Real-Time Visualization of Streaming Text Data: Tasks and Challenges

Christian Rohrdantz, Daniela Oelke, Miloš Krstajić and Fabian Fischer

Abstract— Real-time visualization of text streams is crucial for different analysis scenarios and can be expected to become one of the important future research topics in the text visualization domain. Especially the complex requirements of real-time text analysis tasks lead to new visualization challenges, which will be structured and described in this paper. First, we give a definition of what we consider to be a text stream and emphasize the importance of different real-world application scenarios. Then, we summarize research challenges related to different parts of the analysis process and identify those challenges that are exclusive to real-time streaming text visualization. We review related work with respect to the question which of the challenges have been addressed in the past and what solutions have been suggested. Finally, we identify the open issues and potential future research subjects in this vibrant area.

Index Terms—real-time, text streams, streaming data, dynamic visualization, visual analytics.

1 Introduction

Unstructured and semi-structured text data, such as news articles, Twitter posts or user reviews are being generated in volumes that grow every day. Many analysis tasks require an effective and efficient combination of visualization and automated methods that can be applied on text streams and would assist the user in extracting useful information from this vast content. Very often, the user response time has to be as short as possible and the data can grow in rates that render the common data processing and visualization methods useless.

Sometimes, *text stream* is used to describe: (a) a time-stamped, temporally ordered text collection, or (b) a data stream containing unstructured or semi-structured text. The first definition does not imply that the data analysis and visualization algorithms work with a constantly evolving dataset that has to be processed online. In our work, we consider text data streams that rely on a model described in [4]: (a) "*The data elements in the stream arrive online.*" (b) "*The system has no control over the order in which data elements arrive to be processed, either within a data stream or across data streams.*" (c) "*Data streams are potentially unbounded in size.*" (d) "*Once an element from a data stream has been processed it is discarded or archived it cannot be retrieved easily unless it is explicitly stored in memory, which typically is small relative to the size of the data streams.*"

In this paper, we are addressing the challenges of the visual analysis of real-time text streams and related work in this area. First, in Section 2 we describe three important real-world scenarios for which the visual analysis of text streams in real-time is crucial: Emergency management and response, news and stock market, and server administration and log management. The scenarios differ in terms of user tasks and requirements and rely on text streams with different characteristics, i.e. structure, content, amount and temporal dynamics.

Based on the analysis tasks, in Section 3 different kinds of challenges will be identified and discussed in detail. In addition, general research and concrete solutions from the past will be summarized. First of all, preprocessing and incrementally structuring text in real-time requires special algorithms, which will be further detailed in Section 3.1. Section 3.2 discusses different ways of modeling and conveying the temporal context visually and Section 3.3 shows the implications and challenges of a constantly updated visual environment. Next, Section 3.4 treats the process of sense-making using real-time text visualization and discusses ways to support it. Finally, Section 3.5 identifies those real-time visualization challenges that are special to text data.

Section 4 concludes the paper pointing out open issues in the real-time visualization of text streams. Future research directions are identified and first possible solutions suggested.

2 Tasks

There are many real-time applications where users need to get insights into data streams immediately to facilitate decision making. The implementations of such applications and their visualization techniques are strongly influenced by the overall tasks and analysis goals. According to Thomas et al. [22] three significant problems are to provide situational awareness, show changes and fuse different data sources. This leads to the main tasks for the visual analysis of streaming text data, which are monitoring, decision making, change and trend detection, event tracking, historical retrieval and exploration of data items to eventually achieve situational awareness [10]. Table 1 describes and gives examples for each of these tasks.

To cover a broad field of real-world applications and tasks dealing with text streams, we highlight three important scenarios, which are (1) *emergency management and response*, (2) *news and stock market* and (3) *server administration and log management*.

2.1 Emergency management and response

A typical case for homeland security and catastrophe response is the occurrence of earthquakes. Thousands of seismic stations all over the world track movements of the earth's surface. However, these sensors cannot provide information about the actual situation of the people living in the affected regions. Mining user-generated textual data streams (e.g., Twitter, reports from experts) can reveal heavily affected areas and can help to describe and understand the situation and its change in real-time. The same is true for other local or global catastrophic or terrorist events (e.g., power blackouts, hurricanes, accidents, attacks), which can effectively be monitored using user-generated data streams. The overall goal of this scenario is to provide better response to emergencies.

In general, such short user-generated messages have little structure, grammar and context and are provided as high-frequency multilingual stream containing a high percentage of non-meaningful and irrelevant messages. Temporal and geographic dynamics do result in unexpected and periodic peaks, depending on time, weekday and population density of the monitored region.

