Sharpening Analytic Focus to Cope with Big Data Volume and Variety:
Ten strategies for data focusing with temporal event sequences

Ben Shneiderman (ben@cs.umd.edu) & Catherine Plaisant (plaisant@cs.umd.edu)
University of Maryland
Draft (12/2/2014)

The growing volumes of data available from sensors, social media sources, web logs, or medical histories present remarkable opportunities for researchers and policy analysts. While Big Data resources can provide valuable insights to understand complex systems, which lead to better decisions for business, national security, cybersecurity, and healthcare, there are many challenges to dealing with the volume and variety of data.

While *data cleaning* and *data wrangling* (Kandel 2011) has received some attention with the development of application tools (e.g., OpenRefine, 2014), *data focusing* to sharpen the analytical focus remains a challenge. An admirable example of pre-processing strategies to clean and prepare data is the 5 or 6 step process used in many NASA remote sensing projects (NASA, 2014).

This data focusing problem is being addressed in familiar relational databases, large network (graph) databases, and elsewhere, but this paper emphasizes temporal event sequences in which streams of *point* and *interval events* are organized into records. For example, patient histories might include point events, such as diagnoses, tests or surgery, and interval events, such as medication episodes, dieting plans, or hospitalizations. Each patient history, with a large variable number of events is considered to be one record. In addition to events, patients have attributes, such as gender or age, and events may also have attributes, such as which physician ordered the medication or which hospital provided care.

A growing number of visual analytic and statistical software tools are being built and rapidly refined to deal with temporal event sequences. These tools often have difficulty in dealing with two problems:

1) Volume of records: the number of records may grow to hundreds of millions, making it difficult to load or apply operations to the data
2) Variety of patterns: Within each record there may also be thousands or millions of events, coded using thousands or tens of thousands of event categories. For example, medical histories may include diagnoses that come from the 90,000+ ICD9 codes or medications that come from the 31,000+ medications listed in RxNorm. Most records are unique and the variety makes it difficult to see global patterns such as relationships, clusters or gaps, as well as to identify errors or anomalies.

In other domains such as social media log analysis a Twitter user may post thousands of times and also retweet, reply, mention, or take other actions. Web log analysis for shopping sites may include thousands of website visits, often recorded by which types of products are viewed and of course what purchases are made.

While it is tempting to desire an overview of the variety in the data, analysts need useful ways of sharpening the analytic focus, leading to useful visualizations of global patterns and anomalies of

interest. Just as camera images need to be in focus on objects or faces of interest and telescopes are best when tuned to spectral ranges (visual, ultraviolet, radio, x-ray, etc.), so too analytic tools will be most effective if users can focus their attention. The idea of an analytic pipeline or workflow is well established in mature topics such as pharmaceutical drug discovery or NASA's remote sensing data analysis.

Cleaning the data is critical (Kandel, 2011; Gschwandtner 2012), and needed before the sharpening of focus can occur. Often the first stages are to detect omissions or duplications in the data, then data cleaning begins to deal with dubious values from incorrect processing, human data entry error, failing sensors, etc. This is often far more complex than analysts expect as they discover the vagaries of domain specific data capture. One favorite example is the hospital that was trying to determine the average length of emergency room stays, but did not realize their data was faulty. The most extreme case was the patient who was admitted 14 times, but discharged only twice. In discussing with hospital managers, it became clear that admission is tied to billable events so it is usually accurately collected, but discharges are occasionally skipped. A detailed set of technical reports from the Health & Social Care Information Centre of the UK National Health Service (NHS, 2014), describes their data cleaning process.

Once data analysts begin their work there are many paths and questions, mostly driven by the analytic goal. In some mature application domains the accumulated experience of analysts has led to well-established strategies and workflows for data focusing, but in new domains, such as temporal event sequence analysis, we see a need for novel approaches. Our work in developing systematic yet flexible strategies (Perer and Shneiderman 2008) showed ways to structure analytic workflows - especially for network data. Wang et al. (2011) provided a starting point for developing systematic yet flexible workflows for temporal event sequences, which was pursued by Monroe (PHD 2013), and refined here. Our application of these ideas in EventFlow (www.cs.umd.edu/hcil/eventflow) gave us dozens of case studies from real users who applied our tools for their own investigations. Figure 1 provides one example of the dramatic difference from the initial view even with a small dataset of 215 patient records, which we refer to as the "confetti" view to a focused version that clearly reveals the 105 patients who received the correct treatment sequence (shown on the top right side), and the many ways that the required protocols were violated. The hospital managers are currently investigating the impact of these variations and training staff more carefully.

