other characteristic. For example, in the Galaxies view [17], the documents can be viewed through magic lens in order to discover their temporal dimension (all documents from a certain time interval change their view).

When designing a visualization system, it is common to combine the afore-mentioned techniques with an option to select data based on some non-visual parameters. The user should be able to choose a designated time-interval, thematic categories, persons involved, or any other available structured information upon which conditions can be set.

Interaction techniques can be advanced in the sense that animations are enabled. For now, animation has been used in

methods oriented on current stream change. In [11], the user can select a term and the visualization animates to make the selected term a central point in the view. Next, in [13] the visualization dynamically changes in real time as texts arrive.

V. Evaluation

Design of evaluation methods and methodologies for information visualizations is still an ongoing effort of the research community. All difficulties concerning visualization evaluation are related to human factor that is inherently involved in the process of using a visualization. Evaluation is hard since not all variables can be observed and controlled, in contrast to purely computational processes. A suggested approach to evaluation methodology has an inspiration in social sciences [66]. This section is a brief summary of a recent overview on this topic written as a book chapter by S. Carpendale [9].

Evaluation is important since the users and researchers need to compare and validate visualization methods. In the past, usability of a visualization was only emphasized without systematic research since an easily understandable visualization speaks for itself. As this field developed, a plethora of methods has been created so an empirical approach was needed. At the present, the methods are usually evaluated by university students on small data sets and oriented towards simple tasks. Such evaluations are easy to conduct. In contrast to current practice, the evaluations should be done by real users on real data sets and oriented towards real tasks. When choosing such realistic settings, it might be more difficult to obtain large enough sample sizes, to control all variables or to get precise measurements.

Many of the challenges in information visualization are common to empirical research. Related empirical research disciplines include: human-computer interaction, perceptual psychology, and cognitive reasoning. The main challenge is how to address the *insight* that occurs within the user of a visualization. Such categories highly depend on the participants' interest, motivation, prior knowledge, and mental capabilities.

There are three important factors of evaluation methodologies:

1) *Generalizability*—can the results be extended to a wider scope of users or data sets?
2) *Precision*—to which extent can the measurements be trusted?
3) *Realism*—can the result be considered realistic from the aspect of the context it was studied in?

As discussed in [66], evaluation methodologies in social sciences can be structured in the following instances: laboratory experiment, experimental simulation, field experiment, field study, computer simulation, formal theory, sample survey, and judgment study. Existing methodologies enable the actualization of one or at most two of the given factors what suggests that more than one methodology is preferable to be used.

Evaluation methodology can be quantitative or qualitative. Quantitative methodology has evolved through centuries of experimental research. The data used in within the quantitative approach is expressed using numbers. The challenges in front of this methodology are: conclusion validity, type I and II errors, internal validity, construct validity, external validity, and ecological validity.

The data used within the qualitative approach is expressed in natural language. The techniques can be categorized as:

1) Observation techniques—the participants' behavior is observed and described.
2) Interview techniques—the participants give answers to a set of questions.

The challenges for qualitative methods are: sample size, subjectivity, analysis. The analysis of qualitative methods can be quantitative or qualitative.

Evaluation of information visualization is still a work in progress. It is a labor-intensive task since human work is evaluated and little automatization can be employed. A number of researchers point out the problem of evaluation and some of them encourage others for more evaluation methodologies to be presented in the future works.

VI. Applications

Text stream visualization has a wide range of applications: in all situations where texts have a time stamp or where the collections are not static. In the works cited in this report, various successful applications are mentioned. Those include analyzing news texts, scientific publications, institutional documents, company filings, patent claims, private and enterprise correspondence, web content, and historical archives. The intended users of text visualization suites are media analysts, historians and other scientists from all fields, public policy makers, intelligence officers, journalists, business managers, and private users.

In the future we can expect more applications of text visualization on the task of social media analysis. Also, apart from task-specific and data-specific applications, general applications are being developed. An example of such a general suite is Microsoft's Pivot Project [67] which aims to enable seamless information visualization and interaction.

VII. Conclusion

This work presented a general survey of text visualization methods. Related areas of research areas are presented. Notable works are listed and their methods are discussed. Suggested approaches to evaluation are structured and a wide scope of applications and users is noted.

This field of study has a great potential since the volume of digitally available texts has become huge. Advanced applications are expected to arise from using cutting-edge information extraction techniques developed within the computational linguistics community. Technical advances might include novel methods or solving scalability issues. Applicative advances might include integration with collaborative visualization methods. In order to even better establish these methods and their implementations, more efforts need to be invested in research and deployment of evaluation methodologies. Cognitive and psychological aspects need to be included in the research.

BIBLIOGRAPHY

[1] G. Scott Owen, G. Domik, T.-M. Rhyne, K. W. Brodlie, and B. S. Santos, "Definitions and rationale for visualization," http://www.siggraph.org/education/materials/HyperVis/visgoals/visgoal2.htm, accessed in February 2010.