Monitoring and *decision making* are important tasks in this scenario. The acceptable monitoring delay, which is an important requirement, depends on the current situation: During an initial catastrophe response, there are many analysts available to continuously and closely monitor the situation in command and control centers. In such situations a *short latency and delay* is required, while in other cases,

• *All authors are with the University of Konstanz, Germany. E-mail: firstname.lastname@uni-konstanz.de*

High-Level Task	Possible Question	Example
Monitoring	What is the current situation in the affected geographic areas based on the real-time text streams?	Scenario 2.1
Decision Making	Which geographic area has the highest priority and needs immediate help first?	Scenario 2.1
Change & Trend Detection	Are there any news reports which might influence the stock market?	Scenario 2.2
Event Tracking	How does an identified and relevant situation evolve over time?	Scenario 2.2
Historical Retrieval	Have there been similar events or log entries in the past?	Scenario 2.3
Exploration	Is there further information that helps to reason about the cause of an interesting event?	Scenario 2.3
Situational Awareness	What is the current status of the whole computer network? Are there any ongoing threats?	Scenario 2.3

Table 1. Overview of important high-level tasks with possible task-related scenario questions.

monitoring might only be needed from time to time. Often a fully automated alerting system is needed, to enable early communication and de-escalation of unusual events. However, keyword-based emergency detection cannot be used to cover all those events. It is not possible to use a fixed set of keywords, because it is not possible to expect the unexpected. This also leads to the requirement of *data source fusion* to correlate different user-generated text streams automatically.

2.2 News and stock market

Complex mathematical and stochastic models aim to help analysts to predict stock performance. However, this is not always possible. Unexpected behavior is often due to some unexpected news, which heavily influences the situation.

Thus, real-time *change and trend detection* in and reaction to financial news (e.g., Reuters news stream) help to make better decisions and to prevent severe financial losses. *Event tracking* is needed to follow topics which might become relevant later and develop over time.

During business hours this information is extremely time-critical and should be responded to in a matter of seconds. This highlights the important requirement of having a *short latency and delay*. Opposed to the previous scenario, larger documents and news articles are to be expected. Unexpected and heavy peaks and follow-up news articles focusing on the same event are very likely to occur. To relate and put these items into the *temporal context* is another crucial requirement within this scenario.

2.3 Server administration and log management

The analysis of log data (e.g., textual status, error, debug messages and alerts) of server systems is critical from a security and business point of view. Each process or service running on servers is producing a vast amount of textual data to log current activity or error messages, which may contain arbitrary textual contents written by the application developers. An early detection of problems and server outages is important to any company. Log files contain valuable information to recognize suspicious behavior or even attacks in their early stages. Having peaks of thousands of events per minute and more is common in many computer networks.

To be aware of such events and not be overwhelmed with the large amount of data, *situational awareness* is needed. To investigate root causes of problems and bugs, *historical retrieval* and *exploration* are other important tasks.

Similar to the first scenario, there are situations where continuous monitoring is needed, while in other cases regular checks of the most important events on a daily basis is sufficient. To provide a *general overview* is therefore an important requirement for the visualization of such system log events.

3 Challenges

Section 2 shows that the different analysis scenarios are challenging and quite diverse with respect to both tasks and requirements. The demands they make on automatic and visual methods cover several main aspects: Technical and algorithmic challenges, the challenge of providing temporal context as well as challenges of dynamic visualization and the support of sense-making. While some of these challenges apply to the visual analysis of different data streams as well, others are exclusive to text streams. The following subsections will discuss the different challenges in detail.

3.1 Technical and algorithmic challenges

The technical and algorithmic challenges include all kinds of database related issues [4]. Most importantly, fast algorithms and data structures are needed that enable real-time processing and can deal with incremental updates. Additionally, methods should not depend on apriori assumptions regarding data load, arrival rate or thresholds, because streaming data may have unpredictable contents. Moreover, we assume that streams cannot be stored completely because of their enormous size and that consequently on-the-fly processing and visualization is required. Usually, approaches either process each incoming data item individually or store data items in a buffer according to a predefined time frame, and then process the buffer content in regular intervals. However, it is not quite clear how to come up with suitable time frame sizes and consequently some approaches enable the user to modify this parameter dynamically [1, 14].

There are different ways to address some of the outlined challenges. Wong et al. [24] suggest to make the algorithmic analysis dependent on the volume of incoming data. If in the short term there is a high data load, the trade-off is between being less accurate, as in [24], and temporarily buffering data as done by Alsakran et al. [3] who buffer "document items during peak times and handle them in idling periods."