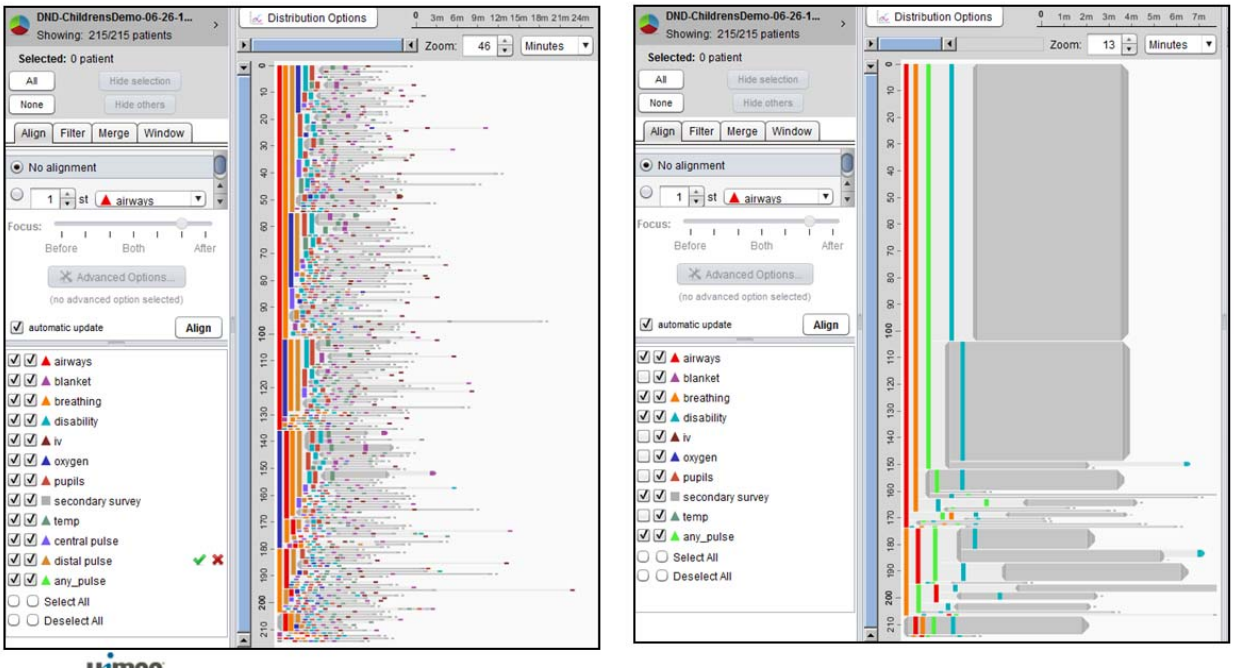

Figure 1: Before (left side) and after (right side) view of data focusing.

A taxonomy of analytic focusing strategies

This taxonomy of analytic focusing strategies for temporal event sequences, is based on our experience working with dozens of case studies using EventFlow (Monroe, 2014) and the knowledge of other visual analytics tools for sequence analysis (e.g. Rind et al., 2013, Gotz et al., 2014). Most strategies can lead to reduction in volume and variety simultaneously.

Extraction strategies

1) **Goal-driven record extraction:** For most projects dealing with large datasets, we find that the question at hand needs only a fraction of the records. For example, the US Army Pharmacovigillance Center has 15M patient histories, but the current project was to examine medications taken for asthma in the past six years, which produced a set of 182,000 records. Furthermore, for each record, only asthma related medications and events were extracted. For Washington Hospital Center with 1M+ records, only 3600 had been given the treatment that was being studied. Traditional query and extraction tools are needed before the focused analysis can begin.