[2] M. Friendly and D. Denis, *Milestones in the history of thematic cartography, statistical graphics, and data visualization*, 2008, vol. 9.

[3] J. J. Thomas and K. A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005.

[4] J. Risch, A. Kao, S. Poteet, and Y. Wu, "Text Visualization for Visual Text Analytics," *Lecture Notes In Computer Science*, vol. 4404, pp. 154–171, 2008.

[5] D. A. Keim, F. Mansmann, and J. Thomas, "Visual Analytics: How Much Visualization and How Much Analytics?" in *ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery - VAKD '09*. ACM Press, 2009.

[6] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley (Reading MA), 1977.

[7] A. J., *Topic Detection and Tracking, Event- based Information Organization*. Kluwer Academic Publishers, 2002.

[8] D. Luo, J. Yang, J. Fan, W. Ribarsky, and H. Luo, "EventRiver: Interactive Visual Exploration of Streaming Text," *Symposium A Quarterly Journal In Modern Foreign Literatures*, vol. 28, no. 3, 2009.

[9] S. Carpendale, "Evaluating information visualizations," pp. 19–45, 2008.

[10] E. R. Tufte, *Visual Explanations*. Graphics Press, 1997.

[11] J. Benson, D. Crist, and P. Lafleur, "Agent-based visualization of streaming text," Raleigh, 2008.

[12] M. Weskamp, http://newsmap.jp/; http://marumushi.com/projects/newsmap, 2004.

[13] C. Albrecht-Buehler, B. Watson, and D. A. Shamma, "Visualizing live text streams using motion and temporal pooling," *IEEE Computer Graphics and Applications*, vol. 25, no. 3, pp. 52–59, May/Jun. 2005.

[14] Y. Mao, J. Dillon, and G. Lebanon, "Sequential document visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1208–1215, 2007.

[15] "The new york times annotated corpus." [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19

[16] J. Leskovec, L. Backstrom, and J. M. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*. ACM, 2009, pp. 497–506.

[17] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," in *Proc. IEEE Symp. Information Visualization, InfoVis*, N. D. Gershon and S. Eick, Eds. IEEE Computer Society, 1995, pp. 51–58.

[18] N. E. Miller, P. C. Wong, M. Brewster, and H. Foote, "TOPIC ISLANDS-a wavelet-based text visualization system," in *Proceedings of the 9th IEEE Conference on Visualization VIS 1998*. IEEE, Oct. 1998, pp. 189–196.

[19] B. Berendt and I. Subasic, "STORIES in time: A graph-based interface for news tracking and discovery," in *Web Intelligence/IAT Workshops*. IEEE, 2009, pp. 531–534.

[20] A. Kontostathis, L. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps, *A Survey of Emerging Trend Detection in Textual Data Mining*, 2003.

[21] Y. Yang, L. Akers, T. Klose, and C. B. Yang, "Text mining and visualization tools - impressions of emerging capabilities," *World Patent Information*, vol. 30, no. 4, pp. 280 – 293, 2008.

[22] G. Salton, A. Wong, and A. C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, pp. 229–237, 1975.

[23] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, Mar. 2002.

[24] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005*, R. Grossman, R. J. Bayardo, and K. P. Bennett, Eds. ACM, 2005, pp. 198–207.

[25] M.-F. Moens, *Information Extraction, Algorithms and Prospects in a Retrieval Context*. Springer, 2006.

[26] M. Gregory, N. Chinchor, and P. Whitney, "User-directed sentiment analysis: Visualizing the affective content of documents," in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. Sydney: Association for Computational Linguistics, 2006.

[27] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computing*, vol. 5, no. 18, p. 401–409, 1969.

[28] X. Lin, D. Soergel, and G. Marchionini, "A Self-organizing semantic map for information retrieval," in *Proc. 14th. Ann. Int. ACM/SIGIR Conf. on R & D In Information Retrieval*, 1991, pp. 262–269.

[29] M. Chalmers and P. Chitson, "Bead: Explorations in information visualization," in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, Copenhagen, 1992.

[30] E. Rennison, "Galaxy of news: An approach to visualizing and understanding expansive news landscapes," in *ACM Symposium on User Interface Software and Technology*, 1994, pp. 3–12.

[31] G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie, "Knowledge mining with vxinsight: Discovery through interaction," *J. Intell. Inf. Syst*, vol. 11, no. 3, pp. 259–285, 1998.

[32] S. Kaski, K. Lagus, and T. Kohonen, "WEBSOM - self-organizing maps of document collections," *Neurocomputing*, vol. 21, pp. 101–117, 1998.

[33] J. Risch, D. Rex, S. Dowson, T. Walters, R. May, and B. Moon, "The starlight information visualization system," in *Proc. IEEE Conference on Information Visualization*. IEEE CS Press, 1997, pp. 42–49.