One important issue is performing topic modeling in real-time. Ishikawa and Hasegawa [14] cluster documents in real-time incorporating a "document forgetting model", i.e. older documents have less influence on the clusters. The clustering is an incremental k-means algorithm, which has the drawback that the parameter k has to be to predefined. Zhang et al. [25] introduce an evolutionary hierarchical Dirichlet process that allows the number of clusters to vary over time, but they do not comment on a potential real-time capability of their approach. Rose et al. [21] cluster keywords for each time interval using a hierarchical agglomerative clustering algorithm in order to learn themes without having to predefine the number of clusters. Themes of adjacent time intervals are grouped into "stories" according to their similarity. These stories can split and merge over time.

Another issue is querying in real-time. Hetzler et al. [13] allow the user to define a network combining different dynamic queries and visualize the corresponding result sets in the same structure.

For many approaches, there is no explicit description of how they behave in the initialization phase, i.e. when the stream starts and the first data items come in. While some algorithms might not need data from the past others heavily depend on it and will require some time to become stable and meaningful.

3.2 Providing temporal context

Not only the algorithmic part is demanding when working with streaming data but also the visual representation. An inherent property of streaming data sets is that new data is constantly arriving (at irregular or regular intervals). This results in several challenges: (a) Approaches have to deal with a constantly growing amount of data. (b) The new data has to be shown in the context of older data. (c) If the data arrives at irregular intervals, fluctuations in the amount of (new) data are to be expected.

Therefore, algorithms can be distinguished with respect to what portion of the data stream is explored. Most approaches work with a time buffer[1] that contains the most recent data (see e.g., [17, 20, 13,

[1] Also called time frame, time window, temporal pooling, or time interval.

1, 9]). Dubinko et al. [9] report different results of the analysis based on the size of the buffer. Since choosing the right buffer size is difficult and may be subject to the task, [1] and [14] suggest to let the user control the time frame. Alsakran et al. [3] use two time frames (buffers) with different levels of detail for in-depth and context analysis. In contrast to displaying only the currently arriving data (snapshots), working with time intervals comes with the advantage that some context can be provided but at the same time the amount of data that has to be displayed is still assessable.

Rose et al. [21] state that "Accurately identifying and intelligently describing change in an information space requires a context that relates *new* information with *old*." Providing temporal context is important for many analysis tasks. We can distinguish between implicit and explicit representation of context. Implicit representations show new data in the context of older events but do not allow to reconstruct the evolution over time. Examples for implicit context representation include [1, 3] that add new data to a dynamic graph structure whose content and layout is influenced by past data. Color is used to further emphasize the age of the represented documents (see also [13]). In contrast to this, explicit context is provided when the documents are represented along a time line like in [8, 19, 11, 17, 14, 21]. Some of these approaches distort the time and give more space to recent data. Distorting the time dimension comes with the advantage that more data can be displayed and that certain time ranges are emphasized.
Furthermore, we can distinguish between techniques that are interval-based (e.g., [14, 21]) and those that visualize context as a continuous flow (e.g., [19, 17]).

3.3 Dynamic visualization challenges

Visualizations must necessarily change over time to reflect changes of the stream. There are two different approaches to visually express changes in the data. The first and more straight-forward solution is to create individual views for different time intervals and sequentially provide this series of views. Here, the challenge is to relate one view to the next view and moreover assure that the user is able to perceive the differences between both. Usually, some sort of transition is used like the animation of appearing and disappearing words in the Parallel Tag Clouds [7]. Many standard text visualizations can potentially be adapted to this scenario.
The second and more demanding possibility is to provide a continuously evolving visualization reflecting a continuous data flow. In such a scenario the current view is constantly modified in accord with the newly arriving data. The modification is triggered by an update and usually expressed through motion. A nice example for the use of motion is [1]. The authors review the perception of motion and derive design guidelines for visualizing changes in live text streams. More recent approaches go in a similar direction: Dörk et al. [8] chose to "use primarily motion to represent data change" and Alsakran et al. [3] point out that "the dynamic procedure of this change is critical for reducing change blindness". However, even if changes are nicely traceable using motion, too many changes at once will overwhelm the user. The system of Alsakran et al. [3] addresses this issue by having at most one label change when a new document arrives and buffering documents during peak times at the cost of introducing delays.
The problem of having a lot of change in the display on updates is also given in the case of visualizations that display the arrival order of items by position. Hao et al. [12] discuss different ways of shifting displayed items on the arrival of new items minimizing the movement in the display. Chin et al. [6] discuss how standard visualizations can be extended to support the analysis of dynamic data and experiment with different ways of updating their dynamic spiral timeline keeping the display more ore less constant.
In the case of frequently or even continuously updating displays, support for interactive exploration is a further challenge. While a lot of approaches enable some user interactions like filtering and highlighting search queries on-the-fly, in-depth exploration is difficult. Interaction with a visualization while the stream keeps coming in, brings the problem of "adapting its display to stream content and user interaction" [1]. Those are two potentially conflicting triggers for changing the display. What if the user might like to explore a peculiarity further? Does he have to worry that it might disappear in the next moment when new data comes in (and could be lost forever)? One solution to prevent this is to freeze a certain point in time for exploration, the user can pause the stream for exploration [13, 3] or take interactive snapshots to save the current situation for a later off-line exploration as in [13]. Many other approaches, however, do not address this problem explicitly or assume that the whole stream can be saved and retrieved.