2) **Goal-driven event extraction:** Some studies require all the records in a database, but may cover only a fraction of the events, such as those related to prostate cancer radiation treatment. Removal of other events, such as eye or dental exams or even the procedure's details will trims the dataset and greatly focus the analysis, e.g. if the goal is finding what durations and intensity of radiation produce the fewest bone fractures, yet still curtail the cancer. We found that successful analysts start with very few event types, then

progressively add more as needed to refine the analysis. Tools to guide the selection of those initial events can be provided.

3) **Temporal windowing:** In many cases only a small window of time matters. The selection might be arbitrary (e.g. only recent data) or goal driven, e.g. only the week before a purchase, or a month before and two months after surgery. Extraction tools are needed that limit the range of events, either using absolute dates (e.g. the holiday shopping period) or relative dates after alignment by a selected event (e.g. surgery).

4) **Random sampling**: If the previous strategies fail to sufficiently reduce the volume of data, random sampling may become a reasonable strategy if there is benefit in getting some indication of the prevalence of the patterns being sought (Fisher et al., 2012). Random sampling of events within records does not seem useful.

Folding strategy

5) Temporal folding: Some datasets have records that are long streams (one person entire twitter history with 1000s of events), which may be more successfully analyzing by folding (or splitting) each record into yearly, monthly, or weekly units. A study of Interpersonal Violence (IPV) had detailed 90-day records of drugs and alcohol use, as well as incidents of arguments, physical violence, sexual abuse, etc. Relationships were difficult to extract until the records were broken into weekly records thereby revealing the patterns of weekend conflicts and drug use. Temporal folding may not address data volume issues as the number of records increases - while keeping the number of events constant. However, once the temporal folding is done, the variety of patterns may be reduced and other strategies to reduce dataset size may become applicable.

Pattern simplification strategies

6) Grouping event categories (aggregation): With the explosion of the number of event categories aggregation becomes necessary. For example, there may 400 types of lung cancer and 200 types of bone cancer, making it impossible to see global patterns. Replacing all lung cancers with a single event category and all bone cancers with a single one as well will reduce the variety of patterns. While the number of events remains the same, the simplification sharpens the analytic focus and allows the analysts to determining that lung cancers are more likely to spread to bones than bone cancers spread to the lungs. Dynamic aggregation (i.e. undoing the grouping) is needed but certain combinations of data transformations may restrict the guarantee of reversibility)

7) Selecting sentinel events in a stream: In social media log analysis, such as the use of Twitter, sharpening the focus of analysis might require thoughtful selection of sentinel events. A typical strategy is to keep only the dates of the 1^{st}, 10^{th}, 100^{th}, and 1000^{th} tweets, which dramatically reduces the clutter of tweets. Then the relationship to retweets, mentions, replies, etc. becomes clearer. Similarly analysts might choose to retain only the dates of the 1^{st}, 10^{th}, 100^{th}, and 1000^{th} followers.

8) **Converting multiple point events into a single interval event:** When dealing with patient histories a major simplification is to convert multiple point events, such as 25 normal blood pressure readings over 12 months into a simpler more meaningful single interval event that shows normal blood pressure of the 12 months.

9) **Converting multiple interval events in a single longer interval:** The Pharmacovigilance project raised this issue for patients who received repeated 90-day prescriptions for the same medication. However, patients might refill the prescription early, which appeared as an overlap of two intervals, or delay their refill, which appeared as a gap between the intervals. Analysts simplified the patterns by merging intervals with overlaps of less than 15 days or gaps of less than10 days resulting in long intervals indicating a the drug "episode".

10) **Identifying hidden complex events:** In many application domains some high level event such as suspecting that the patient had a heart attack or planning for surgery may consist of 20-100 events that all happened within a time period (certain blood tests, imaging etc.) These component events may not be relevant to the analysis, so all of them can be identified and replaced by a single event.

These pattern simplification strategies are either domain specific (e.g. event category aggregation can use domain ontologies) or goal driven (i.e. shaped by the specific question the analyst is trying to answer (e.g. how to merge events into large intervals is application and goal specific). Domain experts who visually inspect sample data, will be able to tune parameters and see the effect of the simplifications.