[34] S. Havre, E. G. Hetzler, and L. T. Nowell, "Themeriver: Visualizing theme changes over time," in *INFOVIS*, 2000, pp. 115–124.

[35] A. Kabán and M. Girolami, "A dynamic probabilistic model to visualise topic evolution in text streams," *Journal of Intelligent Information Systems*, vol. 18, no. 2-3, pp. 107–125, 2002.

[36] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann, "The infosky visual explorer: exploiting hierarchical structure and document similarities," *Information Visualization*, vol. 1, no. 3-4, pp. 166–181, 2002.

[37] P. C. Wong, H. Foote, D. Adams, W. Cowley, and J. Thomas, "Dynamic visualization of transient data streams," in *INFOVIS*. IEEE Computer Society, 2003.

[38] B. Fortuna, M. Grobelnik, and D. Mladenic, "Visualization of text document corpus," *Informatica (Slovenia)*, vol. 29, no. 4, pp. 497–504, 2005.

[39] F. V. Paulovich and R. Minghim, "Text map explorer: a tool to create and explore document maps," in *IV*. IEEE Computer Society, 2006, pp. 245–251.

[40] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant, "Discovering interesting usage patterns in text collections: integrating text mining with visualization," in *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*. ACM, 2007, pp. 213–222.

[41] M. Ghoniem, D. Luo, J. Yang, and W. Ribarsky, "NewsLab: Exploratory Broadcast News Video Analysis," *2007 IEEE Symposium on Visual Analytics Science and Technology*, pp. 123–130, 2007.

[42] F. V. Paulovich, M. C. F. de Oliveira, and R. Minghim, "The projection explorer: A flexible tool for projection-based multidimensional visualization," in *SIBGRAPI*. IEEE Computer Society, 2007, pp. 27–36.

[43] A. B. Alencar, M. C. F. de Oliveira, F. V. Paulovich, R. Minghim, and M. G. Andrade, "Temporal-pex: Similarity-based visualization of time series," 2007.

[44] I. Y. and H. M., "T-scroll: Visualizing trends in a time-series of documents for interactive user exploration," *Lecture Notes in Computer Science*, vol. 4675, pp. 235–246, 2007.

[45] M. Terachi, R. Saga, Z. Sheng, and H. Tsuji, "Visualized technique for trend analysis of news articles," in *New Frontiers in Applied Artificial Intelligence, 21st International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2008, Wroclaw, Poland, June 18-20, 2008, Proceedings*, ser. Lecture Notes in Computer Science, vol. 5027. Springer, 2008, pp. 659–668.

[46] S. Petrovic, B. D. Basic, A. Morin, B. Zupan, and J.-H. Chauchat, "Textual features for corpus visualization using correspondence analysis," *Intell. Data Anal*, vol. 13, no. 5, pp. 795–813, 2009.

[47] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. A. Keim, and O. Deussen, "Document cards: A top trumps visualization for documents," *IEEE Trans. Vis. Comput. Graph*, vol. 15, no. 6, pp. 1145–1152, 2009.

[48] R. Prabowo, M. Thelwall, and M. Alexandrov, "Generating overview timelines for major events in an RSS corpus," *J. Informetrics*, vol. 1, no. 2, pp. 131–144, 2007.

[49] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, Jun. 1990.

[50] J. E. Jackson, *A User's Guide to Principal Components*. John Willey, New York, 1991.

[51] M. J. Greenacre, *Correspondence analysis in practice*. Chapman and Hall, 2007.

[52] http://irthoughts.wordpress.com/2007/05/03/demystifying-lsa-lsi-svd-pca-and-is-acronyms/.

[53] J. B. Kruskal and M. Wish, *Multidimensional Scaling*. Beverly Hills, CA: Sage Piblications, 1978.

[54] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1, pp. 143–175, Jan. 2001.

[55] P. Agarwal, S. Har-Peled, and K. Varadajan, "Geometric approximations via coresets," 2005.

[56] J. York, S. Bohn, K. Pennock, and D. Lantrip, "Clustering and dimensionality reduction in spire," in *In AIPA Steering Group. Proceedings of the Symposium on Advanced Intelligence Processing and Analysis*, Washington DC: Office of Research and Development, 1995.

[57] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz, "Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 3, pp. 564–575, 2008.

[58] T. Kohonen, *Self-Organizing Maps*. Springer, Berlin, 1995.

[59] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[60] R. Minghim, F. V. Paulovich, and A. A. Lopes, "Content-based text mapping using multidimensional projections for exploration of document collections," in *IS&T/SPIE Symposium on Electronic Imaging - Visualization and Data Analysis*, San Jose, CA, USA, 2006.

[61] B. Shneiderman, "Treemaps for space-constrained visualization of hierarchies," http://www.cs.umd.edu/hcil/treemap-history/index.shtml.

[62] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD Workshop*, 1994, pp. 359–370.