3.4 Supporting sense-making in complex analysis tasks

The methods and techniques described in the previous sections in the end all have the goal to enable the user to perform his sense-making and analysis tasks. While the text data can be automatically preprocessed, structured and visually conveyed, the semantics of texts in the end will have to be interpreted by the user. Several issues make the visual analysis of text streams especially demanding. Compared to numerical data streams, the number of *topics*[2] can dynamically change. These *topics* could be, for example, topics appearing in online news streams that report on real world events. In the field of text mining, a lot of research has recently been done on online topic mining, with three main directions: Topic detection and tracking [2], dynamic topic modeling [5], and evolutionary clustering [25]. The real world events that are represented by the topics in the text stream appear and disappear over time, but they can also overlap, split and merge, showing *episodic* and *braided* nature [16]. Kleinberg states that "A fundamental problem in text data mining is to extract meaningful structure from document streams that arrive continuously over time." [15] The events in text streams usually appear as *bursty* patterns with latent hierarchical structure. These complex characteristics pose challenges in the area of visualization such as displaying the temporal evolution of a large number of topics incrementally and displaying topic relationships in a temporal context. Some approaches like EventRiver [19] or TIARA [23] have already been suggested, but all in all little research has been done to visualize topic evolution in a streaming environment.

3.5 Special challenges of real-time streaming TEXT visualization

Many of the before mentioned challenges do not only apply to the analysis of streaming *text* data but also to other kinds of data streams. However, in some respects text is truly special. Usually, the findings in text streams are of a semantic, rather qualitative nature whereas in other streams events may also be formally described. This complicates computational support but also the evaluation of the methods because there are no gold standards about what should be found in a stream. Also, it is difficult to generate artificial data that faithfully mimics a real-world scenario.

Obviously, the automatic algorithms that are applied have to be suitable for text data and there is no easy way to represent text visually. Furthermore, a special challenge of this data type is that the data space is unbounded, i.e. certain words may appear that did not even exist when the monitoring process was started since language may contain neologisms and is prone to change. Similarly, often there is no fixed number of monitored sources (e.g., the number of topics may change dynamically) as is usually the case for sensor data, for example.

4 Open Issues

Although already a lot has been done in the domain, real-time streaming text visualization can still be considered as a young and expanding field with exciting challenges waiting to be solved. In the following we are going to outline some issues for which we see the potential of becoming interesting future research projects.

1. **Advanced definitions of interestingness**
   So far interestingness is mostly put on a level with being novel. However, it might as well be defined differently. Also, in an analysis task a certain interval in the past may suddenly become interesting in the light of current events. Novel challenges here

[2] Also called stories, themes, events, or event episodes.

include both algorithms for detecting documents that must be considered interesting with respect to a task but also visualizations that can deal with the requirement that at the same time an old and a new item may be in the focus of the analysis. Dubinko et al. [9] suggest to allow "random access into the time stream" and let the user reposition the current focus point. But to the best of our knowledge so far no solutions have been presented of how multiple focus points could be shown in a single visualization in a real-time stream.

2. **Dealing with fluctuations**
   Usually, the data in a text stream does not arrive at regular intervals. This causes fluctuations with respect to the number of documents that have to be displayed in different time intervals. First steps are taken in [18] that could be built on in future work.

3. **Supporting interactive exploration**
   Exploration per definition involves going back and forth in a data set. However, this is challenging if the visualization is changing permanently. Hetzler et al. [13] state that "Retaining interaction, analytic, and reasoning context over changing data remains a critical goal." Currently, in-depth exploration is taken offline and the live interaction is mainly limited to filtering, searching and highlighting (see section 3.3). Both for the visualization and interaction new ways need to be found to support exploration within the live display without interrupting the stream.

4. **Collaborative analysis of streaming data**
   Another open issue for future research is how the analysis of streaming data can be done in a collaborative environment. Often, the analysis questions are rather complex and require reasoning that involves the collaboration of multiple analysts.