Multi-step strategies

To address the volume of records issue the extraction strategies can be implemented in traditional database systems and the extracted results converted so they can be then loaded in the visual analytic tools. Our experience indicates that extraction strategies continues to be useful inside the visual analytics tools as users eliminate errors and outliers or select groups of records of interest.

To address the variety of patterns issue, visual inspection of the patterns or the result of the transformations greatly improves the sharpening of focus so it needs to be conducted within the visual analytics tool, possibly on a sample of records as first. For example the initial analysis of a random sample or the last month of data is usually extremely useful to identify data cleaning strategies (e.g. removing all records deactivated in the last month), or new extraction strategies (e.g. identifying event categories of interest) or devising meaningful pattern simplification strategies (e.g. selecting the 1st and 5th abnormal test). In our experience the pattern simplification strategies usually reduce data volume. We have implemented most of the strategies described above in EventFlow, but if the goal is to reduce the volume of records so that all needed records can be loaded in the visual analytics tool, those transformations will have to be executed in the source database or a separate data focusing tool.

Conclusion

There are certainly other temporal analytic focusing strategies beyond extraction, folding or simplification. Some will be generalizable, while others will remain specific to event sequence analysis. Some of those strategies might become an integral part of established domain specific workflow (e.g. for the analysis of drug usage patterns) or task specific (exploratory analysis of the patterns leading to a fixed outcome). Understanding which strategies are relevant in each situation requires experience with the data and problem at hand. We believe that these strategies can be combined to sharpen analytics processes and enable users to deal with ever larger datasets. In time the accumulated experience of analysts with particular application domain will lead to recommended or partially automated workflows.

References

Fisher, D., Popov, I., Drucker, S. and schraefel, mc, Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster., *Proc. CHI 2012*, ACM Press, New York (May 2012), 1673-1682.

Gotz, D. and Stavropoulos, H., DecisionFlow: Visual analytics for high-dimensional temporal event sequence data, *IEEE Trans. on Visualization and Computer Graphics 20*, 12 (December 2014), 1783-1792.

Gschwandtner, T, Gärtner, J, Aigner, W, Miksch, S, A taxonomy of dirty time-oriented data, In Multidisciplinary Research and Practice for Information Systems, *Lecture Notes in Computer Science*, 7465 (2012) 58-72

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Ham, F., Riche, H., Weaver, C., Lee, B., Brodbeck, D., Buono, P. , Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data, *Information Visualization*, 10, 4 (2011) 271-288

Monroe, M., Interactive Event Sequence Query and Transformation, Ph.D. Dissertation University of Maryland Department of Computer Science (June 2014). Available at: http://www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2014-17

NASA Technical Reports http://omniweb.gsfc.nasa.gov/html/omni2_doc_old.html
https://earthdata.nasa.gov/data/data-tools
http://swift.gsfc.nasa.gov/quicklook/swift_process_overview.html
http://heasarc.gsfc.nasa.gov/docs/suzaku/analysis/abc/abc.html
(Accessed November 15, 2014).

National Health Service (UK), Health & Social Care Information Centre, Technical Reports http://www.hscic.gov.uk/article/1825/The-processing-cycle-and-HES-data-quality (Accessed November 14, 2014).

Open Refine, http://openrefine.org/ (Accessed November 15, 2014).

Perer, A. and Shneiderman, B., Systematic yet flexible discovery: Guiding domain experts during exploratory data analysis, *Proc. ACM Conference on Intelligent User Interfaces*, ACM, New York (January 2008), 109-118.

Rind, A., Wang, T., Aigner, W., Miksch, S., Wongsuphasawat, K., Plaisant, C., Shneiderman, B. Interactive Information Visualization for Exploring and Querying Electronic Health Records: A Systematic Review, *Foundations and Trends in Human-Computer Interaction*, Vol. 5, No. 3 (2013) 207-298.

Wang, T. D., Wongsuphasawat, K., Plaisant, C., and Shneiderman, B., Extracting insights from Electronic Health Records: Case studies, a visual analytics process model, and design recommendations, *Journal of Medical Systems* 35, 5 (2011), 1135-1152.