[63] E. J. Keogh, S. Lonardi, C. A. Ratanamahatana, L. Wei, S.-H. Lee, and J. Handley, "Compression-based data mining of sequential data," *Data Min. Knowl. Discov*, vol. 14, no. 1, pp. 99–129, 2007.

[64] M. Hearst, *User Interfaces and Visualization*. Addison-Wesley Longman Publishing Company, 1999.

[65] D. M. Eler, F. V. Paulovich, M. C. F. d. Oliveira, and R. Minghim, "Coordinated and multiple views for visualizing text collections," in *IV '08: Proceedings of the 2008 12th International Conference Information Visualisation*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 246–251.

[66] J. E. McGrath, *Methodology matters: doing research in the behavioral and social sciences*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995.

[67] "Microsoft pivot project," http://www.getpivot.com/.

APPENDIX - VISUALIZATION EXAMPLES

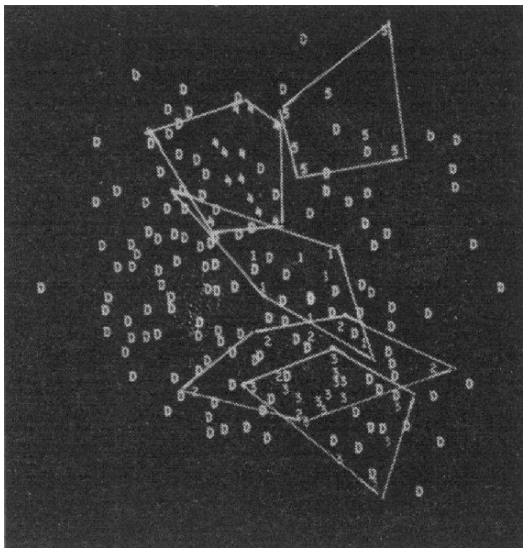

Figure 1. *Sammon* visualization [27]

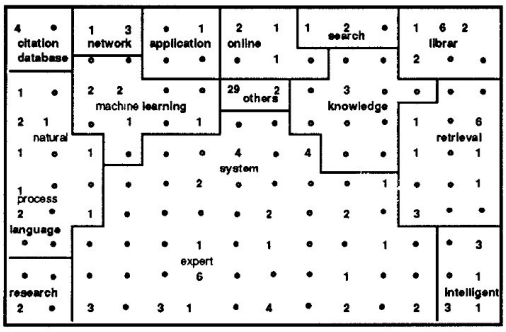

Figure 2. *Lin et al.* visualization [28]

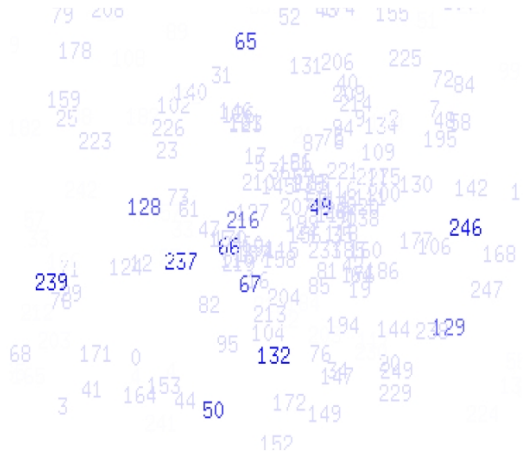

Figure 3. BEAD visualization [29]

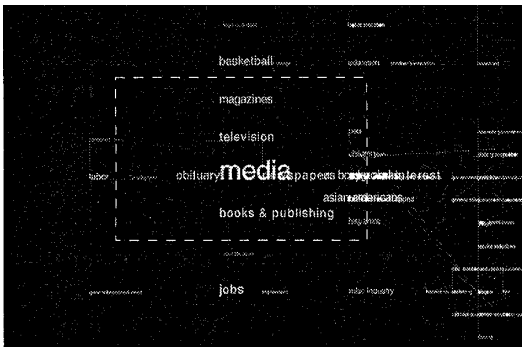

Figure 4. Galaxy of News visualization [30]

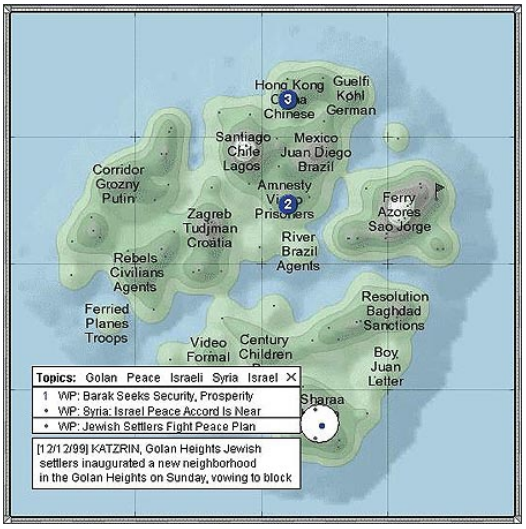

Figure 5. SPIRE / IN-SPIRE visualization [17]