5. **Algorithmic challenges**
   In this paper we put a focus on the visualization challenges. But of course also many algorithmic challenges exist that need to be addressed in the future. This includes issues like temporal alignment of multiple streams, supporting an appropriate, domain-dependent parametrization of advanced text processing algorithms, or preparing standard document analysis methods for real-time requirements.

In conclusion it can be said that real-time text visualization is an evolving topic with interesting research questions being left for future work. Due to the important application scenarios that require the analysis of streaming text data, we hope that many exciting papers in the domain are to be expected in the future.

Acknowledgments

The German Research Society (DFG) under grant GK-1042, "Explorative Analysis and Visualization of Large Information Spaces" partially supported this work.

References

[1] C. Albrecht-Buehler, B. Watson, and D. A. Shamma. Visualizing Live Text Streams Using Motion and Temporal Pooling. *IEEE Computer Graphics and Applications*, 25:52–59, 2005.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking Pilot Study Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[3] J. Alsakran, Y. Chen, Y. Zhao, J. Yang, and D. Luo. STREAMIT: Dynamic visualization and interactive exploration of text streams. *IEEE Pacific Visualization Symposium*, 0:131–138, 2011.

[4] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '02, pages 1–16, New York, NY, USA, 2002. ACM.

[5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[6] G. Chin, M. Singhal, G. Nakamura, V. Gurumoorthi, and N. Freeman-Cadoret. Visual analysis of dynamic data streams. *Information Visualization*, 8:212–229, June 2009.

[7] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel Tag Clouds to explore and analyze faceted text corpora. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pages 91–98, 2009.

[8] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale. A Visual Backchannel for Large-Scale Events. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Information Visualization 2010)*, 16(6):1129–38, 2010.

[9] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. *ACM Transactions of the Web (TWEB)*, 1, August 2007.

[10] M. Endsley. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human*, 37(1):32–64, 1995.

[11] D. Fisher, A. Hoff, G. Robertson, and M. Hurst. Narratives: A Visualization to Track Narrative Events as they Develop. In *IEEE Symposium on Visual Analytics and Technology (VAST 2008)*, pages 115–122, 2008.

[12] M. C. Hao, D. A. Keim, U. Dayal, D. Oelke, and C. Tremblay. Density Displays for Data Stream Monitoring. *Computer Graphics Forum*, 27(3):895–902, 2008.

[13] E. G. Hetzler, V. L. Crow, D. A. Payne, and A. E. Turner. Turning the Bucket of Text into a Pipe. In *Proceedings of the 2005 IEEE Symposium on Information Visualization*, pages 89–94. IEEE Computer Society, 2005.

[14] Y. Ishikawa and M. Hasegawa. T-Scroll: Visualizing Trends in a Time-Series of Documents for Interactive User Exploration. In L. Kovács, N. Fuhr, and C. Meghini, editors, *Research and Advanced Technology for Digital Libraries*, volume 4675 of *Lecture Notes in Computer Science*, pages 235–246. Springer Berlin / Heidelberg, 2007.

[15] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101. ACM, 2002.

[16] J. Kleinberg. *Temporal Dynamics of On-Line Information Streams*. Springer, 2006.

[17] M. Krstajic, E. Bertini, and D. A. Keim. CloudLines: Compact Display of Event Episodes in Multiple Time-Series. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis '11)*, 2011.

[18] M. Krstajic, E. Bertini, F. Mansmann, and D. A. Keim. Visual analysis of news streams with article threads. In *StreamKDD '10: Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, pages 39–46, New York, NY, USA, 2010. ACM.

[19] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. A. Keim. EventRiver: Visually Exploring Text Collections with Temporal References. *IEEE Transactions on Visualization and Computer Graphics*, 2010.

[20] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 227–236. ACM, 2011.

[21] S. J. Rose, S. Butner, W. Cowley, M. L. Gregory, and J. Walker. Describing story evolution from dynamic information streams. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST '09)*.

[22] J. Thomas, K. Cook, I. Electrical, and E. Engineers. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society, 2005.

[23] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. TIARA: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 153–162, New York, NY, USA, 2010. ACM.

[24] P. C. Wong, H. Foote, D. Adams, W. Cowley, and J. Thomas. Dynamic visualization of transient data streams. In *Proceedings of the Ninth annual IEEE conference on Information visualization*, InfoVis'03, pages 97–104, Washington, DC, USA, 2003. IEEE Computer Society.

[25] J. Zhang, Y. Song, C. Zhang, and S. Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1088, New York, NY, USA, 2010. ACM.