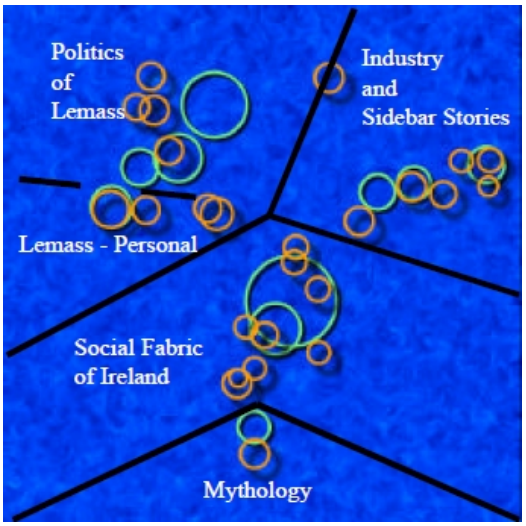

Figure 6. TOPIC ISLANDS visualization [18]

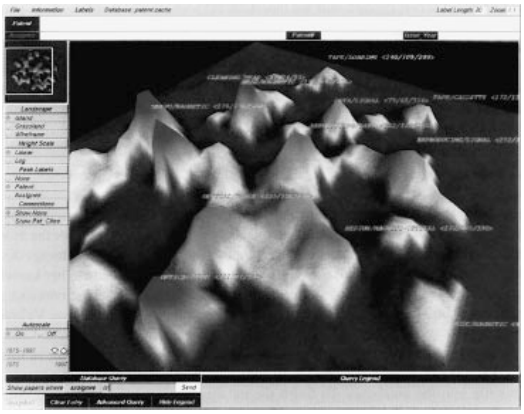

Figure 7. VxInsight visualization [31]

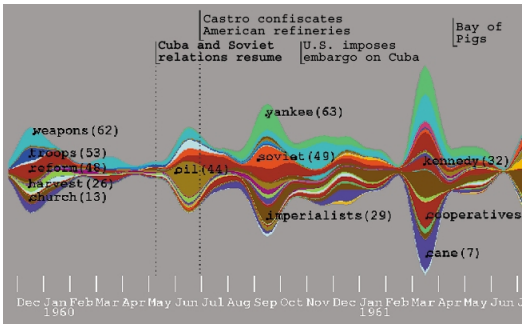

Figure 10. ThemeRiver visualization [34]

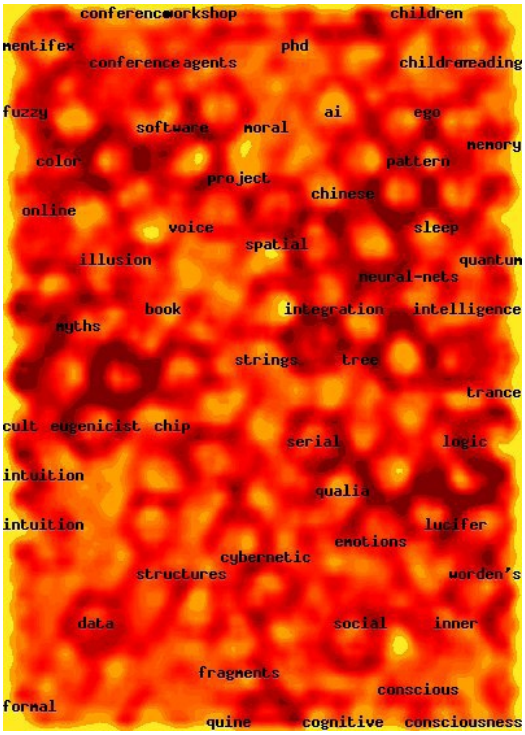

Figure 8. WEBSOM visualization [32]

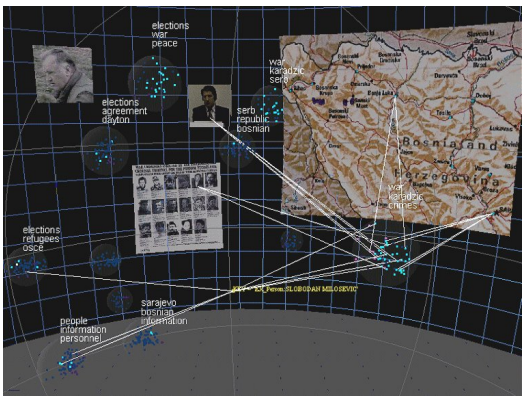

Figure 9. Starlight visualization [33]

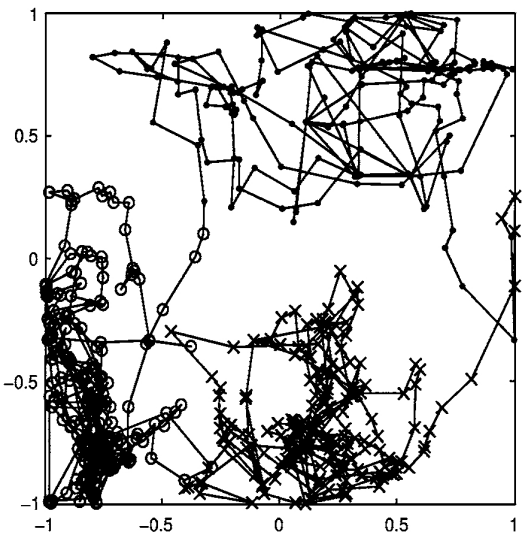

Figure 11. *Kaban and Girolami* visualization [35]

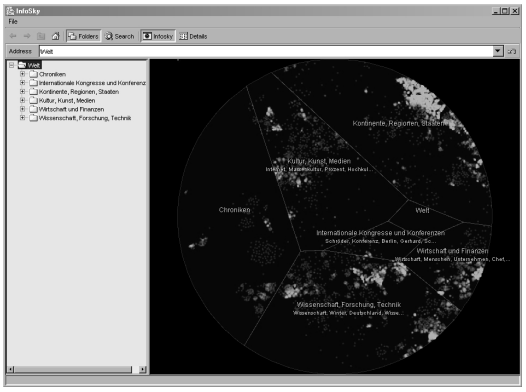

Figure 12. InfoSky visualization [36]

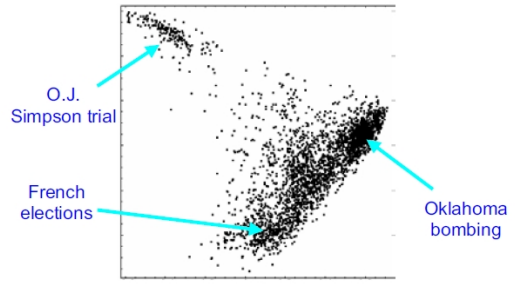

Figure 13. *Wong et al.* visualization [37]

Figure 14. NewsMap visualization [12]

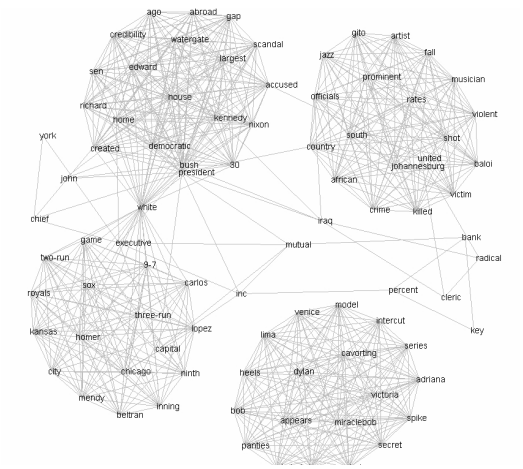

Figure 15. TextPool visualization [13]

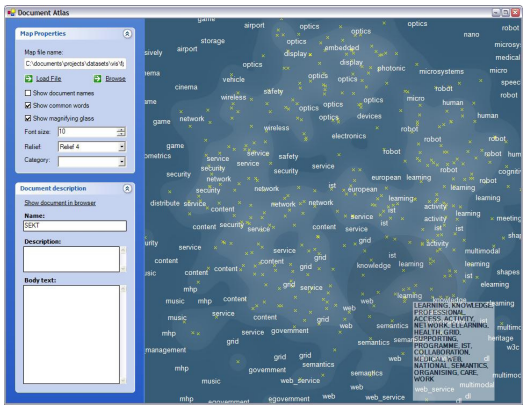

Figure 16. Document Atlas visualization [38]

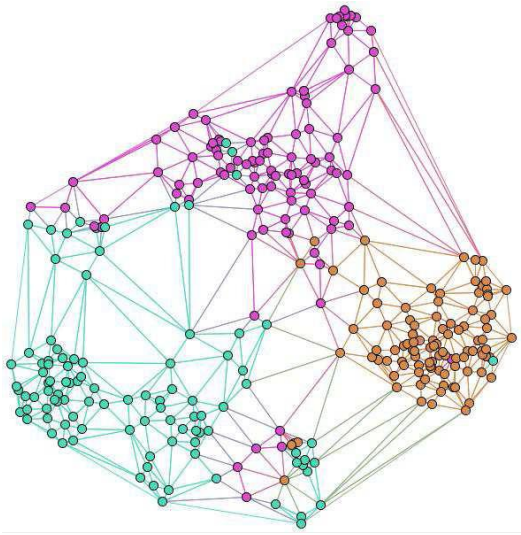

Figure 17. Text Map Explorer visualization [39]

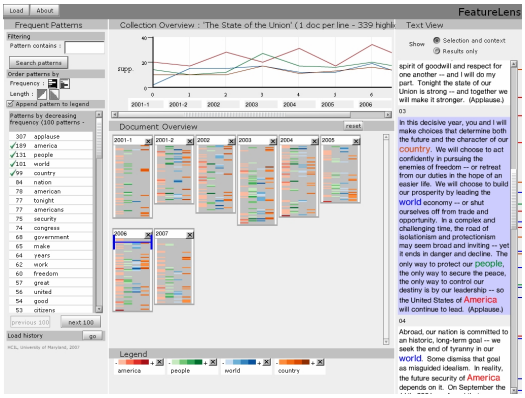

Figure 18. FeatureLens visualization [40]

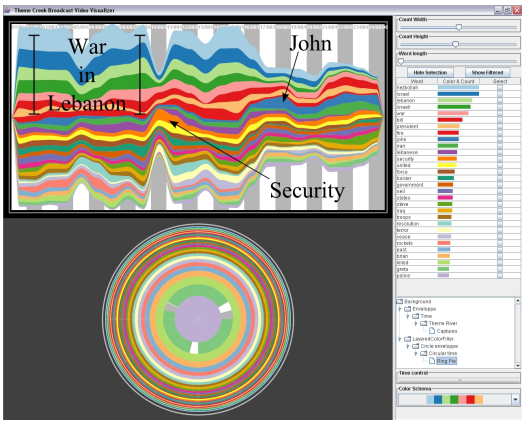

Figure 19. NewsRiver, LensRiver visualization [41]

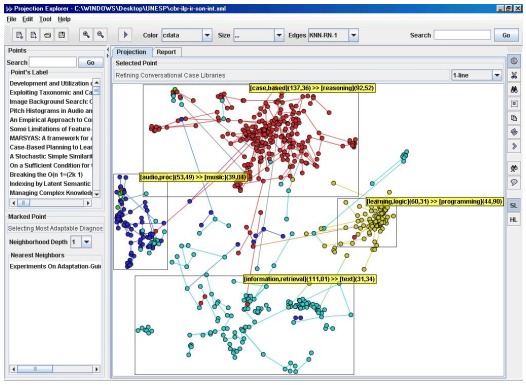

Figure 20. Projection Explorer (PEx) visualization [42]

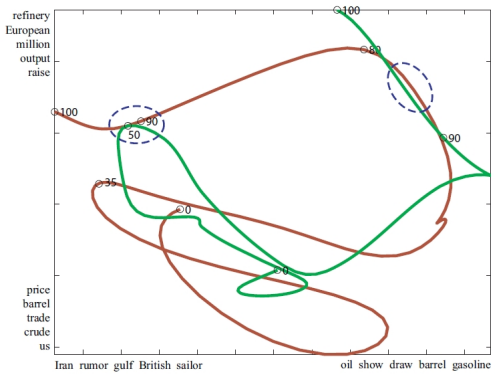

Figure 21. SDV visualization [14]

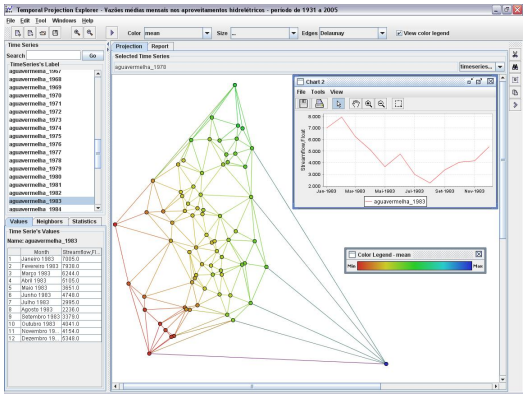

Figure 22. Temporal-PEx visualization [43]

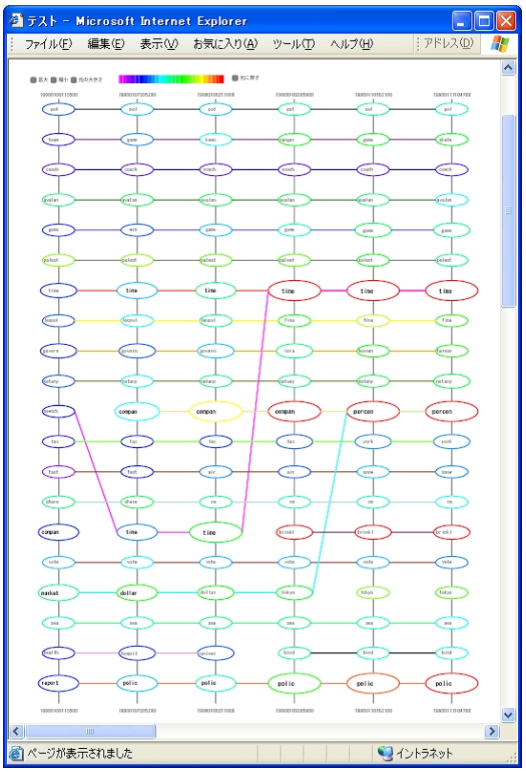

Figure 23. T-Scroll visualization [44]

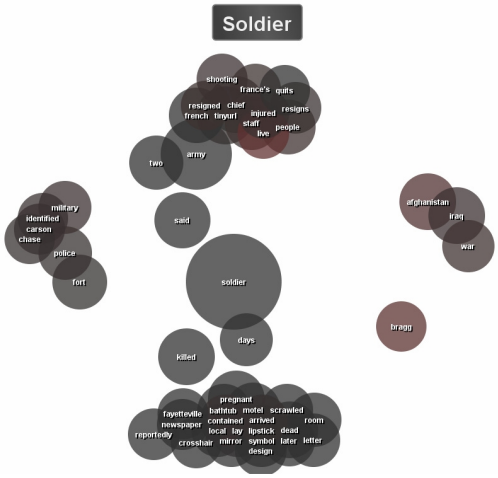

Figure 24. *Benson et al.* visualization [11]

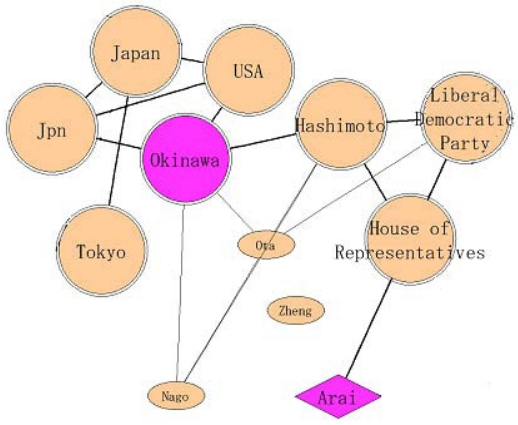

Figure 25. FACT-Graph visualization [45]

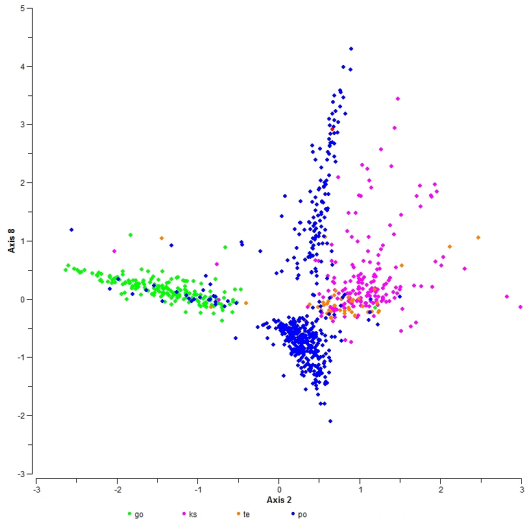

Figure 26. *Petrović et al.* visualization [46]

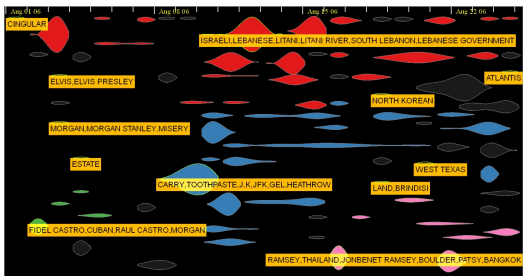

Figure 27. EventRiver visualization [8]

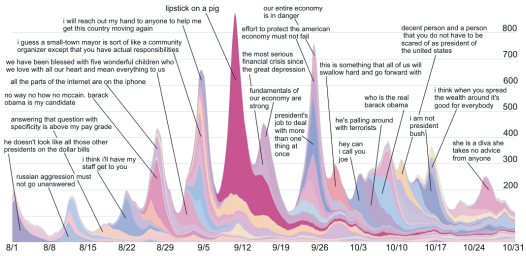

Figure 28. MemeTracker visualization [16]

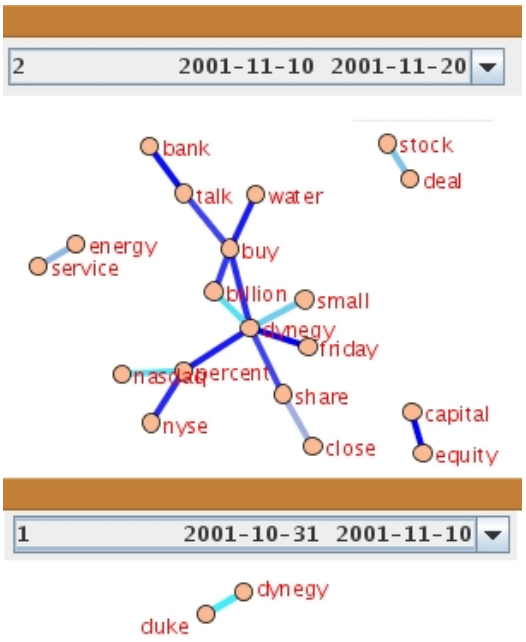

Figure 29. STORIES visualization [19]